STAT 443 Project - Analyzing Vancouver Climate Time Series

Kevin Yang, Manni Zhang, Sienna Lee, Kejin Ren (Group C5)

Introduction

Our research question is as follows: Has the average temperature in Vancouver increased since 2003? Many problems occur as a result of higher temperatures, (e.g. global warming, rising sea levels, abnormal climate, etc.) and it affects both humans and animals. Many recognize the consequences that will arise from global climate change, but often do not recognize the severity of it due to lack of first hand experience. According to The Proceedings of the National Academy of Science, if the average global temperature rises by 2 degrees Celsius before industrialization, the earth is expected to enter a warm period. A warm period puts 20~30% of the world's species at risk of extinction, greatly impacting the world's ecosystem. Moreover, as deforestation destroys forests around the world, less carbon dioxide is absorbed, and the effects of greenhouse gases will become more pronounced. In short, there will be catastrophic consequences if the global temperature rises by more than 2 degrees Celsius.

How can we raise awareness of this issue? Many are aware of the consequences, but lack the motivation to make eco-friendly decisions in their every day lives. In order to address this issue, we want to conduct a time series analysis on temperature data to investigate whether there is an upward trend in temperature over the years.

The Government of Canada provides public temperature data. We use a dataset from the Government of Canada containing measurements such as precipitation, max gust, and the variable of interest in this study, mean temperature for Vancouver from January 1, 2003 to December 31. The mean temperature for each day is obtained by acquiring measurements in the morning, afternoon, and evening, and averaging values together. Due to the presence of missing data in 2020, we only perform the analysis up to March 31, 2020. Any other missing values are set to be the mean of the entire series. The original dataset records daily temperature observations and consists of 6575 rows and 31 columns. For each month, we average the mean daily temperatures, producing a mean temperature for each month. The adapted dataset has 207 rows and 1 column for mean monthly temperature.

Exploratory Data Analysis

The plot of the time series (Figure 1) strongly suggests a seasonal component of one cycle per year. Temperature is seasonal in nature so this is not surprising. A consequence of the series being seasonal is that the series is most likely non-stationary. Differencing can be used to address the non-stationarity of the time series.

There does not seem to be a clear upward trend in the series, but there could be some trends within seasons. For example, the winter seasons after 2015 appear to be changing more dramatically than prior years.

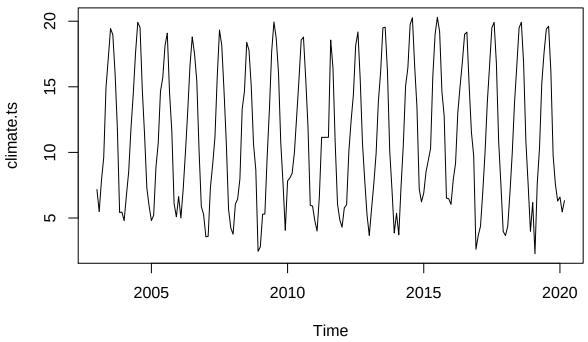


Figure 1: Plot for the Vancouver temperature time series

The correlogram of the time series (Figure 2) exhibits a periodic characteristic and is consistent in its' wavelength and amplitude throughout. This could indicate a seasonal component to the time series.

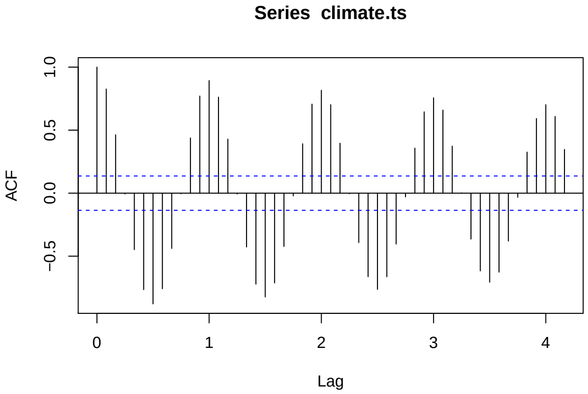


Figure 2: Correlogram for the Vancouver temperature time series

The differenced series (Figure 3) no longer has a clear seasonal component and looks stationary. The differenced series also does not have a clear trend and oscillates around 0. Taking the first difference appears to be sufficient for subsequent portions of the analysis.

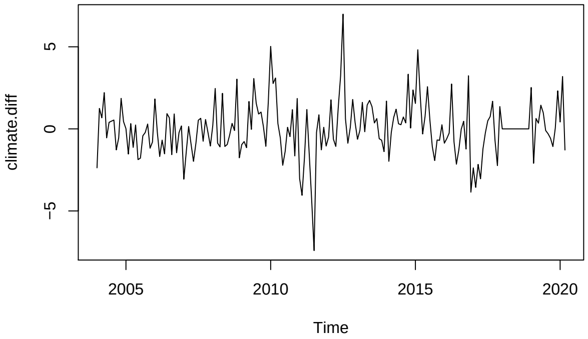


Figure 3: Plot of the first difference for the Vancouver climate time series

Analysis Results

From the Exploratory Data Analysis, since a clear seasonal component was observed and differencing was used to remove non-stationarity, a SARIMA model for the time series seems appropriate. In order to determine the best SARIMA model, the correlogram and partial correlograms are used (Figures 4 & 5). Note: each lag constitutes an entire cycle which corresponds to 1 year.

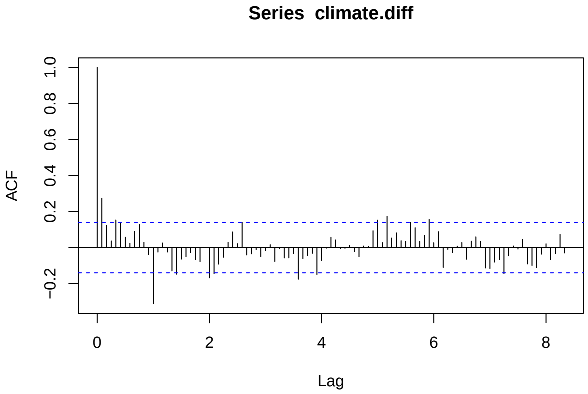


Figure 4: Correlogram of the first differenced series for the Vancouver climate time series

For earlier lags, the correlogram appears to cut off at lag 1 possibly indicating an MA(1) component. For lags that are multiples of 12 which correspond to the frequency, the correlogram appears to cut off at lag 12 suggesting Q=1.

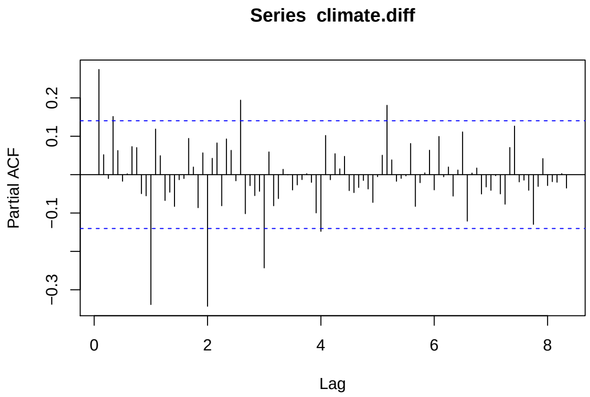
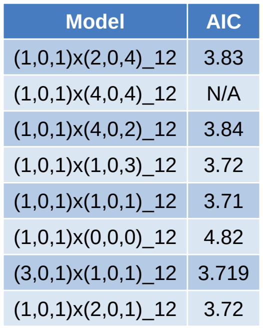


Figure 5: Partial correlogram of the first differenced series for the Vancouver climate time series

For earlier lags, the partial correlogram appears to cut off also at lag 1 possibly indicating an AR(1) component.

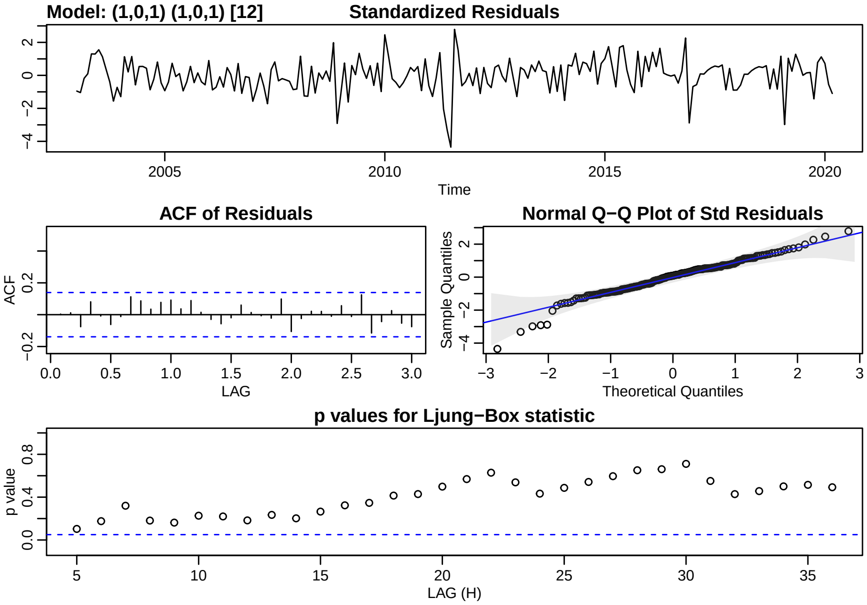
For lags that are multiples of 12 which correspond to the frequency, the partial correlogram appears to cut off at lag 24 suggesting P=2.

Multiple models were fitted. The ACF of the residuals, the Q-Q plot of standardized residuals, the p-values for the Ljung-Box statistic, and AIC values were all used to determine the best model. The table below records the models explored and their corresponding AIC values.



Model	AIC
(1,0,1)x(2,0,4)_12	3.83
(1,0,1)x(4,0,4)_12	N/A
(1,0,1)x(4,0,2)_12	3.84
(1,0,1)x(1,0,3)_12	3.72
(1,0,1)x(1,0,1)_12	3.71
(1,0,1)x(0,0,0)_12	4.82
(3,0,1)x(1,0,1)_12	3.719
(1,0,1)x(2,0,1)_12	3.72

The lowest AIC corresponds to the SARIMA(1,0,1)x(1,0,1)_12. The ACF of the residuals do not have any significant values at 95% confidence, the Q-Q plot suggests normality of the standardized residuals, and the p-values for the Ljung Box statistic are all significant. All these indicate that the model is a good fit to the series. Due to the many models attempted, only the coefficients and complete results for the winning model (SARIMA(1,0,1)x(1,0,1)_12) is displayed below.



Since the time series is seasonal, the Holt Winters Exponential Smoothing method is appropriate to forecast temperatures for 2020. Figure 6 below shows a plot including the forecasted data points for the series. The fitted line appears to be similar to prior years, but the upper and lower bounds do fluctuate quite a bit.

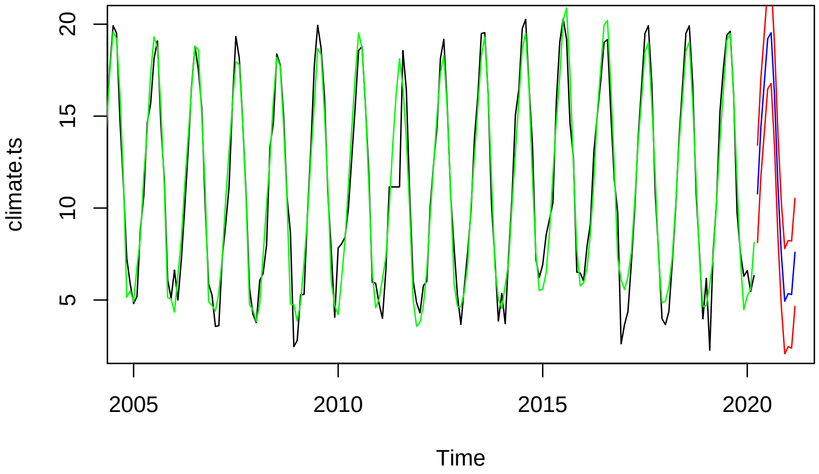


Figure 6: Forecasting using Holt Winters Exponential Smoothing method. The blue line corresponds to fitted values, and red lines to the upper and lower bounds

The Seasonal Mann-Kendall Trend Test (Results shown below) is used to determine if there is a trend for Vancouver's mean monthly temperature. The null hypothesis is that there is no trend for the time series, and the alternative is that there is an upward trend for the time series.
The results give reason to believe that there is an upward trend in the time series. The seasonal Kendall statistic S is 190 and the p-value is 0.0045, both providing evidence to reject the null hypothesis at the 5% significance level. Therefore, there is reason to believe that temperature has increased since 2003.

Conclusions

The analysis results suggest that the SARIMA(1,0,1)x(1,0,1)__12 best models the monthly average temperature in Vancouver. The Holt Winters Exponential Smoothing Method is used to forecast temperatures for 2020. The monthly temperature is forecasted not to increase dramatically in 2020. However, to investigate whether the average temperature in Vancouver has increased since 2003, a Seasonal Mann-Kendall Trend Test is performed on the time series. The test results provide sufficient evidence that Vancouver's mean monthly temperature has increased since 2003. However, a limitation of this test is that the results are greatly influenced if conflicting trends exist within seasons (some upward trends and others downward). This may cause issues in the analysis because much current research in the area of climate change say that temperatures are becoming more extreme (ex. summertime gets hotter and wintertime cooler) instead of merely rising on average. Since the trends of seasons may be in different directions, the results of this test might be misleading. Further investigation can be done in this area.