MAS6019-200234131-TIME SERIES PROJECT

200234131

Contents

1. Introduction

A data set consists of daily maximum temperatures (degrees celsius) in Melbourne covering a period of 1 January 1981 to 31 December 1990. The data will be split into two sets, one for training the models and one for testing. The training dataset will contain the dates $01/01/1981-31/12/1988$ and the test dataset which will contain the dates $01/01/1989-31/12/1990$.

The training data will be used the build the model. The training dataset will be used to compare how accurate our models forecast into the year after our train data 1990. First, the data needs to be changed to monthly data rather than daily. This reduces the noise of the data and makes the models a lot easier to work with. The plots from the data will also be a lot clearer. This is done by taking the average temperature from each month. Fortunately, there are no missing values in the data. In regards to extreme values in the data, there were no observations which seemed to go against the trend and seasonality of the data. Because of this, all observations were included in the analysis of the time series.

Figure 1 shows the plot of the monthly data against the 12 month moving average. This shows us that the data is seasonal with highs in January (Australian Summer) and lows around July/August (Australian Winter). The 12 month moving average can be shown in red. It is shown that there are lows in the monthly moving average in the periods $1983-84$ and $1986-87$. This is also confirmed in Figure 2, where we see the decomposed data. The second graph shows the trend of the data and there are clear drops in the data in these periods.

2. Classical Decomposition

The data can be broken up into trend, seasonal/cyclic and residual/random. To estimate the trend component and seasonal component of a seasonal time series that can be described using an additive model, we can use the "decompose()" function in R. This function estimates the trend, seasonal, and irregular components of a time series that can be described using an additive model.

The trend shows how the maximum temperature changes over time disregarding the seasonality of the data. The seasonal/cyclic component shows the seasonal pattern of the data. The data that is left after the trend and seasonality components are removed is called the residual component and can be related to background noise. Decomposition is important because these components are worth studying on their own and if the models that are fitted perform poorly due to a trend or a seasonal component, we only need to rectify the problem in that specific component and not the whole model, which makes the problem a lot easier to solve. These components can be seen in Figure 2.

3. Stationarity

Stationarity is an important concept in time series analysis. Stationarity means that the statistical properties of a time series (or rather the process generating it) do not change over time. Stationarity is important because many useful analytical tools and statistical tests and models rely on it.

In Figure 3 it is shown that there is a clear year on year trend which shows a significant decrease in average maximum temperature from May to September. This strongly agrees with our claim that the data had a seasonality component. The variance and the mean value in July and August is much higher than the rest of the months. Hence, we have a strong seasonal effect with a cycle of 12 months or less.

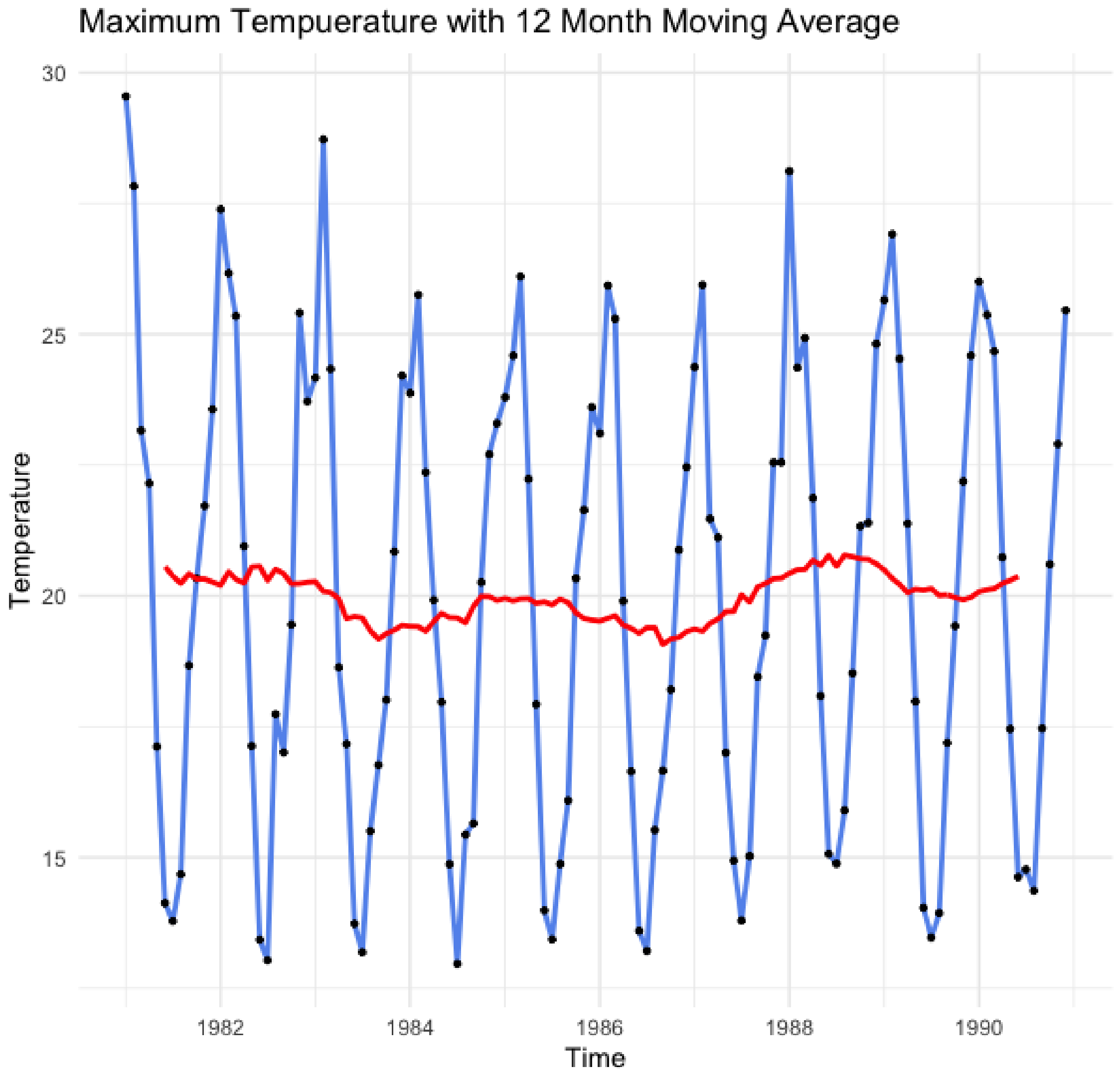

Figure 1. Temperature average throughout the months (green) with moving average (red)

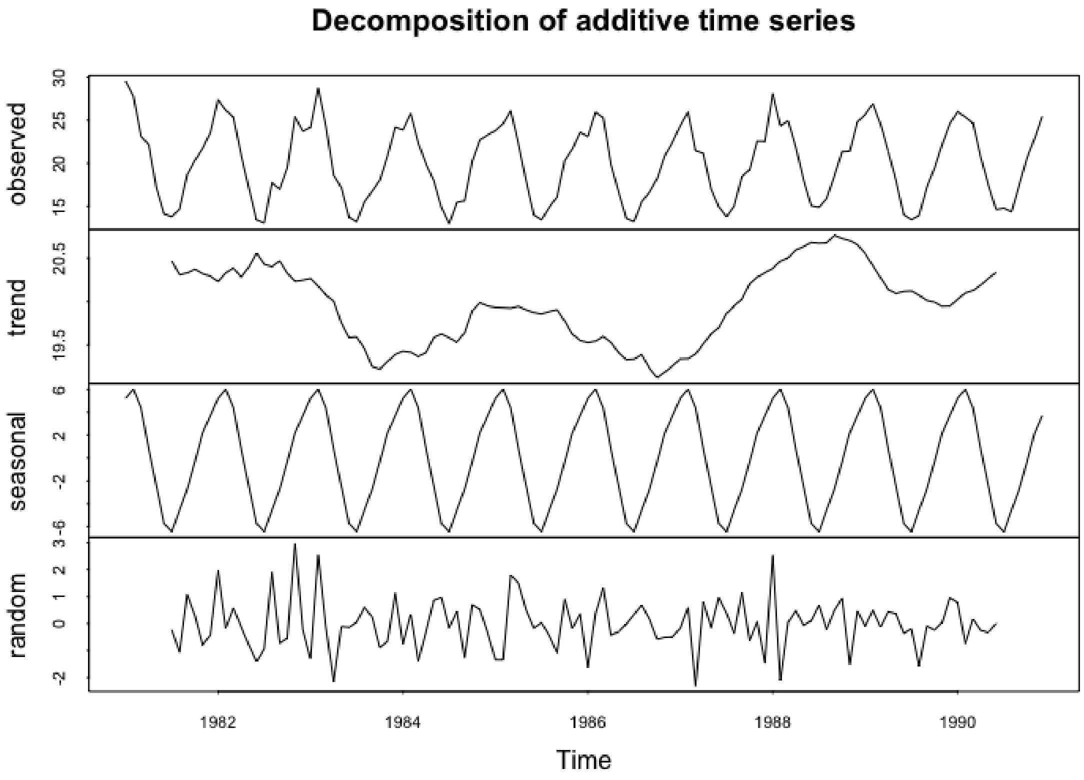

Figure 2. The classical decomposition of the data

You can see from Figure 4 that the seasonal variation has been removed from the seasonally adjusted time series. The seasonally adjusted time series now just contains the trend component and an irregular component. Once the data has been deseasonalised, the Augmented Dickey-Fuller (ADF) test is applied to test to assess if the data is stationary. The hypotheses under consideration are;

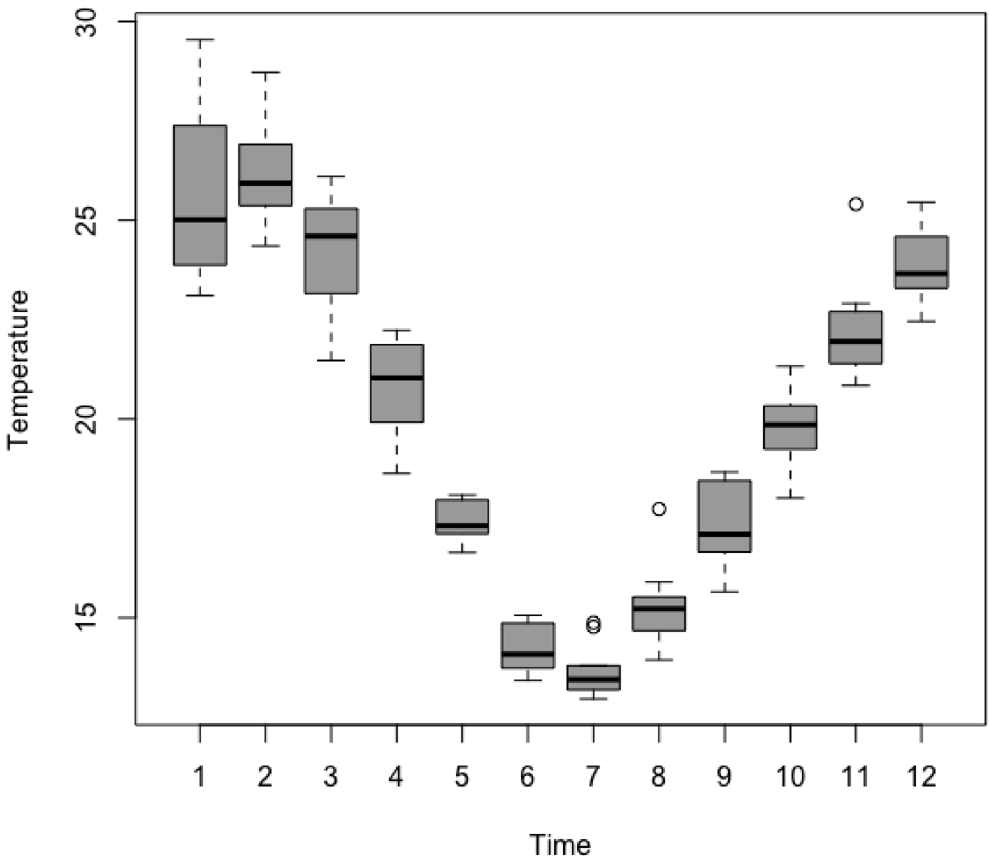

FIGURE 3. Box plot showing the seasonal component of the data

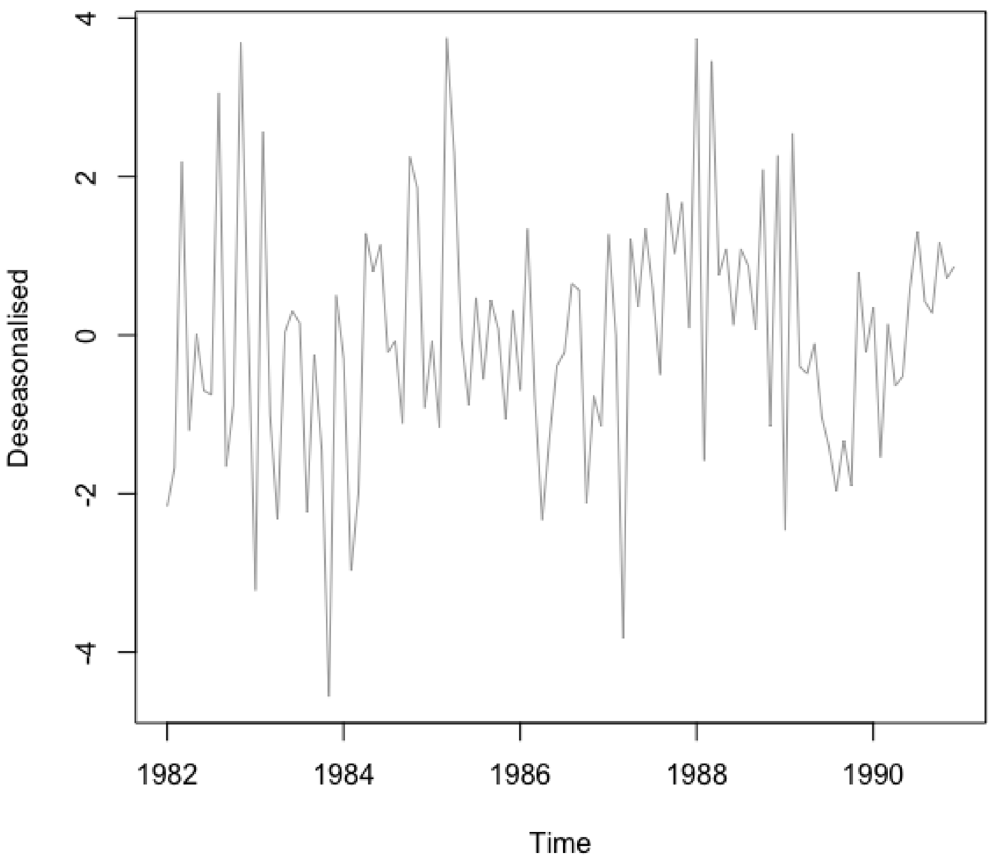

FIGURE 4. The classical decomposition of the data

H_0: The data is stationary vs $H1$: The data is explosive. The p-value threshold for decided if

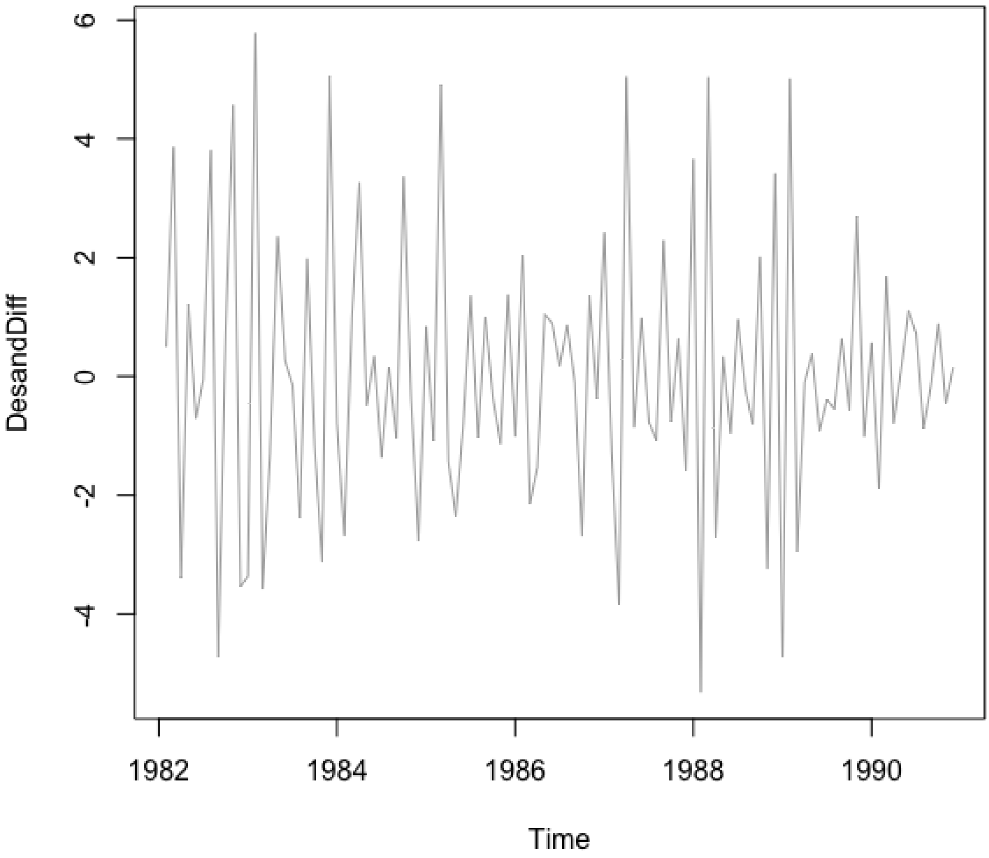

FIGURE 5. The classical decomposition of the data

the data is stationary is < 0.05. The p-value for the deseasonalised data is 0.1986 which suggests that the data is not stationary. Continuing, the data is differenced as seen in Figure 5. The ADF test is again applied and returns a p-value of < 0.01, which presents evidence in favour of the null hypothesis suggests that the data is now stationary enough to continue with the time series modelling.

4. Box Jenkins Method

In order to study and investigate which model will best fit the data, the Box-Jenkins method is applied. A Box-Jenkins approach uses differencing to ensure stationarity, which has been completed. Then running model diagnostics to ensure that the model fit is acceptable. The ACF and PACF plots give a deeper insight into which model is most likely to fit the data well.

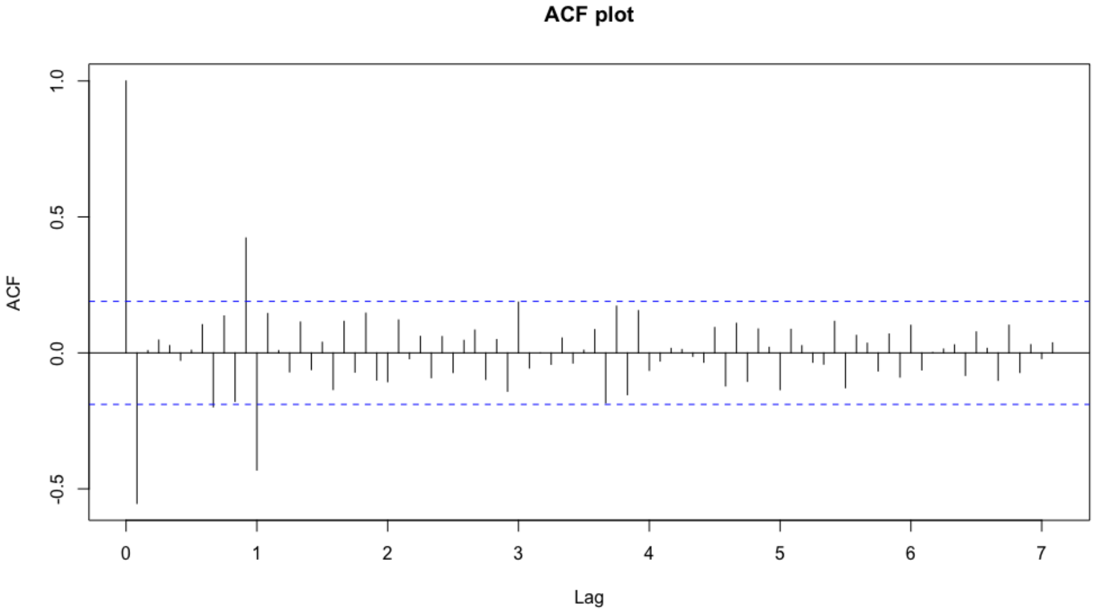

FIGURE 6. PACF plot of the deseaonalised and differenced data

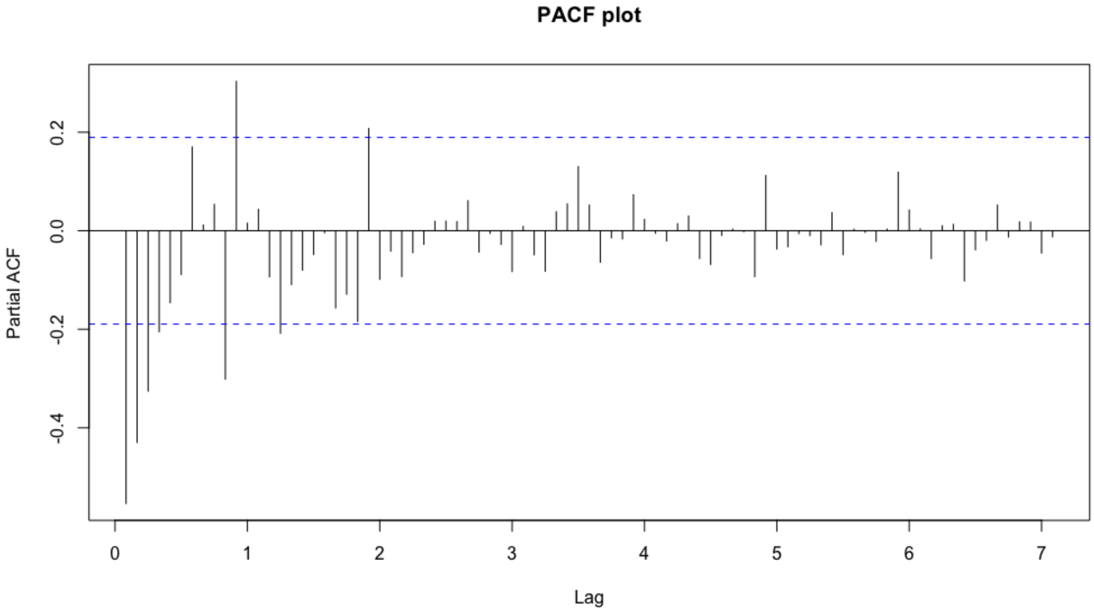

FIGURE 7. PACF plot of the deseaonalised and differenced data

In Figures 6 and 7, the plots indicate that if the PACF shows multiple significant lags and the ACF plot decreases after the first lag. The PACF plot must cut off at lag p and the ACF plot should decrease exponentially if an AR (autoregressive) model is to chosen to fit the data, this is not seen in these plots. Figures 6 and 7 also fail to meet the criteria for fitting an MA model as it is shown that the PACF plot does not decay exponentially.

Due to this, the ARIMA model is considered. The ACF tails off to zero after lag 1 and the PACF tails off to zero after lag 2, so the criteria for the ARIMA model is met and therefore the model is a suitable choice for this data.

5. Selecting the SARIMA models

If the time series is stationary, or if it has been transformed to a stationary time series by differencing d times, the next step is to select the appropriate SARIMA model, which means finding the values of most appropriate values of p and q for a SARIMA(p,d,q) model. To do this, we examine the correlogram and partial correlogram of the stationary time series. From these, it is seen that both the ACF and the PACF decay to zero.

In order to find the model that fit the training data most accurately, the auto SARIMA function was used. This function loops through a range of SARIMA models to get the best order and seasonal parameters. This function was run once based on BIC/AIC score and a second time without. The models that were chosen to be the most suited were SARIMA$(0, 1, 1)(0, 1, 1)[12]$ and SARIMA$(0, 1, 2)(0, 1, 1)[12]$ respectively.

6. Forecasting

The models that were selected are then used to forecast the years 1989 and 1990 and the results are then compared against the test data set. The residuals will give an indication into which of the four models had a higher accuracy of modelling the time series data. Figures 8 to 19 show the forecast, the ACF plot and the histogram of residuals for each fitted model. All of the forecast plots seem to capture the seasonality of the data as seen from the blue line. The grey area around the predicted values is the 95% confidence interval.

The ACF plot was plotted for each of the four models, none of which showed any significant spikes implying that the models were a good fit for the data.

Figure 9 shows that the distribution of forecast errors for the SARIMA$(1, 1, 1)(1, 1, 1)[12]$ model is roughly centred on zero, and is more or less normally distributed, although it seems to be slightly skewed to the left compared to a normal curve. However, the right skew is relatively small, and so it is plausible that the forecast errors are normally distributed with mean zero.

The model SARIMA$(0, 0, 0)(2, 1, 0)[12]$ had a large proportion of residuals at 0 and the histogram shows a normal distribution, suggesting that the distribution may have a mean of zero. The other three models seem to be left skewed which may indicate that these models are not as suited to the data (this is confirmed in the results).

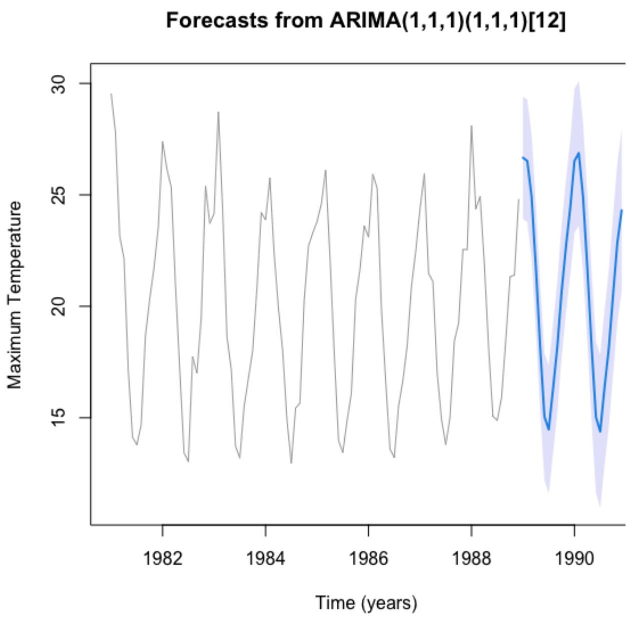

FIGURE 8. Forecast using the SARIMA(1, 1, 1)(1, 1, 1)[12] model.

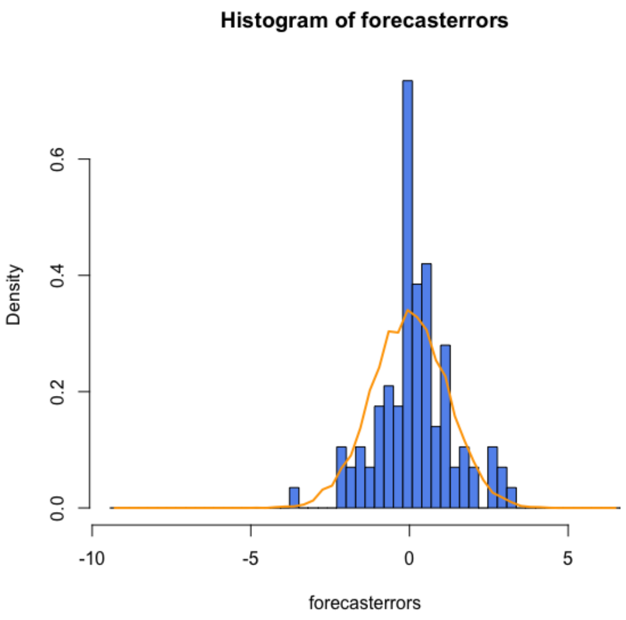

FIGURE 9. Forecast using the SARIMA(1, 1, 1)(1, 1, 1)[12] model.

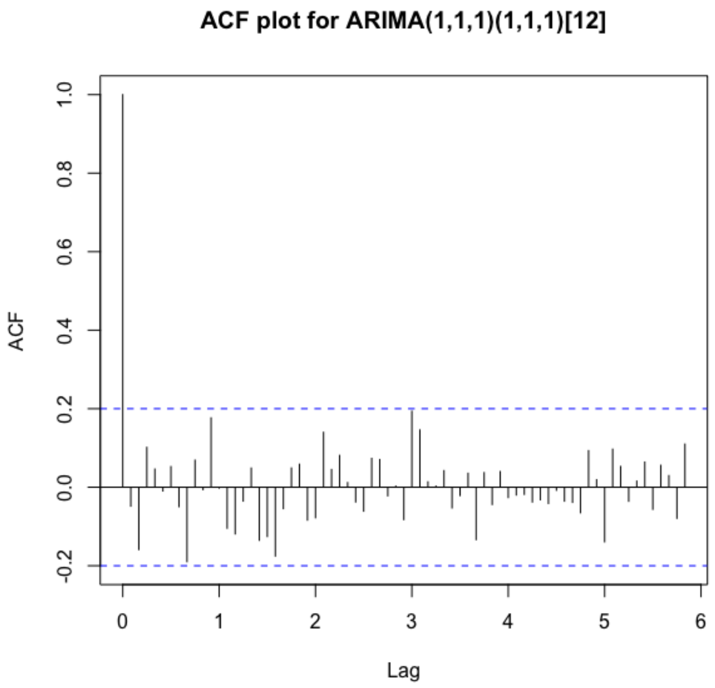

FIGURE 10. ACF plot of the residuals using the SARIMA(1, 1, 1)(1, 1, 1) model

7. RESULTS

The Ljung-Box test is applied to each of the four models and none of the results are significant enough to suggest a rejection of the null, that our model does not show lack of fit. It is seen in Table 1 that the model that has the best results in predicting the test data is model 2, SARIMA(0, 0, 0)(2, 1, 0) with an RMSE of 0.902. This suggests that trend parameters > 0 model the data not as accurately. In Figure 2 it is seen that there is no real pattern to the trend and so agrees with this statement. The two auto Arima had a larger RMSE than model 2. This could be due to what the functions classes as the 'best' model and if the best model is just based on AIC

Model	**ME**	**RMSE**	**MAE**	**ACF1**	**Box-Ljung**
sarima(1,1,1)(1,1,1)	-0.702	0.992	0.800	0.243	0.412
sarima(0,0,0)(2,1,0)	0.096	0.902	0.738	0.189	0.615
sarima(0,1,1)(0,1,1)	-0.590	0.919	0.727	0.339	0.282
sarima(0,1,2)(0,1,1)	-0.858	1.146	0.951	0.413	0.435

TABLE 1. Model results

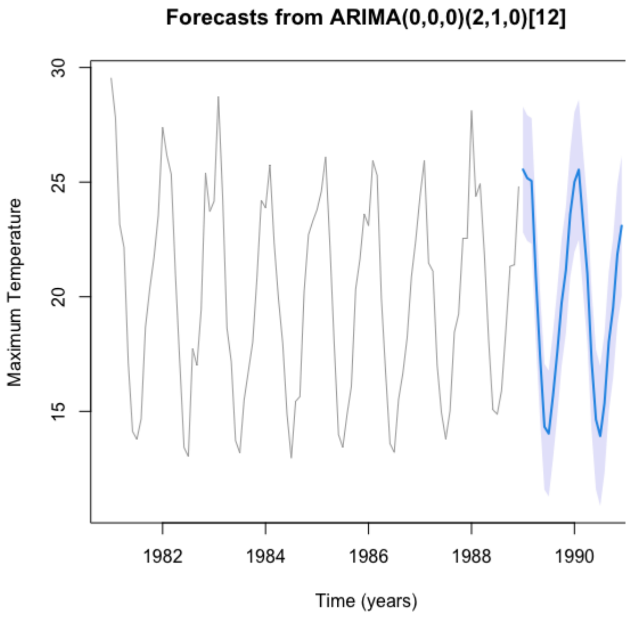

FIGURE 11. Forecast using the SARIMA$(0,0,0)(2,1,0)[12]$ model.

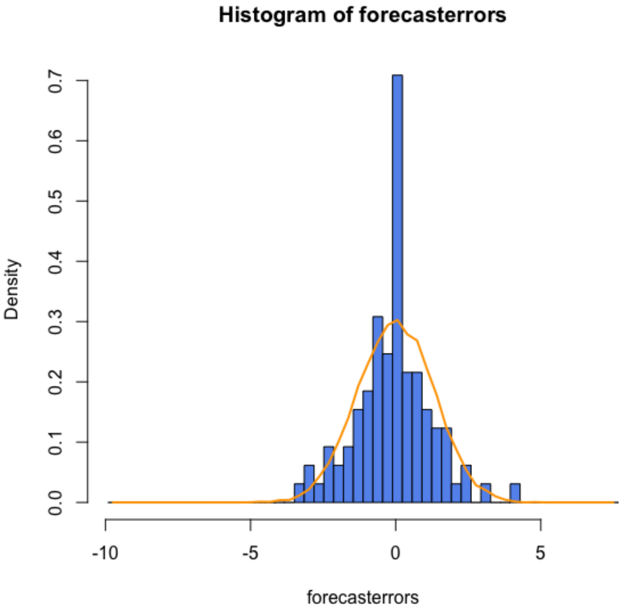

FIGURE 12. Forecast residuals using the SARIMA$(0,0,0)(2,1,0)[12]$ model.

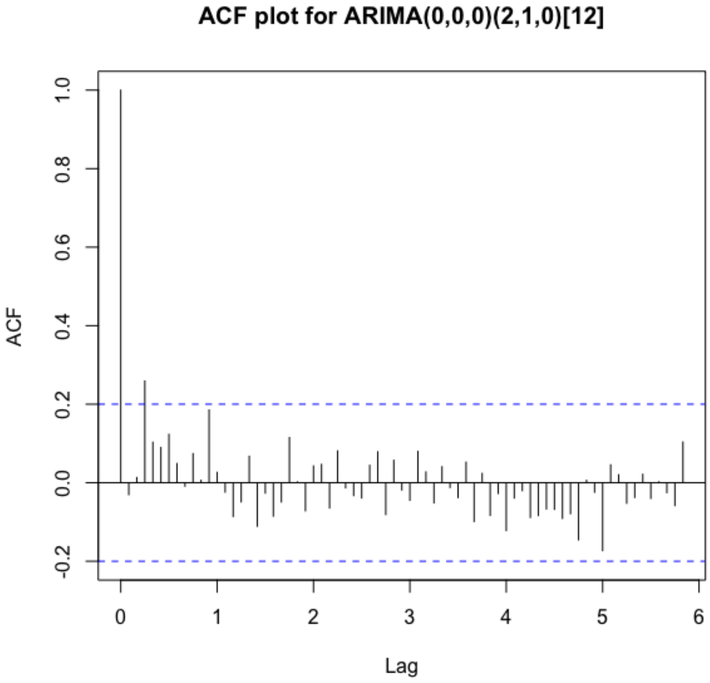

FIGURE 13. ACF plot of the residuals using the SARIMA$(0,0,0)(2,1,0)$ model

and BIC. In order to achieve better results, more data from Melbourne could be used when fitting the models again.

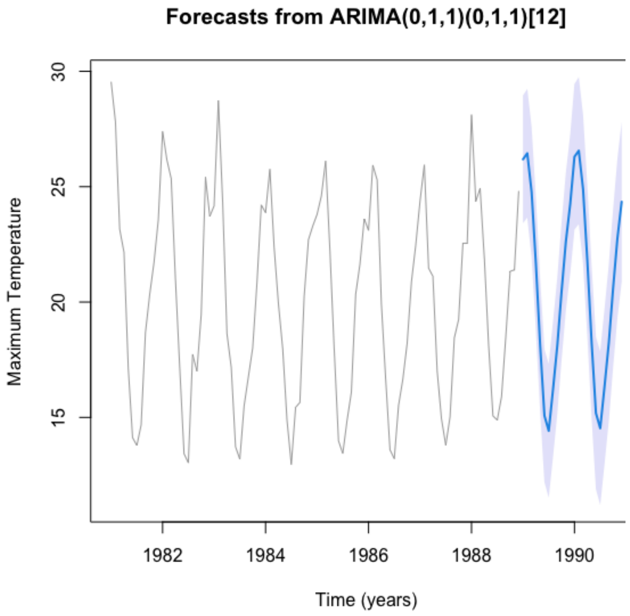

FIGURE 14. Forecast using the SARIMA$(0,1,1)(0,1,1)[12]$ model.

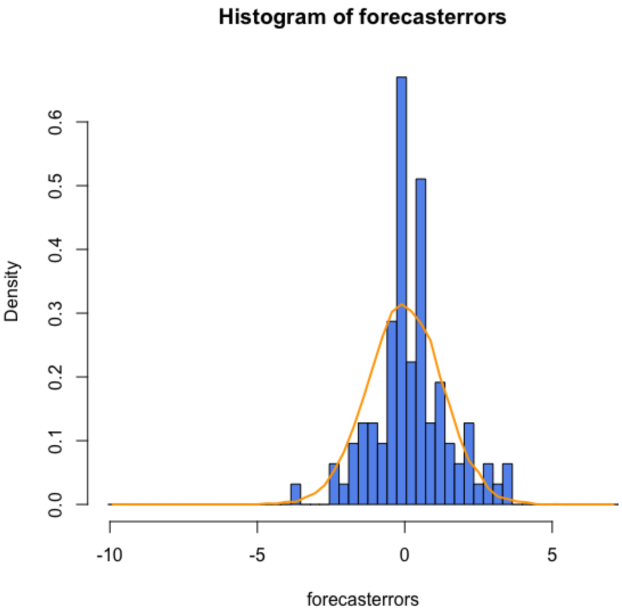

FIGURE 15. Forecast using the SARIMA$(0,1,1)(0,1,1)[12]$ model.

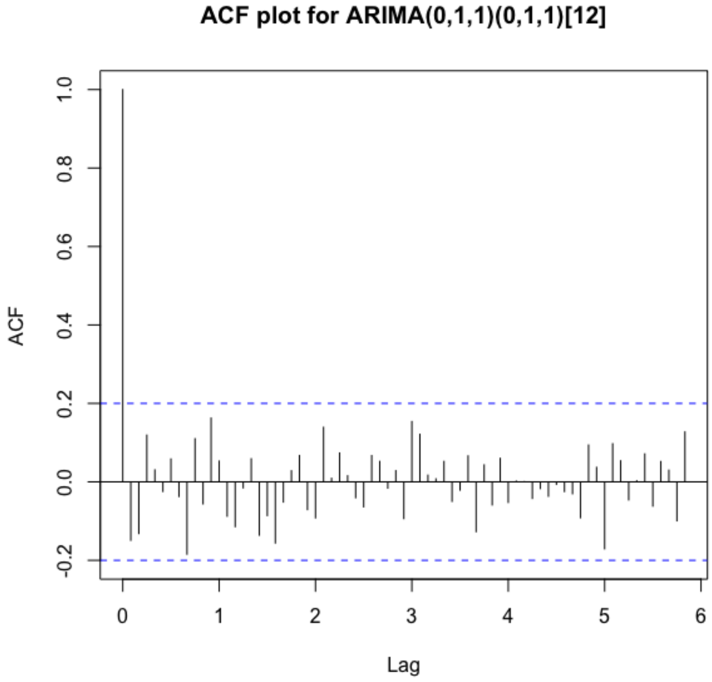

FIGURE 16. ACF plot of the residuals using the SARIMA$(0,1,1)(0,1,1)$ model

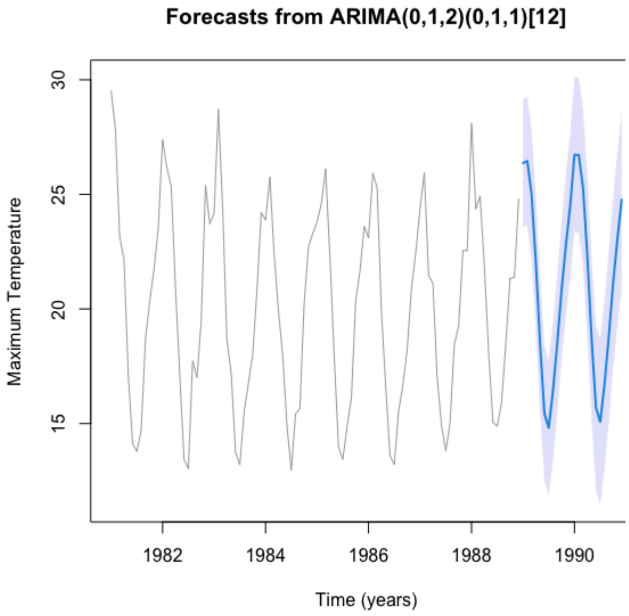

FIGURE 17. Forecast using the SARIMA$(0,1,2)(0,1,1)[12]$ model.

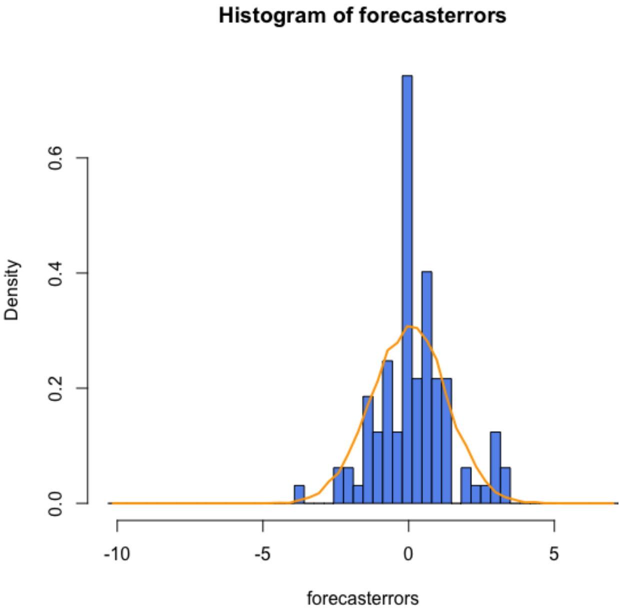

FIGURE 18. Forecast residuals using SARIMA$(0,1,2)(0,1,1)[12]$.

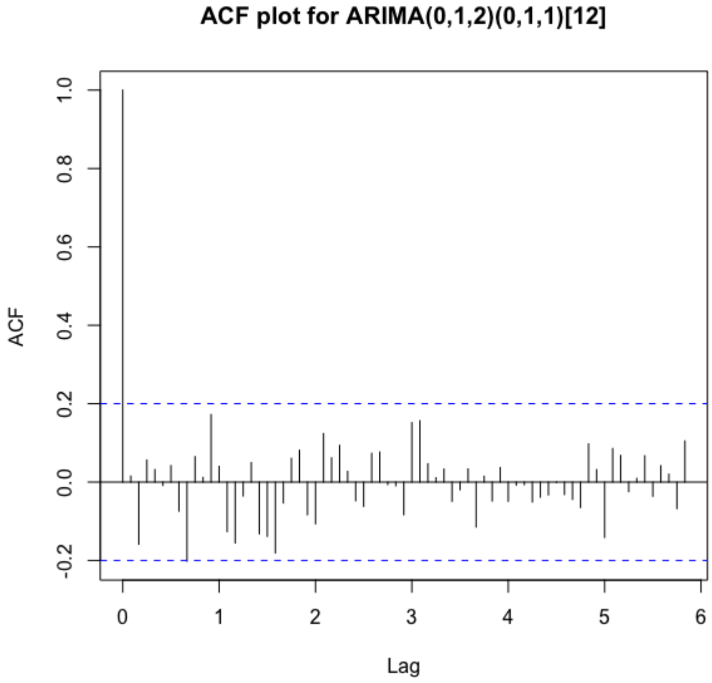

FIGURE 19. ACF plot of the residuals using the SARIMA$(0,1,2)(0,1,1)$ model