Time Series Analysis of Chemical Lab Data

Katherine Schmitzer
University of California, Santa Barbara

PSTAT 274
Final Project
June 12, 2019

Abstract

Science is an ever-expanding and increasingly important industrial field. Scientific data analysis is used to find cures, solve crimes, as well as explain the natural phenomena of the world around us. The goal of this project is to utilize time series analysis to test whether future chemical process readings can be predicted based on past readings. For this bi-minutely chemical process data, an ARMA model is constructed, tested and used for the forecasting of 10 future values. In the course of obtaining such a model, several statistical and time series techniques were used including: data differencing and transforming to obtain a stationary series, diagnostic checking to ensure the residuals resemble white noise, periodograms for periodicity detection, confidence intervals for coefficient estimates, and prediction intervals for the predicted values.

The forecasted values obtained by using the constructed ARMA model were not exactly the same as the actual data values. However, nearly all of the values fall within the 95% prediction interval. Depending on the necessary scientific accuracy of the specific chemical experiment, time series analysis might not be the best option for forecasting. It might be helpful to have a range of plausible values, such as the prediction interval; but if the data is critical and predictions need to be exact, it might be better to perform the experiment and record the results as opposed to trying to predict values.

Introduction

This project aims to predict 10 future temperature values of a chemical process using techniques used in time series analysis as well as the functions and computational tools included in R. The data set used for this project was obtained from Rob Hyndman's Time Series Data Library (TSDL). This data titled "Chemical process readings every two minutes" was originally featured in Montgomery and Johnson's "Forecasting and Time Series Analysis" in 1976 (Montgomery 276). Data values were collected every two minutes of an unspecified chemical process and recorded in degrees Fahrenheit.

After exploratory data analysis, the time series data was transformed to create a stationary series to work with. Using ML estimation along with AICc minimization, 4 possible models were chosen. After diagnostic checking, a final ARMA model was chosen and used to forecast 10 data points ahead.

To test the accuracy of the predictions, the last ten values of the original data set were initially removed and not used in the creation of the ARMA model. After forecasting, the ten predicted values were compared to the ten original data set values. The forecasted values strayed from the actual data values, but the 95% prediction interval included all but one of the true data values.

Even though this data was collected half a century ago, the practice of data analysis on any type of scientific data is important. Whether or not this specific chemical process is still being tested today, the results and predictions of this data can be used as an example for future scientific research.

Exploratory Data Analysis

After removing points 91-100 of the original dataset, the model-building data set consists of 90 values. From plotting the data, it can be seen that there is likely no trend, and seemingly no seasonality. The values range from around 130 degrees to 185 degrees and appear to nearly resemble white noise as depicted in *Figure 1*.

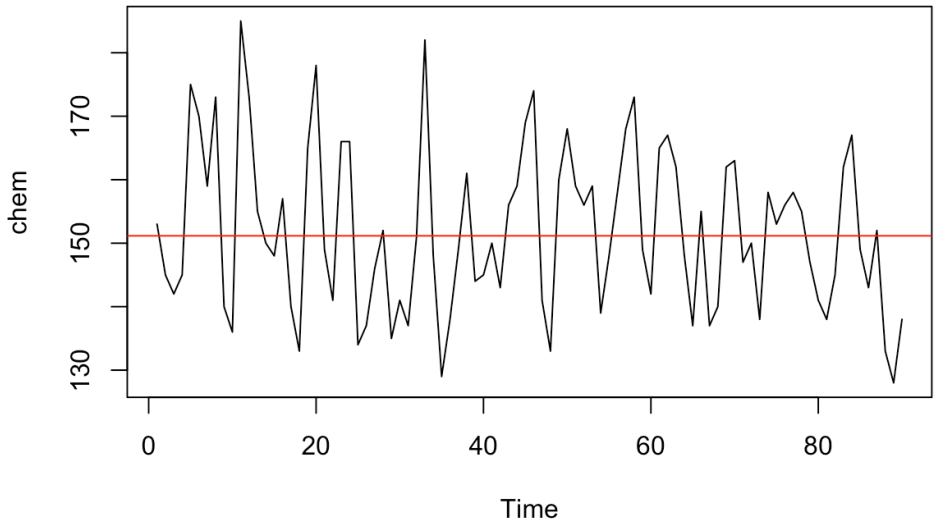


Figure 1. Chemical Process Readings Every Two Minutes in °Fahrenheit

Unlike white noise, however, there appears to be a funneling pattern. The data seems to have a higher variance in the first half of the set as opposed to a smaller variance in the last half of the set. This phenomenon might be able to be resolved through data transformation. There are no sharp changes in behavior, but the largest drop in temperature occurs between lags 33 and 34, from 182°F to 148°F. The mean of the data is 151.9°F and the variance is 167.6865°F.

The plot of ACF shows significant values at lags 1, 2, 13, and 38, *Figure 2*. The plot of PACF shows significant values at lags 1, 2, 20, *Figure 3*. In both of these graphs, the two most prominent, non-zero values are at lags 1 and 2. It is difficult to tell whether the lags cut off or tail off, which suggests that a mixed ARMA model might fit the data.

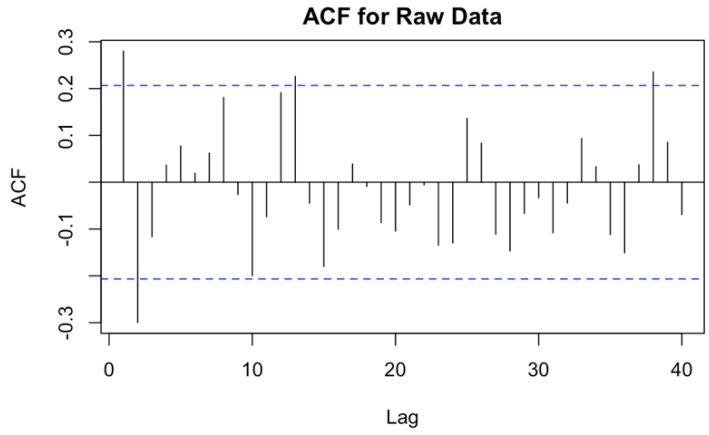


Figure 2. ACF of Raw Data

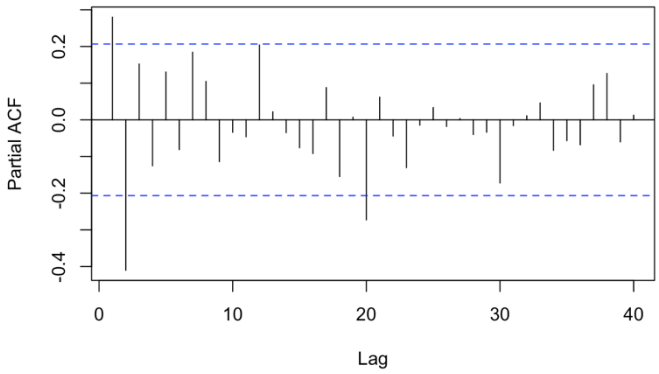


Figure 3. PACF of Raw Data

Data Transformation and Differencing

Since the mean and variance of the data are so large and the variance appears to be changing, a transformation of the data might be useful in helping the data become stationary. Using the boxcox() function in R, it can be seen in *Figure 4*, that there is a rather large confidence interval of lambda values for possible transformations. Included in the 95% confidence interval are $\lambda = -1,0,1$.

Before performing these transformations, we test for the "optimal" lambda value using the which() function. This shows that the optimal lambda value is -1.191919, which tells us that $\lambda = -1$ might be a better transformation than $\lambda = 0$.

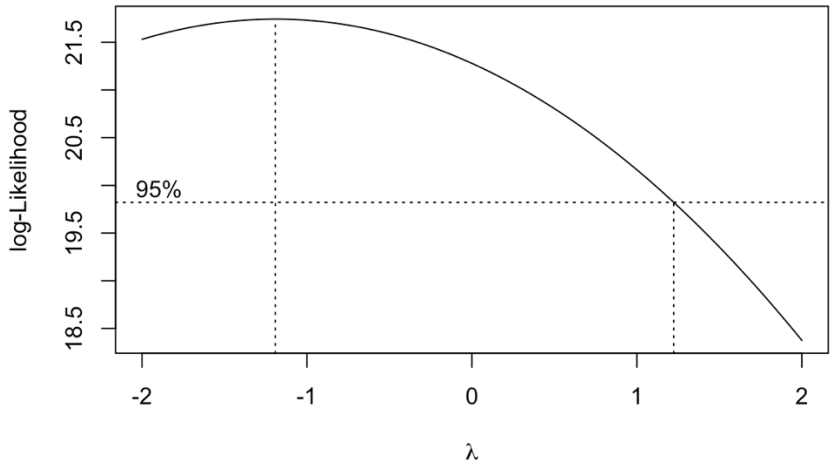


Figure 4. Box-Cox Transformation Test

After testing both lambda values, the following results are obtained and compared to the non-transformed data, *Table 1*.

	Raw Data $Y_t = X_t$ $\lambda = 1$	Log Data $Y_t = \log(X_t)$ $\lambda = 0$	$Y_t = \frac{X_t^{\lambda} - 1}{2}$ $\lambda = -1$
Mean	151.9	5.019678	0.9933702
Variance	167.6865	0.007126066	3.073924e-07
Significant ACF lags	1, 2, 13, 38	1, 2, 13, 38	1, 2, 20
Significant PACF lags	1, 2, 20	1, 2, 20	1, 2, 20

Table 1. Comparing Transformations

Both the log transformation and the $\lambda = -1$ transformation improved the ACF, PACF, and lowered both mean and variance. Because the $\lambda = -1$ transformation substantially lowers the variance to nearly 0 and there are less non-zero lags, this is chosen as the appropriate transformation for the model.

Moving forward, the transformed time series Y_t, pictured in *Figure 5*, will be used to select a model and complete diagnostic checking. During forecasting, the transformation back to X_t will be used to predict non-transformed values of the time series. The transformed times series Y_t shows no linear trend and appears to have constant variance over time, implying stationarity.

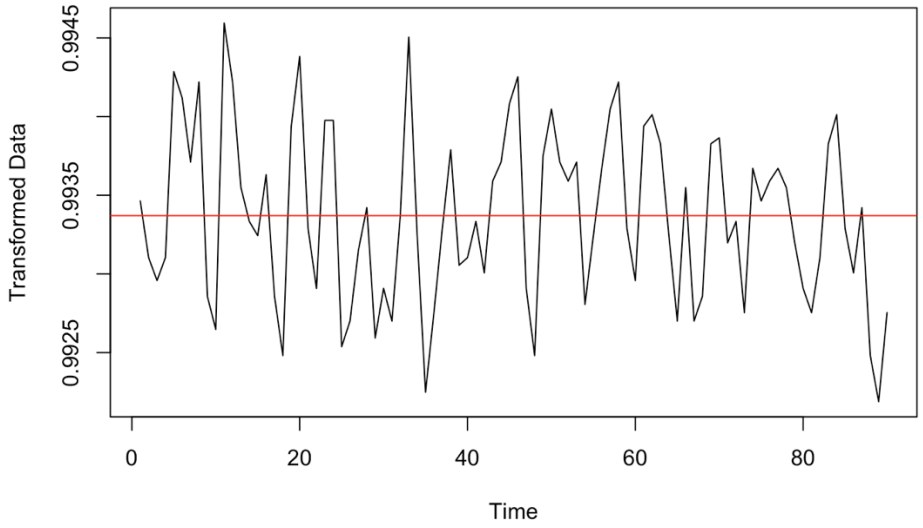


Figure 5. Transformed Time Series Data

While analyzing the original data, there did not appear to be a trend. After also looking at the plot of the transformed data in *Figure 5*, there still appears to be no trend. To be sure, we try differencing at lag 1 and inspect the analytics. After differencing, the variance increased slightly from $3.073924 * 10^{-7} to\ 4.363868 * 10^{-7}$. The graph of PACF, in *Figure 6*, also shows lags with non-zero values that did not appear before. With this, we conclude that the stationary time series that will be used to forecast values of this data set is just Y_t.

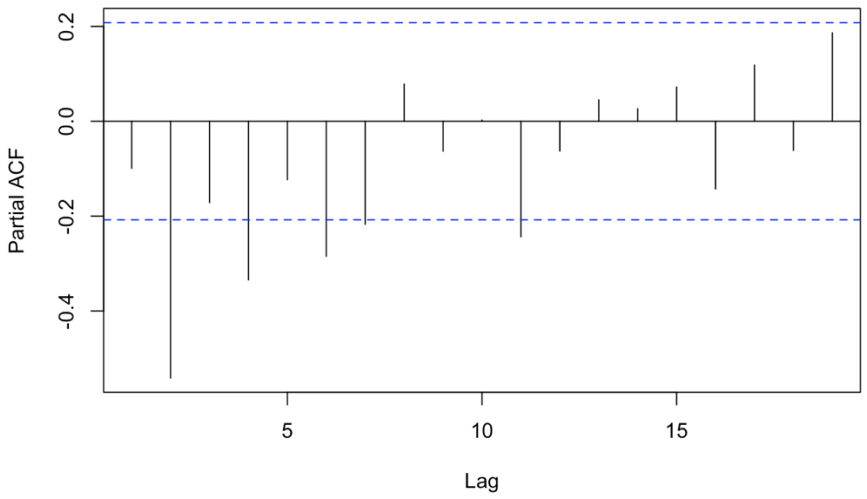


Figure 6. PACF of Transformed Time Series Data

Model Selection

As discussed in the Exploratory Data Analysis section, the ACF and PACF suggest that this data is best represented by a mixed ARMA model. In *Figure 2* and *Figure 3*, it is unclear if there is a point at which the values cut off or tail off. The ACF has sporadic non-zero values all the way past lag 38, and the PACF appears to initially tail off until lag 20, where there is a significant non-zero spike. Even after transformation, the behavior of the ACF and PACF does not strongly suggest a pure AR or pure MA model. So, our initial hypothesis, is that there will be both MA and AR parts of an appropriate fitted model.

At first, in the interest of parsimony, we use the ar() function to fit our transformed data to an AR model. In using both Yule-Walker and ML estimation, the ar() function found the best representative model to be the AR(2) models, depicted in *Table 2*.

Yule-Walker Estimation	ML Estimation
$Y_t = 0.4085Y_{t-1} - 0.4117Y_{t-2} + Z_t$	$Y_t = 0.4041Y_{t-1} - 0.4211Y_{t-2} + Z_t$

Table 2. Fitted AR(2) Models

Next, we use the arima() function in conjunction with AICc to see which ARMA models are suggested for this data using ML estimation. The models with the lowest AICc values, calculated using ML estimation are ARMA(1,3), ARMA(2,1) , and ARMA(0,3) with ARMA(1,3) having the lowest AICc value: -1115.87. There was only slight difference between each model's corresponding AICc value and diagnostic checking will need to be completed in order to determine which model is best.

In examining the estimates for the ARMA(0,3) model, the confidence interval for the MA(3) coefficient contained 0. So, an ARMA(0,2) model was tested. Similarly, when testing the ARMA(0,2) model, the MA(2) coefficient estimate's confidence interval contained 0. So, an ARMA(0,1) model was tested. Finally, the MA(1) coefficient estimate's confidence interval did not contain 0. So, instead of the ARMA(0,3) model, an ARMA(0,1) model was to be among those in model selection. These models are depicted in *Table 3*.

ARMA(1,3)	ARMA(2,1)	ARMA(0,1)
$Y_t = -0.6922Y_{t-1} + Z_t + 1.2634Z_{t-1} - 0.4177Z_{t-3}$	$Y_t = -0.2791Y_{t-2} + Z_t + 0.5145Z_{t-1}$	$Y_t = 0.6349Z_{t-1} + Z_t$
$(1 + 0.6922B)Y_t = (1 + 1.2634B - 0.4177B^3)Z_t$	$(1 + 0.2791B^2)Y_t = (1 + 0.5145B)Z_t$	$Y_t = (1 + 0.6349)Z_t$

Table 3. Fitted ARMA(p,q) Models

Additionally, we try using the auto.arima() function, available in R, to see what model is suggested. Using this function, a model that best represents the data is given as ARMA(1,3). This was expected, since the function decides the best model based on AICc values, and we already found that ARMA(1,3) had the lowest AICc value using the arima() function. Thus, the three models, depicted in *Table 3*, along with the AR(2) model will be analyzed and tested with a variety of diagnostic checks in order to determine which model best represents the data.

Model Diagnostic Checking

For each model, the first diagnostic check conducted was taking a look at the ACF and PACF plots of the residuals. If there are significant non-zero values at any lags other than 1 for ACF, that is a sign that the residuals do not resemble white noise. To further determine whether the residuals resemble white noise, Ljung-Box, Box-Pierce and McLeod-Li tests are performed. Furthermore, a Shapiro-Wilk test is conducted to test if the residuals are normally distributed. Then, histograms and Q-Q plots are graphed in order to display if how closely the residuals resemble a normal distribution. The results from each of these checks is listed below in *Table 4* and the corresponding histograms and Q-Q plots are pictured in *Figures 7-10*.

	AR(2)	ARMA(1,3)	ARMA(2,1)	ARMA(0,1)
Significant ACF Lags	8	N/A	8	8, 38
Significant PACF Lags	8	N/A	8	2, 8, 16
Ljung-Box Test p-value	0.3112	0.4941	0.3375	0.1494
Box-Pierce Test p-value	0.3777	0.5565	0.4131	0.1954
McLeod-Li Test p-value	0.1278	0.5914	0.01664	0.01084
Shapiro-Wilk Test p-value	0.9352	0.2828	0.8003	0.676

Table 4. Results of Diagnostic Checks

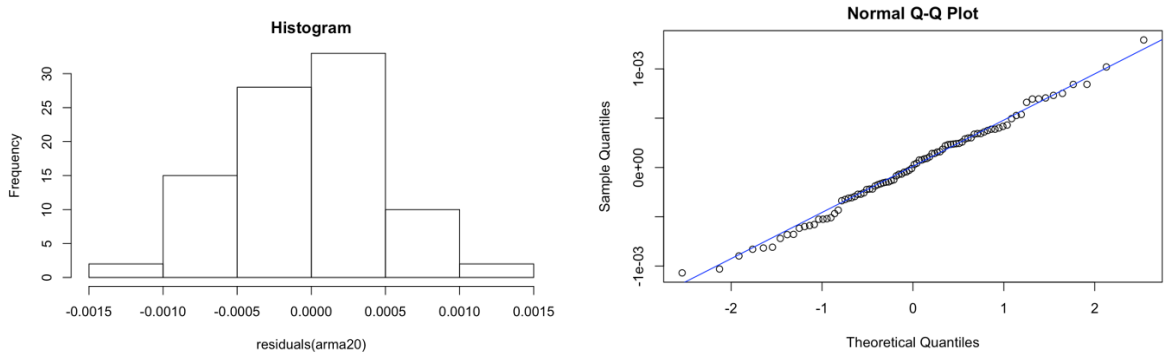


Figure 7. Histogram and Q-Q Plot of AR(2) Model

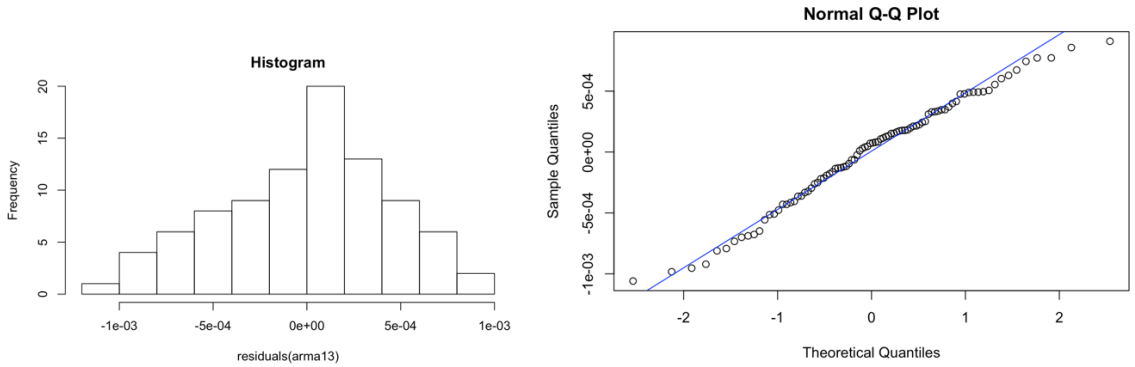


Figure 8. Histogram and Q-Q Plot of ARMA(1,3) Model

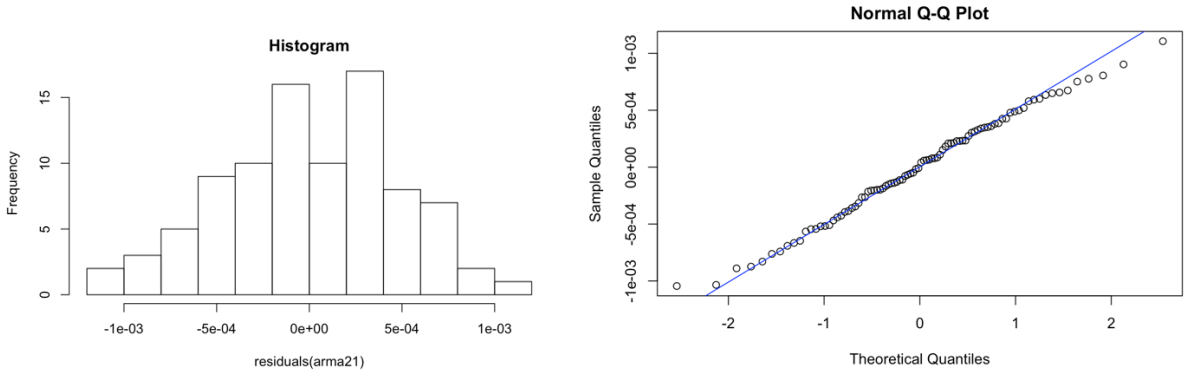


Figure 9. Histogram and Q-Q Plot of ARMA(2,1) Model

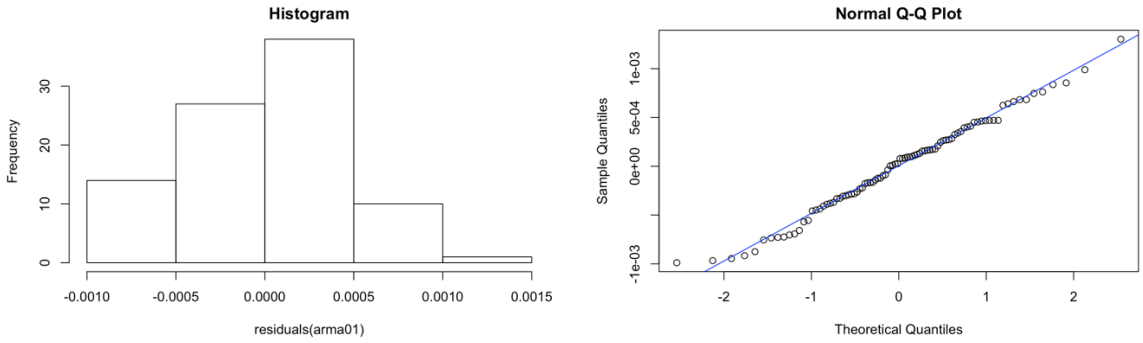


Figure 10. Histogram and Q-Q Plot of ARMA(0,1) Model

It is noticed that both ARMA(2,1) and ARMA(0,1) models do not pass the McLeod-Li Test, which tests whether the squares of the residuals are correlated. Furthermore, both have significant non-zero ACF and PACF values at one or more lags. Looking at the two models left, we notice that AR(2) has significant ACF and PACF values at lag 8, while the ARMA(1,3) model has all lags within the 95% confidence interval. For the AR(2) model, the Q-Q plot displays a more normal distribution and the Shapiro-Wilk test has a higher p-value than that of the ARMA(1,3) model. However, with only 90 data points, we are not as concerned with the data being or looking perfectly normal. The ARMA(1,3) model still passes the Shapiro Wilk test even though the histogram and Q-Q plot do not look as nice the one for the AR(2) model. The histogram for the ARMA(1,3) model is slightly skewed, but again, we are not as concerned with this since there are only 90 data points. Since the ARMA(1,3) model passes all the diagnostic checks while also having all values of ACF and PACF being within the 95% confidence interval, we choose this as the final model.

Forecasting

Using the ARMA(1,3) model, we first forecasted on the transformed data, which can be seen in *Figure 11*. The blue dots represent the 10 forecasted values and the orange dotted lines are the corresponding prediction intervals for the predictions.

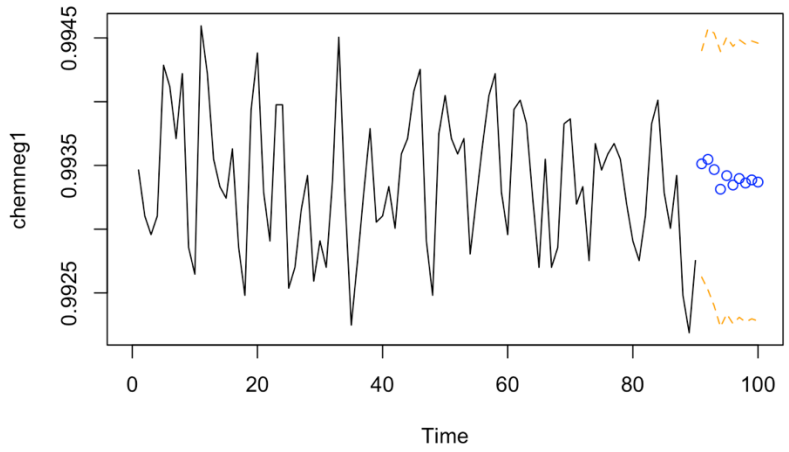


Figure 11. Forecasting on Transformed Data

Next, we transformed the data back to be the original data in order to forecast values of the raw data, producing the graph in *Figure 12*. Similarly, the red dots represent the 10 forecasted values and the orange dotted lines are the corresponding prediction intervals for the predictions.

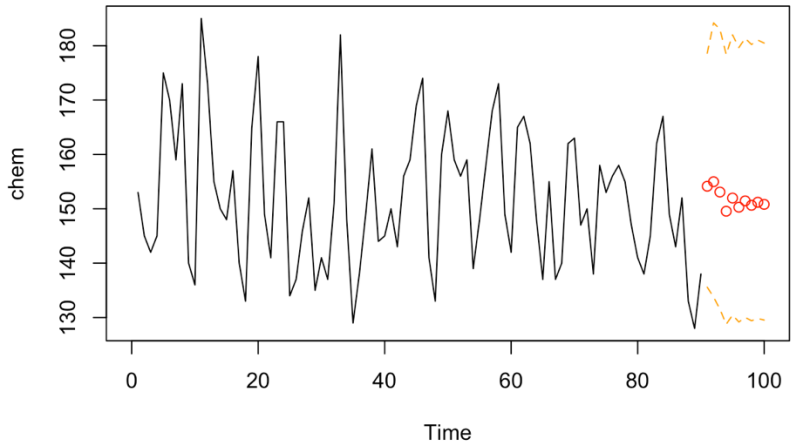


Figure 12. Forecasting on Raw Data

To test how accurate the results were, we compared the forecast on raw data to the original data including the points 91-100 pictured in *Figure 13*.

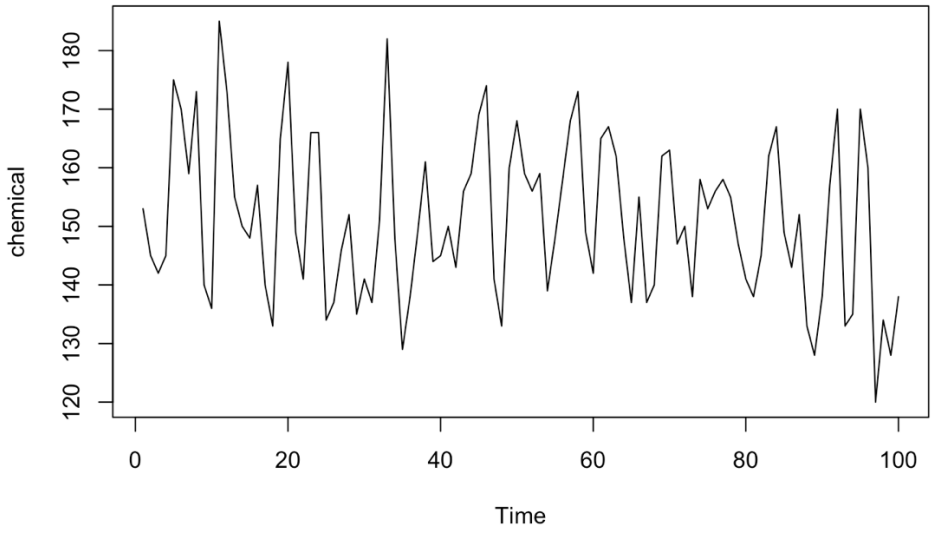


Figure 13. Original 100 Data Points of
Chemical Process Readings Every Two Minutes in °Fahrenheit

Comparing *Figure 12* and *Figure 13*, it can be seen that the predicted values are not the same as the actual data values. However, the actual data values are within the prediction interval for the predictions. Taking a closer look at the prediction interval, it can be seen that almost any of the first 90 given data values would fall in between the upper and lower bounds of the prediction interval. To display this, *Figure 14* shows the forecasts on raw data from along with the actual data values shown as black dots.

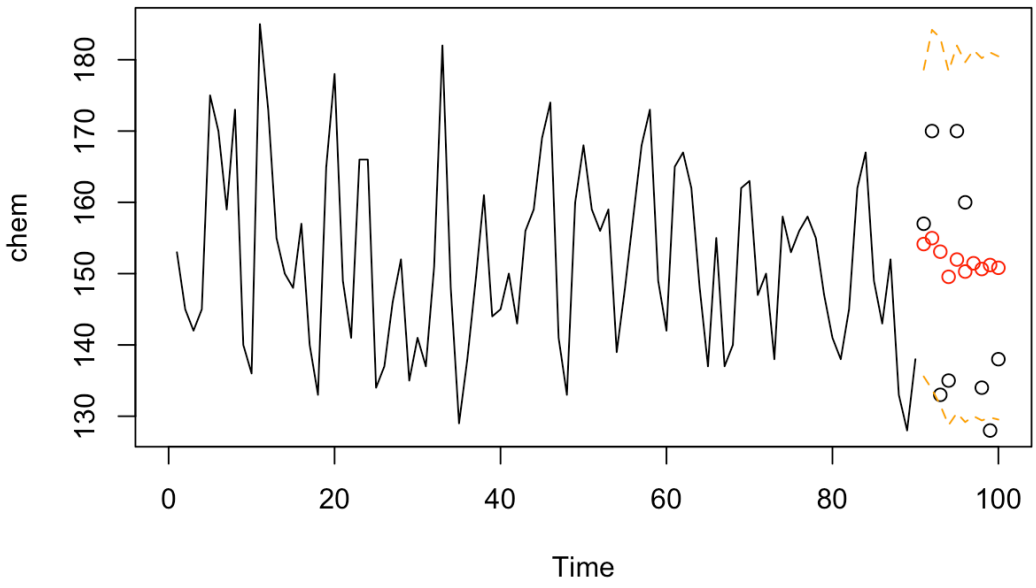


Figure 14. Forecast on Raw Data with Actual Data Points 91-100

Spectral Analysis

The final process of this project was to perform spectral analysis of the ARMA model chosen. This consists of three checks: the Kolmogorov-Smirnov Test on residuals, Fisher's Test on residuals, as well as producing a periodogram to determine frequencies of sin and cosine.

Firstly, conducting the Kolmogorov-Smirnov Test on residuals, it can be seen that the residuals of the ARMA(1,3) model resemble Gaussian white noise. All the values plotted fall within the values given by the cpgram() function, seen in *Figure 15*.

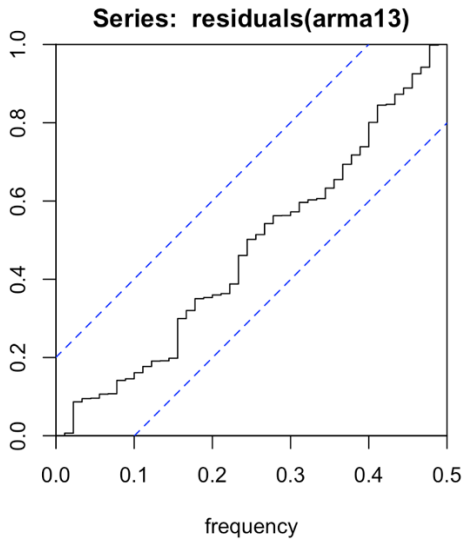
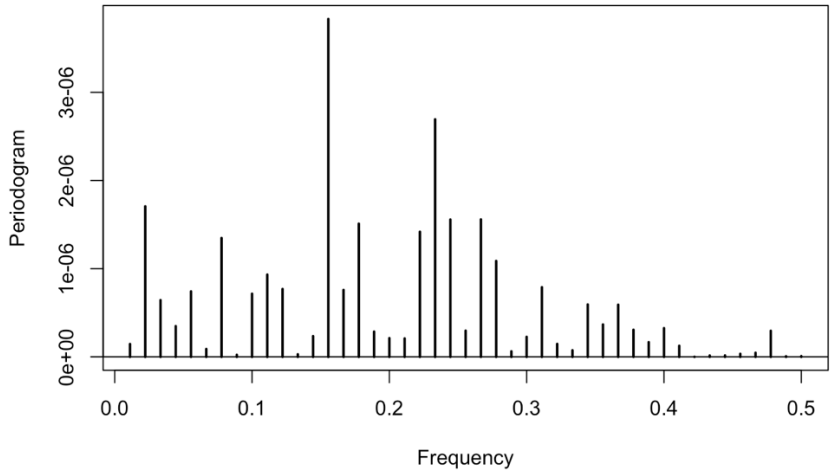


Figure 15. cpgram() - Kolmogorov-Smirnov Test on Residuals

Next, a Fisher's test was applied to the residuals of the ARMA(1,3) model. The resultant p-value was 0.6914845. Given that the p-value is greater than 0.05, we can assume with 95% confidence, that the residuals pass as Gaussian white noise.

Finally, a periodogram of the stationary data Y_t was executed. There was one significant frequency, 0.1555556. However, looking at the scale on the y-axis of the periodogram, the values are exponentially small and close to 0. This frequency would be the frequency of either the cosine or sine function used to represent Y_t. With such a small frequency value, it suggests that there is no periodicity, which is the same conclusion we came to during the data transformation and differencing step.



Conclusion

In conclusion, the values forecasted using the ARMA(1,3) model, written as $(1 + 0.6922B)Y_t = (1 + 1.2634B - 0.4177B^3)Z_t$ **,** strayed from the actual data values. While this model produced a prediction interval that included most of the actual data points, it is difficult to say whether or not this time series analysis was successful. The prediction interval ranged from around 130 to 180. Nearly all of the first 90 actual data values fall in this range, meaning it is less notable for the true data values to be in that same range. It is possible that there is another methodology besides Box-Jenkins that still allows for the utilization of times series analysis while simultaneously providing more accurate predictions.

Throughout the course of this project, I consulted with Catherine Miao, Zheng Jing, as well as Joseph Kinderman.

References

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2008). Introduction to time series analysis and forecasting. Hoboken, NJ: Wiley-Interscience.

Hyndmans, R. (n.d.). Time Series Data Library. Retrieved May 25, 2019.

Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting. Switzerland: Springer.

Appendix

R Code Attached.