# Applied Forecasting: Analysing Google Searches for “Hajj”

## Visualization Tools

## Applications of this Forecast

Hajj is the second largest gathering of Muslims annually, with over 1 million pilgrims partaking in this sacred ceremony. Due to this massive population influx into Saudi Arabia, enormous strain is placed on citizen’s safety over the condensed 5-day period. Stampedes, fires and collapsing buildings have claimed the lives of over 4,000 Muslims since 1987 (CNN, 2017). It was suggested in 2015 that the large population, confined in a cramped area, combined with the intense heat and urgency of pilgrims to complete rituals was the root cause of over 715 deaths (Yan, 2015). Over £200bn has been invested into redesigning the infrastructure of the Hajj to improve safety (Anne Templeton, 2015), but I believe the key to avoiding such tragedy’s is to focus on crowd psychology. Google Trends allows us to examine how many people are interested in the Hajj pilgrimage and their location. By forecasting this we can better predict crowd behaviour and hope to eliminate some security threats. Grouping pilgrims using a social identity approach, such as their country of birth, may form “psychological crowds” in which members feel safer due to a shared common self-definition (Anne Templeton, 2015).

This forecast would not only have social benefits, but economic benefits as airline schedules are modified to accommodate demand. THAI airlines announced 31 special flights for the Hajj period in 2017, which included specifically prepared Halal meals, and estimated to take 9,000 Thai Muslims to Saudi Arabia. (Thailand Travel News, 2017). Furthermore, Jeddah airport has a specific terminal dedicated to handle the large crowds (Harrison, 2004). In this way, this forecast has been used to improve quality of life and increase airline profits.

Relevant Graphs

My initial examination of plots produced from the Hajj timeseries has led me to 3 significant conclusions on the patterns occurring in the time series.

1. Seasonal with a period of 12 Months: The Autocorrelation Function, as shown in Appendix [1,1] shows large positive peaks at lags 12, 24 and 36. This implies that monthly figures are strongly correlated each year, i.e. that August 2017 is strongly correlated to both August 2016 and August 2015. Conversely, lags 3, 15 and 27 have negative ACF values as troughs ten to occur roughly 9 months before peaks. This is confirmed by the season plot [1.2] which shows significant increases from August through October annually. Again, the seasonal component of the decomposition plot [1.3] indicates that this series oscillates uniformly year on year.
2. Gradual increase in mean from 2008 to 2018: The trend component of the Decomposition plot [1.3] indicates that there is an increase in the number of people searching Hajj, signalling the series has a positive trend.
3. ­­Error Terms: The randomness illustrated in plot appears seasonal, and is normally distributed around a mean of zero, as illustrated by the histogram [1.4]. This means that all the data is explained by the trend and seasonality, and not random noise.

Possible Models

Holt Winters: Considering this series has both seasonality and trend, a SES or DES model would be unsuitable. Although the variance does not seem to be constant over time, it is not directly increasing with time. Therefore, I imagine a Holt Winters Additive Model would best forecast this data.

ARIMA: I predict a seasonal Arima model with s =12 would be most suitable. As previously discussed, the Hajj series contains trend which must be removed using differencing before implementing Arima Techniques.

Preparation of the Time Series

All data represented on Google Trends is previously adjusted to ease comparison. It is likely that there would be more searches on google in March as opposed to February due to the increased number of days. Google trends accounts for this by dividing the total searches by the time range it represents. Additionally, data points are divided by the total searches in a specific area to ensure places with the greatest search volume are not constantly ranked the highest (Google, 2018).

## Holt Winters Algorithm

The Hajj series does not become “wider” with an increase in time, as can be seen by the Time Series Plot [1.1]. This indicates that an additive model will suffice in forecasting our data. When examining the results of the decomposition function, the series was confirmed to be of type additive [2.1]. To satisfy my curiosity I fit both an additive and multiplicative model to the Hajj series, and found that the additive model was more appropriate as it had a lower SSE of 11082.23 compared the multiplicative SSE of 11804.05 [2.2].

Alpha, Beta and Gamma

The additive Model gave values of 0,0 and 1 for Alpha, Beta and Gamma respectively [2.3]. Smoothing is controlled by these three parameters which are estimates for the level, slope of trend component and slope of seasonal component at the current point in time (Coghlan, 2017).

The value of alpha (0) indicates that the forecasts are based solely on recent observations. The slope of the trend component of the series (0) is not updated over time, and is constantly set to the initial value. By examining the Hajj time series plot [1.1], it is clear the trend is increasing very marginally, and the slope of the trend is constant over time. Thus, it seems logical that beta would be 0, and no emphasis should be placed on recent changes in trend.

Conversely, the value of gamma (1) is very high, indicating that recent seasonal components have a large influence on the current seasonal component. When reviewing the Additive Seasonal Holt Winters forecast for the Hajj series versus the actual values [2.4], I noticed that the forecasted peaks were almost identical to the peak from the same period the previous year. In 2015 the Hajj series exhibits an unusually large peak, which is mirrored by the forecasted series in 2016. This is due to the large weight on more recent observations. It is also evident that the Holt-Winters exponential method is successful in predicting seasonal peaks, which occur roughly in July/August each year, due to high positive values of 11.35 and 38.22 for s7 and s8 respectively [2.3].

Forecasts

My next step was to forecast 2 years of the Hajj time series using this model, and the results are given in Appendix [2.5]. The level as predicted by the model is 25.78 [2.3], and adding that to s1 gives our first forecast, which is roughly 17 [2.3]. The area in pale blue shows the 95% confidence interval for forecasted values.

From the Histogram of residuals [2.6] after fitting this model we see that errors are roughly normally distributed around a mean of zero. This suggests that the Holt Winters exponential smoothing provides an adequate predictive model for the number of people searching Hajj each day. However, the time plot of residuals [2.7] does not indicate constant variance, indicating that the model could be improved upon, perhaps through a transformation of the original series.

The Holt Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong

Advantages to the Holt Winters Algorithm include simplicity, low data storage, easy automation and effective adaptation to changes in trends/seasonality. Holt Winters has been successful in extending to deal with issues such as the existence of outliers, multiple cycles and the need for prediction intervals. Sarah Gelper developed a mechanism that identifies outliers and replace them with more “moderate” figures that are still unusually large/small. The inclusion of such outliers, but on a less dramatic scale, led to increased accuracy in forecasts. Taylor extended the conventional HW to deal with double/triple cycles, using an additional smoothing equation and constant for each extra cycle, again increasing accuracy. Finally, Bermudez used a Bayesian framework to represent the uncertainty around soothing constants and initial values as probability distributions. Although complex, this method gives accurate results. Holt Winters remains impressive in it’s ability to produce reliable forecasts without sacrificing simplicity, and has stood the test time with it’s successful adaptation (Goodwin, 2010).

## Arima

Seasonal Arima can only deal with series that are stationary in variance. Before estimating a seasonal Arima model for the Hajj series, I fit an auto-Arima model and examined the residuals. As can be seen from figure [2.8], the plot became “wider” with an increased in time, indicating that this series is not stationary in variance. I applied various transformations including square root, cubed root, log and negative inverse, and found the negative inverse had the least variation as it was between 0.005 and 0.0012, and was therefore the appropriate transformation to use for Arima (See plots [2.09] to [2.12]).

Arima 1: (0,0,0)(0,0,0)[12]

This series is seasonal with a period of 12. I have commenced with a seasonal model of s=12 . Upon examining the residual time plot [3.1], I decided to introduce non-seasonal differencing to remove the downward trend present.

Arima 2: (0,1,0)(0,0,0)[12]

The residual time plot produced [3.2] no longer has an obvious trend. The ACF plot produced has large, positive peaks at s, 2s and 3s, thus I am fitting an SAR model next.

Arima 3: (0,1,0)(1,0,0)[12]

Peaks at 12, 24 and 36 are still quite large [3.3], indicating that the SAR model fitted is not sufficient. I increased the model from SAR(1) to SAR(2), and examine the resulting AIC’s.

Arima 4: (0,1,0)(2,0,0)[12]

As you can see from figure [3.4], there is no longer a large peak at lag 24, indicating that this SAR model is sufficient. The AIC has improved from -605. 39 to 613.415, further confirming that SAR(2) is a better fit. In the PACF there is an exponential decay at lag =0, and the ACF has a sharp peak at lag=12 before descending to zero. These are characteristics of a MA(1) model.

Arima 5: (0,1,1)(2,0,0)[12]

Again, we see a decrease in the AIC from -613 to -654 [3.7], indicating that an MA(1) model is appropriate. The ACF still peaks at 12, but now has another slight peak at 24, and there is evidence of exponential decay in the PACF [3.6]. I will increase order of q to 2, and examine the resulting AIC.

Arima 6: (0,1,2)(2,0,0)[12]

From both the ACF and PACF in figure [3.8] I cannot find any patterns that would lead me to believe another term should be added. Therefore, this is my final Arima model which gives me an AIC of -657.

Auto-Arima

R produced an Auto-Arima model of Arima(3,1,1)(2,0,0)[12]. The auto-Arima AIC and BIC are -654.67 and -635.22 respectively, both higher than the AIC and BIC from the Arima model above, which I will now be producing forecasts from [3.11].

Forecast

I Forecasted 2 years of the Hajj Negative Inverse Series, with the point estimates converted from negative Inverse to their original scale [3.12]. The forecast succeeds in encompassing the seasonality of the data, with peaks in roughly month 8/9. From plot [3.13] we see that the prediction intervals are rather wide, and the overall trend of the data is going down.

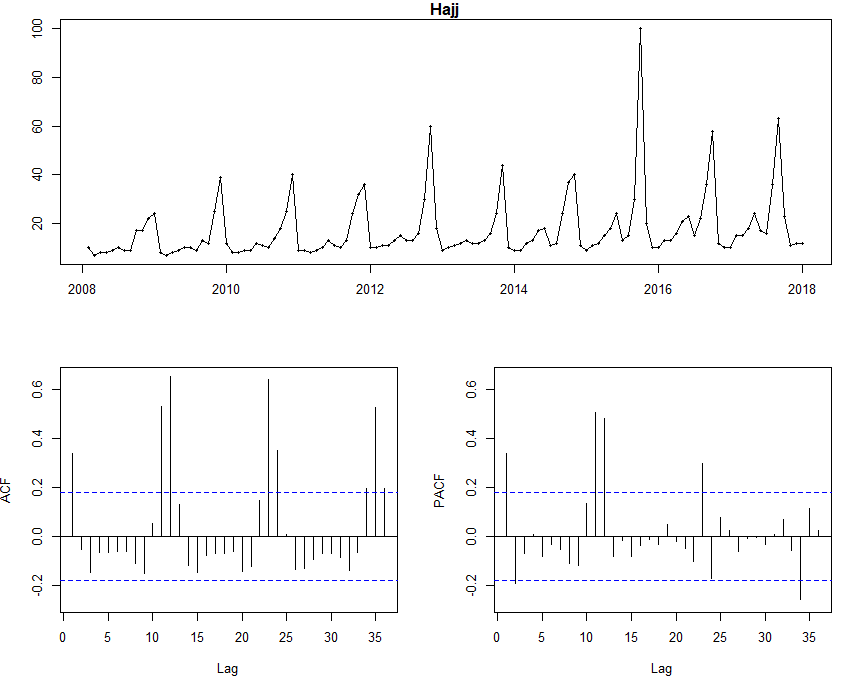
The histogram of residuals [3.10] is normally distributed and centred around a mean of zero, indicating this model is a good fit for the data.

Limit/Danger of Mathematical Techniques for Forecasting

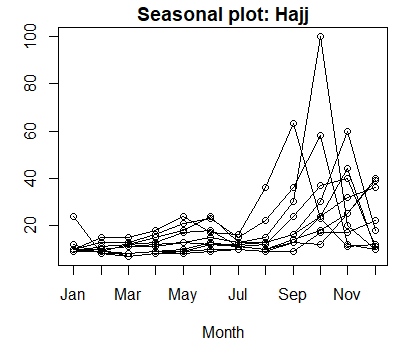
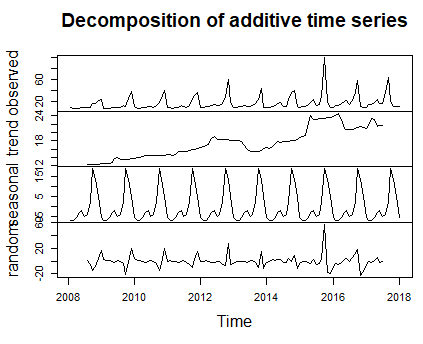
Mathematical forecasting is based on the continuity assumption that some aspect of the past will continue in the future. Forecasting is dependant on how well we understand contributing factors and the available data. Purely quantitative models eliminate the human element and focus solely on data, and if the forecast of interest depends on some categorical variable, it may be difficult to mathematically predict future outcomes. Mathematical models cannot predict unexpected or random events, as an expert might., and rely solely on old data and trends. Forecast’s cannot predict their own impact, a contributing factor, and could cause people to focus on historical data rather than current trends newly occurring (Beattie, n.d.).

# Appendix

### [1.1] TSDisplay Plot with Time Series (2008-2018), ACF and PACF



## [1.2] Hajj Season plot 2008-2018 [1.3] Decomposition Plot Hajj 2008-2018



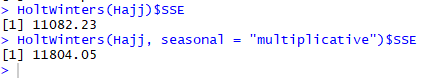
### [1.4] Histogram of Random Errors in Hajj series

### 

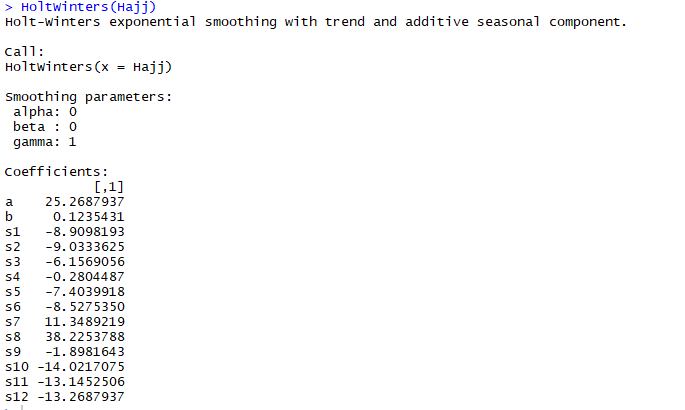
### [2.1] Decompose Function Type

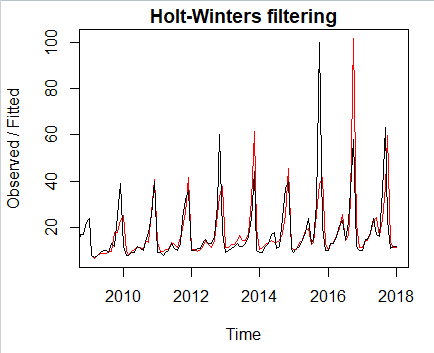


[2.2] Additive and Multiplicative SSE

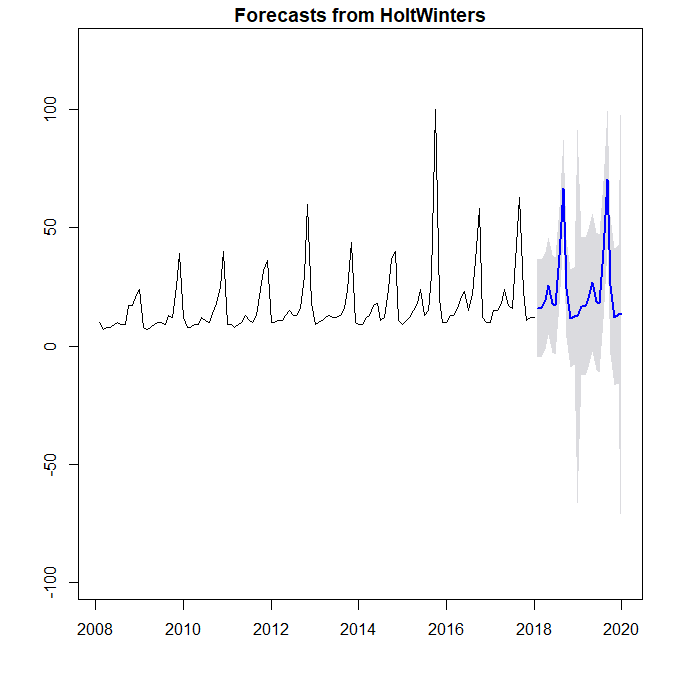


[2.3] Additive Holt Winters results

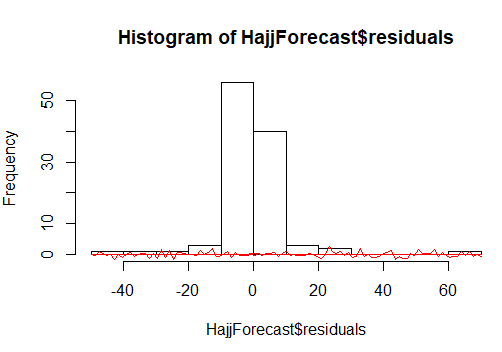


[2.4] Additive Holt Winters plot

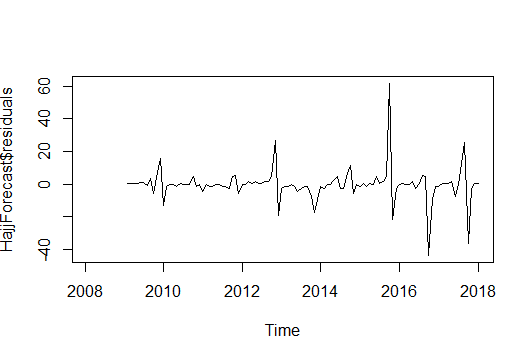
[2.5] Additive Holt Winters Forecast 2018-2020

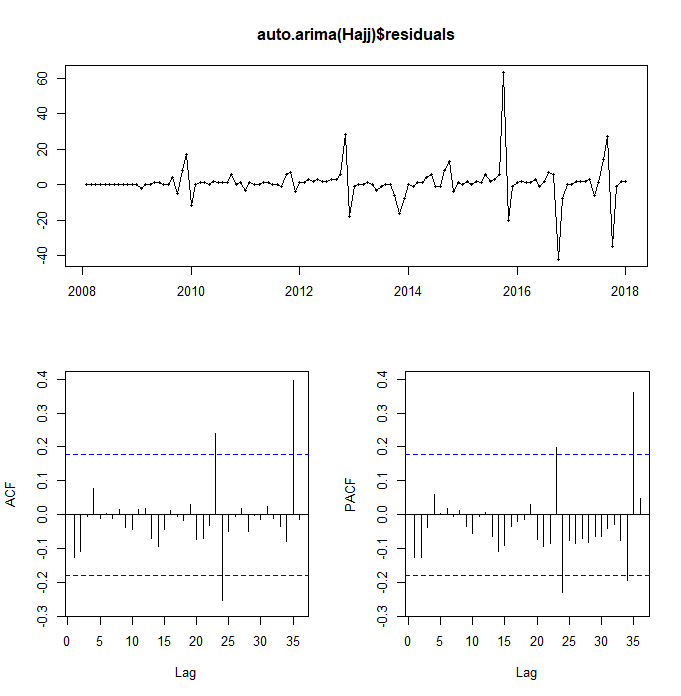


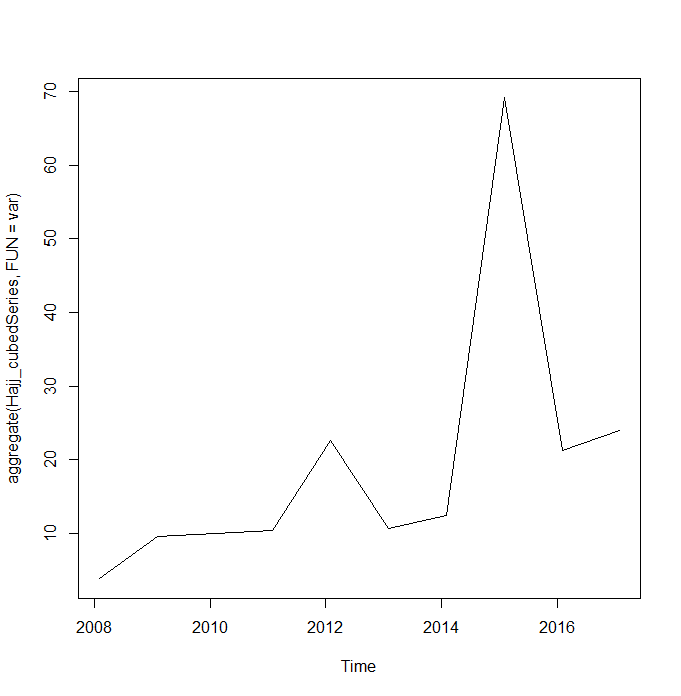
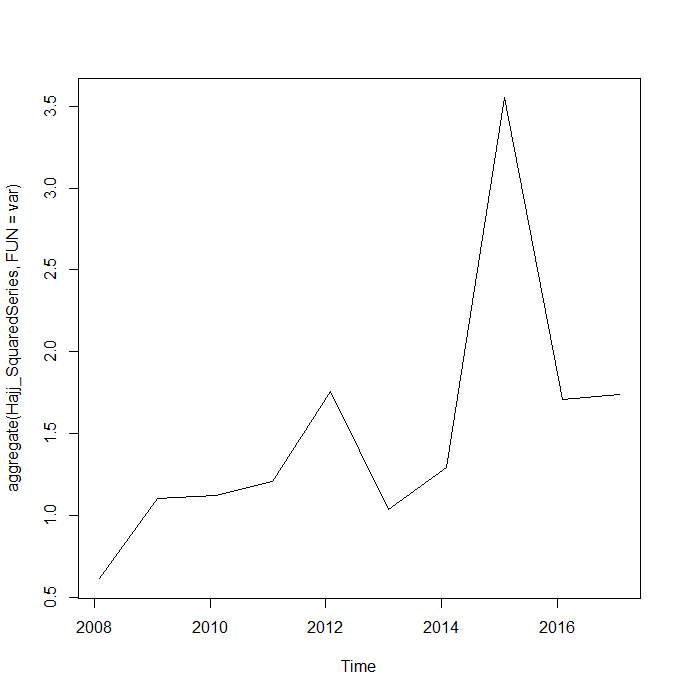
[2.6] Histogram of Residuals from Additive SHW Forecast

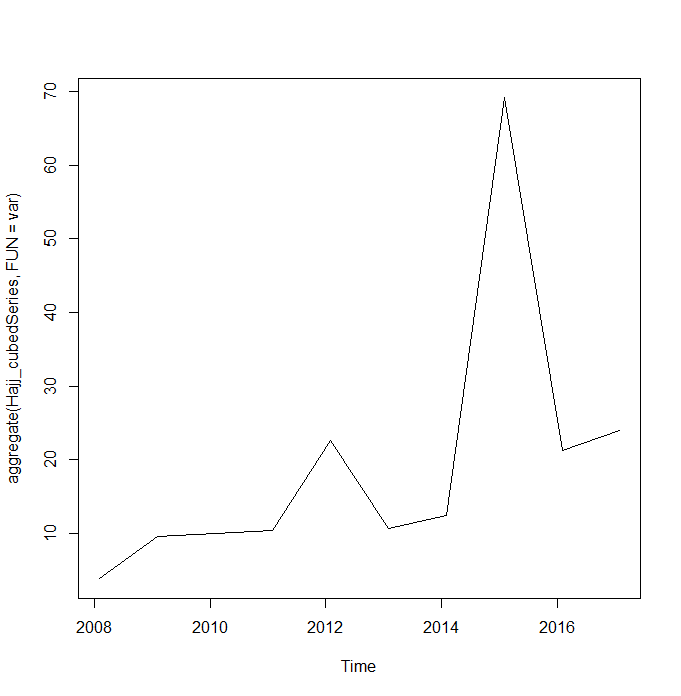
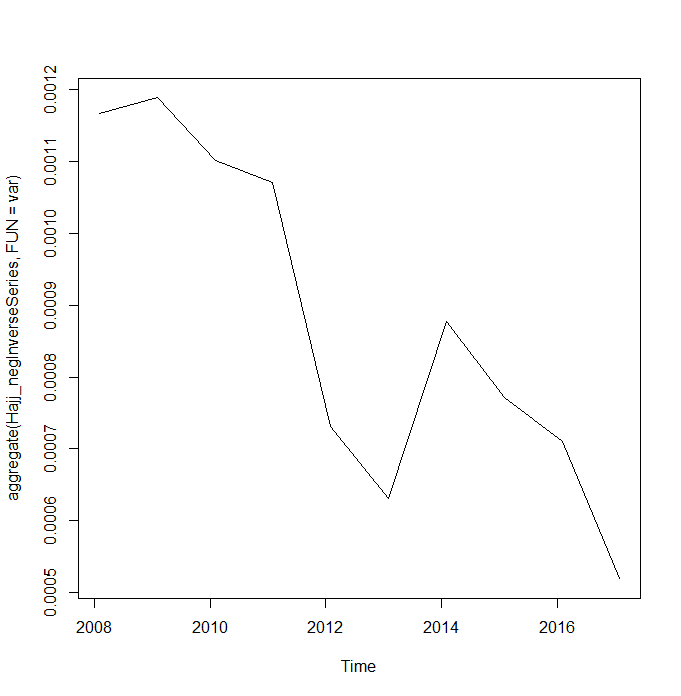


[2.7] Time plot of residuals

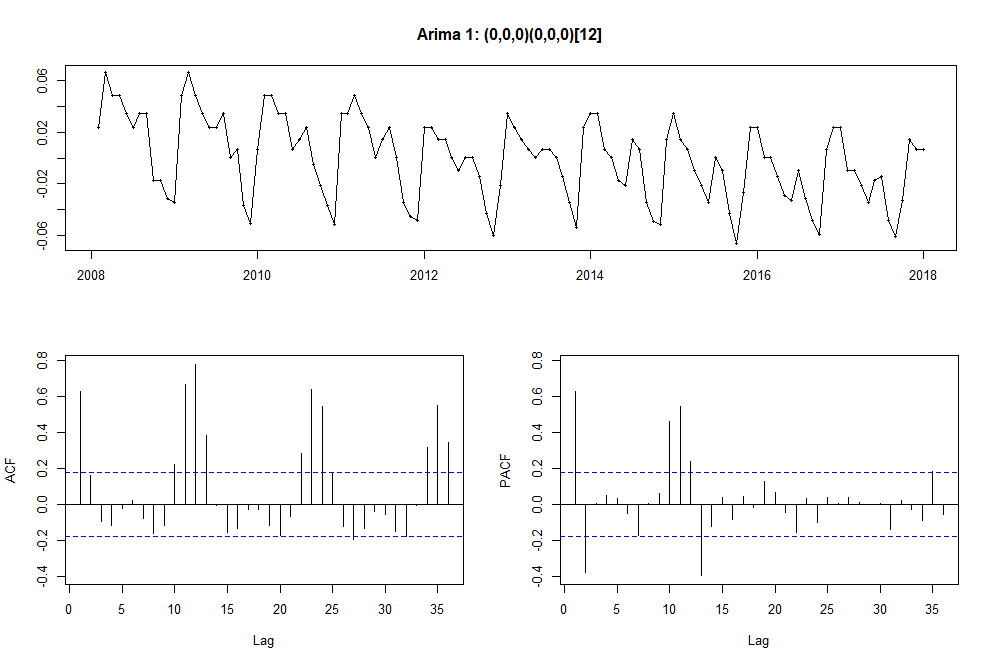


[2.8] Hajj auto-Arima residuals

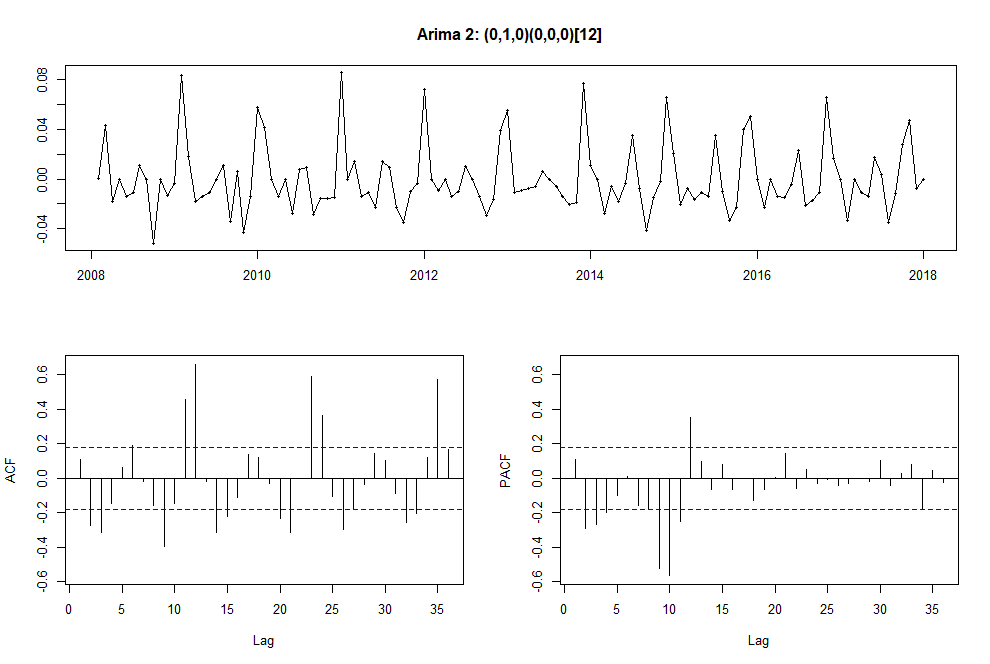
[2.9] Variance in Square Root Hajj Series [2.10] Variance in Cubed Root Hajj Series

[2.11] Variance in Log Hajj series [2.12] Variance in Negative Inverse Hajj Series

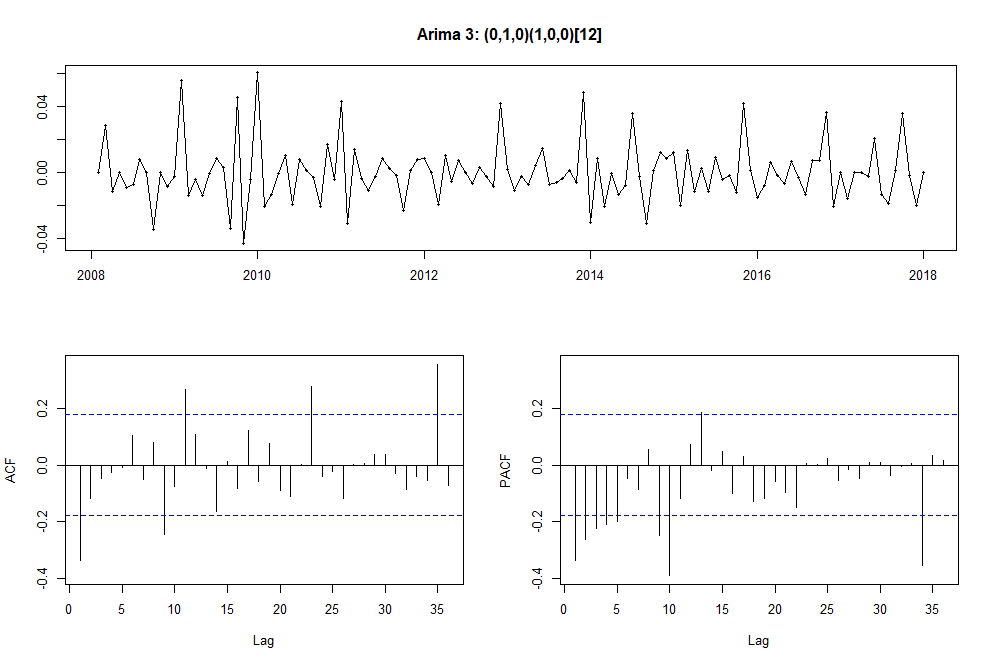
[3.1] Arima (0,0,0)(0,0,0)[12] Model



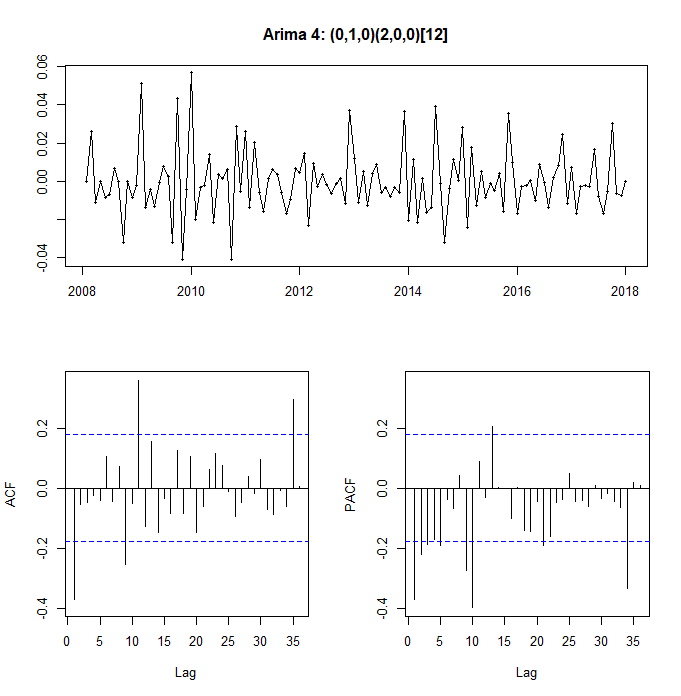
[3.2] Arima (0,1,0)(0,0,0)[12] Model



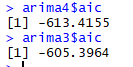
[3.3] Arima 3: (0,1,0)(1,0,0)[12]



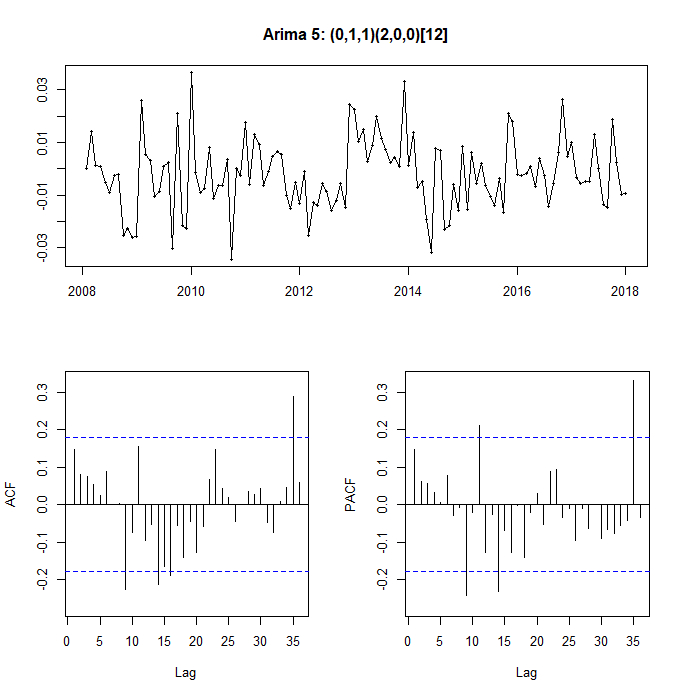
[3.4] Arima 4 : (0,1,0)(2,0,0)[12]



[3.5] AIC for arima4 and arima3

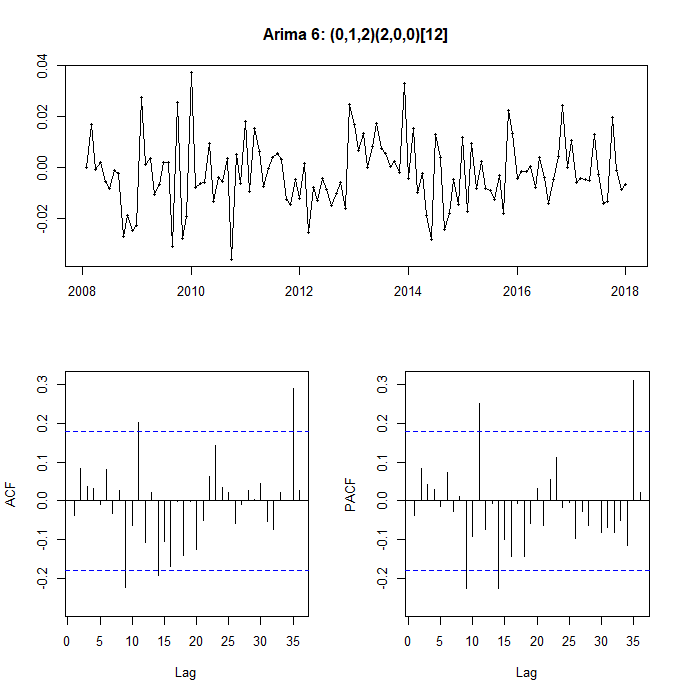


[3.6] Arima 5: (0,1,1)(2,0,0)[12]



[3.7] AIC Arima5

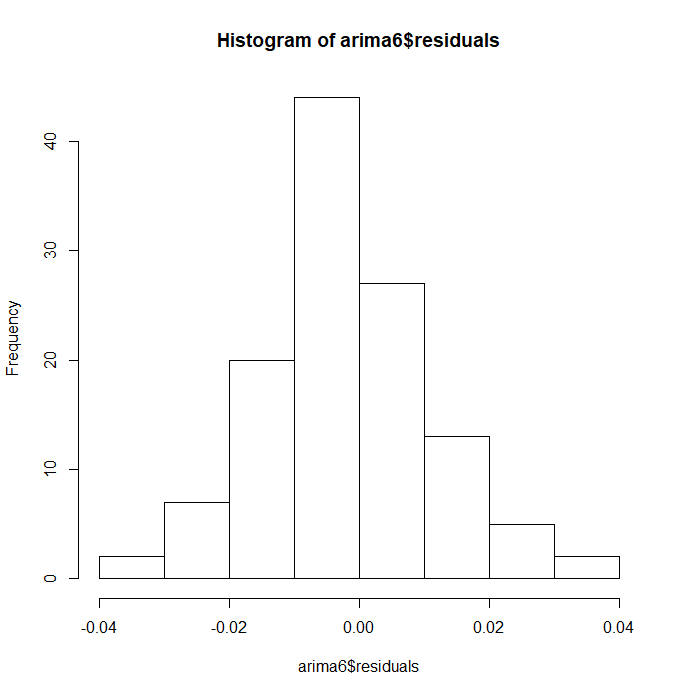


[3.8] Arima 6: (0,1,2)(2,0,0)[12]

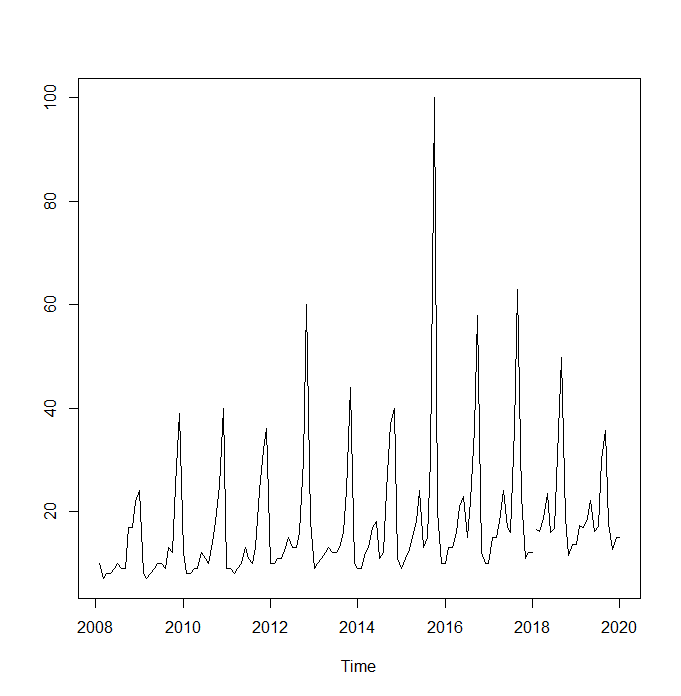
[3.9] Arima 6 AIC

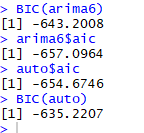


[3.10] Arima 6 Histogram of residuals

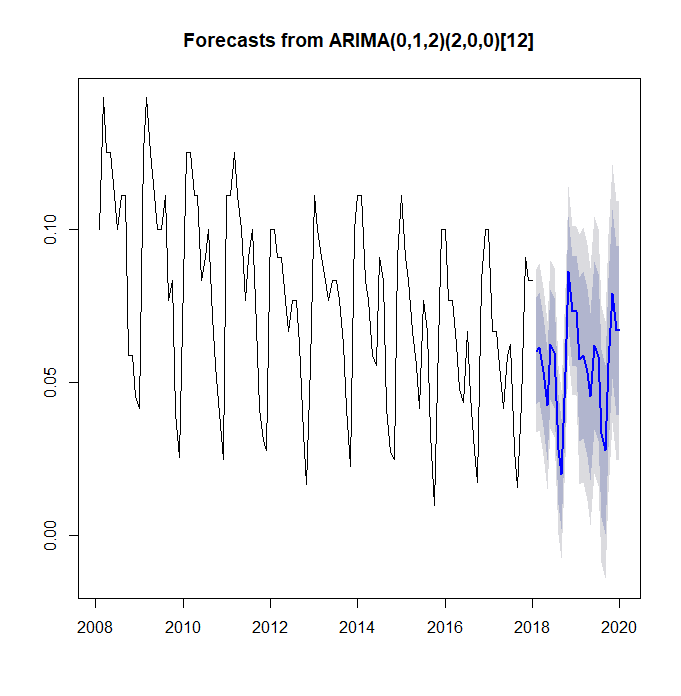


[3.11] Auto Arima vs Arima 6 AIC and BIC [3.12] Arima6 Forecast





[3.13] ARIMA (0,1,2)(2,0,0)[12] Forecast

­

# Bibliography

Anne Templeton, J. D. (2015, September 28). *Analysis: Here's how to make the Hajj safer – by better understanding crowd psychology*. Retrieved from University of Sussex: http://www.sussex.ac.uk/broadcast/read/32242

Beattie, A. (n.d.). *The Basics of Business Forecasting*. Retrieved from Investopedia: https://www.investopedia.com/articles/financial-theory/11/basics-business-forcasting.asp

CNN. (2017, November 22). *Hajj Pilgrimage Fast Facts*. Retrieved from CNN: http://edition.cnn.com/2013/06/21/world/hajj-fast-facts/index.html

Coghlan, A. (2017, June 15). A Little Book of R for Time Series Analysis. Cambridge, UK.

Goodwin, P. (2010). The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong.

Google. (2018). *How Trends Data is Adjusted*. Retrieved from https://support.google.com/trends/answer/4365533?hl=en

Harrison, D. (2004). In D. Harrison, *Tourism and the Less Developed World* (p. 156). Wallingford, UK: CABI Pub.

Thailand Travel News. (2017, July 26). *THAI arranges 31 special flights to Saudi Arabia for 2017 Hajj Pilgrimage*. Retrieved from TATNews: https://www.tatnews.org/thai-arranges-31-special-flights-saudi-arabia-2017-hajj-pilgrimage/

Yan, H. (2015, September 26). *717 people dead: What cause the Hajj stampede?* Retrieved from CNN: http://edition.cnn.com/2015/09/25/middleeast/hajj-pilgrimage-stampede/index.html