ST3010 Assignment

Ross Finnegan (15320532)
Trinity College Dublin, Ireland
8 January 2018

For this assignment, I have chosen to investigate the trends in the worldwide search of the Premier League, the most popular football league in the world. I have obtained this dataset from Google Trends. The league runs from August to May every year where the title is awarded to the best-performing team and the bottom three teams are relegated to the division below. My initial thoughts would therefore suggest a seasonal component present in the data. As well as that, the league has become a lot more popular around the world in recent years, particularly in Asia, which may lend to an upward trend in the number of searches.

I. Visualisation Tools

It is very useful for advertising companies and media outlets to predict future values for Premier League google searches as it indicates the fluctuating popularity of the league. To prepare the data, I used the time-series function to convert the data into a recognisable time-series object. This will allow me to visualise the Google Trends data and apply R's Holt Winters Functions and ARIMA models.

Figure 1 presents the time plot, ACF and PACF of time series Premier League. The time plot shows a significant increase over time in popularity as well as a sustained, unique seasonal pattern. Each year there is a drop in the number of searches during the months of June and July; this is not surprising as these are the two months of the year in which the premier league does not run.

In Figure 2, the decompose function shows the individual components of the time series. It is evident that there is an upward trend and a repeated seasonal pattern as well as noise. Further, Figure 3 demonstrates the extent of the seasonality by plotting each year against each other. There are a variety of peaks (April, August and December). Each of these are the major points of interest for the Premier League. April is when the championship is decided, August is when the new season begins and December is the busiest time of the year for the Premier League, with matches on St. Stephen's Day and New Year's Eve.

As there is both trend and seasonality present in my time series it is clear that the SES and DES algorithms will not be useful in this instance and a Seasonal Holt Winters Algorithm will be required.

The shape of my ACF and PACF, suggest that integrated differencing will be required when applying the ARIMA models.

II. Holt Winter Algorithms

As my time series contains a seasonal component, the Seasonal Holt Winters (SHW) Algorithms must be applied. In order to determine which algorithm is the best fit, I ran both the additive and multiplicative SHW and compared the Sum of Square Errors (SSE). As can be seen in Figure 4, the additive SHW returned SSE of 6227. The multiplicative SHW returned SSE of 4875. This is significantly lower so is the best choice of Holt Winters Algorithm for my time series. However, this result is quite a high SSE which suggests that perhaps the ARIMA models may be better suited for my time series.

Once this algorithm has been chosen the parameters (α, β, γ) have been set. This allows me to generate forecasts. Figure 5 shows the forecasts generated on R. This shows that the upward trend looks set to continue and for seasonality to remain prominent. This function also generates 80% and 95% prediction intervals (P.I.) which can be seen on Figure 5. The dark blue line represents the forecast. The darker shading represents the 80% P.I. and the lighter shading represents the 95% P.I. These intervals signify the potential range of actual observations at a given confidence level (80%/95%). The plot shows that as time goes on these intervals get wider. This makes sense as the further into the future you look, the harder it is to predict the true value with accuracy.

The article "The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong" written by P. Goodwin outlines the extensive uses of the Holt-Winters (HW) Algorithm, fifty years after its invention. Goodwin highlights how the HW Algorithm is a toolbox on which new adaptations can be developed. He outlined three recent issues that have been addressed by tweaking the HW Algorithm: Outlier Handling, Multiple Seasonal Cycles & Prediction Intervals. In these cases, researchers have modified the original HW Algorithm to solve the problem. Goodwin suggests that this is a testament to the method's reliability, simplicity and longevity.

III. ARIMA Models

The patterns shown in Figure 1 suggest that both seasonal and non-seasonal differencing is required for my time series. Therefore, I fit an ARIMA$(0,1,0)(0,1,0)_{12}$ and visualised the residuals as can be seen in Figure 6. This looks like quite a good fit however I looked to reduce the number of significant values in the ACF and PACF by fitting the ARIMA$(0,1,0)(1,1,1)_{12}$. A visualisation of these residuals

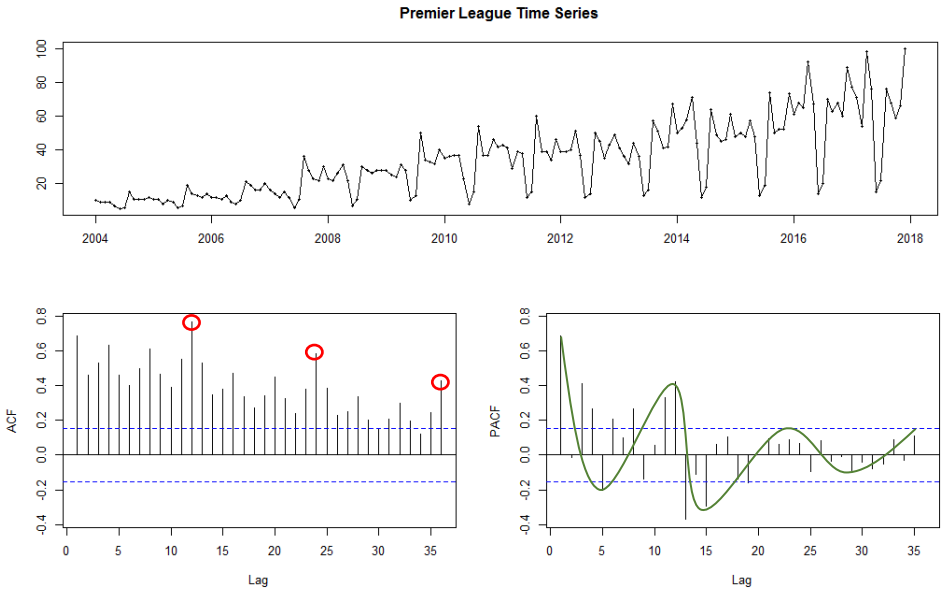
seen in Figure 7 show that this does improve the model. This has greatly reduced the number of significant ACF and PACF coefficients, therefore visualisation tools cannot be used further to improve the choice of the selected model. The AIC of this model is 1029.88. The auto.arima function identified the ARIMA$(0,1,0)(1,1,0)_{12}$ as the best model with a marginally improved AIC of 1000.00. The AIC is only one measure of goodness of fit and if we were to consider the Bayesian Information Criterion (BIC) both methods have BIC = 1012.20.

Similar to the forecast of my chosen Holt-Winters Algorithm, having decided on the ARIMA Model I wanted to use, I generated the plot of the forecast. As in Figure 5, Figure 8 shows the dark blue forecast curve as well as the shaded 80% and 95% Prediction Intervals. One point of interest when comparing the two forecasts is that Figure 8 (the ARIMA model) has far narrower prediction intervals. This suggests that the ARIMA model I have chosen is a much better fit than the Holt-Winters Algorithm as predicted accuracy is much greater.

Quantitative Forecasting must be considered with the potential errors in hypotheses and limitations on the techniques in mind. Often erroneous conclusions can be drawn from forecasts when considered in isolation. For example, although Google searches might be declining, this does not necessarily mean something is declining in popularity as there may be new avenues to access the site/product. Historical numerical data must be available but the most dangerous limitation is that the continuity assumption must hold. Should the contributing factors deviate from the historical behaviour, this can cause a serious error in future forecasts.

Appendix: Figures & Tables

Figure 1:



Timeplot of premier league: every year there is a similar pattern of peaks and troughs. Upward trend also observed.

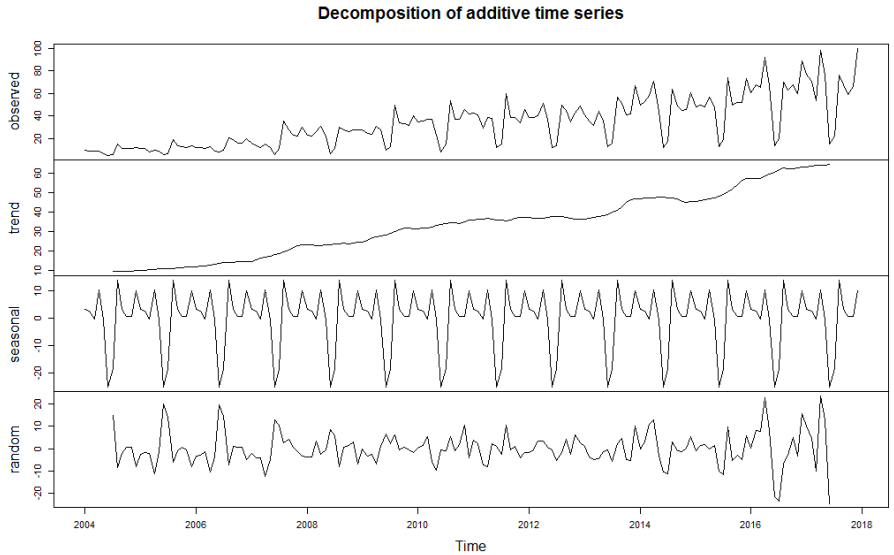
Seasonality period 12 (months)

ACF significant at all levels (differentiate) but spikes at 12,24,36.

PACF resembles loose sine wave shape.

Figure 1: visualisation of the time series premier league

Figure 2:



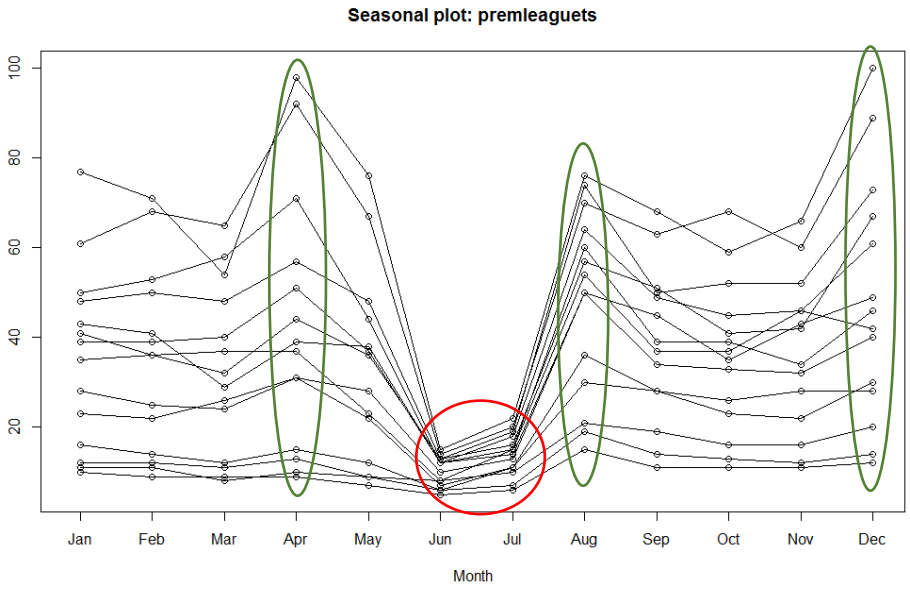
Upward Trend

Seasonal Pattern

Noise Present

Figure 2: Premier League time series broken into components: trend, seasonality, noise

Figure 3:



Peaks in April (End of Competition), August (Start of Competition), December (Busiest run of fixtures)

Troughs in June and July (Off season)

Figure 3: Premier League season plot

Figure 4:

```
> additive$SSE
[1] 6227.865
> mult$SSE
[1] 4875.124
```

Figure 4: R output comparing additive and multiplicative HW Algorithms

Figure 5:

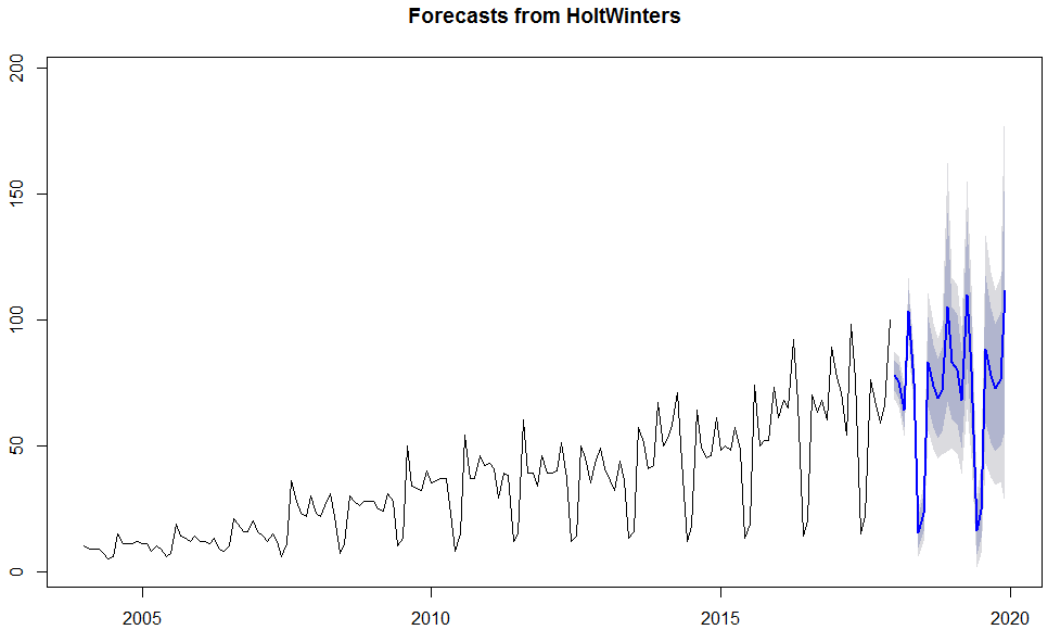


Figure 5: Computed Forecast using Multiplicative SHW. 80% and 95% prediction intervals present

Figure 6:

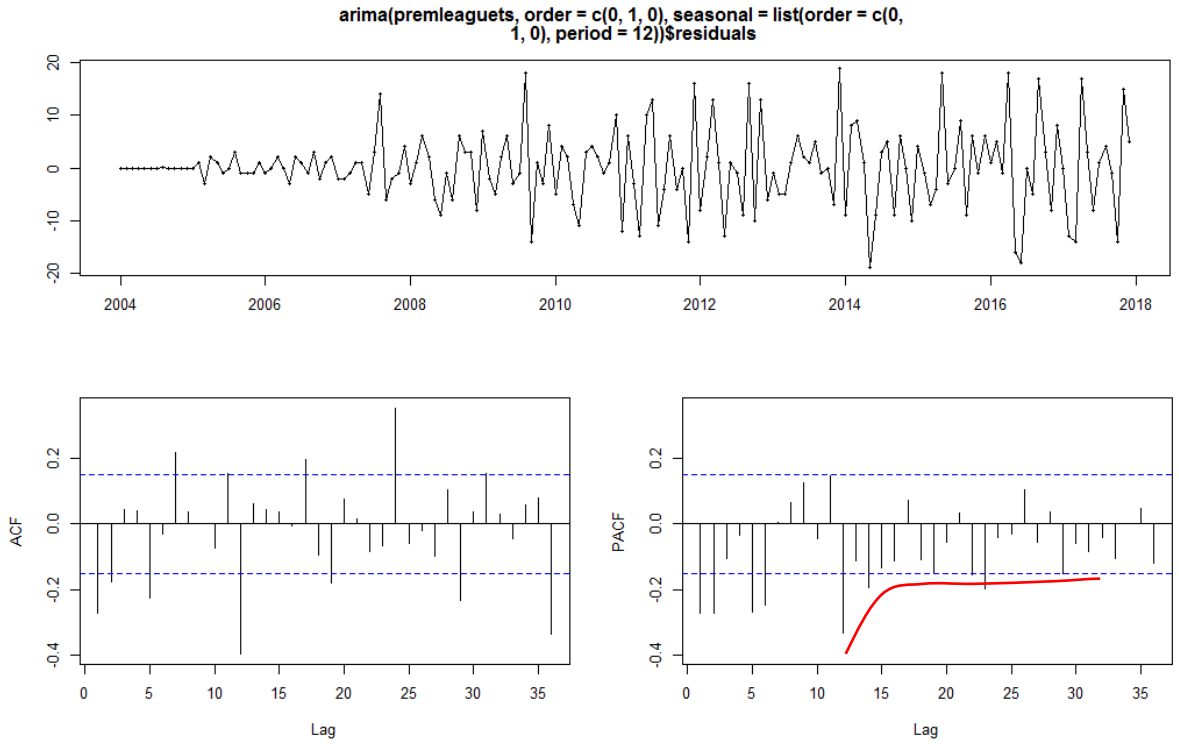


Figure 6: Visualisation of residuals after fitting ARIMA(0,1,0)(0,1,0)$_{12}$

Figure 7:

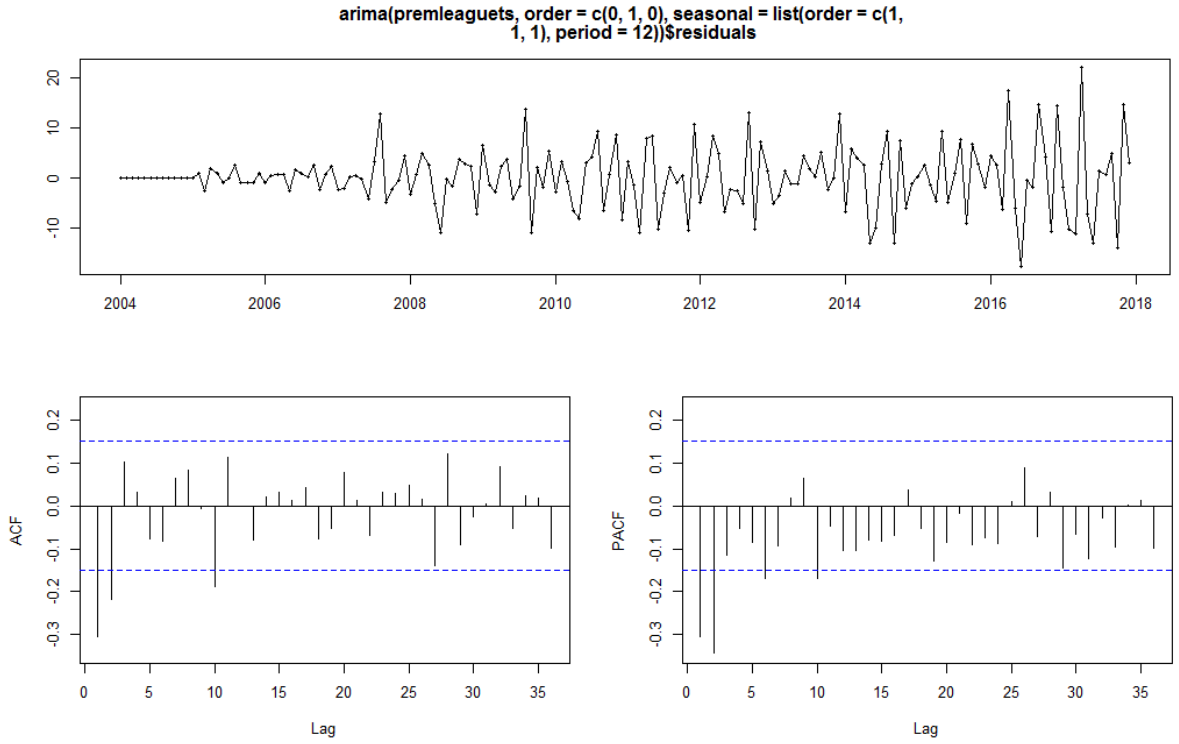


Figure 7: Visualisation of residuals after fitting ARIMA(0,1,0)(1,1,1)$_{12}$

Figure 8:

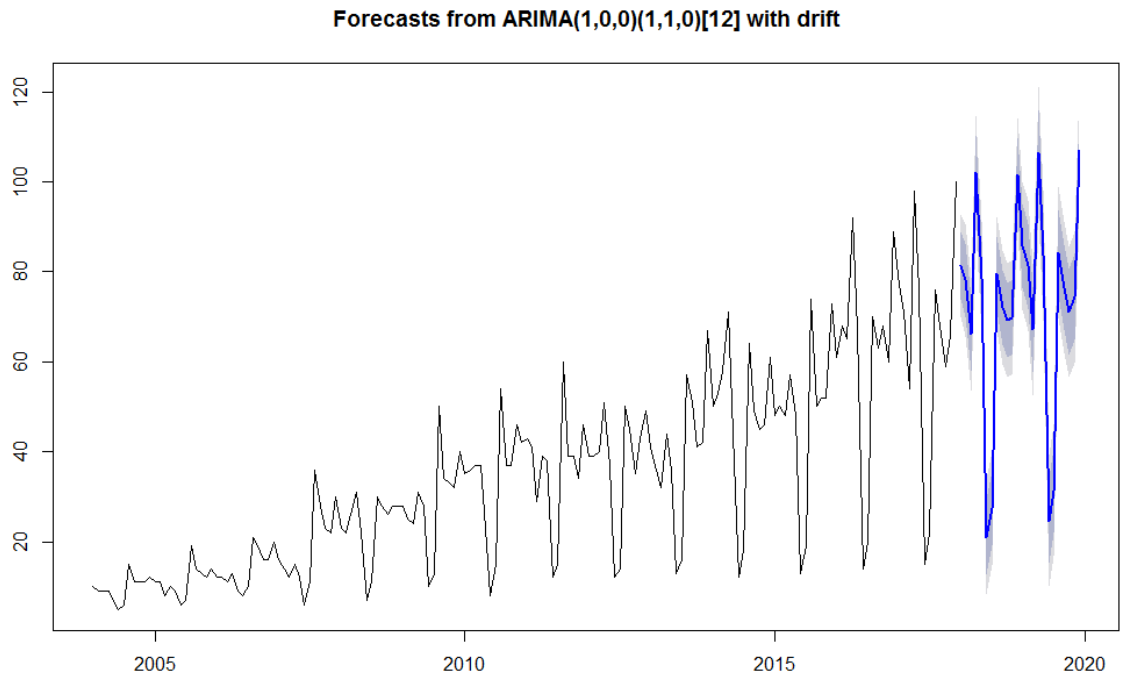


Figure 8: Computed Forecast using ARIMA(0,1,0)(1,1,0) $_{12}$. 80% and 95% prediction intervals present