Time Series Analysis For Long Memory Process Of Air Traffic Using Arfima

Manohar Dingari, D. Mallikarjuna Reddy, V. Sumalatha

Abstract: In the present study, the time series models ARIMA and ARFIMA or FARIMA models have been fitted to Air India domestic air passengers, which considered as self similarity and Long Range Dependence (LRD). In such case ARFIMA model is expected to be superior to ARIMA. We fitted ARIMA and ARFIMA models to air traffic data and compared. Then the best model has been identified using RMSE, MAE and MAPE values. This model can be useful to analyze the air traffic flow and revise the services of Air India. The analysis was carried out using time series data on number of passengers travelling by Air India domestic flights during January 2012 to December 2018.

Index Terms: Air Traffic, ARFIMA, ARIMA, Long Range dependence, Quality of Service, Self-Similarity and Time Series Models.

1. INTRODUCTION

In recent years Air transport is rapidly growing service sector in India. It is important to know the nature of air traffic flow, to analyze and forecast the air traffic which intern helps the air service providers to design and develop their Quality of Services (QoS). Time series models have been proven as best for analyzing and forecasting the time series. In this study we fitted ARIMA model to air traffic data. In our previous study, the nature of the Air India air traffic data has been shown as self-similar and the data was exhibited Long Range Dependence (LRD). From the fore studies when the data has Long Range Dependence ARFIMA models are expected to be superior to ARIMA models, so we fitted ARFIMA models to the Air India air traffic data and compared with ARIMA models. The subject of the analysis is Air India with number of passengers travelled by Domestic flights being the focus. The research period covers the years from 2012 to 2018. Data were presented in monthly cycles. Two research methods were compared for analyzing the data: ARIMA model and ARFIMA model.The main aim of this paper is, to prove when the data has self-similar behavior or Long range Dependence (LRD), the ARFIMA model gives the best result than ARIMA models. The remaining paper has been organized as follows: Time series models are discussed in II. Concept of self similarity and Long Range Dependence (LRD) are discussed in III. Analysis of Air India air traffic data are presented in IV and conclusions are placed in V.

2 TIME SERIES MODELS

2.1 ARIMA Model

The ARIMA model introduced by Box and Jenkins (1976) includes Autoregressive as well as Moving Average parameters and explicitly includes differencing in the formulation of the model. Specifically, the three types of parameters in the model are: the autoregressive parameters (p), the number of differencing passes (d), and moving average parameter (q). In the notation introduced by Box and Jenkins, models are summarized as ARIMA (p, d, q).The foremost step in the process of modeling is to check for the stationary of the time series data. This is done by observing the graph of data or Autocorrelation and the Partial Autocorrelation functions. Another way of checking stationarity is to fit the first order AR model to the raw data and test whether the coefficients Ø is less than one.The next step is to identify an appropriate sub-class of the general ARIMA model.

$$\emptyset(B)\nabla^d \ z_t \ = \theta(B)a_t \qquad (2.1.1)$$

Which may be used to represent a given time series. The general procedure is

i) To difference z_t as many times as is needed to produce stationary.

$$\emptyset(B)\,\omega_t \ = \theta(B)\,a_t \qquad (2.1.2)$$

Where

$$\omega_t \ = (1-B)^d \ z_t = \nabla^d z_t \qquad (2.1.3)$$

ii) To identify the resulting ARIMA process.

The Autocorrelation and Partial Autocorrelation functions will be used as main tools in attaining (i) and (ii). Stationarity in time series means a constant mean, variance and Autocorrelation through time. Generally the hypothetical series will not be stationary. Therefore the series needs to differenced until it is stationary (usually at most two differences will be sufficient to make the series stationary).

$$\omega_t = \nabla^d \ z_t \qquad (2.1.4)$$

After differencing the series to identify the sub class of ARIMA model, which is suitable for the series, the components of Autoregressive (p) and Moving Average (q) should be identified. The number of parameters of Autoregressive (p) can be identified by using correlogram of Partial Autocorrelations and the number of parameters of Moving Average (q) can be identified by using correlogram of Autocorrelations. Then p parameters of Autoregressive $Ø_1$, $Ø_2$, - - - , $Ø_p$ and then q parameters of Moving Average θ_1, θ_2, - - - θ_q have to be estimated by using Least Squares (LS) or Maximum Likelihood (ML) methods.

2.2 ARFIMA Model

Autoregressive Integrated Moving Average (ARIMA) model analyzes the short memory process. Long memory processes are analyzed by Autoregressive Fractionally Integrated Moving Average (ARFIMA) model. Long memory processes are

- *Manohar Dingari, Research Scholor, Department of mathematics, GITAM University, Hyderabad,502329 India. E-mail: manohar.dingari@gmail.com*
- *Dr.D. Mallikarjuna Reddy, Asst Professor, Department of mathematics, GITAM University, Hyderabad,502329 India. E-mail: mallik.reddyd@gmail.com*
- *V.Sumalatha, Research Scholor, Department of Statistics, OSMANIA University, Hyderabad, India, E-mail: sumanu05@gmail.com*.)

stationary. ACFs of long memory process decay more slowly than short memory process. This long Range Dependence (LRD) can be best analyzed by ARFIMA (or FARIMA) models which also satisfies the principle of parsimony. The ARFIMA models are the generalization of ARIMA models. Where 'd' is allowed to take non integer values, $-0.5 < d < 0.5$. This fractional integration parameter 'd' is used to capture the long run effects. While the ARMA parameters capture the short run effect. According to Hosking (1981), Baillie (1996) and others the ARFIMA process with $d < 0$ indicates intermediate memory process and $d > 0$ indicates long memory process. According to Box-Jenkins, Reinsel (2008, 429) the ARFIMA process with $d < 0$ indicates long memory process as they decay slowly in a hyperbolic rate. Thus we follow in this study the ARFIMA process with $-0.5 < d < 0.5$ indicates long memory process.

The ARFIMA model is

$$\emptyset(B)(1-B)^d Z_t = \theta(B)a_t \qquad (2.2.1)$$

Where

$\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$ is AR operator, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is MA operator and d is the fractional integrating parameter, $-0.5 < d < 0.5$.

$$(1-B)^d = \sum_{i=0}^{\infty} (-1)^i \frac{\Gamma(d+1)}{\Gamma(i+1)\Gamma(d-i+1)} B^i \qquad (2.2.2)$$

Γ is the gamma function.

ARFIMA models are extensively used in diverse fields such as hydrology and economics. Long Range Dependence (LRD) was first introduced by Hurst (1951) in the field of hydrology. Hosking (1981) represented Long Range Dependence in hydrology by ARFIMA. Granger and Joyeux (1980) used ARFIMA to represent Long Range Dependence in economics. LRD can be conveniently measured by Hurst parameter 'H'.

3 SELF- SIMILARITY AND LONG RANGE DEPENDENCE

The definition of exact second-order self-similar process is given as follows. Arrival instants are modeled as point process. Split the time axis into disjoint intervals of unit length and let $X = \{X_t : t = 1, 2, \ldots.\}$ be the number of points (arrival) in the t^{th} interval. Let X be a second order stationary process with variance σ^2 (or SD σ) and the ACF $\gamma(k), k \geq 0$ is given as:

$$\gamma(k) = \frac{Cov\ (X_t, X_{t+k})}{Var\ (X_t)} \qquad (3.1)$$

For every $m = 1,2,3,....,$ let a new time series $X_t^{(m)}$ is obtained aggregating the original time series X over non-overlapping blocks of size m .

That is

$$X_t^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(t-1)m+i}, \ t = 1, 2, \ldots \qquad (3.2)$$

This new series $X_t^{(m)}$, for each m , is also a second order stationary process with autocorrelation function $\gamma^{(m)}(k)$.

1. The process 'X' is said to be exactly second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$\gamma(k) = \frac{\sigma^2}{2}\left[(k+1)^{2H} - 2k^H + (k+1)^{2H}\right], \forall k \geq 1 \qquad (3.3)$$

2. The process 'X' is said to be asymptotically second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$\sum_{m \to \infty} \gamma^{(m)}(k) = \frac{\sigma^2}{2}\left[(k+1)^{2H} - 2k^H + (k+1)^{2H}\right], \forall k \geq 1 \qquad (3.4)$$

3. In variance terms, self- similar process is defined as follows: The process 'X' is said to be exactly second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$Var\ (X^{(m)}) = \sigma^2 m^{-\beta}, \forall m \geq 1 \qquad (3.5)$$

Now we shall differentiate long range dependence (LRD) and short range dependence (SRD) processes. For $H \neq 0.5$, from the Eq. (3), we can see that

$$\gamma(k) = H(2H-1)k^{2H-2} \text{ as } k \to \infty \qquad (3.6)$$

and we have

$$\sum_k \gamma(k) \sim c\sum_k k^{-\beta}, \quad c = H(2H-1). \qquad (3.7)$$

The series $c\sum_k k^{-\beta}$ is divergent if $0.5 < H < 1$ or $0 < \beta < 1$ otherwise they are convergent, being a positive term series. Accordingly the left hand series $\sum_k \gamma(k)$ is divergent if $0.5 < H < 1$ or $0 < \beta < 1$, otherwise they are convergent. That is, for $0.5 < H < 1$, the autocorrelation functions decays slowly, that is hyperbolically. In this case, the process X is called LRD. The process X is SRD if $0 < H < 0.5$ and the autocorrelation function is summable.

3.1 Hurst Index-Self-Similar Process

The intensity of self-similarity is given by Hurst index H . The index H was named after the hydrologist H.E. Hurst who spent many years to investigate the problem of water storage and also to determine the level patterns of the Nile River. Hurst index is perfectly well defined mathematically, measuring if it is a problematic one. The data must be measured at high lags or low frequencies where fewer readings are available. The index H has range $0.5 \leq H \leq 1$. Computing index

H is a task. Several approaches are available to estimate degree of self-similarity in a time-series and forecasting analysis. We calculated the Hurst index by Correlogram method, Variance-time analysis and percentiles method.

i) Correlogram Method

In time series analysis, plot of ACF (autocorrelation function) is known as correlogram where the estimated correlation can be given in terms of auto-covariance function $\gamma(k)$

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} \quad (3.1.1)$$

It has already been observed that slow decay of correlation, which is proportional to K^{2H-2} for $\frac{1}{2} < H < 1$ indicates the long-memory process [11]. Therefore, the plot of the sample autocorrelation should exhibit this property. A much better plot for the handling of long-range dependence is the plot of ACF in logarithmic scale. If the asymptotic decay of the correlation is hyperbolic, then the points in the plot should be approximately scattered around a straight line with a negative slope of $2H-2$ for the long memory processes but for short memory, the points should tend to diverse to minus infinity at an exponential rate. If the time series is long enough or if the series has strong long-range dependence, then this log-log correlogram is useful. Correlogram is useful as a preliminary heuristic approach to the data. Some pitfalls of sample correlation which are less known can be found in Mandelbrot [12, 17]. Even though it is neither widely used nor attractive method for estimation, still H , the self-similarity parameter, can be estimated by this method deriving an equation of the form

$$\rho(k) = \hat{H}(2\hat{H}-1)K^{2\hat{H}-2} \quad (3.1.2)$$

Using this method, the obtained value of H in this case is 0.831.

ii) Variance-Time Analysis

This method, variance time analysis [18] is very popular and is based on property of slowly decaying variance of self-similar processes undergoing aggregation. The m -averaged process $X^{(m)} = (X_1^{(m)}, X_2^{(m)}, \ldots)$ of a discrete-time stationary parent process $X_1, X_2, \ldots$ as:

$$X_j^{(m)} = \frac{1}{m}\sum_{(j-1)m+1}^{m} X_i, \qquad j = 1, 2, \ldots, \frac{N}{m} \quad (3.1.3)$$

Where m and j are positive integers.

The variance is defined as:

$$Var[X^{(m)}] = \frac{1}{N/m}\sum_{j=1}^{m}(X_j - \bar{X})^2 \quad (3.1.4)$$

The variances of the aggregated processes $X^{(m)}(m = 1, 2, 3, \ldots.)$ decrease linearly (for large m):

$$Var[X^{(m)}] = Var[X]m^{\beta} \quad (3.1.5)$$

The variance-time plot is obtained by plotting $\log Var[X^{(m)}]$ against $\log m$ and by fitting a sample least squares line through the resulting points in the plane, ignoring the small values for m . The estimated slope by sample least squares is $-\beta$, values of the estimate of the asymptotic slope between –1 and 0 suggest self-similarity and an estimate for the degree of self-similarity is given by

$$H = 1 - \frac{\beta}{2} \quad (3.1.6)$$

Using this method, the obtained value of H in this case is 0.792.

iii) Percentile Method

A percentile is the value of a variable below which a certain percent of observations fall, like partition values of a process such as quartiles and deciles. There is no exact definition for the percentile [10], however all definitions yield similar results when the number of observations is very large. One definition of percentile, often given in texts, is that the P^{th} percentile $(1 \le P \le 100)$ of N ordered values is obtained by first calculating the rank.

$$n = \frac{P * N}{100} + \frac{1}{2} \quad (3.1.7)$$

Given data set or time series $(t, Z_t)\ (t \ge 0)$. First we can find the percentiles $(P_i,\ i = 1, 2, \ldots 100)$ for a given time series or real time data using

$$P_i = \frac{i * N}{100} + \frac{1}{2};\ i = 1, 2, \ldots, 100. \quad (3.1.8)$$

P_i = i^{th} percentile, this a special type of average such as partition values in descriptive statistics like quartiles (Q_1, Q_2, Q_3) . Draw a scattered Plot percentile number against percentiles on log scales. A linear equation $Z_t = \beta t + c$ (say) is obtained with the slope (β) . The Hurst parameter (H) is then computed by Eq. (3.1.6). Using this method, the H value is computed for the data. The pertaining scattered data and trend line with the slope $\beta = 0.406$. The obtained value of H in this case is 0.797. One paper [11], explained how the 95-percentile depends on the aggregation window size, and how this phenomenon justifies the mathematical definition of self similarity or LRD. The advantages of this method are: This method is matter of a simple empirical formula, unlike other two methods. Data however large it may be is divided into hundred parts (partition values) and the plotting involves only 100 points (percentile versus percentile number).

4 ANALYSIS OF THE DATA

Both ARIMA and ARFIMA (FARIMA) models have been estimated by using the same data that refer to number of passengers travelling monthly by Air India on scheduled Domestic services for the years 2012 to 2018. The data has been taken from the Directorate General of Civil Aviation (DGCA) website. SPSS and STATA softwares were used for this analysis.

4.1 ARIMA MODELLING

To check the stationarity sequence chart has applied to the data. The graph appeared was non-stationary.

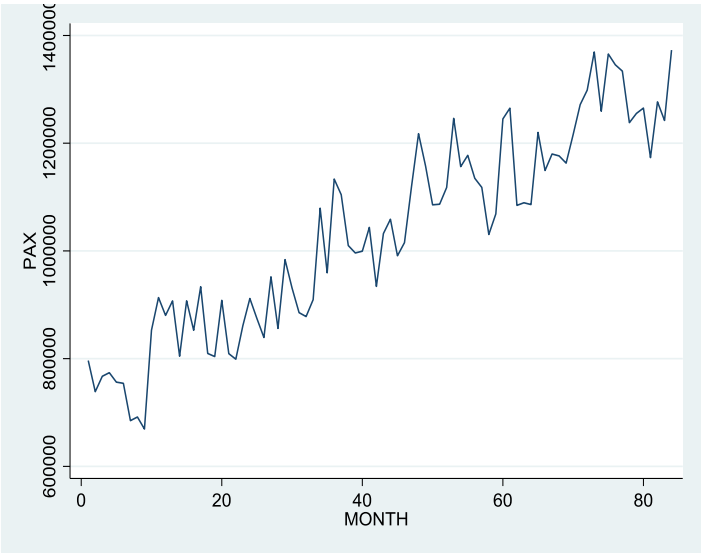

Fig1. *Plot of Air Traffic Data*

To achieve stationarity in mean the series has been differenced once, to remove non-stationarity in variance the time series has been transformed by using the Natural log transformation.

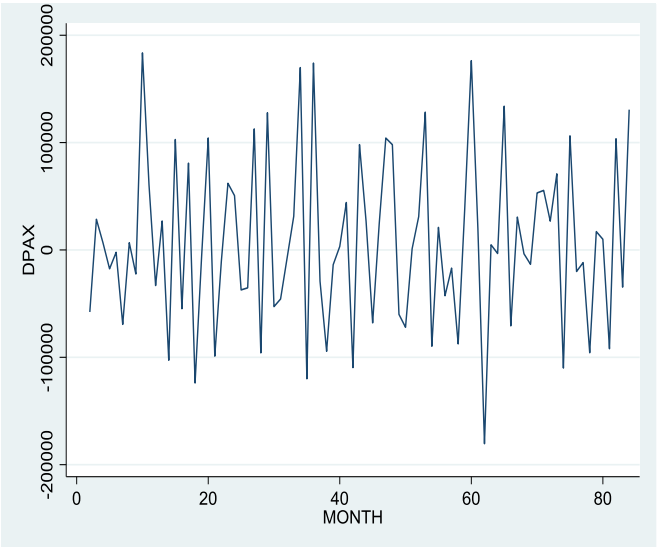

Fig2. *Stationary Air Traffic Data*

To identify p and q values Correlograms of Autocorrelation and partial Autocorrelations were constructed.

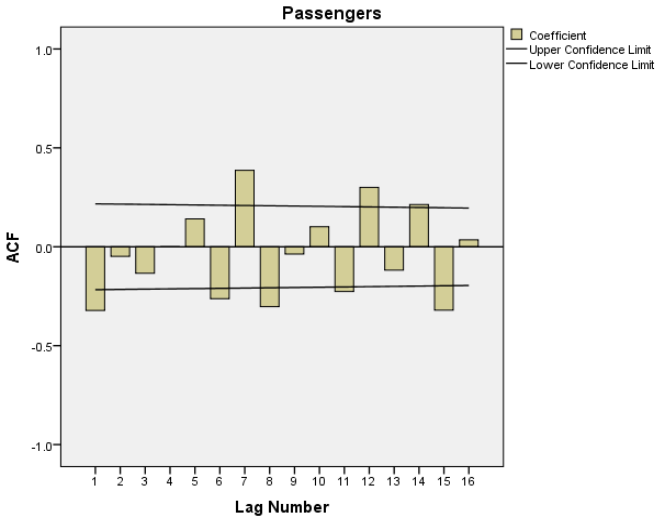

Fig3. *ACF of Air Traffic Data*

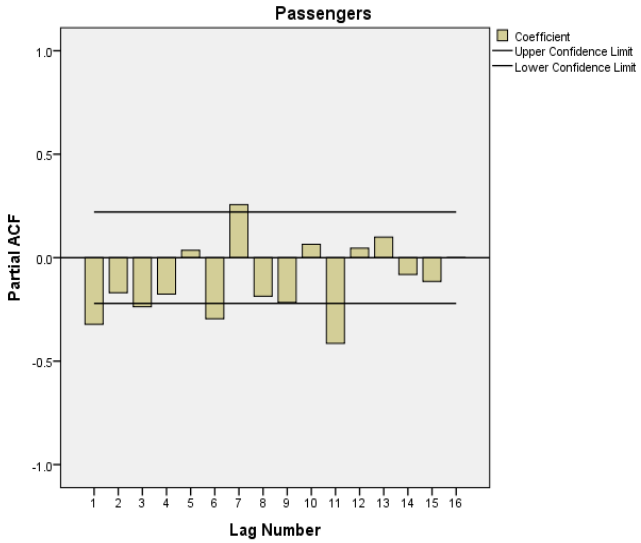

Fig4. PACF of Air Traffic Data

The estimate of the partial Autocorrelation coefficients shows that only Π_1 does not fall within the two standard error bounds

$\pm 2/\sqrt{N}$, therefore the order 1 has been chosen for the AR Component similarly MA component has been chosen from the correlogram of the ACF as 1. Then the model is identified as ARIMA (1, 1, 1). And the coefficients of AR and MA estimated as in the table 1.

TABLE1. ARIMA MODEL PARAMETERS

		Parameter	Estimate
ARIMA(1,1,1)	AR1	Ø	0.355
	MA1	Θ	1.000
	Constant	C	-0.027

The AR coefficient Ø was estimated to be 0.355; MA coefficient θ was estimated to be 1.000.

4.2 ARFIMA Modeling

The processes with Long Range Dependence (LRD) are stationary processes whose autocorrelation functions decay more slowly than Short Range Dependent (SRD) processes. The ARFIMA (FARIMA) model gives a parsimonious model for a long memory process. They ARFIMA model is a generalization of ARIMA model which allows for a fractional differences $-0.5 < d < 0.5$ to model long run effects. In 3.1 the Air India air traffic data has been shown to have long range dependence. Thus ARFIMA model has been fitted to the data. The STATA software has been used for this analysis. ARFIMA parameters have been estimated using Modified Profile Likelihood (MPL) method. The fractional difference parameter estimated by MPL has less small sample bias than the Maximum Likelihood Estimator (MLE).

TABLE 2. ARFIMA MODEL PARAMETERS

		Parameter	Estimate
ARFIMA(p,d,q)	AR1	Ø	0.657
	MA1	Θ	-0.999
	Fractional difference	d	-0.347

The AR coefficient Ø was estimated to be 0.657; MA coefficient θ was estimated to be -0.999 and the fractional difference parameter d was estimated to be -0.347.

4.3 Model Diagnostic Check

It is concerned with testing the goodness of fit of the model

TABLE3.
ARIMA AND ARFIMA PARAMETERS

Model	p	d	q
ARIMA	1	1	1
ARFIMA	1	-0.347	1

The adequacy of these two models checked by the R-square, Root Mean square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

TABLE3.
MODEL FIT STATISTICS

	ARIMA (1,1,1)	ARFIMA (1,-0.347,1)
R-squared	0.876	0.948
RMSE	65859.068	43588.99
MAE	51711.14	36854.83
MAPE	5.116	3.248

In practice, the low value of RMSE, MAE and MAPE indicate a good fit for the model and the high value of R-squared indicate a perfect prediction. Also a MAPE value lower than 10% suggest a forecast likely to be very good, lower than 20% likely to be good and above 30% likely to be in accurate. For this data, both the models ARFIMA and ARIMA have MAPE values less than 10%, which shows both are potentially very good for forecasting but MAPE of ARFIMA model is 3.248 and of ARIMA is 5.116. RMSE of ARFIMA model is 43588.99 and of ARIMA is 65859.068. MAE of ARFIMA model is 36854.83 and of ARIMA is 51711.14. R-squared value of ARFIMA model is 0.948 and of ARIMA is 0.876. Thus RMSE, MAE and MAPE are low for ARFIMA model than ARIMA. Also R-squared value is greater for ARFIMA model than ARIMA, which shows ARFIMA model is superior to ARIMA for this data. The plot of Actual and Forecasted values by using ARFIMA (1,-0.347,1) is presented in figure.5.

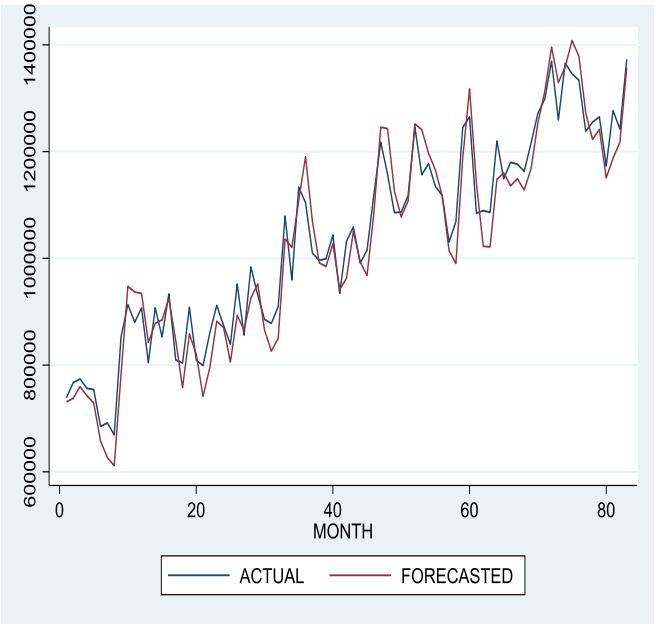

Fig5 *Plot of Actual and Forecasted Air Traffic*

5 CONCLUSION

In this study ARIMA and ARFIMA models have been fitted to the air traffic data regarding number of passengers travelled monthly by Air India scheduled domestic flights during January 2012 to December 2018. Since the air traffic data has exhibited Long Range Dependence (LRD), ARFIMA model was expected to be better than ARIMA. The two models were compared using model fit statistics RMSE, MAE and MAPE values. From those values ARFIMA model has been identified as best model for this data. Hence this model can be used to forecast the future air traffic of Air India domestic carriers. This study is helpful for Air India to revise their services.

REFERENCES

[1] Chen pu, Li ni, Xu jie, Zhao Ting, Liu Chen, " An Experiment on the Hurst Exponent based on FARIMA" Advances in Engineering, Volume 126.

[2] Christos katris, Sophia Daskalaki, " Prediction of Internet traffic using Time Series and neural Networks" adfa, p. 1, 2011. © Springer-Verlag Berlin Heidelberg 2011.

[3] Fei Xue, Jiakun Liu, Yantai Shu, Lianfang Zhang, "Traffic Modeling Based on FARIMA Models"

Engineering Solutions for the Next Millennium. 1999 IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.99TH8411).

[4] Omekara C. O.*, Okereke O. E., Ukaegeu L. U. " Forecasting Liquidity Ratio of Commercial Banks in Nigeria" Microeconomics and Macroeconomics 2016, 4(1): 28-36, DOI: 10.5923/j.m2economics.20160401.03

[5] Alberto Montanari and Renzo Rosso Murad S. Taqqu, "Fractionally dilferenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation" WATER RESOURCES RESEARCH, VOL. 33, NO. 5, PAGES 1035-1044, MAY 1997

[6] Alberto Andreoni, Maria Nadia Postorino*, "Time Series Models To Forecast Air Transport Demand:A Study About A Regional Airport", 11th IFAC Symposium on Control in Transportation Systems Delft, The Netherlands, August 29-30-31, 2006.

[7] Etebong P. Clement, "Using Normalized Bayesian Information Criterion (Bic) to Improve Box - Jenkins Model Building" American Journal of Mathematics and Statistics 2014, 4(5): 214-221 DOI: 10.5923/j.ajms.20140405.02.

[8] Dr.N.Chitra , Ra.Shanmathi , Dr.R.Rajesh, "Application Of Arima Model Using Spss Software - A Case Study In Supply Chain Management" International Journal of Science, Technology & Management www.ijstm.com Volume No 04, Special Issue No. 01, April 2015.

[9] Gooijer, J. G., Hyndman, R. J., 25 years of time series forecasting. International Journal of Forecasting, Vol. 22, No. 3, 2006, pp. 443-473.

[10] Lane,David."Percentiles".http://cnx.org/content/m10805/latest. Retrieved 2007-09-15.

[11] Web hosting talk Forum: 95th Percentile billing polling interval, http://www.webhostingtalk.com, Last accessed 09/23/2008.

[12] Beran J., Statistics for Long-Memory Processes, Chapman and Hall, 1994.

[13] Mallikarjuna Reddy Doodipala , Malla Reddy Perati K. Raghavendra; H. K. Reddy Koppula; Rajaiah Dasari 'Self-Similar Behavior of Highway Road Traffic and Performance Analysis at Toll Plazas" 1234 Journal Of Transportation Engineering © Asce October 2012.

[14] Pushpalatha Sarla, D.Mallikarjuna Reddy, Manohar Dingari, "A Study on Self Similarity Analysis of Web Users Data at Selected Web Centers" Proceedings of International conference on Mathematics ICM- 2015.

[15] Manohar Dingari D.Mallikarjuna Reddy V. Sumalatha "Self -Similar Performance of Passengers Arrival Pattern of Air India-Domestic Services " © 2019 IJRAR June 2019, Volume 6, Issue 2.