Fitting a Time Series with a Proper Model

34643397

February 2020

Abstract

Given data of the monthly mean maximum temperature in Cumbria, a model is set out to be found by deeply analysing its time series features, such as the periodogram, the autocorrelation, the partial autocorrelation and the residuals. The data is then modified based on previous discoveries in order to be properly modelled and hence compared by two model candidates. The preferable model is found and then used to forecast future data.

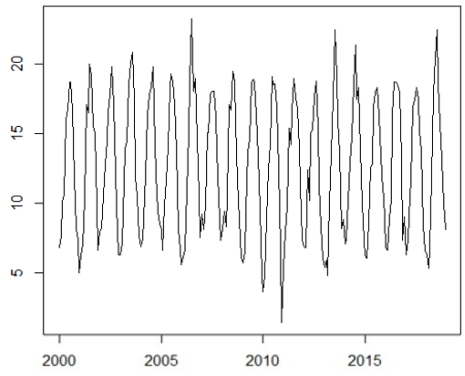
1 Introduction

The time series data on the monthly mean maximum temperature in Cumbria has a range of values from January 2000 to December 2019, and as it is a monthly series there are 240 values in the data. In order to be able to forecast the data, 5% of the last given values, so 12 points (one year), have been omitted so that one may later compare the original data to the predictions of the created model. Hence the purpose of this project is to build a model that fairly forecasts the original data in future results.

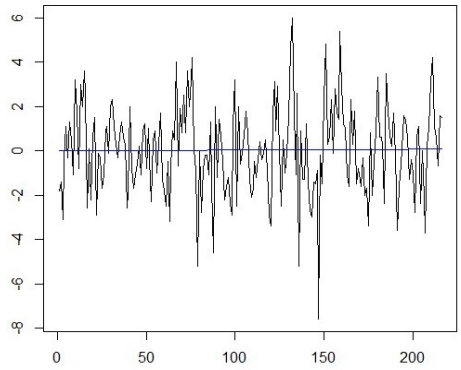
2 Theory

2.1 Analysing the Data

By plotting the temperature in Cumbria data, its details can be analyzed.



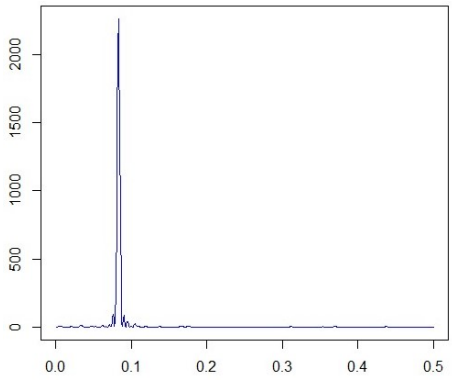
(a) Plot of Cumbria data with 5% of points omitted. January 2000 to Dec 2018.



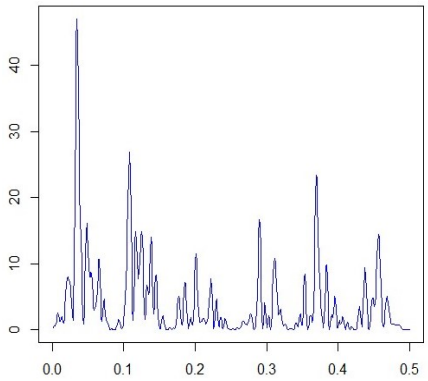
(b) Plot of Cumbria data differenced and with additional linear regression.

Figure 1

Figure (1a) shows the data with 12 points omitted, where it demonstrates that the time series doesn't seem to increase nor decrease throughout the 19 years; it doesn't exhibit a trend. Furthermore, the plot evidently displays monthly seasonality, therefore the frequency is $\frac{1}{12}$. By inspecting the real values in the data, it appears that the maximum temperature is usually in July and August, which is rational as it is Summer. Analysing the autocorrelation function of the data, grants a plot that looks like a sinusoidal function which is a definite hint for seasonality. For a time series to be stationary its statistical properties cannot change throughout time. As we have seasonality we don't have stationarity. Hence we remove the seasonality to proceed to modelling the series. In Figure (1b) the original data has been differenced in order to remove seasonality. Seasonal differencing is the difference between an observation and the previous observation from the same season. If we hadn't removed the seasonality then it would affect the pattern of the sample acf, concealing statistical features. The series has also been linear detrended and plotted in the same figure to confirm that there isn't a trend in the series.
The plot of the periodogram of both the original and the differenced series is then inspected.



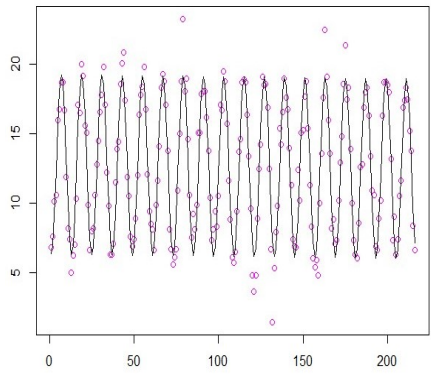
(a) Periodogram of untouched Cumbria data.



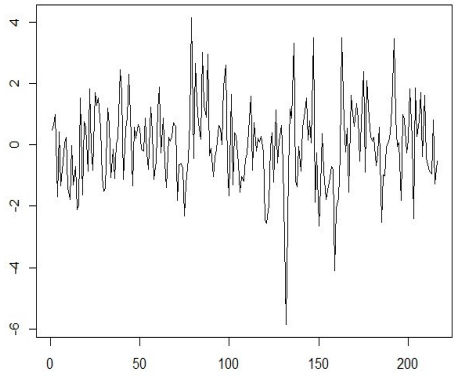
(b) Periodogram of residuals after differencing the Cumbria series.

Figure 2

There are two types of periodogram plots that can be analyzed, the linear and log scale. The linear scale periodogram is more relevant since it is best for peak spotting, whereas the log scale is preferable for recognizing white noise looking data. In Figure (2a), the periodogram (linear scale) of the original data is shown, where a prominent peak is revealed at $\frac{1}{12}$. This peak is the fundamental frequency as discussed previously, this frequency is logical as the monthly seasonality hasn't been erased. In Figure (2b) the periodogram (linear scale) of the residuals is displayed and it can be observed that the peak from before has drastically diminished. In the first periodogram the peak was placed at about 2262.791 and on the second periodogram the peak decreased into roughly 0.107. Hence approximately 100% of that peak has been modelled, however, new peaks have emerged, yet they are small in comparison to the first peak.
Next, we apply a harmonic regression to the series.



(a) Plot of weather in Cumbria and harmonic regression (fundamental only).

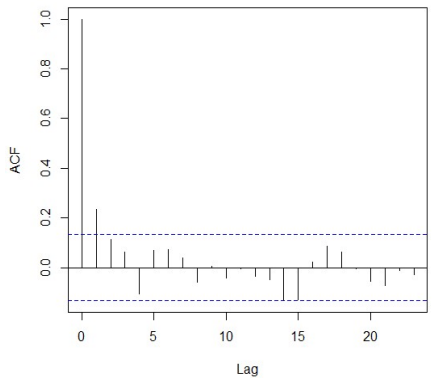


(b) Plot of residuals after harmonic regression (fundamental only).

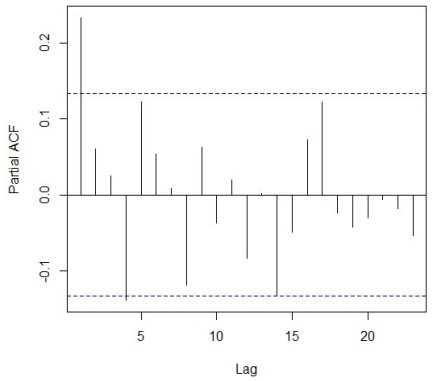
Figure 3

A harmonic regression is applied to the series in Figure (3a), to the original and linear detrended series. It does a decent job modelling the series — it fails to model some values but overall it is not an apparent bad model. Looking at the estimates of the *cos* and *sin*, the results reveal that these estimates are significantly larger than the standard errors, implying that the regression is signified. When plotting the residuals, Figure (3b), the plot appears like white noise. Therefore it is not a bad start for a model.

Then, the plots of the acf and pacf of the residuals after harmonic regression were analyzed. In Figure (4a), the acf of the residuals after adding a harmonic regression is evidently showing that at lag 1 the value lies outside the confidence bound. The remaining lags seem fairly close to zero. The lag at 1 might be a hint into testing a moving average model, MA(1), as it is the MA(1) characteristic property that the autocorrelation maybe vanishes at lag 2 and higher. On the plot on the right, Figure (4b), the pacf of the residuals demonstrate that the lag 4 lies outside the confidence bound. And similar to the acf, the remainder of the lags lie inside the bounds, the lag 14 might also be relevant as it is almost breaching the bounds. However the lag 4 might be encouraging to fit an autoregressive model, AR(4), as the characteristic property of an AR(4) model is that its pacf has a cut-off at lag 4.



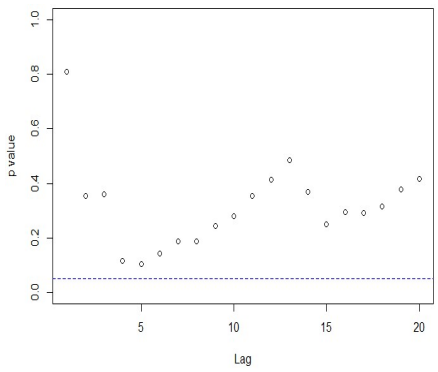
(a) Plot of acf of residuals after harmonic regression.



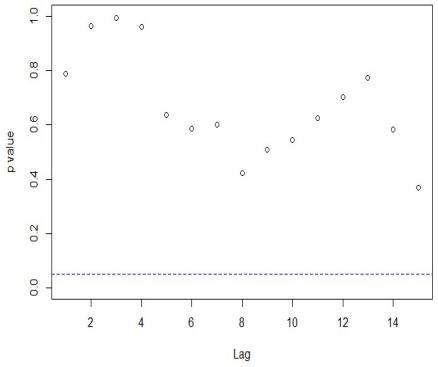
(b) Plot of pacf of residuals after harmonic regression.

Figure 4

Therefore comparing these two models will give guidance into finding the best fit. To begin, the acfs of the MA(1) and AR(4) models were plotted. With regard to the acfs, the optimal plot would have its lags remain close to zero. In both of the acfs, the lags seem to remain inside the confidence bounds except for lag 4 in the MA(1) model, which is notably near to crossing its bound. When plotting the pacfs, they exhibit a more significant change than they did with the acfs. In the MA(1) a lag appears outside the bounds, again at lag 4. However, in the pacf of the AR(4) all the lags remain inside the confidence bounds.
In order to identify a model is a good fit, the p values of the Ljung-box test were plotted.



(a) Plot of p values with Ljung-Box for residuals from harmonic regression plus the MA(1) with $\theta = 0.209$.



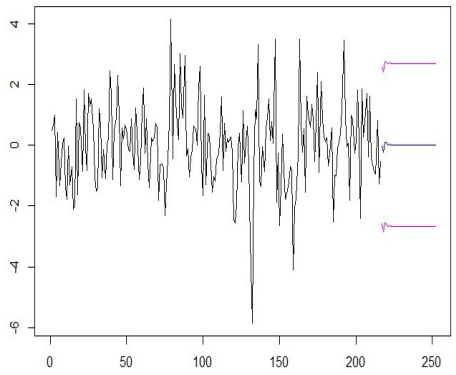
(b) Plot of p values with Ljung-Box for residuals from harmonic regression plus the AR(4) with $\phi_1 = 0.221$, $\phi_2 = 0.063$, $\phi_3 = 0.057$, $\phi_4 = -0.141$.

Figure 5

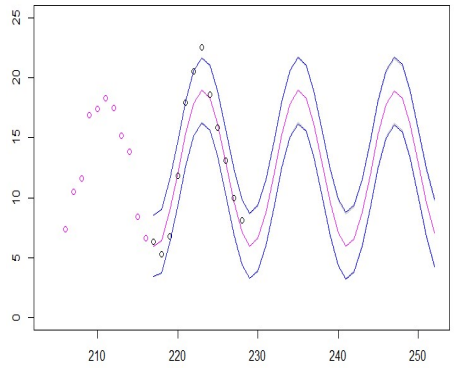
In Figure (5) the p values with Ljung-box for both competing models can be seen. Both have their values above the significance line, hence the residuals resemble white noise. Therefore both models are valid. Were the values to be too low, it would mean that the model is rejecting the null hypothesis of white noise in favor of an alternative model which has correlation. As the values for AR(4) are higher than the values for MA(1), the autoregressive model is chosen to proceed with forecasting.

2.2 Forecasting the Data

Having chosen the AR(4) model, the predictions of the AR(4) fit were plotted, and subsequently, a forecast of the original data is produced.
Figure (6a) shows the forecast of the AR(4) component which has correctly placed bounds with no oscillations. In Figure (6b), the enlarged plot of the forecast of the original data is shown where the pink lines are the forecast, the pink dots are the given data and the black points are the values that were omitted in the beginning of this project.



(a) Plot of Forecast of AR(4) component.



(b) Plot of Forecast of original data from model plus uncertainty.

Figure 6

The black lines underneath the blue lines are uncertainty. As observed only one of the dots is outside the uncertainty bounds, hence about 91.6% of values are well forecasted. It is shown that cycle of the data is being well captured in this model. The forecast of uncertainty from the deterministic component is also plotted with blue lines and are seen to be departing from the black lines in the future, hence it has a wide uncertainty. These uncertainty lines were only added by assuming that the uncertainties are independent from each other.

3 Conclusion

The purpose of this project was to build and fit a model to the given time series, and subsequently, have the model forecast future values. The given time series was thoroughly analyzed and modified in order to be fitted by a proper model. Two models were then examined and compared, so that one may choose the best model to advance to forecast future values. The winning model was then forecasted with uncertainty bounds, appearing to do a decent job modelling the data as all but one value remained inside the uncertainty bounds.