Time series analysis: Monthly temperatures at Newton Rigg Weather Station

Jacob Holmshaw

Abstract

This analysis examines monthly temperature data for Newton Rigg weather station in Cumbria from January 2000 to December 2019 with 10% of the data initially held back. The preliminary check on the data using a periodogram showed it had seasonality at 1/12 suggesting the series was not stationary. Fitting the model with harmonic regression removed this deterministic trend. Assessing the residuals using ACF and PACF plots as well as Ljung-Box statistics, an MA(4) model was found to be adequate to model these residuals. Forecasts were then produced using the model and compared to the data held back, with most of the values falling in our uncertainty bounds.

Introduction

In this analysis, the monthly mean daily maximum temperature at Newton Rigg Weather Station in Cumbria from January 2000 to December 2019 will be studied. The aim of the analysis is to model the data as an appropriate time series by choosing an appropriate model fit. This will be done by transforming our series to a stationary time series and then examining the model fit using time series diagnostics such as ACF and PACF plots, Ljung-Box statistics and periodograms. Some of the most recent data will be held back and future values will be forecasted, allowing comparison of these values later.

Time series techniques

Firstly, Pickup (2016) sets out some time series techniques and how they can be used to fit a model to the data as well as how to assess it so it can be referred back to later. **Transforming a time series** is necessary to transform the data to a stationary time series, as the models are for stationary data. This will be done by fitting a periodic trend but can alternatively be achieved by taking logs or differencing the data. By Derryberry (2014), **Periodograms** are used to search for cycles, trends and seasonality in the data and at what frequency. This is useful as if these are deterministic, they will dominate the pattern of the sample ACF plot, causing incorrect conclusions to be drawn. This will be used more than once until any trends that are deterministic are removed. The **ACF and PACF plots** are important as these will give us information about model fit, and also whether the residuals are characteristic of white noise. If the ACF plot displays a cutoff at lag k, meaning the values at lags higher than k are close to zero or within the 95% confidence regions, while the PACF shows a slow decays to zero, an MA(k) model can be considered. If instead the PACF has a cutoff at lag k, and the ACF slowly decays to zero, an AR(k) can be considered. If both the ACF and PACF show a slow decay to zero, an ARMA model may be appropriate. Lastly, from Glen (2014), **Ljung-Box statistics** are used to assess data fit. If the p values of the plot are high, the errors and residuals are white noise and are independent, so the model fits the data well. These techniques will now be used to analyse this data.

Initial analysis

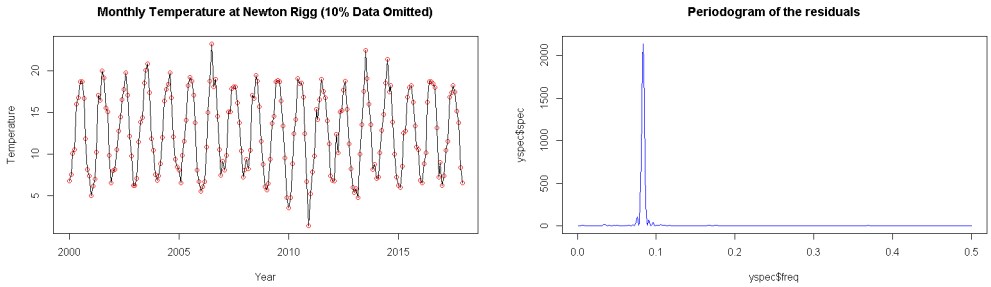

Figure 1: Showing (a) The plot of the original data minus 10% and (b) the periodogram of the data

From figure 1, the initial time series plot of Newton Rigg's monthly temperature displays seasonality. This is evident by the regular variations that occur monthly every year. After fitting a linear trend, assessment of the data using a periodogram will show us any deterministic trend. As explained earlier, the periodogram is useful for detecting cycles in a time series. The periodogram in figure 1 displays a large peak at frequency 1/12, due to seasonality. Continuing to assess our data without removing this would result in the trend dominating the ACF and PACF plots and therefore incorrect conclusions. Therefore, this peak must be removed by regression and this can be done using the seasonal factor model or with harmonic regression.

Model fitting

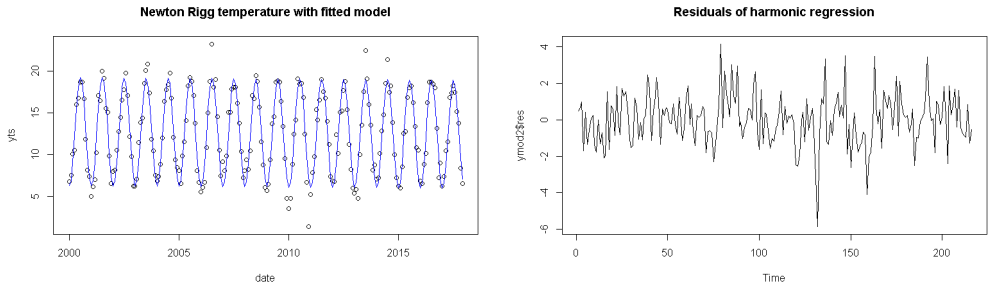

Figure 2: Showing (a) The fit of the harmonic regression to the data and (b) the residuals after the harmonic fit.

After the periodogram confirmed initial thoughts that the data follows a seasonal trend, this trend can be removed in two ways. This can be done using a seasonal regression model, or by using harmonic regression. The seasonal regression model repeats itself at regular intervals of 12. The monthly series can be modelled by $y_t = c + bt + M_{t,1}\alpha_1 + \ldots + M_{t,12}\alpha_{12} + e_t$ with $M_{t,i}$ corresponding to the 12 months. This model can be produced equivalently using the harmonic regression model $y = c + \cos 2\pi t/12 + \sin 2\pi t/12$ where t is time. Harmonic regression is considered instead of the seasonal model as it is possible to obtain a good fit using fewer than the full set of regressors which

arises when the seasonal effect is confined to just a few months in the factor model and if the seasonal cycle is very smooth in the harmonic regression.

Figure 2 shows that the harmonic regression is a relatively good fit to the data. It has not covered the highest and lowest values but this may be because these values do not follow the trend. The residuals at this point look relatively stationary but diagnostics can be used to assess whether they are white noise or not. From figure 3, the periodogram of the harmonic regression has removed the

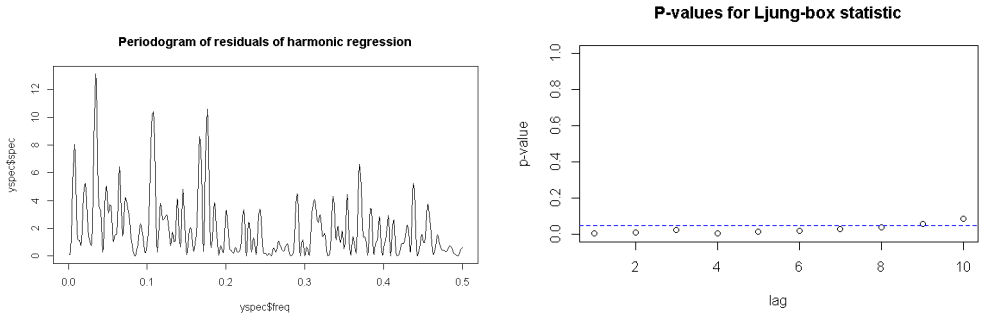

Figure 3: Showing (a) The periodogram after the harmonic fit and (b) The Ljung- Box statistics of the second harmonic regression

peak at 1/12 meaning it will not dominate the series or affect the ACF and PACF. The periodogram now has a scale of up to 12 and has a random fluctuations of peaks, indicating they are white noise. Our model is now $y_t = 12.85 - 0.002t - 5.35\cos\frac{\pi t}{6} - 3.68\sin\frac{\pi t}{6}$. The standard errors on these coefficients are 0.19, 0.002, 0.13, 0.13 respectively. Again, looking at figure 3, all the p values are below the 95% confidence line at this point. This indicates the residuals are dependent and so the model lacks a good fit to the data. To improve these values, an ARIMA model can be used by first looking at the ACF and PACF.

Assessing the ACF and PACF of the residuals

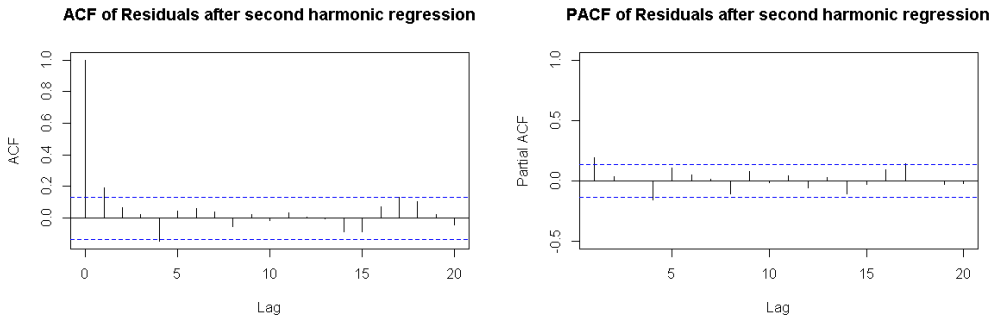

Figure 4: The ACF and PACF plots after harmonic regression

By judging the ACF and PACF plots of the residuals from the second harmonic regression shown in figure 4, a model fit can be proposed for the residuals. As the autocorrelation cuts off and

decays to zero after lag 4, an appropriate model would be an MA(4) model. This is supported by the PACF plot which slowly decays to zero. There is also evidence towards an MA(1), but the fourth coefficient has a large value compared to it's standard error, proving it is important to the fit and therefore result in a better fit for the residuals. However, it may be proved that this is not needed as the model still works with the MA(1) fit, which suggests MA(4) may be an overfit. Despite this, the analysis is continued with the MA(4), the coefficients are $\theta_1 = 0.2645$ with standard error 0.0669, $\theta_2 = 0.1201$; standard error = 0.0696, $\theta_3 = 0.1117$; standard error = 0.0665, $\theta_4 = -0.2093$; standard error = 0.0733. Note that the coefficients are all almost at least 2 times their standard errors which indicates these coefficients are needed. This gives the final model $y_t = 12.85 - 0.002t - 5.35\cos\frac{\pi t}{6} - 3.68\sin\frac{\pi t}{6} + e_t - 0.2645e_{t-1} - 0.1201e_{t-2} - 0.1117e_{t-3} + 0.2093e_{t-4}$.

Assessing the residual model fit

The ARIMA model can now be checked diagnostically to assess how appropriate it is for the data. Firstly, the residuals look closer to white noise and the y scale is much smaller, which indicates they are independent. The ACF now has values which are all within the 95% confidence bounds so this also indicates the fit is a good one. Figure 5 shows Ljung-Box statistics, which indicate all values

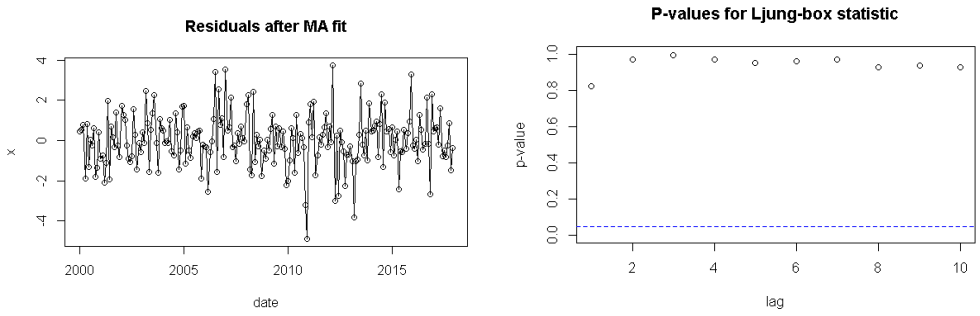

Figure 5: Ljung - Box Statistics of MA(1) fit

are significantly different from zero, which means the errors are independent and the residuals are close to white noise, so the model is a good fit. This compares to the MA(1) fit which is not shown which had lower p values and the lowest value at lag 4. Concerning weaknesses of a moving average model, future values are calculated based on past values, so our forecasts will not account for any drastic changes.

Forecasting

In figure 6, the residuals and the original data has been forecasted. The residuals forecast shows they eventually level out meaning the residuals become stationary which is desired.

The forecast of the original data using the second harmonic regression model produces a wave similar to the rest of the data. When compared to the data points that were originally omitted, the curve appears to fit the trend of the points very well. The uncertainty of the model and linear trend that were added capture the range of the data other than one reading, which may be because the data that was omitted followed a different trend. When the uncertainty from the linear trend is

added, it is noticeably extremely close to the uncertainty from the model but slightly wider, which may suggest there is more uncertainty from the linear trend.

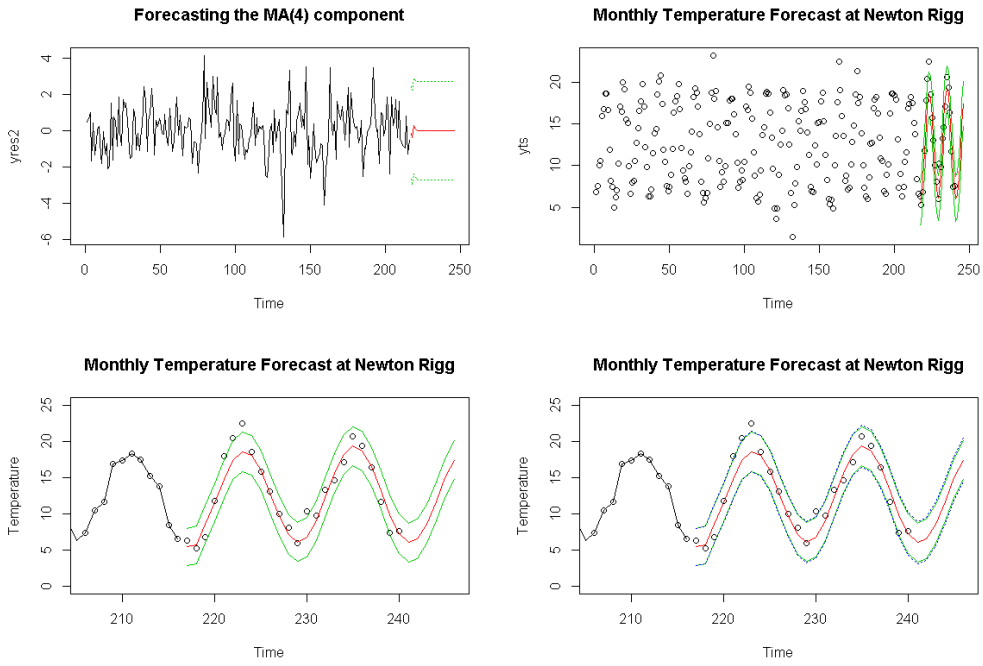

Figure 6: Showing (a) the forecast of the residuals, (b) the forecast of the original data, (c) the forecast of the original data zoomed in and (d) the forecast with linear trend uncertainty

Conclusion

In summary, it was found that the time series could be transformed using a harmonic regression model to remove seasonality plus an MA(4) for the residuals. This led to forecasting the data which showed a similar temperature pattern is expected to what has occured over the last 20 years, with some variation. This analysis addressed the aims of the study by modelling the data as a time series and assigning the harmonic regression plus MA fit to forecast the data. The limitations of the data included that the residuals did not produce ACF and PACF plots that clearly pointed to one model, but a model was still found to be a good fit.

References

Derryberry, D. R. (2014). *Basic data analysis for time series with R*. Wiley, Hoboken, New Jersey. 1

Glen, S. (2014). *Ljung Box Test: Definition*. Available at: `https://www.statisticshowto.datasciencecentral.com/ljung-box-test/`. 1

Pickup, M. (2016). *Introduction to time series analysis*. Quantitative applications Introduction to time series analysis. SAGE, Los Angeles. 1