**A Time Series analysis to forecast the departure delay in US Airlines - R**

A report by Raja Sekhar Singh Bondili

As part of degree course project for MA 611- Time Series Analysis

**Bentley University McCallum Graduate School of Business**

**Abstract**

Flight delays affect passenger travel satisfaction and increase airline costs. In the year 2007 alone, airline delays cost the US Economy 31.2 Billion dollars. To better understand the trend for departure in different airline carriers we decided to explore deeper with a focus on their delays based on various time series forecasting models. Aviation daily data were used in the analysis and model development. By performing a cluster analysis of all the carriers, time series analysis was carried on for any particular trend or seasonality and later for forecasting future departure delays different forecasting models are studied and validated. Differential analysis in the time series prediction models for airline delay suggests variations in airline efficiencies though at the same airport.

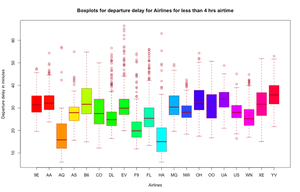
**Data**

The dataset chosen is from the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) which tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled and diverted flights appears in DOT's monthly Air Travel Consumer Report, published about 30 days after the month's end, as well as in summary tables posted on this website. BTS collects details on the causes of flight delays starting June 2003.

Summary statistics and raw data are made available to the public at the time the Air Travel Consumer Report is released. We had 1,936,758 observations and 29 variables. [<https://www.kaggle.com/giovamata/airlinedelaycauses>]

In the study, we will apply time series for model building and to discover the daily flying patterns for the carriers**.**

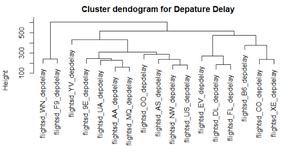
**Data Wrangling**



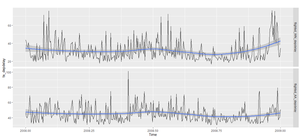
It was cleaned by removing the outliers for the variables airtime and departure delay. As few rows had zero airtime, does the removal takes into account that there is no mis-interpretation for any departure delay when the air time was zero, or the flight was cancelled. Also it removes bigger departure delay which accounts for cancellation of flights.

**Time series analysis:**

To study and build a common forecasting model to predict the departure delay, a cluster analysis was performed for all the 18 carriers departure delay time series objects using Euclidean distance to check for any similarity between their features.



By looking at the dendrogram, two different clusters can be observed. To analyze these two clusters, one airline from each cluster were selected. Hence the following model building and forecasting will be for the American airlines (AA) and Southwest Airlines (WN).



For the South west airlines, the time series seems more chaotic compared to American airlines. When the entropies are calculated for both the time series, the WN is less chaotic with an entropy value of 1.56 compared to 1.88 for American Airlines.

**Seasonality Study**

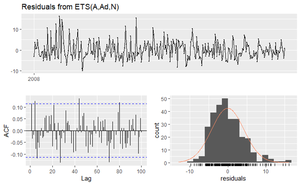
Inorder to check for any weekly seasonality, the polar plots and heat-maps were also analyzed for the two carriers. Looking at the polar map and heat map, we see no clear seasonality for American Airlines. The blue line on Tuesday might be an outlier.

**Model****Building**

For model building we have divided our data set into training and test set. Our training set consists of 300 days of data and remaining 65 days as test set. This is consistent across all models and for both departure delay <4 hrs and 4-10 hrs.

**Model 1: ETS Model**

The ETS model choose an ETS(AAdN) for both the airlines, showing the error is additive and trend is additive(damped) with no seasonality components. The residual test was performed using the Ljung-Box test. The model seemed to have failed the residual test with a p-value of 0.09. There is no particular pattern for the residuals and the data is almost normal.





**Model 2: ARIMA**

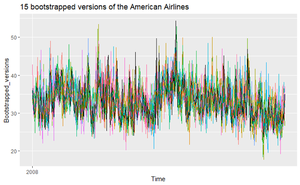
The best model selected was ARIMA (5,1,0) for flights with airtime less than 4 hours. The MAPE values fares better when compared to the ETS model for both the in-sample and out of sample data.



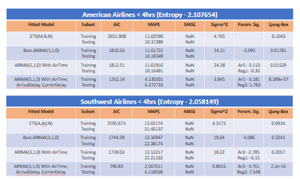
**Model 3: Bootstrapping**

As the ETS and ARIMA model did not pass some of the validity and utility test for all the time series objects, we further studied the bootstrapping model on the time timeseries objects. As bootstrapping model does not depend on the model and is more data dependent, hence bootstrapping would be a better choice to use for all the carriers instead of using individual models for each time series. We get the below graph, accuracy measures and prediction intervals.





**CONCLUSION**



From the above analysis, we were unable to find a model that would help in forecasting all 4 time-series (AA and WN with <4 and 4-10 hours airtime). To overcome this issue a single model we went ahead and constructed a bootstrapped model. There were 2 reasons to adopt a bootstrapped model:

1. We had 4 different models for 4 times series which made working with it very complex and each model did well on different aspects but not as a whole.
2. We were depending on the model for forecasting where as bootstrapping gives the data more flexibility to model itself.

Even though the accuracy measure(MAPE values) were slightly higher, it did well when compared to other ARIMA models. We can compromise on high MAPE values because of high entropy value and we are letting the data decides to forecast.

Depending on the question one wants to answer, we can modify the model and perform our analysis. If the airline would like to understand reasons for departure delay they may use regressors in their model. Through the bootstrapped model, we were able to address the departure delay question on a very general level.