Ensembled Econometric based ARIMA model for Forecasting Demand in Passenger Air Traffic

Sydney International Airport

PROJECT REPORT

submitted towards the partial fulfillment of

the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Harsha Vardhan Miryala **15114045**

Sanju Prabhath Reddy **15114042**

Sri Harsha Majeti **15114044**

Under the guidance of

Professor Durga Toshniwal

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE-247667

APRIL 2019

We declare that the work presented in this thesis with title "**Ensembled Econometric based ARIMA model (E-square ARIMA) for Forecasting Demand in Passenger Air Traffic Sydney International Airport**" towards the partial fulfillment of the requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted to the Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee is an authentic record of our own work carried out during the period from August 2018 to April 2019 under the supervision of **Dr.Durga Toshniwal**, Dept. of Computer Science and Engineering, IIT Roorkee.

PLACE: .....

DATE: .....

.....	.....	.....
Harshavardhan Miryala	Sanju Prabhath Reddy	Sri Harsha Majeti
15114045	15114042	15114044

This is to certify that the declaration made by the students is correct to the best of my knowledge and belief.

DATE: .....

.....

Dr. Durga Toshniwal

ACKNOWLEDGEMENTS

First and foremost, we would like to express our sincere gratitude towards our guide **Dr. Durga Toshniwal** ma'am, Professor, Dept. of Computer Science and Engineering, IIT Roorkee for their ideal guidance throughout the period. Their advices, insightful discussions and constructive criticisms certainly enhanced our knowledge and improved our skills. Their constant encouragement, support and motivation have always been key sources of strength for us to overcome all the diffi-
cult and struggling phases.

We would also like to thank **Department of Computer Science and Engineering, IIT Roorkee** for providing lab and other resources for this project, as well as for being a perfect catalyst to our growth and education over these past four years. We also extend our gratitude to all our friends, for keeping us motivated and providing us with valuable insights through various interesting discussions.

1. Introduction and Motivation

1.1 Introduction

Civil aviation benefits from and contributes to the economic development of all nations by complicated interaction with other econometric sectors. The demand for air services is expanding as incomes and production rises. Future tourism, trade and jobs could therefore be predicted. Therefore. Civil aviation is an important tool of economic development and, through facilitation of international agreements and understanding, air transport provides intangible benefits.

On the other side, it is the expedition and flexibility provided by the global air transport system that is responsible for the role of air transport as a catalyst for general economic and social development. It has expanded the markets for various product types and has also promoted the interaction and exchange of ideas, professional experience and skills between different countries. Since the role of air transport is increasingly important for countries ' and regions ' economic developments and air transport is the forefront of industry, it is important to take due account of the economic and social benefits that can be provided by an effective air transport system and to ensure that future needs for air transport and associated financial and social factors are properly evaluated. Air traffic forecasting will be used to estimate the number of potential air passengers.

In order to build, operate and further expand airports, a significant initial and on-going investment can be required, which is generally paid for in large quantities with public money. As such, it is important that future demands for aviation services are forecast for each airport to estimate the potentially required additional investment in capacity or services in order to meet these requirements. For effective airport planning and decision-making and efficient capacity provision, accurate forecasting is essential.

The type of forecast and extent of effort required to generate it depends largely on the purpose of the forecast. Short-term aviation forecasts are required for operational planning and are frequently used as a means to assess airport staff requirements or to assess the need for incremental improvements or expansions in landside and terminal facilities, air cargo facility, general aviation hangar, etc.

The objective of this analysis is to provide a realistic prediction based on the latest available data reflecting current airport conditions supported by data in the study which provide an adequate justification for airport planning and development.

1.2 Problem Description

The prediction is at the heart of airplane terminal planning and configuration. The airport's forecasts are fully based on aircraft terminals, runways, cargo shelters, parking spaces and also the route to and from the air terminal. Traveler volume forecasts are space requirements for terminal building installations, while aircraft development forecasting reflects the requirements for runway, runway and airport regulations, as well as the airport regulatory framework requirements. There are countless qualities to be evaluated in the arrangement and planning of the aircraft terminal: an absolute number of travelers in time for the structure, domestic, passenger and international proportions in pinnacles of hours, seasonal demand changes, the exchange or transit travel volumes of each type of traffic, number of travelers and bags, dispersion of residence.

These forecasts are utilized to decide the space prerequisite for new terminals or the expansion of existing offices. It is apparent that the forecasting procedure can be the highest critical factor in the advancement of the airplane terminal (Howard, 1974). Mistakes made in this period of the procedure can be all around exorbitant and harming for nearby economies. Underestimating demand can prompt increased congestion, delay, and lack of storage facilities, as it occurred in Venezuela in 1974. The revelation of oil brought about a sensational and unexpected increment of the cargo volumes dealt with by the Caracas air terminal. The arranged storage facilities were inadequate to deal with this expanded interest, thus the load was put away in territories where it was either crushed or stolen. Overestimating demand could likewise make huge issues. Forecasts of traveler demand for the Newark airplane terminal were high to the point that the newly built air terminal stayed void for various years (de Neufville, 1976). Errors in the forecasting procedure can prompt long deferrals or to exhaust terminals: both these cases can make huge damage to the economy of an air terminal's hinterland that relies upon the successful operation of that air terminal.

The aim of this analysis is to provide a realistic prediction based on available data to reflect the current airport conditions, supported by information presented in the study, which provides sufficient justification for the development and planning of the airport.

The aim here is to develop a model that can accurately predict the volume of air traffic in Sydney International Airport using the dataset that is aggregated from various available sources.

We develop a machine learning approach using different techniques to forecast the airport passenger traffic. We implement time series data prediction and provide an econometric modelling on the data obtained from various sources available on the internet.

1.3. Organisation

Literature Survey discusses different traditional forecasting techniques which include trend projection techniques and econometric forecast techniques in previous forecasting studies. Trend surveys discuss several decomposition and smoothing methods.Econometric forecast further discusses the standard procedure for applying this technique, relevant factors that can contribute to the forecast econometrically. This section also explains why those factors were chosen in previous research studies. It also explains regression techniques that are generally used to build models for the forecast.

Proposed Methodology discusses the proposed model that is derived from extensively studying the previously done effluent research papers.

Experiments and Discussion presents the validated proofs of the proposed methodology. It shows results of the individual models used along with proposed model and discusses the advantages of using the proposed model compared to individual components of the model.

Conclusion and Future Work discusses the conclusion of our work presented in the previous sections. It also dwells into further potential future work that could extend our proposed methodology. It also discusses various other techniques, which are not presented in our work, that can be explored.

References contain all the sources from which we referenced our literature survey, techniques,

explanation for relevant factors used in forecasting and other studies to validate our work.

2. Literature Survey

2.1 Traditional forecasting techniques

Three traditional air traffic projection models are: trend projection, economy and market and industry surveys. The following are the following: These methods depend on the quantity of data and statistical analysis necessary and the extent to which selective assessment is important.

2.1.1 Forecasting by trend projection

As a first step in the prediction of air traffic, historical data (time series) are usually studied and the trend for traffic development is determined. With regard to the medium-to long-term forecast, the aviation trend means that air traffic trends are isolated from fluctuation in traffic levels over a long time period, or smoothly. If a medium or long-term forecast is derived from the trend of traffic, it can continue to operate in future, as in the past, and its impact may be gradually changed.

2.1.2 Specification of trend surveys

Various mathematical formulations can be used for different types of trend surveys. The dependent variable "Y," in each case, consists of air traffic, the independent variable' T' consists of the time normally measured in years, and a, b, c are all constants called coefficients, with data estimates.

- Linear Y= a+bT
- Exponential Y= a (1+b)^T
- Parabolic $Y= a+bT+cT^2$

In the case of macroeconomic data for forecast, different methods of traditional time series include decomposition methods and smoothing methods. This section shows the methods and formulas. the Orhunbilge Notation (1999) explains the methods of time series[2].

2.1.3 Decomposition Methods

Methods of decomposition are used in the determination of secular trends, seasonal variation, conjuncture and random fluctuation in series of time series components (irregular variation). Annual data were used for this study. This part of this study therefore referred to three important trend features, including linear, quadratic and growth.

2.1.3.1 Least Squares Method for Determining Trend

One popular way of determining a trend is the least square method. In yt= f(x) function, X is the time variable (year, month, etc.). The estimates of the model parameters can be shown as the following formulae if the value of the time series variable (X) is identified as zero. The yt trend can be determined using least square method. The function which we should use as a trend is not easy to decide. The appropriate trend functions can be found by testing several functions and finding minimum residual quadrants.

$$\sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} \left(y_t - y_t' \right)^2 \Rightarrow \min$$

2.1.3.2 Linear Trend Function

The linear trend function is shown as below:

$$y = a + bx + e_t$$

When the least squares method is applied the linear trend function, the equations below are obtained.

$$\sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} \left(y_t - y_t' \right)^2 = \sum_{t=1}^{n} \left(y_t - a - bx \right)^2$$

For determining the minimum of this function the first level derivatives should be done regarding to a and b parameters.

$$\sum y_t = na + b\sum x$$
$$\sum xy_t = a\sum x + b\sum x^2$$

By solving these equations the parameters a and be can be found as follows:

$$a = \frac{\sum y_t}{n}$$
$$b = \frac{\sum xy_t}{\sum x^2}$$

2.1.3.3 Quadratic Trend Function

If the observed data has a curved figure (in quadratic trend function the mean of the data is increasing first than start decreasing or reverse) than quadratic trend function can be used.

$$y = a + bx + cx^2 + e_t$$
$$\sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} \left(y_t - y_t'\right)^2$$
$$= \sum_{t=1}^{n} \left(y_t - a - bx - cx^2\right)^2 = 0$$

First order derivatives of the equation according to a, b and c parameters should be solved for writing the quadratic trend function with using least squares

method. The equations below are the normal equations. Three unknown can be found by solving these three equations.

$$\sum y_t = na + b\sum x + c\sum x^2$$
$$\sum xy_t = a\sum x + b\sum x^2 + c\sum x^3$$
$$\sum x^2 y_t = a\sum x^2 + b\sum x^3 + c\sum x^4$$
$$b = \frac{\sum xy_t}{\sum x^2}$$

2.1.3.4 Growth Trend Function

If the change of the y variable is nearly constant in time, growth trend function can be used for this kind of data. The growth trend function is shown below.

$$y_t = ab^x + e_t$$
$$\sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} \left(\log y_t - \log y_t'\right)^2$$
$$= \sum_{t=1}^{n} \left(\log y_t - \log a - x \log b\right)^2 = 0$$
$$\sum \log y_t = n \log a + \log b \sum x$$
$$\sum x \log y_t = \log a \sum x + \log b \sum x^2$$
$$\log a = \frac{\sum \log y_t}{n}$$
$$\log b = \frac{\sum x \log y_t}{\sum x^2}$$
$$\log y_t = \log a + x \log b$$

2.1.4 Smoothing Methods

Random or/and coincidental fluctuations in weekly, monthly, seasonal or annual time series data can be removed or softened by smoothing methods. Six smoothing methods including single moving averages, Brown's simple exponential smoothing method, linear moving averages, Brown's linear exponential smoothing methods with single parameter, Holt's linear exponential smoothing with two parameters and Brown's quadratic exponential smoothing methods are mentioned in this part of the study [1].

2.1.4.1 Single Moving Averages

Estimation can be done by using arithmetic mean of number of certain (k) prior period of data. Single moving average method gives the same importance level to the past data for estimating future values.

$$y'_{t+1} = \frac{(y_t + y_{t-1} + \ldots + y_{t-k+1})}{k}$$

$$y'_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^{t} y_i$$

$$y'_{t+1} = \frac{y_t}{k} - \frac{y_{t-k}}{k} + y'_t$$

2.1.4.2 Brown's Simple Exponential Smoothing Method

It is a suitable method for time series that $y_1, y_2, \ldots\ldots\ldots, y_n$ has no significant trend or seasonal fluctuations. y'_t is the estimation value for the time t. y_{t-1} is the observation data for the time t-1. α is a smoothing constant. The constant α has the value between 0 and 1.

$$y'_t = \alpha y_{t-1} + (1-\alpha) y'_{t-1}$$

$$y'_t = y_{t-1} + \alpha(y_{t-1} - y'_{t-1})$$

$$y'_t = y'_{t-1} + \alpha e_t$$

2.1.4.3 Linear Moving Averages

The estimates are always below actual values when the moving averages method applies the data with an important trend. "Linear moving averages" method has been developed to deal with this situation. The main idea is the second moving average calculation.

$$y'_t = \frac{y_t + y_{t-1} + y_{t-2} + \ldots + y_{t-k+1}}{k}$$

$$y''_t = \frac{y'_t + y'_{t-1} + y'_{t-2} + \ldots + y'_{t-k+1}}{k}$$

$$a_t = y'_t + (y' - y''_t) = 2y'_t - y''_t$$

$$b_t = \frac{2}{k-1}(y'_t - y''_t)$$

$$\hat{y}_{t+m} = a_t + b_t m$$

The coefficient “m” is the forecast period to be estimated.

2.1.4.4 Brown’s Linear Exponential Smoothing Method with Single Parameter

Brown's Linear Smoothing Method has some similarity to the linear moving averages method with a single parameter. However, in the first smoothing value the difference between first and second smoothing values is added.

$$y'_t = \alpha y_t + (1-\alpha) y'_{t-1}$$

$$y''_t = \alpha y'_t + (1-\alpha) y''_{t-1}$$

$$a_t = y'_t + (y'_t - y''_t) = 2y'_t - y''_t$$

$$b_t = \frac{\alpha}{1-\alpha}(y'_t - y''_t)$$

$$\hat{y}_{t+m} = a_t + b_t m$$

2.1.4.5 Holt's Linear Exponential Smoothing Method with Two Parameter

Holt (1957) and Winters (1960) have extended Holt's seasonality method. The seasonal Holt-Winters method consists of the predicted equation and three sweetening equations–one for t level, one for bt and one for st, with sweetening parameters for the seasonal component. m is used to indicate the seasonality period, that is, the number of seasons in one year. For instance, m=4 and m=12 for on a quarterly and monthly data.

The nature of the seasonal component differs from two variations to this method. The additive method is preferred when the seasonal variations are approximately constant over the series while when the seasonal variations vary proportional to the series level, the multiplicative method is preferred. The seasonal component with the additive technique is expressed in absolute terms in the measurement of the series observed and seasonally adjusted by subtraction of the seasonal component in the level equation. The seasonal component amounts to approximately zero every year. The seasonal components are expressed in relative terms (percentages), with the multiplicative procedure, and the series are adjusted seasonally by the seasonal component. The seasonal component amounts to around m within each year.

It seems similar to previous method (Brown's Single Parameter Linear Smoothing Method). However, second smoothing is not used in the Holt linear exponential smoothing method. Tendency values are directly smoothed. This makes the method more flexible. α and γ parameters range from 0 to 1.

$$y'_t = \alpha y_t + (1-\alpha)(y'_{t-1} + b_{t-1})$$
$$b_t = \gamma (y'_t - y'_{t-1}) + (1-\gamma) b_{t-1}$$
$$\hat{y}_{t+m} = y'_t - b_t m$$

The parameters α and γ are the smoothing constants. They should be optimized for minimizing the sum of error squares.

2.1.4.6 Brown's Quadratic Exponential Smoothing Method

When the time series are curved shape (quadratic, third order or more) this technique is suitable for estimation. Third parameter is added in addition to the first two to the model. The equations for quadratic exponential smoothing are below:

$$y_t' = \alpha y_t + (1-\alpha) y_{t-1}'$$
$$y_t'' = \alpha y_t' + (1-\alpha) y_{t-1}''$$
$$y_t''' = \alpha y_t'' + (1-\alpha) y_{t-1}'''$$
$$a_t = 3y_t' - 3y_t'' + y_t'''$$

$$b_t = \frac{\alpha}{2(1-\alpha)^2}\left[(6-5\alpha) y_t' - (10-8\alpha) y_t'' + (4-3\alpha) y_t'''\right]$$

$$c_t = \frac{\alpha^2}{(1-\alpha)^2}\left(y_t' - 2y_t'' + y'''\right)$$

Estimation equation can be shown as below:

$$\hat{y}_{t+m} = a_t + b_t m + \frac{1}{2} c_t m^2$$

The selection of α coefficient can be done as the selection in previous methods.

2.2 Econometric Forecasting

The aviation traffic forecast derived from projections of post trends does not specifically take into account the ways in which the development of air traffic affects different economic, social and operational conditions. With the changes underlying these factors, it is desirable to try to consider these changes. Econometric predictions include the quantitative identification of the basis of historical data. On the one hand, the link between traffic on the other hand is used as a basis for estimating the underlying variables for the purposes of the traffic forecast by more important factors or variables influencing the traffic level. Five steps in an econometric forecast should be identified, although they are not carried out without regard to future components of an intertwined process.

- Relevant factors selection(independent variables)
- Data Collection
- The type of functional relationship between dependent variables and independent variables is specified.
- Statistical assessment and testing of the proposed link between adjusting and independent variables.
- Future development forecast of the variables that will be the result of the traffic forecast.

The type of functional relationship for a econometric traffic forecast needs to be developed through assessment and experimentation, and only through tests against actual historical data can the property of the relationship be determined empirically. Three alternative forms are suggested below in each case “y” is traffic, “x1” and “x2” are independent variables, and “a”, “b” and “c” are constant coefficient.

- Linear Y= a+bx1+cx2
- Multiplicative Y=ax1^b * x2^c]
- Linear-logY= loga + b*log x1 + c*logx2

The econometrics model is one of the common models in transport demand forecasting (Wadud, 2011). Furthermore, as Hill et al.'s (2001) postulated, in the last few decades the econometrics model has made several progress to include advanced models based on activity that utilize random utilities. Despite the popularity of the transport economy model, the model was not effectively used in

the aviation industry despite the Dios Ortuzar & Simonetti (2008) assumptions. The factors of attraction between two points include the **Gross Domestic Product (GDP), employment opportunities** in that area.

In addition, travel demand between two cities depends on the factors of attraction between the two points, the impedance among the two points and another dimension of demand projection when a country and a city is destination (Wadud, 2011; 2013). In addition, some of the exploratory factors affecting demand for air travel, such as GDP or income, air flights, and travel time, are similar to the study undertaken.

He added that two important factors could have a bearing on the forecasts, with regard to people, including exchange rates, population, the frequency of flights and the factors related to export and import. He also explained that population growth (the annual percentage of population) and population growth could mean between 15 and 64 years of age. However, the GDP and income parameter are also highlighted by Wadud(2013) as the main factors in demand estimates using a econometric method because they represent the size of a country economy. The demand for air travel between the two points also has an influence on parameters such as consumer price levels and exchange rates between two points.

In addition to the above factors, we suggest that several other factors be added to the economic model in previous literature. The demand for air travel is affected by airfare, as we have discussed. The fuel price of airfare is affected. Different econometric factors such as international dollar fuel prices, economic strength are further affected by the fuel price. Factors such as the country's Gross Domestic Product, the growth in GDP, GDP growth per capita, currency exchange rates relative to USD can measure economic strength at any time.

Another factor of attraction is **Consumer Price Index** (CPI). CPI measures price changes of the commodity and household purchasing market basket. The CPI is a statistical estimate based on the prices of a sample of representative goods, the price of which is regularly collected. This affects the country's macroeconomics and the power parity of purchase. It therefore helps to alter the population's affordability, leading to change in demand projections.

Adjusted net national income per capita (annual % growth) which takes depreciation effects into account can change the growth of GDP which in turn affects the demand forecast. Further factors such as unemployment rate and percentage of working population at any given time will also influence the

demand of air travel between the two points.

Age dependency ratio (% of working-age population) the ratio of dependents--people younger than 15 or older than 64--to the working-age population--those ages 15-64. Dependency ratios capture variations in the proportions of children, elderly people, and working-age people in the population that imply the dependency burden that the working-age population bears in relation to children and the elderly.Patterns of development in a country are partly determined by the age composition of its population. Different age groups have different impacts on both the environment and on infrastructure needs. Therefore the age structure of a population is useful for analyzing resource use and formulating future policy and planning goals with regards infrastructure and development.

Employment to population ratio, 15+, total (%) and **Fuel exports and imports** have been taken into consideration.

Regression techniques are used to built econometric based models given that we have to map the relevant factors collected to estimate the passenger demand.

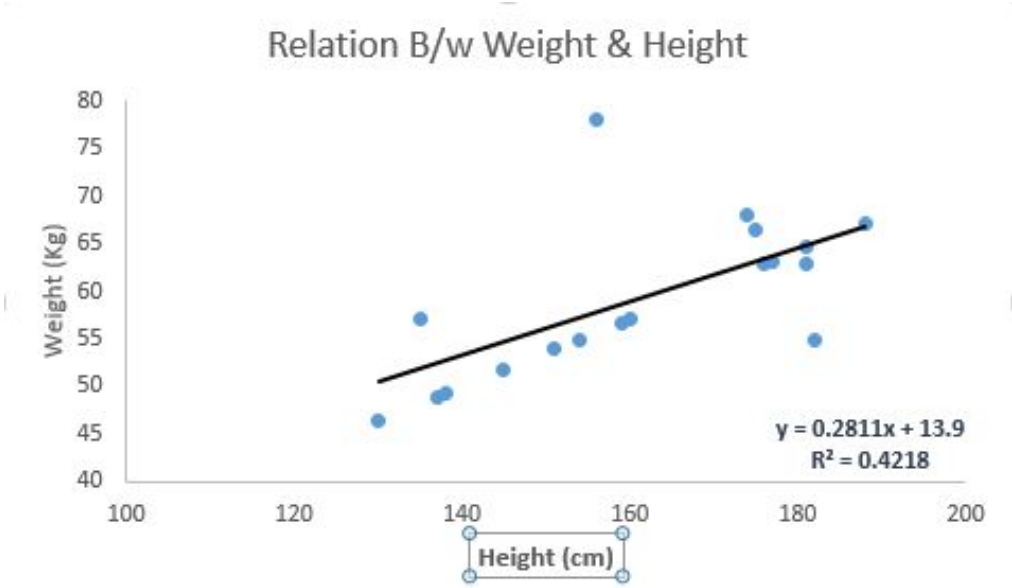
Traditional Regression techniques used are the following:

2.2.1 Linear Regression

This Modelling Technique is one of the most common. This technique involves continuous dependent variable; independent variable(s) can be continuous or discrete.

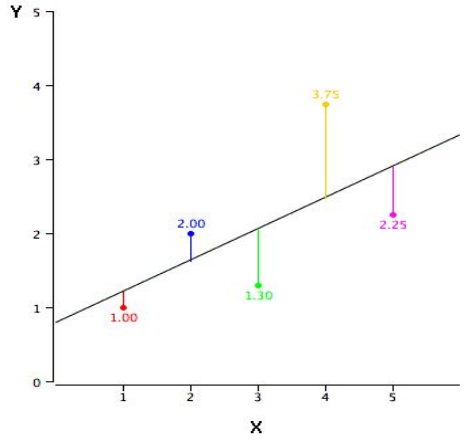
Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the target variable value based on independent values.



By Least Square Method, we get the best fit line. It is the best known way to fit a regression line. The best fit line for the data is calculated by minimizing the sum of the vertical deviations of squares from each data point to the line. If the differences first are squared, then the positive and negative values are neutralized if added.

$$\min_{w} ||Xw - y||_2{}^2$$



Using Root mean square error we can evaluate the performance of the model. There must be a

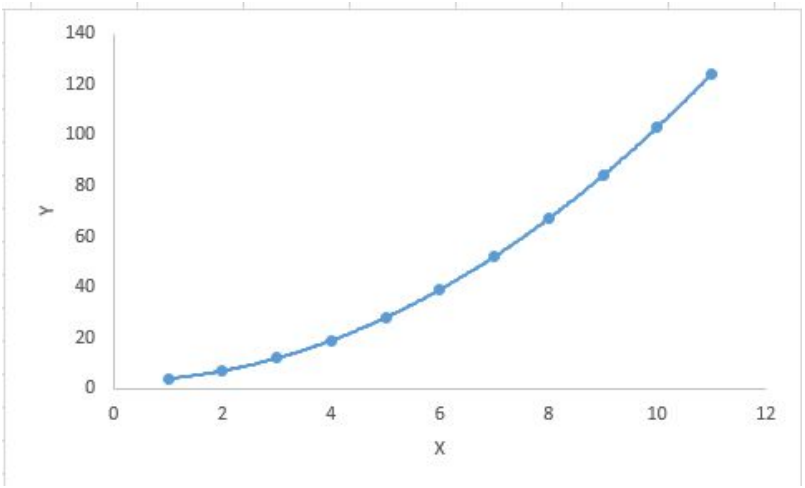
linear relation between independent and dependent variables in order to achieve this technique. It is a highly sensitive technique for outliers. The regression line and eventually the expected values can be dramatically affected. For several independent variables, a backward removal for the selection of the most important independent variables can be performed.

2.2.2 Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1.

$$y=a+b*x^2$$

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.



While there is a likelihood to fit a higher degree polynomial to get lower error, this can lead to overfitting where the model might not work on test data.

2.2.3 Ridge Regression

Ridge regression is a method that is employed when the data is highly correlated with multi-linearity. Although the least square estimates are impartial, in multicollinearity their variances are big, which differ far away from the true value.

By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

The complete equation of linear regression:

y=a+b*x+e (error term)

[e is the value needed to correct for a prediction error between the observed and

predicted value]

In a linear equation, errors can be decomposed into two components. First is due to the **bias** and second is due to the **variance**. Here, we'll discuss about the error caused due to variance.

Ridge regression solves the multicollinearity problem through shrinkage parameter λ (lambda).

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

In this equation, we have two components. First one is least square term and other one is λ (lambda)of the summation of β2 (beta- square) where β is the coefficient. This is added to least square term in order to shrink the parameter to have a low variance.

2.2.4 Lasso Regression

Similar to Ridge Regression, Lasso also penalizes the absolute size of the regression coefficients (Least Absolute Shrinkage and Selection Operator). It is also able to reduce the variability of

linear regression models and to improve their accuracy.

$$= \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

The regression of Lasso differs from the regression of the ridge so that absolute values instead of squares are used in the cost function. This leads to penalization (constraining of the total value of the estimates), which leads to the precise zero output of some parameter values. The larger the penalty, the lower the value, the lower it is. This results in the selection of the variable from given n variables.

3. Proposed Methodology

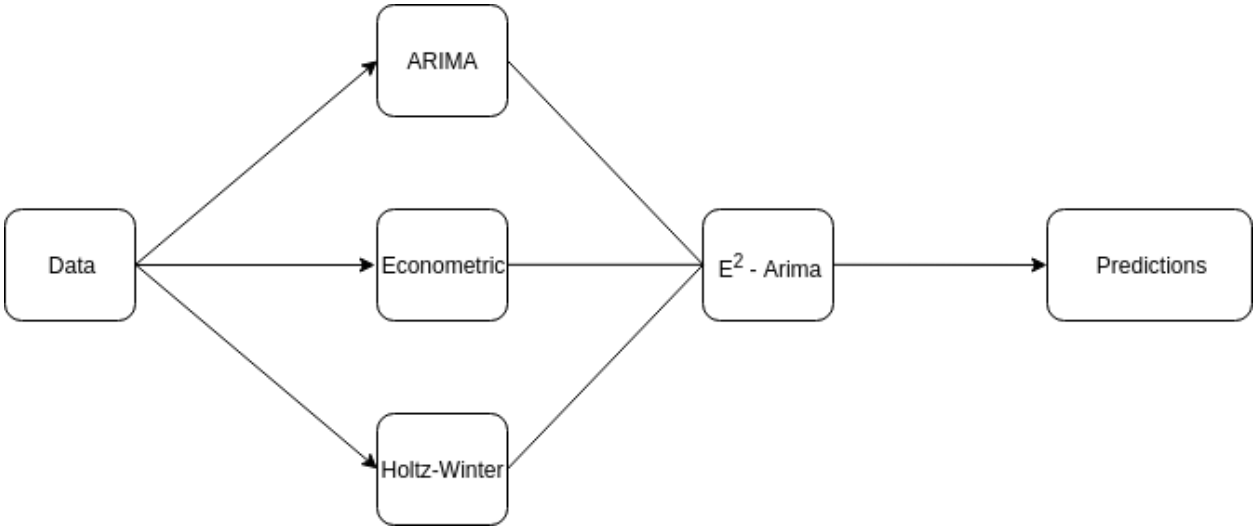
We propose an ensemble model which combines several individual models. We call this proposed model as **E-square ARIMA** which stands for Ensembled Econometric based AutoRegressive Integrated Moving Average model.

The forecasting methods employed for projecting activities at the airport should reflect not only the passenger activity time dependence structure, but also the underlying demographic and economic causal relations that drive passenger transport. When measuring passenger activity levels, demand and supply factors need to be accounted for.

Ensemble modelling is the process of synthesizing the results with two or more associated but different analytical models, in order to be more precise.

We tried two ensemble models -

1. ARIMA + Econometric
2. ARIMA + Holtz Winter + Econometric

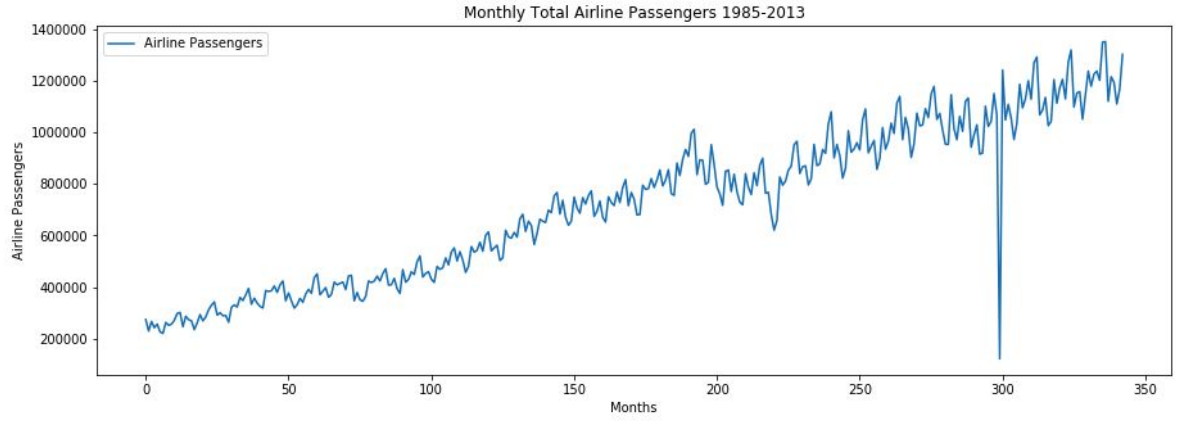


4. Experiments and Discussion

4.1 Time Series Model

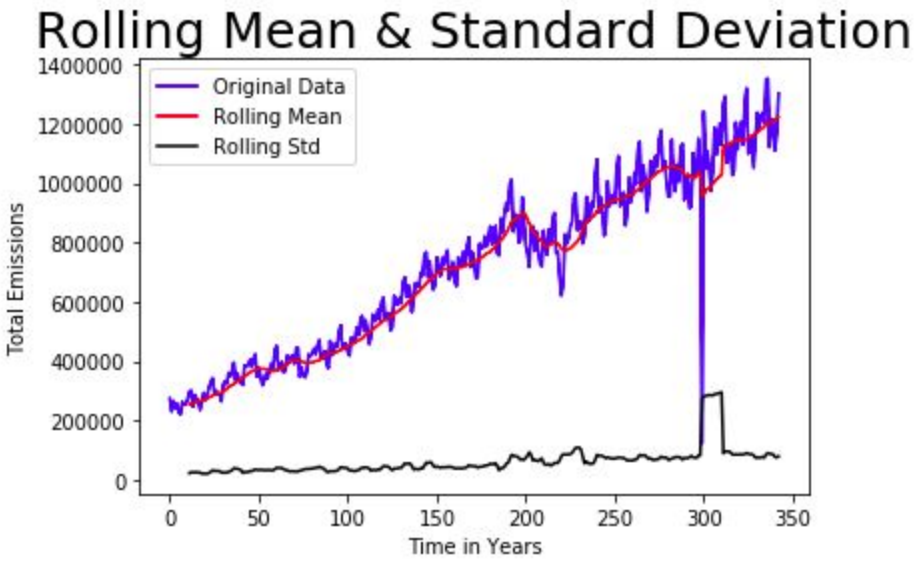
4.1.1 Dataset & Stationarity test

We can observe from the plot below that the monthly passenger traffic from 1985 to 2013 in Imperial terminal at Sydney International airport is fairly seasonal with a slight upward trend.



To implement the model, the parameters have to be found. This method is called as Box Jenkins method. The goal is to make the graph stationary and thereby find out the differencing and averaging done in the process and thereby set the parameters.

We can also see the trend in mean and standard deviation of the original data in the below graph.



Next we shall perform the Augmented Dickey Fuller test and Kpss test to see if the trend and level of the passenger traffic is stationary or non stationary.

4.1.2 Augmented Dickey Fuller Test

data: timeseries_data

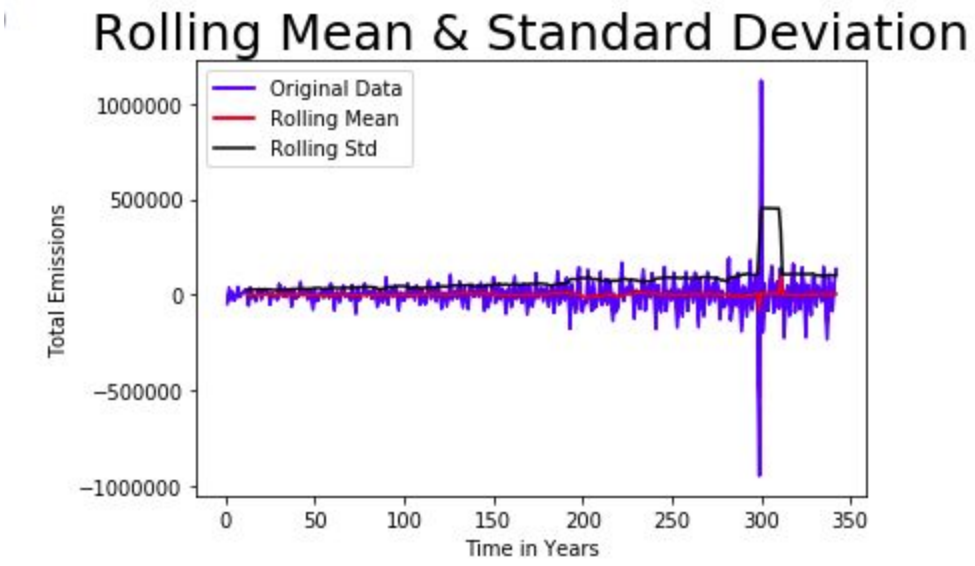
```
[ ]  df = df.dropna()
     TestStationaryAdfuller(df.iloc[:,0])
```

```
Test Statistic                  -0.138803
p-value                          0.945434
#Lags Used                      11.000000
Number of Observations Used    331.000000
Critical Value (1%)             -3.450262
Critical Value (5%)             -2.870312
Critical Value (10%)            -2.571443
dtype: float64
Weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary
```

The p-value of Augmented Dickey fuller test is 0.945434 and statistic test as -0.1388 showing it is non-stationary.

After single differencing the data the following graphs are obtained.

Mean and Standard Deviation of single differenced data can be observed below.

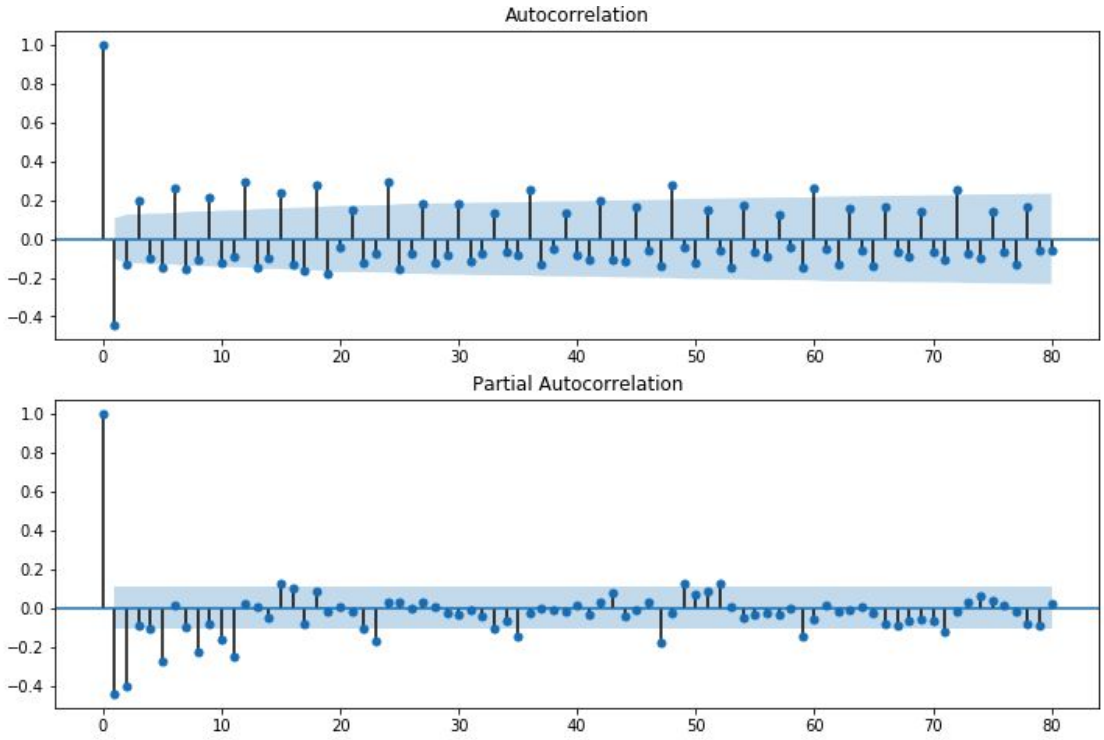


After single differencing, the p-value is less than 0.05 which implies the data is stationary. Hence we applied time series analysis on this data.

```
Test Statistic                -1.142099e+01
p-value                        6.874571e-21
#Lags Used                     1.000000e+01
Number of Observations Used    3.310000e+02
Critical Value (1%)           -3.450262e+00
Critical Value (5%)           -2.870312e+00
Critical Value (10%)          -2.571443e+00
dtype: float64
Strong evidence against the null hypothesis, reject the null hypothesis. Data has no unit root, hence it is stationary
```

4.1.3 ACF and PACF plot Evaluation

ACF and PACF plots of single differenced data are shown below.



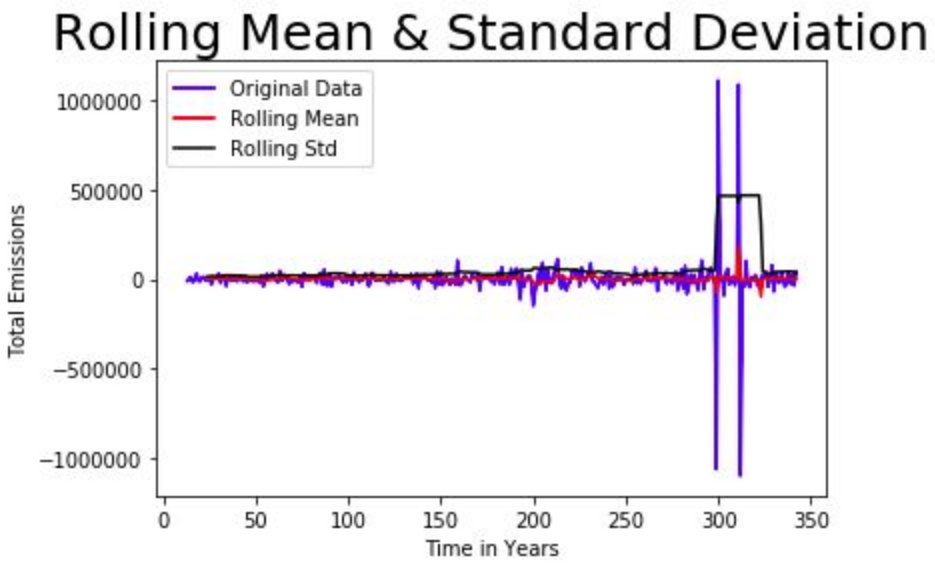
Similarly, plots for double differenced data are also obtained.

The p-value in ADF test of double differenced data is also <0.05.

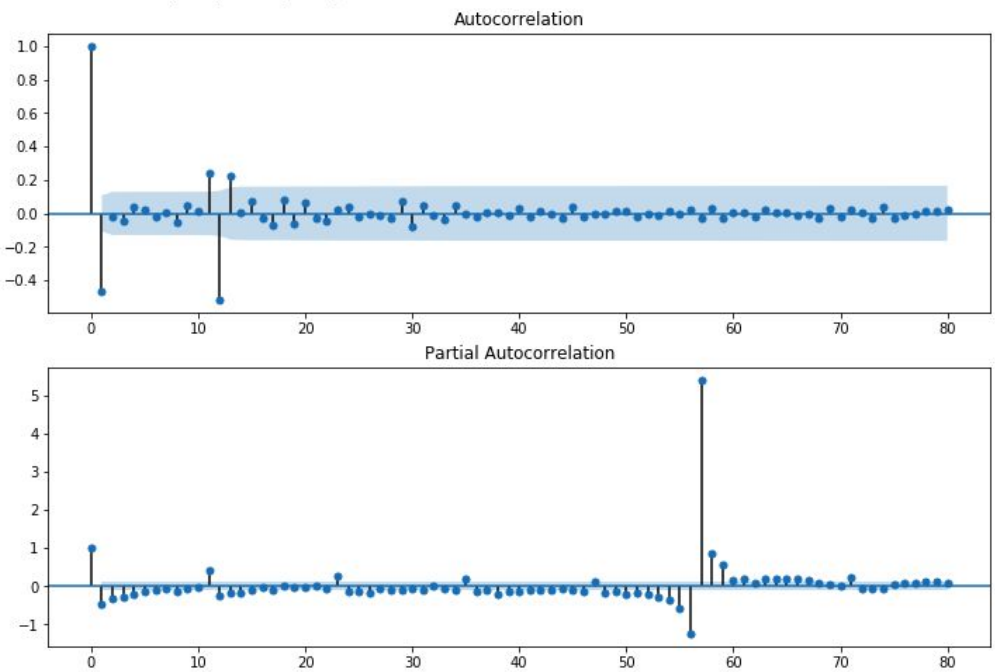
```
[ ]  TestStationaryAdfuller(diff12.iloc[:,0].dropna())

Test Statistic                -7.255687e+00
p-value                        1.733352e-10
#Lags Used                     1.400000e+01
Number of Observations Used    3.150000e+02
Critical Value (1%)           -3.451281e+00
Critical Value (5%)           -2.870760e+00
Critical Value (10%)          -2.571682e+00
dtype: float64
Strong evidence against the null hypothesis, reject the null hypothesis. Data has no unit root, hence it is stationary
```

Double differenced data plot & Mean and Standard Deviation plots
for the double differenced data are shown below:



ACF and PACF plots for double differenced data:



The Adfuller Test gave good results for double differencing also. Now, from the acf and pacf graphs, we found that the value of p=0,d=1,q=1.

The next part that comes into picture is the seasonality. As the data is monthly data, we need to do seasonal differencing and the differencing value is 12.

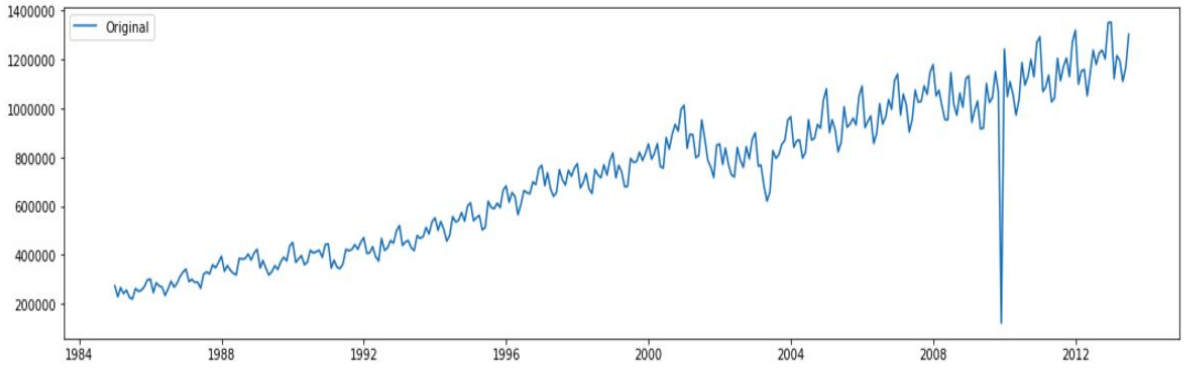
4.1.4 Candidate Model selection

The ACF and PACF of the double differenced data suggests that the following ARIMA model could be the best candidates : **ARIMA(0,1,1)[0,1,1][12]**

We have now built the model and need to perform residual diagnosis before we move on to predict using the model.

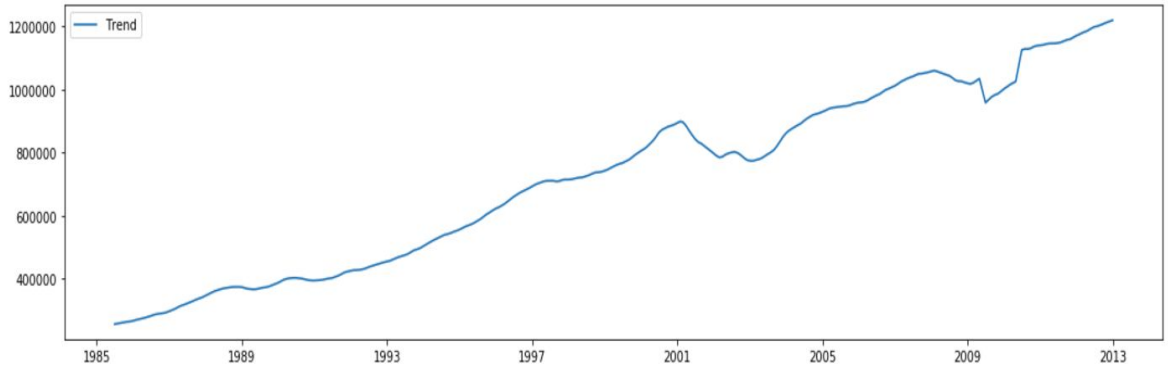
In order to perform residual diagnosis we need to get the residual component of the data.

We decomposed the data into 3 components which are trend, seasonality, residual.



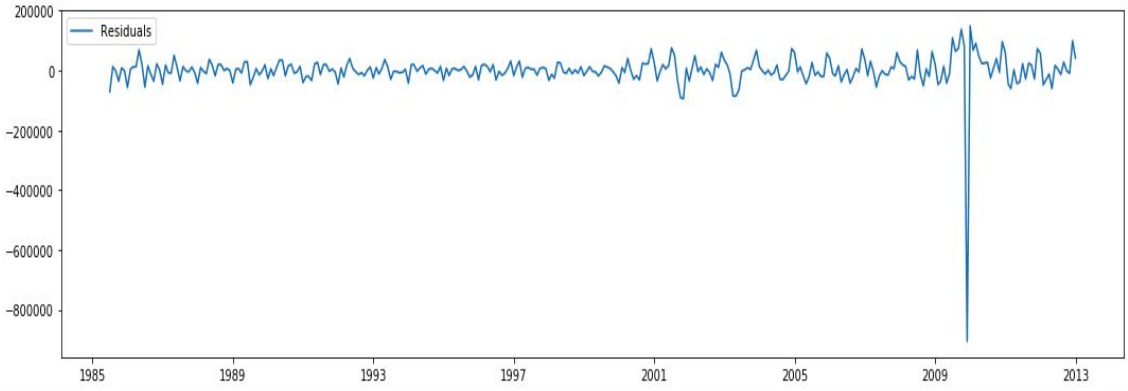
The graphs for each components can be seen below:

Trend component:

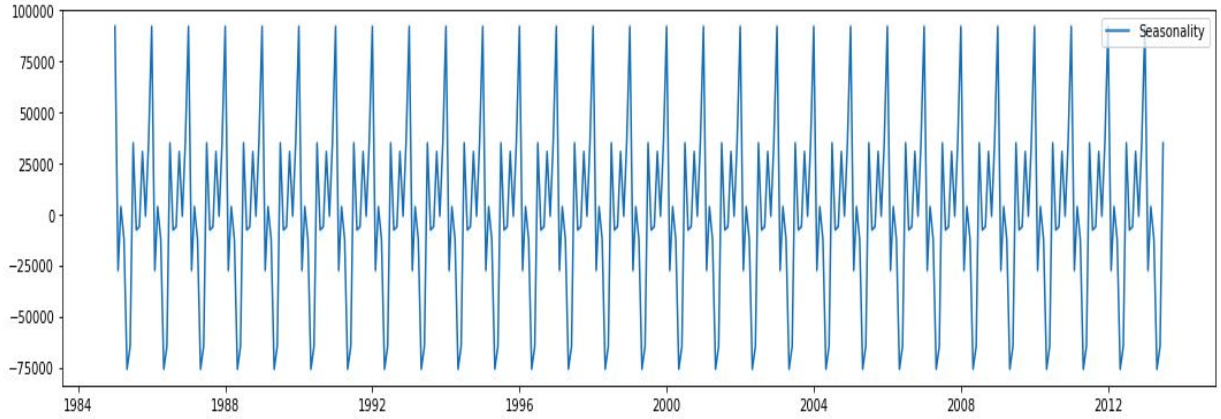


As we can see, the trend is an upward slope - naturally the demand for air transport has been increasing.

Residual component:



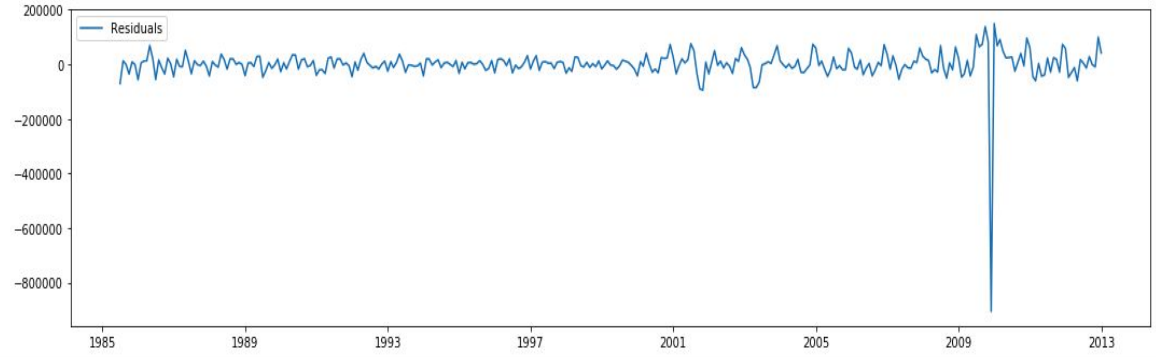
Seasonal component:

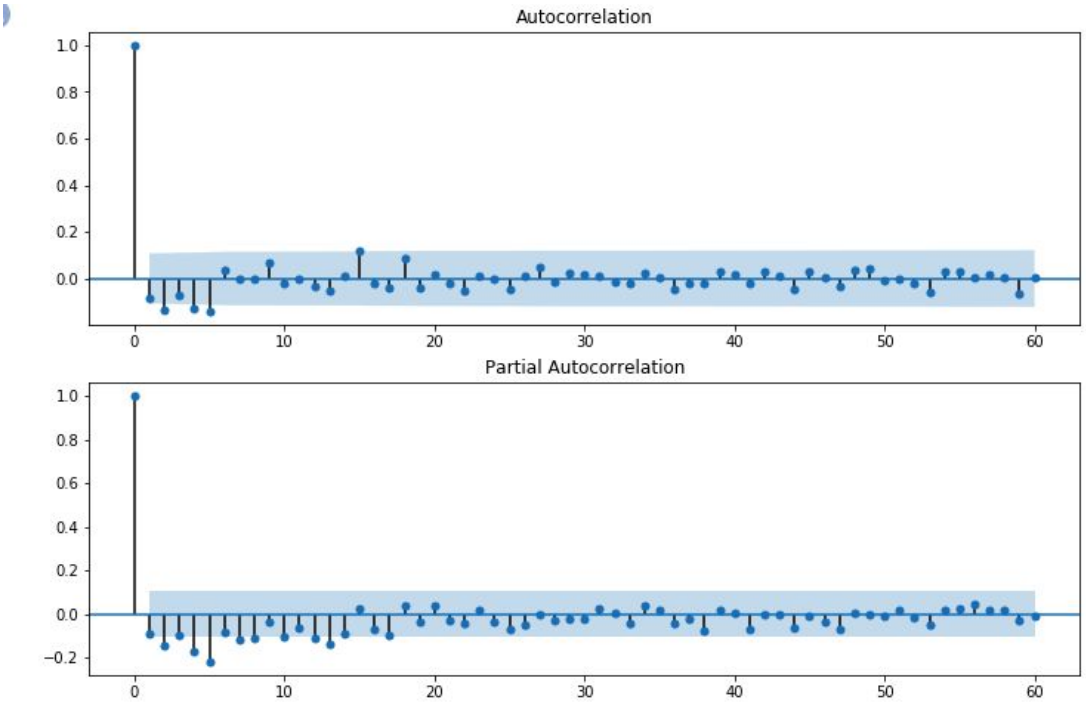


The seasonal graph has some ups and downs showing that there are peak times where a lot of tourists fly from other countries and local population also commutes more.

4.1.4 Residual Diagnosis of our model

The residuals seem fairly linear in distribution.





The residuals do not show any significant autocorrelation which means that our model is adequately built.

Let us further examine the residuals for test of significant autocorrelation by performing the ADF test.

```
[ ] TestStationaryAdfuller(residual.iloc[:,0].dropna())
```

```
Test Statistic                -7.738096e+00
p-value                        1.080073e-11
#Lags Used                     1.600000e+01
Number of Observations Used    3.140000e+02
Critical Value (1%)           -3.451349e+00
Critical Value (5%)           -2.870789e+00
Critical Value (10%)          -2.571698e+00
dtype: float64
Strong evidence against the null hypothesis, reject the null hypothesis. Data has no unit root, hence it is stationary
```

The P-value of the ADF test is low suggesting that the residuals are not auto correlated.

Let us go ahead and forecast with our model.

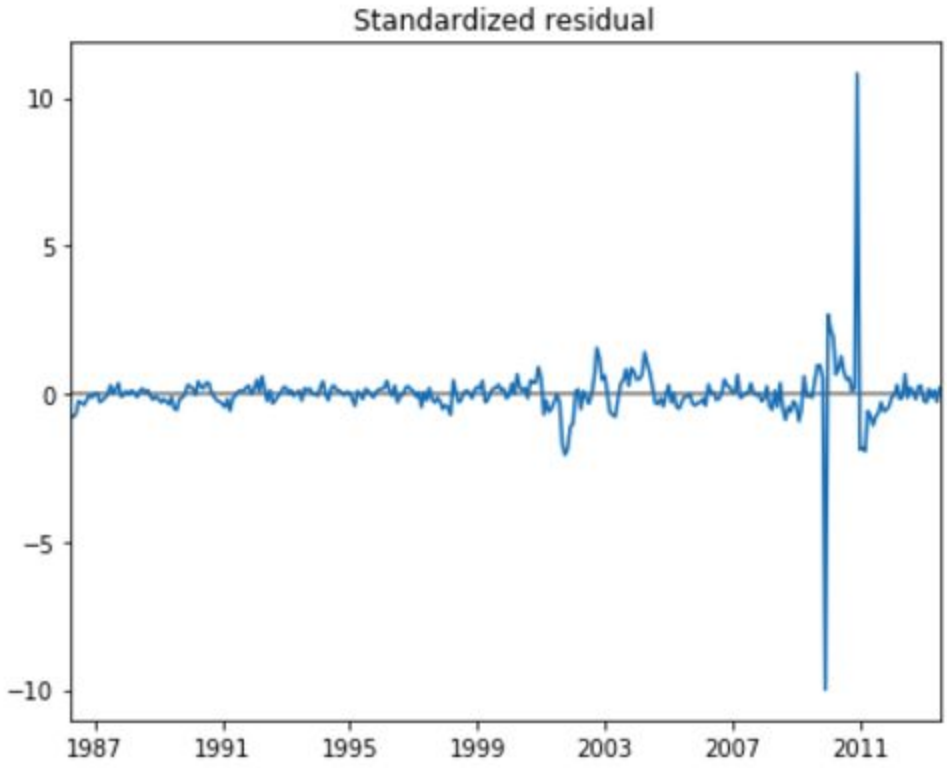
Statespace Model Results

Dep. Variable:	Total	No. Observations:	343
Model:	SARIMAX(0, 1, 1)x(0, 1, 0, 12)	Log Likelihood	-4234.227
Date:	Thu, 18 Apr 2019	AIC	8472.454
Time:	05:14:52	BIC	8480.040
Sample:	01-01-1985	HQIC	8475.481
	- 07-01-2013		
Covariance Type:	opg		

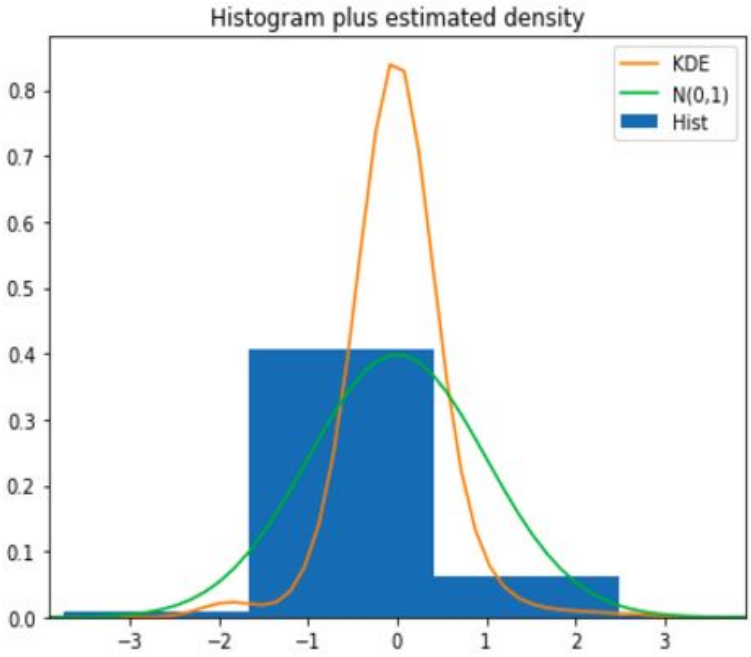
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8222	0.019	-42.854	0.000	-0.860	-0.785
sigma2	1.037e+10	3.06e-13	3.39e+22	0.000	1.04e+10	1.04e+10

Ljung-Box (Q):	105.17	Jarque-Bera (JB):	94515.37
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	38.05	Skew:	0.98
Prob(H) (two-sided):	0.00	Kurtosis:	86.14

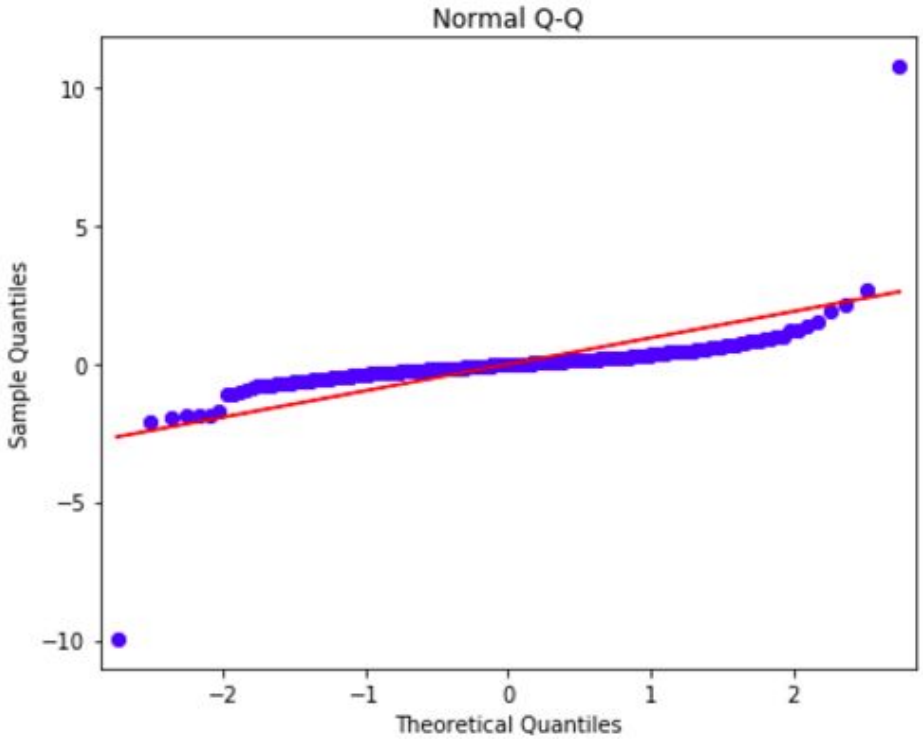
Standardised Residual



Histogram of model Residuals

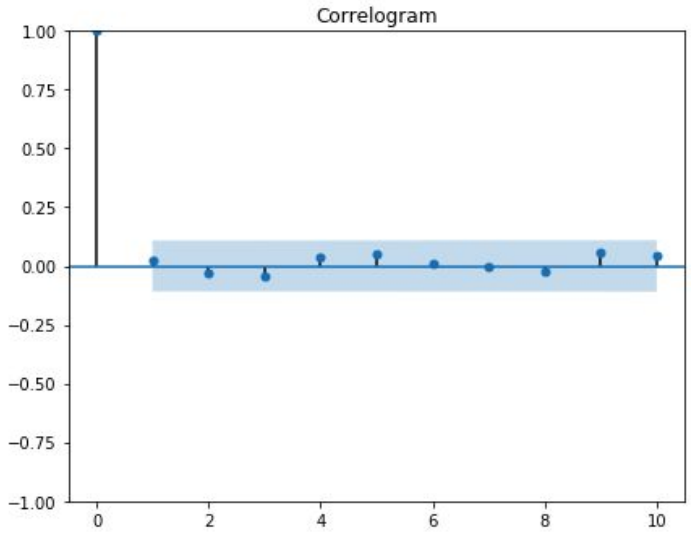


Normal Q−Q Plot

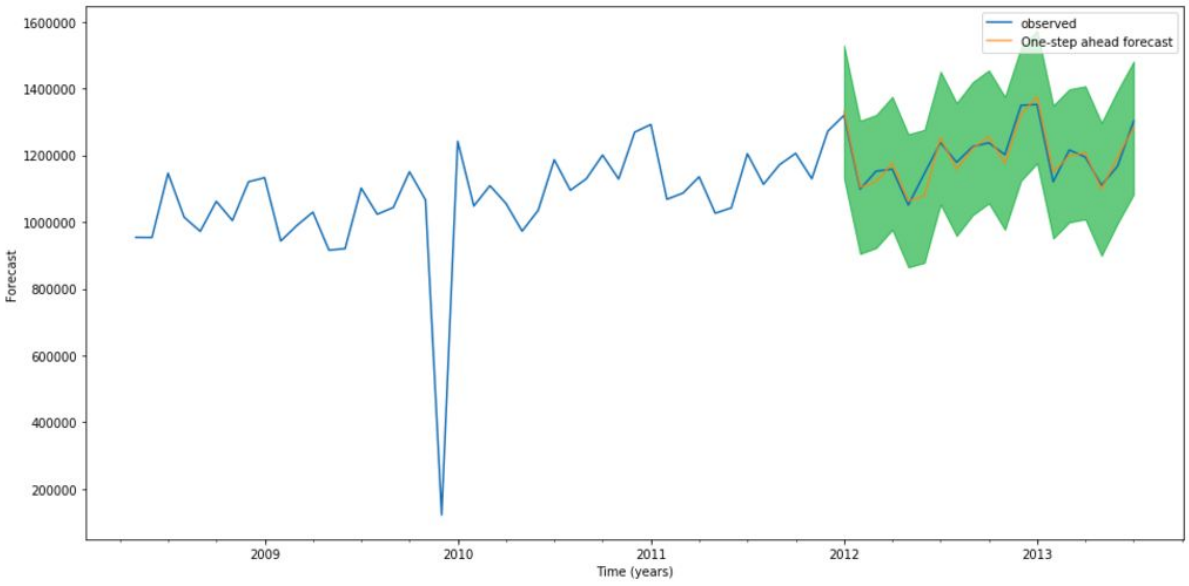


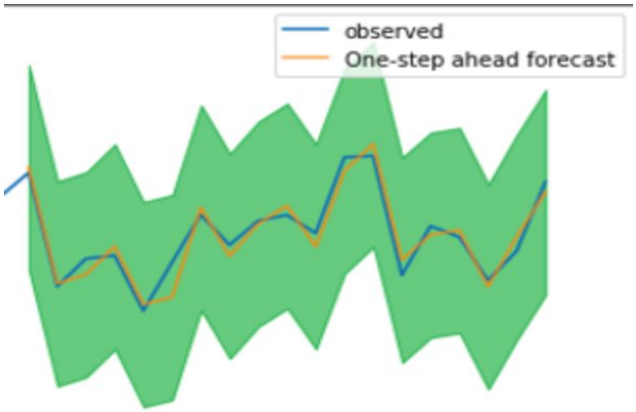
The standard assumption in linear regression is that the theoretical residuals are independent and normally distributed. We can see from the above histogram and the Q-Q plot, that the residuals confirm to this assumption of normality.

Correlogram



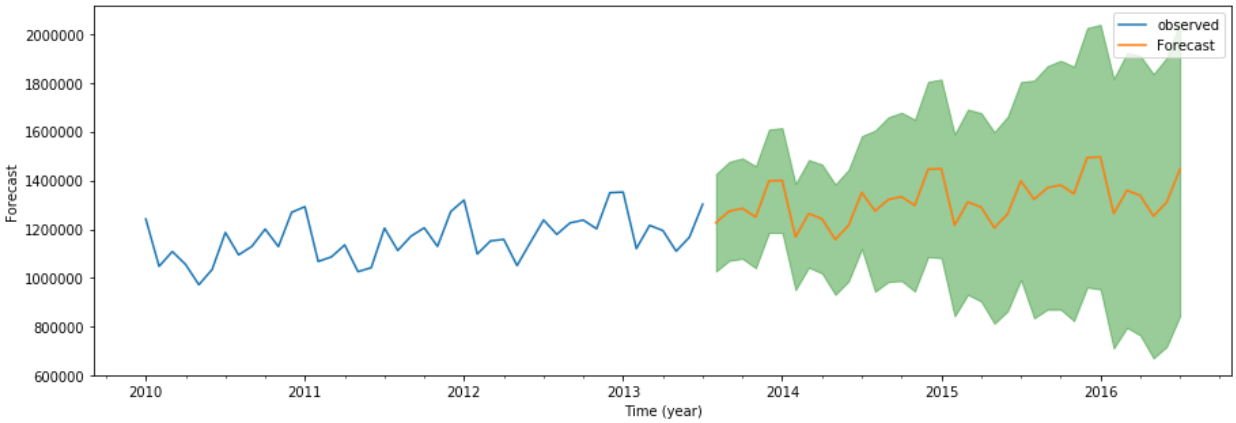
4.1.5 Passenger traffic forecast with 95% confidence intervals





We can see the forecast which is very close to the observed values.

The green shade represents the deviation of forecast values. It conforms to 95% confidence interval.



Above diagram shows the forecast for the coming 3 years.

4.1.6 Model Summary

```
The Mean Squared Error (MSE) of the forecast is 617864761.09
The Root Mean Square Error (RMSE) of the forcast: 24856.8856
The Normalised Root Mean Square Error (NRMSE) of the forcast: 0.0207
```

4.2 Econometric Modelling with Econometric Variables

4.2.1 Dataset

or_Name	Adjusted net national income per capita (annual % growth)	Age dependency ratio (% of working-age population)	Consumer price index (2010 = 100)	GDP per capita growth (annual %)	GNI per capita growth (annual %)	Imports of goods and services (% of GDP)	Inflation, consumer prices (annual %)	Population ages 15-64 (% of total)	Population growth (annual %)	Urban population (% of total)	Unemployment Rate	Fuel Price	Exchange Rate
85	0.427975	51.186393	40.816857	3.819880	3.653830	17.309665	6.734694	66.143516	1.367346	85.582	8.258	1.2600	0.701
86	1.215482	51.012782	44.510926	2.363583	2.026483	18.086141	9.050351	66.219557	1.638989	85.552	8.117	1.2300	0.670
87	0.169411	50.603101	48.309053	0.958029	0.905815	17.094011	8.533022	66.399693	1.520987	85.522	8.108	1.2015	0.701
88	5.832533	50.107027	51.795005	3.965963	3.875367	16.644749	7.215940	66.619136	1.636207	85.493	7.208	1.0049	0.790
89	2.984112	49.742783	55.697190	2.124362	1.229374	16.979151	7.533903	66.781180	1.692567	85.463	6.158	0.9605	0.788
90	-0.112070	49.592036	59.781478	2.043105	1.269443	17.070363	7.333022	66.848477	1.479978	85.433	6.942	1.2120	0.779
91	-1.547560	49.577589	61.680541	-1.649127	-2.158434	16.191721	3.176675	66.854934	1.274578	85.403	9.592	1.0480	0.778
92	-0.963033	49.759815	62.304891	-0.768781	-0.592351	16.418939	1.012231	66.773589	1.213391	85.285	10.742	1.0240	0.731
93	3.676336	50.030597	63.397503	3.026189	4.248443	17.894616	1.753653	66.653071	0.978337	85.157	10.858	0.9880	0.678
94	3.060820	50.215168	64.646202	2.883717	3.354991	18.521413	1.969635	66.571174	1.058509	85.028	9.717	0.9540	0.734

We have collected dataset for each econometric variable from various sources available on the internet.

4.2.2 Dataset Description

Indicator_Name: year of which the collected data represents.

Adjusted net national income per capita (annual % growth):

Adjusted net income complements gross national income by providing a broader measure of domestic income, which represents the depletion of natural resources, in assessing economic progress. The national adjusted net income is calculated by withdrawing from GNI a charge for fixed capital consumption and natural resource depletion. The reduction in natural resource depletion, which covers net forest depletion, energy depletion and mineral depletion, reflects the decline in asset values associated with natural resource extraction and collection. This is similar to fixed assets depreciation. Adjusted National Net Revenue growth rates are calculated from the

consistent price series deflating with the deflator of gross domestic (formerly domestic) expenditure.

Age dependency ratio (% of working-age population):

The population of working age is characterized by 15 to 64 years of age. The key indicator for the company is the extent of the workforce aged 15-64, which are used. The age dependency ratio is the share of the working-aged population (individuals younger than 15 or more than 64).

Consumer price index (2010 = 100):

The CPI is a statistical estimate based on the prices of a sample of representative items whose prices are regularly collected. Sub-indices and sub-sub-indices are calculated by combining the total Index to produce the weights reflecting their share in the total consumer expenses covered by the index for various categories and sub-categories of goods and services. The price indices are among several that the majority of national statistical agencies have calculated. The change in the CPI's annual percentage is used as inflation measure. The real value of wages, payouts, pensions, for price control and for deflating monetary magnitudes to display changes in real values can be used as an index (i.e. adjustment to effect inflation). In most countries, the CPI is one of the most closely monitored national economic statistics, together with the population census.

GNI per capita growth (annual %):

GNI per capita annual growth rate based on constant local currency. Constant 2010 U.S. dollars are the basis for the aggregates. GNI per capita is gross, mid-year national income. GNI (formerly GNP) is the value added amount for all resident producers, plus any product taxes (less subsidies) that are not included in evaluation of output plus net primary (employee and property income compensation) receipts from abroad.

Unemployment Rate :

A key indicator for job market performance is the unemployment rate. Under the U.S. When a worker loses his employment, his families lose their wage and, for the goods and services that could otherwise have been produced, the nation as a whole loses their contribution to the economy, Bureau of Labor Statistic (BLS). In order to forecast passing activity in the airport under consideration, the unemployment rate is used as an indépendent / explainable economic variable.

GDP Per Capita Growth (annual %):

GDP refers to the total value, for example quarterly or yearly (in contrast to provisional or progressive) of final goods and services produced within the borders of a country over a certain

calendar period. Although GDP is a country's economic activity, per capita GDP is a better indicator of the standard of living of a country since it adapts for its people.

Jet_Fuel :

In the assessment of the projections for passenger activity, the volatility associated with the price for jet fuel is an important supply side factor. In 2000, the price per gallon of jet fuel was $0,76 and is currently $1,91 per gallon.

Imports of goods and services (% of GDP):

The value of all products and services received from the rest of the world is represented by imports of goods and services. These include the value of goods and services, such as communications, construction, financial, information, business, personal and government services, transfers, travel fees, royalties and other services. They exclude employee and investment revenue compensation and transfer payments (formerly referred to as factor services).

Inflation, consumer prices (annual %):

An annual percentage cost change for the aperture consumer to purchase a panel of products and services that can be fixed or changed at specific intervals, as is the case every year, refers to the inflation as measured by a consumer price index.

Exchange Rate:

The rate at which one currency is exchanged for another currency is exchanged. It is also seen in relation to another currency as the value of one country's currency.

Urban population (% of total):

The urban population, defined by country, describes the percentage of the total urban population.

Population ages 15-64 (% of total):

The total population between 15 and 64 years of age represents a percentage of the total population. Population is based on the de facto population definition, which includes all residents irrespective of their legal status or citizenship.

Population growth (annual %):

Year t annual population growth is the exponential mid-year population increase of yeart-1 to t, as a percentage. The populations are based on a population de facto definition, which covers all residents irrespective of their legal status or citizenship.

Population, total:

Increases in human populations can influence natural resources and social infrastructure, either as a result of immigration or more births than deaths.

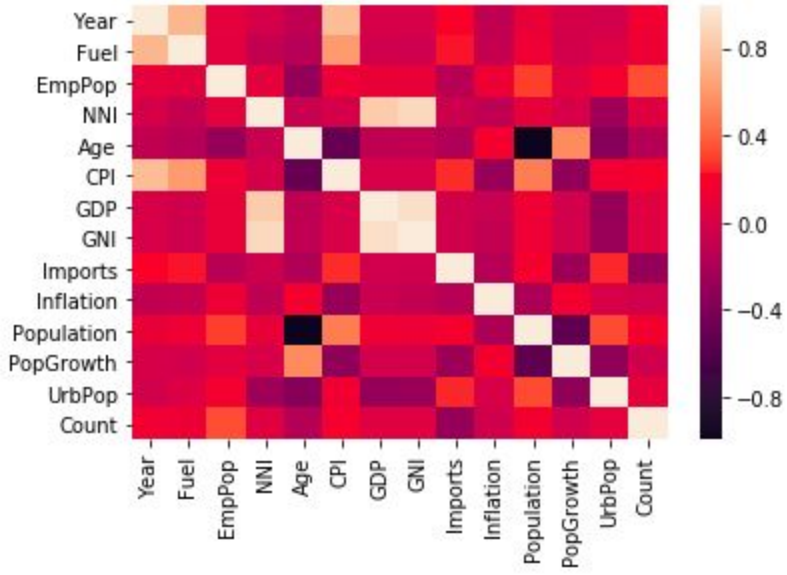
Fuel exports (% of merchandise exports)

We were unable to get the fuel price of the cost of jet fuel for individual countries when making the model for multiple countries. So, instead we collected the fuel exports of a particular country which might in a way differentiate between countries

Fuel imports (% of merchandise imports)

This data was collected and the explanation for it is same as before.

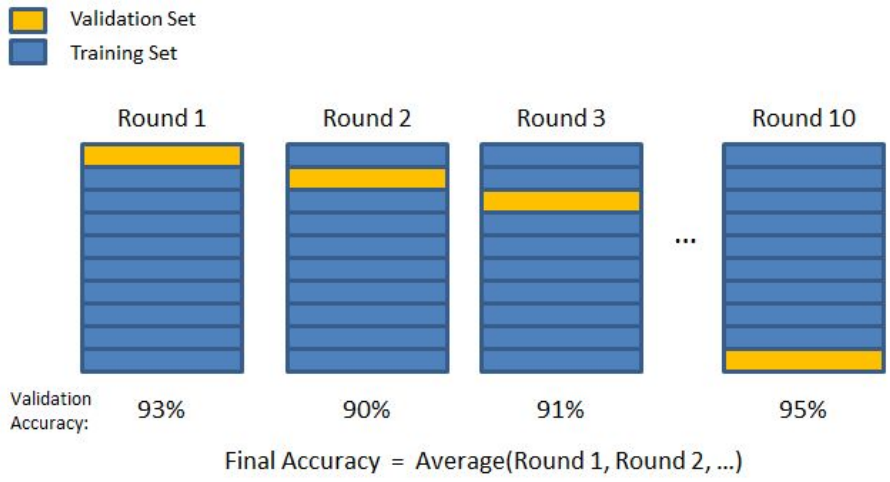
4.2.3 Correlation Matrix



4.2.4 Cross Validation

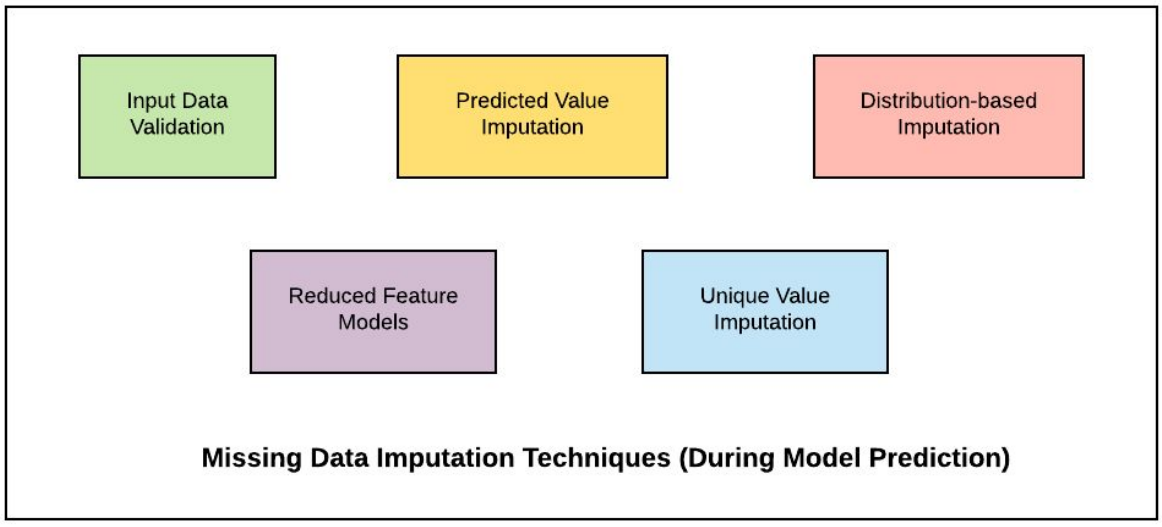
By reducing the data from training we can lose important data set patterns, which in turn increases bias-related errors. K-fold cross validation is applied in order to have ample data for model training and validation. We took K=10 in our case.

Each data point is validated exactly once and has to be in a set of k-1 times. This reduces deficiency considerably because we make use of most of the fitting data and also reduces the overfitting considerably, as most data are also used as validators.



4.2.5 Missing Value Imputation

The data was compiled from different sources and some of the values were missing over the years. There are different techniques to deal with missing data some of them are mean median imputation etc..



Some of the above mentioned attributes which have missing values are **Adjusted net national income per capita, Exchange rate etc.**
Now for a year which has more than 2 missing values have been removed and not used
And the others are replaced with mean of the value of that particular country

4.2.6 Data Normalization

Normalization is a technique that is frequently used in machine learning data preparation. The goal of normalization is to change the values of numeric columns on a common scale in the dataset without distorting the values. Every data set requires no normalization for machine learning. Min-max standardization was used.

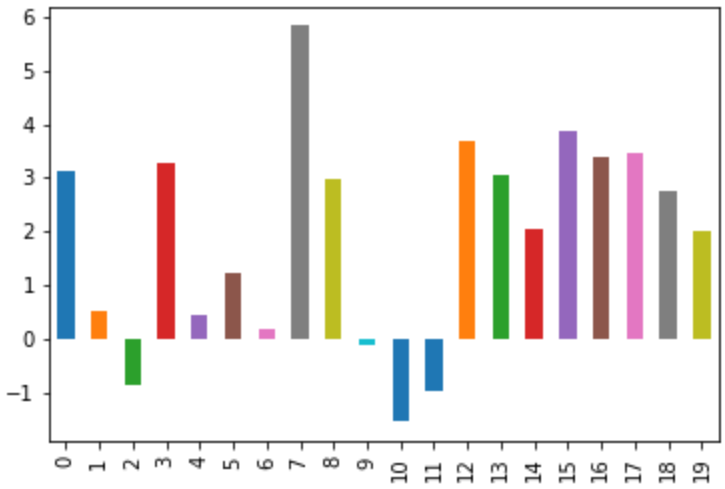
Min-max normalization is one of the most common method of data normalization. Each function is converted into a minimum value of 0, the maximum value is transformed into a 1 and each other is transformed into a decimal between 0 and 1.

For example, if the minimum value of a feature was 20, and the maximum value was 40, then 30 would be transformed to about 0.5 since it is halfway between 20 and 40. The formula is as follows:

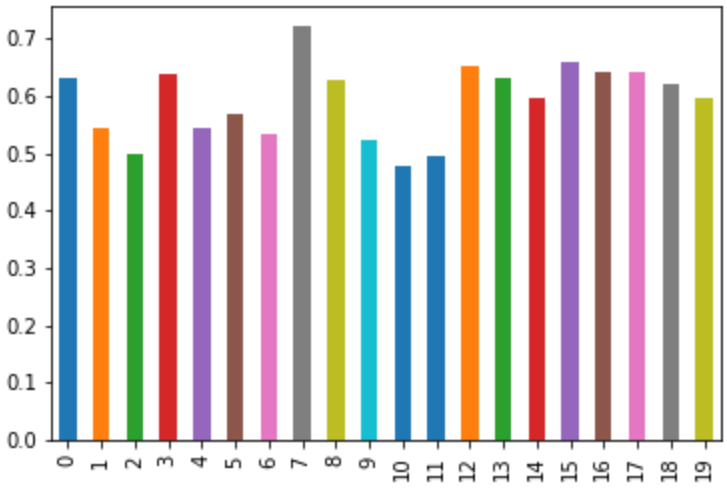
$$x\_norm = (x - min\_value) / (max\_value - min\_value)$$

There is a quite significant downside to Min-max normalization: it does not deal with outliers very well. Let's look at one attribute before and after normalization in the above data.

Here this is the original data of **Adjusted net national income per capita** over 20 years and we can see some of them are negative as well



Now after normalization we have the same values all lying between 0 and 1 as it is min-max normalization.



But since all the attributes we used in the final model are of within range of each other our results of prediction on unnormalized data were better.

4.2.7 Model summary

We performed multiple variants of econometric model. First, we applied the econometric model only one airport of Australia (Sydney). As the data is very small, the chances of the model not working properly is very high.

Hence, a thought popped out and we built the model for multiple countries whom we thought are alike and going forward with the assumption that the weights of each factor is same.

We took the data of Australia, United States and United Kingdom and applied the model on it using ridge regression.

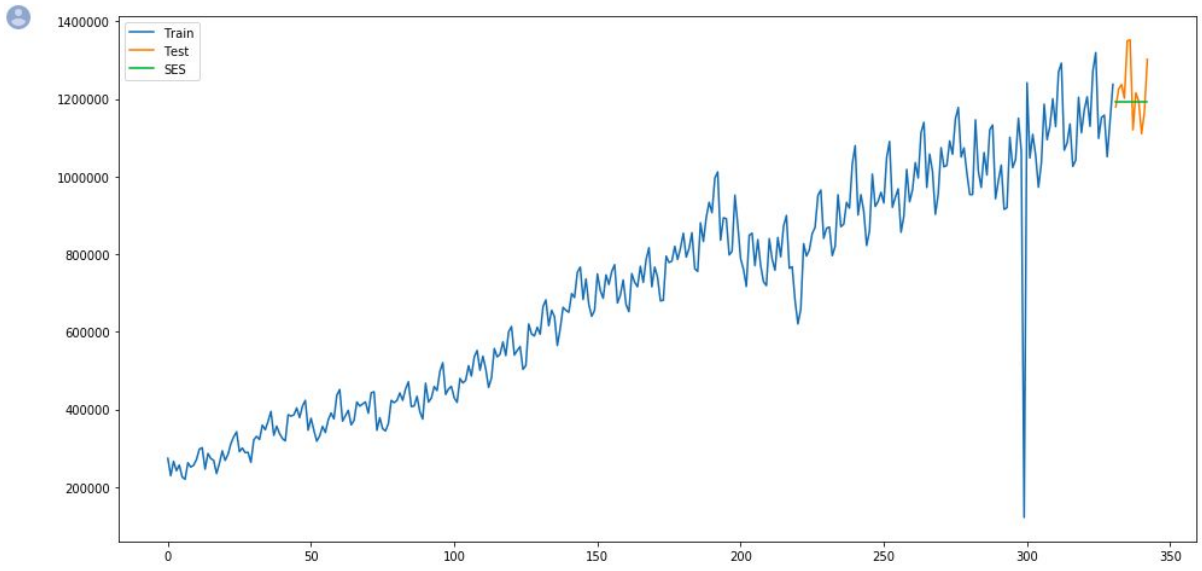
After doing this, from historical data, we found out an airport's share of the total country aviation and then find out the estimate of the airport.

Now, we just got the annual prediction for an airport. To get the monthly data, we have previously extracted the seasonality plot before in arima. So, we find out each month's share by normalising the seasonal graph and find out each month's data as we did in ARIMA.

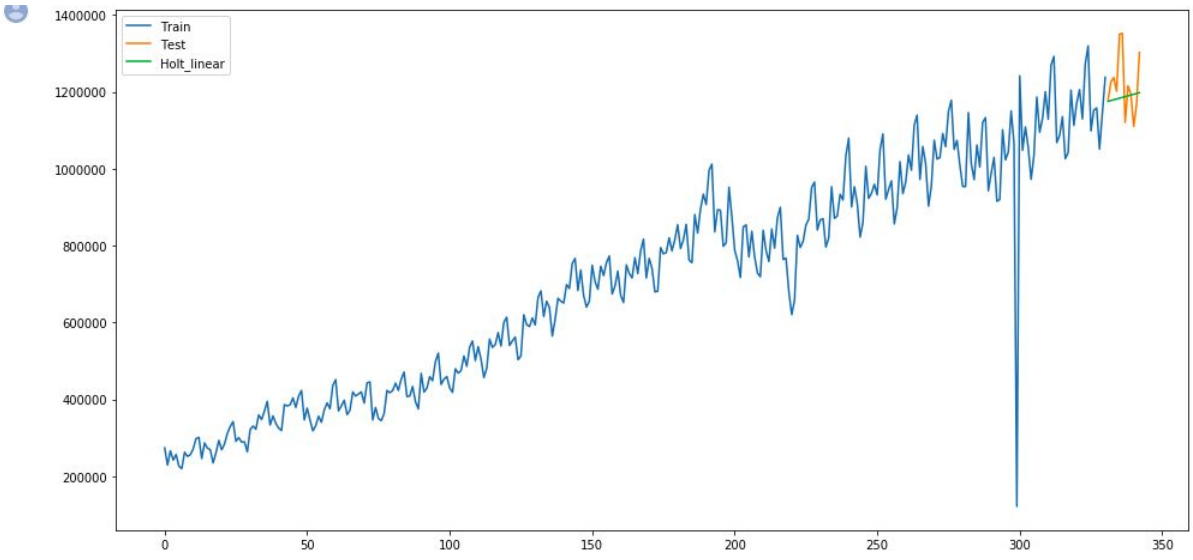
We used Ridge regression analysis to build the model.

4.3 Smoothing Methods

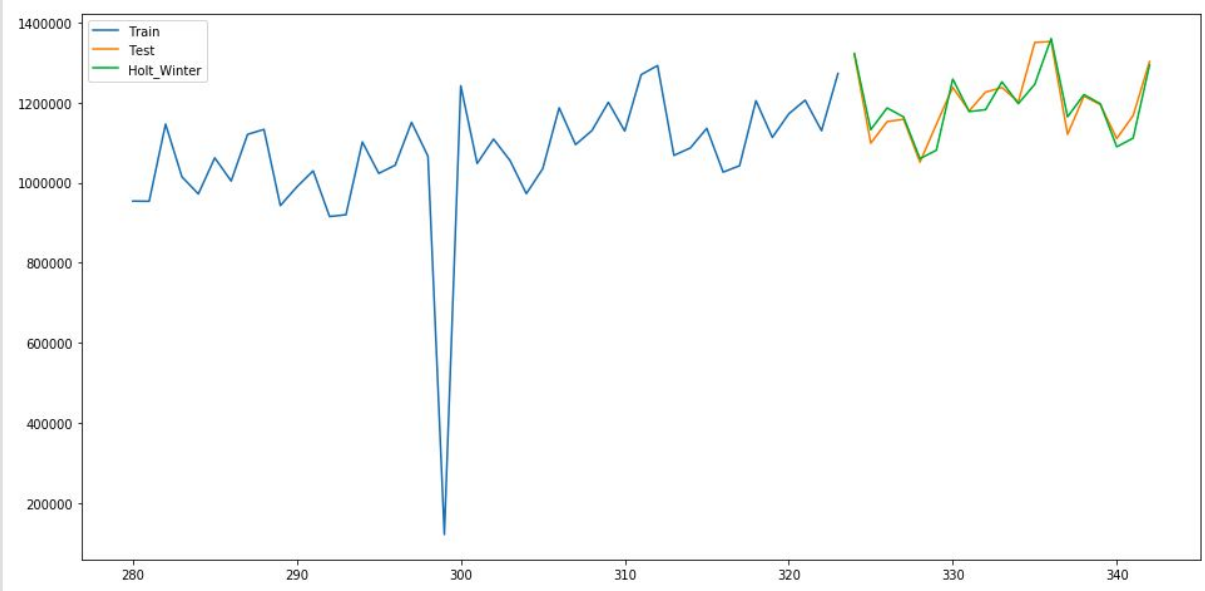
4.3.1 Simple Exponential Smoothing



4.3.2 Holtz Linear



4.3.3 Holtz Winter Smoothing



4.4 Results

Model Used	**Normalised RMSE Value**
Seasonal ARIMA	0.022
Holtz Winter	0.034
Econometric Model (single airport)	0.102
Econometric Model (share analysis)	0.070
Ensemble of ARIMA + Econometric	0.0173
Ensemble of ARIMA + Econometric + HOLTZ	0.0199

Econometric Modelling Measures:

Measure	Value
Root Mean Squared Error (RMSE)	16302829.985037228
Mean Squared Error (MSE)	437438150798814.56
Explained Variance	1.3284523332145874
Mean Squared Log Error	0.1598784465768743
R2	7.298794464399833

5. Conclusion and Future Work

A simple averaging ensemble model that takes the individual forecasts from Arima and Econometric modelling, averages them to produce an estimated forecast is so far the best method in terms of cross validation rmse.

We think that the ensemble model gave out the better results because both of the models were lagging with some drawbacks and the averaging model was able to minimise it.

The drawback of time series model is, it will not be able to find out the sudden dips or peaks due to sudden changes in the economy, hence it needs some outside data.

The drawback of econometric model is, it will be not be able to capture the seasonal behaviour when we want to find the predictions for monthly data as some of the economic factors that we are using are annual.

The final factors that have been used in econometric model are employment ratio, consumer price index, GDP per capita, Age dependency ratio (working class), Adjusted net national income per capita, Fuel imports and exports (% of merchandise), Inflation of consumer prices, Population aging in 15-64, Population growth and Urban population percentage.

Further research can be done on this and our aim was to focus on just the passenger data and hence, if we do the similar analysis on the airport freight data, with the forecast for the coming years, renovation plans can be done. Airport Personnel and FAA tower staffing requirements can be managed accordingly.

Bibliography

[1] Adrangi, B., Chatrath, A. and Raffiee, K., (2001). The demand of US air transport service: a chaos and nonlinearity investigation, Transportation Research, Part E, 37, 337-353

[2] Orhunbilge, N., 1999, Time Series Analysis,Forecasting and Price Index. Istanbul University, School of Business Press, Publication No: 277, 11–130. (In Turkish)

[3]AIRPORT_COOPERATIVE_RESEARCH_PROGRAM_FAA-
http://www.trb.org/Publications/Blurbs/158684.aspx

[4]Önder, E. and Kuzu, S. (2014). Forecasting air traffic volumes using smoothing techniques. Journal of Aeronautics and Space Technologies, 7(1): 65-85.

[5]https://www.analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/

[6]Wadud, Z., 2011. Modelling and Forecasting Passenger Demand for a New Domestic Airport with Limited Data.Transportation Research Record:Journal of the Transportation Research Board, pp. 59-68.

[7] de Dios Ortúzar, J. & Simonetti, C., 2008. Modelling the Demand for Medium Distance Air Travel with the Mixed Data Estimation Method. Journal of Air Transport Management, 14(6), pp. 297-303.

[8] Wadud, Z., 2013. Simultaneous Modelling of Passenger and Cargo Demand at an Airport. Transportation Research Record: Journal of the Transportation Research Board, pp. 63-74.

[9] Profillidis, V. A., 2000. Econometric and Fuzzy Models for the Forecast of Demand in the Airport of Rhodes. Journal of Air Transport Management, Volume 6, pp. 95-100.

[10] Suryan Viktor. Econometric forecasting models for air traffic passenger of indonesia. Journal of the Civil Engineering Forum, Vol 3,Iss 1, Pp 33-44 (2017), (1):33, 2017.

[11] Rupantar Rana . FORECASTING AVIATION ACTIVITY Los Angeles International airport