International Journal of Engineering &Technology, 7 (3.12) (2018) 239-243

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Airport Trends Analytics Engine using the ARIMA Model

Chitransh Rajesh[1*], Yash Jain[2], J. Jayapradha[3]

[1,2]Bachelor of Technology – Computer Science and Engineering
[3]Assistant Professor – Department of Computer Science and Engineering
[1, 2, 3]SRM Institute of Science and Technology, Chennai-603203, India
**Corresponding Author Email: [1]chitransh010@gmail.com, [2]yashfit2012@gmail.com, [3]jayapradha.j@ktr.srmuniv.ac.in*

Abstract

Data Analytics is the process of analyzing unprocessed data to draw conclusions by studying and inspecting various patterns in the data. Several algorithms and conceptual methods are often followed to derive legit and accurate results. Efficient data handling is important for interactive visualization of data sets. Considering recent researches and analytical theories on column-oriented Database Management System, we are developing a new data engine using R and Tableau to predict airport trends. The engine uses Univariate datasets (Example, Perth Airport Passenger Movement Dataset, and Newark Airport Cargo Stats Dataset) to analyze and predict accurate trends. Data analyzing and prediction is done with the implementation of Time Series Analysis and respective ARIMA Models for respective modules. Development of modules is done using RStudio whereas Tableau is used for interactive visualization and end-user report generation. The Airport Trends Analytics Engine is an integral part of R and Tableau 10.4 and is optimized for use on desktop and server environments.

Keywords *- ARIMAmodel, Time series analysis, Airport Trends prediction, Air Cargo Movement, Air Passenger Traffic, Long-termprediction.*

1. Introduction

In the world of incessant evolution, one's desire to innovate new things step-up with increasing needs. As a result, need for forecasting and prediction grew on a much higher note. That is why prediction is still among the most interesting and rousing field of research with the desire of the researchers to improve the contemporary predictive models. The major reason is that our excessive dependency on future.

Year after year we see air traffic increasing at the exponential rate. Passengers, cargo movement and the subsequent flight delays lead to this hike in air traffic. Other major factors include weather problems, like fog, low visibility, etc., to cope up with these increasing problems, we need to have a proper predictive model which can analyze and predict using datasets. To accomplish this task we use time series concept along with ARIMA model which is also used for predicting future. Fig. 1 depicts the schematic representation of the architecture of the engine.

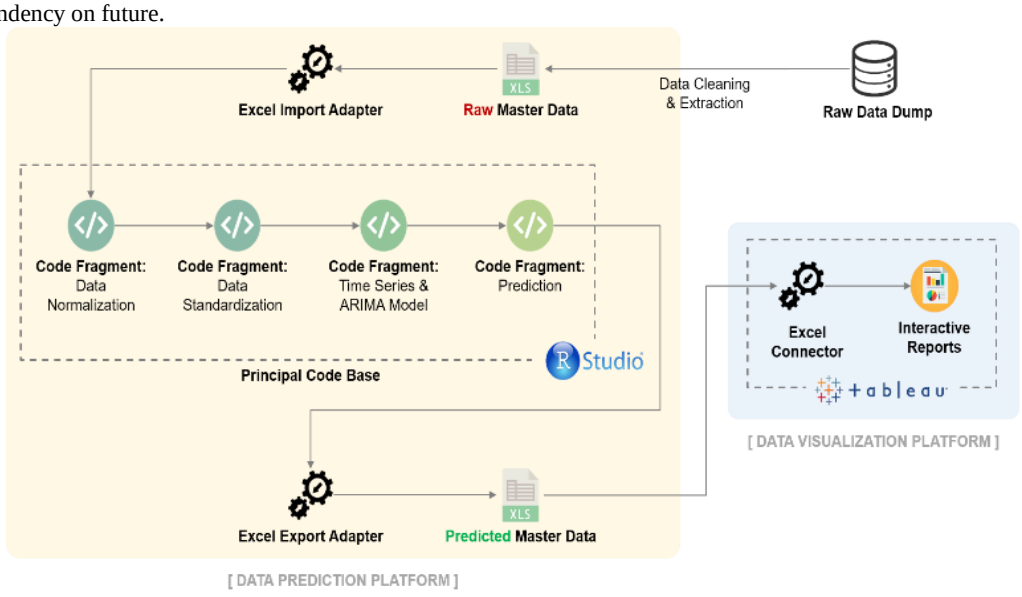

Fig. 1: Architecture of the Engine

2. ARIMA Model

ARIMA Model was introduced by Box and Jenkins in the year 1970. It mainly comprises of a set of activities for prediction and estimation along with time series analysis [1]. This model is considered to be one of the most efficient models which are capable enough to generate both short term and long term forecasts and predictions. The roots of this majorly arise from autoregressive model AR (p), the moving average model MA (q) and the duet of both the AR (p) and MA (q) [1,2]. The ARIMA (p,d,q) model is basically arranged in this following formation:

$$\Phi(B)\nabla^d x_t = \Theta(\mathrm{B})\varepsilon_t$$

$$\mathrm{E}(\varepsilon_t) = 0, \mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2, \mathrm{E}(\varepsilon_t \varepsilon_s) = 0, \mathrm{s} \neq \mathrm{t}$$

$$Ex_s\varepsilon_t = 0, \forall s < t \quad [2]$$

Here, p and q are the orders of the MA model and the AR model and d refers to the differentiation of the respective series. p,d,q are integral values .ε_t stands for estimated residual at each period of time. σ_ε^2 stands for variance of residuals [2].

ARIMA model being a very complicated process can be summarized as the following four steps:

1) Identify p,d,q in ARIMA model.
2) Estimation of the coefficients.
3) Fitting test to be performed on the estimated residuals.
4) Predicting the future from the past data [2].

There are majorly two functions used in the above steps: autocorrelation (ACF) and partial autocorrelation (PACF) functions. The basic need to use these two functions is the identification of the models as they preview various features of the functions [2].

The ARIMA model is efficient in providing the forecasts in both upper limit as well as lower limit and the predicted values. The confidence interval of 1-α is provided by the upper and lower limits [2, 4].

3. Time Series

Time series analysis is a very important model for prediction and forecasting which enhances the practicality of various domains. In the last two decades, many innovative models have been proposed to refine the time series analysis[3].

Among these models, the model which have been used most evidently and reliably is the ARIMA model. In this model, we assume that: time series being linear follow a known statistical distributive data[5].Though being a very powerful model to implement time series analysis, it follows some limitations: The fact that time series analysis follows linearity only makes impractical for different situations[3].

Therefore, many nonlinear models have already been developed by various scholars to overcome these impracticalities into consideration. These parameters may vary due to various interests: data recorded in traffic at equal intervals of time, passenger traffic in various airports periodically, data by various sensors taken down periodically [3,5].The rate of data recording and observation may vary from as low as one millisecond to as long as several years. One of the major problem which occurs is handling of the huge data which is generated over wide span due to which efficient data warehouse techniques must be used for the optimum functioning of the system[3].

Pre-data transformations: The data taken down from the real world are generally distorted and carry some noise which is inadequate in processing for further functions. So, various transformations have to be carried out of linearizing the dataset for the smooth functioning of the system. These transformations are called as pre-processing transformations. There are generally four of these transformations which are mostly used: offset translation, amplitude scaling, removing linear trends, removing noise [3,4].

Offset translation: In this transformation, the existing data set is translated into same time intervals. To do this, the simplest way is to subtract the means from each data values[3].

Amplitude scaling: This transformation is performed when two or more data exist at different amplitudes. The standard way to pursue this translation is by subtracting the mean values and then by dividing these differences by standard deviation[3].

Removing noise: The real world data contain various forms of noises which are undesirable. The most often used method to remove these noises is the moving averages method. The essence is to average each data values according to two or more of its neighbours[3].

4. Methodology

In this study, ARIMA model for forecasting various airport trends are developed and their process is explained in sections below. The tools used for implementation are RStudio and Tableau [6]. Data used in research is historical monthly passenger traffic data, cargo movement, temperature, humidity, dew point etc. So basically, this research focuses majorly on these aspects,

-Air Passengers Traffic Prediction

-Cargo Movement Prediction

Data of various airports around the world are analyzed and for each model, only one airport will be explained in detail below.

To develop most efficient and perfect ARIMA model, the following criteria are used:

- Bayesian or Schwarz Information Criterion (BIC) or should be relatively small.
- Standard error of regression (S.E. of regression) should be relatively small.
- Autocorrelation functions (ACFs) and partial autocorrelation functions (PACFs) of the residuals should have no substantial pattern left which is analysed with the help of Q-statistics[1].

A. ARIMA (p,d,q) for Air Passenger Traffic

In this study, Perth airport passenger movement datasetranges from January 2006 to December 2015 having a total of 120 observation. Following graph i.e. Fig. 2 shows the original pattern and provides the general overview of the passenger traffic with the progressing time.

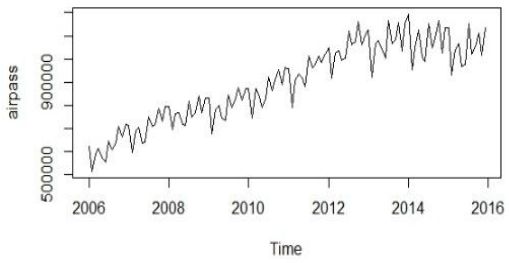

Fig. 2: Graphical Representation of Perth Airport Passenger Movement

Following graph represents autocorrelation and partial autocorrelation factors of the dataset:

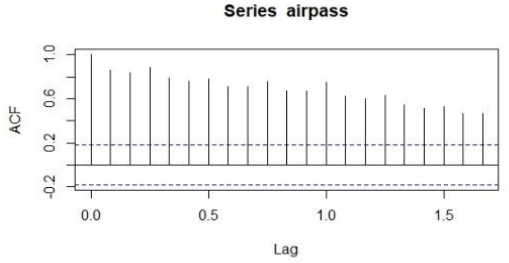

Fig. 3: ACF plot for the dataset

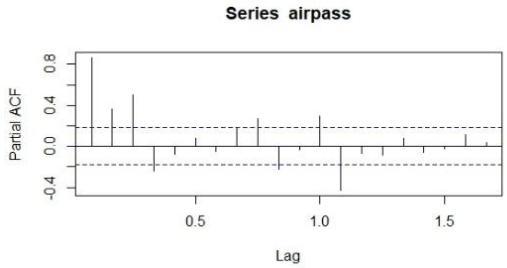

Fig. 4: Partial ACF plot for the dataset

By seeing Fig. 2, Fig. 3 and Fig. 4, we can determine that historical data is not stationary. To make it stationary, so that the pattern of growth is understandable, we proceed as per the following steps:

- Take the log
- Now, do the differentiation of values obtained after taking log

By doing this we get a stationary graph i.e. Fig. 5.

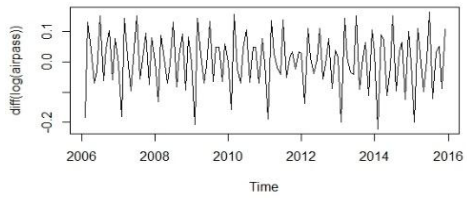

Fig. 5: Conversion to stationary plot by differentiating

After obtaining this, values of (p,d,q) are calculated using BIC/AIC (as shown in Fig. 6). The best fit ARIMA model developed for data of Perth airport is ARIMA (1,1,0).

```
Call:
arima(x = log(airpass), order = c(1, 1, 0), seasonal = list(order = c(0, 1,
    0), period = 12))

Coefficients:
          ar1
      -0.3489
s.e.   0.0900

sigma^2 estimated as 0.0009013:  log likelihood = 223.23,  aic = -442.46
```

Fig. 6: sigma^2, likelihood and aic values for ARIMA (1,1,0)

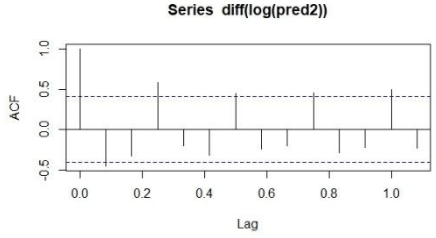

Fig. 7: ACF plot for stationary series

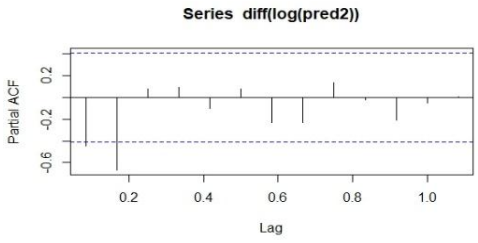

Fig. 8: Partial ACF plot for stationary series

By observing Fig. 7 and Fig. 8, it can be noticed that there are not substantial spikes. When a model is considered good, then the residuals ofa model are just the consequent random errors. As observed there are not many spikes in ACF and Partial ACF plot, which means that residuals are just the white noise. Therefore, there isn't any need of considering any AR(p) and MA(q).

Best model selected for forecasting formis expressed as:

$Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \in_t$

where, $\in_t = Y_t - \widehat{Y_t}$

(i.e., the difference of actual and the forecast value of series) [1]

B. ARIMA (p,d,q) for Air Cargo Movement

In this study, Newarkairport cargomovement datasetranges from January 2006 to December 2015 having a total of 120 observation. Following graph i.e. Fig. 9 shows the original pattern and provides general the overview of the cargo movement (in tonnage) with the progressing time.

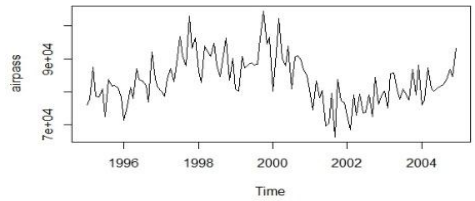

Fig. 9: Graphical Representation of Newark Airport Cargo Movement

Following graph represents autocorrelation and partial autocorrelation factors of the dataset:

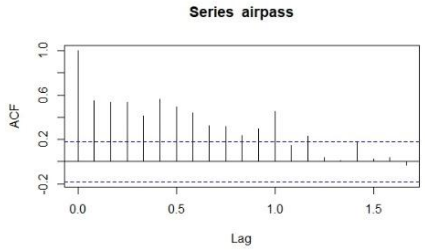

Fig. 10: ACF plot for the dataset

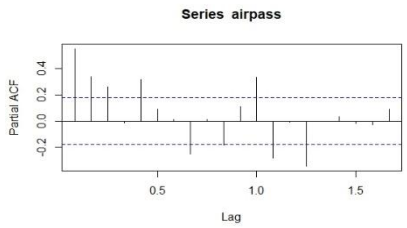

Fig. 11: Partial ACF plot for the dataset

By seeing Fig. 9, Fig. 10 and Fig. 11, we can determine that historical data is not stationary. To make it stationary, so that the pattern of growth is understandable, we proceed as per the following steps:

- Take the log
- Now, do the differentiation of values obtained after taking log

By doing this we get a stationary graph i.e. Fig. 12.

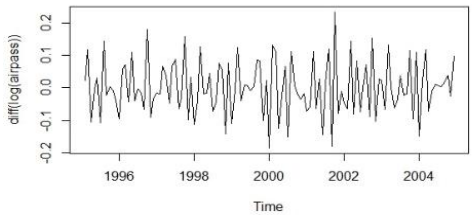

Fig. 12: Conversion to stationary plot by differentiating

After obtaining this, values of (p,d,q) are calculated using BIC/AIC (as shown in Fig. 13). The best fit ARIMA model developed for data of Newark Airport Cargo Movement is ARIMA (2,1,0) with the seasonality of order (2,1,0).

```
ari:
rima(x = log(airpass), order = c(2, 1, 0), seasonal = list(order = c(2, 1,
   1), period = 12))

oefficients:
         ar1      ar2    sar1     sar2     sma1
     -0.7291  -0.4994  0.1453  -0.0259  -0.9999
.e.   0.0875   0.0858  0.1066   0.1142   0.1826

igma^2 estimated as 0.002228:  log likelihood = 162,  aic = -312
```

Fig. 13: sigma^2, likelihood and aic values for ARIMA (1,1,0)

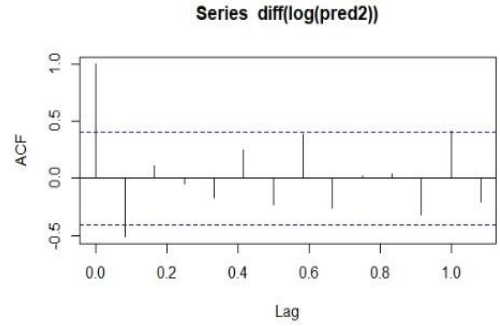

Fig. 14: ACF plot for stationary series

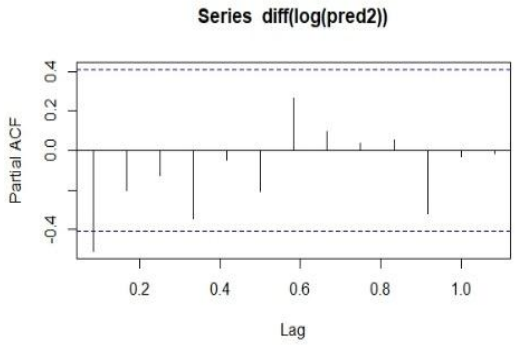

Fig. 15: Partial ACF plot for stationary series

By observing Fig. 14 and Fig. 15, it can be noticed that there are not substantial spikes. Whena model is considered good, then the residuals of a model are just the consequent random errors.Residuals are just the white noise as there are only a few spikes in ACF and PACF plot. Therefore, there isn't any need of considering any AR(p) and MA(q).
Best model selected for forecasting formis expressed as:
$Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \epsilon_t$
where, $\epsilon_t = Y_t - \hat{Y}_t$ [1]
(i.e., the difference of actual and the forecast value of series)

5. Results and discussion

A. Result of ARIMA Model for Air Passenger Traffic Prediction

Table 1 depicts the comparison between actual and predicted values of a sample period, calculated using ARIMA (1,1,0). Fig. 16 shows the graphical representation of both the values and contrasts the accuracy. It is evident that the predicted values are very close to the actual ones with minimal errors, which is quite impressive.

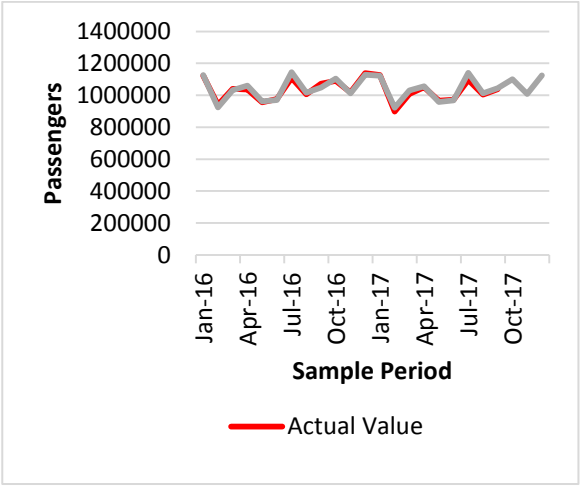

Fig. 16: Graphical representation of original and predicted values

Table 1: Comparison of original and predicted values

Sample Period	Actual Value	Predicted Value
Jan-16	1122583	1126414
Feb-16	939225	924659
Mar-16	1039280	1033028
Apr-16	1036602	1060162
May-16	956091	961842
Jun-16	976538	970826
Jul-16	1108324	1145286
Aug-16	1005953	1015078
Sep-16	1072661	1048616
Oct-16	1091243	1103709
Nov-16	1018529	1011986
Dec-16	1137632	1128092
Jan-17	1127293	1122305
Feb-17	897848	921285
Mar-17	1007158	1029260
Apr-17	1049476	1056295
May-17	964995	958333
Jun-17	972245	967285
Jul-17	1096856	1141108
Aug-17	1004850	1011375
Sep-17	1039122	1044791
Oct-17	NA	1099683
Nov-17	NA	1008294
Dec-17	NA	1123977

B. Result of ARIMA Model for Air Cargo Movement Prediction

Table 2 shows the monthly comparison between actual cargo movement and predicted cargo movement which is calculated using ARIMA (2,1,0) with the seasonality of order (2,1,0) after some modification in the autoregressive (p) and moving average (q) parameters. Fig. 17 gives the graphical representation of the two values to show the correlation accuracy. There is minor deviation at some stages as values might depend on various environmental factors.

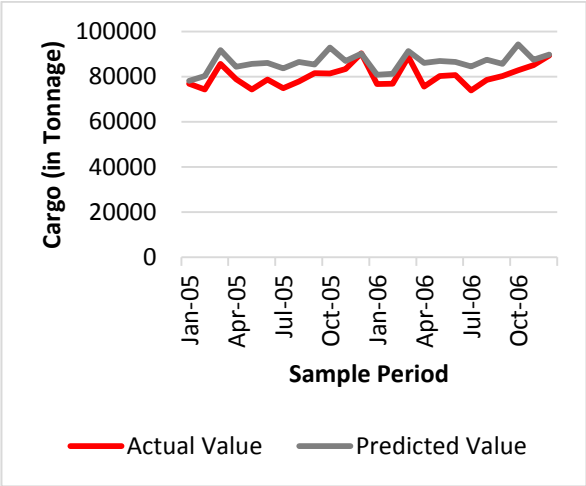

Fig. 17: Graphical representation of original and predicted values

Table 2: Comparison of original and predicted values

Sample Period	Actual Value	Predicted Value
Jan-05	76905	78161
Feb-05	74280	80243
Mar-05	85631	91748
Apr-05	78775	84389
May-05	74316	85628
Jun-05	78649	86005
Jul-05	74809	83646
Aug-05	77776	86516
Sep-05	81534	85321
Oct-05	81412	92914
Nov-05	83289	86918
Dec-05	90227	90084
Jan-06	76631	80851
Feb-06	76831	81208
Mar-06	88673	91355
Apr-06	75559	86098
May-06	80200	86879
Jun-06	80707	86533
Jul-06	73914	84508
Aug-06	78596	87447
Sep-06	80307	85581
Oct-06	82769	94301
Nov-06	85049	87433
Dec-06	89293	89779

6. Conclusion

Time series analysis is used in this paper along with ARIMA model to analyze and predict various airport trends. The predicted values obtained by using best ARIMA model are very close to that of original values with minimal error. Results obtained are accurate on long-term forecasting and can help in distant future. This could help and guide airport authorities in improving there infrastructure and facilities at given period of time.

References

[1] Ayodele A. Adebiyi., Aderemi O. Adewumi and Charles K. Ayo, "Stock Price Prediction Using the ARIMA Model", UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, pp. 105-111, 2014.

[2] Peng Chen, Hongyong Yuan and Xueming Shu, "Forecasting Crime Using the ARIMA Model", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 627-630, 2008

[3] Mirjana Ivanović and Vladimir Kurbalija, "Time series analysis and possible applications", 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 473-479, 2016

[4] Xing-qiang Zhang, Xue Yang and Shi-qing Dong, "Study on composite forecasting model of air passenger capacity based on air partition", pp. V9-66–V9-69, 2010

[5] M O D Rizwan, R. Jeberson Retna Raj and M Vasudev, "A Novel Approach For Time Series Data Forecasting Based On Arima Model For Marine Fishes", International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), 2017

[6] Richard Wesley, Matthew Eldridge and Pawel T. Terlecki, "An Analytic Data Engine for Visualization in Tableau", SIGMOD '11 Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pp. 1185-1194, 2011