TIME SERIES

&

FORECASTING

(Project)

Forecasting number of International Passengers Flying to USA via foreign carriers over a month using the methods of time series modeling.

ABSTRACT

To gain the practical insights of forecasting concept & test the reliability of ARMA modeling technique as a forecasting tool. We will use a data set based on number of passengers traveling to US through foreign carriers over a month for certain years. We will predict the number of passengers flying to US in future over a month based on the previous numbers from the dataset. We will first extract the deterministic part from the raw data identifying the periodicity and nature of trend present in dataset. Then we will fit the ARMA model over the residuals of deterministic part.

The primary goal is to reduce the residuals as much as possible. We will be using F test to obtain the best model over a range of different patterns of trends & periodicity. We will perform necessary conditions testing on auto regressive & moving average in order to get the optimum fit for the stochastic part of the data. We will perform the joint optimization of deterministic & stochastic part. Then we will do the forecasting on the 10% of the data available & find the errors between predicted and actual values. The results will give the accuracy of the model. It is not necessary that the prediction should be accurate if the model fitting is perfect. The accuracy depends upon the availability of the maximum possible significant parameters. The problem of over fit data points can also disturb the accuracy of the forecast. We should keep this in mind throughout the whole process of modeling.

1. INTRODUCTION

The ability to model and perform decision modeling and analysis is an essential feature of many real-world applications ranging from emergency medical treatment in intensive care units to military command and control systems. Existing formalisms and methods of inference have not been effective in real-time applications where tradeoffs between decision quality and computational tractability are essential. In practice, an effective approach to time-critical dynamic decision modeling should provide explicit support for the modeling of temporal processes and for dealing with time-critical situations.

With Tourism becoming one of the essential contributors to country's economy. It is important to concentrate & facilitate Tourists & Transportation. For most of the international trips, we invest the major money for transportation across the different destination. With the stiff competition across Aviation industry, no company can afford to lose their passengers. In fact, to stay ahead in the competition, one needs to be updated about all the relevant data & necessary details. The future data predictions for number of passengers, their favorite destinations, period preference, etc. can definitely help a company to set their best bets & improve their business.

1.1 Structure of Report

2. Data Selection
3. Analysis of raw data: Plotting & evaluation
   Splitting the data in Training & testing data
4. Modeling Procedure
   - 4.1. Fitting the Deterministic Trend of the data
   - 4.2. Specifying and Modeling the Stochastic Part
   - 4.3. Jointly Optimizing the Integrated Model
5. Model Diagnostics on the Integrated Model
   - 5.1. Checking stability of jointly estimated ARMA(2,2) model
   - 5.2. Comparison of the Jointly Optimized Model and Actual Data
   - 5.3. Analysis of Residuals Obtained from the Integrated Model
6. Forecasting
7. Conclusion
8. References

For analysis, modeling & forecasting procedure we have used "RStudio" platform.

2. DATA SELECTION

- To gain the practical insights and test the reliability of ARMA modeling technique as a forecasting tool. We will be submitting the project based on the Forecasting over a certain data set.
- We have selected a raw data of the international Airline flyers landing in USA month over month from the US government portal of transportation data & statistics.
  Link: - **http://www.transtats.bts.gov/Data_Elements.aspx?Data=1**
- These forecasted numbers can help Airline companies to prepare a schedule, arrange the number of flights and related things. Forecasting also can help companies to save upon their money by increasing or decreasing number of flights, Number of People working on flights, etc.

We will perform a step by step extraction of patterns for the best forecasting, by obtaining a proper fit for the given data based on trends, seasonality and other factors integrated in the raw data.

3. ANALYSIS OF RAW DATA

3.1 Plotting and Evaluation of raw data set

The figure below shows the plot of raw data points (148 data points in total). The data is recent, and it is spread over years starting from 2004 to 2015. With initial analysis we can see that there is increasing trends with certain seasonality between the data points year over year.

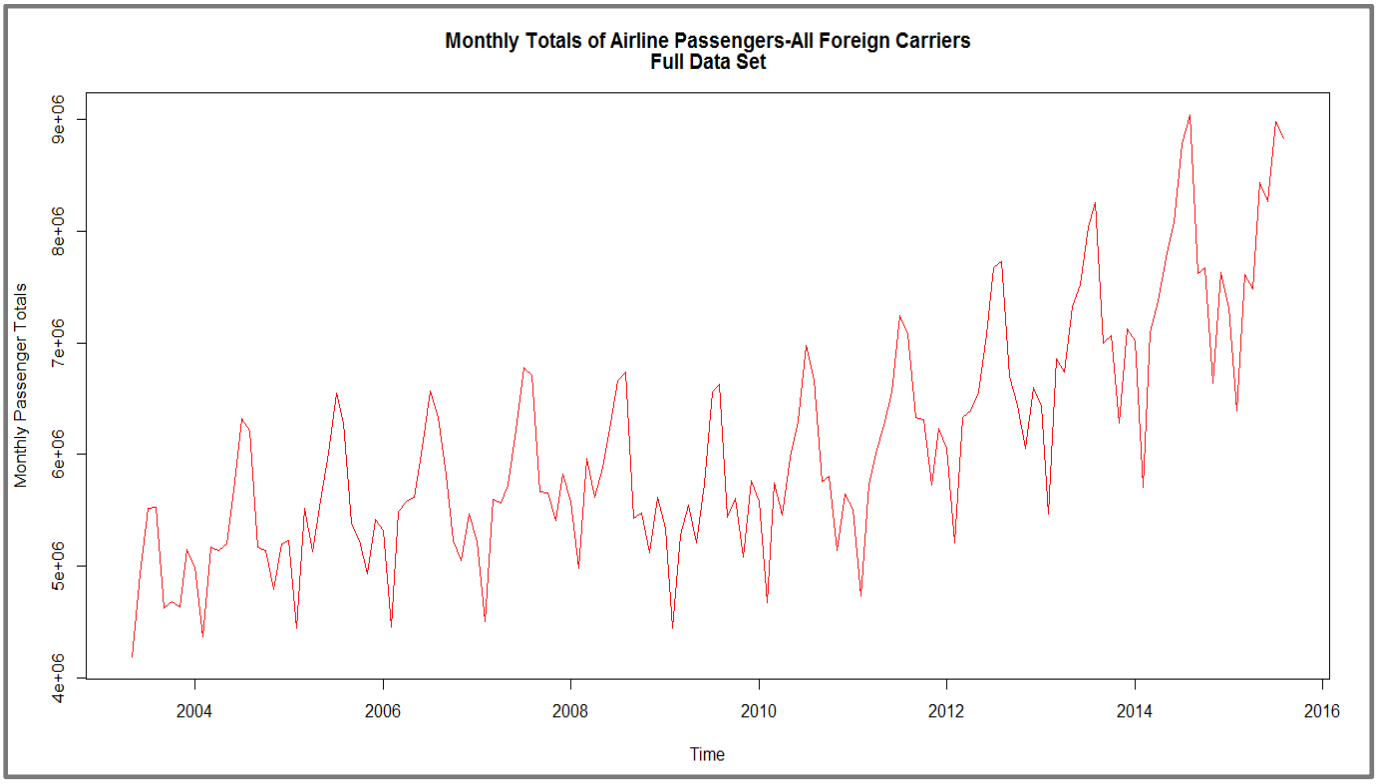

Fig 3.1.1: Full data set plot

It is important to evaluate forecast accuracy using genuine forecasts. That is, it is invalid to look at how well a model fits the historical data; the accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when estimating the model. When choosing models, it is common to use a portion of the available data for testing and use the rest of the data for estimating (or "training") the model. Then the testing data can be used to measure how well the model is likely to forecast on new data.

Thus, we have split our raw data into two parts:

- **Training data**: We will use this data for modeling procedure. Typically, 90:10 split ratio is considered to be the optimum for modeling & forecasting.
- **Testing data**: This part of the data will be used for forecasting based on the modeling we have done using training data.

The figure below shows the 90% of the raw data (133 data points) that we have used for modeling purpose.

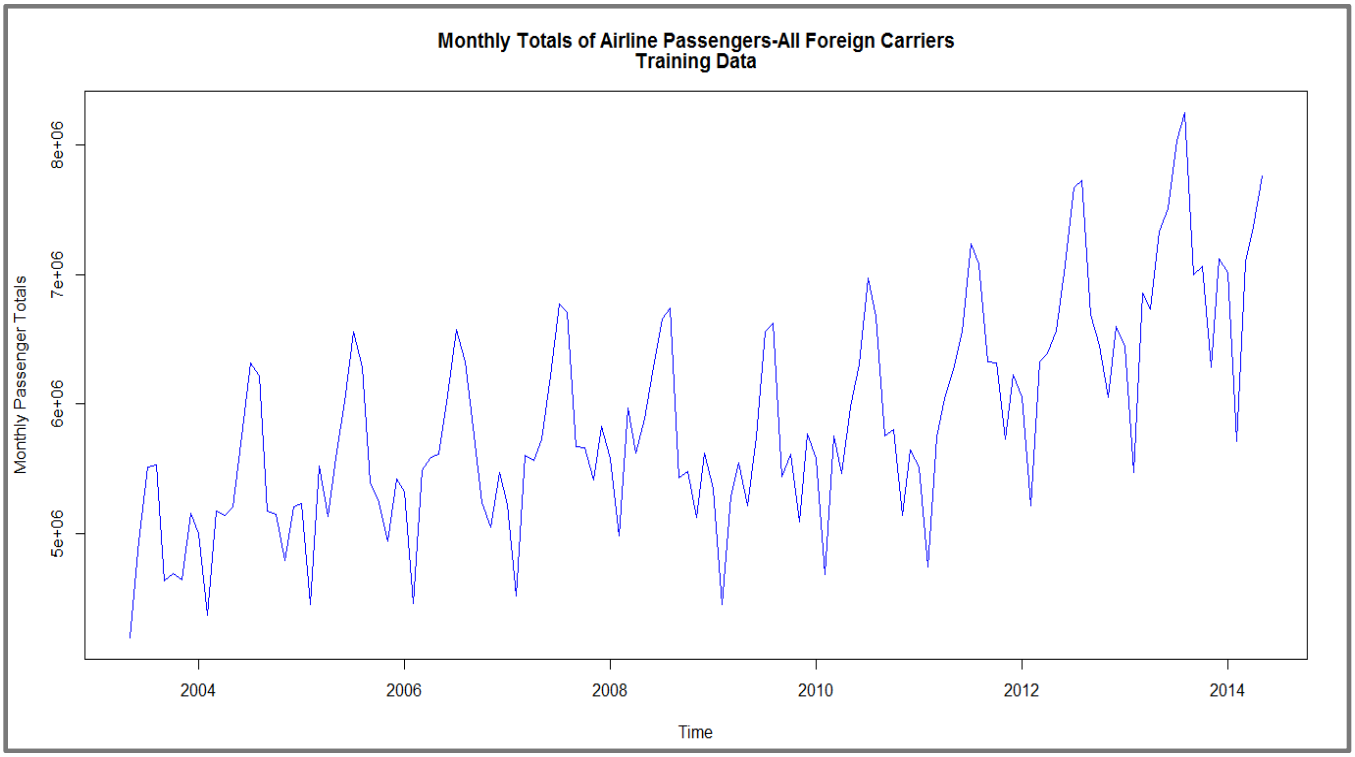

Fig 3.1.2: Training data set plot

3.2 Evaluation of raw data

For initial analysis purpose we tried different plots using some RStudio methods. *Fig 3.2.1* shows month plot of the total number of passenger travelling to US from all foreign carriers. With a close look at this month plot, we can understandthat the number of passengers flying in the month of July and August are at the highest peaks and on the other side less people travel in the month of February. Some of the months shows almost equal amount of passengers travelling to US via foreign carrier.

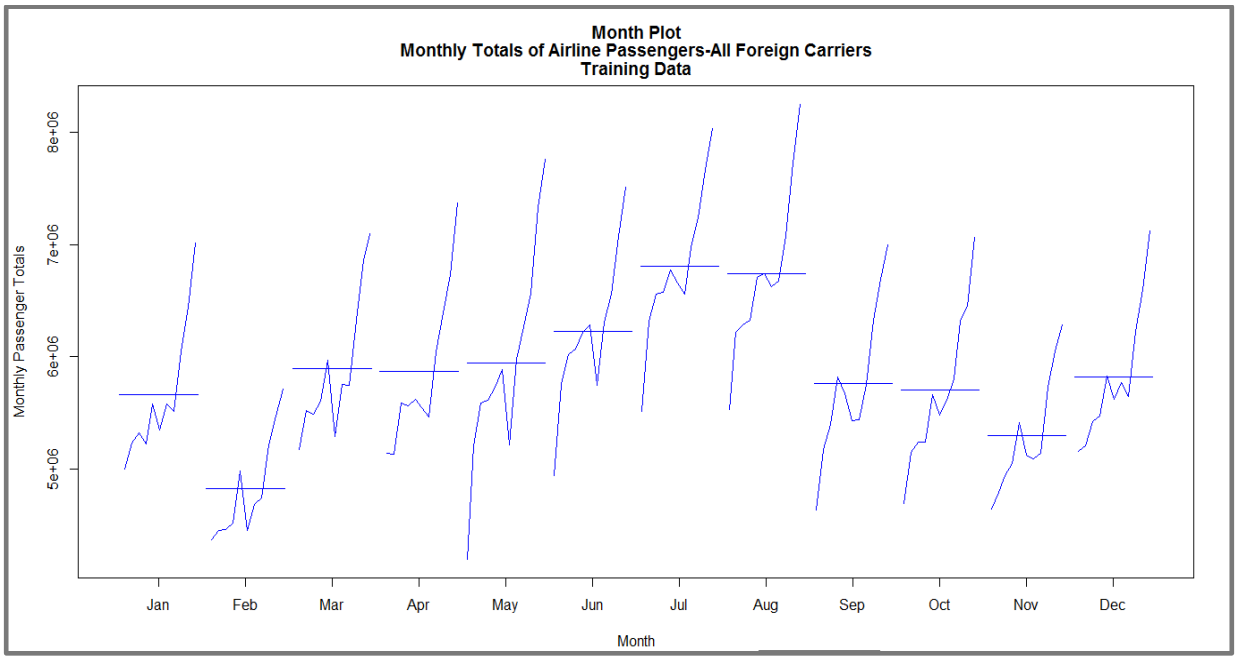

Fig 3.2.1: Month plot of passenger totals

The *Fig 3.2.2* below shows season plot of the total number of mothly passengers over the years from 2004 to 2015. The trend suggests that there is a certain increase every year for the number of passengers. It also shows that there is almost a similar pattern for monthly data over the years. The month of February doesn't show the significant increase every year, where as July and August seems to have significant growth in number of passengers over the years.

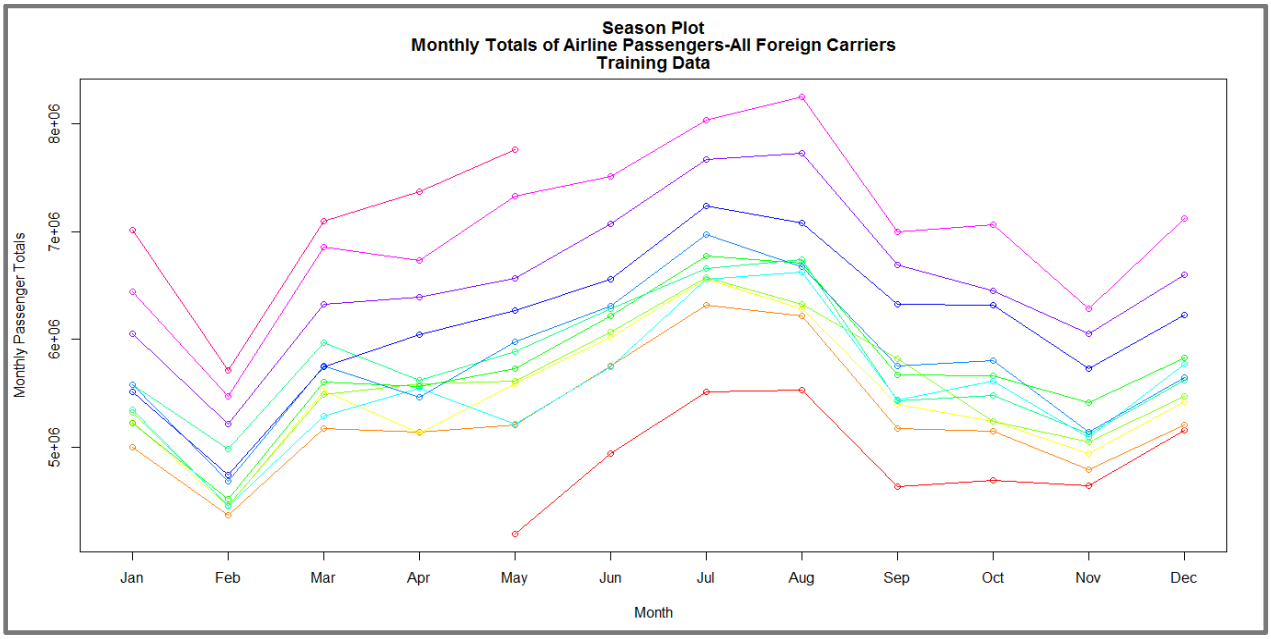

Fig 3.2.2: Season Plot of passenger totals

For a closer initial diagonistics we have used RStudio's Seasonal-trend decomposition *stl()* function to extract the ratios in which deterministic trend and seasonality are present in our training data. The *Fig 3.2.3* below shows the data in four different parts:

1) Raw data = 100%
2) Seasonal Portion = 28.1%
3) Trend Portion = 55.4 %
4) Residuals = 11.9 %

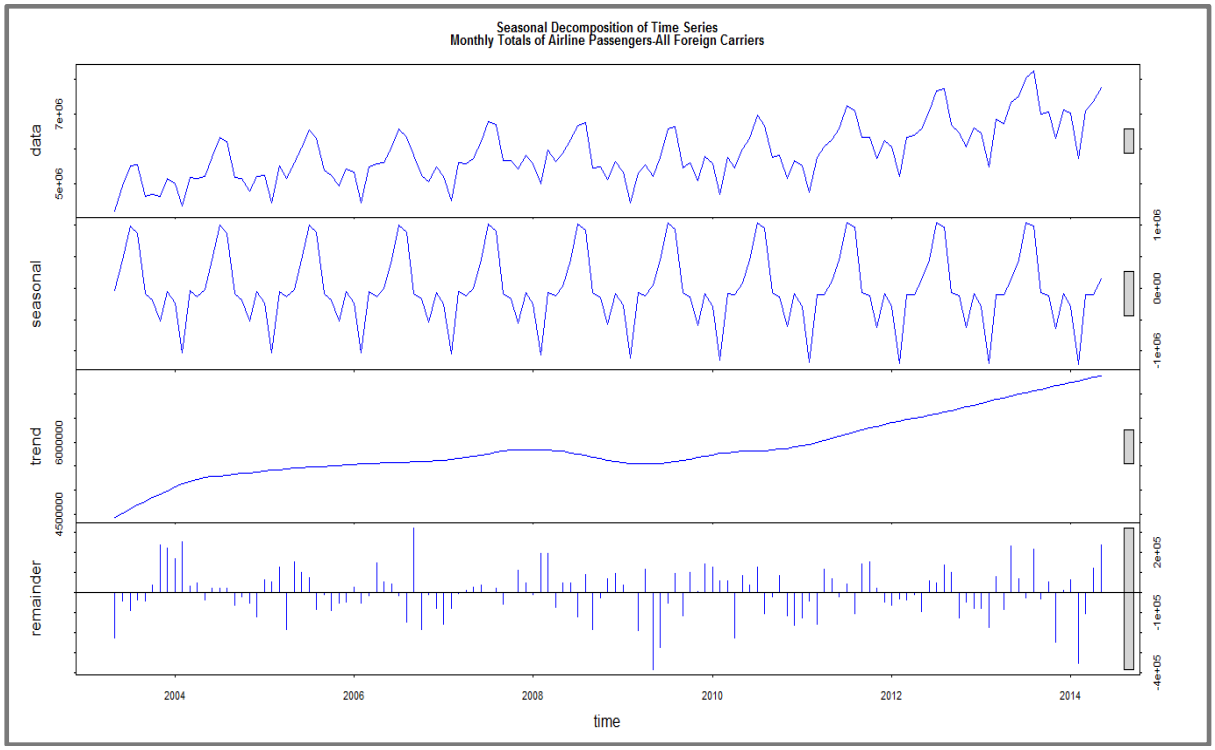

Fig 3.2.3: Seasonal Decompositin of Time Series

The graph indicates that there is periodicity in the data integrated with deterministic trend. The combined deterministic part accounts for almost 83.5% of the raw data, which indicates large explanatory power of the deterministic model (which will be further supported in the analysis of joint optimization in Section 4.3). The residual part is the stochastic part on which we will fit an ARMA model at a later stage.

After observing the trend from the figure, it seem to have a polynomial trend because the growth is not continuous and exponential.We will try and compare different fitted model over the training data and select the optimum one with lesser residuals.

4. MODELING PROCEDURE

For non-stationary series with deterministic trends & seasonality, we have:

$$y_t = f(t) + X_t$$

Where, y_t = Actual data
$f(t)$= Deterministic part with growth trend and periodic trend
X_t = Stochastic part

We followed four steps in producing our model:

1. Identify and model the data trend including deterministic periodicity;
2. Calculate the residuals from the de-trended data;
3. Identify the adequate order of the ARMA(n,m) model, using the standard (2n, 2n-1) approach; and,
4. Jointly optimize the parameters for an integrated model with the trend identified in [1] and the ARMA(n,m) identified in [3].

The raw data may or may not contain sine cosine periodicity along with polynomial or exponential trends. Because of the yearly repetition of the patterns we are sure about the seasonality present in the raw data along with the deterministic trend. We will use the F-test to compare and choose between the different orders of trends and seasonality. We will stop at a particular polynomial or exponential order if the F-test suggests that the reduction in residual sum of squares is insignificantly small. The model equation will be of the following form, where *time* is the index of time in the data:

$$y_t = \sum_{j=0}^{l} \beta_j * time^j + \sum_{j=1}^{i} \{\delta_{0,j} \sin\left(\frac{2\pi * time}{12} * j\right) + \delta_{1,j} \cos\left(\frac{2\pi * time}{12} * j\right)\} + X_t \qquad \text{(Eqn. 4.1)}$$

4.1 Modeling of deterministic part

4.1.1 Polynomial Growth Trend

At first place we started fitting exponential of 1st order to the growth trend of deterministic part using following equation:

$$y_t = R_1 * e^{r_1 * t} + \varepsilon_t \qquad \text{(Eqn. 4.1.1)}$$

The initial guesses for the parameters R_1 and r_1 were obtained from the plot as R_1 = 4191039 (when t = 0, $R_1 \approx y_0$ = 4191039) and r_1 = 0.005747 (when t = 1/2r_1, $y_t \approx R_1 e^{0.5}$ = 6909855, and *Fig 3.1.2* gives an average value of t = 87 months. Thus, r_1 = 0.005747). With these initial estimates, the nonlinear least squares routine yields R_1 = 4808000, r_1 = 0.003044. The resulting RSS was 6.73945×10^{13}.

We then fitted the growth trend again, but now by fitting polynomial of 1st order using following equation:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t \qquad \text{(Eqn. 4.1.2)}$$

The results for this fit are presented in *Table 4.1.1.1*. The RSS obtained by this fit is 5.19139×10^{13} which is less than the RSS obtained from the 1st order exponential fit.

When we compared the residuals for next higher order of exponential fit with polynomial fit, we observed that always a higher RSS resulted by fitting exponential trend. Thus, we restricted our analysis to fitting polynomial

orders to the growth trend for de-trending the data and then based on F-test determine the adequate order of polynomial. The equation for polynomial fit we used was,

$$y_t = \sum_{j=0}^{l} \beta_j * time^j + \varepsilon_t \quad \text{(Eqn. 4.1.3)}$$

With this equation we tried fitting different order polynomials. We compared the different orders by using F-test. The results of the F-tests and estimates of polynomial co-efficient with 95% confidence bounds are listed in table below:

Polynomial	**First order (l=1)**	**Second order (l=2)**	**Third order (l=3)**	**Fourth order (l=4)**
Total Number of parameters	2	3	**4**	5
β_0	4927428 ± 215188.988	5234084 ± 319481.9404	**4816612 ± 419923.336**	4591730± 532444.0156
β_1	14185 ± 2786.6888	556 ± 11007.0268	**37256 ± 27037.6316**	69727 ±54619.1632
β_2		102 ± 79.576	**-580 ± 467.9108**	-1661 ±1648.7716
β_3			**3 ± 2.2932**	16 ± 18.4632
β_4				0 ± 0.0588
Residual sum of Squares	5.19139×10^{13}	4.95228×10^{13}	**4.64972×10^{13}**	4.58542×10^{13}

Table 4.1.1.1: Estimation of parameters of model (Eqn. 4.1.1)

Polynomial order	**F-statistic**	**F-critical**
l = 1	-	-
l = 2	6.2766	3.84
l = 3	8.3943	3.84
l = 4	1.7948	3.84

Table 4.1.1.2: F-values

Based on the results of F-test which are listed in *Table 4.1.1.2*, when comparing 3rd (l=3) and 4th (l=4) order polynomial we get F-statistic value as 1.7948, which is less than the F-critical value of $F_{0.95,[1,\infty)}$= 3.84. This indicates insignificance of using 4th order polynomial to fit the growth trend of the deterministic part. Thus, the growth trend appears to be best modeled with polynomial order 3.

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \varepsilon_t \quad \text{(Eqn. 4.1.4)}$$

The *Fig 4.1.1.1* below shows that, the fitted polynomials of different orders start from first order (shown with red) to fourth order (shown with light blue). The 3rd order polynomial (shown with dark blue) shows a good fit over the actual data.

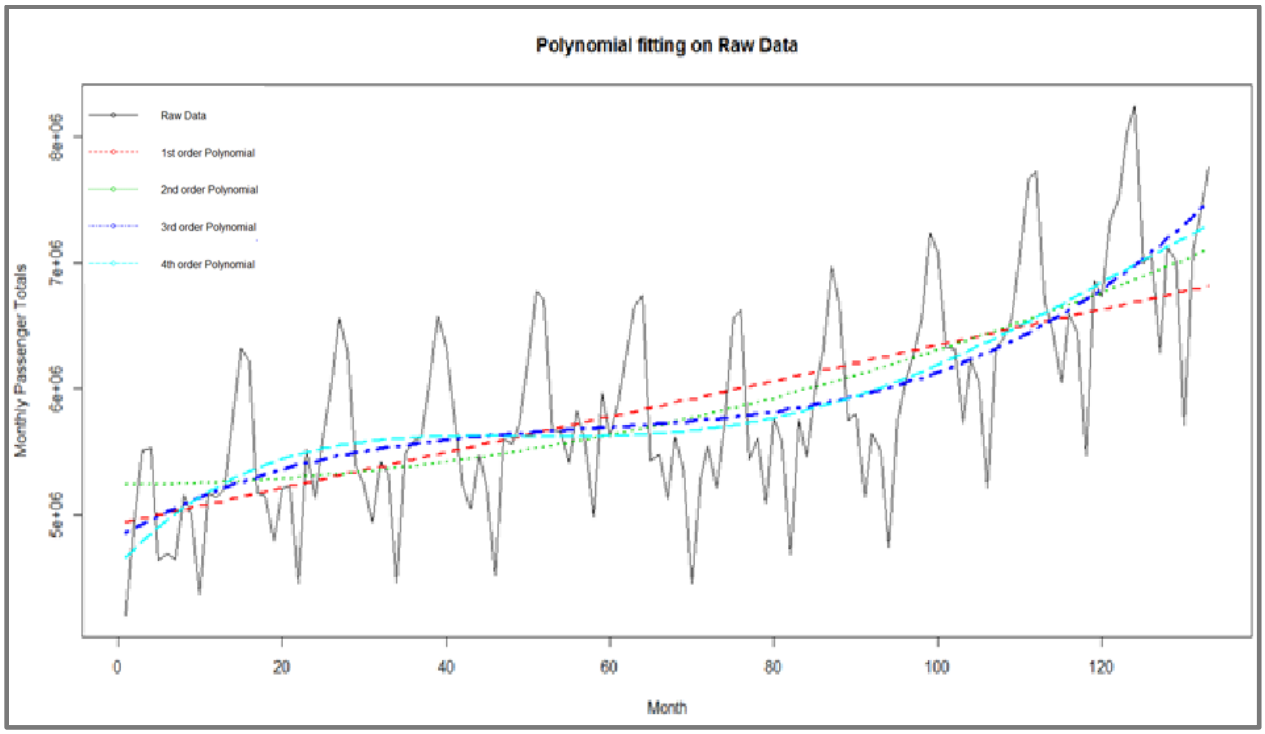

Fig 4.1.1.1: Polynomial fitting on raw data

4.1.2 Addition of Periodic Trends:

The residuals from the model *Eqn. 4.1.4* are shown in *Fig 4.1.2.1*. The polynomial growth present in the original data has now been removed and the residuals only have periodic trends with dominant period of 12 months. The plot also suggests possible existence of some harmonics of the dominant period.

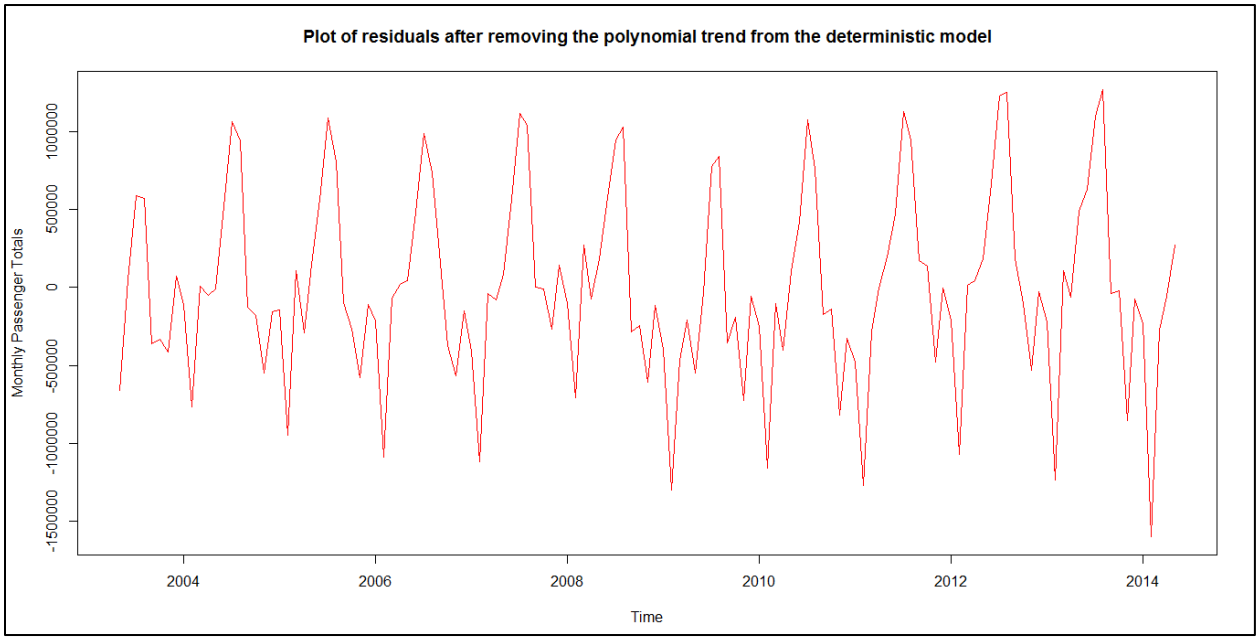

Fig 4.1.2.1: Plot of residuals from deterministic Eqn. 4.1.4

Taking l = 3 in the model *Eqn. 4.1*, we now started adding the periodic trends. The first model we started fitting is therefore,

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \delta_{0,1}\sin\left(\frac{2\pi * t}{12}\right) + \delta_{1,1}\cos\left(\frac{2\pi * t}{12}\right) + \varepsilon_t \qquad \text{(Eqn. 4.1.5)}$$

Where,

$$\frac{2\pi * t}{12} = \omega$$

is the dominant frequency in radians per month for the major peaks clearly visible in the residuals of the model *Eqn. 4.1.4* as also in the original data.

The parameters were estimated by generalized least method in RStudio using *gls()* function by fitting polynomial order 3 along with sine and cosine functions. The results of estimation with 95% confidence bounds are shown in 2nd column of *Table 4.1.2.1*.

The residual sum of squares was drastically reduced from 4.64972×10^{13} to 1.99863×10^{13} by inclusion of only one period of 12 months. This reduction in RSS is also supported by *Fig 3.1.2* and *Fig 4.1.2.1* which suggests that the yearly period can account for a large part the variation in the series.

Since the improvement in RSS is large, we successively fit the models for values l = 3 and i = 2, 3, 4, . . . until the reduction in the RSS is insignificantly small. To test this adequacy of the model we performed F-test between a higher and a lower period model. The estimation results are shown in the *Table 4.1.2.1*. The F-statistic for the F-test performed between periods i = 5 and i = 6 was calculated as 0.67384 which is less than F-critical, $F_{0.95,[2,\infty)}$= 3 which indicated that adding the sixth period does not result in significant improvement and therefore we stop at i = 5. The calculated F-statistic and corresponding F- critical are shown in *Table 4.1.2.2*.

i	1	2	3	4	5	6
Total no. of parameters	6	8	10	12	**14**	16
β_0	4642116 ± 278815.7	4617139 ± 250980.5	4641940 ± 207262.3	4651502 ± 185613.7	**4641879 ± 139869**	4642261 ± 140275.2
β_1	48286 ± 17953.6	50073 ± 16160.63	48578 ± 13344.8	48176 ± 11950.1	**48652 ± 9004.338**	48613 ± 9029.72
β_2	-765 ± 310.6796	-798 ± 279.6528	-774 ± 230.9076	-770 ± 206.78	**-776 ± 155.8004**	-775 ± 156.8
β_3	4 ± 1.5288	4 ± 1.372	4 ± 1.1368	4 ± 1.0192	**4 ± 0.7644**	4 ± 0
$\delta_{0,1}$	635645 ±96072.1636	637722 ±86384.4716	637580 ±71304.8	638738 ±63845.3732	**638061 ±48105.28**	637824 ±48239.52
$\delta_{1,1}$	-340740 ±95386.7712	-33605 ±85771.4228	-31771 ±70803.334	-29595 ±63397.964	**-31196 ±47770.34**	-301910 ±47967.08
$\delta_{0,2}$		24455 ±85635.7124	25540 ±70691.32	27674 ±63297.7296	**26216 ±47694.88**	241320 ±48012.16
$\delta_{1,2}$		-247562 ±86007.348	-246375 ±70994.66	-245096 ±63566.3476	**-246054 ±47895.05**	-245006 ±48090.56
$\delta_{0,3}$			-109326 ±70537.38	-106834 ±63161.392	**-108579 ±47593.27**	-111346 ±48057.24
$\delta_{1,3}$			259546 ±71069.2864	259569 ±63631.6352	**259511 ±47943.15**	259670 ±48076.84
$\delta_{0,4}$				-137442 ±63273.11	**-138971 ±47676.35**	-137759 ±47900.44
$\delta_{1,4}$				121697 ±63497.02	**122541 ±47842.82**	122779 ±47976.88
$\delta_{0,5}$					**-108820 ±47837.68**	-109224 ±47974.92
$\delta_{1,5}$					**-210257 ±47674.53**	-209699 ±47816.16
$\delta_{0,6}$						128402 ±27783.59
$\delta_{1,6}$						-160800 ±41310.92
Residual sum of squares	1.99863×10^{13}	1.59029×10^{13}	1.06618×10^{13}	0.840801×10^{13}	$\mathbf{0.469419\times10^{13}}$	0.464074×10^{13}

Table 4.1.2.1: Estimation of parameters of model Eqn. 4.1

Period	F-statistic	F-critical
i = 1	-	-
i = 2	16.0478	3
i = 3	30.2322	3
i = 4	16.2173	3
i = 5	47.07336	3
i = 6	0.67384	3

Table 4.1.2.2: F-values

As we can see, the trend appears to be best modeled with an order 3 polynomial, plus sine and cosine functions with periodicity 12, 6, 4, 3 and 2.4. The introduction of these five periods accounts for yearly, half-yearly, 4 monthly, quarterly, and 2-2/5 monthly periods. It was therefore natural that periods smaller than 2-2/5 monthly may not be necessary.

Upon inspection of the parameters of the selected model (i = 5) and their respective p-values, we observed that cosine parameter of 1st period ($\delta_{1,1}$) and sine parameter of 2nd period ($\delta_{0,2}$) had p-value greater than 0.05 indicating their insignificance. *Table 4.1.2.3* below shows estimation of parameters of the reduced model.

	i=5	**P value**	**i=5 (Eliminating $\delta_{1,1}$ and $\delta_{0,2}$)**	**P value**
β_0	4641879 ± 139869	0.0000	4647563 ± 140147.2	0.0000
β_1	48652 ± 9004.338	0.0000	48266 ± 9021.331	0.0000
β_2	-776 ± 155.8004	0.0000	-769 ± 156.0748	0.0000
β_3	4 ± 0.7644	0.0000	4 ± 0.7644	0.0000
$\delta_{0,1}$	638061 ± 48105.28	0.0000	637756 ± 48253.73	0.0000
$\delta_{1,1}$	**-31196 ± 47770.34**	**0.2030**	-	-
$\delta_{0,2}$	**26216 ± 47694.88**	**0.2835**	-	-
$\delta_{1,2}$	-246054 ± 47895.05	0.0000	-245951 ± 48045.13	0.0000
$\delta_{0,3}$	-108579 ± 47593.27	0.0000	-108633 ± 47737.23	0.0000
$\delta_{1,3}$	259511 ± 47943.15	0.0000	259598 ± 48094.7	0.0000
$\delta_{0,4}$	-138971 ± 47676.35	0.0000	-138993 ± 47822.35	0.0000
$\delta_{1,4}$	122541 ± 47842.82	0.0000	122612 ± 47987.22	0.0000
$\delta_{0,5}$	-108820 ± 47837.68	0.0000	-108827 ± 47820.32	0.0000
$\delta_{1,5}$	-210257 ± 47674.53	0.0000	-210198 ± 47674.53	0.0000
RSS	0.469419×10^{13}		0.480334×10^{13}	

Table 4.1.2.3: Re-estimation of parameters after eliminating $\delta_{1,1}$ and $\delta_{0,2}$

Thus we again performed F-test between this reduced model and model with i = 5 to ascertain whether the smaller model was adequate. The F-statistic we calculated came out to be 1.38342 which was less than the corresponding F-critical indicating the significance of the lower model. Thus, we ended up eliminating the cosine function with 12 month periodicity and the sine function with 6 month periodicity.

So now we have adequately modeled the growth trend and periodic trend of the deterministic part. *Fig 4.1.2.2* shows the comparison of the deterministic part and the actual data. The fit of this deterministic component is pictorially represented by a solid blue line, the original data are shown by black dots. This figure also confirms the adequacy of the deterministic part of the model.

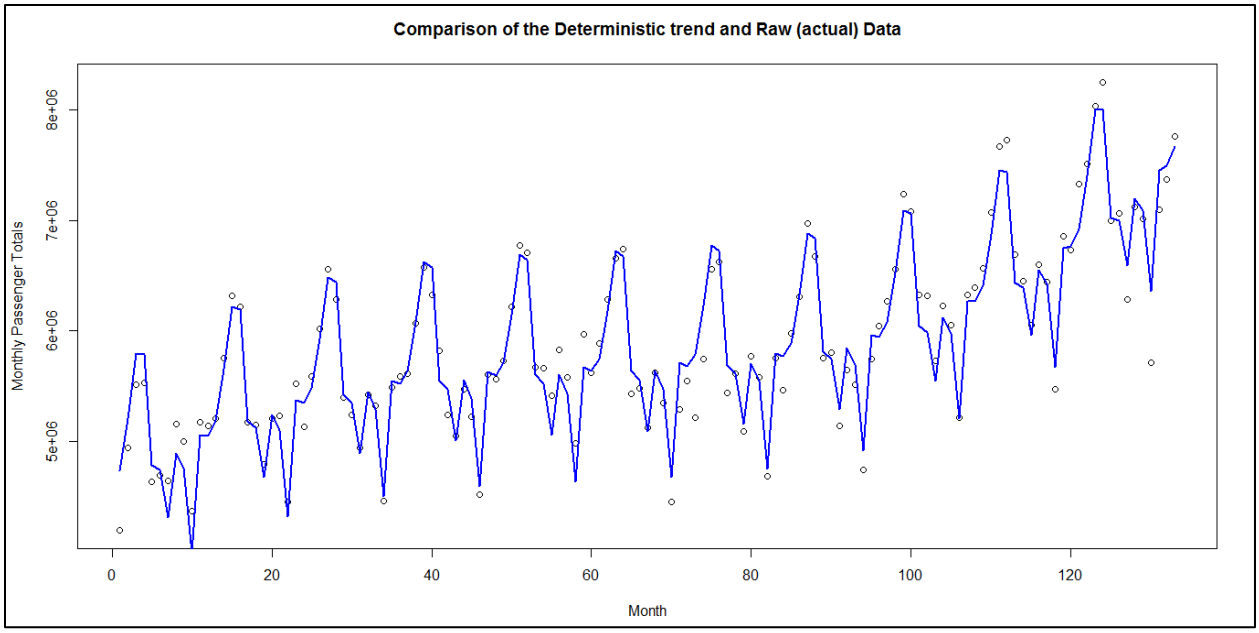

Fig 4.1.2.2: Comparison of the Deterministic trend and actual data

4.2 Specifying and Modeling the Stochastic Part

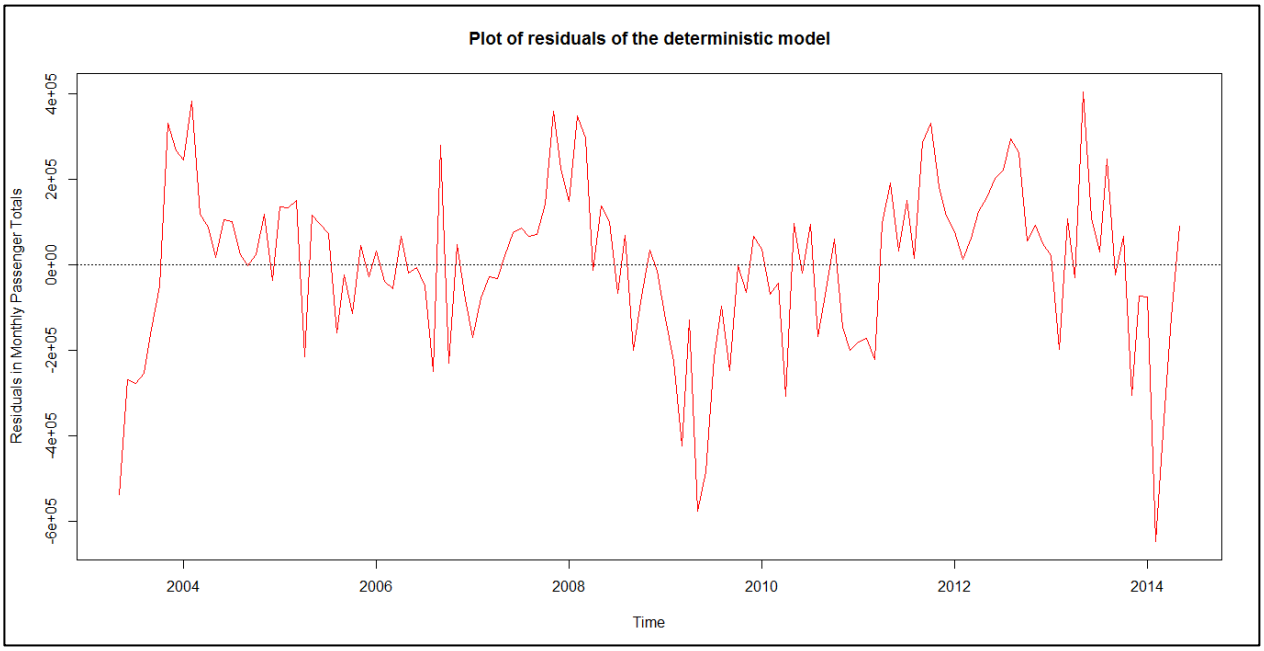

Fig 4.2.1: Plot of residuals of the deterministic model

The residuals from the model *Eqn. 4.1*, shown in *Fig 4.2.1* above, appear to be stationary series without persistent nonstationary trends. In order to check this stationarity, we graphed Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots as shown in *Fig 4.2.2* and checked significant lags.

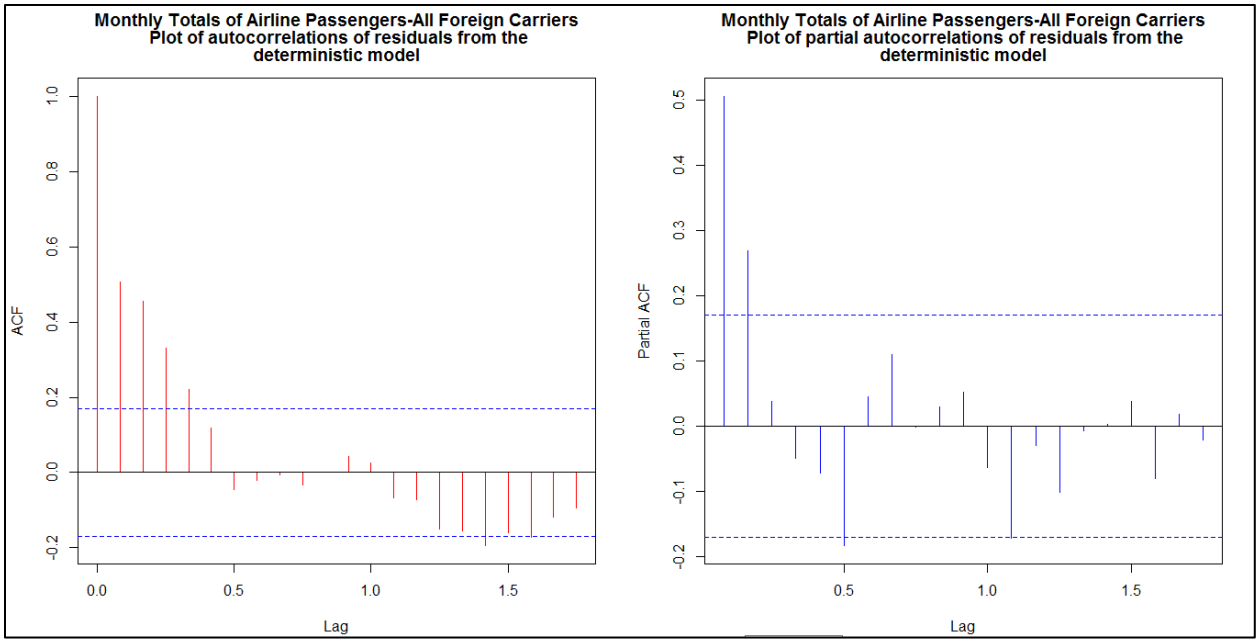

Fig 4.2.2: ACF and PACF plots of residuals from deterministic model

Both graphs as seen have a few significant lags, but the autocorrelations die out asymptotically, so we can conclude that our series is stationary. In fact, the partial autocorrelation shows almost no significance after the second lag. Based on this, we expected a small order in our ARMA model.

In addition, to support stationarity, we performed Augmented Dicker-Fuller (ADF) t-statistic test in RStudio using function *adf.test ().* The alternative hypothesis for this test is stationarity. Based on this test performed on residuals from the deterministic model, we obtained t-statistic = -4.4549 and a p-value = 0.01. The small p-value suggests that the residuals are stationary.

Thus, the residuals can be considered as the stochastic part of the final model and the modeling procedure of ARMA is now applicable.

4.2.1 ARMA Modeling on Stochastic Part

After the appropriate fitting of deterministic trend from the raw data with help of R studio, we are left with only a small amount of stochastic variance in the data. Here, we will try to further reduce the RSS of 4.80334 $\times 10^{12}$ obtained after fitting the deterministic part.

For ARMA modeling, we will start by fitting ARMA (2n, 2n-1) with n=1. For every increase of "n" by one, we will check the improvement in the residual sum of squares of a_t's by the F-criterion. Thus, to find an adequate model for the stochastic part, we first fitted ARMA(2,1) and ARMA(4,3) models in RStudio using *arima() method.* This modeling procedure is briefly detailed below;

We first compared the model,

- ARMA(2,1) vs ARMA(4,3) → F-test (significant) → Select higher model ARMA(4,3)
- ARMA(4,3) vs ARMA(6,5) → F-test (insignificant) → Select lower model ARMA(4,3)

- Check whether ϕ_{2n} and/or θ_{2n-1} includes zero → if yes, drop these parameters
- ϕ_4 and θ_3 includes zero → Drop both parameters and check significance with ARMA(3,2)

- ARMA(3,2) vs ARMA(4,3) → F-test (insignificant) → Select lower model ARMA(3,2)

- Dropping MA parameter check significance of ARMA(3,1) with ARMA(3,2)
- ARMA(3,1) vs ARMA(3,2) → F-test (significant) → Select higher model ARMA(3,2)
- ϕ_3 includes zero→ Drop ϕ_3 and check significance with ARMA(2,2)

- ARMA(2,2) vs ARMA(3,2) → F test (insignificant) → Select lower model ARMA(2,2).
- No further reduction in ARMA order was significant

Table below shows estimation of parameters of the ARMA models.

Parameters	**ARMA (2,1)**	**ARMA (4,3)**	**ARMA (6,5)**	**ARMA (3,2)**	**ARMA (2,2)**
Φ_1	0.4896 ± 0.473	1.3634 ± 0.958	0.0305 ± 1.116	1.5757 ± 0.222	1.5169 ± 0.234
Φ_2	0.2346 ± 0.298	-0.6955 ± 1.539	-0.0987 ± 0.903	-1.0066 ± 0.533	-0.7354 ± 0.217
Φ_3		-0.148 ± 1.125	0.2544 ± 0.914	0.1969 ± 0.382	
Φ_4		0.188 ± 0.318	-0.0828 ± 0.879		
Φ_5			-0.3387 ± 0.777		
Φ_6			0.0942 ± 0.409		
θ_1	0.1137 ± 0.476	1.0777 ± 0.966	-0.3118 ± 1.098	1.2582 ± 0.134	1.2427 ± 0.206
θ_2		-0.6963 ± 1.264	-0.5216 ± 0.519	-0.8567 ± 0.319	-0.7262 ± 0.193
θ_3		-0.2319 ± 0.959	-0.1043 ± 1.074		
θ_4			-0.4550 ± 0.668		
θ_5			-0.6628 ± 1.034		
Mean (μ)	-10790.5 ± 83412.41	-3299.62 ± 71402.49	-2596.49 ± 64878.94	-1668.48 ± 64655.22	1521.14 ± 56504.6
Residual Sum of Squares	3.20881×10^{12}	2.84846×10^{12}	2.73525×10^{12}	2.97485×10^{12}	3.01396×10^{12}

Table 4.2.1.1: Parameter Estimates of ARMA model

Sample F-test calculation between ARMA (2, 1) & ARMA (4, 3):

The F-criterion is given by;

$$F = \frac{A1-A0}{s} \div \frac{A0}{N-r} \sim F(s, N\text{-}r)$$

where,

A_1 = RSS of lower model = RSS of ARMA(4,3) = **3.20881×10^{12}**

A_0 = RSS of higher model = RSS of ARMA(2,1) = **2.8485×10^{12}**

s = Number of additional parameters = (4+3) – (2+1) = **4**

r = Number of unrestricted parameters = (4 + 3) + 1 = **8**

N = number of observations = **133**

The residual sum of squares of the ARMA(4,3) model is smaller than that of ARMA(2,1) model and the F-criterion shows

$$F = \frac{(3.20881 - 2.8485)}{4} \div \frac{2.8485}{133 - 8} \sim F(s, N\text{-}r)$$

F = **3.9533**

F-critical = $F_{0.95,[4,\infty)}$ = **2.37**

We can see that F-statistic is greater than F-critical, i.e. F-test shows significance for ARMA(4,3). Similarly, other F-tests were performed and the results from F-test are listed below.

ARMA	F-statistic	F-critical
(2,1)	-	-
(4,3)	3.9533	2.37
(6,5)	1.2521	2.37
(3,2) vs (4,3)	2.7731	3
(3,1) vs (3,2)	9.5629	3.84
(2,2) vs (3,2)	1.6698	3.84

Table 4.2.1.2: F-values

ARMA(2,2) turns out to be the adequate model for the stochastic part from the above analysis.

4.3 Jointly Optimizing the Integrated Model

We are now in a position to formulate the final adequate combined model with a deterministic plus stochastic part and estimate all the parameters simultaneously. The combined model has 3rd order polynomial trend (l=3), 5th order sine-cosine periodic trend (i = 5, excluding $\delta_{1,1}$ and $\delta_{0,2}$) and the appropriate ARMA order (2, 2).

This model has the following form:

$$y_t = \sum_{j=0}^{l=3} \beta_j * time^j + \sum_{j=0}^{i=5} \{\delta_{0,i}\sin\left(\frac{2\pi * time}{12} * j\right) + \delta_{1,i}\cos\left(\frac{2\pi * time}{12} * j\right)\} + X_t \qquad \text{(Eqn. 4.3.1)}$$

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

We started with estimates ϕ_1, ϕ_2, θ_1, θ_2, β_0, β_1, β_2, β_3, δ_{01}, δ_{12}, δ_{03}, δ_{13}, δ_{04}, δ_{14}, δ_{05}, δ_{15} in the separate models obtained above as the initial values. The nonlinear least squares routine minimizing the sum of squares of a_t's in the model *Eqn. 4.3.1* yielded the following estimates as shown in column 3 of *Table 4.3.1* below.

Parameter	Initial Estimates	Joint Estimates	Std. Error	p-value
$\boldsymbol{\phi_1}$	$1.5169\times10^{+00}$	$1.51823\times10^{+00}$	1.2271×10^{-01}	1.4695×10^{-22}
$\boldsymbol{\phi_2}$	-7.354×10^{-01}	-7.30071×10^{-01}	1.1763×10^{-01}	2.0339×10^{-08}
$\boldsymbol{\theta_1}$	$1.2427\times10^{+00}$	$1.2365\times10^{+00}$	1.1242×10^{-01}	2.5707×10^{-19}
$\boldsymbol{\theta_2}$	-7.262×10^{-01}	-7.1393×10^{-01}	1.0459×10^{-01}	1.0251×10^{-09}
$\boldsymbol{\beta_0}$	$4.6475\times10^{+06}$	$4.5961\times10^{+06}$	$1.2105\times10^{+05}$	1.7849×10^{-67}
$\boldsymbol{\beta_1}$	$4.8266\times10^{+04}$	$5.19852\times10^{+04}$	$7.7848\times10^{+03}$	2.1188×10^{-09}
$\boldsymbol{\beta_2}$	$-7.690\times10^{+02}$	$-8.3758\times10^{+02}$	$1.3426\times10^{+02}$	1.7466×10^{-08}
$\boldsymbol{\beta_3}$	$4.000\times10^{+00}$	$4.67977\times10^{+00}$	6.5612×10^{-01}	2.2346×10^{-10}
$\boldsymbol{\delta_{01}}$	$6.3775\times10^{+05}$	$6.4566\times10^{+05}$	$3.7805\times10^{+04}$	3.5537×10^{-33}
$\boldsymbol{\delta_{12}}$	$-2.4595\times10^{+05}$	$-2.4847\times10^{+05}$	$1.2199\times10^{+04}$	6.3168×10^{-40}
$\boldsymbol{\delta_{03}}$	$-1.0863\times10^{+05}$	$-1.0634\times10^{+05}$	$1.5132\times10^{+04}$	3.7770×10^{-10}
$\boldsymbol{\delta_{13}}$	$2.5959\times10^{+05}$	$2.581\times10^{+05}$	$1.5255\times10^{+04}$	7.6368×10^{-33}
$\boldsymbol{\delta_{04}}$	$-1.3899\times10^{+05}$	$-1.3698\times10^{+05}$	$1.6205\times10^{+04}$	2.4325×10^{-13}
$\boldsymbol{\delta_{14}}$	$1.2261\times10^{+05}$	$1.2057\times10^{+05}$	$1.6256\times10^{+04}$	5.3050×10^{-11}
$\boldsymbol{\delta_{05}}$	$-1.0882\times10^{+05}$	$-1.0766\times10^{+05}$	$1.6659\times10^{+04}$	6.0081×10^{-09}
$\boldsymbol{\delta_{15}}$	$-2.1019\times10^{+05}$	$-2.1289\times10^{+05}$	$1.6603\times10^{+04}$	1.2956×10^{-23}
RSS	3.01396×10^{12}	$\mathbf{3.00062\times10^{12}}$		

Table 4.3.1: Jointly optimized estimates

We note that these final jointly optimized estimates are not much different from initial estimates, primarily because the deviations from the deterministic part of the "noise" are much smaller in magnitude. This is the result of the large explanatory power of the deterministic model. Also, the RSS reduces slightly as expected.

The last column of *Table 4.3.1* shows the p-value for jointly optimized estimates. These p-values are extremely small indicating significance of all the parameters in the integrated model.

5. MODEL DIAGNOSTICS ON THE INTEGRATED MODEL

5.1 Checking stability of jointly estimated ARMA(2,2) model

The ARMA(2,2) model from integrated model is:

$$X_t = 1.51823X_{t-1} - 0.730071X_{t-2} + a_t - 1.2365a_{t-1} + 0.71393a_{t-2}$$

The Green's functions for the above model were calculated upto j = 30 as shown in *Table 5.1.1* below,

j	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Gj	1	0.28173	0.41159	0.41921	0.33597	0.20403	0.06448	-0.0511	-0.1246	-0.15187	-0.13961
j	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Gj	-0.10108	-0.0515	-0.00445	0.03085	0.050085	0.053518	0.04469	0.02877	0.01106	-0.00422	-0.01448
j	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**		
Gj	-0.01889	-0.01812	-0.01372	-0.0076	-0.0015	0.00323	0.006004	0.00676	0.005875		

Table 5.1.1 ***Green's functions for ARMA(2,2) model***

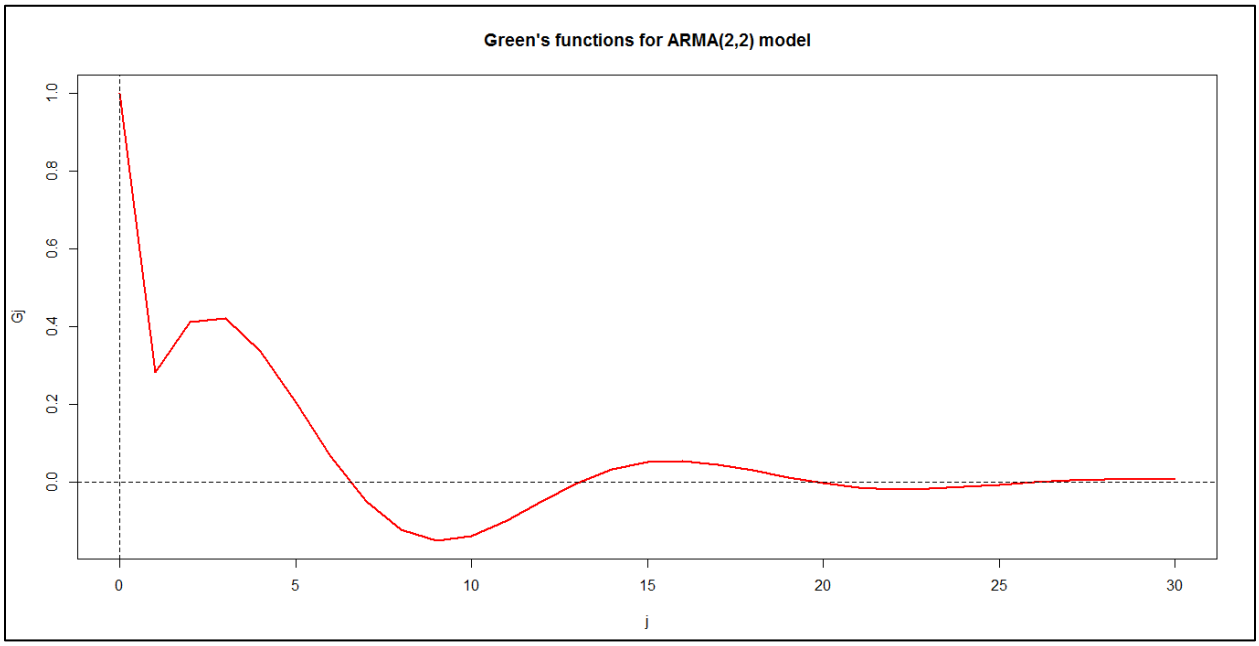

Fig 5.1.1: Green's functions plot

Fig 5.1.1 shows Green's functions plot which indicates the system asymptotically tend to zero. Also, the characteristic roots for jointly estimated ARMA(2,2) model were calculated as:

$$\lambda_1, \lambda_2 = 0.759115 \pm i(0.39219)$$

Since, $|\lambda 1| = |\lambda 2| = 0.8544 < 1$, the system is asymptotically stable. This is also supported by inverse AR roots plot shown in *Fig 5.1.2*. Also, looking at the inverse MA roots plot we can say that the system is also invertible since the inverse MA roots lie inside the unit circle.

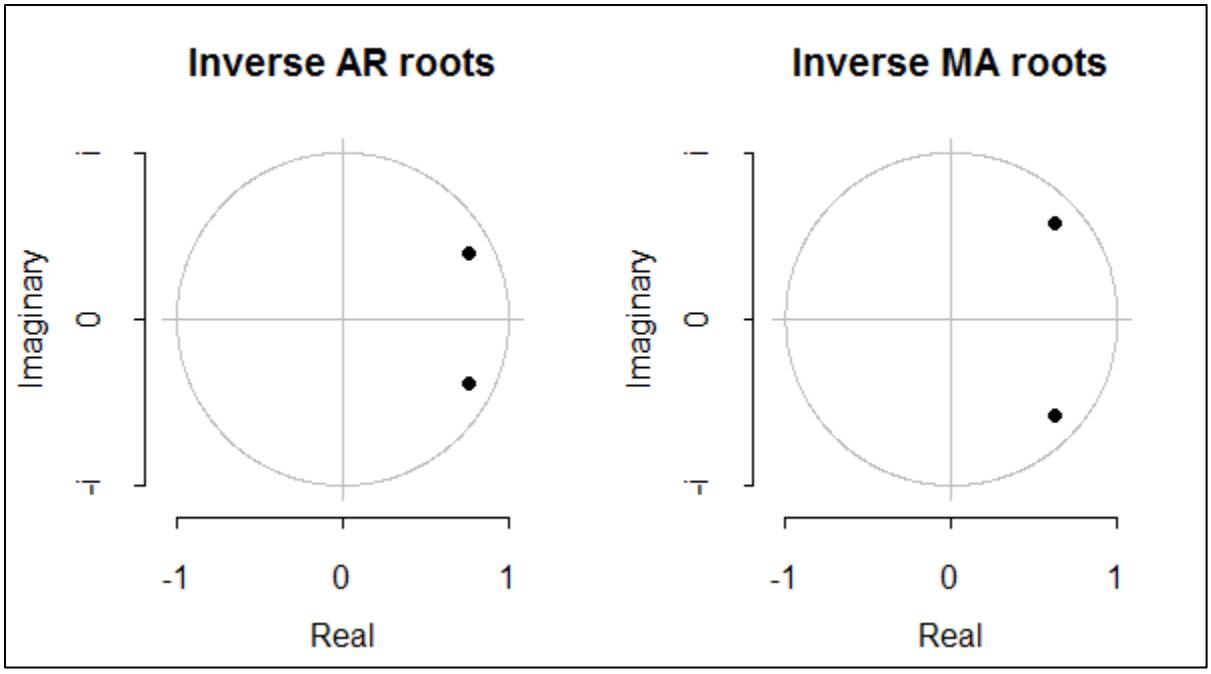

Fig 5.1.2: Inverse AR and MA roots of ARMA(2,2) model

Our complex roots λ_1, λ_2 indicated we have no stochastic seasonality, since the length of the root vector is not much close to unity. Thus, no parsimonious representation of this ARMA model was desired.

5.2 Comparison of the Jointly Optimized Model and Actual Data

The plot of this jointly optimized model (integrated model) against the actual data shown in *Fig 5.2.1* is visually identical to the plot of the deterministic trend model versus the data which was shown in *Fig 4.1.2.2*. The fit of this integrated model is represented by a solid blue line and the actual data are shown by black dots.

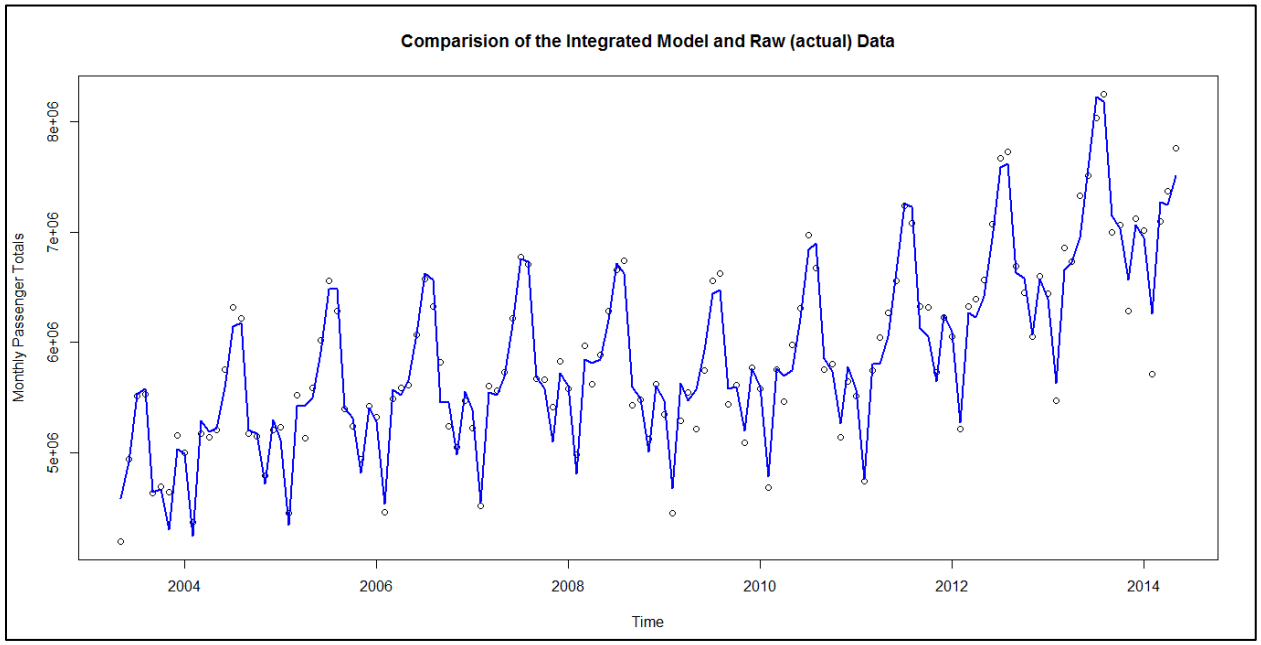

Fig 5.2.1 Comparison of the integrated model and actual data

5.3 Analysis of Residuals Obtained from the Integrated Model

The residual a_t's from the model *Eqn. 4.3.1* are shown in *Fig 5.3.1*. They appear to be scattered and do not show any trends characteristic of periodic components or autocorrelation in the data (i.e. residuals have constant variance). The plot of the autocorrelations of a_t's as shown in *Fig 5.3.2* also confirmed that the a_t's were uncorrelated, all the $\hat{\rho}_k$'s were within $\pm 2/\sqrt{N} = 0.1734$ bounds (threshold limits) indicating that the residuals from integrated model behave like white noise.

To further investigate, we performed the Box–Pierce and Ljung–Box test statistic for examining the null hypothesis of independence in RStudio using *Box.test( )* method. Both these tests resulted in large p-value of 0.9121 and 0.9037 respectively, thereby, indicating independence of residuals.

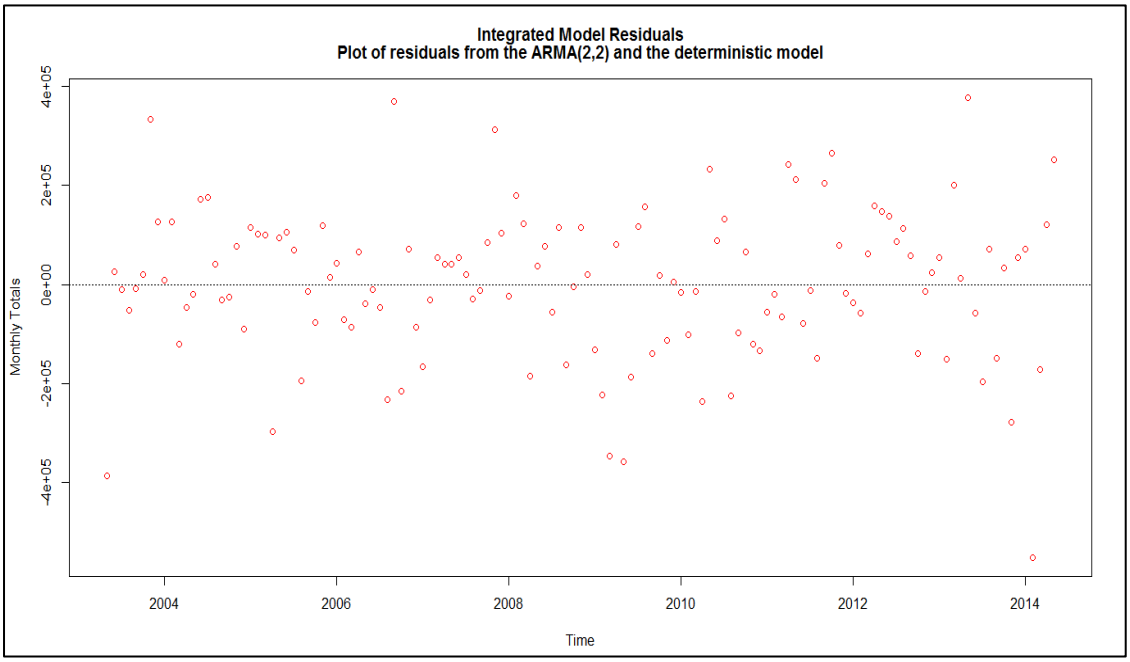

Fig 5.3.1: Plot of residuals from Integrated model

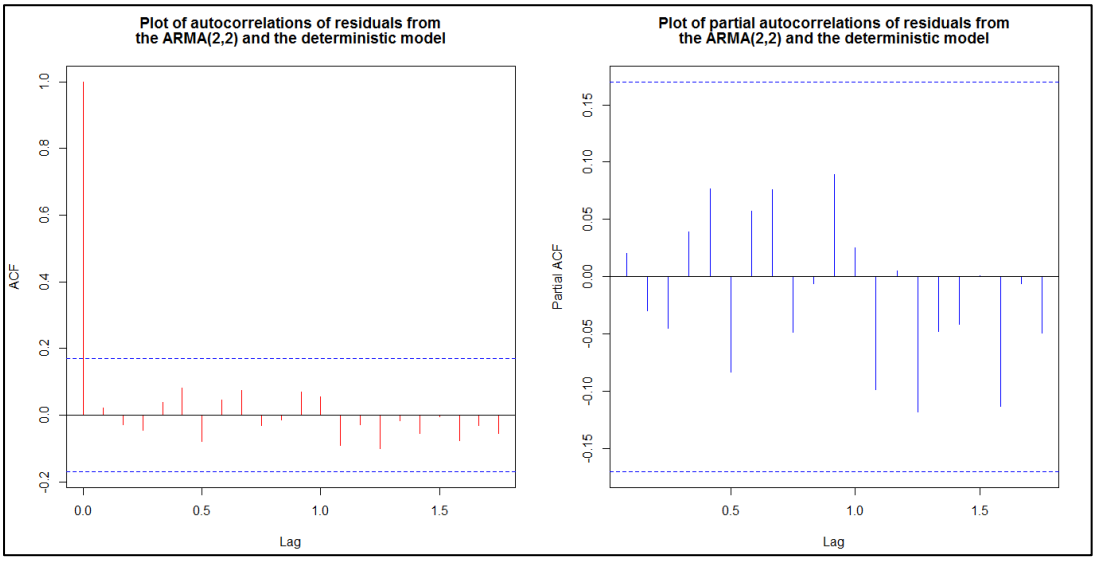

Fig 5.3.2: ACF and PACF plots of residuals from Integrated model

We then checked whether the residuals were normally distributed with mean zero, by making a histogram (shown with red bars) with overlaid normal curve (shown with blue line) as shown in *Fig 5.3.3* below,

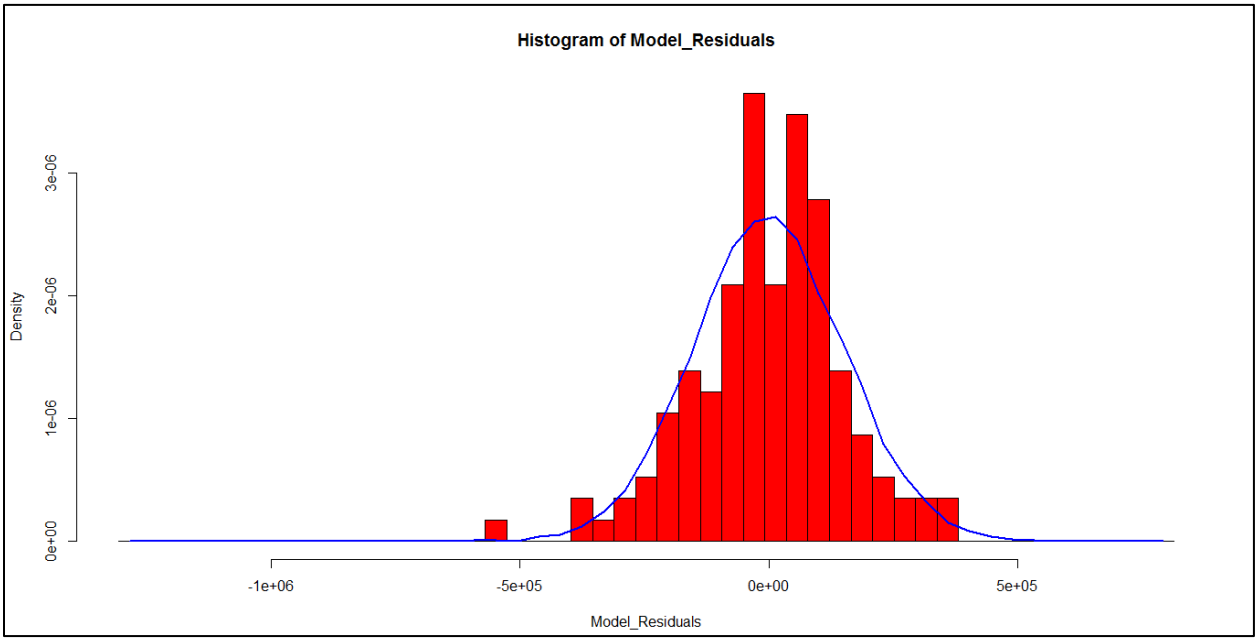

Fig 5.3.3: Histogram plot of Integrated model residuals

The histogram plot shows that the distribution of residuals is roughly centered on zero, and so it is plausible that the residuals are normally distributed with mean zero.

To further investigate, we performed Shapiro-Wilk normality test on our model residuals for examining the null hypothesis of normality in RStudio using *shapiro.test( )* method. This test resulted in a p-value of 0.06557, thereby, failing to reject null hypothesis indicating normality of residuals.

So, we now have an integrated model that passes the required checks and is ready for forecasting.

6. FORECASTING

To check the model adequacy, we will do forecasting over the testing data using joint optimization equation. The equation is an integrated form of deterministic & stochastic part together. We have performed forecasting for next 15 months based on previously modeled 133 data points. The right most part in the *Fig 6.1* below shows the forecasting part with the upper & lower 95% confidence limits. The close-up on this forecast interval is shown in *Fig 6.2*. Notice how the forecasts follow the recent trend in the data

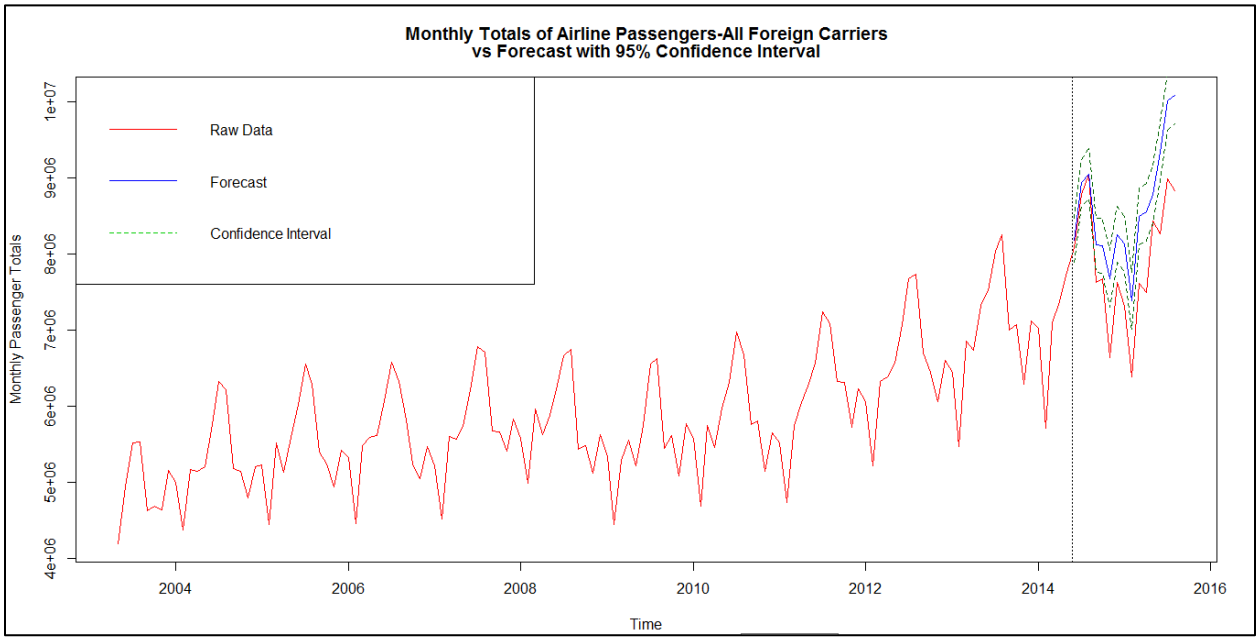

Fig 6.1: Forecasts from the model for the next 15 months

The figure below gives closer insights of the prediction interval. In this figure the blue line indicates the forecasts & the red line shows the actual figures for the monthly airline passengers traveling to US. The dotted lines show 95% confidence interval on forecasts.

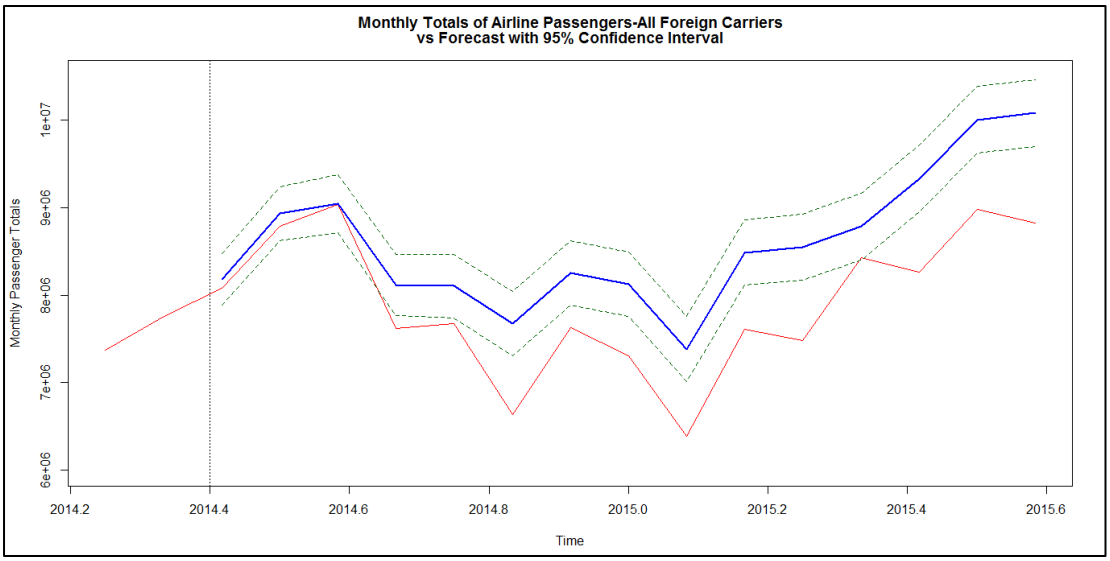

Fig 6.2: Close-up on forecast period

Our forecast performs moderately well for the initially 4 months of the test period, however, the predicted and actual results for the remaining months in the test period show increasing deviation. This indicates the presence of significant errors in the prediction of the later part of the test period. This is actually not surprising, since, in most forecasting situations, the further ahead we forecast, the more uncertain we are. This indicates that the forecasting accuracy decreases as we go farther from the initial points of testing data. The effect of training data decreases with increase in future prediction points. We further performed analysis of the errors present in the forecasts and the forecast accuracy measures are computed as shown in the table below.

Month	Actual Data	Forecast	Error	Percentage Error (%)
06/2014	8079300	8176056	-96756	-1.1976
07/2014	8783220	8930582	-147362	-1.6777
08/2014	9038255	9046370	-8115	-0.0898
09/2014	7619465	8113469	-494004	-6.4834
10/2014	7673417	8102393	-428976	-5.5904
11/2014	6638808	7672084	-1033276	-15.5642
12/2014	7626128	8253149	-627021	-8.222
01/2015	7308423	8123713	-815290	-11.1555
02/2015	6387635	7383198	-995563	-15.5858
03/2015	7608260	8486925	-878665	-11.5488
04/2015	7484515	8544834	-1060319	-14.1668
05/2015	8431816	8783959	-352143	-4.1764
06/2015	8260661	9333629	-1072968	-12.9889
07/2015	8978879	10004471	-1025592	-11.4223
08/2015	8823253	10081040	-1257787	-14.2554
			MAE = 686256	**MAPE = 8.9417 %**

Table 6.1: Forecast accuracy measures

The table above shows the error present between forecasted values & actual values of the number of passengers. The mean absolute error (MAE) and mean absolute percentage error (MAPE) were calculated as 686256 and 8.9417% respectively. We can see that the MAPE for first 4 predictions (2.3619 %) is comparatively less than the overall MAPE (8.9417 %), which means our initial forecasts have good accuracy. In order to obtain accurate forecasts for later months in the testing period we need to regularly update the model and determine updated forecasts. The farther we go with updating the forecast & adjusting the errors between previous predictions& actual data, the percentage error will decrease simultaneously. The more parameters we have for the future prediction, lesser will be the error rates for the future prediction.

Although we desire accurate forecasts, we wanted to prevent over-fitting the model to the training data as this may detrimentally affect the out-of-sample performance of the model. We expect the MAPE to decline if the training set and test set are increased and decreased respectively. The use of this model would likely be accompanied by monthly updates and long-term out-of-sample forecasts would not be required.

7. CONCLUSION

The results of deterministic part were fairly optimum and was consistent with our initial analysis. Starting with the modeling procedure, we extracted the deterministic part present in training data. We tried fitting various order polynomial trends. We have got 4^{th} order polynomial as the optimum fit using F test & residuals for selection. As there was visible periodicity present in the raw data, we tried to fit the seasonal trend for different periodicity. The outcome was "i=5", as the best fit based on the same sequence of testing over the residuals.

After the extraction of deterministic part of the data we did the ARMA modeling for the stochastic part. Following the standard ARMA modeling steps, comparing the various models with F test & dropping the parameters based on the provided conditions. We got ARMA (2, 2) as the best performing model. Now we should be left with the minimum possible residuals in the system.

After identifying the deterministic & stochastic part from the training data, we have done the joint optimization using "R studio" platform. The results which we have got after the joint optimization were very close to our initial estimates as shown in the table above. This indicates that the model is optimum and properly fitted.

The forecasting was done on the testing data with joint optimization approach. The forecasted values seem to be well within the range of actual data for 95% confidence interval. The significant deviation towards the end part of the testing data is because further forecasted values are not based on the updated previous values. The further we go in the future the less significance will be the impact of training data on the forecast. This can be improved by updating the forecasted data points.

This model can help determine the expected number of international passengers traveling with foreign carrier. This forecasted number can be used to target a certain population of passengers or travelling routes. Aviation companies can decide about the most popular routes and keep the best bets on that route to improve the revenues of company. This numbers can be useful for other distinct purposes benefiting a particular company or city or even a country as whole.

8. REFERENCES

1. **Data Selection**
   http://www.transtats.bts.gov/Data_Elements.aspx?Data=1 as on date 10/20/2015.

2. **RStudio**
   https://www.otexts.org

3. **Text book**
   S. M. Pandit and S. M. Wu, "Time Series and System Analysis with Application", J. Wiley & Sons, 1983 (Also, Robert Krieger Publishing Company, 1990, Reissued 1993, and Reissued 2001).