MEF UNIVERSITY

THE PASSENGER LOAD FACTOR PREDICTION OF AIRLINE TRANSPORT

Capstone Project

Kalender Karakoç

İSTANBUL, 2017

MEF UNIVERSITY

THE PASSENGER LOAD FACTOR PREDICTION OF AIRLINE TRANSPORT

Capstone Project

Kalender Karakoç

Advisor: Associate Professor Şuayb Ş. Arslan

İSTANBUL, 2017

EXECUTIVE SUMMARY

THE PASSENGER LOAD FACTOR PREDICTION
OF AIRLINE TRANSPORT

Kalender Karakoç

Advisor: Associate Professor Şuayb Ş. Arslan

SEPTEMBER, 2017, 30

Turkish Airlines is one of the most preferred leading European air carriers with global network coverage thanks to its strict compliance with flight safety, reliability, product line, service quality and competitiveness. Turkish Airlines maintains its identity as the flag carrier of Turkey. Turkish Airlines, a Star Alliance member, is a four star airline with a fleet of over 300 aircraft flying to over 290 destinations around the world. Turkish Airlines has been named "Best Airline in Europe" at the 2016 Skytrax World Airline Awards. This is the sixth year in a row that the global airline has been awarded this prestigious award based on the votes of the traveling public. Turkish Airlines was also given the "Best Business Class Dining Lounge" and "Best Business Class On-Board Catering" for its service excellence.

With this study, it was aimed to forecast the passenger load factor by using the information of two years reservation, group sales data, calendar information, weekly dates, trend difference between current year and previous year , past load factor information, load factor information of the same period of the previous year. When each flight is thought to be its own characteristic, there is a need to find a solution for this work by a method that can reflect both the flight profile and the flight time dimension. Panel data regression method will be used for finding a solution of the problem.

When the flight has not yet departed, a preliminary structure for both the economy and the business class will be obtained. In terms of revenue management, it is expected to optimize income, change capacities, efficiency of flight routes, forecasts for special days and certain flight days and months.

Key Words: panel data; passenger load factor; capacity utilization

ÖZET

HAVA YOLU TAŞIMACILIĞINDA YOLCU YÜKÜNÜN TAHMİNİ

Kalender Karakoç

Tez Danışmanı: Doçent Şuayb Ş. Arslan

EYLÜL, 2017, 30

Türk Hava Yolları, uçuş emniyetine, güvenilirliğine, ürün gamına, hizmet kalitesine ve rekabet gücüne sıkı sıkıya bağlı kalmasıyla birlikte global şebeke kapsama alanına sahip en önde gelen Avrupalı hava taşıyıcılarından biridir ve Türkiye'nin bayrak taşıyıcısı olarak kimliğini sürdürmektedir. Star Alliance üyesi olan Türk Hava Yolları, dünyanın dört bir yanındaki 290 noktaya uçan 300'den fazla uçağın bulunduğu dört yıldızlı bir havayolu şirketidir. Türk Hava Yolları, 2016 Skytrax Dünya Hava Yolu Ödülleri'nde "Avrupa'nın En İyi Havayolu" seçildi. Bu küresel hava yolu şirketi, seyahat eden halkın oylarına dayanan bu prestijli ödülü altı yıldır almaktadır. THY'ye servis mükemmellik için "En İyi İş Sınıfı Yemek Salonu" ve "En İyi İş Sınıfı Araç Üstü Yemek Servisi" verildi.

Bu çalışmada, iki yıllık rezervasyon, grup satış verileri, takvim bilgileri, haftalık tarihler, güncel yıl ile geçmiş yılın değişim trendi, geçmiş yük faktörü bilgileri ve bir önceki yılın aynı döneminin yük faktörü bilgileri bilgilerini kullanarak yolcu yük faktörünü tahmin etmeyi amaçladık. Her uçuşun kendi karakteristiği olduğu düşünülürse, hem uçuş profilini hem de uçuş süresi boyutunu yansıtabilen bir yöntemle bu problem için bir çözüm bulma ihtiyacı vardır. Problemin çözümü için panel veri regresyon yöntemi kullanılacaktır.

Uçuş henüz gerçekleşmeden hem ekonomi hem de iş seyahati sınıfı için bir ön yapı elde edilecektir. Gelir yönetimi açısından, geliri optimize etmek, kapasiteleri değiştirmek, uçuş rotalarının etkinliği, özel günler ve belirli uçuş günleri ve aylar için tahminler yapılması beklenmektedir.

Anahtar Kelimeler: Yolcu yükü tahmini, panel veri regresyon, uçuş kapasitesi

TABLE OF CONTENTS

TABLES

FIGURES

1. INTRODUCTION

The global airline industry serves each and every country of the world. This industry is important for the growth of global economy, supports mutual economic integration and has an important influence on facilitating the financial dynamics worldwide (Belobaba, P., Odoni, A. And Barnhart, C.,2009).

A PLF (Passenger load factor) measures the capacity utilization for airlines. It indicates the efficiency of the airline; filling seats and generating revenues (Bruckner and Whalen, 2000). Passenger load factor is an important key performance indicator for the performance of any transport system. It measures capacity utilization transport services like airlines, railways, bus services. If an airline had charged the same price for each seat on a flight, it would not have made the highest of their revenue. In order for a flight to be productive, both the seats should be available, and the plane should be as full as possible. It is fully concerned with supply and demand optimization. Just in this case passenger load factor is used as a metric. It is important to adjust the pricing accordingly by managing the demand in a balanced way. At this stage, the Passenger Load factor provides an important insight. Typically, 80% of PLF is considered as the standard in the domestic airline industry. The improvement of PLF has a direct impact on the revenue management. The bettering of the PLF affects the plan and costs of complementary functions; such as workforce, fuel, catering and ground services.

PLF is highly affected by seasonality, unpredictable demand and even political & economical issues. Considering that the airline industry dynamics include cancellations of reservations, multi-leg flights complexity and the openings of new flight routes, forecasting PLF becomes even harder. Yet it is attainable by building airline-specific models to forecast the aggregate passenger traffic in a certain time frame, region or an individual flight. The model delivers an optimum revenue and enables the business units to cleverly act on price and campaigns down the road.

PLF is a percentage of seats filled by passengers and is a complex metric to compute. It is generally used to measure performance and sufficiency. In the past, stochastic models are developed for obtaining load factor, which is best fitted trend for Europe's North Atlantic and Mid-Atlantic flights in the Association of European Airlines (Tesfay, Y. Y, 2016). Load factor has periodic and serial correlations. For periodic and serial correlations, two different models

are developed. Dynamic time effects for periodic correlation and the Prais-Winsten methodology for serial correlations were added later in the models (Tesfay, Y. Y, 2016).

A choice model of consumer behavior (buy-up, buy down etc.) is a critical problem for revenue management. A single-leg reserve management is analyzed and many trials have been done to understand the impact of consumer behavior and develop heuristic methods. Buy-up and buy-down is considered to be an important phenomenon which shall have an effect on the final revenue optimization (Croissant, Y. and Millo, G., 2008).

1.1. Statement of The Need and Aim of This Project

PLF is calculated by dividing Revenue Passenger Kilometers (RPK) by the Available Seat Kilometers (ASK). Hence, assuming the capacity of an airline remains the same, an increase in RPK is directly proportional to an increase in PLF. Its formula is given by

PLF = Revenue Passenger Kilometers / Available Seat Kilometers

Turkish Airlines uses PLF for capacity utilization and revenue management. The set of objectives of this project can be itemized as follows.

- After the flights are opened for sale, THY want to predict the PLF on boarding times,
- Understanding the seasonality,
- Defining the special days effects,
- Defining the past flights behaviors,
- Board-off, country, area based predictions,
- Cabin class based PLF predictions

1.2. Project Inputs and Concepts

PLF shortly called Load Factor (LF) is a key performance indicator of airline's carrying capacity. It is most commonly used to describe the performance of the airline. Having high load rates and capacity utilization is essential for the airline's profitability. In this study, load factor will be anticipated by four aspects. First one is the past load factor rates. Second one is the seat sold on the remaining days. Third one is calendar including day, month, year effects. Last one is official holidays.

2. LITERATURE REVIEW

Load factor is a metric that measures the success of an airline´s capacity and demand management efforts. In order to manage ticket sales demands on airway transportation, ticket sales are accomplished gradually in each route plan. Thus, with this gradual sale, the price and demand balance is taken into consideration.The airline will adjust the number of seats allocated to each fare class. When one class has been sold, the sale price will leap to the next one. Supply can only be produced in units equivalent to the capacity of whichever aircraft type is available to operate the flight-legs and routes. They are designed to serve targeted origin and destination markets and is broadly fixed in the short run. Moreover, the requirements to uphold both the high flight completion rate and integrity of network connections and aircraft and crew assignments might impede a scheduled passenger carrier from cancelling a striking number of its lightly loaded flights (Bruckner and Whalen, 2000).

Load factor is the issue of revenue management and one of the main metrics used to maximize revenue in the aviation sector. The Load Factor, which is directly related to the airline's revenue management, is also an important driver for tariff strategy, route planning, active capacity management. The achievement of the airline is considered to make revenues higher than its unit costs with key performance indicators (i.e., the product of yield and load factor).

In airline operation, it is very important to have knowledge of the future values of the load factor that is agreed upon to identify the line. There are many factors in determining this line. One of them is that aviation sector operations are managed in a loop and load factor is not stable within the aviation sector (Ľubomír and Jakub, 2013). Another one is airlines implement dynamic capacity management to meet and absorb demand and can suddenly change the type of aircraft (Bertsimas and Popescu, 2003; Li et al., 2007). The intense competition between the end-market free-market airlines and the Online Reservation System (ORS) makes the customer strong. This leads to dynamic decisions governed by airway efficiency affecting the load factor (Doganis, 2010; Cento, 2009; Kahn, 1988).

Two important stochastic numerical methods are quite instrumental in forecasting studies: (1) Time series and (2) regression analysis. There may be variable values for other nominal variables within each time unit in time series runs. These nominal variables are also called factor. For example, region, country, occupation, province etc. . For example, income

per capita, unemployment rates in different countries, income on a job basis, and so on. A time series analysis does not adequately measure the factor effect. Regression analysis on the other hand does not measure the dynamic effect of external variables on internal variables. Panel data seems to be very appropriate to solve these two problems of influence (Pearl, 2000). The reason why the panel data model is suitable for solving these two problems is explained in Section 2.1 with the benefits of the panel data model.

2.1. Panel Data Benefits

Panel data allows individual heterogeneity control. Panel data suggests that every flight, flight destination, flight departures and arrivals are heterogeneous. (Hsiao, 2003). Time-series and cross-section studies not controlling this heterogeneity run the risk of obtaining biased results, e.g. see Moulton (1986, 1987). Baltagi and Levin (1992) consider cigarette demand across 46 American states for the years 1963–88. Consumption is modeled as a function of lagged consumption, price and income. These variables vary with states and time.

Panel data gives more informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency (Baltagi, B. 2008). In airline data example, remaining days of the departure within the reservation sales data have high collinearity. In the case of demand for cigarettes, there is high collinearity between price and income in the aggregate time series. Panel data works better to study the dynamics of adjustment. Although cross-sectional distributions are stable, they contain lots of changes.

Unlike cross-sections, it shows that it includes changes for individuals or households. It allows us to observe how the individual living standards change during the development process. It enables us to determine who is benefiting from development. It also allows us to observe whether poverty and deprivation are transitory or long-lived, the income-dynamics question. Panels are also necessary for the estimation of intertemporal relations, lifecycle and intergenerational models. In fact, panels can relate the individual's experiences and behavior at one point in time to other experiences and behavior at another point in time. For example, the lack of evaluation of training programs observes the absence of a group of participants and participants before and after the implementation of the training program (Deaton 1995).

3. DATA SOURCES

We will try to create a metadata by putting together multiple sets of data for use in the modeling phase. Our primary and most important source of data will be our ***lag data frame***, which holds all the past 365 days of sales data until the date of departure of a flight. The second source will be group sales data. There is information on how much of the booking sales data comes from group sales. Third, 365 days lag of load factor information for completed flights will be included. This is the fourth calendar data, and the month, year, and day of week dummy variables will be derived.

3.1. Reservation Lag Data Frame

Reservation lag data frame contains 11 different board area, 118 different countries, 314 different airport information. ***Reservation lag data frame*** is a daily time-data frame format from the first day of booking to the last day. We will not be able to observe the variation of different flight areas, countries and airports in the modeling phase in this format as the time series modelling. For this reason, we will talk about how this data will be converted into panel data format at a later stage.

Table 1:Reservation Lag Data Frame

ID_ORIGIN_YMD	ID_FLIGHT_NUMBER	ID_BOARD_POINT	ID_OFF_POINT	ID_COMPARTMENT	Remaining Day	Seat_Sold
20160101	2328	IST	ADB	Y	0	138
20160101	2328	IST	ADB	Y	1	130
20160101	2328	IST	ADB	Y	2	118
20160101	2328	IST	ADB	Y	3	109
20160101	2328	IST	ADB	Y	4	88
20160101	2328	IST	ADB	Y	5	75

3.2. Group Sales Data Frame

Group sales as can be seen in Table 2, contains 11 different board area, 118 different countries, 314 different airport information. ***Group sales data frame*** is a daily time-table format from the first day of booking to the last day. The group sales table is a variant that directly affects the load factor of flight data. As with individual purchases, they occur in bulk during flights. These collective purchases have an upward effect on flight overtime.

Table 2: Group Sales Data Frame

ID_ORIGIN_YMD	ID_FLIGHT_NUMBER	ID_BOARD_POINT	ID_OFF_POINT	ID_COMPARTMENT	Remaining Day	Group_Seat_Sold
20160101	2328	IST	ADB	Y	0	138
20160101	2328	IST	ADB	Y	1	130
20160101	2328	IST	ADB	Y	2	118
20160101	2328	IST	ADB	Y	3	109
20160101	2328	IST	ADB	Y	4	88
20160101	2328	IST	ADB	Y	5	75

3.3. Lag of Load Factor Information

The Table 3 is derived from ***Reservation Lag Data Frame***. Although booking information is very effective at the launch of a plane, the general departure rates of the same plane are also important. To give an example, let's consider two different flights with 120 days to departure. Consider that 95% of these flights are predominantly located in the province, while the rest remain at around 70%. We expect that the model we developed will give different results for these two flights. The data that will provide us with this information is the ***lag of load factor information***.

Table 3: Lag of Load Factor Data Frame

ID_ORIGIN_YMD	ID_FLIGHT_NUMBER	ID_BOARD_POINT	ID_OFF_POINT	ID_COMPARTMENT	LF
20160101	2328	IST	ADB	Y	0.844
20160102	2328	IST	ADB	Y	0.882
20160103	2328	IST	ADB	Y	0.973
20160104	2328	IST	ADB	Y	0.844
20160105	2328	IST	ADB	Y	0.553
20160106	2328	IST	ADB	Y	0.439
20160108	2328	IST	ADB	Y	0.826
20160109	2328	IST	ADB	Y	0.726
20160110	2328	IST	ADB	Y	0.939
20160111	2328	IST	ADB	Y	0.769

3.4. Calendar Data Frame

Calendar data frame is a data frame showing the date of the flight, the day of the week, and the month. It should be noted with this data frame that the official holiday information of Turkey is in not excluded. It is assumed that the model will also have the effect of these special days. At work comprehension meetings held, the main subject matter experts agree that aircraft flying on special days are definitely higher than those flying on other days. (Wesonga, R., Nabugoomu, F., & Masimbi, B. 2013).

4. DATA PREPARATION AND METADATA CREATION

4.1. Data Preparation

We start the data preparation process with the Reservation Lag Table. The first thing we do is to start by selecting specific reservation days to use in the panel format from the time series format. If we wanted to create a time series model with all reservation days, we would have to build a model of about 1.8 billion rows of data. At this point, we would not have to wait for days for a model with a single server that does not distribute such a model with classic data frame-like structures. Together with developing technologies, we plan to solve such a constraint in future studies. Distributed computing systems or cloud computing systems are among our future efforts to process data. The following reservation dates have been selected to avoid and limit such a constraint.

("RES1","RES2","RES3","RES4","RES7","RES10","RES15","RES20","RES30","RES45","RES60","RES90","RES120","RES340")

RES1 : *1 Day Remaining Flight Load Factor*
RES2 : *2 Days Remaining Flight Load Factor*
RES3 : *3 Days Remaining Flight Load Factor*
RES4 : *4 Days Remaining Flight Load Factor*
RES7 : *7 Days Remaining Flight Load Factor (between 4 days and 7 days remaining)*
RES10 : *10 Days Remaining Flight Load Factor (between 8 days and 10 days remaining)*
RES15 : *15 Days Remaining Flight Load Factor (between 11 days and 15 days remaining)*
RES20 : *20 Days Remaining Flight Load Factor (between 11 days and 15 days remaining)*
RES30 : *30 Days Remaining Flight Load Factor (between 21 days and 30 days remaining)*
RES45 : *45 Days Remaining Flight Load Factor (between 31 days and 45 days remaining)*
RES60 : *60 Days Remaining Flight Load Factor (between 46 days and 60 days remaining)*
RES90 : *90 Days Remaining Flight Load Factor (between 61 days and 90 days remaining)*
RES120 : *120 Days Remaining Flight Load Factor (between 91 days and 120 days remaining)*
RES340 : *340 Days Remaining Flight Load Factor (between 121 days and 340 days remaining)*

The days starting with the *RES* we have mentioned above indicate the number of days left on the flight. So *RESn* indicates the data for *n* days before the flight. The seat load values for these selected days are calculated by the load factor calculation below and the booking load factor rates for each booking day are calculated. With such a transformation, a unique structure is provided for each departure day.

$$RES1 = \sum_{i=1}^{t} \left(\frac{(Number\ of\ Carried\ Passenger(RES1)_t * Distance_t)}{(Available\ Seat)_t * Distance_t}\right)$$

$$RES2 = \sum_{i=1}^{t} \left(\frac{(Number\ of\ Carried\ Passenger(RES2)_t * Distance_t)}{(Available\ Seat)_t * Distance_t}\right)$$

•

$$RES340 = \sum_{i=1}^{t} \left(\frac{(Number\ of\ Carried\ Passenger(RES340)_t * Distance_t)}{(Available\ Seat)_t * Distance_t}\right)$$

After converting the data we gathered, the data frame structure shall look like as shown below.

```
$ ID_ORIGIN_YMD   : int  20160911 20161203 20161008 20160118 20160720 20161201 20160414 20160220 20161030 20160329 ...
$ ID_FLIGHT_NUMBER: chr  "2054" "7001" "2943" "669" ...
$ ID_BOARD_AREA   : chr  "DO" "DO" "DO" "EW" ...
$ DS_BOARD_AREA   : chr  "TURKIYE" "TURKIYE" "TURKIYE" "WESTERN AFRICA" ...
$ ID_OFF_AREA     : chr  "DO" "DO" "DO" "EW" ...
$ DS_OFF_AREA     : chr  "TURKIYE" "TURKIYE" "TURKIYE" "WESTERN AFRICA" ...
$ ID_BOARD_CNTRY  : chr  "TR" "TR" "TR" "CM" ...
$ DS_BOARD_CNTRY  : chr  "TURKEY" "TURKEY" "TURKEY" "CAMEROON" ...
$ ID_OFF_CNTRY    : chr  "TR" "TR" "TR" "CM" ...
$ DS_OFF_CNTRY    : chr  "TURKEY" "TURKEY" "TURKEY" "CAMEROON" ...
$ ID_BOARD_POINT  : chr  "IST" "ADB" "BJV" "NSI" ...
$ ID_OFF_POINT    : chr  "KZR" "ESB" "SAW" "DLA" ...
$ ID_COMPARTMENT  : chr  "Y" "Y" "C" "Y" ...
$ RES1            : num  0.78 0.698 0.333 0.541 0.83 0.063 0.9 0.733 0.75 0.063 ...
$ RES2            : num  0.735 0.624 0.417 0.548 0.803 0 0.667 0.661 0.75 0.25 ...
$ RES3            : num  0.682 0.571 0.417 0.504 0.707 0 0.6 0.57 0.75 0 ...
$ RES4            : num  0.606 0.545 0.417 0.481 0.66 0 0.6 0.436 0.75 0 ...
$ RES7            : num  0.553 0.492 0.417 0.459 0.524 0 0.453 0.297 0.625 0 ...
$ RES10           : num  0.515 0.349 0.167 0.437 0.429 0 0.36 0.236 0.625 0 ...
$ RES15           : num  0.402 0.249 0 0.37 0.401 0 0.26 0.206 0.625 0 ...
$ RES20           : num  0.318 0.175 0 0.326 0.388 0 0.1 0.188 0.5 0 ...
$ RES30           : num  0.091 0.058 0 0.259 0.306 0 0.067 0.188 0.563 0 ...
$ RES45           : num  0.038 0 0 0.252 0.224 0 0 0.179 0.313 0 ...
$ RES60           : num  0.038 0 0 0.185 0.122 0 0 0.173 0.188 0.188 ...
$ RES90           : num  0.008 0 0 0.104 0.085 0 0 0.167 0.063 0 ...
$ RES120          : num  0 0 0 0.074 0.03 0 0 0.154 0 0 ...
$ RES340          : num  0 0 0 0.007 0 0 0 0 0 0 ...
```

4.2. Merging of Data Frames and Metadata Creation

Once the above data frame is created, we can move to the metadata creation stage. In this phase, combined with the group sales data frame, the lag of load factor information, and the calendar data frame, final metadata will be created for analysis and modeling. The final metadata structure will be as follows.

```
$ ID_ORIGIN_YMD   : int  20160911 20161203 20161008 20160118 20160720 20161201 20160414 20160220 20161030 20160329 ...
 $ ID_FLIGHT_NUMBER: chr  "2054" "7001" "2943" "669" ...
 $ ID_BOARD_AREA   : chr  "DO" "DO" "DO" "EW" ...
 $ DS_BOARD_AREA   : chr  "TURKIYE" "TURKIYE" "TURKIYE" "WESTERN AFRICA" ...
 $ ID_OFF_AREA     : chr  "DO" "DO" "DO" "EW" ...
 $ DS_OFF_AREA     : chr  "TURKIYE" "TURKIYE" "TURKIYE" "WESTERN AFRICA" ...
 $ ID_BOARD_CNTRY  : chr  "TR" "TR" "TR" "CM" ...
 $ DS_BOARD_CNTRY  : chr  "TURKEY" "TURKEY" "TURKEY" "CAMEROON" ...
 $ ID_OFF_CNTRY    : chr  "TR" "TR" "TR" "CM" ...
 $ DS_OFF_CNTRY    : chr  "TURKEY" "TURKEY" "TURKEY" "CAMEROON" ...
 $ ID_BOARD_POINT  : chr  "IST" "ADB" "BJV" "NSI" ...
 $ ID_OFF_POINT    : chr  "KZR" "ESB" "SAW" "DLA" ...
 $ ID_COMPARTMENT  : chr  "Y" "Y" "C" "Y" ...
 $ RES1            : num  0.78 0.698 0.333 0.541 0.83 0.063 0.9 0.733 0.75 0.063 ...
 $ RES2            : num  0.735 0.624 0.417 0.548 0.803 0 0.667 0.661 0.75 0.25 ...
 $ RES3            : num  0.682 0.571 0.417 0.504 0.707 0 0.6 0.57 0.75 0 ...
 $ RES4            : num  0.606 0.545 0.417 0.481 0.66 0 0.6 0.436 0.75 0 ...
 $ RES7            : num  0.553 0.492 0.417 0.459 0.524 0 0.453 0.297 0.625 0 ...
 $ RES10           : num  0.515 0.349 0.167 0.437 0.429 0 0.36 0.236 0.625 0 ...
 $ RES15           : num  0.402 0.249 0 0.37 0.401 0 0.26 0.206 0.625 0 ...
 $ RES20           : num  0.318 0.175 0 0.326 0.388 0 0.1 0.188 0.5 0 ...
 $ RES30           : num  0.091 0.058 0 0.259 0.306 0 0.067 0.188 0.563 0 ...
 $ RES45           : num  0.038 0 0 0.252 0.224 0 0 0.179 0.313 0 ...
 $ RES60           : num  0.038 0 0 0.185 0.122 0 0 0.173 0.188 0.188 ...
 $ RES90           : num  0.008 0 0 0.104 0.085 0 0 0.167 0.063 0 ...
 $ RES120          : num  0 0 0 0.074 0.03 0 0 0.154 0 0 ...
 $ RES340          : num  0 0 0 0.007 0 0 0 0 0 0 ...
 $ GROUP_LF        : num  0 0 0 0 0 0 0 0.152 0 0 ...
 $ WAIT_COUNT      : num  0 0.243 0 0.015 0 0 0 0 0.125 0 ...
 $ LF_365          : num  0.879 0.757 0 0 0.886 0.167 0 0 0.188 0 ...
 $ month1          : num  0 0 0 1 0 0 0 0 0 0 ...
 $ month10         : num  0 0 1 0 0 0 0 0 1 0 ...
 $ month11         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ month12         : num  0 1 0 0 0 1 0 0 0 0 ...
 $ month2          : num  0 0 0 0 0 0 0 1 0 0 ...
 $ month3          : num  0 0 0 0 0 0 0 0 0 1 ...
 $ month4          : num  0 0 0 0 0 0 1 0 0 0 ...
 $ month5          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ month6          : num  0 0 0 0 0 0 0 0 0 0 ...
```

$ month7 : num 0 0 0 0 1 0 0 0 0 0 ...

$ month8 : num 0 0 0 0 0 0 0 0 0 0 ...

$ month9 : num 1 0 0 0 0 0 0 0 0 0 ...

$ dow1 : num 1 0 0 0 0 0 0 0 1 0 ...

$ dow2 : num 0 0 0 1 0 0 0 0 0 0 ...

$ dow3 : num 0 0 0 0 0 0 0 0 0 1 ...

$ dow4 : num 0 0 0 0 1 0 0 0 0 0 ...

$ dow5 : num 0 0 0 0 0 1 1 0 0 0 ...

$ dow6 : num 0 0 0 0 0 0 0 0 0 0 ...

$ dow7 : num 0 1 1 0 0 0 0 1 0 0 ...

$ LF : num 0.75 0.772 0.417 0.533 0.887 0.125 0.969 0.842 0.75 0.063 ...

5. PANEL DATA CREATION

After creating the metadata above, we can create the panel data structure. In the panel data structure as in Table 4 provided with an example, an item and time variables need to be created.

Table 4: Sample Panel Data Frame

Item	Time	Y	X1	X2	X3	
A		20160101	0.85	0.86	0.18	0.43
A		20160102	0.82	0.17	0.28	0.50
A		20160103	0.80	0.79	0.25	0.79
B		20160101	0.87	0.54	0.07	0.44
B		20160102	0.85	0.52	0.39	0.43
B		20160103	0.85	0.78	0.13	0.17

In our example, the panel data frame is formed as follows. As an Item variant, the Board Country and Off Country variables will be combined and the model will be edited to a higher level from the AirPort level. The purpose of doing this is to increase the working performance of the model. The table format is shown on Table 5.

Table 5: The Panel Data Frame of Our Study

KEY	ID_ORIGIN_YMD	ID_FLIGHT_NUMBER	ID_BOARD_POINT	ID_OFF_POINT	ID_COMPARTMENT	LF	RES1	RES2	RES3
Y-TR-TR	20160101	2328	IST	ADB	Y	0.844	0.701	0.735	0.776
Y-TR-TR	20160102	2328	IST	ADB	Y	0.882	0.819	0.819	0.844
Y-TR-TR	20160103	2328	IST	ADB	Y	0.973	0.925	0.925	0.966
Y-TR-TR	20160104	2328	IST	ADB	Y	0.844	0.803	0.735	0.721
Y-TR-TR	20160105	2328	IST	ADB	Y	0.553	0.404	0.389	0.343
Y-TR-TR	20160106	2328	IST	ADB	Y	0.439	0.456	0.377	0.386
Y-TR-TR	20160108	2328	IST	ADB	Y	0.826	0.58	0.435	0.241

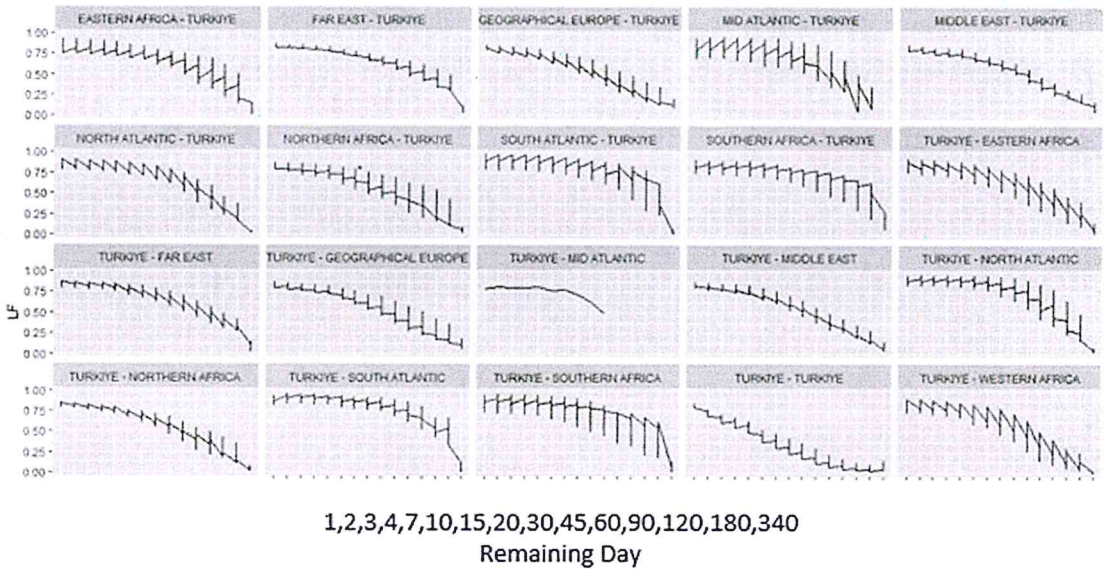
The KEY variable is our item variable and ID_ORIGIN_YMD is our Time variable.

6. EXPLORATORY DATA ANALYSIS

6.1. Reservation Patterns

Figure 1 below shows how the load factor value is changing by the number of days remaining to the flight in different destinations. When the graph is read from right to left, the last point gives us the load factor value. Especially when we think of the distance, the flights to North Atlantic, South Atlantic, Far East and South Africa show that the airplanes are starting to fill up, even as the airplane has more time to take off. We can see very clearly when the flights from Turkey to Turkey, that is to say on domestic flights, the plane is starting to be filled especially recently. As can be seen from these inferences, flights to different destinations show different patterns. The panel data format is crucial for capturing these patterns. The vertical lines show us how the reservation days include variability. For example, if we take the MID Atlantic-Turkey destination, each booking day shows highly variable values. This means that both forecasting will be difficult and that deviations will be dramatic.

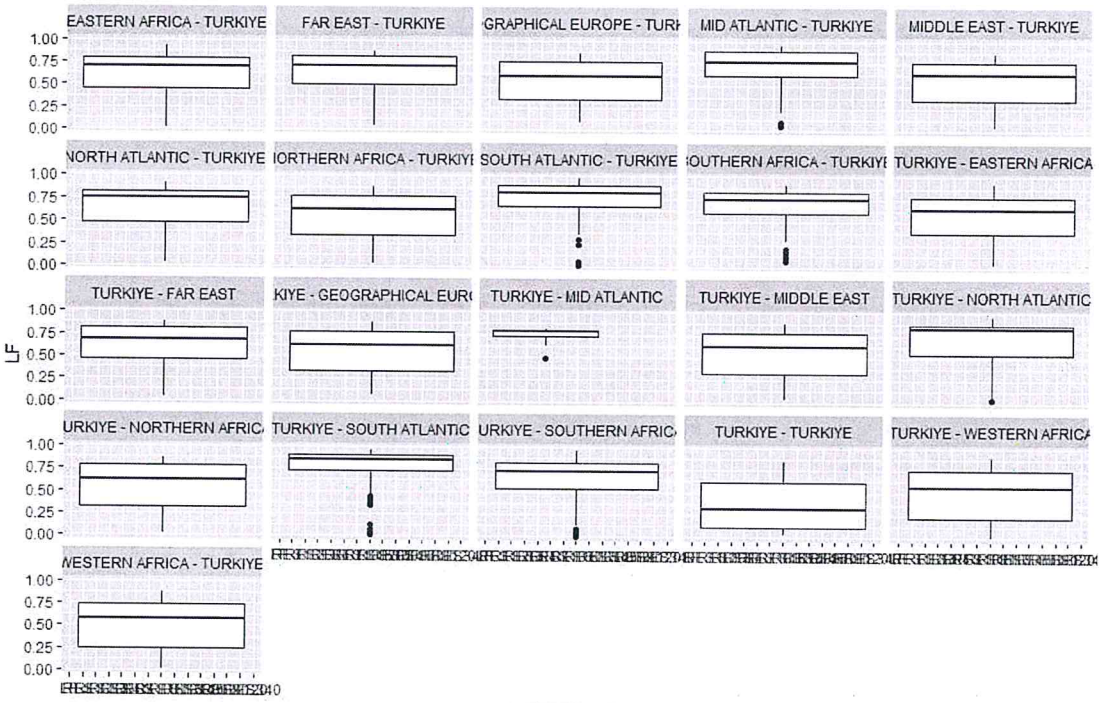
Figure 1: Reservation Patterns



6.2. Load Factor Values and Major Outliers

When we look at the Figure 2 below, we can see how load factor values and reservation values are distributed according to destinations. If we look at this chart and see that the box plot median value is close to zero according to the remaining days, we can deduce that the flights of that destination have recently filled the flight capacity. However, we can conclude that high load factor values do not constitute an outlier value. In some destinations (MID Atlantic - Turkey, South Atlantic - Turkey, and South Atlantic), we can see that it is an outlier. These outliers must be excluded from the model.

Figure 2: Load Factor Values and Major Outliers

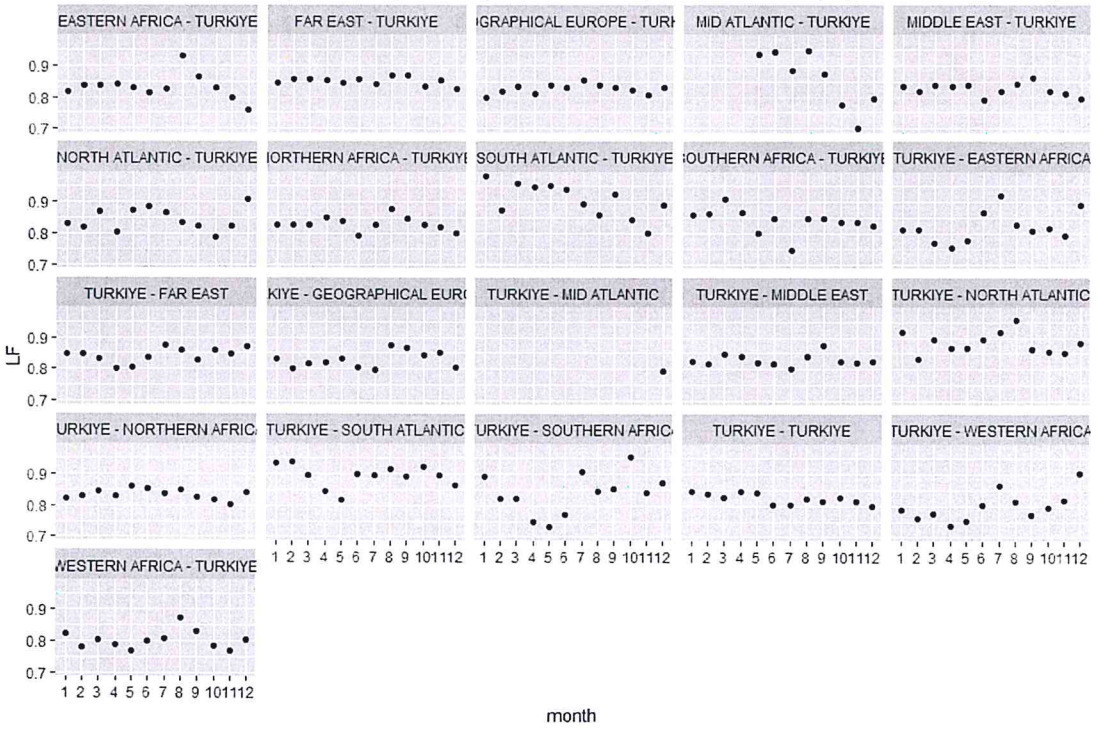


6.3. Month Variable and Load Factor Values

When we look at the Figure 3 below, we can see how load factor values and reservation values are distributed according to destinations. If we look at this chart and see that the box plot is close to the end of the bottom line according to the remaining days, we can deduce that the

flights of that destination have recently filled the flight capacity. However, we can conclude that high load factor values do not constitute an outlier value. In some destinations, for example, we can see that flights from the Mid Atlantic to Turkey and flights from the South Atlantic to Turkey include outliers. These outliers must be excluded from the model.

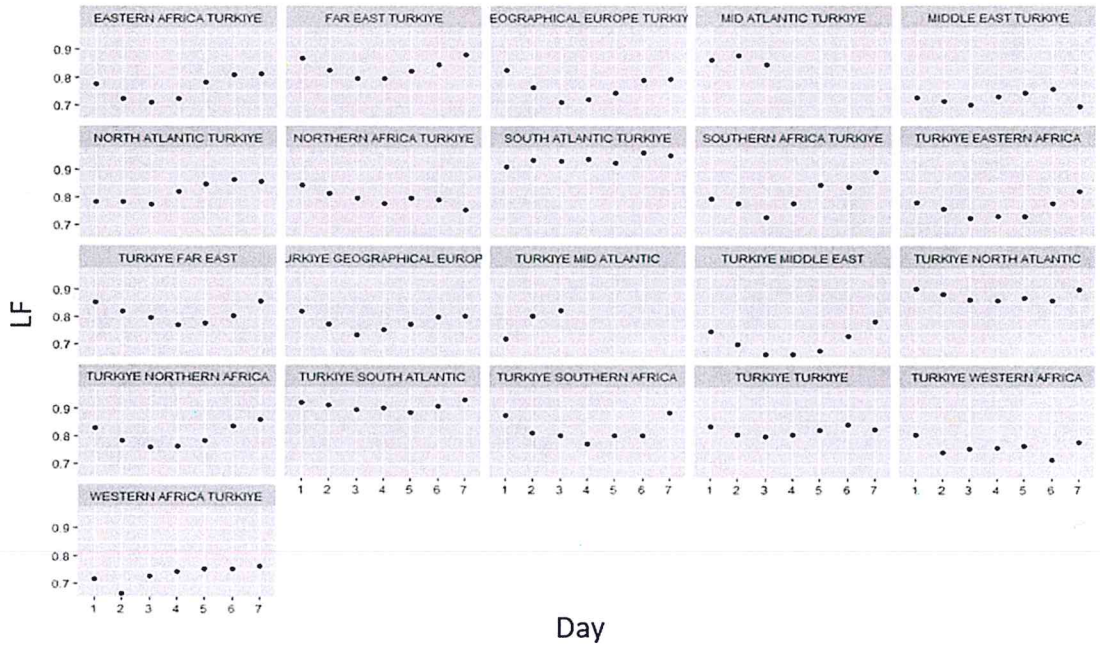
Figure 3: Month Variable and Load Factor Values



6.4. Day Variable and Load Factor Values

When we look at the Figure 4 below, we can see how the load factor values and the days of flight are scattered. The "day" variable on the x axis tells us what day of the week it is. The "day" variable in each region without statistical variation is a variable. Especially on the sixth and seventh days there are observable high values. The "day" variable will be added as a model dummy variable to distinguish between weekend flights and weekday flights.

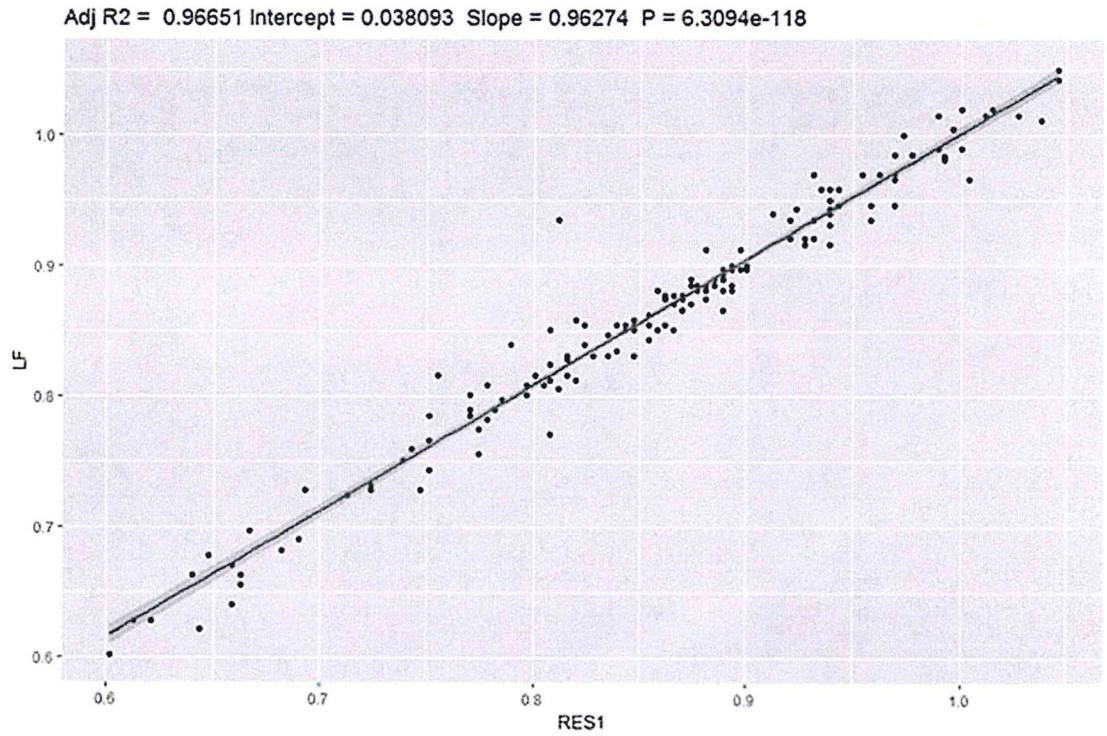
Figure 4: Day Variable and Load Factor Values



6.5. Correlation Between Reservation days and Load Factor values

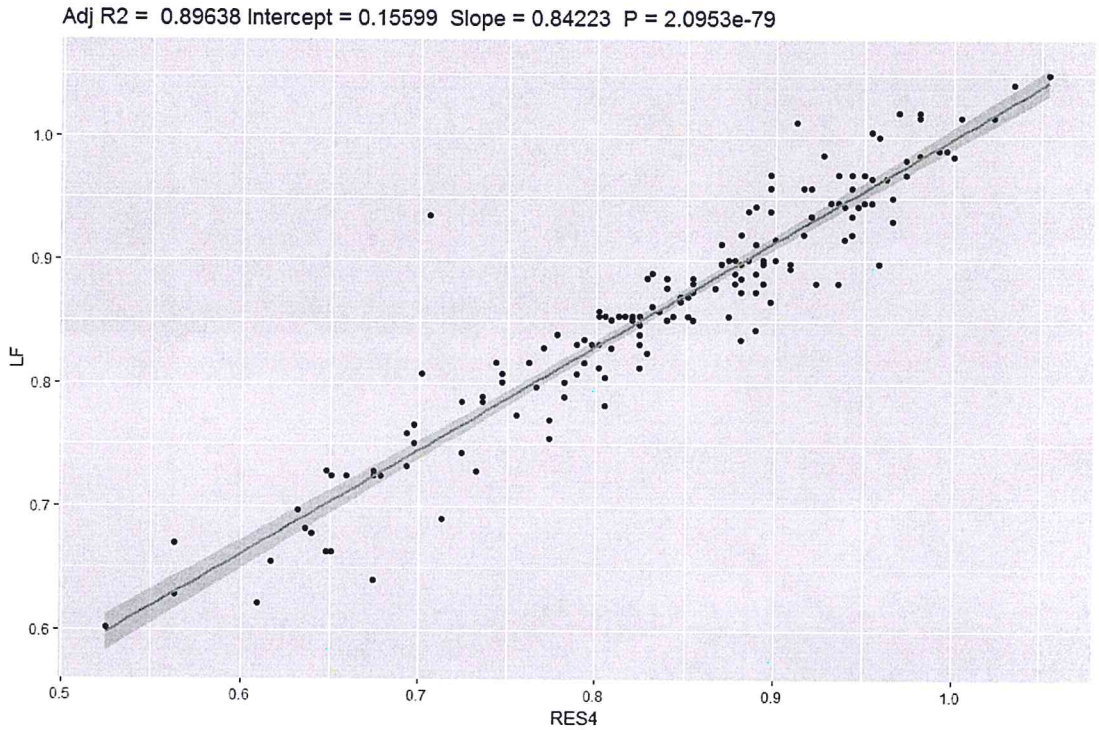
In the Figure 5 below, we will look at how we can explain the booking dates and the Load Factor data. To inspect, we will look at the values of the flight for only 1,4,7,15,30 days. We're looking for information on the following chart, a day before the flight. As we can see, the intersection is very small which is expected.

Figure 5: Correlation Between RES1 and LF



We're looking for information on the following chart Figure 6, four days before the flight. As we have seen, the intersection has increased somewhat, but again we can make accurate forecast decisions.

Figure 6: Correlation Between RES4 and LF



We're looking for information on the following chart Figure 7, seven days before the flight. As we have seen, the intersection has increased somewhat, but again we can make solid forecasts.

Figure 7: Correlation Between RES7 and LF

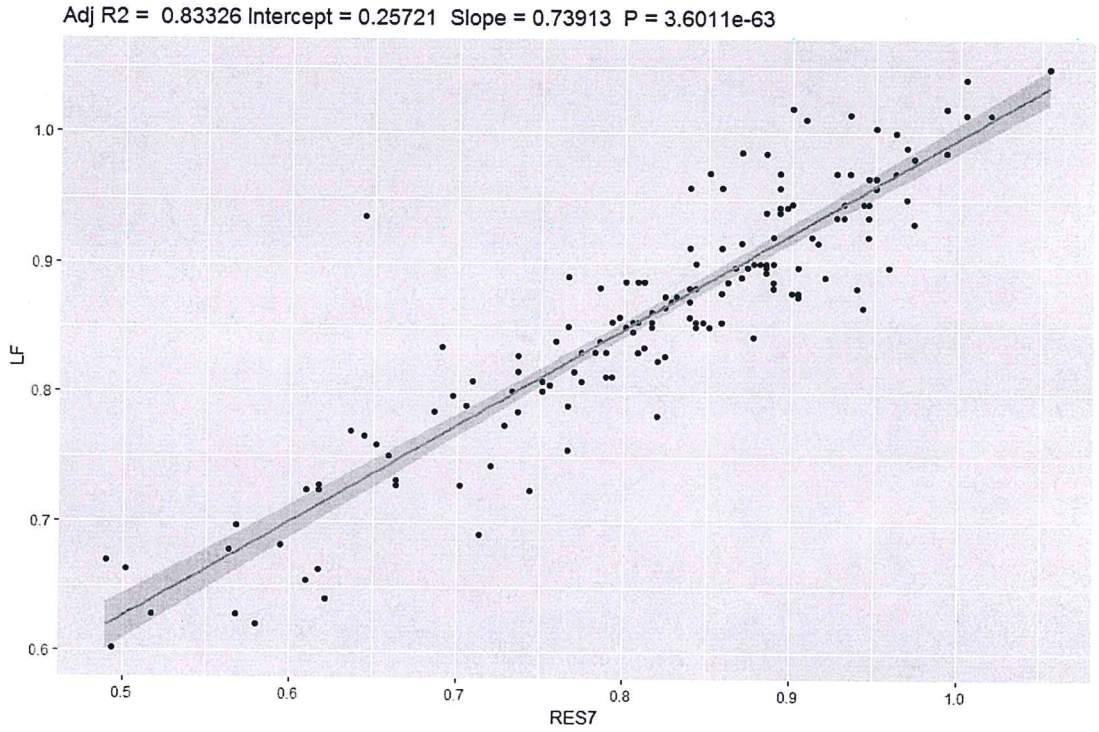
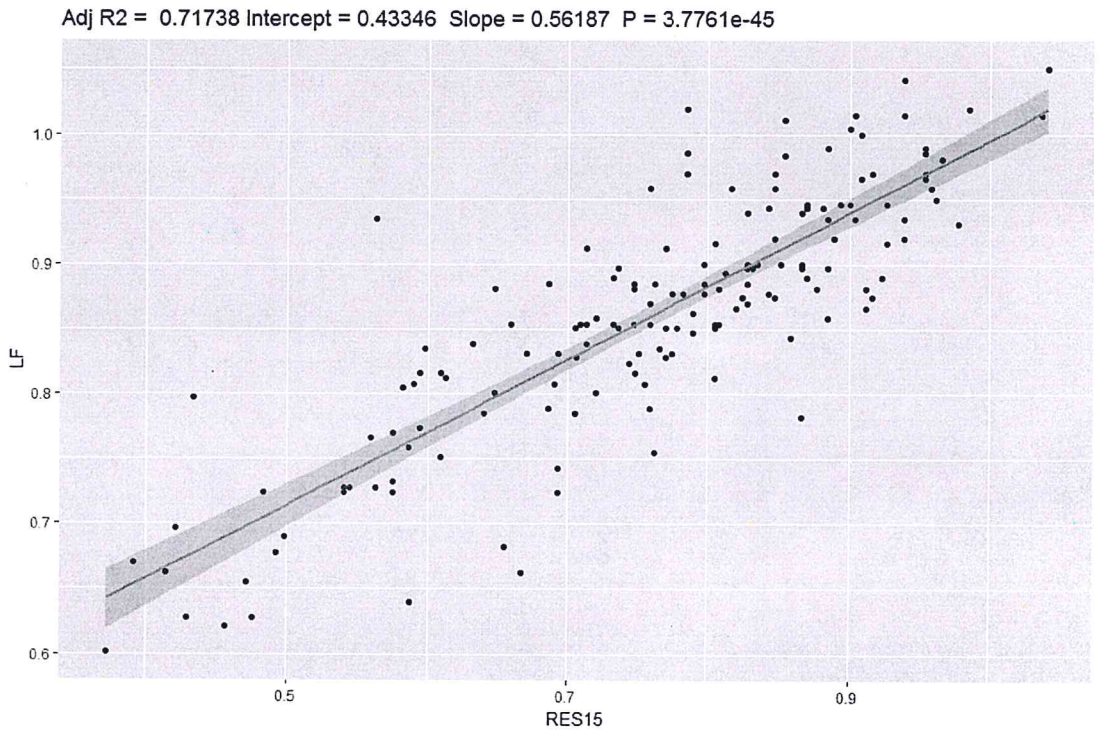
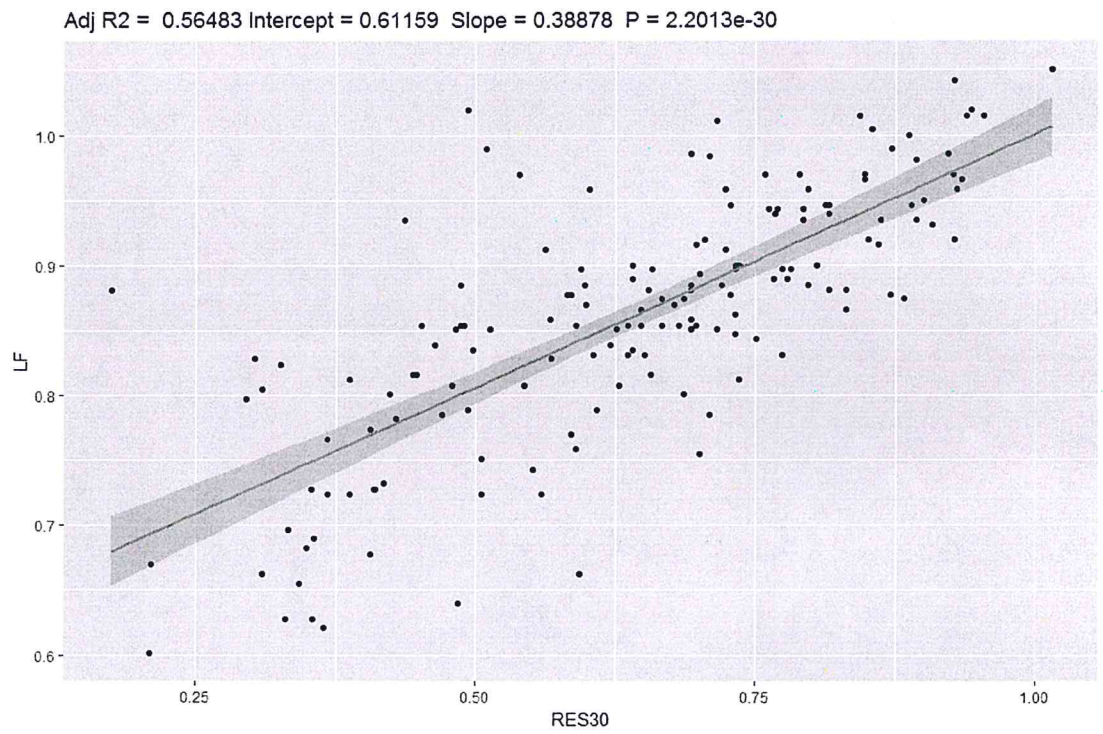


Figure 8: Correlation Between RES15 and LF



We're looking for information on the following charts Figure 8. As you can see, the intersection is extremely high. Other variables need power to make good predictions.

Figure 9: Correlation Between RES30 and LF



We're looking for information on the following charts Figure 8. As you can see, the intersection is extremely high. There is a strong relationship but it is inadequate to explain it on its own.

7. MODELLING

After detailed data analysis, we are in the phase of modeling. The R programming language is used in the modeling phase and in the creation of the visuals. The following section contains detailed information about the R programming language.

7.1. R Programming Language

This work was written in the R Programming Language. R offers extensive statistical data analysis facilities. Linear modeling includes classical statistical tests, classification and clustering. Since it is a programming language, it allows the user to compute with the new function. In addition, other programming languages can be used when necessary for the R programming language.

The R Programming Language is an open source code. Continuous improvement with its free access is obvious. In addition to the built-in libraries for statistical analysis, it also allows the development of new functions and new libraries depending on them. R, being a function-based and an object-based programming language, provides flexibility in many studies.

The R console can connect to databases and offer processing opportunities in a single data environment in different database environments. In this work, a connection will be made to the database with the ODBC protocol. The "RODBC" package will be used to provide these connections. (Ripley, B., Lapsley, M., & Ripley, M. B. 2017).

7.1. Mathematical Formulation of Work

7.1.1. Target Definition

$$Load\ Factor = \sum_{t=1}^{t} \left(\frac{(Number\ of\ Carried\ Passenger)_t * Distance_t}{(Available\ Seat)_t * Distance_t}\right)$$

The formula above gives us the calculation of the Load Factor. This calculation is made for each flight and cabin class and the target variable is created. With this calculation, we solve the ratio problem by weighting kilometers when we encounter multi-leg flights.

7.2. Mathematical Formulation of the Study

During the modeling phase, the entire year 2015 will be used as training data. We will test our model for 2016. Approximately 50% of the training data will be used as a 50% test data Here is the mathematical formula we use to estimate the load factor target variable.

$$LF_{it} = \beta_0 + GLF_{it}\beta_1 + WCLF_{it}\beta_2 + LLF_{it}\beta_3 + BCOCLF_{it}\beta_4 + (\sum_{k}^{340} RES_{it} \sum_{l}^{19} \beta_l) + \sum_{m}^{12} M_{it}\beta_m + \sum_{n}^{4} YR_{it}\beta_n + \sum_{o}^{7} DOW_{it}\beta_o$$

where

$k = 1, 2, 3, 4, 7, 10, 15, 20, 30, 45, 60, 90, 120, 340$ $(RES\ parameter)$

$l = 1, 2, 3, \dots, 14$ $(RES\ length\ parameter)$

$m = 1, 2, 3, \dots, 12 (Month)$

$n = 1, 2, 3, 4$ $(Year)$

$o = 1, 2, 3, \dots, 7 (Day\ of\ Week)$

$\beta_0 = Y - Intercept\ (Base\ Load\ Percentage)$

$\beta_1 = Group\ LF\ Coefficient$

$\beta_2 = LF\ Wait\ Count\ Coefficient$

$\beta_3 = 365\ Days\ Lag\ of\ LF\ Coefficient$

$\beta_4 = Board\ Country - Of\ Country\ LF\ Coefficient$

$\beta_i = RESi\ Coefficient$

$\beta_m = Month\ i\ Coefficient$

GLF: Group LF

WCLF : LF Wait Count

LLF : 365 Days Lag of LF

BCOCLF : Board Country-Of Country LF

RES: Reservation

M: Month

YR: Year

DOW: Day of Week

When estimating the coming days in the modeling, there are different departure days for each flight. Here, the following function has been used to dynamically program. We can open up what we want to achieve with this function: for example, there is 10 days for the flight to take off, in which case *RES15* will be used in calculating the model. For example, there are 25 days for the plane to departure, in which case the *RES30* will enter the model.

In this way different models will be created according to different *RES* values. We will have 14 different model results, *RES1, RES2, RES3, RES4, RES7, RES10, RES15, RES20, RES30, RES60, RES90, RES180, RES340.*

```
SelectReservRange <- function(i){
 if(i <= 1){
  RES.i <- 1
  Rowindex.i <- 1
 }else
  if(i <= 2){
   RES.i <- 2
   Rowindex.i <- c(2:2)
  }else
   if(i <= 3){
    RES.i <- 3
    Rowindex.i <- c(3:3)
   }else
    if(i <= 4){
     RES.i <- 4
     Rowindex.i <- c(4:4)
    }else
     if(i <= 7){
      RES.i <- 7
      Rowindex.i <- c(5:7)
     }else
      if(i <= 10){
       RES.i <- 10
       Rowindex.i <- c(8:10)
      }else
       if(i <= 15){
        RES.i <- 15
        Rowindex.i <- c(11:15)
       }else
        if(i <= 20){
         RES.i <- 20
         Rowindex.i <- c(16:20)
        }else
         if(i <= 30){
          RES.i <- 30
          Rowindex.i <- c(21:30)
         }else
          if(i <= 45){
           RES.i <- 45
           Rowindex.i <- c(31:45)
          }else
           if(i <= 60){
            RES.i <- 60
            Rowindex.i <- c(46:60)
           }else
            if(i <= 90){
             RES.i <- 90
```

```
            Rowindex.i <- c(61:90)
           }else
            if(i <= 120){
             RES.i <- 120
             Rowindex.i <- c(91:120)
            }else
             if(i <= 340){
              RES.i <- 340
              Rowindex.i <- c(121:340)
             }
 else{
  RES.i <- 9999
  Rowindex.i <- c(9999,9999)
 }
 return(c(RES.i,Rowindex.i))
}
```

7.3. Using The F Statistic and P-Value

You can use the F statistic when deciding to support or reject the null hypothesis. In your F test results, you'll have both an F value and an F critical value. The F critical value is what is referred to as the F statistic. In general, if your calculated F statistic in a test is larger than your table F value, you can reject the null hypothesis.It means that your model is significant. However, the statistic is only one measure of significance in an F Test. You should also consider the p value. The p value is determined by the F statistic and is the probability your results could have happened by chance.

7.4. Model Results

As we mentioned in the previous section, you will see 14 different model results at this stage.

Model 1: *RES1* Model Results

Residual standard error: 0.0849 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.9146, *Adjusted* $R-squared$: 0.9145
$F-statistic$: $1.489e+04$ *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 2: *RES2* Model Results

Residual standard error: 0.1041 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.8715, *Adjusted* $R-squared$: 0.8714
$F-statistic$: 9439 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 3: *RES3* Model Results

Residual standard error: 0.1159 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.8406, *Adjusted* $R-squared$: 0.8405
$F-statistic$: 7340 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 4: *RES4* Model Results

Residual standard error: 0.1244 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.8165, *Adjusted* $R-squared$: 0.8163
$F-statistic$: 6191 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 5: *RES7* Model Results

Residual standard error: 0.1419 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.7613, *Adjusted* $R-squared$: 0.7611
$F-statistic$: 4438 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 6: *RES10* Model Results

Residual standard error: 0.1516 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.7276, *Adjusted* $R-squared$: 0.7274
$F-statistic$: 3717 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 7: *RES15* Model Results

Residual standard error: 0.162 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.6889, *Adjusted* $R-squared$: 0.6887
$F-statistic$: 3082 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 8: *RES20* Model Results

Residual standard error: 0.1695 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.6593, *Adjusted* $R-squared$: 0.659
$F-statistic$: 2692 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 9: *RES30* Model Results

Residual standard error: 0.1793 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.6489, *Adjusted* $R-squared$: 0.6186
$F-statistic$: 2260 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 10: *RES45* Model Results

Residual standard error: 0.1881 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.6007, *Adjusted* $R-squared$: 0.5804
$F-statistic$: 1927 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 11: *RES60* Model Results

Residual standard error: 0.1931 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.5981, *Adjusted* $R-squared$: 0.5578
$F-statistic$: 1757 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 12: *RES90* Model Results

Residual standard error: 0.1984 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.5833, *Adjusted* $R-squared$: 0.5329
$F-statistic$: 1590 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 13: *RES120* Model Results

Residual standard error: 0.2011 *on* 763978 *degrees of freedom*
Multiple $R-squared$: 0.5708, *Adjusted* $R-squared$: 0.5204
$F-statistic$: 1512 *on* 549 *and* 763978 *DF*, $p-value: < 2.2e-16$

Model 14: *RES340* Model Results

Residual standard error: 0.2043 *on* 763978 *degrees of freedom*

$Multiple\ R-squared$: 0.5052, $Adjusted\ R-squared$: 0.5048
$F-statistic$: 1421 on 549 and 763978 DF, $p-value$: $< 2.2e-16$

The above model give meaningful results when looking at Rsquare values. The high variability is 60% which is seen in the aviation sector and above that Rsquare gives consolidated results for the 30 day forecasts. The model does not explain the uncertainty very well for the flights remaining 30 days or more before its departure.

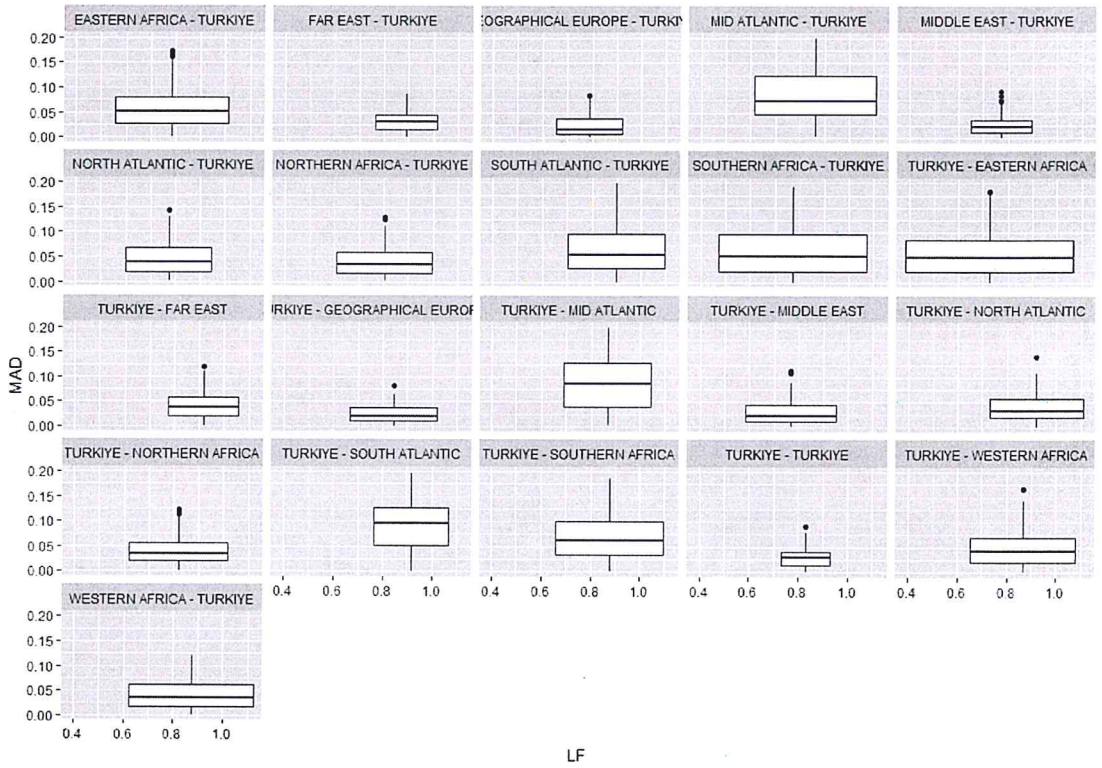
8. MODEL VALIDATION

The model results were obtained using the 2016 data for testing. The mean absolute deviation statistic is used for model validation. In statistics, the mean absolute deviation (MAD) is a robust measure of the variability of an univariate sample of quantitative data. It can also refer to the population parameter that is estimated by the MAD calculated from a sample. The validation results are shown Table 6 and Figure 10.

Table 6: Model Validation MAD Data Frame

DS_BOARD_AREA	DS_OFF_AREA	1	2	3	4	7	10	15	20	30	45	60	90	120	340
EASTERN AFRICA	TURKIYE	0.02	0.02	0.06	0.05	0.05	0.06	0.08	0.09	0.10	0.08	0.07	0.09	0.09	0.10
FAR EAST	TURKIYE	0.04	0.05	0.05	0.07	0.05	0.06	0.06	0.07	0.09	0.08	0.07	0.07	0.08	0.08
GEOGRAPHICAL EUROPE	TURKIYE	0.03	0.05	0.08	0.09	0.08	0.09	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.09
MID ATLANTIC	TURKIYE	0.02		0.07		0.06	0.17	0.08	0.17	0.17	0.13	0.09	0.11	0.08	0.07
MIDDLE EAST	TURKIYE	0.05	0.08	0.08	0.09	0.07	0.09	0.09	0.09	0.09	0.10	0.10	0.10	0.09	0.09
NORTH ATLANTIC	TURKIYE	0.01	0.03	0.05	0.07	0.05	0.06	0.08	0.11	0.11	0.08	0.09	0.07	0.07	0.06
NORTHERN AFRICA	TURKIYE	0.05	0.06	0.08	0.09	0.07	0.07	0.09	0.10	0.10	0.10	0.09	0.09	0.08	0.09
SOUTH ATLANTIC	TURKIYE	0.01	0.03	0.05	0.05	0.07	0.07	0.08	0.07	0.07	0.10	0.05	0.08	0.07	0.07
SOUTHERN AFRICA	TURKIYE	0.01	0.02	0.04	0.02	0.02	0.02	0.08	0.13	0.13	0.12	0.09	0.08	0.08	0.07
TURKIYE	EASTERN AFRICA	0.03	0.04	0.03	0.07	0.07	0.12	0.06	0.10	0.09	0.08	0.08	0.10	0.09	0.09
TURKIYE	FAR EAST	0.03	0.04	0.05	0.06	0.07	0.10	0.08	0.08	0.09	0.09	0.08	0.09	0.09	0.09
TURKIYE	GEOGRAPHICAL EUROPE	0.04	0.05	0.05	0.07	0.07	0.09	0.08	0.08	0.09	0.09	0.09	0.09	0.09	0.09
TURKIYE	MID ATLANTIC	0.01	0.04		0.04		0.17	0.00	0.09	0.10				0.09	0.09
TURKIYE	MIDDLE EAST	0.04	0.06	0.08	0.08	0.08	0.09	0.11	0.11	0.10	0.10	0.10	0.10	0.10	0.09
TURKIYE	NORTH ATLANTIC	0.04	0.02	0.03	0.04	0.05	0.08	0.06	0.06	0.07	0.08	0.07	0.07	0.06	0.06
TURKIYE	NORTHERN AFRICA	0.05	0.06	0.07	0.10	0.07	0.09	0.11	0.09	0.10	0.10	0.10	0.09	0.09	0.09
TURKIYE	SOUTH ATLANTIC	0.03		0.05	0.09	0.06	0.07	0.05	0.08	0.11	0.11	0.11	0.12	0.10	0.09
TURKIYE	SOUTHERN AFRICA	0.02	0.03	0.05	0.02	0.06	0.10	0.03	0.08	0.11	0.08	0.07	0.09	0.09	0.08
TURKIYE	TURKIYE	0.08	0.08	0.08	0.09	0.09	0.09	0.08	0.09	0.08	0.09	0.09	0.08	0.08	0.08
TURKIYE	WESTERN AFRICA	0.05	0.04	0.07	0.09	0.10	0.11	0.08	0.08	0.11	0.09	0.09	0.10	0.09	0.09
WESTERN AFRICA	TURKIYE	0.03	0.03	0.07	0.06	0.07	0.06	0.08	0.08	0.08	0.09	0.08	0.08	0.09	0.08

Figure 10: Model Validation MAD Boxplot, Defining the Deviation



9. MODEL FIT LINE GRAPHS

In the graphs below Figure 11, Figure 12, Figure 13, Figure 14, we clearly observe how the estimate and actual values fit closely.

Figure 11: Sample Model Fit Line Graph

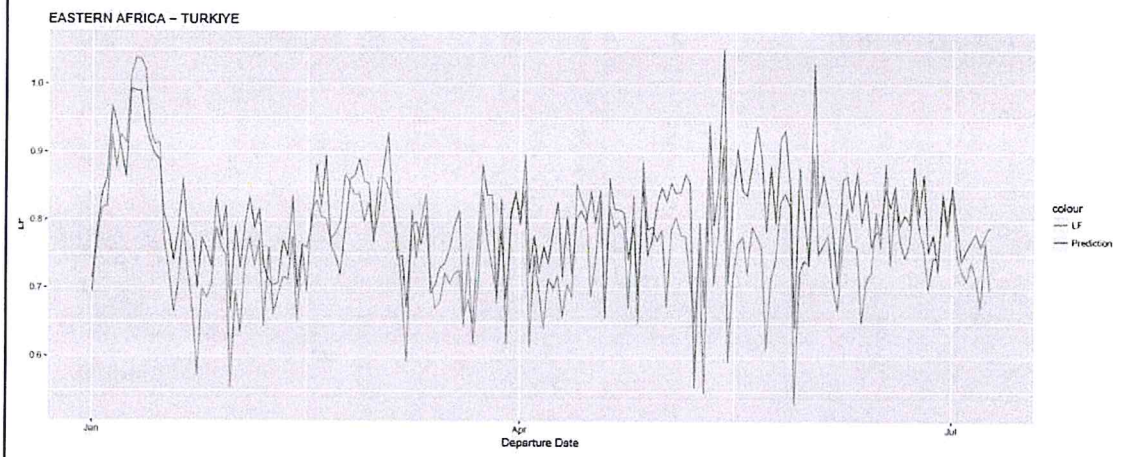


Figure 12: Sample Model Fit Line Graph

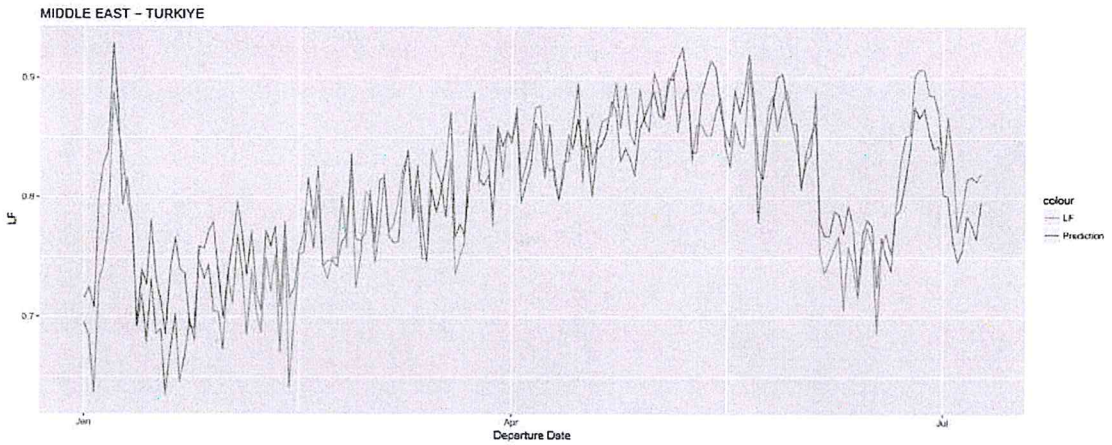


Figure 13: Sample Model Fit Line Graph

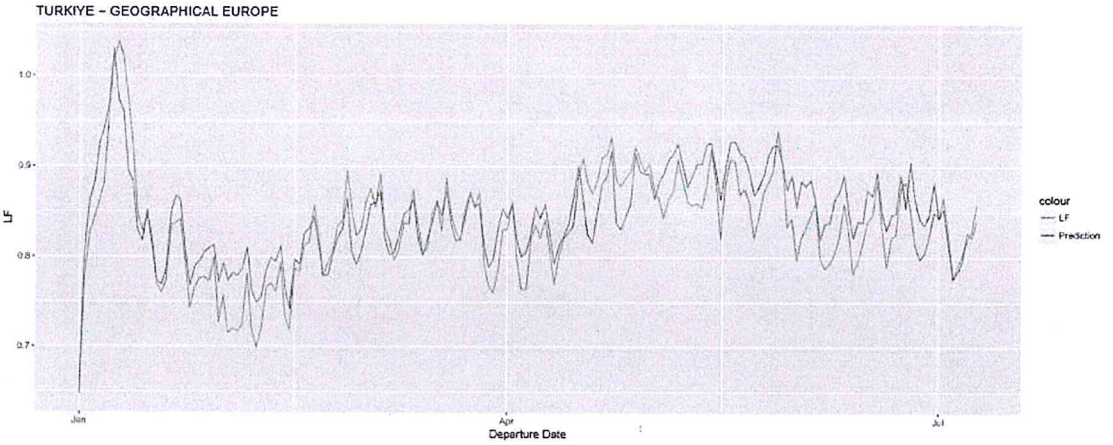
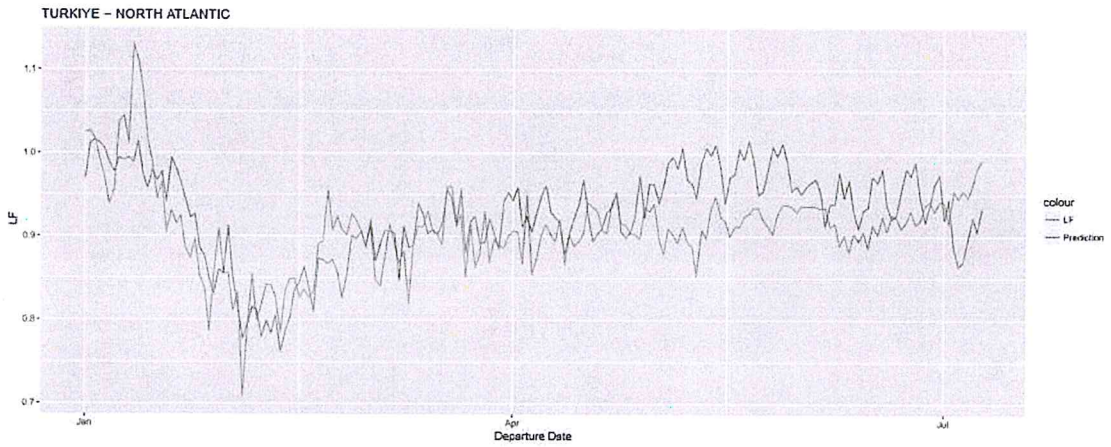


Figure 14: Sample Model Fit Line Graph



10. CONCLUSION

Detailed analysis of this project showed that we obtained the best estimation accuracy with panel data method. With this study, a load factor prediction model has been developed in order to better manage the flight capacity of Turkish Airlines. All analysis and reports have been developed with the popular R programming language for portability. Since a dynamic and conditional model method is chosen, the daily job of the models is trained by parallel processing tools.

Turkish Airlines initially stated that they own a time series model to forecast the load factor. They point out that their model results are accurate for 34% of the flights within a range of 0% to 5% error in thirty day long forecasting. They have mandated that the accuracy of rate of flights, whose error is between 0% and 5%, should be better than 34% in order to integrate the proposed model of this project into their operational system. As a result of our forecasting model results, the rate of flights within the range of 0% to 5% error in thirty days period has been increased from 34% to 64%, a major jump compared to the previous model performance. With the contribution of this thesis' work, the model has been accepted and begun being actively used by the current live system. It has been successfully included in the business processes of Turkish Airlines.

11. FUTURE STUDIES

There are two issues in our work that we will work on in the future. The most important of these is providing an infrastructure that we can process all the days of sales from the reservations. We will focus on using the distributed data structure to process large data faster and increase model performance. The second issue is to increase the accuracy of model results using weather forecast data.

As a continuation of the scope of this study, Spark Framework was established within the Turkish Airlines. Spark technology is a technology that enables multiple servers to be used together to make transactions through memory. It is aimed to do modeling work through machine learning libraries where the Spark framework is provided. With this work, more data will be processed and increasing model accuracy will be our most important goal.

The use of weather forecasts in the modeling phase will be like an early warning system. In the work we did, the load factor values that were low due to the weather opposition were evaluated as outlier. With the weather forecast data being used in the modeling phase, we will be explaining the reason for the low load Factor values due to weather opposition.

REFERENCES

[1] Brueckner, J. K., & Whalen, W. T. (2000). The price effects of international airline alliances. The Journal of Law and Economics, 43(2), 503-546.

[2] Belobaba, P., Odoni, A. And Barnhart, C. (2000). "The Global Airline Industry", Wiley Online Library. (BOOK).

[3] Tesfay, Y. Y. (2016). "Modified panel data regression model and its applications to the airline industry: Modeling the load factor of Europe North and Europe Mid Atlantic flights". Journal of Traffic and Transportation Engineering, 3 (4), pp: 283-295. (BOOK).

[4] Talluri, K. and Ryzin G.(2004). "Revenue Management Under a General Discrete Choice Model of Consumer Behavior". Informs, 50 (1), pp: 15-33.

[5] Kellner, L. (2000). "Building a global airline brand." *2000 Transport Conference. UBS Warburg, London.*

[6] Ľubomír, F., Jakub, H. (2013). Airline pricing strategies in European airline market. Available at: pernerscontacts.upce.cz/30_2013/FeMArco.pdf (accessed 12.03.14.).

[7] Bertsimas, D., Popescu, I. (2003). Revenue management in a dynamic network environment. Transportation Science 37(3), 257–277. (BOOK).

[8] Doganis, R. (2010). Flying off course: Airline Economics and Marketing, fourth ed. Routledge, London.

[9] Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York.

[10] Hsiao, C. (2003). Analysis of Panel Data (Cambridge University Press, Cambridge).

[11] Moulton, B.R. (1986). Random group effects and the precision of regression estimates, Journal of Econometrics 32, 385–397.

[12] Baltagi, B. H., & Levin, D. (1992). Cigarette taxation: Raising revenues and reducing consumption. Structural Change and Economic Dynamics, 3(2), 321-335.

[13] Baltagi, B. (2008). Econometric analysis of panel data. John Wiley & Sons.

[14] Deaton, A. (1995). Data and econometric tools for development analysis. Handbook of development economics, 3, 1785-1882.

[15] Ripley, B., Lapsley, M., & Ripley, M. B. (2017). Package 'RODBC'.

[16] Wesonga, R., Nabugoomu, F., & Masimbi, B. (2013). Assessing Aircraft Timeliness Variations By Major Airlines: Passenger Travel Practice In Uganda. International Journal of Sciences: Basic and Applied Research, 11(1), 75-83.