International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 3 (2019) pp. 646-650
© Research India Publications. http://www.ripublication.com

Air Passengers Occupancy Prediction Using Arima Model

Konda Himakireeti
Department of Information Technology
SreeNidhi Institute of Science and Technology, India.

Tammishetti Vishnu
Department of Information Technology
SreeNidhi Institute of Science and Technology, India.

Abstract

Predicting a number of passengers per trip is an important topic in airline business and travel economy which has spurred the interest of airline companies to develop better predictive models to accommodate in their business model. This paper presents extensive process of predicting the occupancy of the airline seats using the ARIMA model. Published airline passenger data is obtained from RPubs is used with predictive model developed. Results achieved convey that the autoregressive integrated moving average model has a strong potential for prediction and can compete with existing models for this business projects.

1. INTRODUCTION

Though we have sophisticated machine learning models to predict/ forecast a time sensitive / dynamic datasets, since the correlations between variables is time sensitive and the model has to train design algorithms based upon data relation form consecutive timestamp ARIMA model has an upper edge to handle time series datasets as it has a functionality of auto regression which most of the machine learning algorithms lack of.

1.1 Time Series:

A time series is a collection of numerical data points in successive order. A time series tracks the movement of the chosen data points, such as stock's price, over a specified period of time with data points recorded at regular intervals. An example of time series data is monthly electricity bill collected in chronological order over a year. This will be one-year monthly electricity bill time series. A time series data collected over same variable is univariate time series and a time series data collected over more than one variable is multivariate time series.

In general time series can be decomposed into 3 components:

1. Trend: A long term increase or decrease in data, it does not have to be linear.
2. Seasonal: It’s a time series effected by seasonal factors of time like a month in a year.
3. Cyclic: When data exhibits rises and falls that are not a fixed frequency.
4. Noise: An optional variability in the observation that cannot be explained by the model.

1.2 Time series analysis:

The main objective of time series analysis is to develop mathematical models that provide reasonable descriptions from given data. It can be used to know how the value of a given economic variable changes over time and also bemused to examine how the changes associated with the chosen data point compare to shifts in other variables over the same period of time.

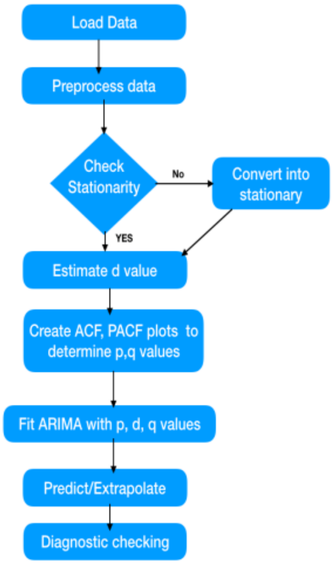


Figure 1: The process flowchart

1.3 Time series forecasting:

Forecasting involves taking mathematical models fit on sample data and using them to predict future. In statistical handling of time series data making predictions is called extrapolation.

2. STATIONARY AND NON-STATIONARY:

A time series is said to be stationary if it satisfies 3 conditions

1. Mean of the time series should be constant over time.
2. Variance of the time series should be constant over time.
3. Autocorrelation should not vary with time.

While a time series which does not follow any of these conditions then it is said to be non-stationary data. For a forecasting method to be applied on the time series data it should be stationary because of the reason that all forecasting methods are developed on the assumption that the given data is stationary because the results obtained by non-stationary time series may be spurious in that they may indicate a relationship between variables where one does not exist. So if we encounter a non-stationary data we need no convert into stationary data before applying the forecasting methods to predict the output.

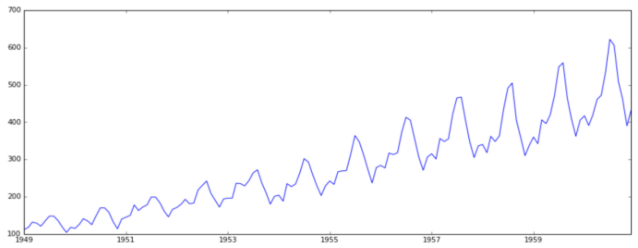


Figure 2. Visualizing given data

Some methods for checking stationarity of time series data like,

1. Look at plot: Plot a run series plot to see anything like trend or seasonal component.
2. Augmented Dicky-Fuller test(ADF): ADF tests the null hypothesis that a unit root is present in time series sample. ADF statistic is a negative number and lesser the value it is stronger the rejection of the hypothesis that there is a unit root.

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

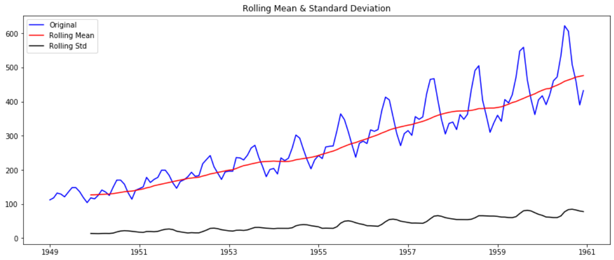


Figure 3. Rolling mean and standard deviation visualization

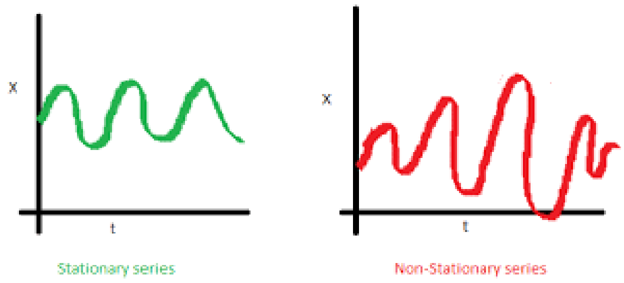


Figure 4. Examples of stationary and non-stationary series

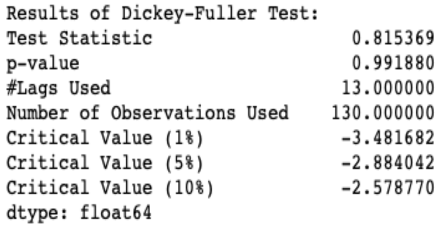
Null Hypothesis (H0): The time series has a unit root, meaning it is non-stationary. It has some time dependent structure.

Alternate Hypothesis (H1): The null hypothesis is rejected; this means the time series does not have a unit root, meaning it is stationary.

p-value > 0.05: Accept null hypothesis, indicating the time series data has a unit root and is non-stationary.

p-value ≤ 0.05: Reject null hypothesis, indicating the time series data does not have a unit root and is stationary.

p-value is associated with a test statistic. It is "the probability, if the test statistic really were distributed as it would be under the null hypothesis, of observing a test statistic [as extreme as, or more extreme than] the one actually observed.



```
Results of Dickey-Fuller Test:
Test Statistic                   0.815369
p-value                          0.991880
#Lags Used                      13.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
dtype: float64
```

So from both of the above results (figure 3 & ADF), we can say that data is non-stationary because the dicky fuller test satisfies the null hypothesis and from the plot we can see that there is an upward trend and mean is changing over time.

2.1 Converting Non-Stationary series in to Stationary series:

We can convert a non-stationary time series to stationary time series generally in two ways:

1. Differencing: Take the difference between the successive data points. For example if original time series was (x1, x2, x3….., x50) then our series with difference with degree 1 becomes (x2-x1, x3-x2, …… , x50 - x49) and this series consists of only 49 values.
2. Logarithmic Transformation: If you cannot make a time series stationary, you can try out transforming the variables. Logarithmic transform is probably the most commonly used transformation, if you are seeing a diverging time series. However, it is normally suggested that you use transformation only in case differencing doesn't work.

Here in this case (Figure 3), we can see that the there is a significant positive trend. So we can apply transformation which penalizes higher values more than smaller values. The data can be taking a log, square root, cube root, etc.

Let's take a **log transform** here for simplicity. After applying the log transform. Notice the Y-axis values they have been decreased.

We can see clearly that there is still some noise present in the data. So we can use some technique to estimate or model this trend and then remove it from the given series. There can be many ways of doing it and some of most commonly used are aggregation, smoothing, differencing etc.,

Now we need to eliminate the seasonality and trend component it can be done now by differencing,

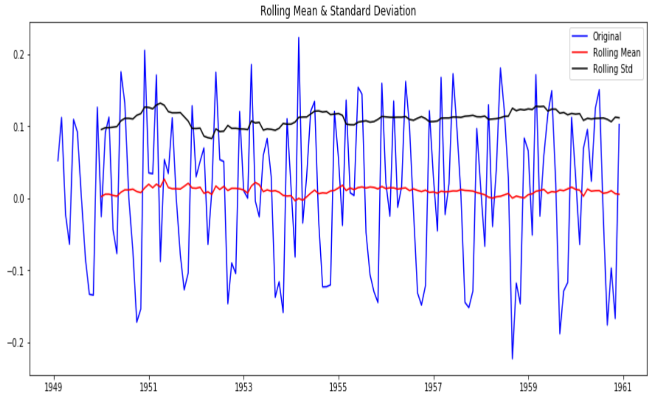
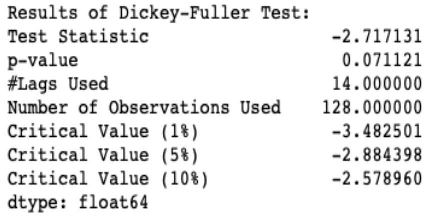


Figure 5. Rolling mean and standard deviation after applying logarithmic transformation

From figure 5, we can clearly see that after applying first order differencing the time series, the mean and standard deviation have small variations overtime.

Let's apply the dicky fuller test to the differenced data.



```
Results of Dickey-Fuller Test:
Test Statistic                  -2.717131
p-value                          0.071121
#Lags Used                      14.000000
Number of Observations Used    128.000000
Critical Value (1%)             -3.482501
Critical Value (5%)             -2.884398
Critical Value (10%)            -2.578960
dtype: float64
```

The test statistic value is less than the 10% critical value this means our data is stationary with 90% confidence. Further differencing might get even better results.

Forecasting:

As it said before forecasting can be done only on stationary data. Forecasting cannot be applied on random walk or on white noise.

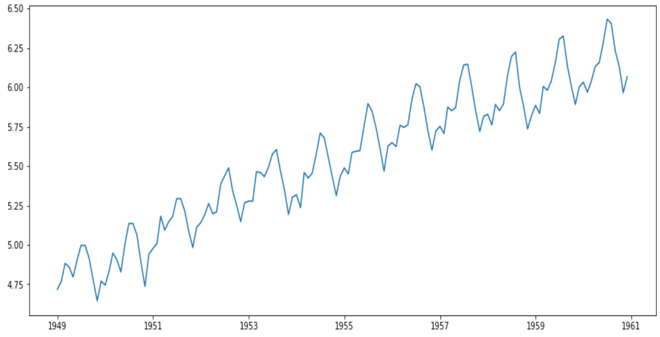


Figure 6. Data visualization after applying log transform

3. ACF AND PACF

3.1 Auto Correlation:

Correlation measures the extent of linear relationship between two different variables, whereas auto correlation is the measure of the linear relationship between lagged values of a time series variable. The auto correlation coefficient represented by r_k this defines the correlation between y_t and y_{t-k}.

$$r_k = \frac{\sum_{t=k+1}{}^{T}(y_t-\underline{y})(y_{t-k}-\underline{y})}{\sum_{t=1}{}^{T}(y_t-\underline{y})^2},$$

The auto correlation coefficients are plotted to show Auto Correlation Function (ACF). This plot is also known as Correlogram.

ACF of trended time series tend to have positive values that slowly decay as the lags increases. When data are seasonal, the autocorrelation will be larger for the seasonal lags than for alternative lags. When data are both trended and seasonal, you see a combination of these effects.

3.2 Partial Auto Correlation:

It is the simple correlation between the TS with a lagged version of itself after eliminating the variations explained by intervening lags. E.g. at lag 4, it will check the correlation by removing the effects already explained by lags 1 to 3.

$$\tilde{y}_t = \phi_{21}\tilde{y}_{t-1} + \phi_{22}\tilde{y}_{t-2} + e_t,$$

The partial auto correlation coefficients are plotted to show partial autocorrelation function (PACF).

4. FORECASTING MODELS

4.1 Auto Regressive model:

In an auto regression model, we forecast the variable using a linear combination of past values of the same variable. The term auto regression indicates that it's a regression of the variable against its previous values.

Thus, an autoregressive model of order p can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t.$$

Where ε_t is white noise and it is normally distributed, it is a regression with lagged values of y_t as predictors and ϕ_k is coefficient for k^{th} lagged value. We can refer this as AR(p) model.

4.2 Moving average model:

We use Moving average models if there are any random jumps in the time series data, these jumps are represented in the error that is calculated. A moving average model uses past forecast errors in a very regression-like model.

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}.$$

Where ε_t is white noise and it is normally distributed. We see this as an MA(q) model, a moving average model of order q.

Finding the order of AR and MA models:

After a time series has been stationarized by transformations, the next step in fitting an ARIMA model is to determine whether AR or MA terms are needed to correct any autocorrelation that remains in the differenced series. Of course, with some sophisticated software like Statgraphics, we can try some different combinations of terms and see which combination works best. But there is an additional systematic way to do this, by looking at the **autocorrelation function (ACF)** and **partial autocorrelation (PACF)** plots of the differenced series, you can tentatively identify the numbers of AR and/or MA terms that are required.

Consideration of model AR and/or MA model condition:

Model	ACF	PACF
AR(p)	Spikes decays towards zero	Spikes cutoff to zero
MA(q)	Spikes cutoff to zero	Spikes decays towards zero
ARMA(p, q)	Spikes decays towards zero	Spikes decays towards zero

The value of p will be equal to value of lag where the coefficient line touching the upper boundary of confidence interval in PACF plot, the value of q will be equal to the value flag where the coefficient line touching the upper boundary of confidence interval in ACF plot. The value of d will be equal to no of times the data is differenced to convert into stationary. In the plots the dotted lines on either side of 0 represents confidence interval. These can be used to find the values of p and q i.e, the order of AR and MA.

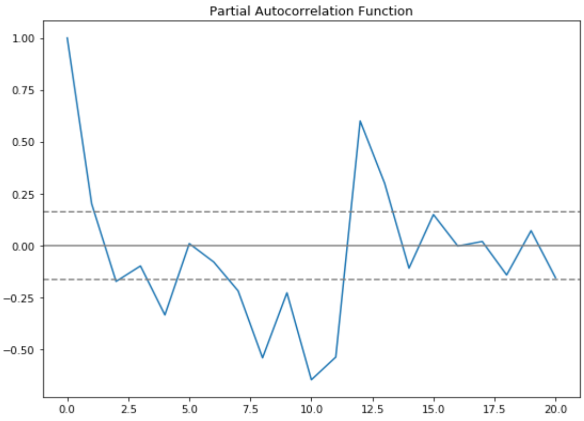


Figure 7. Partial auto correlation plot of transformed data

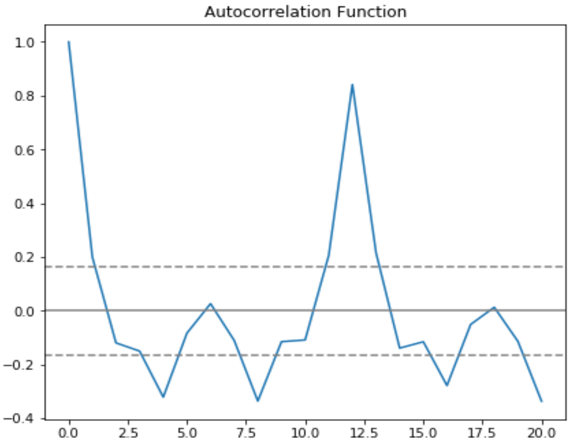


Figure 8. Auto correlation plot of transformed data

Here in our case we need to take both MA and AR into consideration because both ACF and PACF plots are decaying towards zero (no sudden drops are seen). The value of p will be equal to value of lag where the coefficient line touching the upper boundary of confidence interval in PACF plot, the value of q will be equal to the value flag where the coefficient line touching the upper boundary of confidence interval in ACF plot. The value of d will be equal to no of times the data is differenced to convert into stationary.

So our model is ARIMA(2,1,2)

Now we got the order of the model that means now we know how many time lags should be taken into consideration.

AR(2) : $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$

MA(2) : $y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$

Next situation is estimating the value of coefficients in AR and MA models this can be done by

1. Maximum Likelihood Estimation(MLE). Language R uses MLE for the estimation of parameters, this technique finds the values of the parameters which maximize the probability of obtaining the data that we have observed. For ARIMA models, MLE (maximum likelihood estimation) is similar to the least squares estimates that would be obtained by minimizing.

$$\sum_{t=1}{}^{T} \varepsilon_t^2.$$

2. Information criteria : Akaike's Information Criterion (AIC), which was useful in selecting predictors for regression, is also useful for determining the order of an ARIMA model.

$$AIC = -2log(L) + 2(p + q + k + 1),$$

For ARIMA models, the corrected AIC can be written as

$$AICc = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2},$$

A good model is obtained by minimizing AIC/AICc. In some cases we can also refer Bayesian Information criteria (BIC).

After fitting the ARIMA model and predictions are done on the training data set and the graph looks like,

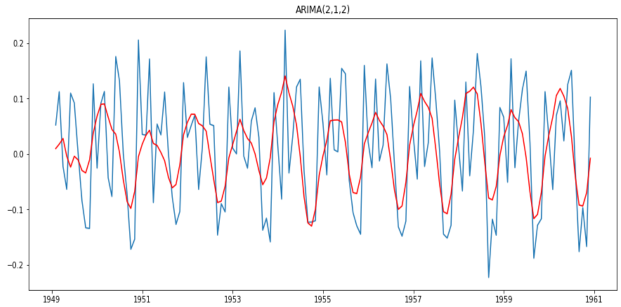


Figure 9. Prediction on train set

Given below are training errors with different parameters,

MODEL	RMS
ARIMA(2,1,0)	1.5023
ARIMA(0,1,2)	1.4721
ARIMA(2,1,2)	1.0292

5. PREDICTIONS

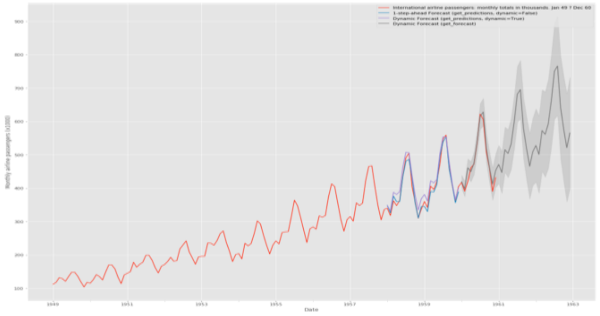


Figure 10. Extrapolating the time series

The grey shaded region in the Figure 10 marks the future predictions of the number of passengers.

CONCLUSION

This paper presents extensive process of building ARIMA model for predicting number of passengers per trip. The test results got with best ARIMA display showed the capability of ARIMA models to foresee on dynamic business problem. This could give airline business a profitable edge. With results obtained ARIMA models can compete reasonably well with emerging forecasting techniques in prediction.

REFERENCES

[1] https://otexts.org/

[2] https://people.duke.edu/~rnau/411diff.htm

[3] https://pdfs.semanticscholar.org/0f08/bcca67b3db328edfa5d3f48331dc71d8789e.pdf

[4] https://www.researchgate.net/publication/890956_Evaluating_Disease_Management_Program_Effectiveness_An_Introduction_to_Time-Series_Analysis

[5] https://sites.google.com/site/econometricsacademy/econometrics-models/time-series-arima-models

[6] http://www.ams.sunysb.edu/~zhu/ams586/UnitRoot_ADF.pdf