YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS
MASTER DEGREE PROGRAM OF ECONOMICS

MASTER DEGREE THESIS

FORECASTING ECONOMIC VARIABLES USING GOOGLE TRENDS

AHMED OMIC
15729018

THESIS ADVISOR
Prof. Dr. HÜSEYİN TAŞTAN

ISTANBUL
2018

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS
MASTER DEGREE PROGRAM OF ECONOMICS

MASTER DEGREE THESIS

FORECASTING ECONOMIC VARIABLES USING GOOGLE TRENDS

AHMED OMIC
15729018

THESIS ADVISOR
Prof. Dr. HÜSEYİN TAŞTAN

ISTANBUL
2018

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS
MASTER DEGREE PROGRAM OF ECONOMICS

MASTER DEGREE THESIS

FORECASTING ECONOMIC VARIABLES USING GOOGLE TRENDS

AHMED OMIC
15729018

Date of Submission: September 19th, 2018
Date of Defense: October 26th, 2018
Thesis was successfully defended through consensus vote

	Name and Surname	Signature
Thesis Advisor	: Prof. Dr. Hüseyin TAŞTAN	.....
Jury Members	: Prof. Dr. Mübariz HASANOV	.....
	Dr. Hasan A. KARADUMAN	.....

ISTANBUL
OCTOBER, 2018

ABSTRACT

FORECASTING ECONOMIC VARIABLES USING GOOGLE TRENDS
Ahmed Omic
October, 2018

This thesis examines the ability of Google Trends (GT), i.e. the free public tool for obtaining data regarding web search activities, in forecasting economic variables of Turkey by using different query selection methods. It reports whether internet search activity improves forecasting of tourism demand, unemployment, and car sales by comparing forecast errors of models with and without GT variable. For each variable of interest, three models are constructed. The first, baseline model, is estimated using only the past values of selected variables. The second model, in addition to the past values of the economic variables, integrates additional regressor constructed from the simple query selection method, where search indices of single keywords related to each of the selected variables are obtained. The third model follows the same approach as the second, but with different query selection method, where composite search index using Principal Component Analysis (PCA) is constructed from the large number of queries related to our economic variables. All models are estimated using the Autoregressive Integrated Moving Average (ARIMA) methodology. Forecast comparisons indicate that models with GT information outperformed the basline models in most of the forecasting experiments. When it comes to performance of the two query selection methods, composite search index, on average, provides better forecasts. The study's results could be beneficial for the policy makers and other stakeholders as selected variables play important role for the Turkish economy. GT offers a new way of tracking economic behavior at almost zero cost. Furthermore, they have ability to get real-time insights regarding economic decisions.

Key Words: Forecasting, ARIMA, Google Trends

ÖZ

GOOGLE TRENDS İLE İKTİSADİ DEĞİŞKENLERİN ÖNGÖRÜLMESİ

Ahmed Omic
Ekim, 2018

Bu tez, web arama aktiviteleri ile ilgili bilgi sunan Google Trends (GT) aracının, farklı sorgu seçim yöntemlerini kullanarak Türkiye'deki ekonomik değişkenleri tahmin etmedeki yeteneğini incelemektedir. GT değişkeni olan ve olmayan modellerin tahmin hatalarını karşılaştırarak internet arama etkinliğinin turizm talebi, işsizlik, ve araba satışlarının öngörü başarısını geliştirip geliştirmediğini ortaya koymayı amaçlamaktadır. Her bir değişken için üç model oluşturulmuştur. Baz model, yalnızca seçilen değişkenlerin geçmiş değerlerini kullanarak öngörü oluşturmaktadır. İkinci model, ekonomik değişkenlerin geçmiş değerlerine ek olarak, seçilen değişkenlerin her biri ile ilgili tek anahtar kelimeye ilişkin arama indekslerinin elde edildiği basit sorgu seçim yönteminden oluşturulan ilave değişkeni içermektedir. Üçüncü model, ikinciyle aynı yaklaşımı izlemekte, ancak arama indeksini oluştururken çok sayıda anahtar kelime kullanmakta ve Temel Bileşenler Yöntemi (Principal Component Analysis, PCA) yöntemini kullanarak yeni bir bileşik indeks oluşturmaktadır. Tüm modeller Autoregressive Integrated Moving Average (ARIMA) metodolojisi kullanılarak tahmin edilmiştir. Öngörü karşılaştırmaları, her iki sorgu seçim yöntemini takiben GT değişkenini içeren modellerin, baz modele göre daha başarılı olduğunu ortaya çıkarmaktadır. İki sorgu seçim yönteminin performansları söz konusu olduğunda ise ortalama olarak bileşik arama indeksinin daha iyi sonuçlar sağladığı görülmektedir. Araştırmanın sonuçları, politika yapıcılar ve diğer paydaşlar için çok yararlı olabilir çünkü seçilmiş değişkenler Türkiye ekonomisi için önemli bir rol oynamaktadır. Bu çalışma onlara neredeyse sıfır maliyetle ekonomik hareketleri takip etmenin yeni bir yolunu sunmaktadır. Ayrıca, ekonomik kararlarla ilgili gerçek zamanlı bilgiler edinme olanağı sağlamaktadır.

Anahtar Kelimeler: Öngörü**,** ARIMA, Google Trends

ACKNOWLEDGEMENT

To my beloved mother Samira and brother Muhamed. I love you!

Istanbul; October, 2018 Ahmed Omic

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AIC : Akaike Information Criterion
AR : Autoregressive
ARIMA : Autoregressive Integrated Moving Average
AR-MIDAS : Autoregressive Mixed-Data Sampling
BIST : Istanbul Stock Market Index
GDFM : General Dynamic Factor Model
GDP : Gross Domestic Product
GPL : General Purpose Loan
GT : Google Trends
ISPAT : Investment Promotion Agency of Turkey
KPSS : Kwiatkowski-Phillips-Schmidt-Shin test
MA : Moving Average
MAE : Mean Absolute Error
OCSB : Osborn-Chui-Smith-Birchenhall test
PCA : Principal Component Analysis
RMSE : Root Mean Square Errors
SARIMA : Seasonal Autoregressive Integrated Moving Average
SVM : Support Vector Machine Regression
USD : United States Dollar
VAR : Vector Autoregression

1. INTRODUCTION

Human beings have always been striving towards development and innovation. Throughout the human history discoveries such as fire, wheel, steam train, electricity, motor vehicle, phone, etc., had an enormous impact on human life. In the 21st century, advances in technologies gave birth to the invention of internet, smartphones, and smart sensors, which are used daily by individuals and companies. These inventions have changed the way we perform our everyday activities. Nowadays, most of our activities are performed online, i.e. communication is done via emails and various social platforms, formal and informal education via various online courses and knowledge platforms, daily news are obtained from online press editions, solutions for our everyday problems are found with the help of search engines, job applications, purchase of various goods and services, etc.

The rising usage of these technologies provides a massive amount of information known as “big data”, which can be defined as “linkable information with large data volume and complex data structure” (Khoury and Ioannidis, 2014). Due to its massive size and complex structure there are several challenges with analyzing, storage, visualization, and information privacy of “big data”. This phenomenon has drawn attention of a business world and academic community in finding ways of the utilization of such a large amount of information. It offers new opportunities to our society as “these vast new repositories of information can provide researchers, managers, and policymakers with the data-driven evidence needed to make decisions on the basis of numbers and analysis rather than anecdotes, guesswork, intuition, or past experience (Song and Liu, 2017)”. In general, one of the most significant applications of “big data” is its ability to predict human behavior and improve forecasts. Daily, millions of people around the globe leave traces of their attitudes, preferences, wants, and needs. This opens an immense possibility for researchers to develop new predictive methods, discover new patterns of behavior, nowcast the present and forecast the future more precisely (Blazquez et al., 2018).

Especially, the application of forecasting is valuable in socio-economic policy and research (Einav and Levin, 2014). On the one hand, socio-economic variables are often hard to model and forecast as they depend on a complex modulus of human behavior. On the other hand, there is the issue of delivering statistics of economic variables with substantial lag. This happens as traditional statistics relies on a highly time-consuming process that gathers data from government statistical agencies, financial reports, and similar documents. While making decisions, policymakers highly rely on the official statistics of the economic indicators, i.e. GDP growth, unemployment rate, etc., and their forecasts.

One type of big data, web search engine data, attracted significant attention of the academicians in the attempt to predict economic variables. The wide spread usage of a search engines enabled the availability of real time information of billions of economic decisions made by people all around the globe. Every time we search something on the internet, we reveal important information regarding our future moves. For example, search of hotels availability in certain country or a car brand indicates the intention of individual to visit that country or buy a car in the near future. By observing millions of queries "researchers can significantly improve the accuracy, granularity, and timeliness of predictions about future economic activities (Wu et al., 2015)". One of the reasons why this type of "big data" gained popularity among researchers is that it can be obtained at nearly zero cost.

Google launched Google Trends (GT) tool as a free public service, where everyone can observe volume index of searches for a specific keyword. The usage of GT for forecasting was first introduced by Choi and Varian (2009). They published the study titled "Predicting the present with Google Trends" that later has become the most widely cited work in this field. They proposed the framework of employing GT for nowcasting various economic variables such as retail sales, tourism demand, automotive and home sales. The main motivation was to "familiarize readers with GT data, illustrate some simple forecasting methods that use this data, and encourage readers to undertake their own analyses (Choi and Varian, 2012)." Indeed, they succeeded in their aim as many academicians, inspired by this work, employed GT for prediction of various variables.

Consequently, the purpose of this thesis is to examine the success of GT in the forecasting selected economic variables of Turkey. The focal research question of the thesis is:

Does Google Trends improve forecasting of tourism demand, unemployment, and car sales in the case of Turkey?

For providing an answer to the research question, the thesis compares forecast errors of models with and without GT variables. For each variable of interest three models will be estimated. The first, baseline model, uses only the past values of selected variables for producing forecasts. For other two models, GT variables, obtained from the two different query selection methods, are added in addition to the past values of selected economic variables.

As one of the biggest challenges of previous studies is query selection process, this study uses two different methods and compares their performances. The first method follows a simple approach, where search indices of single keywords related to each of our economic variables of interest are obtained. The second method uses the Principal Component Analysis (PCA), where a composite search index is constructed from the large number of queries related to our economic variables. The queries are obtained in the following steps. First, initial queries for each of the selected variable are determined using previous studies as reference (unemployment and car sales) or supported by economic theory related to the selected variable (tourism demand). Second, for each selected initial query their related queries are obtained by using GT option to retrieve related queries of each searched keyword. In the final step, for each variable, composite search index is constructed from all obtained queries.

For model estimations, Autoregressive Integrated Moving Average (ARIMA) methodology is used. All models are estimated using *"auto.arima"* function in R (Hyndman and Khandakar, 2008). The algorithm selects the number of differences using Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, seasonal differences using Osborn-Chui-Smith-Birchenhall (OCSB) seasonal unit root test, and autoregressive (AR) and moving average (MA) components by minimizing AIC.

Turkey is one of the fastest growing economies in the world and accurate forecasts of economic variables are important for future decision making and development. Especially, this stands for selected variables. For example, traditionaly tourism sector plays important role for the Turkish economy. According to the Association of Turkish Travel Agencies (TÜRSAB), Turkey should expect 40 million foreign visitors and income of over $30 billion by the end of 2018 (Hurriyet Daily News, [14.05.2018]). When it comes to automotive industry, Turkey is one of the largest cars manufacurers in the world and its automobile penetration level of 165 is well behined the European average of 500 cars per 1,000 people (Invest in Turkey, [17.05.2018]). In the end, unemployment is important variable for every country as it is the major indicator of labor market performance and provides valuable insights regarding the health condition of an economy.

To the best of our knowledge, there is no study in the literature using GT and Turkey for forecasting selected variables, except unemployment. As a matter of fact, there are just three works employing GT and Turkey as a case study. In the first study, GT data is used for nowcasting monthly nonagricultural unemployment rate of Turkey (Chadwick and Sengul, 2015). They used linear regression models and Bayesian Model Averaging for model estimation. The results showed that GT is successful in nowcasting monthly nonagricultural unemployment rate for both in-sample and out-of-sample forecasts. In the second paper, GT is used for nowcasting the credit demand, i.e. national general-purpose loan (GPL) (Zeybek and Ugurlu, 2015). The results confirmed paper's assumption regarding GT ability to nowcast GLP demand. In the last example, GT is used for forecasting of the Istanbul Stock Market Index (BIST) (Bilgic, 2017). In this paper, GT data is used for determining its effect on the BIST volume data and absolute return. The results showed significant effect of the selected keywords in the estimation of transaction volume and absolute return. In addition, this is one of the first studies that compares the performances of different query selection methods.

On the one hand, this means that this thesis may help in the providing more comprehensive answer on whether internet activity in Turkey can be used for the prediction of selected variables. On the other hand, the relevance of work is justified by addressing social and economic indicators for which data may not be distributed on time and that have substantial importance for the economic performance of the

selected country. Additionally, this thesis will allow us to validate consistency of previous findings in this field.

The rest of the thesis is structured as follows. Section 2 presents importance of selected variables for Turkey and literature review of related works. Section 3 describes the data used in this study. Section 4 presents the methodology employed for testing our research question. Results and discussions are presented in the section 5. Finally, section 6 presents conclusions and proposes suggestions for future works.

2. LITERATURE REVIEW

2.1. Importance of selected variables for Turkey

In this section we will briefly explain the reasoning behind selecting unemployment, tourism demand, and car sales for forecasting purposes. There are two main reasons why we selected previously named variables. First, there is high probability that we could find forecasting capabilities from search queries related to these variables. For example, people regularly search for job opportunity via internet which indicates unemployment fluctuations or there is high probability that travelers will search tourism related information regarding the destination they intend to visit. The same reasoning goes with the car sales. People tend to search brand or models of cars before they conduct the purchase.

As a matter of fact, these variables are valuable for the overall performance of Turkish economy, which is the second reason for their selection. Unemployment rate, the share of the labor force without a job but actively seeking work, is important variable for every country as it is a major indicator of labor market performance and provides valuable insights about the health condition of an economy. Unemployment has a cyclical nature; it increases during the economic slowdown and decreases when economy is performing well. This variable is often good confirmation of what other indicators are showing as it is always a lagging indicator. It typically reaches its peak after the recession because companies refuse to provide employment until they are sure economy will perform well. Being able to forecast unemployment is important so that policymakers can set proper monetary policy and other measures that will ensure well performance of economy. Especially, this stands for emerging economies, such as Turkey, as they are subdued to fast shift of economic performance.

Tourism sector traditionally plays an important role for the Turkish economy. With its rich history, favorable climate, location, and developed tourism infrastructure, Turkey is a real magnet for tourists all over the world. According to the most recent

statistics, more than 38.6 million tourists visited Turkey in 2017, which makes this country the tenth most popular tourist destination in the world (Wikipedia, [23.05.2018]). According to the Investment Promotion agency of Turkey (ISPAT), Turkey has set an annual target of 50 million tourist arrivals and revenues of 50 billion USD by 2023 (Invest in Turkey, [24.05.2018]). In reaching these targets important role plays ability to precisely forecast tourism demand. It is very important for the companies like airlines, tour operators, and hotels to be able to know demand for their products and services. In many cases, market failure occurs as management is not able to meet the market demand. An accurate forecast of demand for the products will help business to timily manage their supplies and avoid shortages or surpluses, plan cash expenditures and revenues. It also improves control of internal operation within the businesses as based on anticipated future demand decisions regarding hiring new staff, marketing, and expansion can be made (Johnston, [24.05.2018]).

Automotive is another sector that plays a significant role for the Turkish economy. With the history dating back to 1960, Turkey is one of the biggest manufacturers of motor vehicles. Although most of the production goes for export, domestic car sales have been increasing over years. For example, the average annual sales in the 2000s accounted for around 360,000, while in the 2016 they reached 1,000,000 vehicles. This offers a big opportunity for carmakers in the domestic market. Especially, this statement holds true if we take into consideration that Turkey's automobile penetration level of 165 is well behind the European average of 500 cars per 1,000 people (Invest in Turkey, [17.05.2018]). Consequently, forecasting car sales will be highly beneficial for all stakeholders in the automotive sector.

2.2. Theory behind consumer search behavior

Before making purchase, customers tend to conduct research regarding product's characteristics. This behavior is driven by customer's need to maximize satisfaction and decrease uncertainty during the buying decision process.

The information search behavior is highly affected by the prior knowledge of customers regarding the products of interest and type of products being searched. In general, the more knowledge customer has about product the better he is in finding necessary information and making purchasing decision (Raju al., 1995). However,

when the level of knowledge is high the customer searching activities decreases as there is no need for additional search activities. Furthermore, customers tend to perform more comprehensive information search for the products with higher perceived level of risk. For example, products that are more expensive or require higher involvement, such as purchase of car or house, bring higher level of risk and customers tend to devote more time for information searching activities in order to make right decisions.

Nowadays, great proportion of this process is performed online, and it often leads to product purchasing (Shim et al., 2001). According to Klein (1998), low perceived cost of providing and assessing objective data is one of the main reasons why internet is useful for search purposes. Internet usage for product search and price comparison before purchase is common practice of millions customers across the world. Due to the complex nature of online search process it is hard to determine its pattern. However, the most recent findings suggest that customers on average search 2-3 brands while purchasing online, and they use multiple channels during this process, i.e. search engines, display advertising, price comparison websites and social media (Competition and Markets Authority (CMA), 2017).

As we can see, search statistics can reveal important insights about a customer's intention to make purchasing decisions, which offers the possibility of using them for forecasting purposes. One of the benchmark studies in this field was conducted by Moe (2003), who analyzed the motivation and outcomes of internet search activities. Later, Ettredge et al. (2005) investigated search behaviors of men and women for job search purposes. He pointed that search activities offer insights regarding people's needs and wants. As the importance of internet has been increasing, empirical studies in the field of search engine data also increased. The turning point was launching of GT service in 2008, which resulted in the substantial increase of empirical research. The following subsection discusses the reasons of GT popularity and presents pioneer works.

2.3. Google Trends and first applications

Digital economy, emerged as a result of rapid technology advances, provides new non-traditional sources of socio and economic data. One type of this source is web search data. Every day individuals, companies, and other parties generate tons of information by searching and posting on internet. These online activities leave important traces of our needs and wants that can be utilized for social and economic forecasts.

By launching the GT tool in 2008, Google enabled us to get valuable insights regarding web search data of people worldwide. The tool provides ability to obtain data regarding search activities over different categories and regions. It does not provide the raw level of queries, but rather an index of the volume of search term by geographic location and category (Varian and Choi, 2012). The index is based on a query share where "the total query volume for search term in a given geographic region divided by the total number of queries in that region at a point in time (Varian and Choi, 2012)." The query share is than normalized in the range starting from 0 and ending with the maximum value of 100. Queries can be searched over 27 first level categories and 241 second level categories from 2004 up to present. If the time span is less than five years, data is provided at weekly frequency, otherwise, search queries are presented as monthly data. In addition, GT provides an option to obtain related queries for each searched keyword.

Despite valuable possibilities that GT data offers, there are several drawbacks. First, GT is an index and it does not provide raw search data. Second, as data is computed using a random sampling method (Varian and Choi, 2012), results for the same search keyword vary a few percent from day to day.[1] Third, only queries with the substantial amount of search volume, at least 50 observations, can be obtained.

GT tool has gained huge popularity among researchers in the attempt to predict social and economic behavior.

There are three main reasons for the popularity of GT:

1. **Free of charge:** Since its first introduction in 2008, GT is free of charge and available to everyone.

[1] More information regarding sampling method can be found here: https://support.google.com/trends/answer/4365533?hl=en .

2. **Easy to use:** Google created customer friendly platform, with a simple design and options for comparison over different countries and categories. In the Figure 1 you can see the design and structure of GT tool.
3. **Representative source of human and social behavior:** Google is the absolute leader among web search providers holding the 90% of the market share (Statcounter, [27.05.2018]). Only in the few countries, i.e. Russia, China, South Korea and Japan, it is not the dominant search engine due to the political and language barriers. Generating over 5.5 billion queries daily makes GT excellent platform for observing human search activities as it "offers instant reflection of the needs, wants, demands and interests of its users (Jun et al., 2017)"

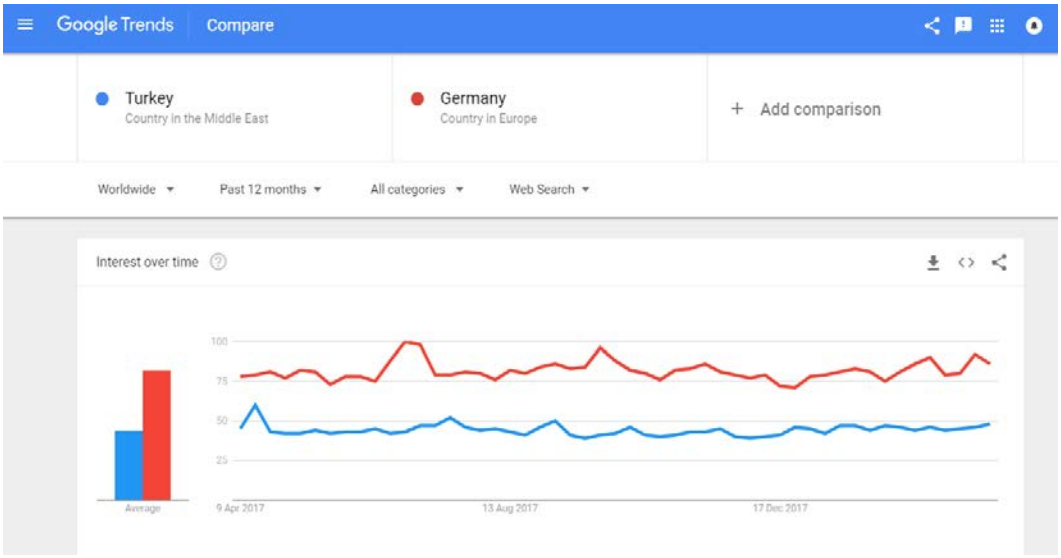


Figure 1: **Example of GT platform**

The first publications regarding the usefulness of GT came directly from Google itself. Ginsberg et al. (2009) first used search data to demonstrate that influenza can be traced with GT. This was the benchmark paper for the GT application in the field of medicine. Second paper came from Choi and Varian, both employees of Google at that time, who published the work "Predicting present with Google Trends", which later has become the most widely cited work among those that utilized GT in the field of economics. In their work, they proposed the use of publicly available web

search volume histories from GT together with time-series regression analysis to nowcast economic variables (Choi and Varian, 2012). Later, Varian and Scott proposed a more advanced methodology in their paper "Predicting the present with Bayesian structural time series". They created the system that provides information regarding which coefficients are more likely to be important (Scott and Varian, 2014). Inspired by these two initial studies, many researchers tested the usefulness of this tool in various fields. The following section reviews these studies with the special focus on the GT application in the field of unemployment, tourism, and car sales.

2.4. Google Trends and unemployment

Internet arrival completely changed the way people are finding job opportunities. Nowadays, numerous platforms exist that facilitate the matchmaking process among employers and employees. Their existence increased users search activities in the domain of labor market demand and supply. These trends have provided researchers insights that they could be used for the prediction of unemployment.

The first two works that utilized this possibility were by Askitas and Zimmerman (2009) and D'Amuri (2009). The first one examined unemployment data for Germany and found predictive power of GT for unemployment. Driven by the theory of consumer search behavior, they assumed that as unemployment occurs people tend to use Google for searching different keywords related to their unemployment status. They found strong correlation with unemployment rate and predictive power of four keywords, i.e. "*unemployment rate*", "*unemployment office or agency*", "*personnel consultant*", and "*most popular job search engines in Germany*". D'Amuri (2009) examined the unemployment data from Italy and found that search query of keyword "*job offers*" ("*offerte di lavoro*") improves out-of-sample forecasts of unemployment compared to previously used leading indicators of unemployment evolution, i.e. employment expectations surveys and the industrial production index.

Work by Vicente et al. (2015) explores the usage of GT in nowcasting unemployment rate of Spain, which experienced sharp decline in unemployment due to the economic crisis. They followed ARIMA methodology for model's estimations and found significant forecasting improvement of models with GT variable.

Additionally, this work pointed out that internet searches can successfully capture the economic shocks that Spain encountered during the economic crisis. Nacarroto et al. (2018) used GT variable for the prediction of youth unemployment in Italy. For model estimations ARIMA and VAR approaches were used. The results confirmed previous findings and indicated positive forecasting impact of search queries obtained from GT.

2.5. Google Trends and tourism demand

The importance of forecasting tourism demand was documented in various studies such as Artus (1972) and Wong and Song (2003). The first book that systematically presented modern econometric techniques for tourism demand analysis is the monograph by Song and Witt (2000). Tourism demand is extremely difficult to predict due to the high seasonality of tourism demand and reliance of variable to external shocks, complexity of tourist's decision to travel, and a large set of potential dependent variables (Douglas et al., 2001). However, due to the rise of internet and search engine popularity, many users are leaving important traces of their travel's preferences online, which can be used for tourism demand predictions.

Pan (2012) used search engine to forecast hotel room demand for city Charleston in South Carolina, USA. This research confirmed the significance of using GT in increasing forecasting accuracy. Bangwayo-Skeete (2015) compared GT variable forecasting performance using Autoregressive Mixed-Data Sampling (AR- MIDAS) and ARIMA approach. AR-MIDAS approach outperformed alternatives in most of the out-of-sample forecasts. Yang (2015) managed to improve forecasting accuracy of Chinese tourist volume by using search engines, Google and Badu. This work is significant as it presented the new query selection method, where composite search index is created using Principal Component Analysis (PCA) from many tourism related queries. PCA is the method that simplifies the complexity in high-dimensional data. It produces fewer dimensions, which represent summaries of data's trends and patterns.

Research by Zeynalov (2017) analyzed the performance of GT data for nowcasting tourism demand to Prague using MIDAS methodology. Results showed usefulness of GT variable for nowcasting purposes. When it comes to the most recent studies, Li et al. (2017) proposed an advanced framework for creating a composite search index

based on the general dynamic factor model (GDFM). GDFM is the extended version of general dynamic model effective in analyzing large number of variables and producing common components. Based on tourism data from Beijing, the proposed method outperformed alternative methods, i.e. traditional model using past tourist arrivals and index created by PCA. Önder (2018) compared forecasting performances between cities and countries using GT web and image indices. Forecasting performance in the case of Vienna, with web and image search indexes, produced the best results comparing to Belgium, Barcelona, and Austria. Lastly, Padhi (2017) used GT information processing method and theory of planned behavior of tourists to construct composite search index in the case of Kerala city in India. The output of models outperformed alternative model based on the ARIMA with exogenous variable (ARIMAX) approach

2.6. Google Trends and car sales

When it comes to GT application for forecasting car sales, Swallow and Labbe (2011) tested index of online interest for car purchases in Chile for nowcasting of the automobile sales. Despite the low rate of internet users in Chile, models with GT variable outperformed baseline model in in-sample and out-of-sample nowcasts. Fantazzini and Toktamysova (2015) proposed multivariate models with GT data for forecasting monthly car sales in Germany. They constructed models for 10 car brands in Germany for the period of thirteen years, 2011-2014. Models with GT outperformed the comparing models for most of the car brands. Figueiredo (2016) used similar approach, where he tested performance of GT data in nowcasting present levels of new vehicle sales in ten Canadian provinces. Results showed that internet search data is significantly helpful for nowcasting car sales in the case of more populous provinces. However, in the case of less populous provinces the improvements are not significant. The most recent work by Wijnhoven and Plant (2017) used interesting approach, where they compared sentiment analysis and GT data for car sales forecasting. Namely, using linear regression they analysed predictive power of more than 500,000 social media posts regarding 11 car models in Netherlands. Later, they compared the outcomes from sentiment analyses with predictive power of GT data for the same 11 car brands. According to the results, the data from social media posts had little predictive power comparing to the GT data.

2.7. Google Trends and other applications

Beside previously mentioned applications, i.e. tourism demand, unemployment, and car sales, GT data is widely used for prediction of other economic variables. For example, Toth and Hajdu (2012) showed that GT can be effective in the prediction of consumption in Hungary. Moreover, Google searches help in determining which of the two products will win in terms of the market share (Prakash et al., 2012). Relation between business performances and GT data is examined by Vaughan and Yang (2013), where significant correlation between these two variables is confirmed.

Usage of GT is also employed in the field of sociology, politics, and medicine field. Mavragani et al (2016) examined usage of GT in predicting the result of 2015 Greek Referendum. Interestingly, results showed strong predictive power of online search data in relation to referendum results. Vaughan and Frias (2014) confirmed the usage of GT in prediction of academic fame, while Stephens-Davidowitz (2014) used GT for analyzing cost of racial animus on a black candidate.

When it comes to medical field applications, several studies confirmed GT usability in monitoring diseases. Pelat et al. (2009) discovered considerable correlation among internet searches and incidence of three infectious diseases, while Seifter et al. (2010) found effectiveness of GT data in determining epidemiology of Lyme disease. GT data helped in the prediction of suicide rate in the fifty American states (Lester and Gunn, 2013).

3. DATA

3.1. Tourism demand

Monthly tourism demand data was obtained for the major source countries, i.e. Austria, Germany, France, Netherlands, and Poland, from the Republic of Turkey Ministry of Culture and Tourism's database (Turkey Ministry of Culture and Tourism, [30.5.2018]). Although the Russian Federation is the main source country of tourism arrivals to Turkey, we omitted it as Google is not the main search engine in this country.

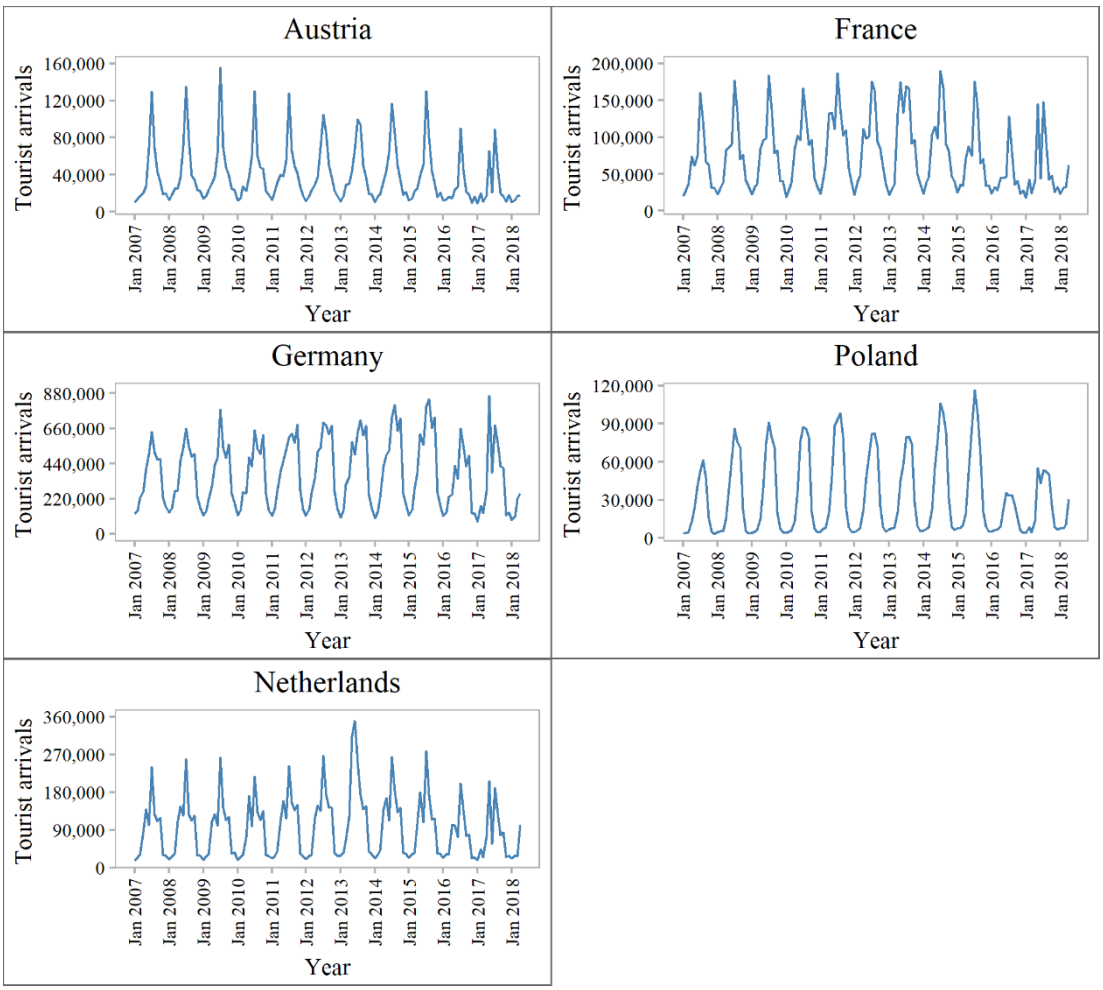


Figure 2: Tourism demand by countries

Data runs from January 2007 to April 2018, making a total of 136 observations per country. Countries are selected as they represent 30% of total tourism arrivals for the time 2007-2018. In addition, high percentage of population within selected countries is using internet and Google holds the highest share of market (Internet World Stats, [06.06.2018]). These features make these countries suitable for the research aim of this study. The Figure 2 shows the time series graphs of tourism arrivals to Turkey from mentioned source countries. Turkey experienced its peak in tourist arrivals in 2014, when around 42 million tourists visited the country, making it sixth most popular tourist destination in the world (Daily Sabah, [06.06.2018]). However, in the following years it experienced substantial decline in tourism arrivals due to the political instability and terrorist attacks. The graphs from selected source countries are showing this decline (Figure 2).

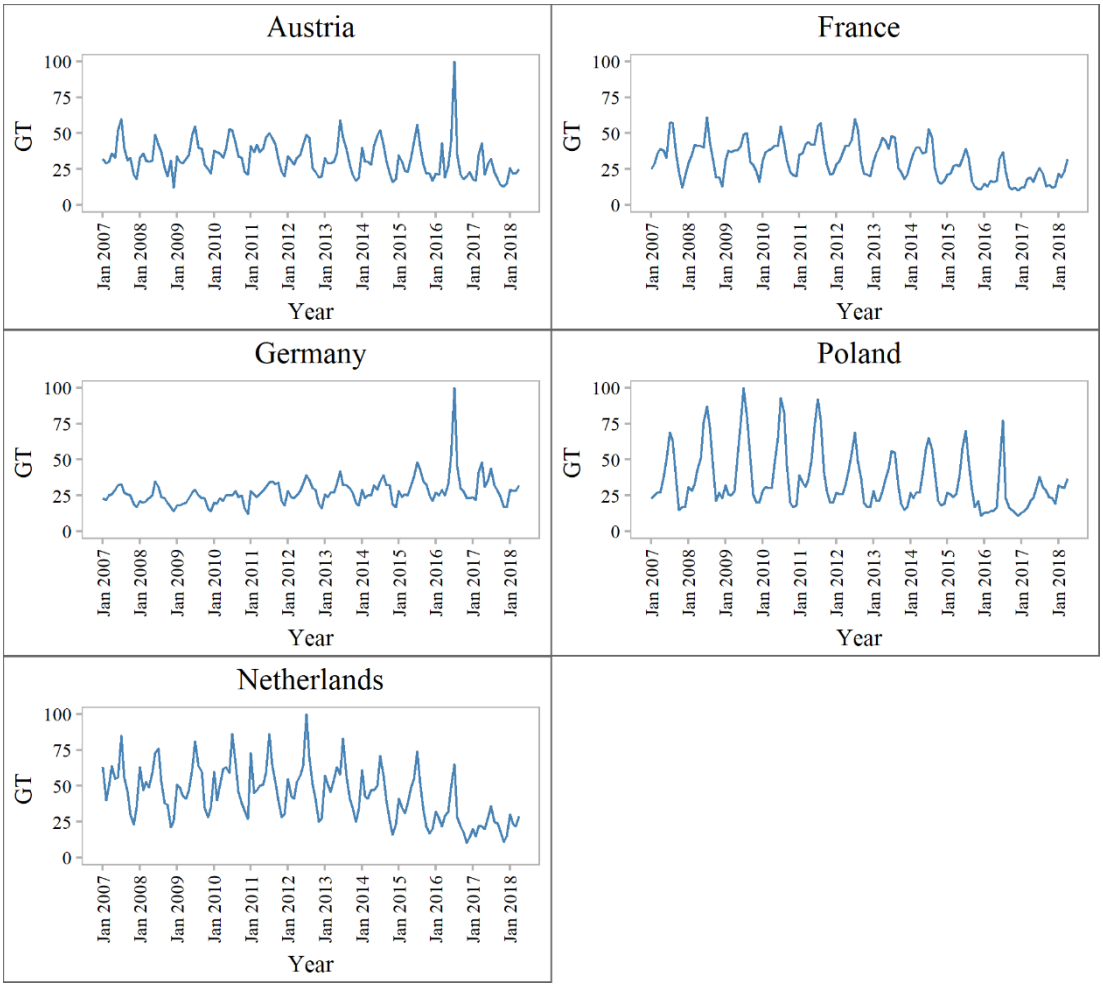


Figure 3: **GT data by countries for word "*Turkey*"**

For the Google searches, two query selection methods are used. First one is based on the simple ad hoc approach, where search index is obtained for the word *"Turkey"* under the travel and each source country's region section of GT tool from January 2007 to April 2018. Word *"Turkey"* is translated to each of the selected countries' languages, i.e *"Turquie"* for France, *"Türkei"* for Germany and Austria, *"Turkije"* for Netherlands and *"Turcja"* for Poland. Figure 3 shows obtained data.

Data from Figure 3 is showing similar pattern like in the Figure 2. Interestingly, Google searches from most of the countries managed to capture the decline in actual tourist arrivals to Turkey. However, this was not the case with the Germany and Austria as we can see large volume of searches for 2016, namely month of July. Possibly this happened due to the coup attack that occurred on 15 of July 2016. This event requires proper modelling, which is presented in the methodology section.

For the second query selection method, composite search index is constructed, using PCA, from large number of tourism related queries. During the query selection process, five major aspect of travelling are considered, i.e. country characteristics, cuisine, weather, transportation, and accommodation. These aspects are based on the customer journey theory in tourism, which has four stages (Figure 4).

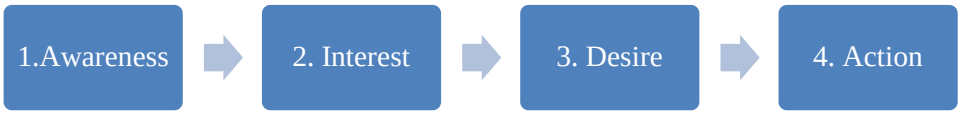


Figure 4: **Customer journey funnel (Lewis, 1903)**

In the first two stages, awareness and interest, Google could be used by customers for conducting information search regarding country characteristics, cuisine, and weather. Found information may create desire to travel and eventually action, which includes planning stage that envelopes information search regarding transportation and accommodation (Rödel, 2017).

Selection of queries for composite search index is conducted in the following steps:

1. For the five basic queries representing major aspect of travelling, i.e. *"Turkey"* and *"Turkey holidays"* for country characteristics, *"Turkey food"* and *"Turkey weather"* for cuisine and weather, *"Turkey airport"* for transportation, and *"Turkey hotel"* for accommodation, GT data under the travel and each source country's region section is obtained. The process is conducted for each source country using the queries translated in the countries's native languages.
2. After entering each of the five basic queries their related queries are retrieved for each source country.
3. For collected related queries duplicates and unrelated queries are removed and for each query GT data is obtained.
4. For each source country composite search index from all retrieved queries is obtained using Principal Component Analysis (PCA).

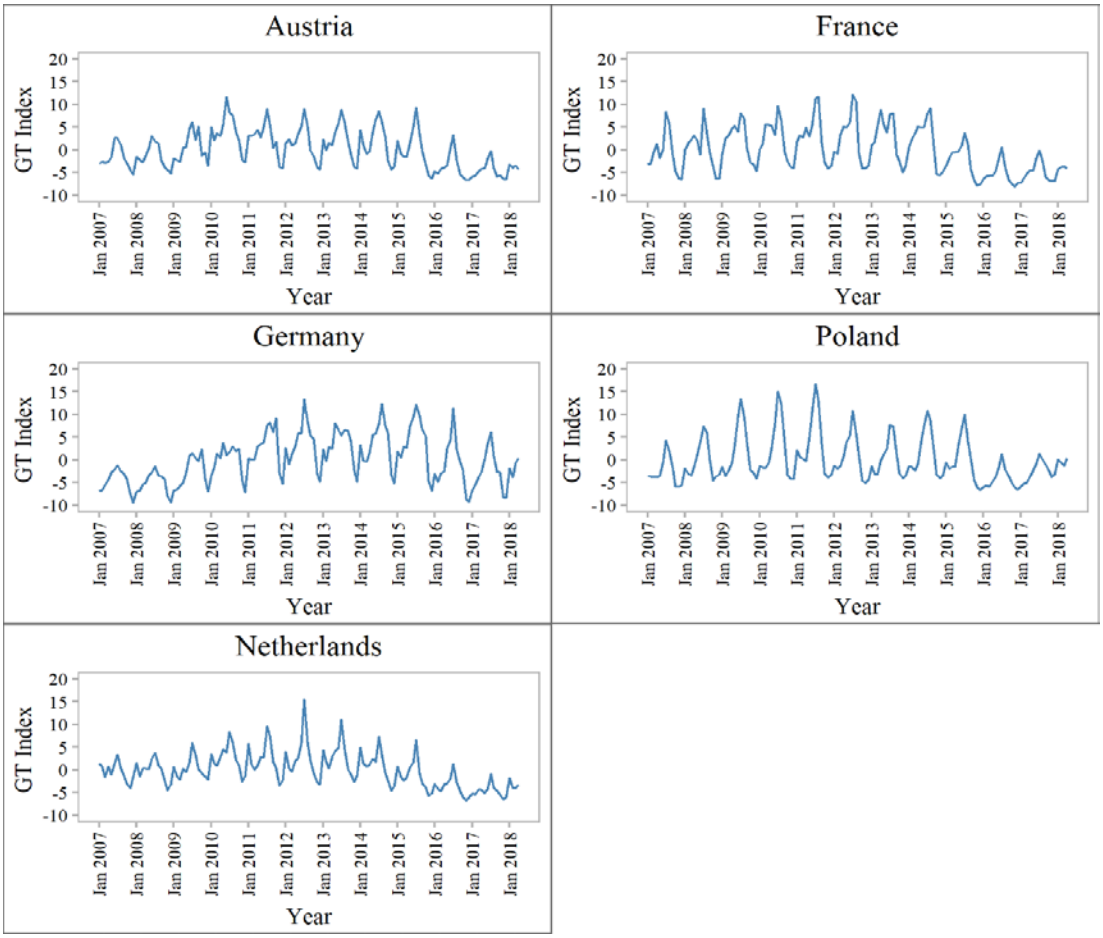


Figure 5: **Composite search index for tourism demand**

Following work of Li et al. (2017), poorly correlated queries with actual tourism arrivals are not removed as we wanted to prevent loss of information. For all queries monthly data from January 2007 to April 2018 is obtained. Obtained queries for each source country and PCA analysis results are provided in the Appendix I and II. The Figure 5 presents composite search index of each source country. Unlike the first query selection method, composite search indexes managed to capture decline in tourism arrivals for all source countries.

3.2. Unemployment

Total number of unemployed in Turkey was collected from Eurostat statistical database. Data from January 2007 to December 2017 is used, making a total of 132 observations. In addition to the total unemployment, we obtained data for youth unemployment in Turkey. The assumption is that young people are more prone to using internet and web searches could provide higher predictive power in the case of youth unemployment (Naccarato, 2018).

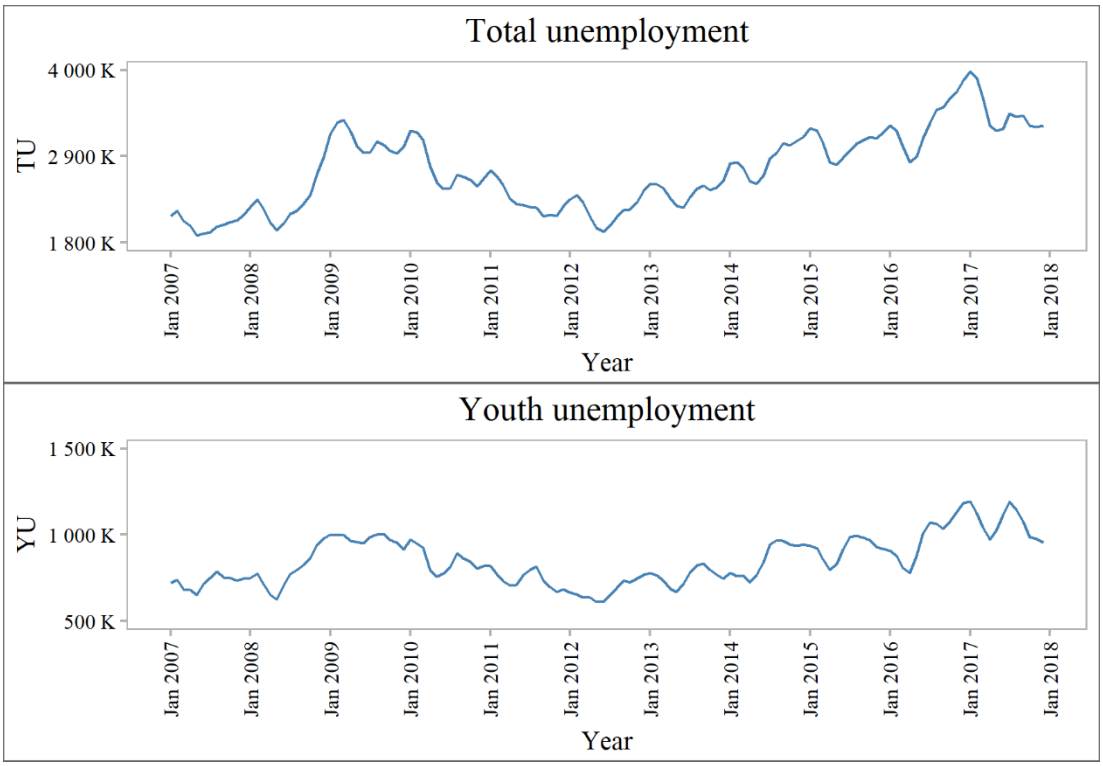


Figure 6: **Total and youth unemployment of Turkey**

The graph above (Figure 6) plots the total unemployment and youth unemployment in thousands over the selected period of analysis. Total unemployment in Turkey averaged 10.05 percent. After reaching its peak in February of 2009, 14.80 percent (3,331 in thousands), it started to decline towards a record low of 7.30 percent (1,935 in thousands) in June of 2012. Since 2012 unemployment had upward trends reaching 13.00 percent (3,985 in thousands) in January of 2017. On the other hand, youth unemployment in Turkey averaged 20 percent for selected period of analysis. The highest youth unemployment rate was detected in February of 2009 reaching 25.4 percent (998 in thousands) and the lowest of 13.9 percent (608 in thousands) in June of 2012.

Following the simple query selection method, GT data for word "*iş ilanları*", under Jobs & Education category and region Turkey, is obtained for selected period of analysis (Figure 7). Selected word, translated in English, means *"job announcements"*. We selected this word based on the assumption that unemployed people are usually conducting web search behavior regarding job announcements. Moreover, previous studies of Askitas and Zimmerman (2009) and D'Amuri (2009) followed similar approach.

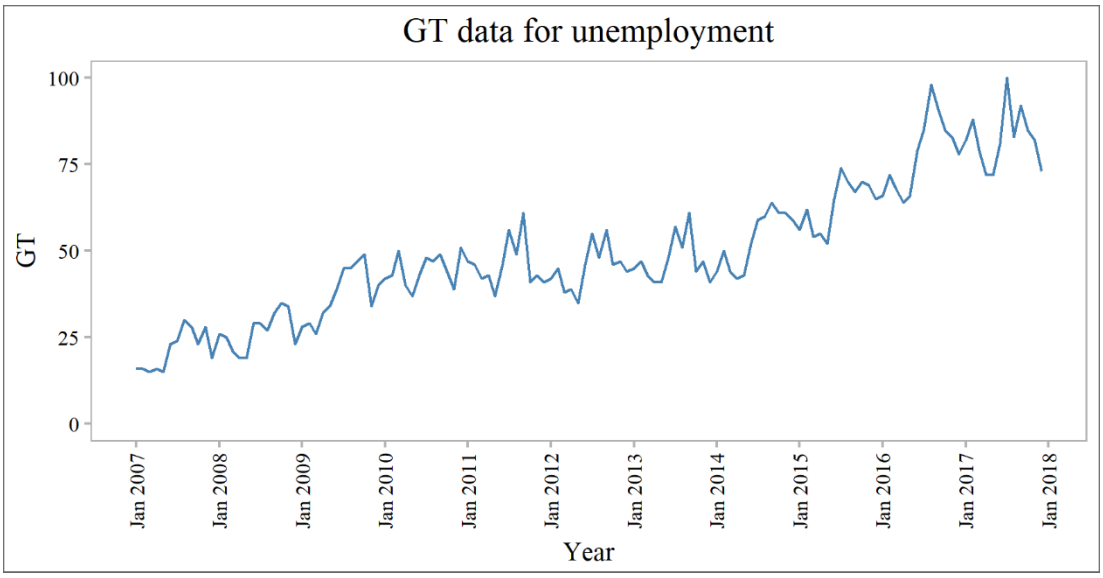


Figure 7: GT data for word "*iş ilanları*"

For the second query selection method following methodology is used (Figure 8):

1. Google data for keyword *"iş ilanları"* under Jobs & Education category and region Turkey is obtained for selected period of analysis.
2. Related queries of word *"iş ilanları"* are obtained.
3. For collected related queries duplicates and unrelated queries are removed and for each query GT data is obtained.
4. From all obtained queries composite search index is created using Principal Component analysis (PCA).

After removing duplicates and unemployment unrelated queries we had total of 25 retrieved queries from which composite search index is created (Appendix I).

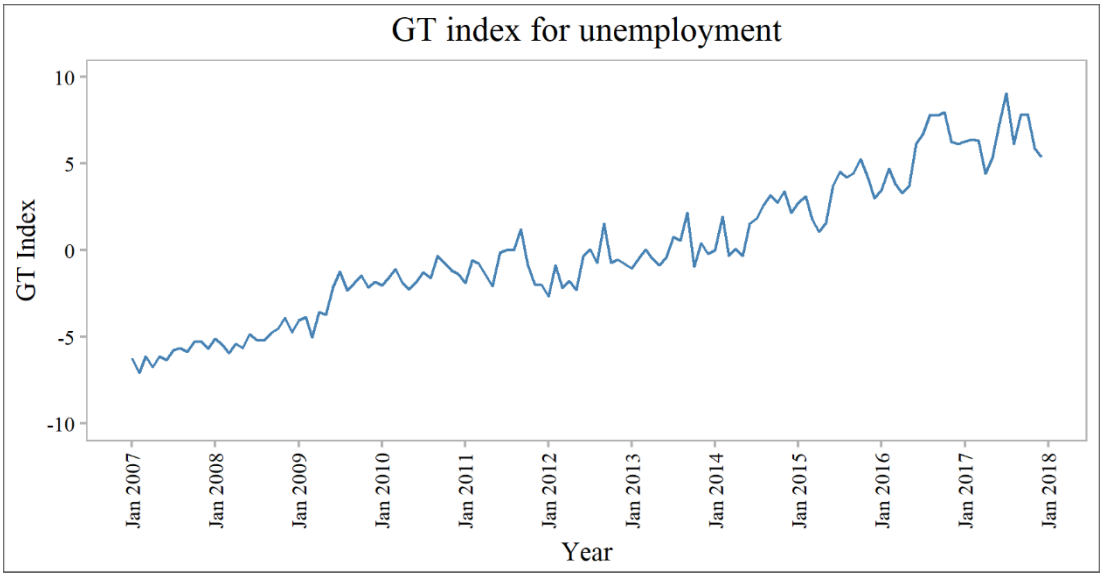


Figure 8: **Composite search index for unemployment**

3.3. Car sales

For car sales, we used number of registered cars as proxy variable for actual volume of car sales. The similar approach is followed in the works by Von Graeventiz, et al. (2016) and Fantazzini and Toktamysova (2015). Monthly data for the four major car brands in Turkey, i.e. Renault, Fiat, Opel, and Hyundai, is used (Statista, [08.06.2018]). Data was obtained from the Turkish Statistical Institute for the period starting from January 2010 to December 2017. In the below graph (Figure 9) the time

series plots of each car brand are presented. For the selected period of analysis, Renault has 901,275 registered cars, which is the highest number comparing to other selected brands. Fiat is the second with 481,975 registered cars, while Hyundai is the third with 480,271 registered cars. The least number of registered cars has Opel, 452,411.

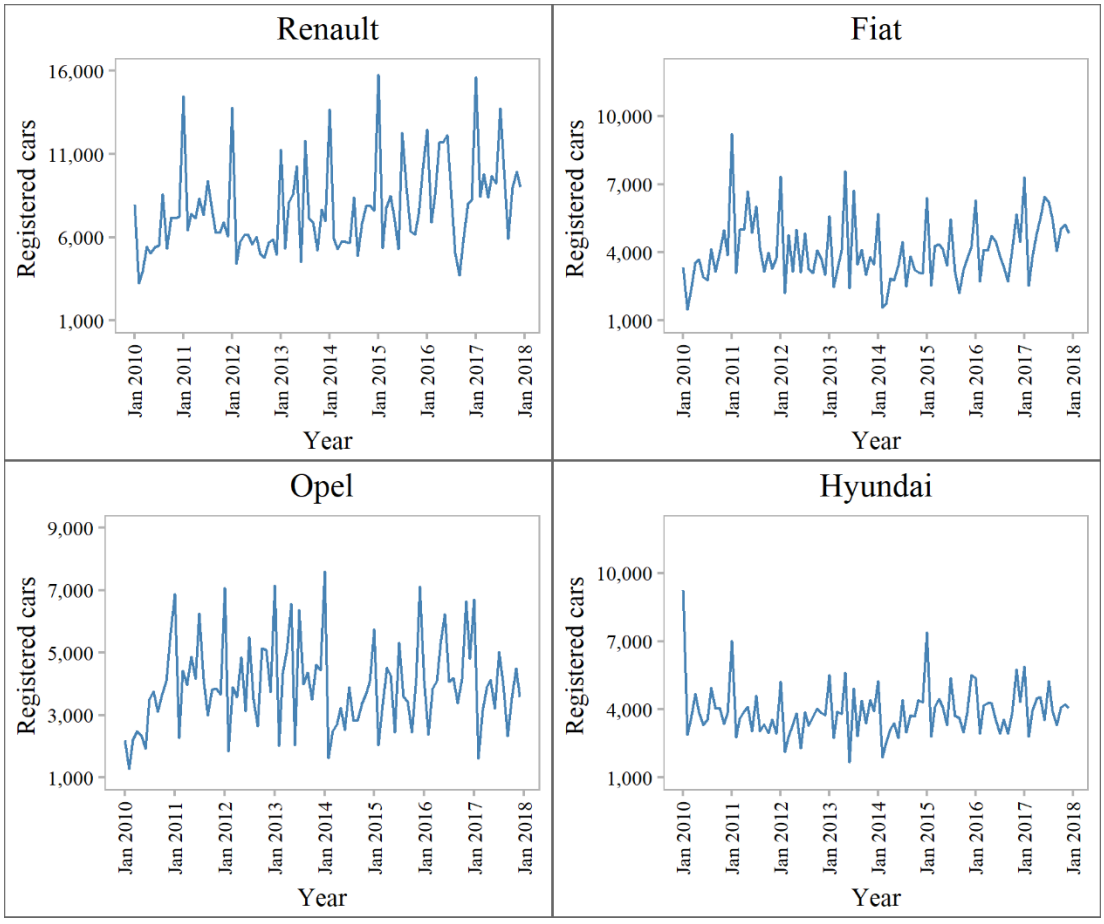


Figure 9: **Number of selected registered car brands in Turkey**

In the case of care sales, GT data is obtained in the following way. For the first query selection method, search index is obtained for the words: "*satılık renault", "satılık fiat", "satılık opel",* and *"satılık hyundai"* under Autos and Vehicles category and region Turkey of GT tool (Figure 10). Data is obtained from January of 2010 to December of 2017. English translation of the word *"satılık"* means "*for sale"* and indicates customer intention to buy, in this case, selected car brands.

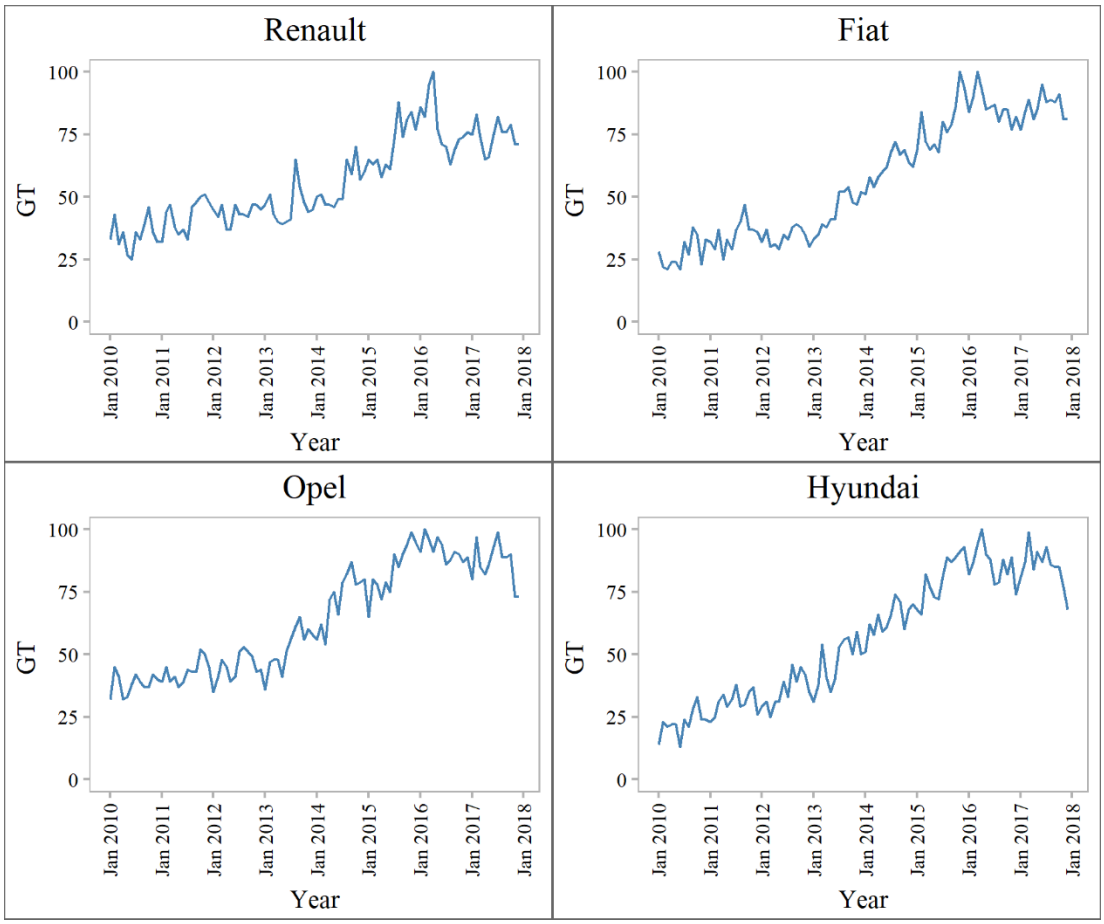


Figure 10: **GT data for the words: *"satılık renault", "satılık fiat", "satılık opel" and "satılık hyundai"***

For the second query selection method following methodology is used:

1. Google data for the initial keywords, i.e. *"satılık renault"*, *"satılık fiat"*, *"satılık opel"* and *"satılık hyundai"*, under Autos and Vehicles category and region Turkey is obtained for the selected period of analysis.
2. Related queries of initial keywords are obtained.
3. For collected related queries duplicates and unrelated queries are removed and for each query GT data is obtained.
4. From all obtained queries composite search index is created for each car brand using Principal Component analysis (PCA).

Following Figure 11 presents results of the second query selection method for car sales.

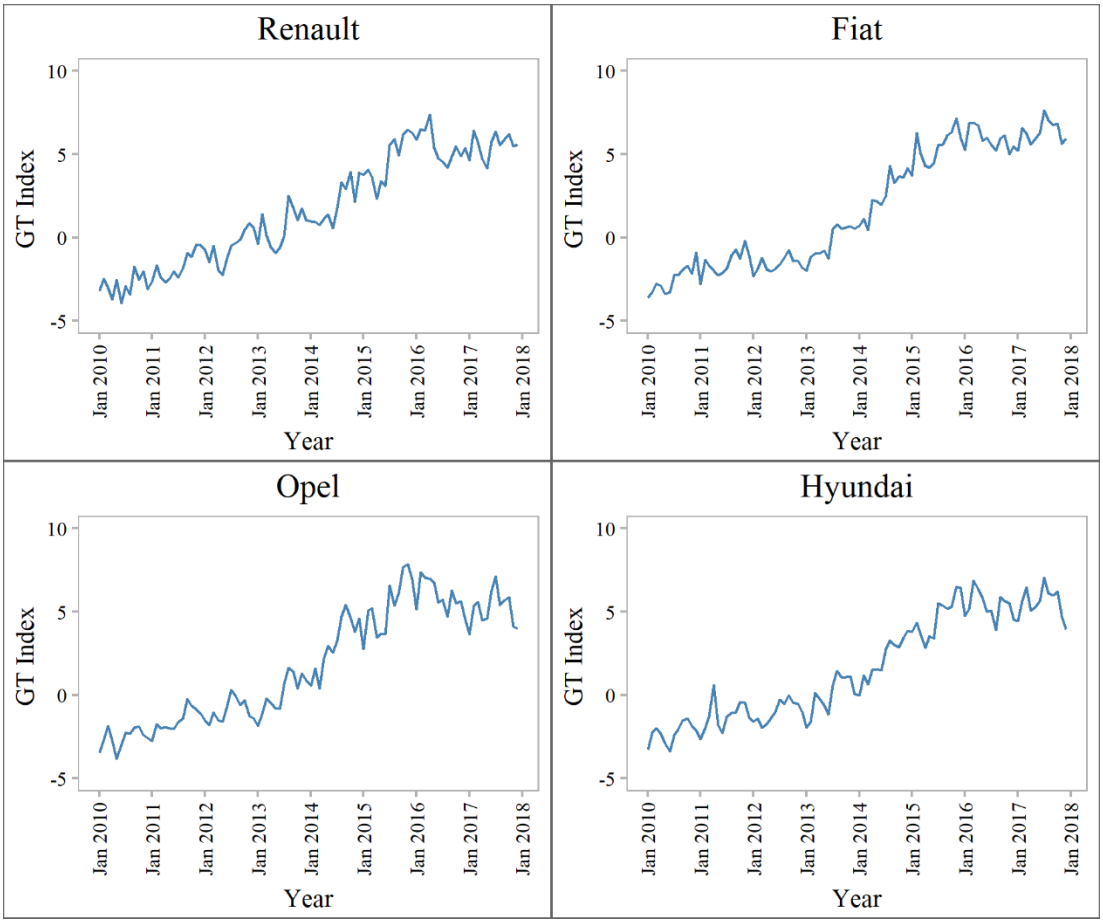


Figure 11: **Composite search index for car sales**

4. METHODOLOGY

In this section, the distinct research steps of this empirical investigation are presented. As we are dealing with time series data and our main goal is to produce forecasts, we need to employ time series analysis that looks for historical patterns in a data, i.e. trends, cycles and seasonal fluctuations, and based on findings produces forecasts. Main assumption under which time series analysis can be used is that patterns identified in the past will also occur in the future (Makridakis, Wheelwright, & Hyndman, 2008). In the past four decades, times series models have been widely used for forecasting economic variables. For this study, Autoregressive Integrated Moving Average (ARIMA) and regression with ARIMA errors are used for the forecasting purpose. In addition, "*auto.arima"* command of R's *"forecast"* statistical package is used for determining orders of ARIMA models. In the following subsections, detailed description of these methods and the framework of determining the effectiveness of GT data are presented.

4.1. ARIMA

The Autoregressive Integrated Moving Average (ARIMA) model is first proposed by Box and Jenkins (1970). Since then it has enjoyed great popularity among academic and business community. It is the most popular linear model for short run forecasts. It minimizes the errors in simulating the past by searching the combination of two forecasting models, i.e. autoregression (AR) and moving average (MA), and their lag parameters p and q (Song et al., 2008). The I in ARIMA, labeled by d, represents the order of integration that indicates number of times the variable is differenced. The process of differencing is necessary for making a non-stationary variable stationary, which implies that the mean, autocovariance, and the variance of a time series are not changing over time.

Using backshift notation, i.e. $By_t = Y_{t-1}$, a general ARIMA model of order (p, d, q) can be written as:

$$\left(1-\phi_1 B \ldots-\phi_p B^p\right)(1-B)^d y_t = c + \left(1+\theta_1 B+\cdots+\theta_q B^q\right)e_t \quad (1)$$

In the equation (1), AR (p) part is denoted by $(1-\phi_1 B \ldots-\phi_p B^p)$, MA (q) by $(1+\theta_1 B+\cdots+\theta_q B^q)$, and I (d) by $(1-B)^d y_t$, while c and e_t represent constant and white noise error term respectively.

If time series exhibits seasonal variation, then ARIMA process is called seasonal ARIMA (SARIMA) with additional seasonal autoregressive notation *P*, seasonal moving average notation *Q*, seasonal order of integration or seasonal differencing *D*, and notation *m* that represents length of seasonal period. A general SARIMA model of order (*p, d, q*) (*P, D, Q*) $_m$ can be written as:

$$\left(1-\phi_1 B \ldots-\phi_p B^p\right)(1-\Phi_1 B^m \ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D y_t = c + \left(1+\theta_1 B+\cdots+\theta_q B^q\right)\left(1+\Theta_1 B^m+\cdots+\Theta_Q B^m\right)e_t \quad (2)$$

In the above equation (2), $\left(1-\phi_1 B \ldots-\phi_p B^p\right)(1-\Phi_1 B^m \ldots-\Phi_P B^m)$ represent AR *(p)* and seasonal AR *(P)* part, $\left(1+\theta_1 B+\cdots+\theta_q B^q\right)\left(1+\Theta_1 B^m+\cdots+\Theta_Q B^m\right)$ MA (*q*) and seasonal MA *(Q)* part, and $(1-B)^d(1-B^m)^D y_t$ non-seasonal *d* and seasonal *D* differences. Again, constant is denoted by *c* while e_t represents white noise error term. Literature suggests that the ARIMA approach is appropriate for forecasting horizons of twelve to eighteen months and when at least fifty observations are available. As this method requires complex and repeated mathematical computations, most statistical software offer automated ARIMA calculation.

4.2. Regression with ARIMA errors

An ARIMA model is purely based on the estimation of past observations of a series and it does not allow inclusion of other variables that could have deterministic power for dependent variable. Other external variables may explain some additional historical variation and improve overall forecasts. One of the extension versions of an ARIMA model, which allows as to incorporate other external variables, is called regression with ARIMA errors or dynamic regression model. It includes dynamic autoregressive (AR) and moving average (MA) components, in addition to theoretical explanatory variables, to explain variations in endogenous variables. A general ARIMA model with regression errors can be written as:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots \beta_k x_{k,t} + n_t \quad (3)$$

$$n_t = \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)}{(1-\phi_1 B \ldots-\phi_p B^p)\,(1-B)^d} e_t \quad (4)$$

From equation (3) and (4) we can see that model has two error terms, i.e. n_t and e_t. The first one is derived from the regression model, while the latter comes from the ARIMA model, where only ARIMA model errors are assumed to be white noise (Hyndman et al., 2018). The main assumption of above model is that dependent and all independent variables are stationary before estimation is conducted, otherwise, coefficients may be biased. However, this assumption does not apply when non-stationary variables are cointegrated, i.e. existence of linear combination between dependent and independent variables (Hyndman et al., 2018).

4.3. Automated algorithm for ARIMA estimation

The usage of ARIMA models for prediction of time series variables is common practice of many businesses across the world (Hyndman et al., 2018). This process can be demanding and time consuming as often hundreds of variables need to be forecasted monthly. In addition, many people, due to the lack of training and nature of time series modelling, have difficulties with properly conducting time series forecasting. Especially, this is the case with using ARIMA as the process of order selection is difficult to apply and subject to a user's personal judgments (Hyndman et al., 2018). These circumstances inspired R. Hyndman and Y. Khandakr to develop algorithm for automated ARIMA modelling incorporated in the "*auto.arima"* function of R's "*forecast*" statistical package. The "*auto.arima"* function works in the following way:

- Number of non-seasonal differences is selected based on KPSS test.
- Number of seasonal differences is based on OCSB unit root test.
- AR and MA components are selected by minimizing AIC or BIC depending on our preferences.

The algorithm is using stepwise search to traverse the model space, which decreases the time for the algorithm to produce the best model. It also allows incorporation of

additional explanatory regressors and estimation of regression with ARIMA errors. In addition, we have the option to manipulate with ARIMA arguments, i.e. *d*, *D*, *p*, *P*, *q*, and *Q*, and restrict model selection. The Figure 12, below, presents comparison over traditional approach of ARIMA modelling and approach using automated "*auto.arima"* function. As you can see, the automated algorithm does not cover all the steps of modelling process, i.e. identification of unusual patterns in a data, testing residuals, and forecasting calculation.

4.4. Framework of modelling procedure

Using Turkey as the case study, the main aim of this study is to determine whether the usage of GT data improves forecasts of economic variables, i.e. unemployment, tourism demand, and car sales, comparing to the baseline models. Second aim is to compare performances of the two different query selection methods. For reaching stated aims, following general procedure is used:

1. Time series of variables are splitted into two periods, i.e. training data used for model estimation of in-sample period and test data used for testing predictive ability of the estimated models (out-of-sample period).
2. For every variable of interest baseline ARIMA model is estimated.
3. For every variable of interest second model is constructed by extending baseline model with the GT variable obtained from the simple query selection method.
4. For every variable of interest third model is constructed by extending baseline model with the GT composite search index variable obtained with PCA.
5. For every variable of interest out-of-sample static forecast is computed and compared based on the Root Mean Square Errors (RMSE) and the Mean Absolute Error (MAE) criteria.

For model estimations modelling process with "*auto.arima"*,which is presented in the Figure 12, is used. For all models, AR and MA components are selected by minimizing AIC criteria. We set "*stepwise"* and "*approximation"* arguments within "*auto.arima"* function as "FALSE". This will exclude named arguments and algorithm will perform searches over all models. In models with GT variables, we always use their contemporary observations in addition with their first lag values.

The lag variables are added as online searches might lead actual purchasing action or arrival with the lag and they are kept if statistically significant from zero. After "*auto.arima*" estimations, diagnostic test is performed. Namely, Ljung-Box test is used for checking whether residuals are showing white noise behavior using the 5% significance level. Following subsections present model specifications for each variable.

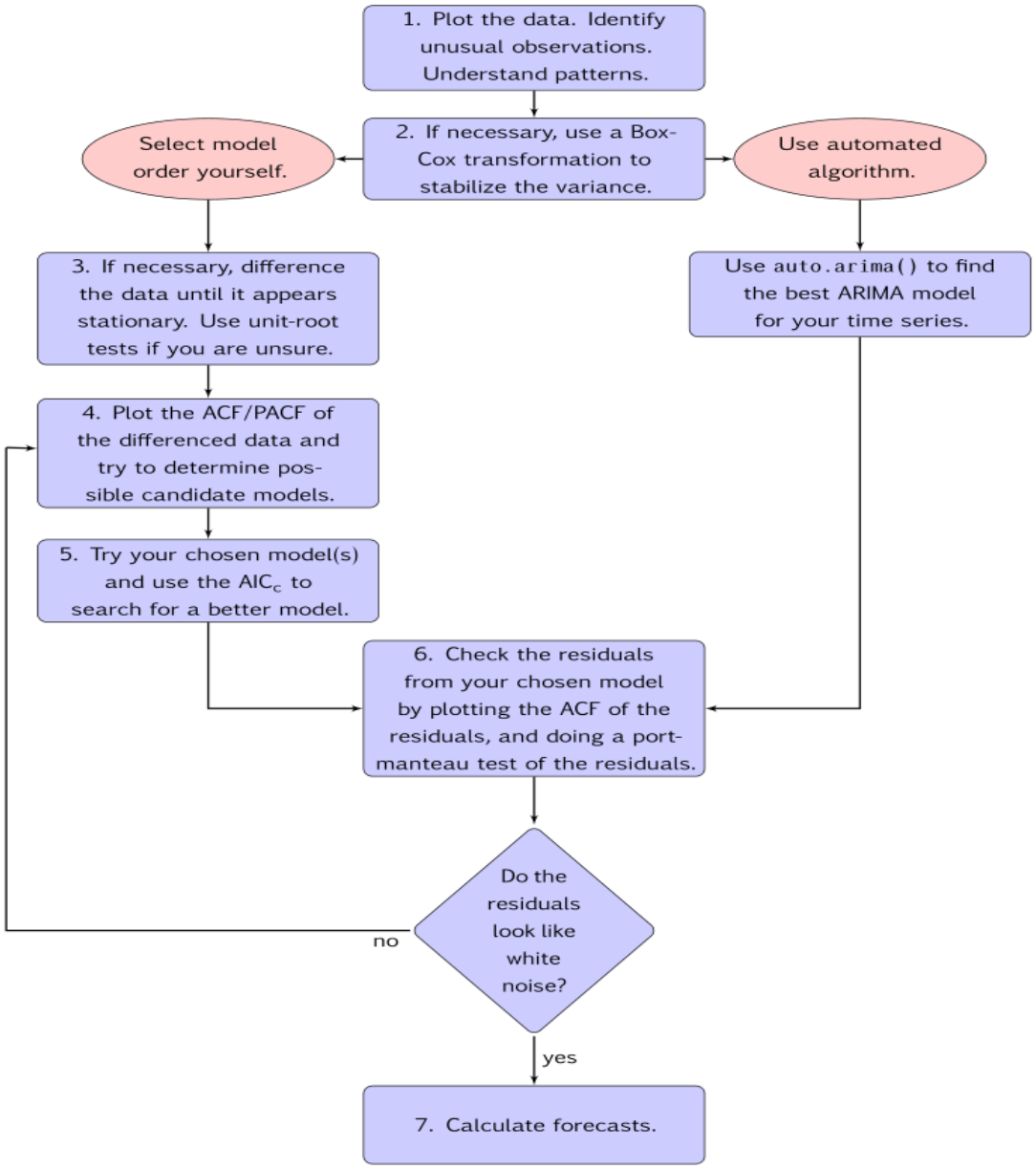


Figure 12: **ARIMA modelling process (Hyndman et al., 2018).**

4.4.1. Tourism demand

For all source countries, variables are splitted into training (in-sample) data and test (out-of-sample) data. Traning data starts from January of 2007 to December of 2016, while last 16 months, from January of 2017 to April of 2018, are left for test data. In order to ensure stable variance, past values of tourism demand are converted to the logarithm form before running the "*auto.arima"* algorithm. The general three models for tourism demand can be written as:

Baseline model: $\ln(TD_{t,i}) = \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

Model 2: $\ln(TD_{t,i}) = \beta_1 GT_{t,i} + \beta_2 GT_{t-1,i} + \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

Model 3: $\ln(TD_{t,i}) = \beta_1 GTx_{t,i} + \beta_2 GTx_{t-1,i} + \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

From models above $TD_{t,i}$ represents tourism demand at time *t* for country *i*, $GT_{t,i}$ represents GT variable for word "*Turkey*" at time *t* for country *i*, $GTx_{t,i}$ represents GT composite search index at time *t* for country *i*, $GT_{t-1,i}$ and $GTx_{t-1,i}$ represent first lags of GT variables at time *t* for country *i*. The rest parts of the equations represent orders of SARIMA model that will be determined by "*auto.arima"* function. GT data, obtained for the word "*Turkey*", in the case of Germany and Austria is affected by the one-time event, i.e. coup attack in July of 2016. For capturing the effect of this outlier, we incorporate the dummy variable, denoted as "*coupd "*, while estimating the second model for these countries.

4.4.2. Unemployment

In the case of unemployment variable, we used seasonally and calendar unadjusted total unemployment and youth unemployment data. Unemployment series is transformed to logarithm form for stabilizing the variance. In-sample data is from January of 2007 to December of 2016. The rest 12 month, from January of 2017 to December of 2017, is used for out-of-sample prediction. The general three models for unemployment can be written as:

Baseline model: $\ln(U_t) = \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

Model 2: $\ln(U_t) = \beta_1 GT_t + \beta_2 GT_{t-1} + \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

Model 3: $\ln(U_t) = \beta_1 GTx_t + \beta_2 GTx_{t-1} + \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

From models above U_t represents total unemployment in Turkey at time *t*, GT_t represents GT variable for word "*iş ilanları*" at time *t*, GTx_t represents GT composite search index at time *t*, GT_{t-1} and GTx_{t-1} represent first lags of GT variables at time *t*. Other parts of the equations represent orders of ARIMA models that will be determined by "*auto.arima*" function. In the case of youth unemployment, we use the same model specification where dependent variable is denoted as YU_t and represent youth unemployment of Turkey at time t.

4.4.3. Car sales

For the car sales we used seasonally and calendar unadjusted total registered number of cars as our dependent variable. For stabilizing the variance, dependent variable is transformed to logarithm form. Data from January of 2010 to December of 2016 is used for in-sample forecasting. The rest 12 month, from January of 2017 to December of 2017, is used for out-of-sample prediction. The general three models for car sales can be written as:

Baseline model: $\ln(CS_{t,i}) = \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

Model 2: $\ln(CS_{t,i}) = \beta_1 GT_{t,i} + \beta_2 GT_{t-1,i} + \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

Model 3: $\ln(CS_{t,i}) = \beta_1 GTx_{t,i} + \beta_2 GTx_{t-1,i} + \frac{c+(1+\theta_1 B+\cdots+\theta_q B^q)(1+\Theta_1 B^m+\cdots+\Theta_Q B^m)}{(1-\phi_1 B\ldots-\phi_p B^p)\,(1-\Phi_1 B^m\ldots-\Phi_P B^m)(1-B)^d(1-B^m)^D}\,e_t$

From models above $CS_{t,i}$ represents registered cars at time *t* for brand *i*, $GT_{t,i}$ represents GT variable for keywords from the first query selection method at time *t* and brand *i*, $GTx_{t,i}$ represents GT composite search index at time *t* for brand *i*, $GT_{t-1,i}$ and $GTx_{t-1,i}$ represent first lags of GT variables at time *t* for brand *i*. The rest parts of the equations represent orders of ARIMA models.

5. RESULTS

In this section main results of described methodology are presented. First, for each variable, outputs of automated algorithm, statistical and diagnostics test results are presented. Later, forecasting performances are introduced.

5.1. Tourism demand

5.1.1. Model selection and estimations

For each source country, i.e. Austria, France, Germany, Netherlands, and Poland, three models are estimated. First one is the baseline model using only the past values of tourism arrivals. The second and third are incorporating GT variables, obtained from two different query selection methods, as additional regressors. The Table 1 shows the output of "*auto.arima*" function for baseline models of each source country.

Table 1: Baseline models for each source country

Country	**Baseline model**
Austria	ARIMA $(0,1,1)\,(1,1,0)_{12}$
France	ARIMA $(4,1,0)\,(0,1,1)_{12}$
Germany	ARIMA $(4,1,0)\,(0,1,1)_{12}$
Netherlands	ARIMA $(0,1,1)\,(2,1,0)_{12}$
Poland	ARIMA $(1,1,1)\,(1,1,1)_{12}$

Before using estimated models for forecasting purpose, we need to make sure that our variables are showing stationarity behavior. From graphs of tourism arrivals, presented in data section, we could notice strong seasonality behavior and possible non-stationarity. The automated algorithm confirmed these assumption as it selected

one non-seasonal difference (*d*=1) and one seasonal difference (*D*=1) for all variables. To test these outputs, we employed Augmented Dickey–Fuller test for stationarity. The ADF test confirmed previous results and all variables became stable after first difference (Appendix III). After estimating baseline models, they are extended by adding GT variables and their first lags as additional explanatory variables. As both sides of equations are differenced, we have models in differences without intercepts, which are lost due to differencing.

In the case of Austria, contemporary observations of GT variables were statistically significant and had positive sign. However, the lagged variables were not statistically different from zero and they were omitted from model. Residuals from all models are white noise as null hypothesis could not be rejected.

Table 2: SARIMA estimations for Austria

Austria			
Variables	Baseline model	Model 2	Model 3
ma1	-0.6838*** (0.0814)	-0.6632*** (0.0843)	-0.6651*** (0.0814)
sar1	-0.2363* (0.0993)	-0.2739** (0.0997)	-0.1720 (0.1049)
GTaustria		0.0069** (0.0026)	
GTaustriax			0.0225** (0.0088)
coupd		-0.1553 (0.2001)	
AIC	-72.13	-76.27	-76.44
Ljung-Box test (lags=24)	p-value = 0.81	p-value = 0.83	p-value = 0.50

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*' , 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 120
Period :2007-2016
Standard errors are in parenthesis.

When it comes to France (Table 3), for both extended models, GT variables and their lags were statisticaly significant. All models satisfy assumption of uncorrelated residuals. In the case of Netherlands (Table 4), for Model 2 only lagged variable of web searches was statistically significant, while for the Model 3 none of the GT index variables were statistically significant. However, we left lagged value as its insignificants was lower, i.e. 0.17 %. According to the Ljung-Box test, residuals of all models are white noise.

Table 3: SARIMA estimations for France

France			
Variables	Baseline model	Model 2	Model 3
ar1	-0.2264* (0.0940)	-0.2658** (0.0945)	-0.2342* (0.0943)
ar2	-0.3636*** (0.0870)	-0.3333 *** (0.0879)	-0.3386*** (0.0897)
ar3	-0.4129*** (0.0870)	-0.4369 *** (0.0894)	-0.4266*** (0.0866)
ar4	-0.2836** (0.1000)	-0.3206 ** (0.1047)	-0.3472*** (0.1041)
sma1	-0.3657** (0.1382)	-0.4814 ** (0.1731)	-0.5144*** (0.1499)
GTfrance		0.0061 * (0.0028)	
L1GTfrance		0.0077 * (0.0030)	
GTfrancex			0.0156* (0.0073)
L1GTfrancex			0.0204** (0.0071)
AIC	-100.07	-103.68	-104.74
Ljung-Box test (lags=24)	p-value =0.55	p-value = 0.57	p-value = 0.37

Level of significance: 0.001 '***', 0.01 '**', 0.05 '*' , 0.1 '.'
Ljung-Box test H0: Residuals are white noise
Sample size: 120
Period: 2007-2016
Standard errors are in parenthesis.

Table 4: SARIMA estimations for Netherlands

Netherlands			
Variables	Baseline model	Model 2	Model 3
ma1	-0.6984*** (0.0822)	-0.7155 *** (0.0788)	-0.7074*** (0.0804)
sar1	-0.5827*** (0.0946)	-0.5961 *** (0.0955)	-0.5742*** (0.0961)
sar2	-0.2006* (0.0926)	-0.2000 * (0.0940)	-0.1873* (0.0938)
L1GTnetherlands		0.0072** (0.0027)	
L1GTnetherlandsx			0.0174 (0.0126)
AIC	-32.05	-35.65	-30.5
Ljung-Box test (lags=24)	p-value =0.88	p-value = 0.89	p-value 0.90

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*' , 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 120
Period: 2007-2016
Standard errors are in parenthesis.

Table 5: SARIMA estimations for Germany

Germany			
Variables	Baseline model	Model 2	Model 3
ar1	-0.6542*** (0.0952)	-0.6663 *** (0.0946)	-0.6563*** (0.0927)
ar2	-0.2040. (0.1081)	-0.1708 (0.1090)	-0.2517* (0.1072)
ar3	-0.2903* (0.1138)	-0.2952 * (0.1178)	-0.3531** (0.1153)
ar4	-0.3470*** (0.0979)	-0.3489 *** (0.1014)	-0.3806*** (0.0978)
sma1	-0.4253*** (0.1069)	-0.4168 *** (0.1070)	-0.4450*** (0.1174)
GTgermany		0.0069* (0.0027)	
GTgermanyx			0.0170*** (0.0047)
coupd		-0.2364 (0.1763)	
AIC	-147.54	-151.87	-157.94
Ljung-Box test (lags=24)	p-value = 0.60	p-value = 0.70	p-value = 0.89

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*' , 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 120
Period: 2007-2016
Standard errors are in parenthesis.

In the Table 5 estimations of models from Germany are presented. Contemporary GT variables were statistically significant, but their lag variables were not, and they were omitted from the model. In addition, all models passed diagnostic test as residuals are showing white noise behavior. For the last source country, i.e. Poland (Table 6), both, contemporary and lagged, GT variables were statistically significant. Also, all models satisfy assumption of residuals being white noise.

Table 6: SARIMA estimations for Poland

Poland			
Variables	Baseline model	Model 2	Model 3
ar1	0.7979*** (0.1366)	0.5924 ** (0.1950)	0.3634**.** (0.1907)
ma1	-0.9257*** (0.0949)	-0.8208 *** (0.1362)	-0.7317*** (0.1419)
sar1	0.3195* (0.1870)	0.2962**.** (0.1694)	0.3273**.** (0.1936)
sma1	-0.7707*** (0.1657)	-0.8658*** (0.2024)	-0.7579*** (0.1754)
GTpoland		0.0045 * (0.0020)	
L1GTpoland		0.0067 ** (0.0020)	
GTpolandx			0.0327*** (0.0085)
L1GTpolandx			0.0253** (0.0085)
AIC	-94.17	-100.94	-113.49
Ljung-Box test (lags=24)	p-value =0.78	p-value = 0.92	p-value = 0.51

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*', 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 120
Period: 2007-2016
Standard errors are in parenthesis.

5.1.2. Forecasting results

In this section, forecasting performance of models described above are presented. In the Table 7 forecasting performance of all models is presented. In general, extended models with GT variables managed to improve out-of-sample fit comparing to the baseline models. When it comes to performances of two query selection methods, composite search index tends to provide better forecasts.

In the case of Austria, both extended models improved in-sample fit comparing to the baseline model by up to 3.8 percent using RMSE and MAE criteria. For the out-of-sample forecasting performance, only Model 3, with composite search index, showed forecasting improvement. For Model 3, RMSE and MAE are decreased by 5 percent and 4.3 percent, respectively, compared to the baseline model.

Using France as the source country, both models with web search variables improved in-sample fit by up to 5 percent comparing to the baseline model. Also, extended models performed better in the out-of-sample case. The Model 2 decreased forecasting errors by 15.9 percent according to RMSE and 18.6 percent in the case of MAE criteria. Model 3, with composite search index, decreased forecasting errors for 21.3 percent in the case of RMSE and 26.4 percent in the case of MAE criteria.

When it comes to Germany, both extended models showed better in-sample fit, where RMSE and MAE decreased by up to 7.5 percent. Out-of-sample results were mixed. Model 3 showed better out-of-sample performance, while this was not the case with the Model 2. According to the out-of-sample results for Model 3, RMSE and MAE decreased by 12.3 percent and 14.1 percent respectively. In the case of Model 2, RMSE and MAE were higher by 6 percent and 8.8 percent, respectively, comparing to the baseline model.

Results from Netherlands showed forecasting improvements of extended models in the in-sample case, where RMSE and MAE improved by up to 2.9 percent. In the case of the out-of-sample performance, RMSE and MAE decreased by up to 1.6 percent and 3.1 percent respectively for Model 2, and 5.6 percent and 7.1 percent respectively for Model 3.

In the case of Poland, both extended models performed better according to the in-sample fit. RMSE and MAE decreased by 7 percent and 11.1 percent for Model 2 and 11 percent and 7.7 percent for Model 3. Accroding to the out-of-sample performance, both extended models performed better comparing to the baseline model. Model 2 decreased RMSE by 8.1 percent and MAE by 4.5 percent, while RMSE and MAE of Model 3 decreased by 4.6 percent and 0.3 percent respectively.

Table 7: Forecasting performance for tourism demand

Source country		In -sample						Out-of-sample					
		RMSE	**Rank**	**%**	**MAE**	**Rank**	**%**	**RMSE**	**Rank**	**%**	**MAE**	**Rank**	%
Austria	*Baseline model*	0.1577	3		0.1192	3		0.5973	2		0.4897	2	
	Model 2	**0.1516**	**1**	**3.8%**	**0.1148**	**1**	**3.7%**	0.6263	3	-4.9%	0.5236	3	-6.9%
	Model 3	0.1533	2	2.7%	0.1157	2	2.9%	**0.5676**	**1**	**5.0%**	**0.4688**	**1**	**4.3%**
France	*Baseline model*	0.1338	3		0.1003	3		0.6258	3		0.5405	3	
	Model 2	0.1275	2	4.9%	0.0959	2	4%	0.5265	2	15.9%	0.4402	2	18.6%
	Model 3	**0.1266**	**1**	**5%**	**0.0957**	**1**	**5%**	**0.4923**	**1**	**21.3%**	**0.3978**	**1**	**26.4%**
Germany	*Baseline model*	0.1069	3		0.0782	3		0.4356	2		0.3751	2	
	Model 2	0.1029	2	3.8%	0.0736	2	5.9%	0.4616	3	-6.0%	0.4082	3	-8.8%
	Model 3	**0.1007**	**1**	**5.8%**	**0.0723**	**1**	**7.5%**	**0.3822**	**1**	**12.3%**	**0.3222**	**1**	**14.1%**
Netherlands	*Baseline model*	0.1852	3		0.1282	2		0.3688	3		0.3001	3	
	Model 2	**0.1798**	**1**	**2.9%**	**0.1261**	**1**	**1.7%**	0.3629	2	1.6%	0.2907	2	3.1%
	Model 3	0.1845	2	0.4%	0.1284	3	-0.1%	**0.3481**	**1**	**5.6%**	**0.2788**	**1**	**7.1%**
Poland	*Baseline model*	0.1366	3		0.0980	3		0.5877	3		0.5062	3	
	Model 2	0.1265	2	7%	**0.0882**	**1**	**11.1%**	**0.5403**	**1**	**8.1%**	**0.4833**	**1**	**4.5%**
	Model 3	**0.1223**	**1**	**11%**	0.0905	2	7.7%	0.5606	2	4.6%	0.5046	2	0.3%

Note: Column denoted with " % " represents percentage change in forecasting accuracy of the second and third model comparing to the baseline model.

5.2. Unemployment

5.2.1. Model selection and estimations

Following previous example with tourism demand, baseline models for total and youth unemployment variables are selected by using *"auto.arima"*. Based on AIC criteria algorithm selected ARIMA (2,0,2) $(1,1,0)_{12}$ with drift for total unemployment and ARIMA (1,0,3) (1,1,0) for youth unemployment (Table 8). According to the KPSS and OCSB test, variables need only seasonal differencing to make them stationary. To check these results, we applied additional Augmented Dickey–Fuller test for stationarity, which results confirmed outputs of algorithm (Appendix III). Additionally, estimated baseline models are extended with two GT variables and their first lags.

Table 8: Baseline models for each unemployment variable

Variable	Baseline model
Total unemployment	ARIMA (2,0,2) $(0,1,1)_{12}$ with drift
Youth unemployment	ARIMA (1,0,3) $(0,1,1)_{12}$

In the case of total unemployment, for Model 2 both GT variables were not statistically significant, but we left lagged variable as insignificant was lower comparing to the unlagged variable. For Model 3 only lagged value of composite search index was statistically significant. All models satisfied diagnostic test of uncorrelated residuals. Other estimated coefficients of models are provided in the Table 9. Models with youth unemployment (Table 10) as dependent variable satisfied diagnostic tests as we could not reject the null of residuals being white noise. For both extended models we kept lagged values of GT variables. In the case of Model 3 first lag was statistically significant, while for Model 2 its insignificance was lower comparing to unlagged values.

Table 9: SARIMA estimations for total unemployment

Total Unemployment			
Variables	Baseline model	Model 2	Model 3
ar1	0.5594*** (0.1230)	0.5629*** (0.1095)	0.5814*** (0.0986)
ar2	0.3882** (0.1238)	0.3814*** (0.1104)	0.3626*** (0.0994)
ma1	0.9636*** (0.0780)	0.9938*** (0.0643)	1.0203*** (0.0662)
ma2	0.8065*** (0.0632)	0.8582*** (0.0796)	0.9537*** (0.1328)
sma1	-0.7438*** (0.1430)	-0.7998*** (0.1669)	-0.8380*** (0.1920)
drift	0.0048. (0.0027)	0.0042. (0.0027)	0.0040. (0.0022)
L1GTU		0.0002 (0.0002)	
L1GTUx			0.0025** (0.0008)
AIC	-479.52	472.19	-476.54
Ljung-Box test (lags=24)	p-value = 0.71	p-value = 0.81	p-value = 0.73

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*', 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 120
Period: 2007-2016
Standard errors are in parenthesis.

Table 10: SARIMA estimations for youth unemployment

Youth Unemployment			
Variables	Baseline model	Model 2	Model 3
ar1	0.9726*** (0.0321)	0.9694*** (0.0341)	0.9698*** (0.0343)
ma1	0.4343*** (0.0901)	0.4884*** (0.0953)	0.5612*** (0.0877)
ma2	0.4066*** (0.1068)	0.4041*** (0.1189)	0.4306*** (0.1046)
ma3	-0.4164*** (0.0900)	-0.4897*** (0.1207)	-0.5244*** (0.0878)
sma1	-0.7445*** (0.1509)	-0.8006*** (0.1609)	-0.8142*** (0.1576)
L1GTYU		0.0006 (0.0004)	
L1GTYUx			0.0005*** (0.0008)
AIC	-399.36	-394	-402.94
Ljung-Box test (lags=24)	p-value = 0.81	p-value = 0.96	p-value = 0.88

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*', 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 120
Period: 2007-2016
Standard errors are in parenthesis.

5.2.2. Forecasting results

In this section forecasting performance of models described above are presented. Table 11 presents forecasting performance of all models. In the case of total and youth unemployment, extended models managed to improve out-of-sample fit comparing to the baseline models. However, there is no clear distinction regarding which of the two query selection methods is providing better results.

In the case of total unemployment in-sample fit improved by both extended models. For Model 2, RMSE and MAE decreased by 2 and 2.1 percent respectively. For Model 3 in-sample performance was slightly better comparing to Model 2, where RMSE and MAE decreased by 5.9 and 5.6 percent respectively. When it comes to the out-of-sample performance, extended models provided better forecasts comparing to the baseline model. For Model 2, RMSE decreased by 14.2 percent and MAE by 14.9 percent, while RMSE decreased by 8.4 percent and MAE by 3.7 percent for Model 3

For youth unemployment results were similar, but GT variables provided significantly better improvements comparing to the total unemployment. This supports our assumption that web searches may better help in explaining variations in the case of youth unemployment as young people are more prone to using internet and web search providers. Again, both extended models provided better in-sample fit. Errors decreased by up to 3 percent in the case of Model 2 and by up to 10 percent for Model 3. When it comes to out-of-sample results, both models with GT variables performed better than baseline model. For Model 2, RMSE and MAE decreased by 24.7 percent and 13.4 percent respectively, while for Model 3 RMSE and MAE decreased by 24.2 percent and 17.3 percent respectively.

Table 11: Forecasting performance for unemployment variables

Unemployment variables		In -sample						Out-of-sample					
		RMSE	**Rank**	**%**	**MAE**	**Rank**	**%**	**RMSE**	**Rank**	**%**	**MAE**	**Rank**	**%**
Total unemployment	*Baseline model*	0.0219	3		0.0176	3		0.1065	3		0.0884	3	
	Model 2	0.0215	2	2.0%	0.0172	2	2.1%	**0.0913**	**1**	**14.2%**	**0.0752**	**1**	**14.9%**
	Model 3	**0.0206**	**1**	**5.9%**	**0.0166**	**1**	**5.6%**	0.0976	2	8.4%	0.0851	2	3.7%
Youth unemployment	*Baseline model*	0.0319	3		0.0242	3		0.0927	3		0.0600	3	
	Model 2	0.0309	2	3.0%	0.0234	2	3.1%	**0.0698**	**1**	**24.7%**	0.0519	2	13.4%
	Model 3	**0.0292**	**1**	**8.4%**	**0.0218**	**1**	**10.0%**	0.0703	2	24.2%	**0.0496**	**1**	**17.3%**

Note: Column denoted with " % " represents percentage change in forecasting accuracy of the second and third model comparing to the baseline model.

5.3. Car sales

5.3.1. Model selection and estimations

First, we determined the baseline models for each car brand by using "*auto.arima"* algorithm. Algorithm selected only one seasonal differencing for all variables. We checked these results by applying additional Augmented Dickey–Fuller test for stationarity (Appendix III). The tests showed that GT variables need additional non-seasonal differencing to make them stationary. Therefore, we applied non-seasonal differencing while estimating our models for each car brand.

Table 12: Baseline models for each car brand

Car brand	Baseline model
Renault	ARIMA (2,1,1) (1,1,0) $_{12}$
Fiat	ARIMA (2,1,0) (2,1,0) $_{12}$
Opel	ARIMA (1,1,1) (1,1,0) $_{12}$
Hyundai	ARIMA (0,1,2) (0,1,1) $_{12}$

In the case of Renault (Table 13) all models satisfied diagnostic test of uncorrelated residuals. When it comes to GT variables, only their lagged values were statistically significant. Models using data for Fiat's registered cars also satisfied condition of residuals being white noise. We kept lagged values of GT variables as they were statistically significant. Other estimated coefficients are provided in the Table 14.

Table 13: SARIMA estimations for Renault

Renault			
Variables	Baseline model	Model 2	Model 3
ar1	-1.2145*** (0.1846)	0.3031* (0.1198)	0.3234** (0.1193)
ar2	0.2533*** (0.1478)	0.2118. (0.1234)	0.2376. (0.1223)
ma1	0.8487. (0.1358)	-0.9999*** (0.0487)	-0.9999*** (0.0471)
sar1	-0.6999*** (0.0836)	-0.7030*** (0.0833)	0.6588*** (0.0921)
L1GTRenault		0.0081* (0.0032)	
L1GTRenaultx			0.0670* (0.0294)
AIC	17.08	9.8	10.58
Ljung-Box test (lags=24)	p-value = 0.51	p-value = 0.55	p-value = 0.56

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*', 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 84
Period: 2010-2016
Standard errors are in parenthesis.

Table 14: SARIMA estimations for Fiat

Fiat			
Variables	Baseline model	Model 2	Model 3
ar1	-0.7017*** (0.1195)	-0.7362 *** (0.1208)	-0.7666 *** (0.1205)
ar2	-0.1256 (0.1179)	-0.1825 (0.1210)	-0.1789 (0.1191)
sar1	-0.6390*** (0.1310)	-0.6306 *** (0.1350)	-0.5856 *** (0.1353)
sar2	-0.3164* (0.1334)	-0.2811 * (0.1384)	-0.2657 * (0.1355)
L1GTCS		0.0080. (0.0047)	
L1GTCSx			0.1013 * (0.0402)
AIC	30.37	29.36	26.21
Ljung-Box test (lags=24)	p-value = 0.76	p-value = 0.74	p-value = 0.89

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*', 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 84
Period: 2010-2016
Standard errors are in parenthesis.

Table 15 presents model estimations for the Opel. For both query selection methods, GT variables and their lagged values were not statistically significant. As with previous cases, we kept their lagged values as insignificance was lower. All models satisfied diagnostics tests.

Table 15: SARIMA estimations for Opel

Opel			
Variables	Baseline model	Model 2	Model 3
ar1	-0.17993 (0.1632)	-0.1813 *** (0.1720)	-0.2205 (0.1788)
ma1	-0.58653 *** (0.1216)	-0.5748 *** (0.1413)	-0.5447 *** (0.1529)
sar1	-0.53633 *** (0.1135)	-0.5515 (0.1115)	-0.5208 *** (0.1193)
L1GTCS		0.0051 (0.0051)	
L1GTCSx			0.0459 (0.0380)
AIC	36.27	32.81	32.29
Ljung-Box test (lags=24)	p-value = 0.93	p-value = 0.9835	p-value = 0.98

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*', 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 84
Period: 2010-2016
Standard errors are in parenthesis.

Table 16: SARIMA estimations for Hyundai

Hyundai			
Variables	Baseline model	Model 2	Model 3
ma1	-0.8574*** (0.1099)	-0.8534*** (0.1117)	-0.8881*** (0.1107)
ma2	0.1201 (0.1048)	0.1381 (0.1095)	0.1484 (0.1092)
sma1	-0.7466*** (0.2177)	-0.7351*** (0.2230)	-0.7857** (0.2476)
GTCS		0.0057 (0.0035)	
GTCSx			0.0471. (0.0270)
AIC	-9.39	-9.99	-10.34
Ljung-Box test (lags=24)	p-value = 0.89	p-value = 0.64	p-value = 0.63

Level of significance: *0.001 '***', 0.01 '**', 0.05 '*', 0.1 '.'*
Ljung-Box test H0: *Residuals are white noise*
Sample size: 84
Period: 2010-2016
Standard errors are in parenthesis.

Estimation results for Hyundai (Table 16) showed following results. All models satisfied diagnostics tests of residuals being white noise. GT variables were not statistically significant except the lagged values of composite search index at the 10 percent significance level. In the case of the first query selection method, we kept unlagged values as their insignificance was lower comparing to the lagged ones.

5.3.2. Forecasting results

After model estimations their forecasting performance is checked. Extended models with GT variables, generally, improved out-of-sample fit comparing to the baseline models. When it comes to query selection methods, results are mixed. In the case of Renault, Fiat and partly Opel (only in the case of MAE criteria) first query selection method provided better results. Composite search index performed better only in the case of Hyundai.

For Renault, in-sample fit improved by up to 9.2 percent for Model 2 and by up to 7.7 percent for Model 3. When it comes to the out-of-sample performance, Model 2 performed the best by reducing forecasting error for 11.3 percent according to RMSE criteria and 18.8 percent following MAE criteria. For Model 3, RMSE reduced by 1.7 percent and MAE by 4.8 percent comparing to the baseline model.

In the case of Fiat, in-sample fit improved by up to 1.8 percent for Model 2 and 3.6 percent for Model 3. According to the out-of-sample fit, extended models performed better comparing to the baseline model, except for the Model 3 following the RMSE criteria. However, Model 2 performed the best for both, RMSE and MAE, criteria. Forecasting error is reduced by 8.1 percent in the case of RMSE and 16.5 percent in the case of MAE. Model 3 performed better only in the case of MAE criteria where error is slightly decreased by 0.9 percent.

In-sample performance of extended model outperformed baseline model in the case of Opel's data. Forecast error is decreased by almost 4 percent and 8.2 percent according to RMSE and MAE respectively. When it comes to out-of-sample fit, Model 3 performed the best according to the RMSE that decreased by 5.2 percent. In the case of MAE criteria, Model 2 slightly outperformed Model 3 reducing the error by 3.8 percent.

In the last case of Hyundai, results are following. In-sample performance improved for both extended models. RMSE and MAE decreased by up to 3.1 percent and 1.3 percent respectively. According to out of -sample fit, RMSE decreased by 14.3 percent and MAE by 17.9 percent in the case of Model 2. Model 3 performed the best, where RMSE reduced by 20.8 percent and MAE by 26.21 percent. Following Table 17 presents forecasting performance for all models.

Table 17: Forecasting performance for car sales

Car brand		In -sample						Out-of-sample					
		RMSE	**Rank**	**%**	**MAE**	**Rank**	**%**	**RMSE**	**Rank**	**%**	**MAE**	**Rank**	%
Renault	*Baseline model*	0.2209	3		0.1558	3		0.2608	3		0.2401	3	
	Model 2	**0.2005**	**1**	**9.2%**	**0.1490**	**1**	**4.4%**	**0.2313**	**1**	**11.3%**	**0.1949**	**1**	**18.8%**
	Model 3	0.2038	2	7.7%	0.1504	2	3.5%	0.2563	2	1.7%	0.2286	2	4.8%
Fiat	*Baseline model*	0.2457	3		0.1788	3		0.1416	3		0.1147	3	
	Model 2	0.2412	2	1.8%	0.1757	2	1.8%	**0.1301**	**1**	**8.1%**	**0.0957**	**1**	**16.5%**
	Model 3	**0.2369**	**1**	**3.6%**	**0.1732**	**1**	**3.1%**	0.1508	2	-6.5%	0.1137	2	0.9%
Opel	*Baseline model*	0.2626	3		0.1921	3		0.3837	3		0.3314	3	
	Model 2	**0.2524**	**1**	**3.9%**	**0.1764**	**1**	**8.2%**	0.3666	2	4.4%	**0.3189**	**1**	**3.8%**
	Model 3	0.2525	2	3.89%	0.1773	2	7.7%	**0.3637**	**1**	**5.2%**	0.3193	2	3.7%
Hyundai	*Baseline model*	0.1828	3		0.1284	3		0.0992	3		0.0859	3	
	Model 2	0.1801	2	1.5%	0.1377	2	0.7%	0.0850	2	14.3%	0.0705	2	17.9%
	Model 3	**0.1771**	**1**	**3.1%**	**0.1374**	**1**	**1.3%**	**0.0786**	**1**	**20.8%**	**0.0635**	**1**	**26.1%**

Note: Column denoted with " % " represents percentage change in the forecasting accuracy of the second and third model comparing to the baseline model.

6. CONCLUSION

This paper tests ability of GT to predict economic variables, i.e. tourism demand, unemployment and car sales, in the case of Turkey. Moreover, it uses two different query selection methods and compares their performances. According to the applied methodology, results indicate that in most cases extended models with GT variables show forecasting improvements comparing to the baseline models. Results are in line with the previous studies that demonstrated effectiveness of search queries in the prediction of various economic variables.

In the case of tourism demand, we focused on the five major source countries of tourist arrivals to Turkey. For all countries, most of the GT variables improved forecasting performance, with the exception of variables following the first query selection method in the case of Austria and Germany. For example, in the case of France improvement goes up to 21.3 percent and 26.4 percent according to the RMSE and MAE criteria respectively. For total and youth unemployment results are similar. However, GT variables provide higher predictive power for the youth unemployment, which confirmed assumption that young people are more prone to using internet for making economic decisions. Furthermore, results for car sales are also in accordance with the previous findings. For all four brands, i.e. Renault, Fiat, Opel and Hyundai, GT information reduced forecasting error. For example, in the case of Hyundai, error reduction goes up to 26 percent.

When it comes to performance of the two query selection methods, there is no clear distinction regarding which of them is better. However, it seems that composite search index provides, on average, better results. In the case of tourism demand, composite search index provided better forecasts for all source countries, except Poland, and it always outperformed the baseline models. In addition, it was better in capturing the effect of one-time event comparing to the first query selection method. When it comes to the unemployment variables, first query selection method performed better using the total unemployment as dependent variable, while for the youth unemployment composite search index provided better results following MAE

criteria. Furthermore, for car sales composite search index was better only in the case of Hyundai and Opel following the RMSE criteria.

In terms of contribution, due to best of our knowledge, there is no found literature that tests usability of GT for selected variables, except unemployment, in the case of Turkey. In addition, this is one of the first studies that compares the performances of different query selection methods. The study's results could be beneficial for the policy makers and other stakeholders as selected variables play important role for the Turkish economy. It offers them a new way of tracking economic behavior at almost zero cost. Furthermore, they have ability to get real-time insights regarding economic decisions and nowcast them.

This study has several limitations. For example, we used very simple methodology for our benchmark models. Further research could address the impact of using more sophisticated forecasting models such as machine learning techniques. In addition, for the sake of simplicity and the main goal of this study, we used GT data as the only exogenous variable. To address this limitation, further research could incorporate additional variables for which deterministic impact on selected variables, i.e., tourism demand, unemployment, and car sales, is confirmed in the previous studies. Inclusion of other factors will enhance generalizability of results regarding GT data. Future studies should also focus on creation of more comprehensive and dynamic query selection methods as there is a need for their standardization. Moreover, the results of this study emphasize the usage of other big data sources, i.e. online reviews and various data from social media, for predicting economic behavior.

REFERENCES

Artus, Jacques R. 1972. An econometric analysis of international travel. **Staff papers.** 19.3: 579-614.

Askitas, Nikolaos, and Klaus F. Zimmermann. 2009. Google econometrics and unemployment forecasting. **Applied Economics Quarterly.** 55.2: 107-120.

Bangwayo-Skeete, Prosper F., and Ryan W. Skeete. 2015. Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. **Tourism Management.** 46: 454-464.

Bilgic, Emre. 2017. Google Trends Search Volume Index in Estimation of Istanbul Stock Market Index (BIST). Master's Thesis. Istanbul Bilgi University Institute of Social Sciences.

Blazquez, Desamparados, and Josep Domenech. 2018. Big Data sources and methods for social and economic analyses. **Technological Forecasting and Social Change.** 130: 99-113.

Carrière-Swallow, Yan, and Felipe Labbe. 2013. Nowcasting with Google Trends in an emerging market. **Journal of Forecasting.** 32.4: 289-298.

Chadwick, Meltem Gülenay, and Gönül Sengül. 2015. Nowcasting the Unemployment Rate in Turkey: Let's Ask Google. **Central Bank Review.** 15.3: 15.

Choi, Hyunyoung, and Hal Varian. 2009. **Predicting initial claims for unemployment benefits.** Google Inc. sl: 1-5.

_____. 2012. Predicting the present with Google Trends. **Economic Record.** 88. s1: 2-9.

Competition and Markets Authority (CMA). 2017. **Online search: Consumer and firm behavior.** London, UK.

Daily Sabah. [06.06.2018]. Tourism, 42 million tourists visit Turkey in 2014. https://www.dailysabah.com/tourism/2015/01/01/42-million-tourists-visit-turkey-in-2014.

Douglas, C. F. 2001. **Forecasting Tourism Demand: Methods and Strategies.** Linacre House, Jordan Hill, Oxford.

Einav, Liran, and Jonathan Levin. 2014. Economics in the age of big data. **Science** 346.6210: 1243089.

Ettredge, Michael, John Gerdes, and Gilbert Karuga. 2005. Using web-based search data to predict macroeconomic statistics. **Communications of the ACM.** 48.11: 87-92.

Fantazzini, Dean, and Zhamal Toktamysova. 2015. Forecasting German car sales using Google data and multivariate models. **International Journal of Production Economics.** 170: 97-135.

Figueiredo, Nigel. 2016. Predicting Current Auto Sales in Canada using Google. Bachelor's Thesis. Diss. University of Victoria.

Francesco, D'Amuri. 2009. Predicting unemployment in short samples with internet job search query data. **MPRA (Munich Personal RePEc Archive).** Paper No. 49382.

Ginsberg, Jeremy, et al. 2009. Detecting influenza epidemics using search engine query data. **Nature** 457.7232: 1012.

Gunn III, John F., and David Lester. 2013. Using google searches on the internet to monitor suicidal behavior. **Journal of affective disorders.** 148.2-3: 411-412.

Hurriet Daily News. [14.05.2018]. Turkey expects 40 million tourists from abroad in 2018: Experts. http://www.hurriyetdailynews.com/turkey-expects-40-million-tourists-from-abroad-in-2018-experts-130313.

Hyndman, Rob. J., and Yeasmin Khandakar. 2008. Automatic time series forecasting: The forecast package for R. **Journal of Statistical Software.** 26.3.

Hyndman, Rob J., and George Athanasopoulos. 2018. **Forecasting: principles and practice.** OTexts.

Internet World Stats. [06.06.2018]. Internet User Statistics for the European countries and regions. https://www.internetworldstats.com/stats4.htm.

Invest in Turkey. [17.05.2018]. Sectors, Automotive. http://www.invest.gov.tr/en-US/sectors/Pages/Automotive.aspx.

Invest in Turkey. [24.05.2018]. Sectors, Wellness and Tourism. http://www.invest.gov.tr/en-US/sectors/Pages/WellnessAndTourism.aspx.

Johnston, Kevin. [06.07.2018]. What Is the Relative Importance of Forecasting?. http://smallbusiness.chron.com/relative-importance-forecasting-35627.html.

Jun, Seung-Pyo, Hyoung Sun Yoo, and San Choi. 2017. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. **Technological Forecasting and Social Change.** 130: 69-87.

Khoury, Muin J., and John PA Ioannidis. 2014. Big data meets public health. **Science.** 346.6213: 1054-1055.

Lewis, Elmo. 1903. **Catch-line and argument.** The Book-Keeper 15: 124.

Li, Xin, et al. 2017. Forecasting tourism demand with composite search index. **Tourism management.** 59: 57-66.

Makridakis, Spyros, Steven C. Wheelwright, and Rob J. Hyndman. 2008. **Forecasting methods and applications.** John wiley & sons.

Mavragani, Amaryllis, and Konstantinos P. Tsagarakis. 2016. YES or NO: Predicting the 2015 GReferendum results using Google Trends. **Technological Forecasting and Social Change.** 109: 1-5.

Moe, Wendy W. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. **J. Consum. Psychol.** 13: 29–39.

Naccarato, Alessia, et al. 2018. Combining official and Google Trends data to forecast the Italian youth unemployment rate. **Technological Forecasting and Social Change.** 130: 114-122.

Önder, Irem. 2018. Forecasting tourism demand with Google trends: Accuracy comparison of countries versus cities. **International Journal of Tourism Research.** 19.6: 648-660.

Padhi, Sidhartha S., and Rupesh K. Pati. 2017. Quantifying potential tourist behavior in choice of destination using Google Trends. **Tourism Management Perspectives.** 24: 34-47.

Pan, Bing, Doris Chenguang Wu, and Haiyan Song. 2012. Forecasting hotel room demand using search engine data. **Journal of Hospitality and Tourism Technology.** 3.3: 196-210.

Pelat, Camille, et al. 2009. More diseases tracked by using Google Trends. **Emerging infectious diseases.** 15.8: 1327.

Prakash, B. Aditya, et al. 2012. Winner takes all: competing viruses or ideas on fair-play networks. Proceedings of the 21st international conference on World Wide Web. ACM.

Raju, Puthankurissi S., Lonial, Subhash.C., Glynn Mangold, W. 1995. Differential effects of subjective knowledge, objective knowledge, and usage experience on decision making: an exploratory investigation." **J. Consum. Psychol.** 4:153–180.

Rödel, Elke. 2017. Forecasting tourism demand in Amsterdam with Google Trends. Master's Thesis, University of Twente.

Scott, Steven L., and Hal R. Varian. 2014. Predicting the present with bayesian structural time series. **International Journal of Mathematical Modelling and Numerical Optimisation.** 5.1-2: 4-23.

Seifter, Ari, et al. 2010. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. **Geospatial health.** 4.2: 135-137.

Shim, S., Eastlick, M.A., Lotz, S.L., and Warrington, P. 2001. An online prepurchase intentions model: The role of intention to search. **J. Retail.** 77: 397–416.

Song, Haiyan, Kevin KF Wong, and Kaye KS Chon. 2003. Modelling and forecasting the demand for Hong Kong tourism. **International Journal of Hospitality Management.** 22.4: 435-451.

Song, Haiyan, and Stephen F. Witt. 2000. **Tourism demand modelling and forecasting: Modern econometric approaches.** Routledge.

Song, Haiyan, and Han Liu. 2017. **Predicting Tourist Demand Using Big Data.** Analytics in Smart Tourism Design: 13.

Song, Haiyan, Stephen F. Witt, and Gang Li. 2008. **The advanced econometrics of tourism demand.** Routledge.

Statcounter. [27.05.2018]. Search engine market share. http://gs.statcounter.com/search-engine-market-share.

Statista. [08.06.2018]. Leading vehicle brands in Turkey based on sales (in 1,000 units). https://www.statista.com/statistics/473806/best-selling-vehicle-brands-in-turkey/.

Stephens-Davidowitz, Seth. 2014. The cost of racial animus on a black candidate: Evidence using Google search data. **Journal of Public Economics.** 118: 26-40.

Toth, Istvan Janos, and Miklós Hajdu. 2012. Google as a tool for nowcasting household consumption: estimations on Hungarian data. 31th CIRET Conference, Vienna.

Turkey Ministry of Culture and Tourism. [30.05.2018]. Turizm Istatistikleri. http://yigm.kulturturizm.gov.tr/TR,9851/turizm-istatistikleri.html

Vaughan, Liwen, and Esteban Romero-Frías. 2014. Web search volume as a predictor of academic fame: An exploration of Google trends. **Journal of the Association for Information Science and Technology.** 65.4: 707-720.

Vaughan, Liwen, and Rongbin Yang. 2013. Web traffic and organization performance measures: Relationships and data sources examined. **Journal of informetrics.** 7.3: 699-711.

Vicente, María Rosalía, Ana J. López-Menéndez, and Rigoberto Pérez. 2015. Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing?. **Technological Forecasting and Social Change.** 92: 132-139.

Von Graevenitz, Georg, et al. 2016. Does online search predict sales? Evidence from big data for car markets in Germany and the UK. **CCP (Centre for Competition Policy).** Working Paper 16-7.

Wijnhoven, Fons, and Olivia Plant. 2017. Sentiment Analysis and Google Trends Data for Predicting Car Sales. Thirty Eighth International Conference on Information Systems. South Korea.

Wikipedia. [23.05.2018]. Tourism in Turkey. https://en.wikipedia.org/wiki/Tourism_in_Turkey.

Wu, Lynn, and Erik Brynjolfsson. 2015. **The future of prediction: How Google searches foreshadow housing prices and sales**. Economic analysis of the digital economy. University of Chicago Press. 89-118.

Yang, Xin, et al. 2015.Forecasting Chinese tourist volume with search engine data. **Tourism Management.** 46: 386-397.

Zeybek, Ömer, and Erginbay UĞURLU. 2015. 2015. Nowcasting Credit Demand in Turkey with Google Trends Data. **Editorial Board**: 293.

Zeynalov, Ayaz. 2017. Forecasting Tourist Arrivals in Prague: Google Econometrics. **MPRA (Munich Personal RePEc Archive).** Paper No. 83268.

APPENDICES

Appendix I: Query selection results

Austria	France	
alanya	aeroport de turquie	vol pas cher turquie
alanya turkei	alanya turquie	vol pas cher turquie istanbul
antalya	all inclusive turquie	vol turquie
antalya turkei	antalya	vol turquie pas cher
antalya urlaub	billet turquie	voyage en turquie
antalya weter	bodrum	voyage en turquie pas cher
belek	bodrum en turquie	voyage organisao turquie
belek turkei	bodrum turquie	voyage turquie pas cher
belek wetter	carte turquie	voyage turquie tout compris
flufghafen antalya	club med turquie	voyages turquie
flughafen istanbul	hotel belek	
hotel titanic	hotel bodrum	
kappadokien	hotel bodrum turquie	
magic life turkei	hotel club turquie	
pegasos hotel turkei	hotel en turquie	
pegasos world turkei	hotel samara bodrum	
side	hotel sultan turquie	
side turkei	hotel turquie	
side wetter	hotel yali turquie	
turkei	istanbul turquie	
turkei all inclusive	kusadasi	
turkei antalya	la turquie	
turkei flughafen	marmara	
turkei hotel	marmara turquie	
turkei last minute	marmara turquie bodrum	
turkei resien	meteo en turquie	
turkei side	meteo turquie	
turkei urlaub	paris turquie	
turkei wetter	sejour turquie	
urlaub in der turkei	sejour turquie pas cher	
urlaub in turkei	side turquie	
urlaub turkei	turquie	
urlaub turkei all inclusive	turquie aeroport	
wetter belek	turquie carte	
wetter belek turkei	turquie hotel	
wetter in antalya	turquie meteo	
wetter in der turkei	turquie pas cher	
wetter in turkei	turquie sejour	
wetter side turkei	turquie voyage	
wikipedia turkei	vacances turquie	
Total= 40	**Total= 50**	

Germany	
ab in den urlaub turkei	turkei essen
alanya	turkei fluge
alanya hotel	turkei flughafen
alanya turkei	turkei hotel
alanya wetter	turkei hotel side
all inclusive turkei	turkei lara
antalya	turkei last minute
antalya flughafen	turkei reisen
antalya turkei	turkei side
belek	turkei urlaub
belek wetter	turkei urlaub gunsting
flug turkei	turkei wetter
flughafen alanya	urlaub in der turkei
flughafen antalya abflug	urlaub in turkei
flughafen antalya turkei	urlaub turkei
flughafen turkei	urlaub turkei all inclusive
flughafen turkei side	urlaub turkei buchen
holidaycheck turkei	urlaub wetter turkei
hotel antalya	wetter alanya turkei
hotel in der turkei	wetter belek
hotel in turkei	wetter belek turkei
hotel lara turkei	wetter in der turkei
hotel side	wetter in turkei
hotel turkei	wetter in turkei side
hotels turkei	wetter side
lara turkei	wetter turkei side
last minute urlaub turkei	
side	
side turkei	
side turkei hotel	
side turkei urlaub	
side wetter	
turkei	
turkei alanya	
turkei all inclusive urlaub	
turkei antalya	
turkei antalya hotel	
turkei belek	
turkei belek hotel	
turkei billing urlaub	
Total=66	

Netherlands	**Poland**	
all inclusive turkije	alanya	wczasy turcja 2017
goedkoop vliegen naar turkije	alanya turcja	wczasy turcja all inclusive
goedkope vakantie turkije	bodrum	wczasy turcja itaka
goedkope vliegtickets turkije	bodrum turcja	wczasy w turcji
het weer in turkije	forum turcja	wiza turcja
hotel turkije	hotel golden beach turcja	wycieczki turcja
klimaat turkije	itaka turcja	wycieczki fakultatywne turcja
last minute turkije	last minute turcja	
last minute vakantie turkije	pogoda turcja	
op vakantie naar turkije	pogoda turcja alanya	
rondreis turkije	pogoda turcja bodrum	
side turkije	pogoda w turcja	
temperatuur turkije	tanie wczasy turcja	
titanic hotel turkije	tui turcja	
turkije	turcja	
turkije all inclusive	turcja alanya	
turkije vakantie	turcja all inclusive	
turkije weer	turcja antalya	
vakantie in turkije	turcja bodrum	
vakantie naar turkije	turcja egejska	
vakantie turkije	turcja hotel	
vakantie turkije 2013	turcja hotele	
vakantie turkije all inclusive	turcja last minute	
visum turkije	turcja last minute all inclusive	
vliegen naar turkije	turcja lotnisko	
vliegen turkije	turcja mapa	
vliegticket turkije	turcja marmaris	
vliegtickets turkije	turcja pamukkale	
vliegvakantie turkije	turcja pogoda	
vlucht turkije	turcja wauluta	
weer in turkije	turcja wczasy	
weer lara turkije	turcja wczasy 2014	
weer turkije	wakacje pl turcja	
zonvakantie turkije	wakacje turcja	
	wczasy turcja	
	wczasy last minute turcja	
	wczasy turcja 2010	
	wczasy turcja 2011	
	wczasy turcja 2012	
	wczasy turcja 2015	
Total=34	**Total=47**	

Note: Tourism demand keywords obtained for PCA analysis using the second query selection method

Unemployment index
acik is ilanlari
anaokulu is ilanlari
antalya is ilanlari
eleman
fethiye is ilanlari
guvenlik is ilanlari
is ilanlari
is ilanlari
is ilanlari ankara
is ilanlari bodrum
is ilanlari bursa
is ilanlari istanbul
is ilanlarii
iskur
iskur is ilanlari
iss ilanlari elazig
izmir is ilanlari
kariyer
mersin is ilanlari
mugla is ilanlari
muhasebe is ilanlari
part time is
part time is ilanlari
samsun is ilanlari
yenibiris
Total: 25

Note: Unemployment keywords obtained for PCA analysis using the second query selection method

Renault	Fiat	Opel	Hyundai
satilik renault	satilik fiat	satilik opel	satilik hyundai
renult kangoo	satilik fiorino	sahibinden satilik araba opel	sahibinden satilik hyundai era
renault clio	sahibindensatilikfiat	sahibinden satilik opel	hyundai accent
satilik clio	sahibindenfiat	satilik astra	sahibinden satilik hyundai
sahibinden satilik renault	fiat doblo satilik	opel astra	sahibinden hyundai accent
satilik renault megane	fiat doblo	sahibinden satilik opel astra	sahibinden satilik araba hyundai
satilik renault kangoo	satilik doblo	satilik opel astra	sahibinden satilik hyundai accent
renault megane	fiat doblo sahibinden	satilik opel vectra	satilik hyundai era
satilik reno	satilik fiat fiorino	opel vectra	satilik hyundai accent era
renault clio sahibinden	fiat doblo sahibinden satilik	sahibinden opel astra	hyundai era
sahibinden satilik renault clio	fiat fiorino	opel corsa satilik	satilik hyundai getz
sahibinden renault	sahibinden fiat albea	sahibinden opel vectra	satilik2 hyundai
renault 12 satilik	fiat albea	satilik opel astra	hyundai getz
renault 12	satilik fiat albea	sahibinden satilik opel vectra	satilik hyundai 100
satilik renault 19	sahibinden satilik fiorino	sahibinden satilik opel corsa	hyundai kamyonet
renault 2. el	satilik albea	sahibinden opel corsa	satilik hyundai kamyonet
renault 19	fiat fiorino sahibinden	satilik araba opel astra	hyundai i20
renault symbol	fiat linea	satilik2 opel	hyundai h100
sahibinden renault kangoo	sahibinden satilik fiat fiorino	sahibinden satilik araba opel astra	satilik hyundai
satilik renault symbol	fiat palio	satilik opel combo	hyundai 2. el
	satilik palio	opel combo	
		sahibinden opel	
Total: 20	**Total:21**	**Total:22**	**Total:20**

Note: Car sales keywords obtained for PCA analysis using the second query selection method

Appendix II: PCA analysis

Tourism demand

Austria				
Principal components/correlation				**Number of obs.** = 136
				Number of comp. = 7
				Trace = 40
Rotation: (unrotated = principal)				**Rho** = 0.7457
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	17.7326	13.8634	0.4433	0.4433
Comp2	3.86924	1.40894	0.0967	0.5400
Comp3	2.46031	.742995	0.0615	0.6016
Comp4	1.71731	.16049	0.0429	0.6445
Comp5	1.55682	.195517	0.0389	0.6834
Comp6	1.3613	.231717	0.0340	0.7174
Comp7	1.12959	.198474	0.0282	0.7457

France				
Principal components/correlation				**Number of obs.**= 136
				Number of comp.= 7
				Trace = 50
Rotation: (unrotated = principal)				**Rho**= 0.7410
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	25.2448	20.6916	0.5049	0.5049
Comp2	4.55319	2.28468	0.0911	0.5960
Comp3	2.26851	.718532	0.0454	0.6413
Comp4	1.54997	.353979	0.0310	0.6723
Comp5	1.19599	.0223769	0.0239	0.6962
Comp6	1.17362	.107951	0.0235	0.7197
Comp7	1.06567	.11177	0.0213	0.7410

Germany				
Principal components/correlation				**Number of obs.** =136
				Number of comp.= 5
				Trace = 53
Rotation: (unrotated = principal)				**Rho**= 0.8527
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	27.476	17.7428	0.5184	0.5184
Comp2	9.73323	5.25273	0.1836	0.7021
Comp3	4.48051	2.34261	0.0845	0.7866
Comp4	2.1379	.772547	0.0403	0.8269
Comp5	1.36535	.512606	0.0258	0.8527

Netherlands				
Principal components/correlation				**Number of obs.** =136
				Number of comp.= 6
				Trace = 34
Rotation: (unrotated = principal)				**Rho** = .7091
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	14.6722	11.3959	0.4315	0.4315
Comp2	3.27631	.806555	0.0964	0.5279
Comp3	2.46976	.90798	0.0726	0.6005
Comp4	1.56178	.454673	0.0459	0.6465
Comp5	1.10711	.0857719	0.0326	0.6790
Comp6	1.02133	.0388986	0.0300	0.7091

Poland				
Principal components/correlation				**Number of obs.** =136
				Number of comp.= 8
				Trace = 47
Rotation: (unrotated = principal)				**Rho** = 0.8135
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	25.3692	21.7396	0.5398	0.5398
Comp2	3.62962	1.36265	0.0772	0.6170
Comp3	2.26696	.451361	0.0482	0.6652
Comp4	1.8156	.265318	0.0386	0.7039
Comp5	1.55028	.278383	0.0330	0.7368
Comp6	1.2719	.0464062	0.0271	0.7639
Comp7	1.2255	.122289	0.0261	0.7900
Comp8	1.10321	.241009	0.0235	0.8135

Unemployment

Unemployment				
Principal components/correlation				**Number of obs.**=132
				Number of comp.= 3
				Trace = 25
Rotation: (unrotated = principal)				**Rho** = 0.7711
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	15.9275	14.0411	0.6371	0.6371
Comp2	1.88639	.423162	0.0755	0.7126
Comp3	1.46323	.527438	0.0585	0.7711

Car sales

Renault				
Principal components/correlation				**Number of obs.**= 96
				Number of comp.= 2
				Trace = 20
Rotation: (unrotated = principal)				**Rho**= 0.7796
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	14.117	12.6413	0.7058	0.7058
Comp2	1.47561	.693699	0.0738	0.7796

Fiat				
Principal components/correlation				**Number of obs.**= 96
				Number of comp.= 1
				Trace=23
Rotation: (unrotated = principal)				**Rho**= 0.8569
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	19.7078	18.8482	0.8569	0.8569

Opel				
Principal components/correlation				**Number of obs.**= 96
				Number of comp.= 2
				Trace = 21
Rotation: (unrotated = principal)				**Rho** = 0.8482
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	16.61	15.4086	0.7910	0.7910
Comp2	1.20136	.305839	0.0572	0.8482

Hyundai				
Principal components/correlation				**Number of obs.**=96
				Number of comp.= 2
				Trace = 19
Rotation: (unrotated = principal)				**Rho**= 0.7899
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	13.7403	12.4718	0.7232	0.7232
Comp2	1.26847	.38249	0.0668	0.7899

Appendix III: ADF test

ADF results for tourism demand			
Tourism demand variables	**y**	**y (d=1, D=1)**	**Lags**
Austria	p-value = 0.99	p-value = 0.00	12
France	p-value = 0.96	p-value = 0.05	12
Germany	p-value = 0.88	p-value = 0.00	12
Netherlands	p-value = 0.78	p-value = 0.01	12
Poland	p-value = 0.28	p-value = 0.01	12
GT variables	**x**	**x (d=1, D=1)**	**Lags**
GTAustria	p-value = 0.57	p-value = 0.01	12
GTFrance	p-value = 0.99	p-value = 0.01	12
GTGermany	p-value = 0.95	p-value = 0.05	12
GTNetherlands	p-value = 0.99	p-value = 0.00	12
GTPoland	p-value = 0.95	p-value = 0.00	12
GTAustria-index	p-value = 0.91	p-value = 0.07	12
GTFrance-index	p-value = 0.98	p-value = 0.03	12
GTGermany-index	p-value = 0.52	p-value = 0.03	12
GTNetherlands-index	p-value = 0.97	p-value = 0.00	12
GTPoland-index	p-value = 0.63	p-value = 0.01	12

Note: ADF test is employed on all variables for the training set period. Lags are selected based on DFGLS test and AIC criteria

ADF results for unemployment			
Unemployment variables	**y**	**y (D=1)**	**Lags**
Total Unemployment	p-value = 0.057	p-value = 0.003	10
Youth Unemployment	p-value = 0.103	p-value = 0.000	10
GT variables	**x**	**x(D=1)**	
GTunemployment	p-value = 0.129	p-value = 0.013	12
GTunemploymenindex	p-value = 0.789	p-value = 0.024	12

Note: ADF test with drift is employed on all variables for the training set period. Lags are selected based on the DFGLS test and AIC criteria.

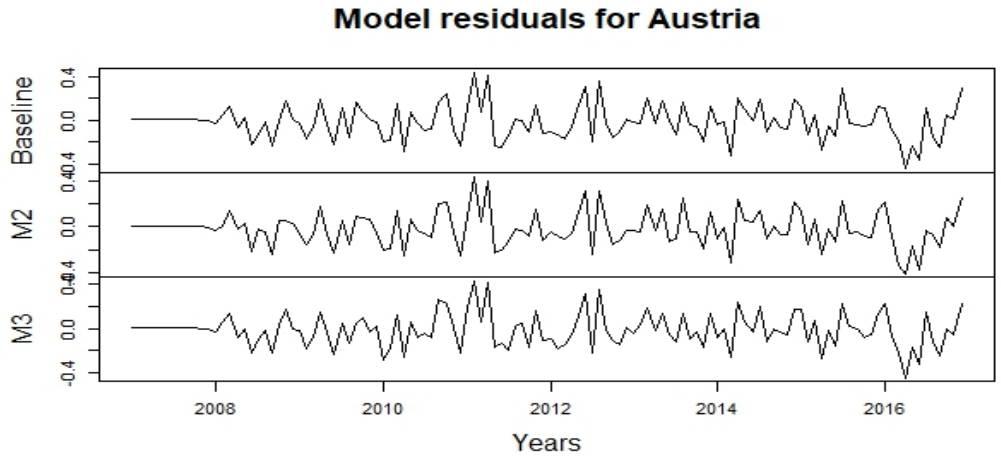
ADF results for car sales			
Car sales variables	**y**	**y (d=1, D=1)**	**Lags**
Renault	p-value = 0.05	p-value = 0.00	5
Fiat	p-value = 0.30	p-value = 0.06	11
Opel	p-value = 0.23	p-value = 0.07	11
Hyundai	p-value = 0.45	p-value = 0.06	11
GT variables	**x**	**x (d=1, D=1)**	**Lags**
GTRenault	p-value = 0.73	p-value = 0.00	4
GTFiat	p-value = 0.84	p-value = 0.00	3
GTOpel	p-value = 0.95	p-value = 0.02	9
GTHyundai	p-value = 0.83	p-value = 0.00	3
GTRenault-index	p-value = 0.80	p-value = 0.00	10
GTFiat-index	p-value = 0.90	p-value = 0.00	3
GTOpel-index	p-value = 0.83	p-value = 0.00	6
GTHyundai-index	p-value = 0.89	p-value = 0.00	3

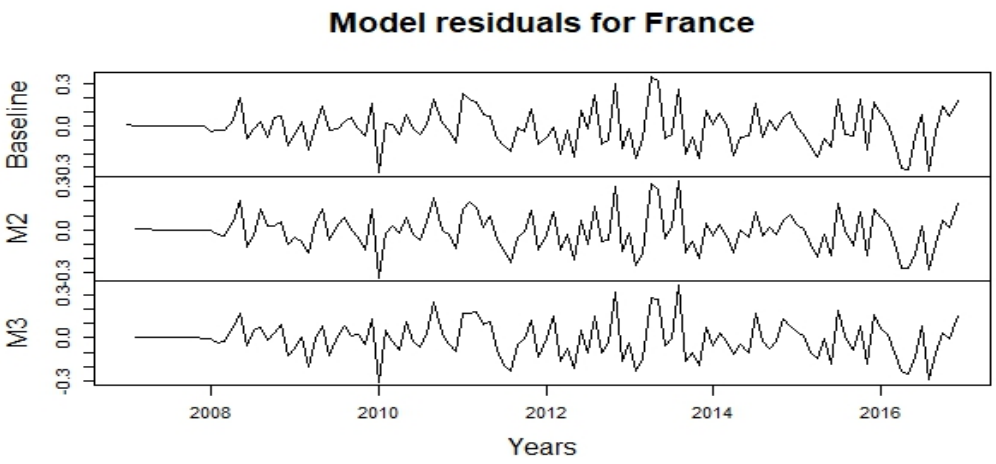
Note: ADF test is employed on all variables for the training set period. Lags are selected based on DFGLS test and AIC criteria.

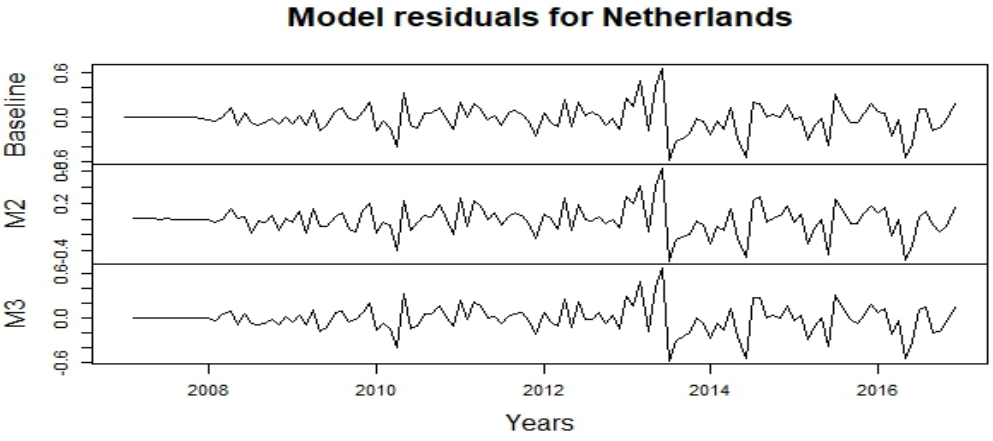
Appendix IV: Residuals and forecast graphs

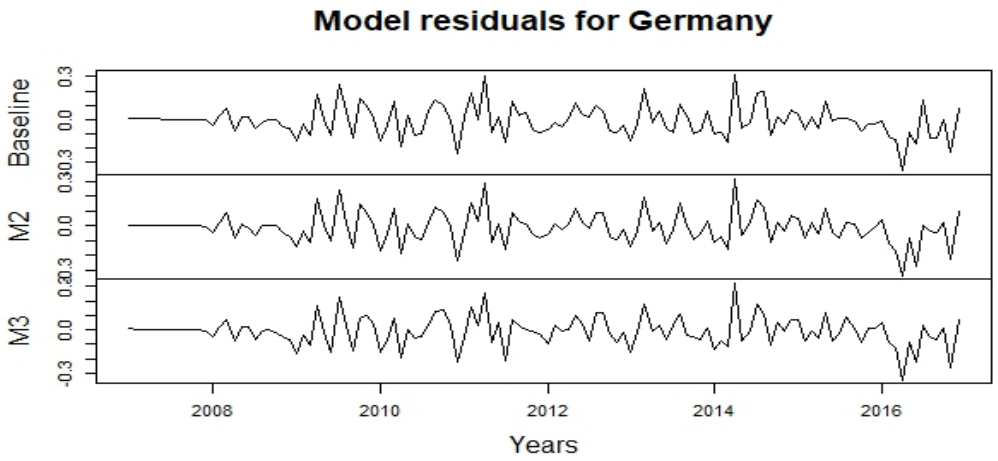
Tourism demand

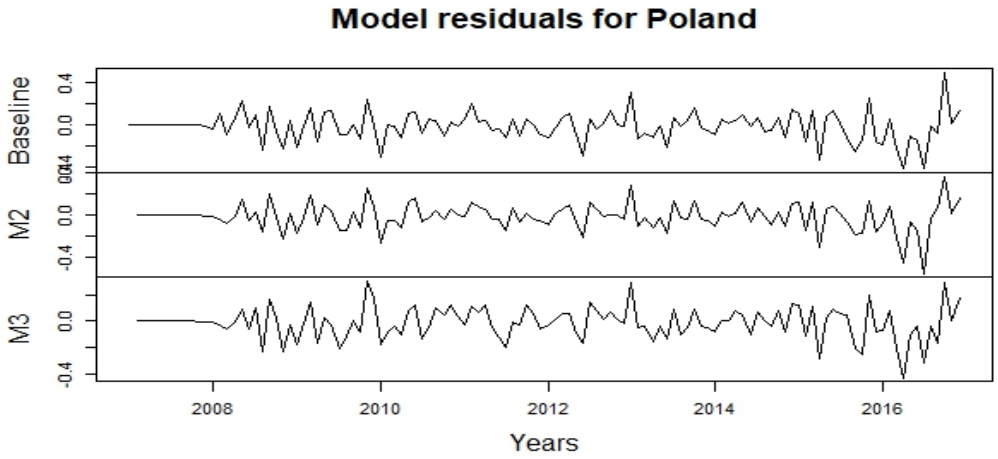
Residuals





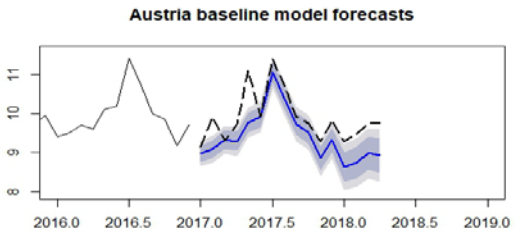


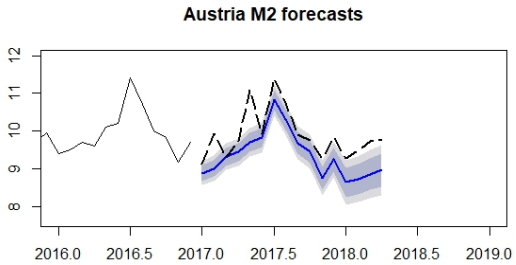


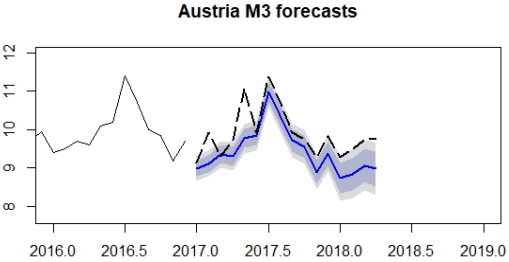


Forecast graphs

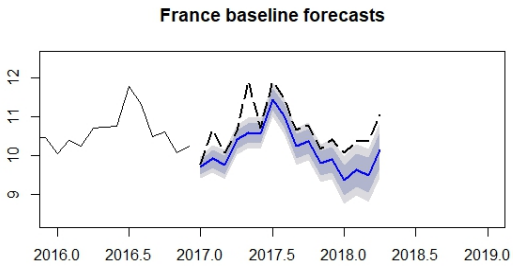
Austria

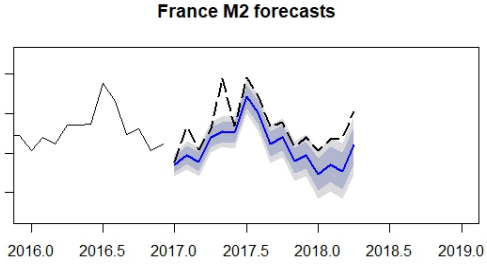


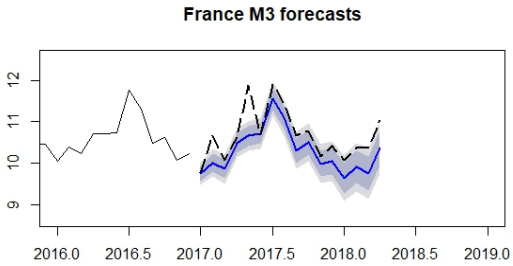




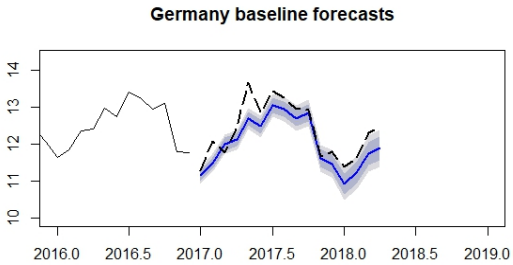
France

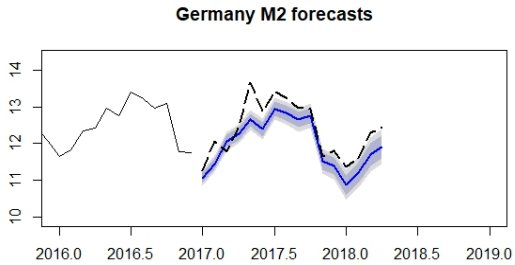


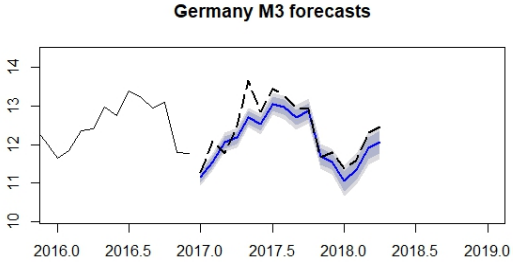




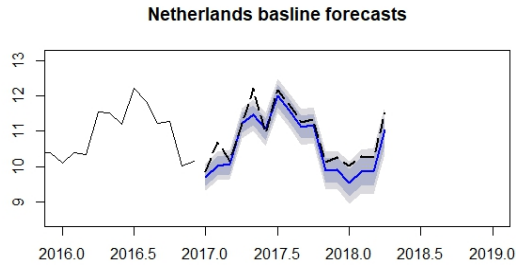
Germany

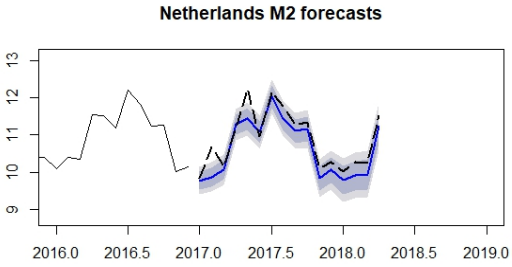


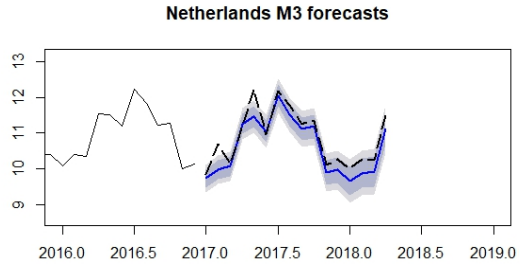




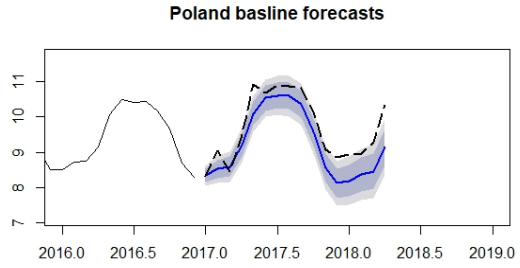
Netherlands

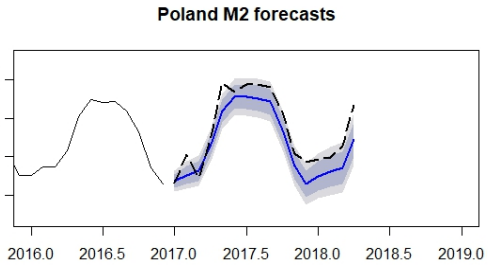


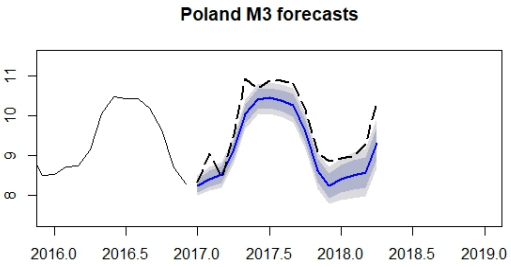




Poland

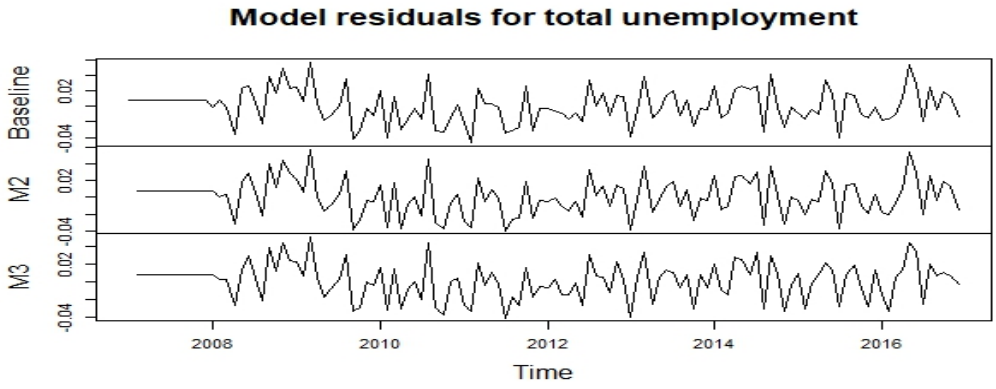


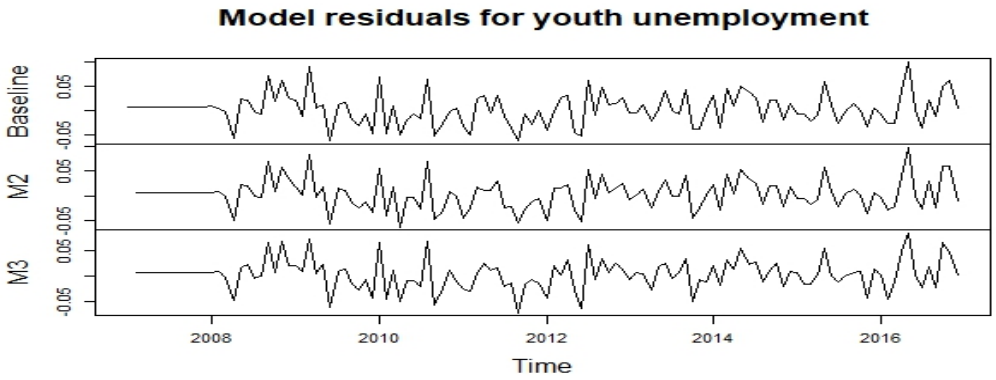




Unemployment

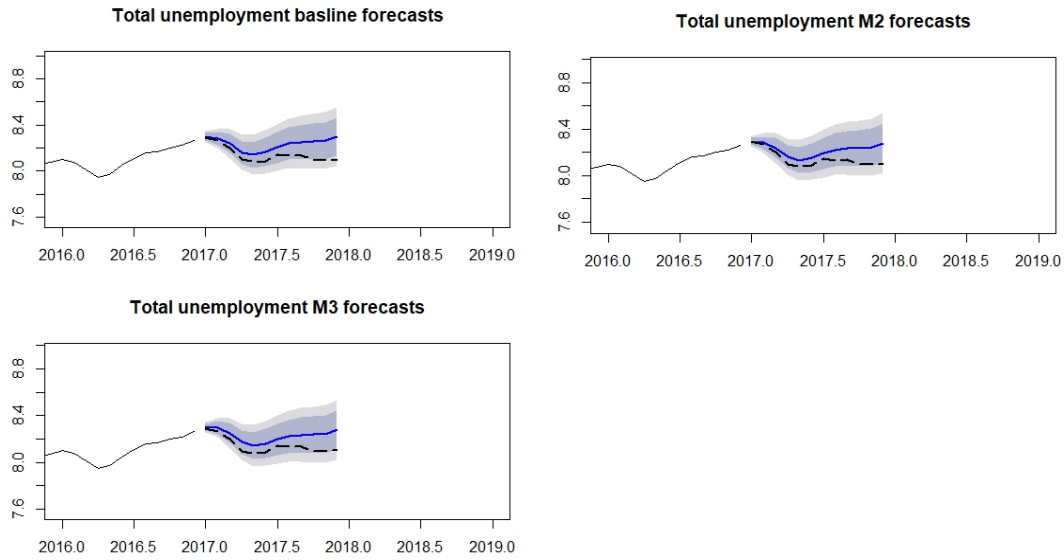
Residuals



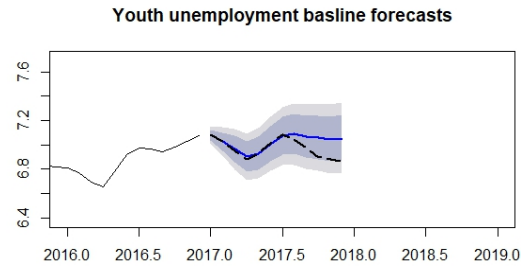


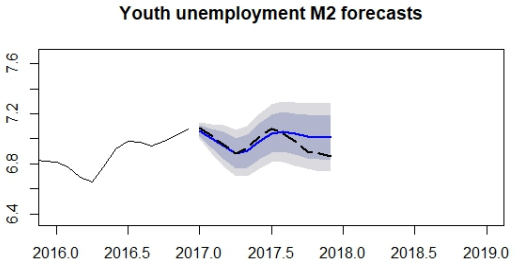
Forecast graphs

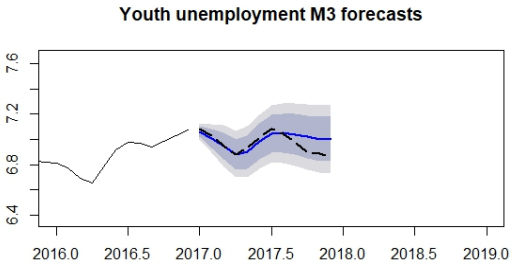
Total unemployment



Youth unemployment

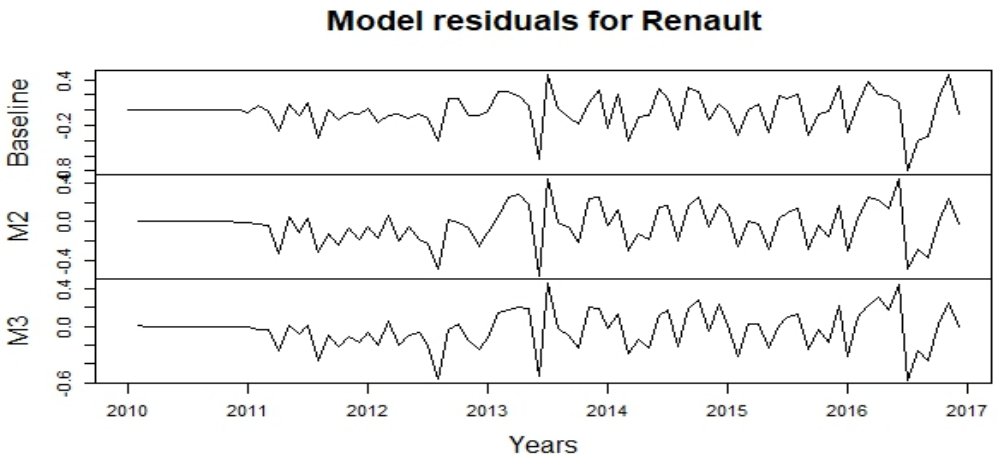


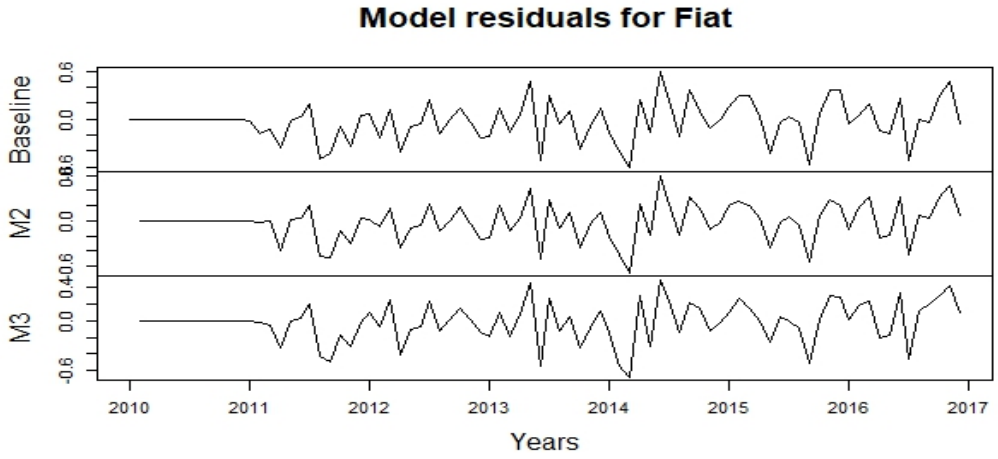


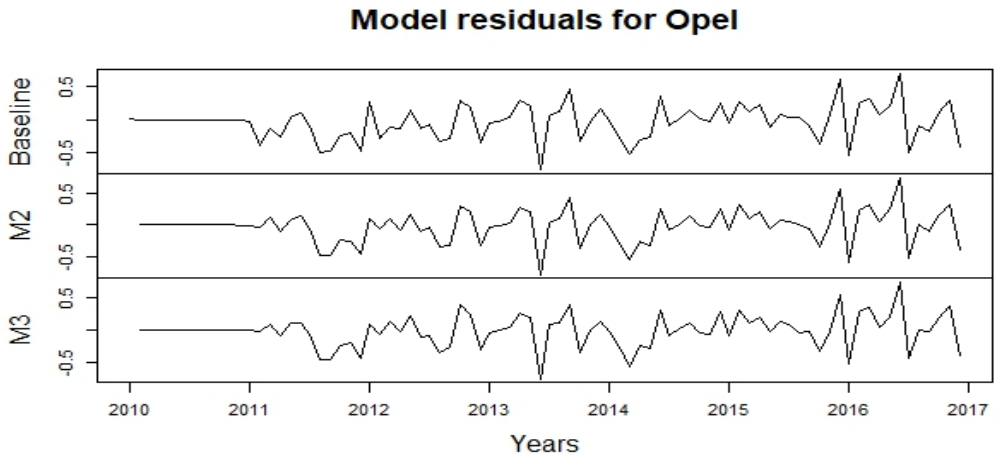


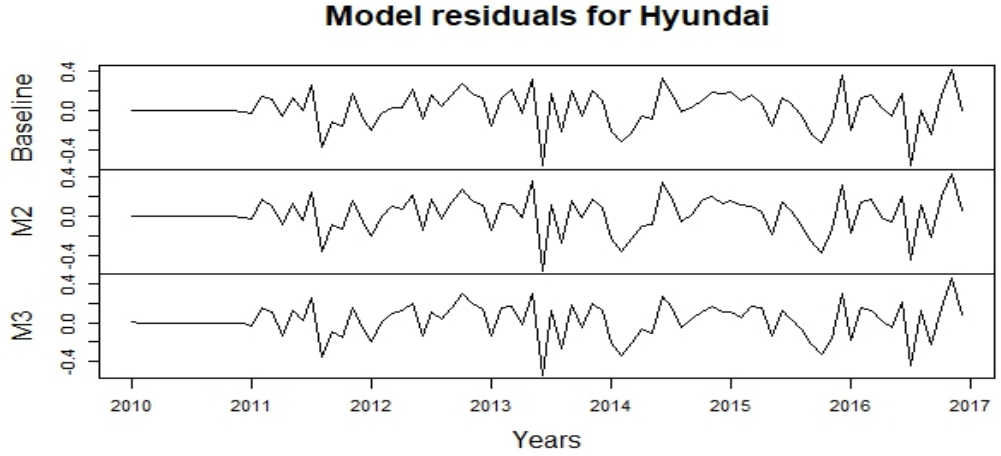
Car sales

Residuals



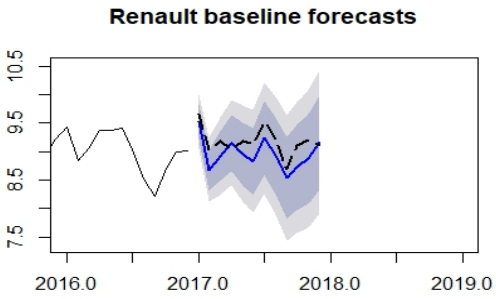

Model residuals for Fiat
Baseline
M2
M3
0.6
0.0
-0.6
2010
2011
2012
2013
2014
2015
2016
2017
Years

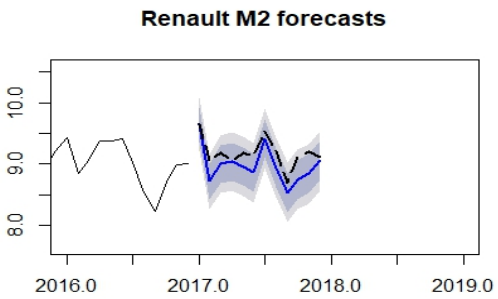

Model residuals for Opel
Baseline
M2
M3
0.5
-0.5
2010
2011
2012
2013
2014
2015
2016
2017
Years

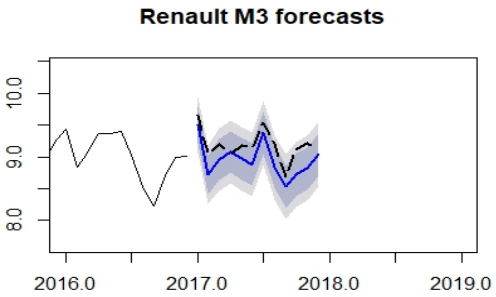

Model residuals for Hyundai
Baseline
M2
M3
0.4
0.0
-0.4
2010
2011
2012
2013
2014
2015
2016
2017
Years

Forecast graphs

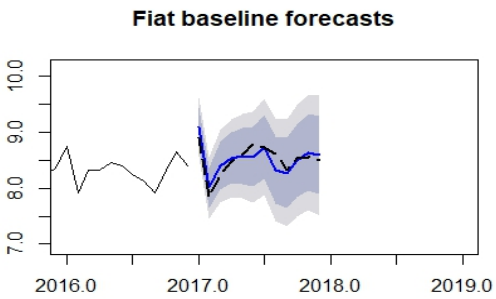
Renault

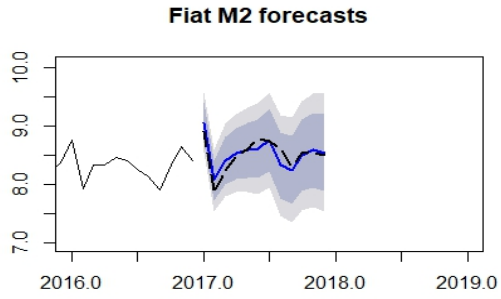


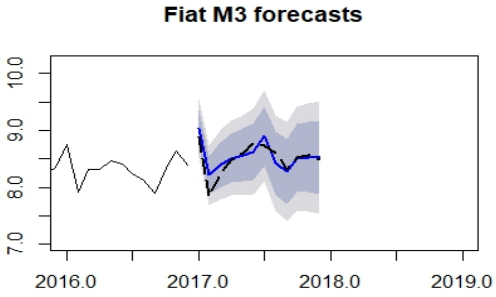




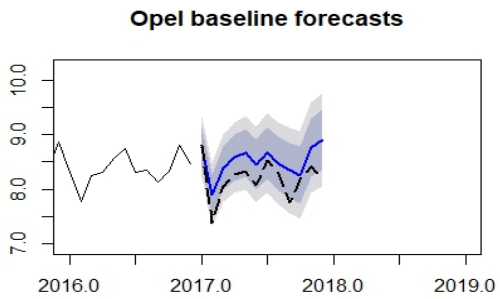
Fiat

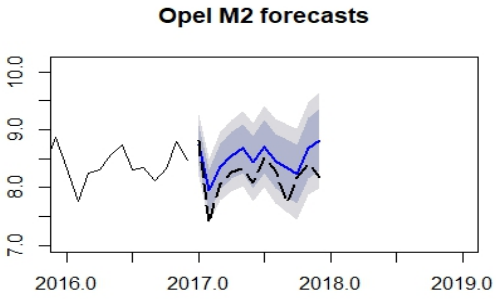


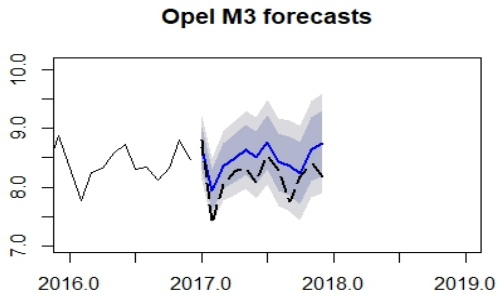




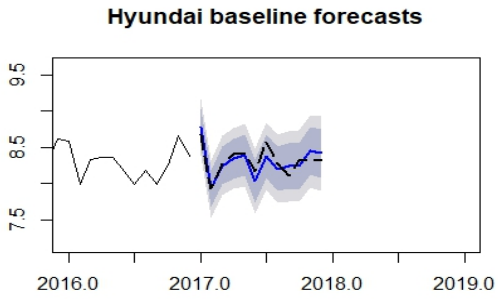
Opel

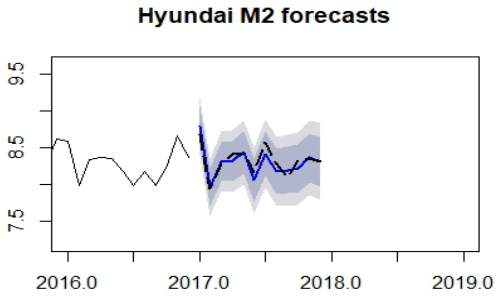


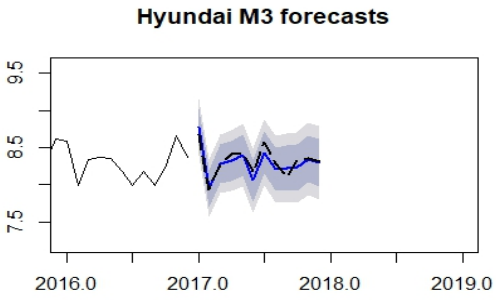




Hyundai







CURRICULUM VITAE (CV)

NAME AND SURNAME: Ahmed Omic

TR: +90 539 853 98 05

ahmed.omic@hotmail.com

Skype ahmed.omic92

Sex Male | Date of birth 04/05/1992 | Nationality Bosniak

WORK EXPERIENCE

11/09/2017–Present

Research Analyst

WAIPA, Istanbul
http://www.waipa.org

- Data collection and analysis
- Conducting qualitative and quantitative research in the field of FDI
- Coordination with stakeholders (IPAs, UNIDO, UNCTAD, World Bank)

01/06/2015–31/08/2015

Internship (Assurance)

EY (Ernst & Young), Sarajevo
http://www.ey.com/ba/en

- Analysing, reviewing and documenting financial statements of clients
- Learning EY audit policies

EDUCATION AND TRAINING

01/10/2016–26/10/2018

Master of Arts (MA) in Economics

Yildiz Technical University, Istanbul (Turkey)
www.yildiz.edu.tr

- Official language of teaching process: English language
- CGPA: 3.43

03/10/2011–13/06/2015

Bachelor of Arts (BA) in Economics

International University of Sarajevo, Sarajevo (Bosnia and Herzegovina)
www.ius.edu.ba

- Official language of teaching process: English language
- Achievements: Four times on Dean's honor list
- CGPA: 3.33

PERSONAL SKILLS

Mother tongue(s) Bosnian

Other language(s)

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
English	C1	C1	C1	C1	C1
Turkish	A2	A2	A2	A2	A2

Levels: A1/A2: Basic user - B1/B2: Independent user - C1/C2: Proficient user
Common European Framework of Reference for Languages

Computer skills

- Stata (Advanced level)
- R (Advanced level)
- SQL (Basic level)

ADDITIONAL INFORMATION

Seminars and certificates

- Data Science and Machine Learning Bootcamp with R, Udemy (2018)
- Managing Big Data with SQL, Coursera (2017)

Publications

- “The empowerment of Investment Promotion Agencies, their potential, and the importance of capacity building publication”, ICEC, 2017