STATS 4A03 Final Project - Monthly Average Temperature in New York State (Albany)

Jessica Zhang (400390028)

2024-04-10

Contents

(NOAA (n.d.))

Section 1: Introduction

The climate data for this project was sourced from the Kaggle website, containing daily, hourly, monthly, and three-hourly temperature readings from January 2015 to May 2022 in Albany, New York State. In this project, we are focusing on forecasting the monthly average temperature using the monthly data file, includes 13 variables and 87 observations in this data set. Thus, 87 months are collected to be used for model determination and the following 24 months temperature will be predict by the founded model.

The purpose of the monthly temperature forecast is to describe a rough chronology which can contribute to decision making in various industries, such as electricity companies and farms. Electric utilities can use these forecasts to predict monthly gains. Farmers can planing their planting schedules due to those forecasts temperature.

Section 2: Modelling

We would like to plot the initial data and find the trend.

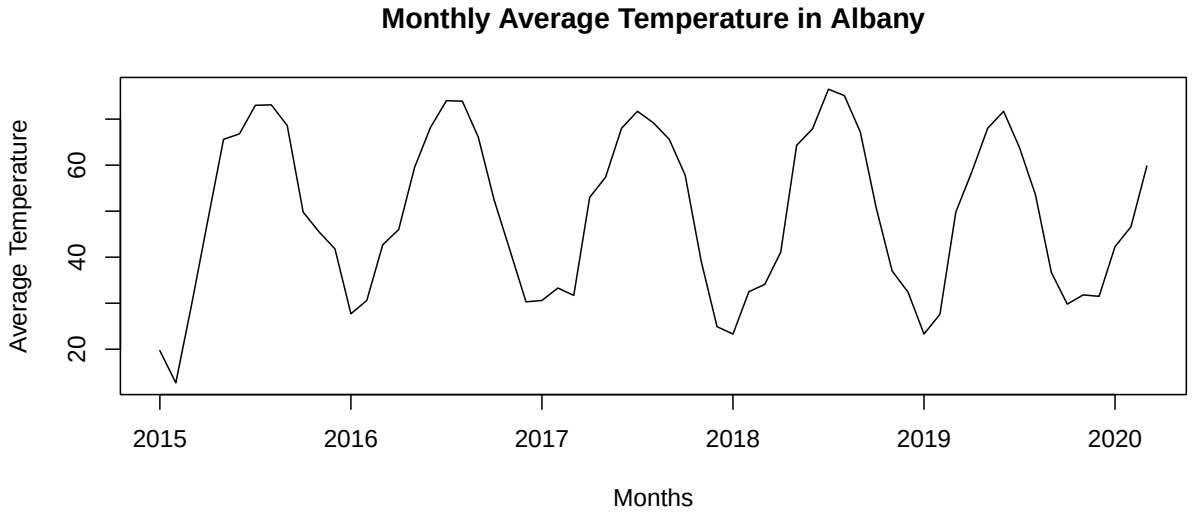


Figure 1: Average temperature in Albany

From the above plot, it is a **non-stationary** series. However, the plot follows a similar pattern after a period of time. To finding the frequency of this pattern, rather directly use the code or visualizing from the plot that the frequent is 12.

After that, we want to determined the model specification based on the sample ACF and sample PACF. This methods is to used for determining the simplest time series model, AR(p)

or MA(q).

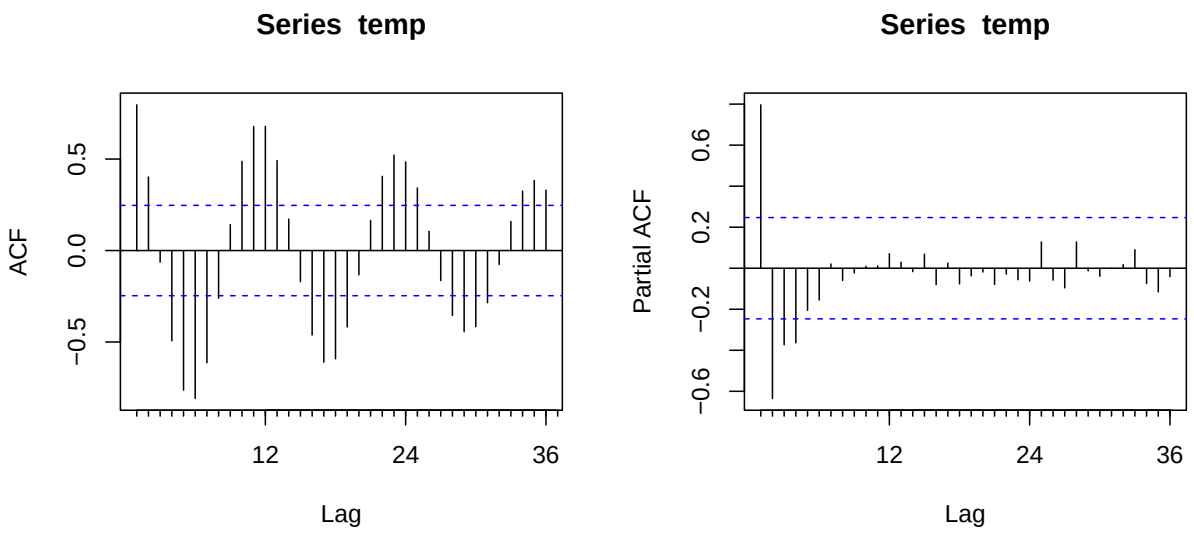


Figure 2: ACF and PACF Plot of Time Series

According to the autocorrelation function (ACF) plot, the first lag is the first peak in this model and the second peak is the 12th lag. Thus, the peaks trend to be 12 quarters apart, and so as the troughs trend. As this result, the ACF of the average monthly temperature in Albany decays to zero asymptotically.

For the partial autocorrelation function (PACF) plot, there is a cut off after the 4th lag. For the PACF plot, it is used to determined the auto regressive (AR(p)) model. Thus, the p value can be considered as 4 ($p = 4$).

Based on the above two conclusion, we obtain the same result that the data are non-stationary. Therefore, we consider applying the seasonal variation as the frequency of this data is 12.

To identified a more accurate seasonal model, we can apply the same process by checking the ACF and PACF plots with a seasonal difference of lag 12 ($s = 12$ and $D = 1$) and check its residual analysis.

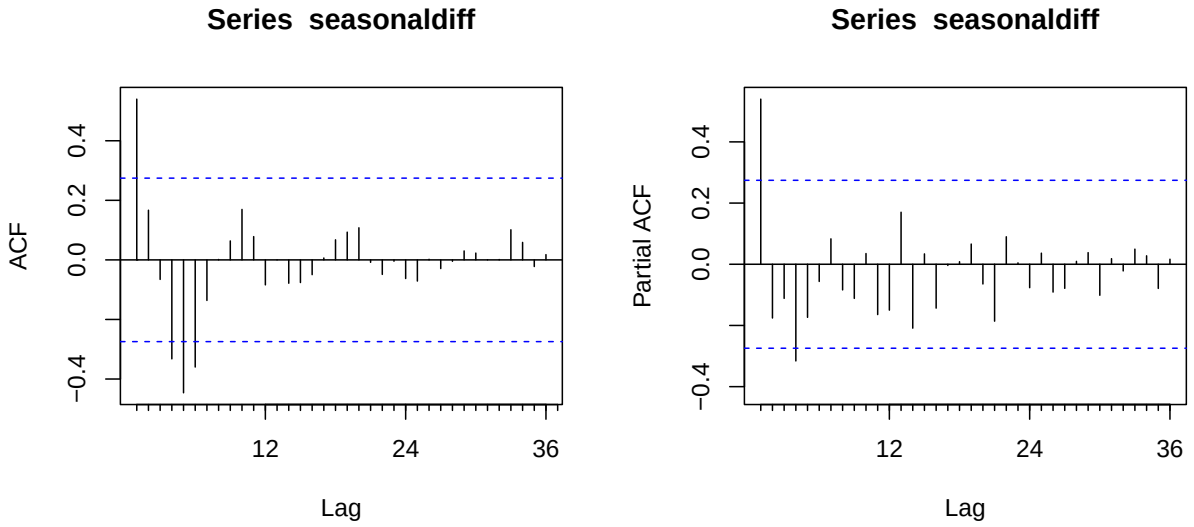


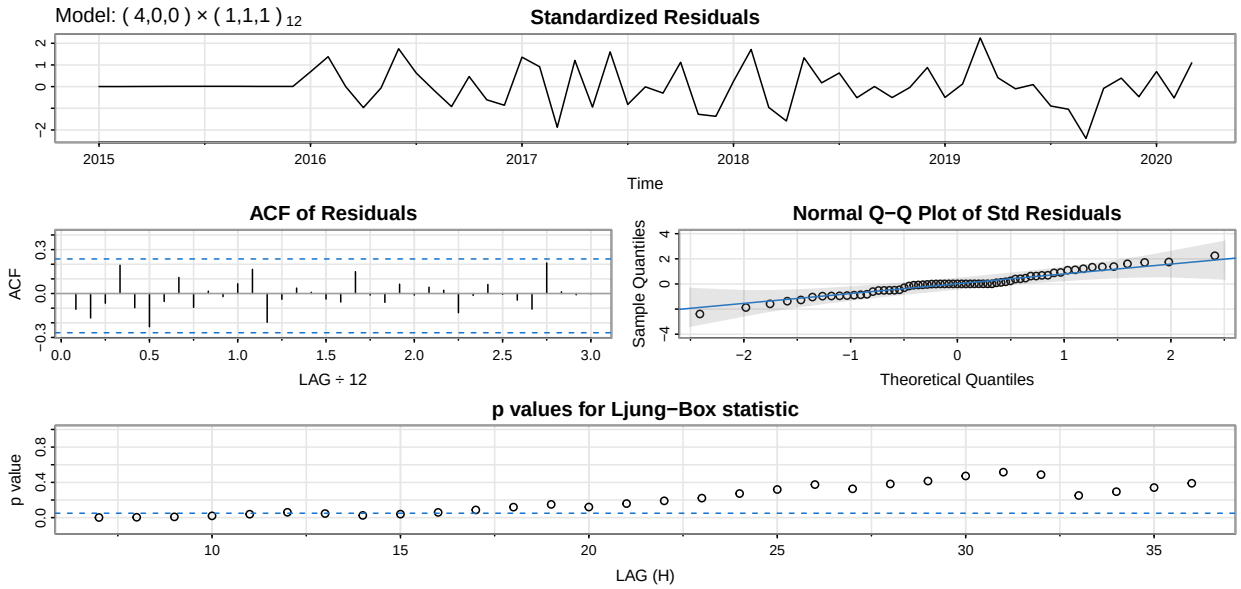
Figure 3: ACF and PACF Plot of Seasonal Differenced Time Series

From the seasonal difference ACF and PACF plots, the first peak for both plot is at lag 1. According to this output, we considered that $P = 1$ and $Q = 1$.

Finding the best fitted model, we will repeating the process of refit the model and check its residual analysis plots. For a precision model, ACF of residuals should be closed enough to the confidence interval and the QQ-plot of standard residuals should be more normalize. At the same time, we also want the p value for Ljung-Box test statistic to be larger than the 0.05 alpha line.

Check whether the model is:

$$Y \sim \text{SARIMA}(4, 0, 0) \times (1, 1, 1)_{12}$$



From the above output, the ACF of residual plot for model $Y \sim \text{SARIMA}(4,0,0) \times (1,1,1)_{12}$ has no spikes greater than the confidence interval with 0.5 significant level. Meanwhile, the QQ-plot of standard residuals is not approximate normal distribution, the points at the center do not fall close to the reference line. Moreover, the p values for Ljung-Box statistic maintain the minimum at the beginning, which reject our null hypothesis that H_0: residuals are independent. As a result, the model fail to be true, and need to be improved.

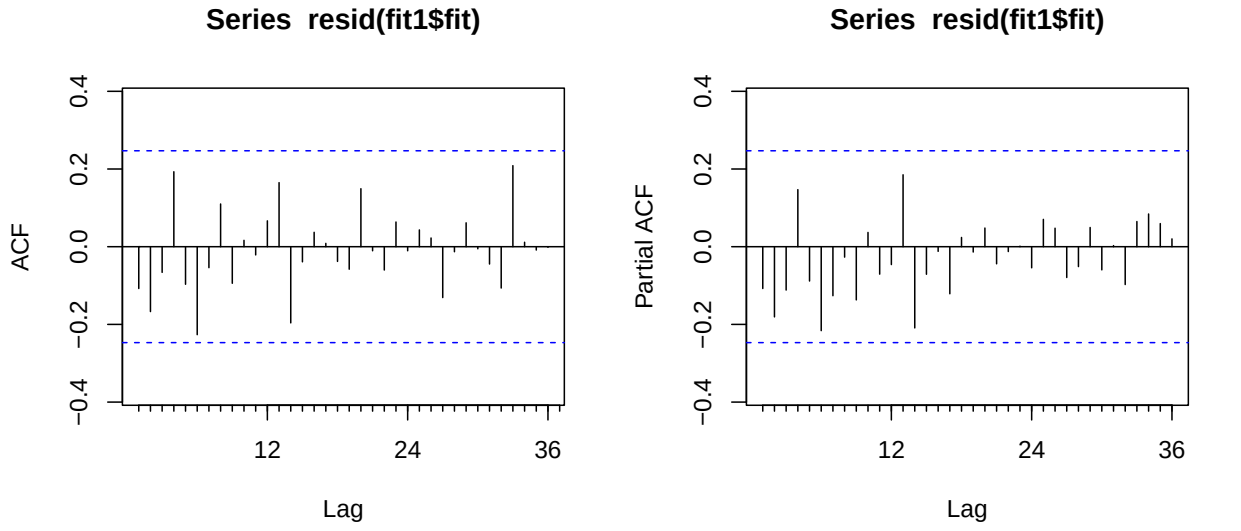


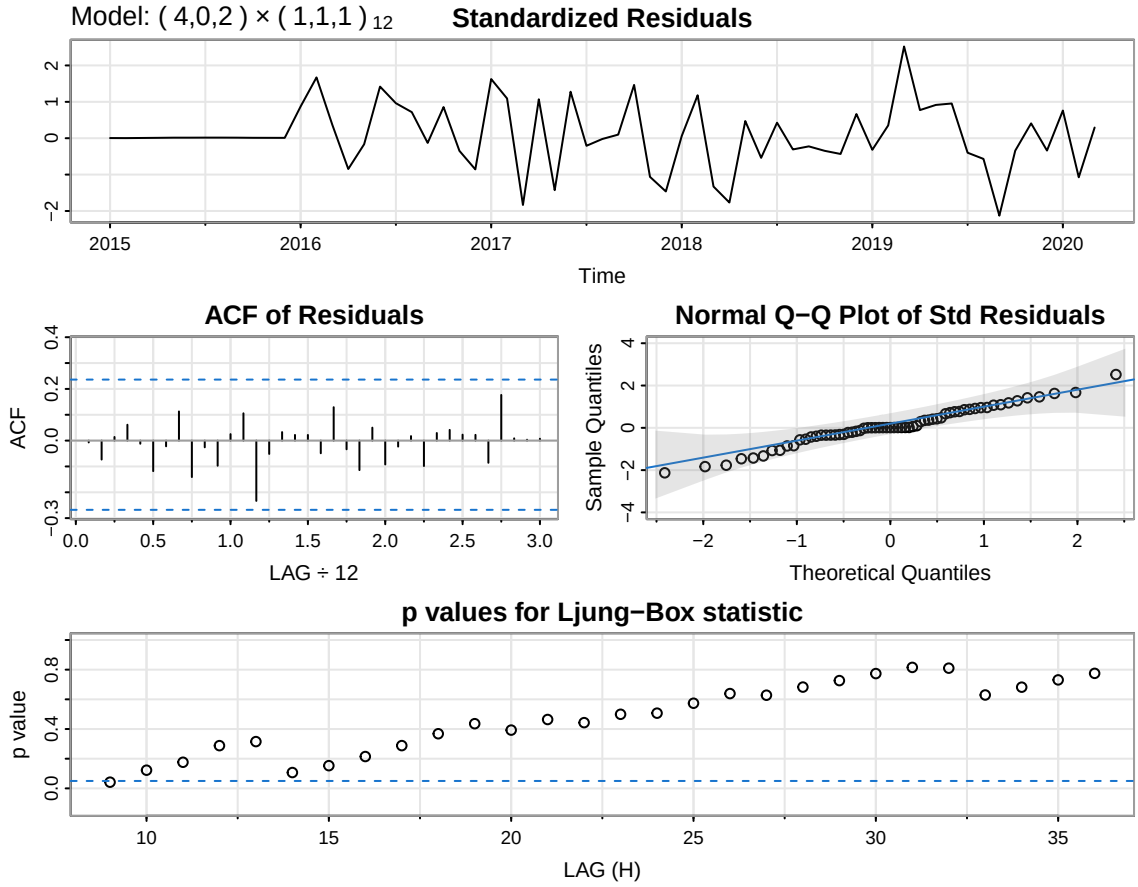
Figure 4: ACF and PACF Plot of the First Fit Seasonal Differenced Time Series

The residual plots are all in between the cut-off lines, thus, this model is relative to be good fitted. However, we want improve our model to get a more normalized normal QQ-plot of standard residuals and increase the p value for Ljung-Box test statistic. According to the

PACF plot of residuals, the model has peaks at $p = 2$.

Check the fitted model:

$$Y \sim \text{SARIMA}(4, 0, 2) \times (1, 1, 1)_{12}$$



The model is better than the previous one, the ACF of residuals are inside the cutoff point. The normal QQ-plot of standard residuals are approximately normal distribution but it is narrow at the end of two sides, but it is better than the first fit model. Additional, the p-value for Ljung-Box test statistic are greater than the 0.05 significant level that we fail to reject the null hypothesis, H_0: the residuals are independent. In conclusion, this model is better than the previous one.

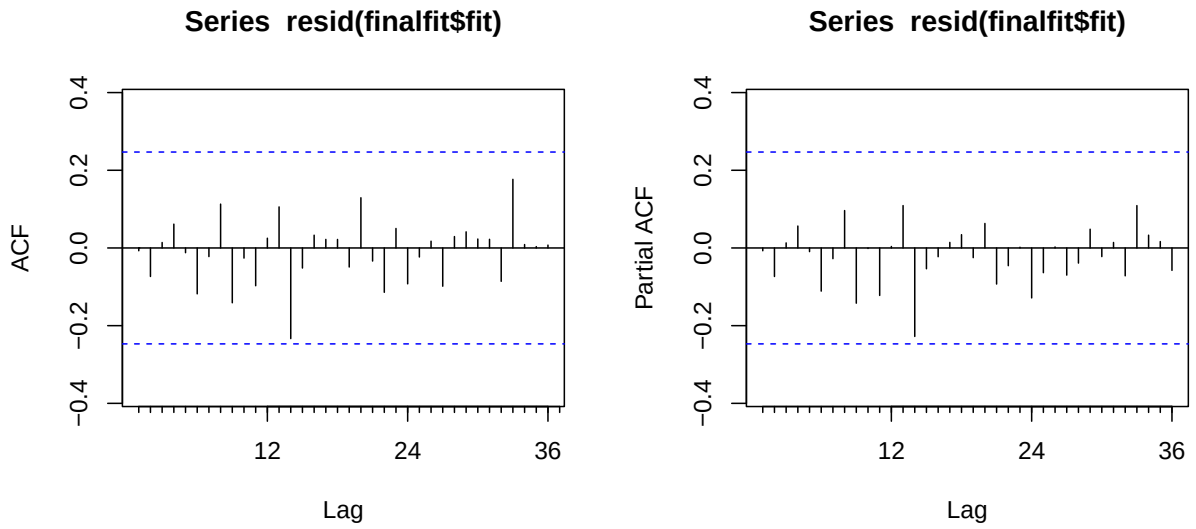


Figure 5: ACF and PACF Plot of the Final Fit Seasonal Differenced Time Series

The model of residual ACF and PACF plots are all in between the cut-off lines. By the combination of all the residual analysis output, the data is now **stationary**. As a result, we considered the monthly average temperature data can be represent as the seasonal difference model: $Y \sim \text{SARIMA}(4,0,2) \times (1,1,1)_{12}$.

Using this model to forecast 24 steps ahead. The forecast period is from April 2020 to May 2022.

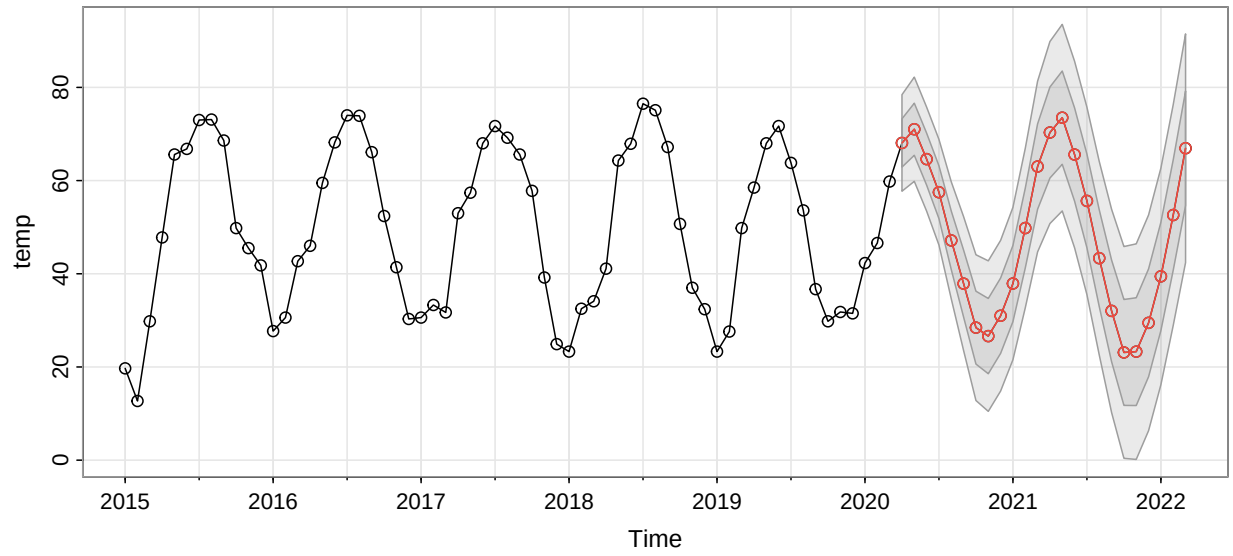


Figure 6: Forecast for monthlly average temperature plot

Section 3: Results

The model of the average monthly temperature in New York States, Albany, is:

$$Y \sim \text{SARIMA}(4,0,2) \times (1,1,1)_{12}$$

We notice that, the range of predicting the 95% confidence interval is pretty narrow, this means our model is confident about our predicting value.

Next, plot the true value of monthly average temperature in Albany and compare the forecast data with the true value.

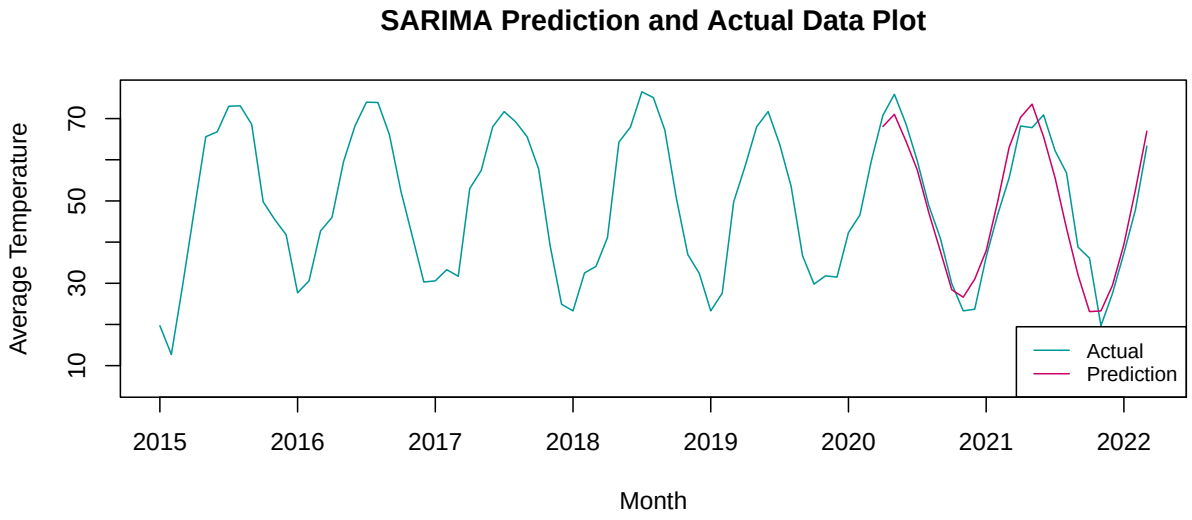


Figure 7: True value for monthlly average temperature plot

Based on our model, the forecast average monthly temperature is closed to the true value. Thus, the model $Y \sim \text{SARIMA}(4,0,2) \times (1,1,1)_{12}$ has a good prediction on the monthly average temperature in Albany.

Section 4: Conclusion

In this project, we use the SARIMA $Y \sim \text{SARIMA}(4,0,2) \times (1,1,1)_{12}$ model to predict the monthly mean temperature for the next two years and use our predicted average temperatures to support the Electric Company's monthly profitability as well as the farm's planting schedule. Based on the average monthly temperatures predicted by this model, the Electric Company expects to see an increase in earnings from its heating system in 2021 and 2022 compared to the winter of 2020. In addition, heating demand will be greater in 2022 compared to 2021. For farms, in the case of barley, for example, the temperature demand is 4°F to 5°F due to sowing barley. Based on the average monthly temperatures predicted by this model, we recommend that farmers seed barley in January of 2021 and 2022.

The model shows the inherent periodicity of the monthly average temperature over a long period of time. Based on the predicted value by our model, the temperature follows an upward trend on a yearly time scale. In fact, due to the greenhouse effect, local temperature increases over time. However, this upward trend cannot be visualized and predicted in our model since the small amount of data we have collected.

Comparing the predicted temperature with the true temperature value, we find that some of these predictions are not particularly accurate. This is a limitation of the predictive temperature model. Although our model have a high accuracy, it is still unable to predict the precision of future weather changes. Since the model is making predictions of temperatures at future points in time through the patterns of past data.

References

NOAA. (n.d.). “Climate Data - New York State.” Kaggle.