STA 137

Project

Prof. Burman

Jared Yu (914640019) Danli Zhang (915011728)
3-16-2019

- **Introduction**

During this past quarter, various concepts and methods within the field of timeseries were covered. We had learned how to model timeseries data, along with different approaches to fitting an appropriate model for the purposes of forecasting. The goal of this project is to thoroughly analyze the dataset using concepts and methods that were learned over the course of the class. We will outline the different methods which we will utilize and explain their concepts along with their output for this dataset. In the end, we will identify a specific model which we feel is best suited for this specific dataset.

- **Materials and Methods**

In the dataset that we are given, we are to analyze a set of annual temperature anomalies for the northern hemisphere from the years 1850-2018. The source of the dataset is from Climate Research Center, University of East Anglia, UK. The dataset that we are given is indexed by years, the points in the data are indexed by a time order, and therefore the data is a timeseries dataset. Since we are using a timeseries dataset, it is therefore appropriate to use a specific set of methods which are designed to best handle this type of data.

To better understand the dataset, a basic start was to first plot the data. As a timeseries dataset, we are interested in modeling the data until it appears stationary. We would like to interpret the following model:

$$Y_t = m_t + s_t + X_t$$

where Y_t is the observed data at time t, m_t is the trend component, s_t is the seasonal component, and X_t is the rough component. It is also possible that there is no seasonal component, or even that the trend is missing also. However, if we would like to model the data, it is important to identify the rough component so that important values such as estimates can be made. When looking at the raw data, it was clear that there was an upward trend. Stationary data should fluctuate about a constant value μ, the size of the fluctuations should remain constant, and the correlations are the same for all possible time lags. An example of stationary data is white noise, which can be expressed as $\varepsilon_t \sim WN(0, \sigma^2)$. It fluctuates around 0 and has a constant variance. The rules for stationarity can be summed as:

$$E(X_t) = \mu, \text{for all } t$$

$$Cov(X_t, X_{t+j}) \text{ is the same for all } t \text{ (for any nonnegative integer } j)$$

Due to the upward trend of the data (shown later) along with different degrees of variance along the x-axis, it is visually clear that the above properties for weak stationarity are not met (there is a stricter form called 'strong stationarity,' but it will not be discussed). Therefore, to transform the data for the purpose of identifying the important rough component, we must first at least detrend the data. It is being noted that the seasonal component will be ignored for this project. The reason is that seasonality is something which can be modeled much more simply if the time series is represents something cyclical (e.g., twelve months per year, quarterly business data, etc.). However, in this dataset, the time is ordered by years, something we don't know how to decompose into seasonal components for the purpose of analyzing annual temperature anomalies. Also, looking visually at the detrended data (shown later), there were no obvious cycles, a feature exhibited by data with an obvious seasonal component.

Before moving on towards discussing the process of detrending the data, we will first discuss the issue of transforming the data. There is a method called Box-Cox transformation, where the observations Y_t is set to the power of λ, or Y_t^{λ}. The use of transformation can help with stabilizing the variance within a set of observations. Later when analyzing a plot of the raw data, it will become apparent that there are problems with constant variance within the data. To overcome this issue, we have considered using Box-Cox to help with the issue. However, the resulting value for λ from using the forecast package in R was extremely close to 1, leading to only minor changes in the data. Also, the preliminary diagnostics were not as good after using the transformed data. For this reason, no transformations were done on the data.

It is also possible to use a simple line plot of the raw observations to observe the movement of the data across time. Looking at the raw observations can show the existence of a trend, cyclical behavior, and the variance of the observations over time. By analyzing the raw observations, we may gain a better perspective of what sort of modeling will be necessary in order to obtain a rough of the model which appears stationary.

Out of the three methods for identifying a trend that have been covered in class: polynomial fit, Loess fit, and two-sided moving average, the Loess fit has been considered the best amongst them for identifying the trend. Also, the guidelines of the report had identified Loess as the choice for if we decided it was necessary to detrend the data. The Loess (locally weighted polynomial regression) uses linear regression within a sliding time window (i.e., we regress Y_s on s when $s = t - q, \cdots, t + q$). Then we have a fitted value $\widehat{m}_t$ for each time point t. The choice of q should be a fraction, and in this case, we will use 0.25, like in a previous assignment where there was a similar number of observations (i.e., 138 in the previous assignment, 169 in the project). When trying to use different spans, the R^2 was used as a metric, which leads to results which are larger than 1 when the span is increasingly small. Therefore, this method was deemed unusable for the purpose of choosing a proper span. Instead, it was said to 'eyeball,' or just use judgement based on the appearance of the rough to determine a span. Starting with 0.25 as the initial span, lower or increasing the span lead to little change, that is unless we choose to make it into a span so large that it would be inappropriate considering the size of the data (e.g., span of 0.5 implies that we are using about 42 neighbors, which is too large for a dataset of only 169 observations). Therefore, the span was left as 0.25.

After identifying the rough through detrending, it is possible to then examine it using various diagnostics. The diagnostics for analyzing the rough that we will be using in the project are as follows: histogram, Normal Q-Q plot, Shapiro-Wilk test, and Ljung-Box test. We will briefly explain the concepts behind these diagnostics below. Many of these tests are concerned with the issue of normality. Like linear regression where we would like the error term to be normally distributed, we would like the same results for the rough within time series analysis. The reason is mainly due to the benefit of performing predictions off a rough that is normally distributed.

We can use a histogram to examine if the distribution of the rough appears like the bell curve of a normal distribution. If the rough appears similar in shape to the normal distribution, this will help with the notion that the rough is normally distributed. The closer that the histogram appears to the normal distribution, the better.

We can also use a Normal Q-Q plot to compare the theoretical quantiles with those of the rough. The quantiles divide the data into even sizes and compares the sample quantiles with those of a theoretical distribution. If the quantiles are similar, then the points should follow closely along the diagonal. Any departure from the diagonal is evidence that the data is instead skewed or exhibiting some other behavior. We can also inspect the correlation of the sample quantiles with those of the theoretical. Higher correlation between the two quantiles can indicate that the data is closely following a theoretical normal distribution.

Lastly, we may use the Shapiro-Wilk test to test the null hypothesis that the data, in this case the rough, is normally distributed. All these tests are important for the normality assumption. Like in linear regression where we are looking for the error term to be normally distributed, having the same result for the rough would benefit for the purpose of methods such as prediction.

Another diagnostic tool for analyzing the rough is to use the Ljung-Box text. We may be interested in finding out whether the sequence is independent and identically distributed, or i.i.d. The Ljung-Box test will test the null hypothesis that $\rho(1) = \cdots = \rho(h) = 0$ against the alternative that at least one of $\rho(1), \cdots, \rho(h)$ is nonzero. The purpose is like looking at the plot of the ACF, to analyze whether the autocorrelations fall within the confidence interval up to lag j, but in this case it would be lag h. If we fail to reject the null hypothesis, then we have more evidence that we have properly identified a stationary rough, and no further model fitting is required.

We would like to next examine the Loess residuals to understand what sort of $ARIMA(p, d, q)$ model it may possibly follow. The reason is that the Loess method may not give an ideal rough but may take us a step closer to finding a model which will help to generate a rough which appears more like stationary data. However, if we can determine a proper $ARIMA(p, d, q)$ model for the data which is not currently stationary, it can come closer to data which is stationary. We will use autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to help understand the process.

The ACF plot is meant to help in identifying the proper q in an $MA(q)$ model. The PACF plot is for identifying the correct p in an $AR(p)$ model. The combination of both can help in identifying an $ARMA(p, q)$ series. If the lags 'tail off' in either, it will indicate that the other model is more suited. However, if both 'tail off,' rather than cutting off so that all tails fall within the confidence interval after a certain lag, it may indicate that the data follows an $ARMA(p, q)$ model. However, these results are not always simple to interpret. In addition to using the ACF and PACF plots to identify a possible range of models from which the rough can be further modeled, we will use the $AICc$ criterion to select a single optimal model, looking for the model with the smallest value for $AICc$. This criterion has been used in the past in the class for selecting models and it was also recommended in the project outline.

The ACF shows the different autocorrelations for lags up to order j. The autocorrelation is denoted as $\rho(j)$ and is a function of the autocovariance $\gamma(j)/\gamma(0)$. The autocovariance $\gamma(j)$ describes the covariance $Cov(X_t, X_{t+j})$ in the time series. Ideally, we would like all the lags in the ACF plot to be within the confidence interval bands. This is equivalent to seeing if the $\hat{\rho}(j)$ are inside or outside the range $\pm 1.96/\sqrt{n}$. If all the tails fall within the confidence intervals, like the behavior of white noise, this would be a sign that the data is stationary. The PACF plot uses a slightly more complicated technique for finding lags. It uses forecasting, backcasting, best linear predictor, and other methods to calculate the PACF. However, the way for interpreting is similar, except that the PACF applies to the $AR(p)$ model.

In addition, we are interested also in looking at the Loess residuals and final model residuals using spectral analysis. Spectral analysis is based on a mathematical formula which says that a stationary series can be represented by a linear combination of sines and cosines. Below is the formula,

$$X_t \approx \sum_{j=1}^{m} \{A_j \cos(2\pi w_j t) + B_j \sin(2\pi w_j t)\}, \quad (1)$$

where m is large, $w_j's$ are distinct frequencies, $\{A_j\}$ and $\{B_j\}$ are random with zero means and $Var(A_j) = Var(B_j) = \sigma_j^2$. Both $\{A_j\}$ and $\{B_j\}$ are uncorrelated. If X_t is in fact stationary, then the spectral density function of it, or f, can be seen below,

$$f(w) = \sum_{h=-\infty}^{\infty} \gamma(h) \exp(-2\pi i w h) = \gamma(0) + 2\sum_{h=1}^{\infty} \gamma(h) \cos(2\pi w h)$$

where $-\frac{1}{2} \leq w \leq \frac{1}{2}$.

Also, there is the periodogram, which is like an estimated version of the spectral density function. The above values for $\{A_j\}$ and $\{B_j\}$ from (1) can be using least squares. The exact formula will not be given, but the least squares estimate or scaled periodogram has the following formula,

$$P(j/n) = \hat{A}_j^2 + \hat{B}_j^2.$$

While the rescaled version has the following formula,

$$I(j/n) = (n/4)P(j/n),$$

and it is referred to as the periodogram. This will be used to estimate the spectral density function $f\left(\frac{j}{n}\right)$, where $w = \frac{j}{n}$. The spectral density function itself can be thought of as akin to modeling theoretical or population results. We can create the spectral density function by giving R a specific set of coefficients and a variance. The periodogram as an estimate, however, can be generated from something such as the residuals from our Loess model. We are interested in seeing which frequencies have the highest spectrum. The reason is that these frequencies contribute most towards the overall variance. This can be expressed in the following formula,

$$Var(X_t) \approx \sigma_1^2 + \cdots + \sigma_m^2.$$

The contribution of variability at frequency w_j is σ_j^2. A goal in mind is to have a smoothed periodogram which results in a flat appearance, where no frequencies have distinctly high spectrums. This would indicate that we have found a rough is not only stationary but is akin to white noise. This would be beneficial for the purpose of prediction.

In addition, we will also be looking at smoothed periodograms. The smoothed periodogram will use modified Daniell's kernel to take a local moving average. The smoothed periodogram has theoretical properties which are favorable to that of the raw periodogram. The formula is as follows,

$$\hat{f}_j = \frac{1}{2k}\left[\left(\frac{1}{2}\right) I_{j-k} + I_{j-k+1} + \cdots + I_j + \cdots + I_{j+k-1} + \left(\frac{1}{2}\right) I_{j+k}\right].$$

We will also use the following formula for selecting the correct k,

$$Q_D(k) = \sum\left(I_j - \hat{f}_j\right)^2 + \frac{1}{2k}\sum I_j^2.$$

The criterion says that wherever Q_D is the smallest at $k = k_D$, then k_D is the number of neighbors to use for smoothing. However, it will become apparent later that a balance is required when considering the number of neighbors in relation to the value of Q_D.

The smoothed periodograms will be used to analyze the Loess residuals and the different model residuals. Additionally, we will overlap the spectral density of the final model with the rough that it was initially trying to fit, which are the Loess residuals. The process of analyzing the smoothed periodograms will allow us to see if the model we fit is producing a rough which has a more constant variance. The process of analyzing the spectral density with the Loess residuals will allow us to see how close the model we choose is to fitting the periodogram of the data itself.

Based upon the initial set of diagnostic tests, we will try to fit a preliminary model. This model will be used on the Loess residuals and generate a new set of residuals which will also require a set of diagnostic tests. After fitting and analyzing the preliminary model, we will also look towards generating a final model of the rough. The process for doing so will be explained next.

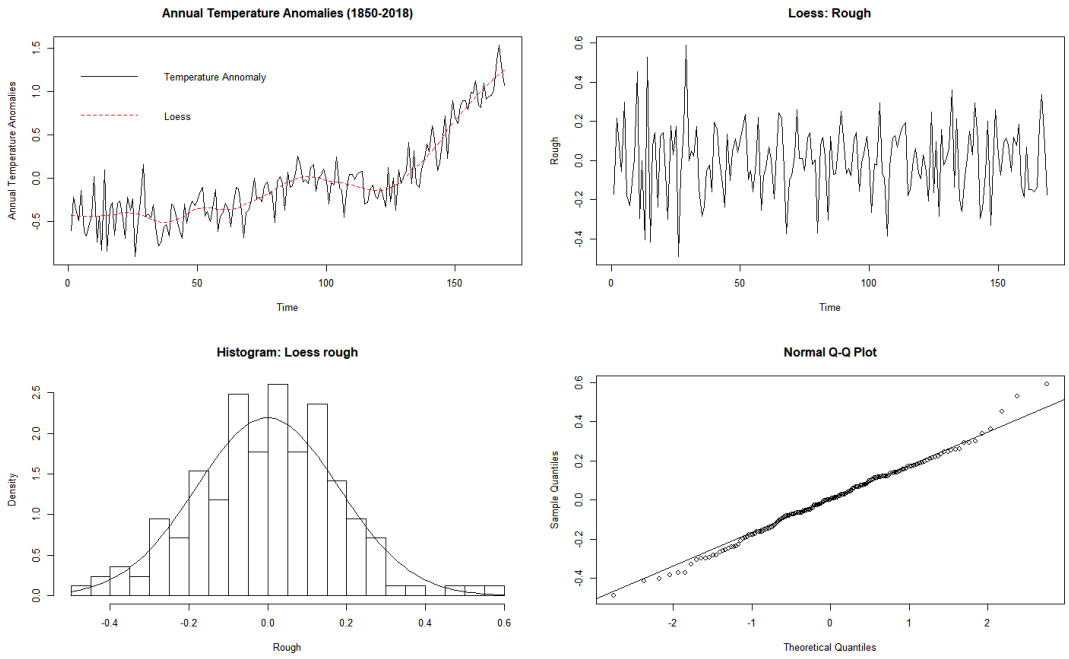
The teaching assistant had mentioned that we had the choice of using either `auto.arima()` or `sarima()` to determine which model is preferable. In the case of `auto.arima()`, the function will automatically try and determine which model is the best according to some criterion such as $AICc$. However, when using `sarima()`, each model must be fitted independently, and their respective $AICc$ values can be compared afterwards. For some reason which has yet to be determined, both methods give different outputs for the $AICc$. This results in both methods being incomparable. Therefore, it was mentioned that it would be fine to use whichever method makes the rough appear more ideal. The results of both methods will be discussed later during the analysis. The assignment had also mentioned it was only necessary to calculate 25 possible models, for each combination where $p = 0, \cdots 4$ and when $q = 0, \cdots, 4$. For that reason, only these models will be tested using the function `sarima()`.

Finally, we would like to try to forecast some data using our final model. Since we have fitted the model using Loess, we will be forecasting the values $\hat{m}_{n-5}, \cdots, \hat{m}_n$ using linear extrapolation and $\hat{X}_{n-5}, \cdots, \hat{X}_n$ using the R function `predict()`. The actual formula which we utilize will be given later, after the model has been determined. The process will require us to remove the last 6 observations and refit the final model to the subset of the data. Then, we will try to forecast the next 6 observations using the above-mentioned techniques. It is not possible to forecast too far out into the future using linear extrapolation, as there is a constant slope for the data based on the last two fitted values in the subset of the data. We will explain in more detail the process of forecasting with previous observations and adding linear extrapolation to determine certain parts of the equation during the analysis.

- **Results**

Initial Diagnostics

The first step in the analysis process was to plot the data. Below is a plot of the raw data along with the Loess fit (clockwise from the top left). Additionally, the rough of the model fitted with Loess is shown below. The other two plots show a Normal Q-Q plot and a histogram of the rough.



The first plot of the raw observations show that the data is trending upwards. The data is quite jagged, moving up and down repeatedly, sometimes with greater movement than others. It seems that in the beginning of the series, there is a larger variance and over time it begins to decrease. The variance doesn't seem to pick up quite as dramatically again until the end of the data, when there is a sharp spike upwards. This inconstant variance may be difficult to overcome during the modeling process.

The fitted values of the Loess model show that the line follows closely with the trend. The span chosen was 0.25. When trying to fit different spans the main method was to try and find a span which makes it such that the plot of the residuals (Loess: Rough) will appear as close as possible to stationary. Therefore, the 0.25 from previous assignments was left as the span, other lengths would cause the residuals to appear more erratic or have a higher variance. The plot of the Loess residuals looks quite like a normal rough plot that has been seen before in this class. The only concern is in the first few points, before time index 50, it is evident that the variance is slightly larger than in the rest of the plot. This may indicate that further modeling will be needed to change the rough later.

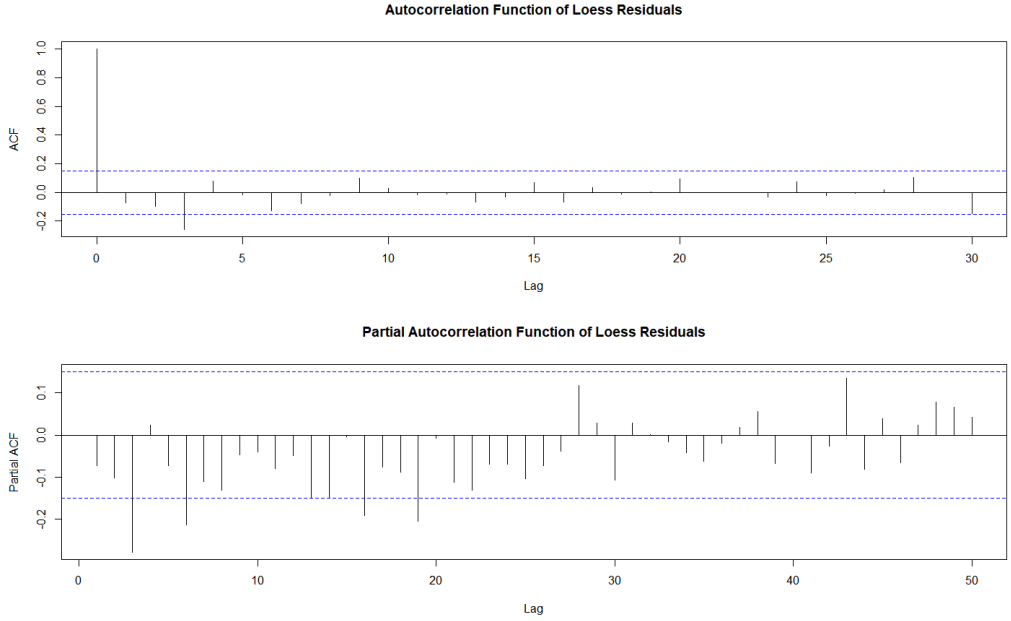
Next, we will move on towards an analysis of the rough from the Loess model. It is worth mentioning that a log model was also fitted. The data includes negative values, so the number 1 was added to the observations before taking the log. The resulting plot showed data which was more skewed than before, and so this method was ignored. Also, a Box-Cox transformation was considered, but due to the negative values it was not possible to use. Therefore, Box-Cox was also not used for the purpose of transforming the data.

The histogram has also been fitted with the curve of a normal density using the `dnorm()` function in R. Ideally, we would like the bars of the histogram to follow closely to the curve. The

histogram seems to show a slight skew towards the right. The distribution is heavier towards the left side in comparison. Also, there seems to be a higher density in the center, as evident through the tall bars which reach above the curve of the normal density. Regarding the normality assumption, we would like the rough to follow a distribution much more like a typical bell curve from the normal distribution. In this case, it may be such that there is not exactly normality, but it could be somewhat close.

The Normal Q-Q plot however seems to also show this skew to some extent. Ideally, the points should follow quite closely to the diagonal line. On the left half of the Normal Q-Q plot, the points seem to follow quite closely with this pattern. However, on the right side it is apparent that the points seem to depart from the line to some extent. This phenomenon takes place on the left and right side; however, it is much more noticeable on the right side. Also, the tail on the right side seems to have a wider spread and is also heavier in comparison to the left tail. The correlation between the x and y values of the points in the Normal Q-Q plot are still high, at 0.9959739.

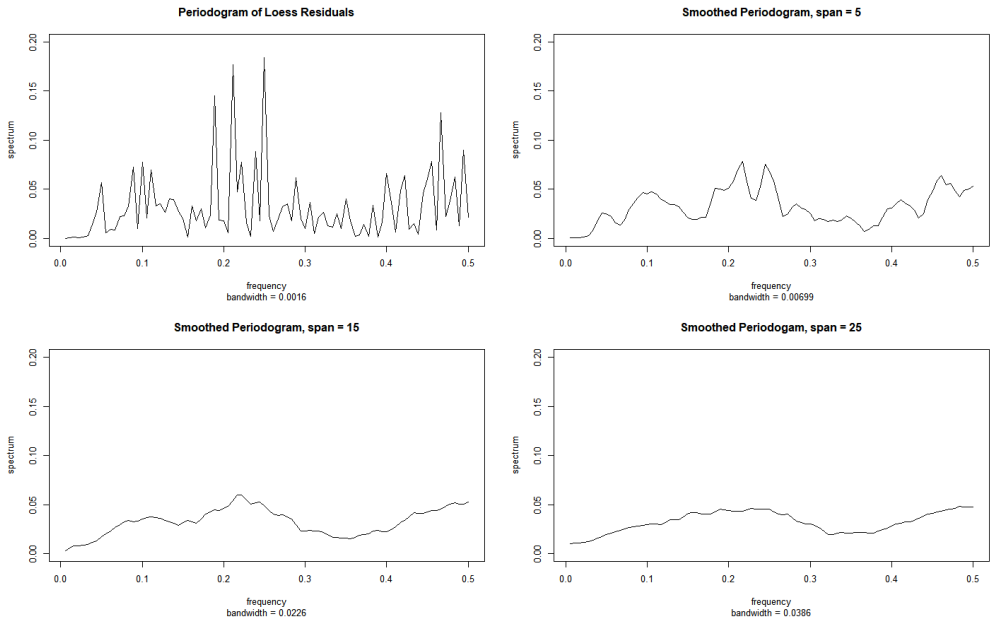
Another step was to use the Shapiro-Wilk test to test the null hypothesis that the data follows a normal distribution. The p-value for this test is 0.5294, so we fail to reject the null hypothesis at the 0.1 confidence level. From the test, we could then say that the data follows a normal distribution. The Ljung-Box test was also utilized as a diagnostic. The result of which is a p-value of 0.01966 at a lag of 10. Therefore, we reject the null hypothesis that the data is independently distribute. The conclusion then is that at least one of $\rho(1), \cdots, \rho(10)$ is nonzero at confidence level 0.05. This result from the Ljung-Box test indicates that the data is not yet stationary, and that further modeling is required.



Next, we will look at the ACF and PACF plots of the rough from the Loess model. The first plot shows the ACF, and it is apparent that the rough mostly falls within the confidence intervals. There seems to be one possibly significant tail at lag 3. This could imply that the data possibly has an $MA(3)$ component. The second PACF plot was plotted up to lag 50. It was apparent that there were still significant lags close to 30, and so further lags were added for the purpose of analysis. There seems to

be significant tails at lag 3, 6, 16, and 19. This could imply that there is an $AR(p)$ for any of the significant lags.

Next, we shall examine a set of periodograms for the Loess residuals. Below is a set of plots showing the raw periodogram and smoothed periodograms for the Loess residuals.

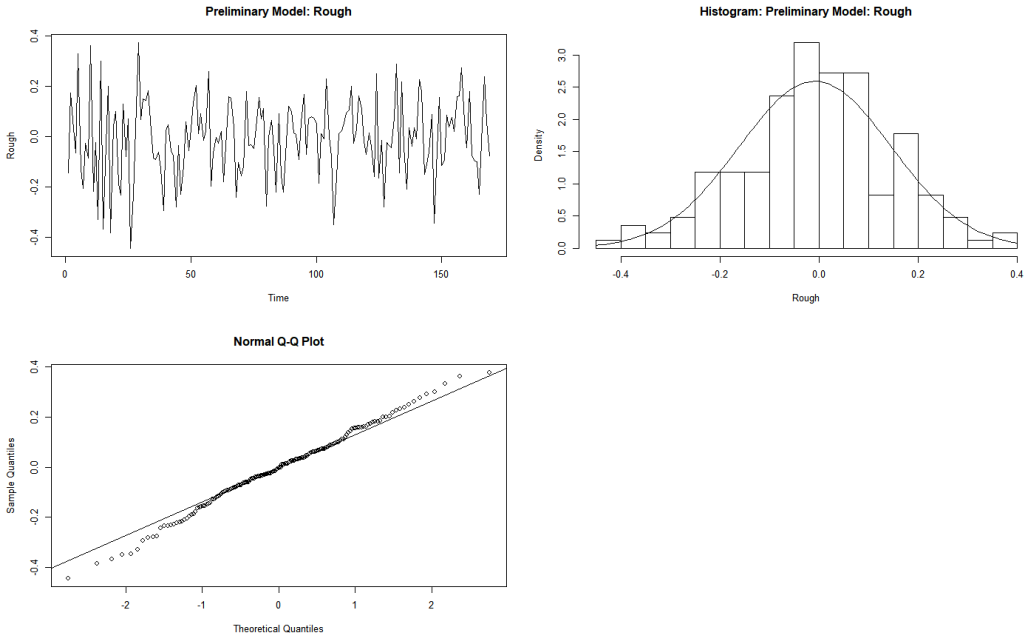


The raw periodogram shows a concentration of spikes around frequency 0.23. Around the edges there are also some spikes, they take place around frequency 0.1 and 0.47. This implies that the variance is greatest around these areas, most significantly in the middle around 0.23. Looking towards the smoothed periodograms, these concentrations of higher spectrums are still visible in spans 5 and 15. However, in span 25, the periodogram has become highly smoothed, and it is starting to look closer to a flat line. When smoothing the data, it becomes apparent that the range of the spectrum narrows down to a rather narrow region, roughly around 0.05 on the y-axis. This is quite close to 0.03598609, the residual variance from the Loess model. This means that the model is becoming close to the spectral density of white noise, and that we are close to stationarity. This concept will be elaborated on later. We are looking to further model the data such that the periodograms will appear flatter, removing the effect of any frequencies that contribute the most towards the overall variance.

Preliminary Modeling

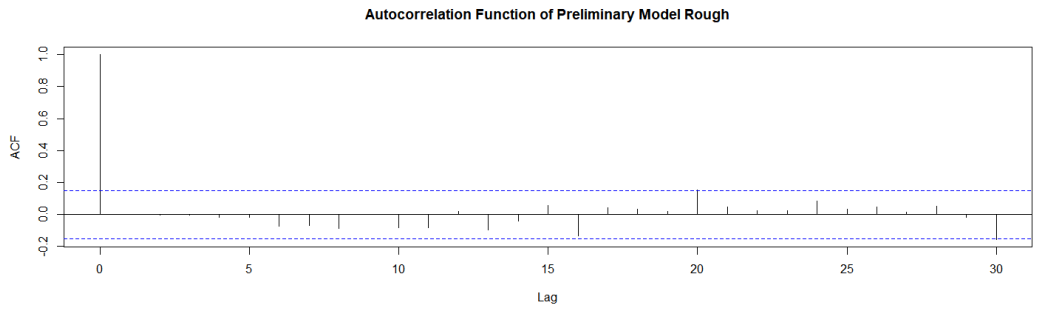
We have been asked to fit a preliminary model based upon our initial diagnostics of the Loess residuals. It has been decided that an $MA(3)$ component looked appropriate based upon the results of the ACF plot. Also, the PACF plot show several possibly significant tails, but $AR(6)$ seemed reasonable, as having a large p for $AR(p)$ seemed too complicated (e.g., trying to fit an $AR(16)$ or $AR(19)$ model). Also, it seemed from the other plots such as the normality tests and the periodograms that the Loess residuals were quite close to stationary, and only needed slight modification. An interesting note is that when fitting the model, an error came about which could be solved by changing the number of iterations to 1,000. The precise details are not completely understood, also the mathematics and programming for the process of fitting the $ARIMA$ model have not been emphasized in the class. Below

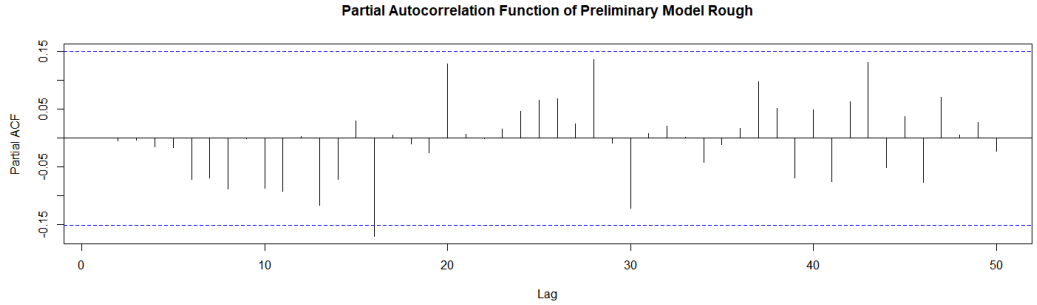
are some plots of the diagnostics from the preliminary $ARIMA(3,0,6)$ model. The label 'Preliminary Model' will be a reference to this specific $ARIMA(3,0,6)$ model.



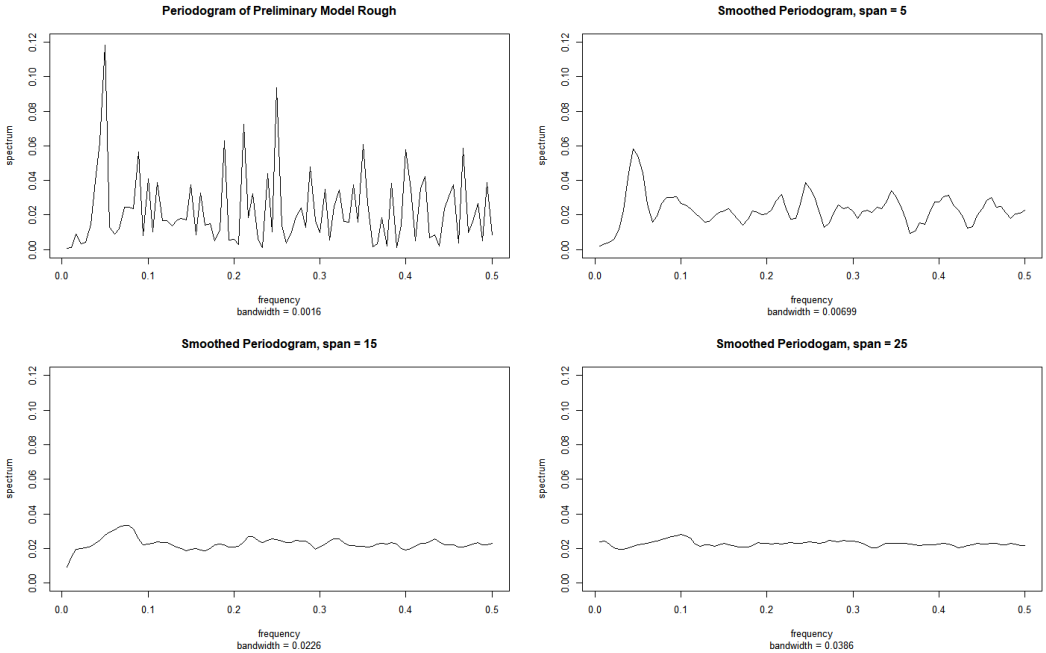
The first set of plots look like the first set during the initial diagnostics. The rough looks somewhat different from before, but it still suffers from a similar issue where there was a larger variance in the beginning. This results in a problem with the requirement that stationary series have equal variance throughout the data. The histogram improved in terms of the balance of the distribution. It appears less skewed than before, and so it looks closer to the bell-curve of the normal distribution. A possible issue is that the histogram has some gaps underneath the curve, this could be a sign that the distribution is still not quite that of the normal distribution. It may be the case this time that there is a higher density around the center than what is ideal. The Normal Q-Q plot looks slightly more balanced, as previously it seemed heavier on the right side. Now, it seems that the points have spread more evenly so the problem of imbalance doesn't look as severe as before. The covariance is 0.9974688, a slight improvement over before.

The p-value for the Shapiro-Wilk test and Box-Ljung test are 0.7539 and 0.9315. The Shapiro-Wilk test has a higher p-value than before, indicating that there is an improvement in the normality of the preliminary modeling. The conclusion is the same, where we fail to reject the null hypothesis that the data comes from a normal distribution. The Box-Ljung test however has improved dramatically and this time we can fail to reject the null hypothesis that $\rho(1) = \cdots = \rho(10) = 0$. Therefore, we can conclude that the rough in the preliminary model is independent and identically distributed at confidence level 0.1.





Looking towards the ACF and PACF plots, it is also apparent that there is an improvement over the initial Loess fit of the data. The ACF doesn't seem to have any significant tails up to lag 30. Previously there was a single significant tail at lag 3. The noticeable difference is in the PACF plot, where there is only one significant lag at $j = 16$. Previously, there were possibly 4 significant lags. This indicates that the $ARIMA(3,0,6)$ modeling has improved the rough, and the residuals are now closer towards stationarity. Next, we will move towards looking at the periodograms for the preliminary model.



The set of periodograms look noticeably different. Before there were three noticeable concentrations which are still evident, but their appearance is greatly reduced. Comparing their heights,

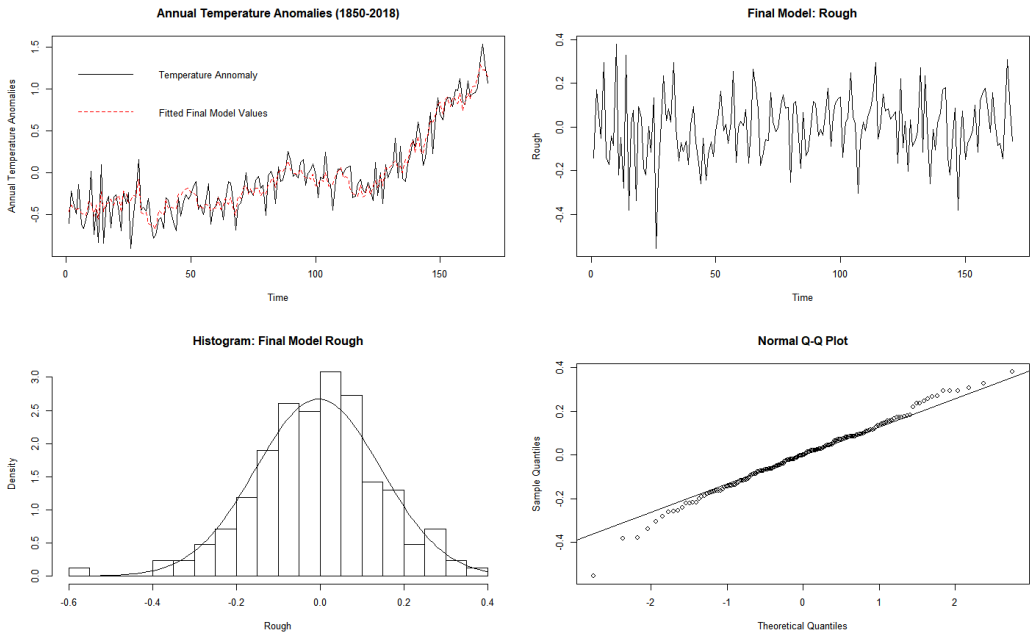
the periodograms in the preliminary model are shorter. Before, the largest spikes were in the middle, around frequency 0.2. This time however, the largest spike is in beginning, close to frequency 0.05. The smoothed periodograms are also much flatter than before. At span 5, the rough shape of the raw periodogram is evident, but there is much less definition in comparison to before. The raw periodogram has become less jagged, and so each of the smoothed periodograms have likewise become simpler in appearance.

The other two spans, 15 and 25, are quite close to being flat. They both have come close to looking quite like the spectral density of white noise. This result is positive, as it indicates that the model is closer than before to stationary. Before, the last two smoothed periodograms still exhibited more curvature and didn't appear quite as flat. In span 15, the concentration around frequency 0.05 is still evident, but it is almost gone in span 25. This helps with the idea that the variance is coming close to being constant throughout the data.

Final Model Selection

In the assignment, it was mentioned that we had the choice of fitting a model automatically using either `auto.arima()` or manually finding the $AICc$ using `sarima()` in R. Using `auto.arima()`, an $AR(3)$ model was determined to be the best. However, using `sarima()`, it was determined that the optimal model was an $ARMA(4,4)$ for the Loess residuals. When looking at the ACF and PACF plots for both models, the `sarima()` method proved to have an improvement over past results, while the plots for the `auto.arima()` showed little change. Therefore, we will continue with modeling based off the results of using the `sarima()` method.

The final model of the rough can be thought of as first detrending the raw observations to find the residuals or an initial rough. Then, we need to model these Loess residuals as an $ARIMA(4,0,4)$ process, which was previously determined based upon the $AICc$ criterion. Now we shall repeat the diagnostics that were performed initially at the start when only the Loess residuals were gathered. Below (clockwise from top left) is a plot of the fitted values with the observations, the rough from the final model, a Normal Q-Q plot, and a histogram of the rough of the final model.



Looking at the first plot of the fitted values with the observations, it is apparent that the fitted values of the final model has a much 'wigglier' shape. The fitted values seem to move up and down much more closely with the observed values. Previously, the fitted values of just the Loess model would have a smooth line fitting through the data. It makes sense however that the final model fits much closer to the observations. The reason being that the final model is more complicated, adding an additional $ARIMA(4,0,4)$ process to the modeling.

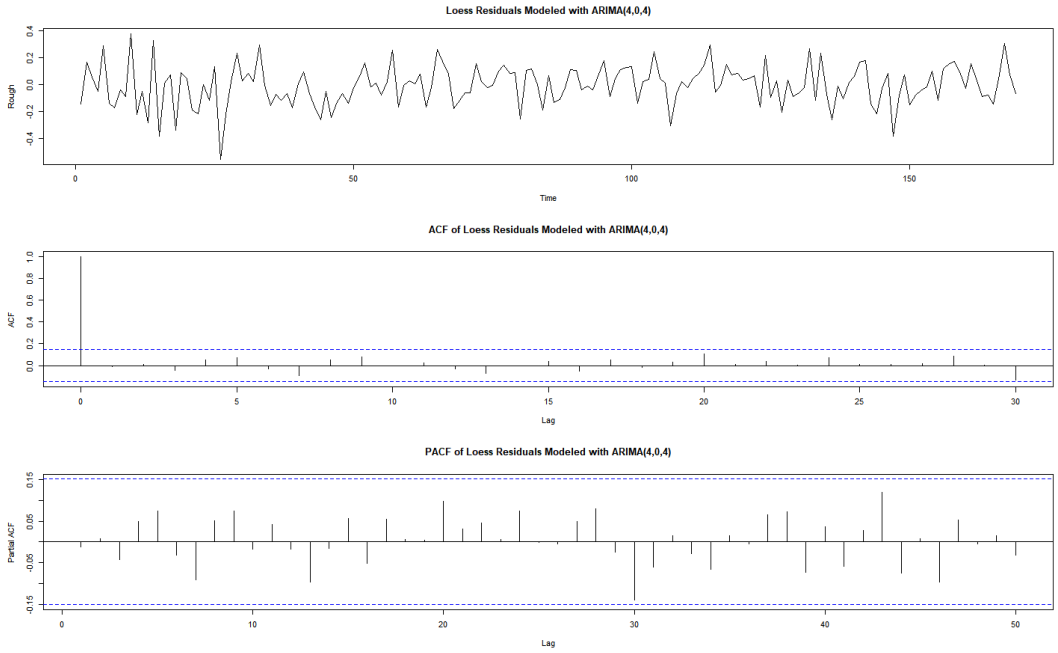
The rough of the final model looks to be an improvement over the original plot of the rough from just the Loess residuals. Previously (also in the preliminary model), it was apparent that there was a significant larger variation in the initial observations, prior to time index 50. The problem still exists in the plot of the rough of the final model, however the issue of fluctuating about a mean and constant variance appears more balanced in this rough. The initial large variance in the beginning also seems slightly more exaggerated in comparison to the other parts of the rough which have become more horizontal.

Looking at the histogram of the rough for the final model, it seems that there still may be some problems with the rough being normally distributed. Ideally, the rough would follow the normal bell-shaped curve. However, there still seems to be some uneven shape of the distribution. The bars seem to follow quite close to that of the normal density curve this time and the appearance is better than any of the previous rough models. Although the appearance has become much more like that of a normal distribution, there is a left skew this time which was not noticed before. This may show that the rough has gotten closer towards a normal distribution, but it is still not perfectly like a normally distributed set of data. It is possible that this skewness is related to the high variance which is evident in the plot of the rough, as they look somewhat similar in terms of their degree of being different from the rest of the data.

We can next look to the Normal Q-Q plot, and it seems this time that the right half has become more balanced. Previously, it was viewed that the right side seemed to be suffering more from heavy

tails. However, this time with the rough of the final model, the left side seems to be less like normality in comparison to the right side. Along the left side, it is evident that the points depart somewhat from the diagonal. However, it does still appear from the plot that the data does closely resemble that of the theoretical quantiles throughout the center of the plot. The correlation between the sample and theoretical quantiles is 0. 9948472, slightly worse than both previous models. An explanation could be the apparent higher variance in the beginning rough observations. As the rest of the data smooths out, this initial area of outliers has a great effect on something such as correlation. Before the heavier tails seemed to balance each other out on the left and right side, but this time there is mostly only the heavy tail on the left side.

Additionally, the Shapiro-Wilk test and Ljung-Box test were repeated. The Shapiro-Wilk test came back with a p-value of 0.3436, so we fail to reject the null hypothesis at the 0.1 confidence level. The conclusion then is that the data comes from a normal distribution. The failure to reject is slightly lower in comparison to the first time and the second time it was done, however it is still plenty insignificant for the purposes of the test. The Ljung-Box test has a p-value of 0.8921, and so we fail to reject the null hypothesis that $\rho(1) = \cdots = \rho(10) = 0$ at the 0.1 confidence level. This implies that the rough for the final model is independent and identically distributed or i.i.d. In the initial diagnostic, we had rejected the null hypothesis, so this time the null hypothesis was not rejected. This is a sign that there was indeed an improvement in the rough in comparison to when it was just the residuals from the Loess model. The p-value is not as high as the preliminary model however, but it still insignificant. We will now move on with analyzing the ACF and PACF plots.



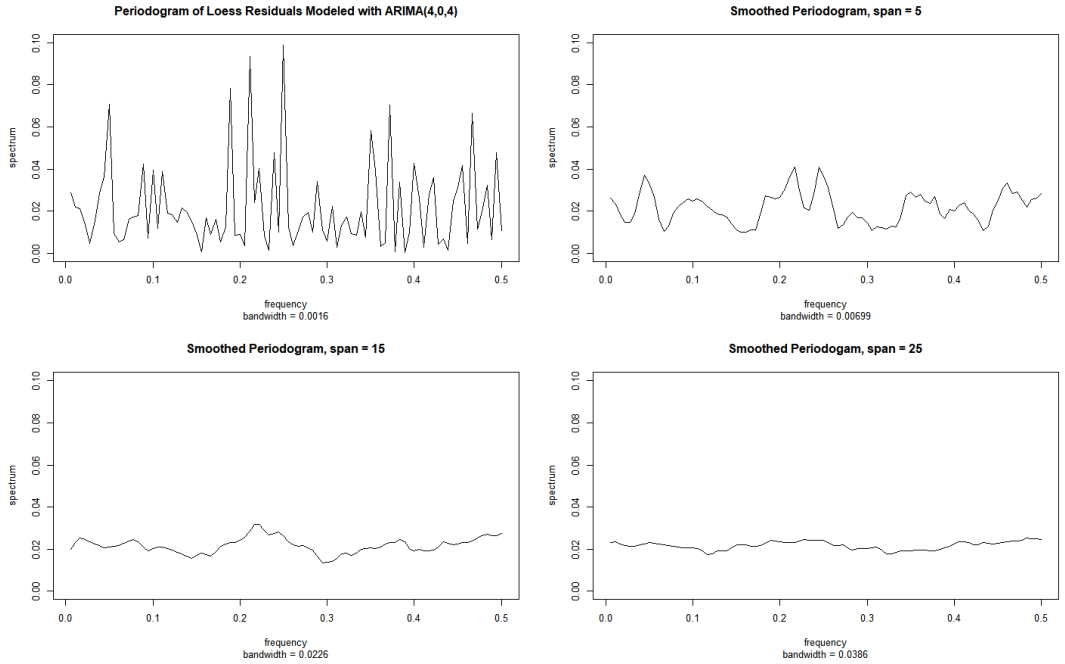
The resulting ACF and PACF plots show that all the tails fall within the confidence intervals. This is a good sign and it indicates that the rough has been properly modeled. It is an improvement over both the initial Loess model and the preliminary model. The implication is that we are now closer towards the

forecasting process. Below are the estimates and standard errors of the parameters from above along with their p-values:

	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
$Coefficients$	0.3829	0.4129	0.4531	-0.4948	-0.6346	-0.6407	-0.6776	0.9676
$Standard\ error$	0.0893	0.0867	0.0721	0.0738	0.0718	0.0414	0.0503	0.0752
$p-value$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

The $AICc$ for the model is -2.680218. Also, the estimated variance or $\hat{\sigma}^2$ is 0.02222. The fact that the p-values for all the coefficients are less than 0.0000 indicates that all the coefficients are important and don't need to be dropped from the model.

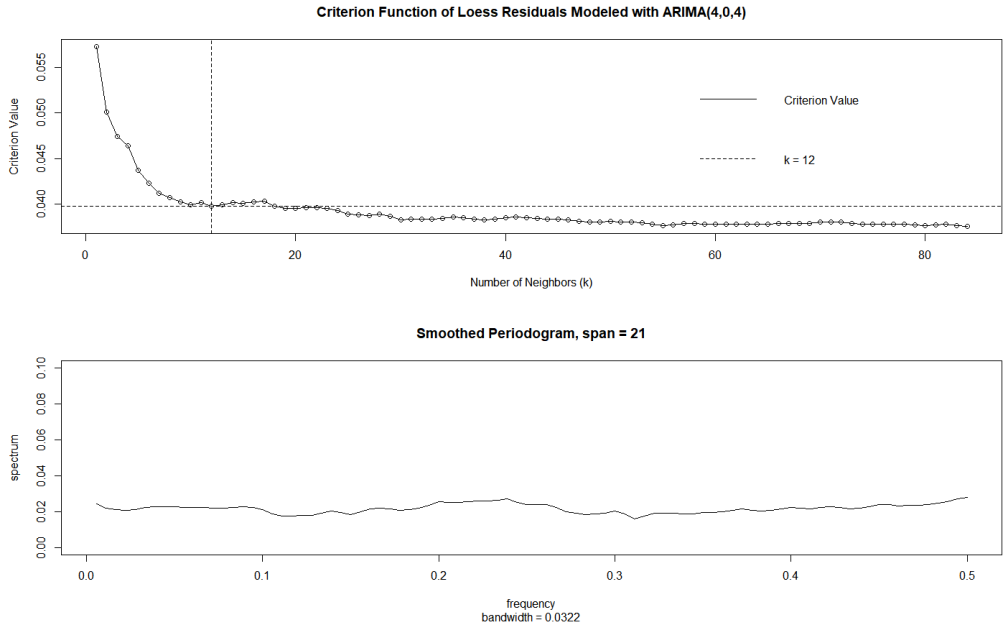
After having chosen a model, we will next move on towards analyzing the periodograms for the final model of the rough. Below is the raw periodogram, along with different smoothed periodograms of the Loess residuals after they have been modeled as an $ARIMA(4,0,4)$ series. The type of smoothing is the modified Daniell periodogram which is accomplished by using the `spec.pgram()` function.



The results look noticeably more like the initial model that was only detrended using Loess than the preliminary model. The top left plot shows the raw periodogram without any smoothing. The raw periodogram shows many jagged peaks. However, their overall height is lower than the previous two sets of periodograms for the Loess and preliminary models. This flattening is a good sign, as it indicates that each of the frequencies are having less of an effect on the overall variance, helping with the requirement of equal variance.

The span of 5 seems to show a rough pattern of what was seen before. This smoothed periodogram seems to show that the spikes in the middle around frequency 0.23 are somewhat distinct from the rest. The span of 15 is much flatter, and much of any pattern has been worn out. The most noticeable similarity is the spike in the spectrum towards the middle of the frequencies. The last span of 25 is quite flat, and much of the data seems to have been smoothed out. Next, we will move on towards

using a metric to analyze the smoothed periodograms. The flatness in the last two spans are like the preliminary model, but it could be said that the preliminary model became flatter than the final model. However, it could also be said that this model is an improvement over the preliminary model because the spectrums aren't as high as before.



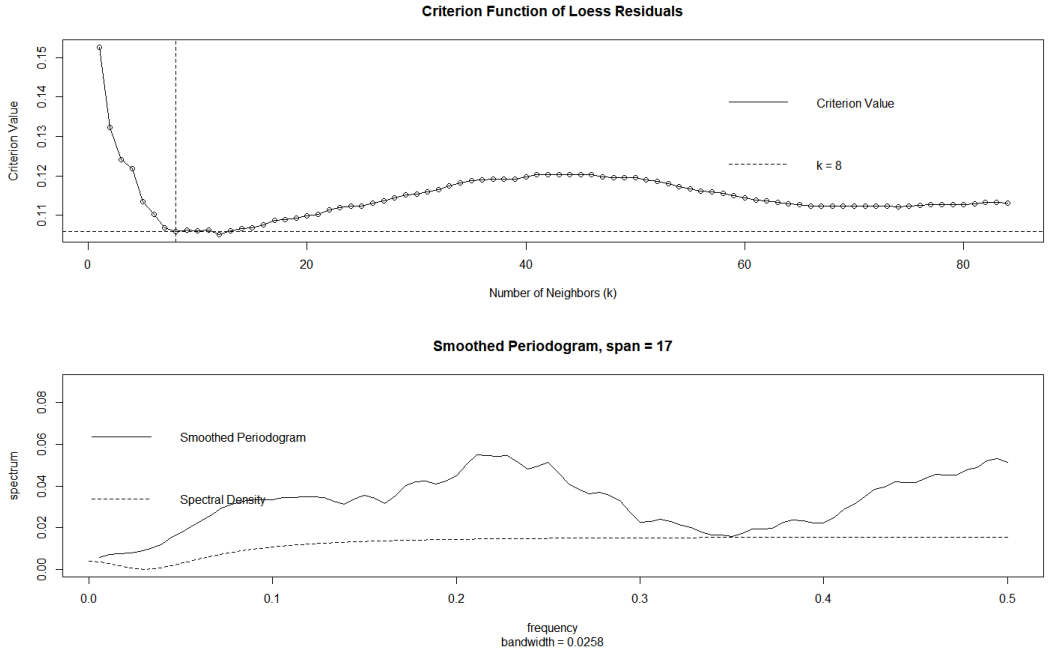
The top plot shows the criterion function which utilizes the `specselect()` function that was introduced to the class by the professor. The function can determine the optimal k within a certain range by using $Q_D(k)$, or the criterion for modified Daniell's smoothing. It is apparent when observing the plot which measures the criterion value for the number of neighbors, that the values become smaller as the number of neighbors become larger. In fact, the optimum k selected by the function is 84, which is the max possible span for the dataset given its size. Ideally, we want a criterion value which is small, but this will come at the sacrifice of making a span which is too large. It is evident that the numbers, although they get smaller, don't get much smaller than a certain range. So, we want to balance the lower criterion value with the number of neighbors. It seems that this can be done if we choose $k = 12$ neighbors, as the criterion value decreases rapidly, and then slows down significantly after. This would result in $span = 2 \times 12 + 1 = 25$. Before 20 numbers, there is a small cup and 12 is the number of neighbors for the minimum value in this cup. Two lines have been drawn on the criterion function plot and they show this point relative to the other points on the plot. It shows that there is another point to the left which is virtually the same criterion level, but with a shorter span. Therefore, we will go with this value which is $k = 10$ neighbors, or $span = 21$.

After smoothing with modified Daniell's method with a chosen span of 21, selected by the corresponding criterion for the same method, we can see a smoothed periodogram which appears quite flat. This is ideal, the reason is that if we think about the spectral density function, when looking at the spectral density of white noise we have the following result,

$$f(w) = \gamma(0) + 2 \times 0 = \gamma(0) = \sigma^2.$$

We can see that since white noise is independently distributed, the covariance of any lag other than 1 will result in 0. Therefore, the spectral density of white noise is equivalent to its variance, or σ^2. This value is a constant and appears as a flat line across the spectrum. This is not too different than what the smoothed periodogram looks like. The interpretation then is that the resulting smoothed periodogram of the model residuals resemble that of white noise, which is a good indication that the model that we have chosen is a matching fit to the data.

It is also important to note that the smoothed periodogram falls within a range that is quite close to 0.02222, the estimated variance that resulted from using `sarima()` to fit an $ARIMA(4,0,4)$ model on the Loess residuals. What this information regarding smoothed periodograms and spectral densities is saying that the final model of the rough that we have chosen closely appears like white noise through spectral analysis. White noise is stationary and having this property for our rough has been the intent as stated from the beginning. Now that we have successfully modeled the rough of the observations as something that appears like a stationary series, we can move towards doing some forecasting with the final model in mind.



Then, we will move on towards analyzing the fit of the chosen model using spectral analysis. The above plot shows first the criterion function used with the residuals from the data fitted with Loess. The optimal k chosen as the minimum of the dataset is at $k = 10$, however there is a similar criterion value to the left at $k = 8$. This is the index which is highlighted by the dotted vertical and horizontal lines. For this reason, the span of 17 was chosen to smooth the periodogram of the residuals from fitting Loess to the data.

In the second plot we can see the smoothed periodogram of the chosen span of the Loess residuals. This is shown as the solid black line which appears somewhat rough. This is the data which we were originally trying to model using $ARIMA(4,0,4)$. The smoothed periodogram didn't seem to have many distinct features, like the sunspot data from the textbook and handout. The sunspot data had a frequency with an incredibly noticeable peak which the spectral density was able to somewhat mimic.

However, in this data, there are only moderate groups of peaks which aren't nearly as distinct as the peak in the sunspot data. Therefore, observing the spectral density overlaid on top of the smoothed periodogram, we can see that the spectral density seems to be following a general trend and direction of the Loess residuals, without overly accounting for any highly unique characteristics that were seen before in the sunspot data.

Forecasting

Now that we have fitted a final model of the rough, along with analyzing the results using spectral analysis, we are interested in trying to forecast some data. In previous assignments, it was possible to utilize the `sarima.for()` function from the astsa package. However, considering that the data is modeled with Loess, it was not understood how to consider this step in the implementation of `sarima.for()`. Therefore, we have done some simple calculations in R to determine the results. This required that we utilized the `predict()` function and linear extrapolation. After taking a subset of the data to remove the last 6 observations, we refit the final model which had previously been determined.

We will first explain the mathematical process, and what we are trying to accomplish. We have ignored the seasonal component, and are left with the following model of the data:

$$Y_t = m_t + X_t$$

We have also modeled the trend using Loess, and so we have estimated $\widehat{m}_t$. The formula can then be thought of as follows:

$$\hat{Y}_t = \widehat{m}_t + \hat{X}_t$$

Since we have also modeled the detrended model as being an $ARIMA(4,0,4)$ process. The X_t part can be written as follows:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \phi_4 X_{t-4} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \theta_4 \varepsilon_{t-4},$$
$$t = 1, 2, \cdots, n-6$$

Here we are considering the forecast model, where we have removed the last 6 observations, where n is the number of observations in the original dataset. Our model currently has $ARIMA(p,d,q)$ components and we will obtain predictions for the rough using the `predict()` function in R. This is a simple way to find the next values of $\hat{X}_{n-5}, \cdots, \hat{X}_n$ in R.

The process of linear extrapolation is somewhat more complicated and we will explain as follows. We are interested in estimating the next 6 observations, and so we will have the following model:

$$\hat{Y}_{n-5} = \widehat{m}_{n-5} + \hat{\phi}_1 X_{n-5} + \hat{\phi}_2 X_{n-6} + \hat{\phi}_3 X_{n-7} + \hat{\phi}_4 X_{n-8} + \hat{\theta}_1 \varepsilon_{n-5} + \hat{\theta}_2 \varepsilon_{n-6} + \hat{\theta}_3 \varepsilon_{n-7} + \hat{\theta}_4 \varepsilon_{n-8}$$

$$\hat{Y}_{n-4} = \widehat{m}_{n-4} + \hat{\phi}_1 X_{n-4} + \hat{\phi}_2 \hat{X}_{n-5} + \hat{\phi}_3 X_{n-6} + \hat{\phi}_4 X_{n-7} + \hat{\theta}_1 \hat{\varepsilon}_{n-4} + \hat{\theta}_2 \varepsilon_{n-5} + \hat{\theta}_3 \varepsilon_{n-6} + \hat{\theta}_4 \varepsilon_{n-7}$$

$$\vdots$$

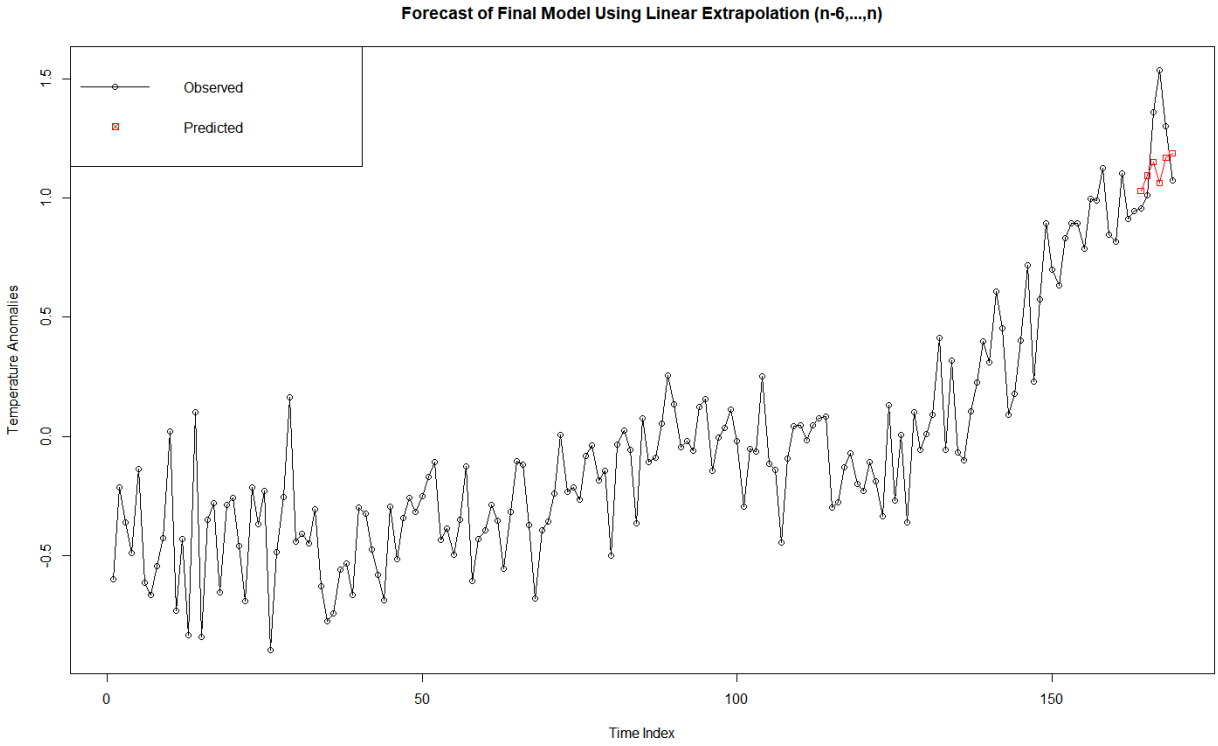
$$\hat{Y}_n = \widehat{m}_n + \hat{\phi}_1 \hat{X}_n + \hat{\phi}_2 \hat{X}_{n-1} + \hat{\phi}_3 \hat{X}_{n-2} + \hat{\phi}_4 \hat{X}_{n-3} + \hat{\theta}_1 \hat{\varepsilon}_n + \hat{\theta}_2 \hat{\varepsilon}_{n-1} + \hat{\theta}_3 \hat{\varepsilon}_{n-2} + \hat{\theta}_4 \hat{\varepsilon}_{n-3}$$

The values of $\hat{X}_{n-5}, \cdots, \hat{X}_n$, are being predicted using the `predict()` function in R. The other values of $\hat{m}_{n-5}, \cdots, \hat{m}_n$ need to be predicted using linear extrapolation. They were initially generated using a Loess model, but since the future observations do not exist, they can't simply be added to the equation. Instead of refitting the entire model with the estimated values of $\hat{Y}_t$, we will use the following formula for linear extrapolation to estimate the future values of $\hat{m}_t$:

$$Y(X^*) = Y_{k-1} + \frac{X^* - X_{k-1}}{X_k - X_{k-1}} \times (Y_k - Y_{k-1})$$

where the * indicates that it is a point which needs to be extrapolated. Also, k refers to the last observation in the subset data, it is equivalent to $t = n - 6$. So, utilizing this formula, we can generate 6 values for $\hat{m}_{n-5}, \cdots, \hat{m}_n$.

The next step to find the predicted values for $\hat{Y}_{n-5}, \cdots, \hat{Y}_n$ is to sum together $\hat{X}_{n-5}, \cdots, \hat{X}_n$ and $\hat{m}_{n-5}, \cdots, \hat{m}_n$. Below is a plot of the forecasted values alongside the actual observed values from the original dataset. The predicted values are on a separate line represented by squares, the actual observations are represented as circles connected by a line.



The plot above shows all the observations from the original dataset, with each point denoted as a small circle connected by a line. The 6 forecasted observations that have just been generated are shown as squares connected by a line at the top right of the plot. The forecasted observations show that they are forecasting an upwards trend. This is overall in line with where the last 6 observations are going. However, it is interesting to note that the last 6 observations had made a large sudden spike upwards, and back down. These sudden jagged movements where data quickly accelerates in one direction and then the other are may not have been completely anticipated by the model.

- **Conclusion and Discussion**

During the initial analysis of the rough as coming from the Loess residuals, it was apparent that the data was close to what was desired, but there were still some issues. The initial diagnostics revealed that the Loess residuals shared some similarities to normality, but it was not quite identical. Looking at the ACF and PACF plots, it was also apparent that there were some serious issues with significant tails. These sorts of deviations from stationarity were expected, as the data comes from a real-world dataset, and it seemed unlikely that the rough of the data would immediately resemble stationarity.

The preliminary modeling where an $ARMA(3,6)$ model was fit to the Loess residuals was surprisingly effective. The diagnostics started to behave in a way that better resembled normality and the spectral analysis showed that the rough was looking less erratic. The most surprising aspect was that the PACF plot went from having four significant tails to just one.

During the process of deciding on a final model, there was some slight confusion due to the differing results based on the $AICc$ criterion. The main decision for choosing the model chosen by independently fitting different values for p and q was that the resulting ACF and PACF plot showed a rough which fit within the confidence intervals for both plots. Other diagnostics were sometimes better and sometimes worse. The formula for the final $ARMA(4,4)$ model of the Loess residuals is as follows:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \phi_4 X_{t-4} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \theta_4 \varepsilon_{t-4}$$

where X_t comes from the following model with $\widehat{m}_t$ as the fitted Loess values,

$$Y_t = \widehat{m}_t + X_t.$$

The periodograms also showed an improvement in that they were overall less tall. The behavior of the smoothed periodograms were such that they were becoming flatter than both the Loess residual and preliminary model periodograms. This was a good sign, as we were looking for each of the frequencies to become more balanced in terms of their contribution to the overall variance. When comparing the models using spectral analysis, it seemed that the final model of the rough exhibited properties most like that of white noise. The spectral density of the final model seemed to show that the smoothed periodogram fit relatively close to each other considering that the smoothed periodogram itself didn't have any highly characteristic features like that of the sunspot data from the handout and the textbook.

It was also interesting to use the criterion for selecting the number of neighbors in a measured way, rather than simply looking for the global minimum. By balancing the number of neighbors and the criterion value, it was possible to choose a span which was reasonable in its criterion value and its span size relative to the data set. Overall, we were able to find a smoothed periodogram where the frequencies seemed mostly uniform in terms of their spectrum levels. This implies that the overall variance is not being overly contributed by any frequency.

During the forecasting, it became apparent that it would be difficult to find a forecast due to the Loess modeling. The `sarima.for()` function allows for simple forecasting when the input doesn't account for a Loess detrending step. Therefore, it was necessary to perform linear extrapolation. This step makes sense, since in our case we are only dealing with a forecast of six observations. Also, the direction of the last observations is predominantly upward, so the linear extrapolation didn't have an

issue anticipating this trend. It is interesting to see however that the last few observations made some sudden movements up and down, different from what was happening before. Looking carefully, it is apparent that the raw observations were narrowing down into the shape of a triangle, before breaking out in a fast reaction. This sort of behavior may not have been completely anticipated by the model, but the forecast was able to understand the general direction of where the observed data was going.

It seems that after having gone through several steps of modeling the rough and trying to understand whether it resembled a stationary series, the test of this important property came through forecasting. By taking a subset of the data, it was possible to compare the forecasted values with the actual observations. It seems that the forecast was able to understand the current trend of the observations. The most interesting aspect is that the predicted values seemed to go downwards first, before moving upwards. This seems to be the actual pattern of what was taking place in the raw observations. However, it is uncertain whether this is something that the model was anticipating.

If we were to spend more time on the project, we would be interested in trying alternate models. Primarily, we would like to test the differencing, using first or second order differences instead of Loess to detrend the data. It is possible that when considering a larger combination of values for p, d, and q for $ARIMA(p, d, q)$, the resulting $AICc$ value could have been an improvement. Also, if doing repeated predictions over time, this model may have more optimal results in terms of accurately predicting the direction of the observations. It would have been interesting to see how well the spectral density would have fit with a first order difference of the data.

Another interesting possibility would be to test different lengths of the data using the final model for prediction. We were only able to test the last six observations; however, it would be interesting to try and somehow test the accuracy of the predictions on different segments of the data. This would allow us to see how effective our final model was in predicting the overall dataset. It is possible that it would show that we would instead need to occasionally refit the model as 'new' observations are added on. Looking at the dataset, there is not a single trend as the data goes sideways then upwards. This could imply that certain models are useful for certain sections of data, while other retuned models are necessary for predicting other sections of data.

Reference:

https://en.wikipedia.org/wiki/Time_series

https://data.library.virginia.edu/understanding-q-q-plots/

http://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html

https://en.wikipedia.org/wiki/Ljung–Box_test

https://en.wikipedia.org/wiki/Extrapolation#Linear

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/predict.arima.html

https://stackoverflow.com/questions/37503612/why-i-got-this-message-in-r-optim-gave-code-1

https://stackoverflow.com/questions/1497539/fitting-a-density-curve-to-a-histogram-in-r

https://stackoverflow.com/questions/10108073/plot-legends-without-border-and-with-white-background