Nima Malboubi

INTRO:

In this study, I will be analyzing qtrain5 data, and will attempt to forecast two years of data based on the current 431 observations. See *Figure 1.1* for a graph of the original data. From this graph, it appears that the first 15 weeks of observations deviate from the rest of the graph and do not follow a similar pattern. Thus, I have removed the first 15 weeks of data because of the lack of consistent trend in the initial data points with the rest of the data (see *Figure 1.2a* and Figure *1.2b)*. *Figure 1.2a* shows the entire data set including the first 15 weeks, which are colored in red. *Figure 1.2b* is the graph of the data after the first 15 weeks have been removed. Notice, how the data now corresponds to the pattern exhibited by the rest of the graph.

VARIANCE OF DATA:

It appears that there is some decreasing variance in the second half of the data, and thus, in order for the data to look reasonably homoscedastic, I used the logarithm function (see *Figure 1.3*) to transform the data.

DIFFERENCING:

Trend and seasonality are evident in the data set. In order to remove the trend, I differenced the data with a time lag of 1. See *Figure 1.4* for the graph of the once differenced data. However, this newly differenced data displays seasonality, which can be clearly seen in the ACF and PACF of the differenced data (see *Figures1.5a* and *Figure 1.5b*). By looking at the ACF, one can clearly see that significant correlation exists approximately every 52 weeks. The PACF also exhibits seasonality at lags 52 and 104.

The plot of the 2^{nd} order differenced data is given (see *Figure 1.6*) as well as the ACF and PACF (see *Figure 1.7a* and *Figure 1.7b*). The second order differenced data looks like white noise; the data points appear to be uncorrelated, except for a few lags, which can be due to randomness. By looking at the ACF (*Figure 1.7a)* with a lag max of 200, there are only 6 significant correlations, and thus the data appears to be stationary.

FITTING MODELS:

I attempted to fit several models to the data using the maximum likelihood method in R. I prescribed these models by utilizing the characteristics of the ACF and PACF of the model after second order differencing. I then used cross validation for each of the models, and the AIC to determine which model would provide a better fit. In order to successfully use cross-validation to determine an accurate prediction, I used the first two years of data to predict the third year, and then calculated the cross-validation error by comparing it to the actual data for year three. Then I used years one, two, and three, in order to predict year four, and again calculated the cross-validation error by comparing it to the actual data. I continued this process through year eight. After completing the cross-validation, I then continued to take the average of the cross-validation error. I chose the model that gave me the lowest average cross-validation error.

In order to verify the correctness of the log transformation of the data, I used cross-validation to analyze the original data without transformations. I then used cross-validation on the logged values; however, before calculating the cross-validation errors of the logged values, I converted them back by taking log base e and raising it to the log transformed data. In this way, I would be able to compare the cross-validation errors of the log transformation with the cross-validation errors of the original data. In

every case, the cross-validations of the logged data were smaller. Thus, I used the logged data to complete my predictions. This can be seen in my R code for Data5 labeled "R Check".

THREE MODELS:

MA (1) Model X Seasonal MA (1)_ 52 Model

After lag 1, there are very few significant correlations in the ACF plot, other than significant correlations every 52 weeks. The rest of the correlations are insignificant, and thus I chose to use an MA (1) model with a seasonal MA(1)_52 model. The average cross-validation error was 0.05938427, and the AIC was -1326.744

.

MA(1) Model X Seasonal AR(1)_52 Model

For this model, the correlation at the lags was not significant after lag 1, so thus I again used an MA(1) model in an attempt to fit the data. The significance of the correlations seemed to decay every 52 weeks, as if the seasonality followed an AR(1) model. Thus, I crossed an MA(1) model with a seasonal AR(1) model with a lag of 52 weeks. I received an average cross-validation error of 0.06046347, and an AIC of -1329.952

AR(4) Model X Seasonal MA(1) Model

After looking at the PACF of the graph, it appears that the correlation between the lags become insignificant after the 4^{th} lag, except for the significant correlations every 52 weeks for seasonality. I chose to cross this with a seasonal MA(1) model, and received an average cross-validation error of 0.06818945. The AIC was 1310.106

CONCLUSION:

The first model: MA(1) X Seasonal MA(1)_52 had the lowest cross-validation error. However, its AIC was not the lowest of the models (-1326.744 vs -1329.952) However, the model with the highest AIC: MA(1) X Seasonal AR(1)_52, did have a slightly higher cross-validation error. Since the cross-validation error was smaller when the predicted data was compared to the actual data, and the AIC is very close to the other in the other model, I decided to choose an MA(1) X Seasonal MA(1)_52 in order to predict the next 104 data points. See *Figure 1.8a* for the graph of the predicted log transformed data and *Figure 1.8b* for the combined log transformed data. The predicted data is in red. See *Figure 1.9a* for the graph of the "de-logged" predicted values, and *Figure 1.9b* for the combined "de-logged" graph of the original data and the "de-logged" predicted values.

Figure 1.1: A plot of the original data

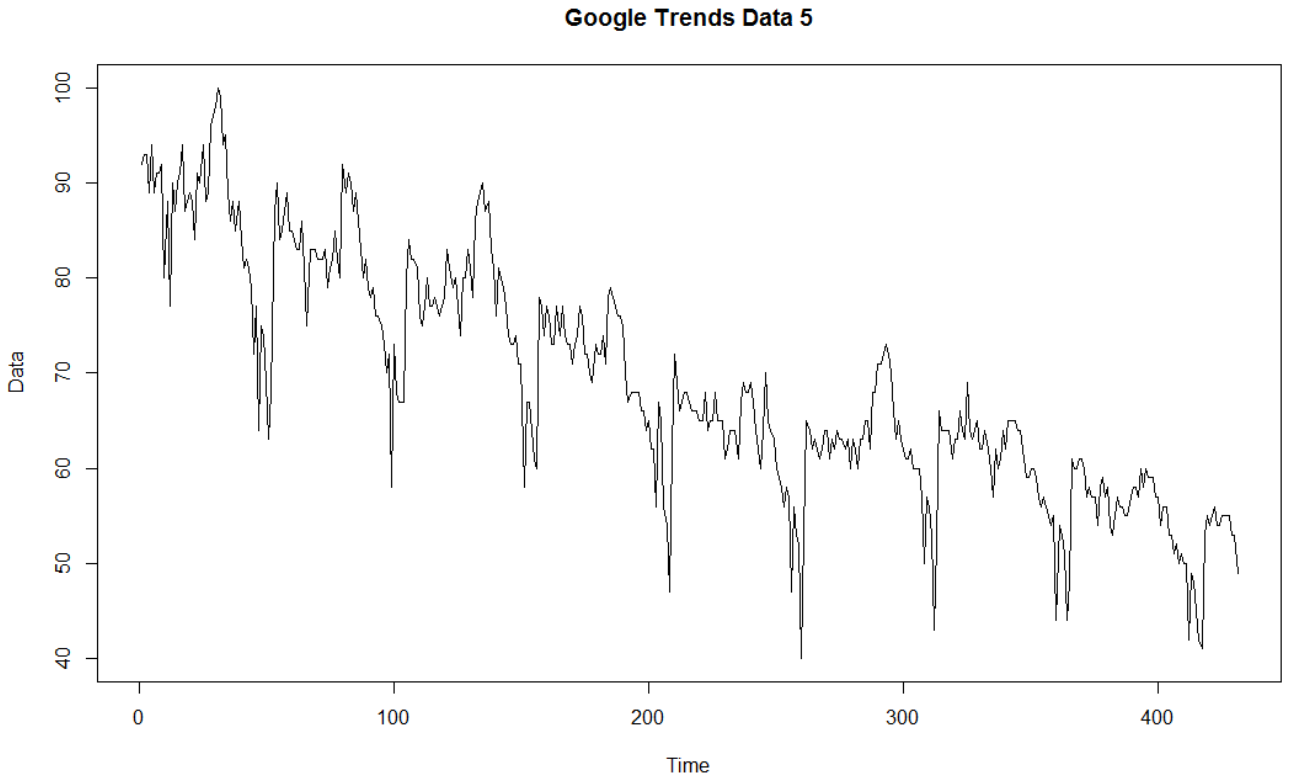


Figure 1.2a: A graph of the original data, including the first 15 weeks, shown in red.

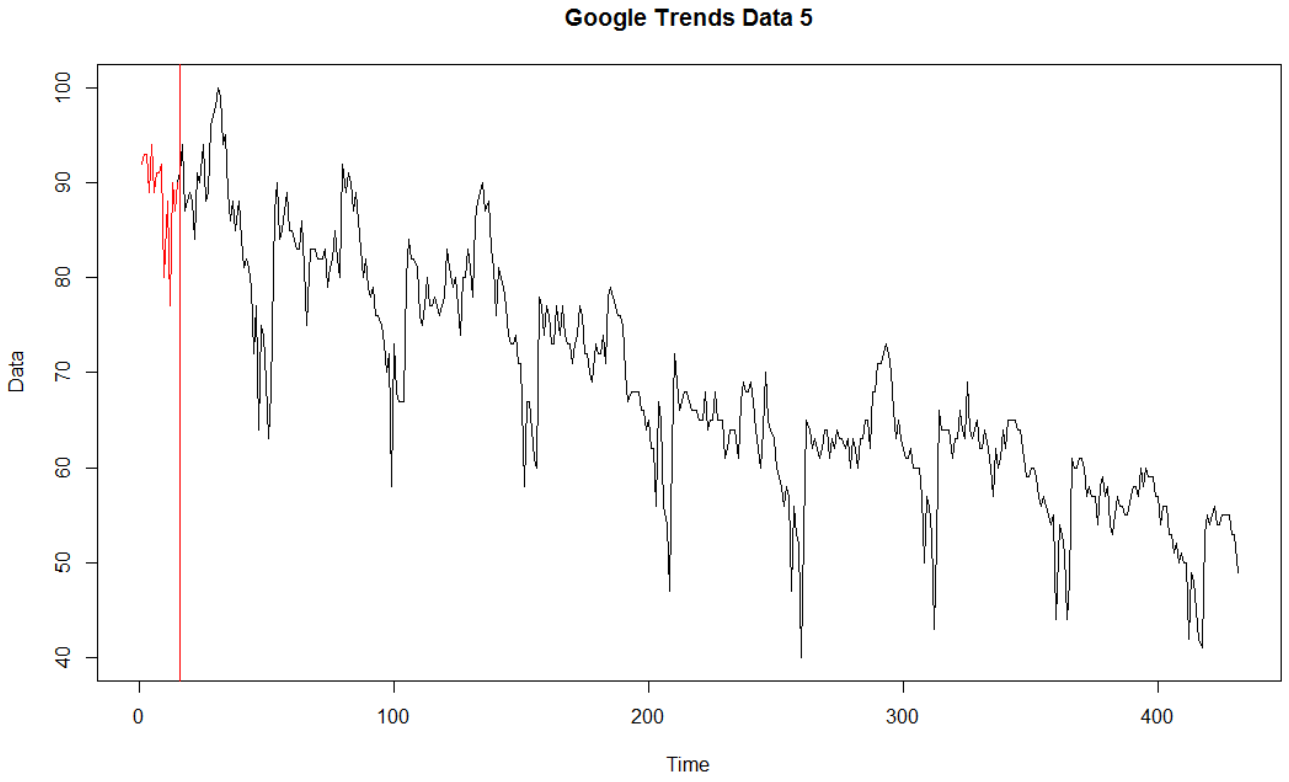


Figure 1.2b. A plot of the original data with the first 15 weeks removed

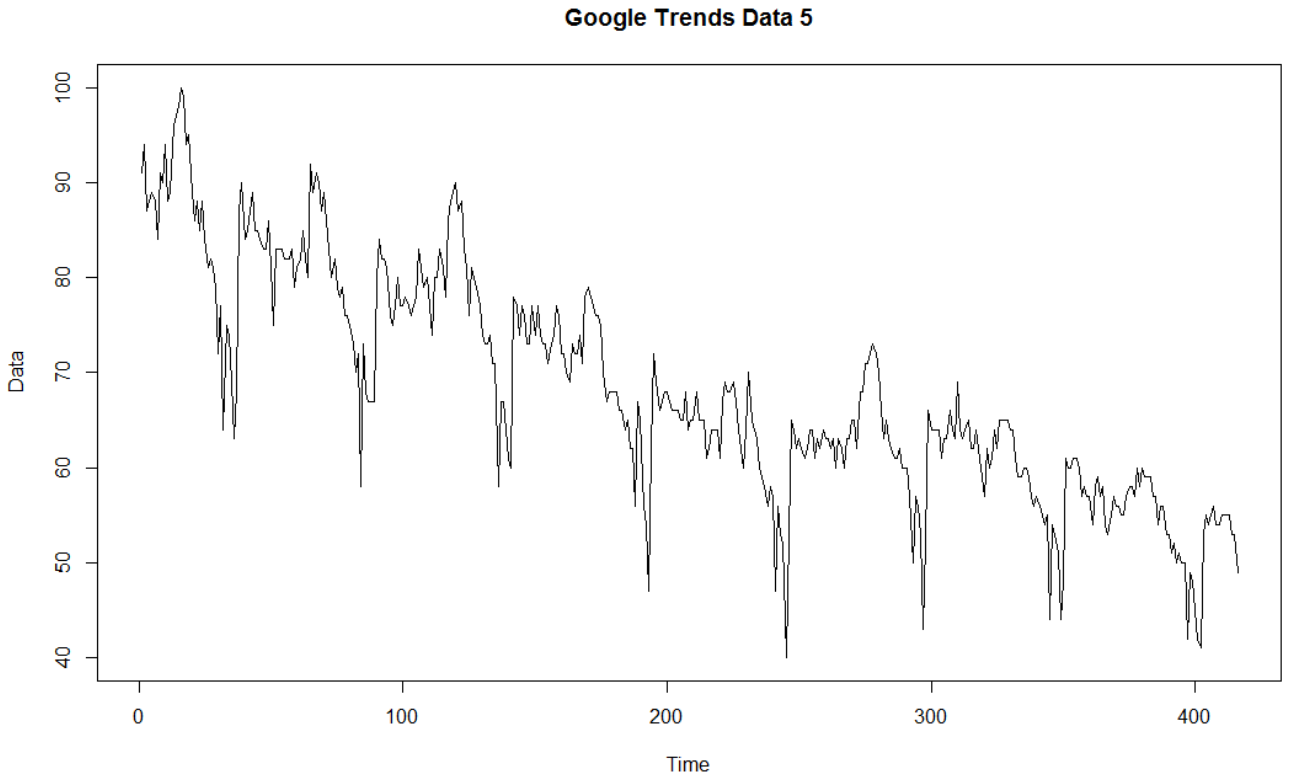


Figure 1.3: Log of the Data

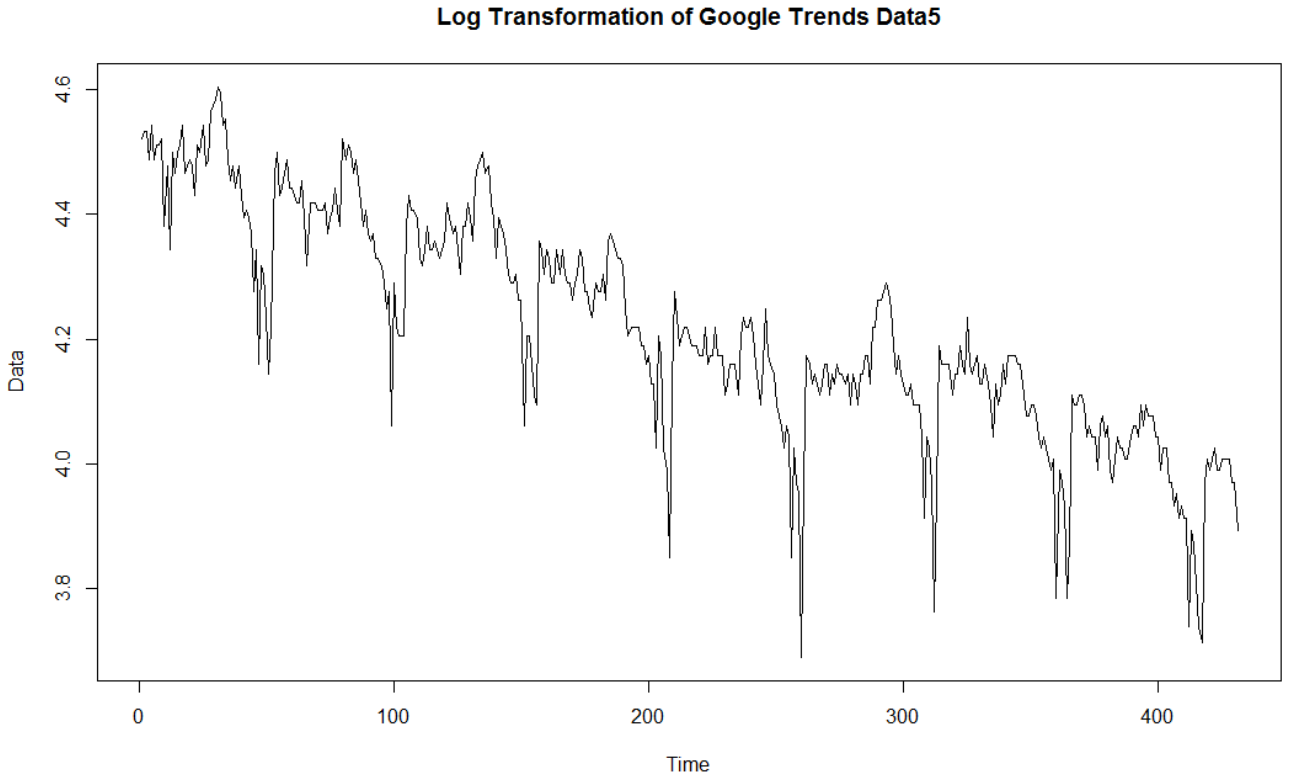


Figure 1.4: Data after Differencing Trend

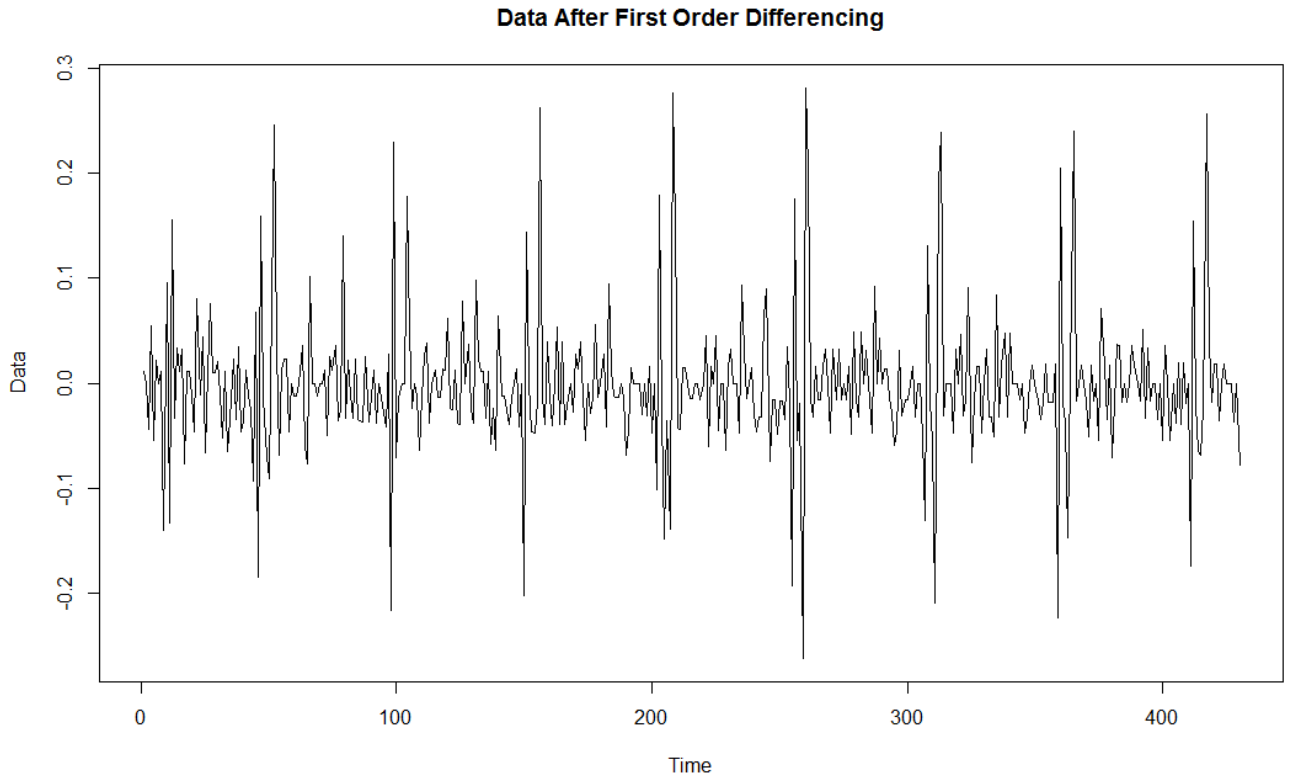


Figure 1.5a: ACF of Differenced Data

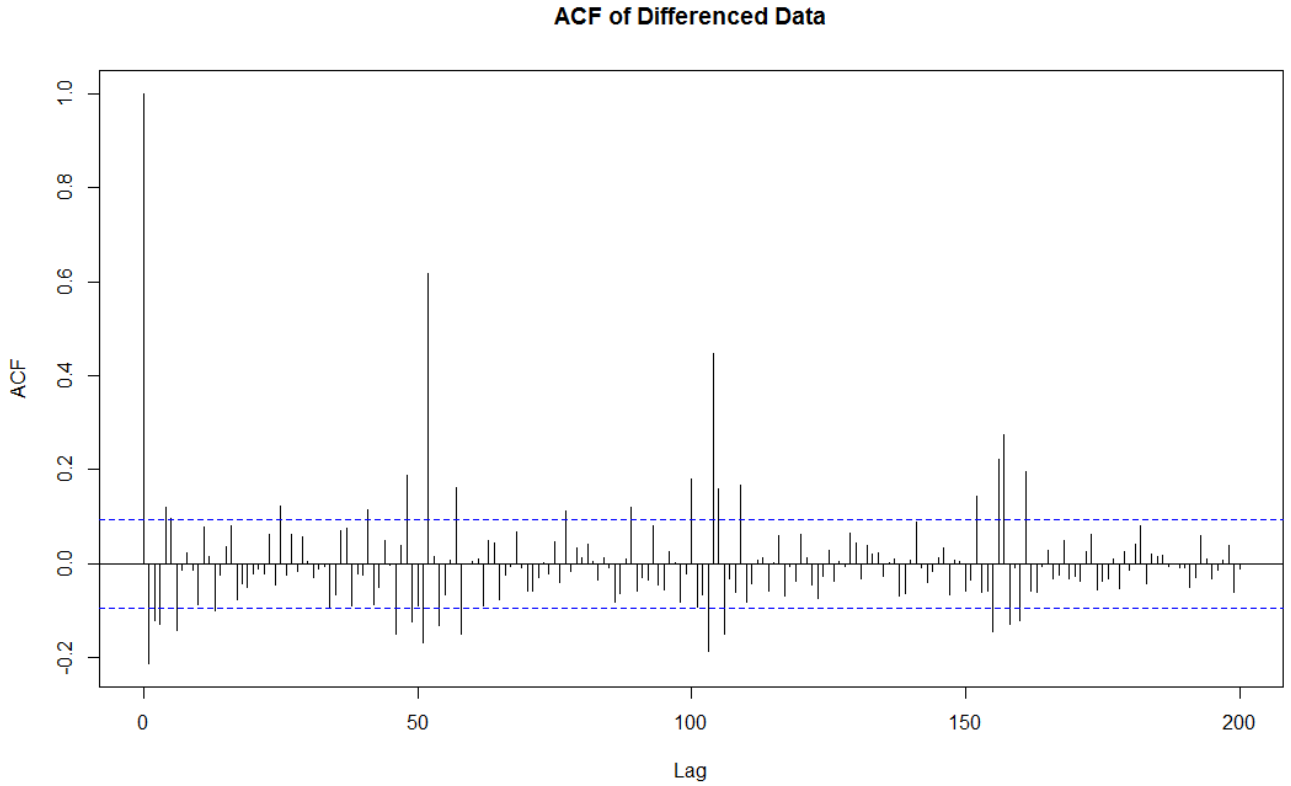


Figure 1.5b: PACF of the first order differenced data

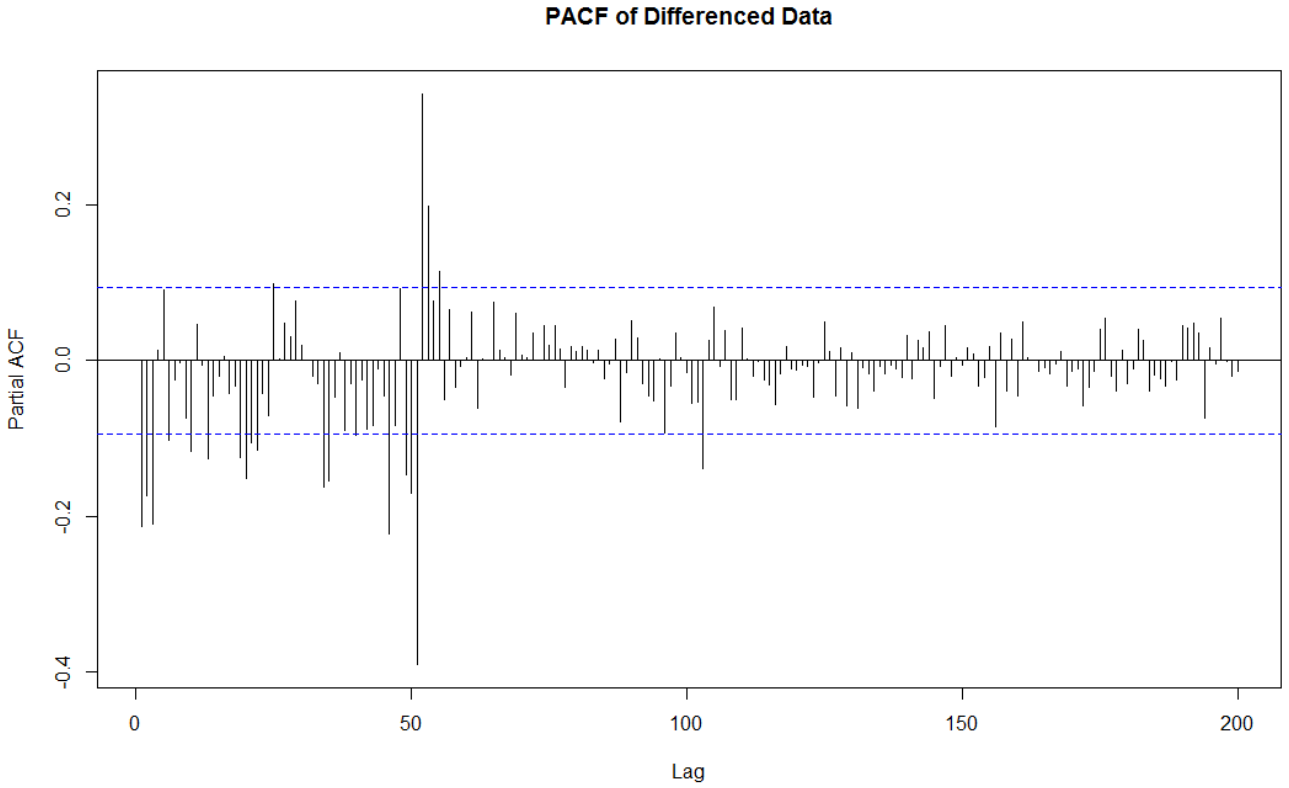


Figure 1.6: Graph of data after differencing again for seasonality

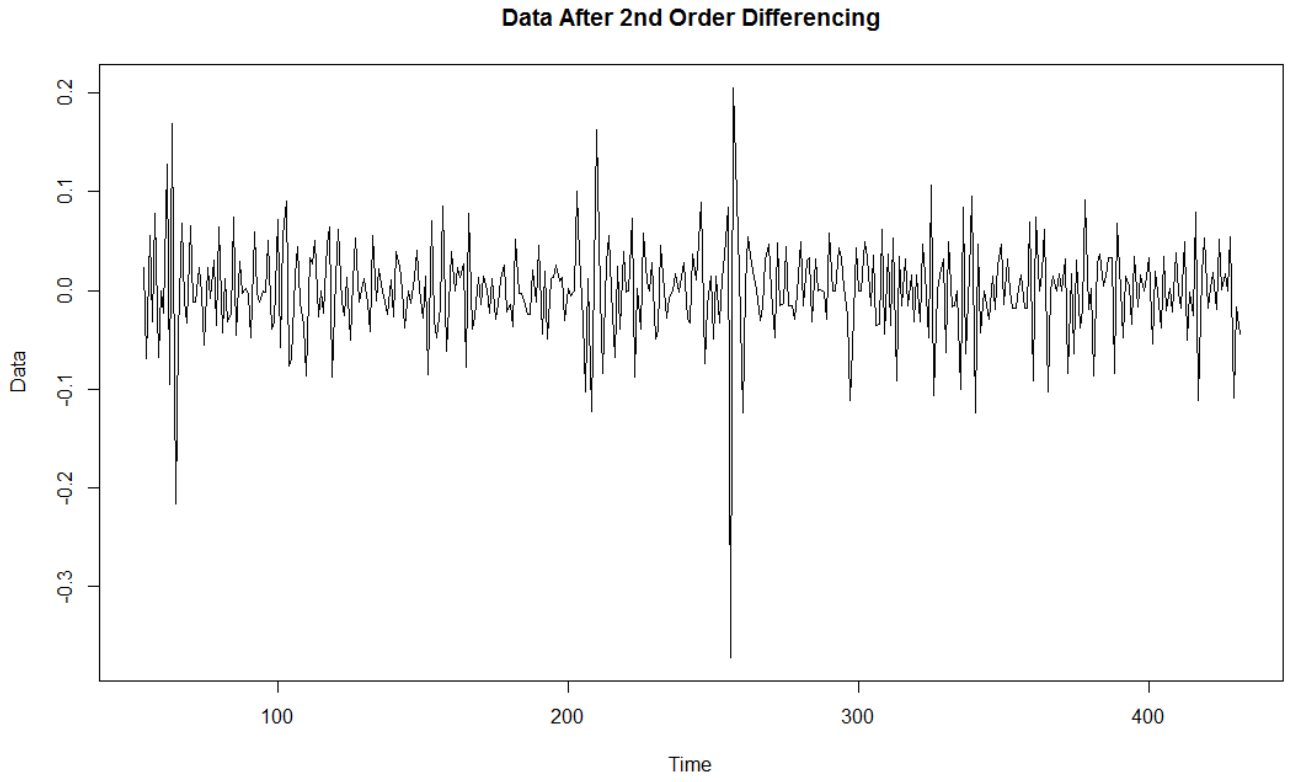


Figure 1.7a: ACF of the data after differencing again for seasonality

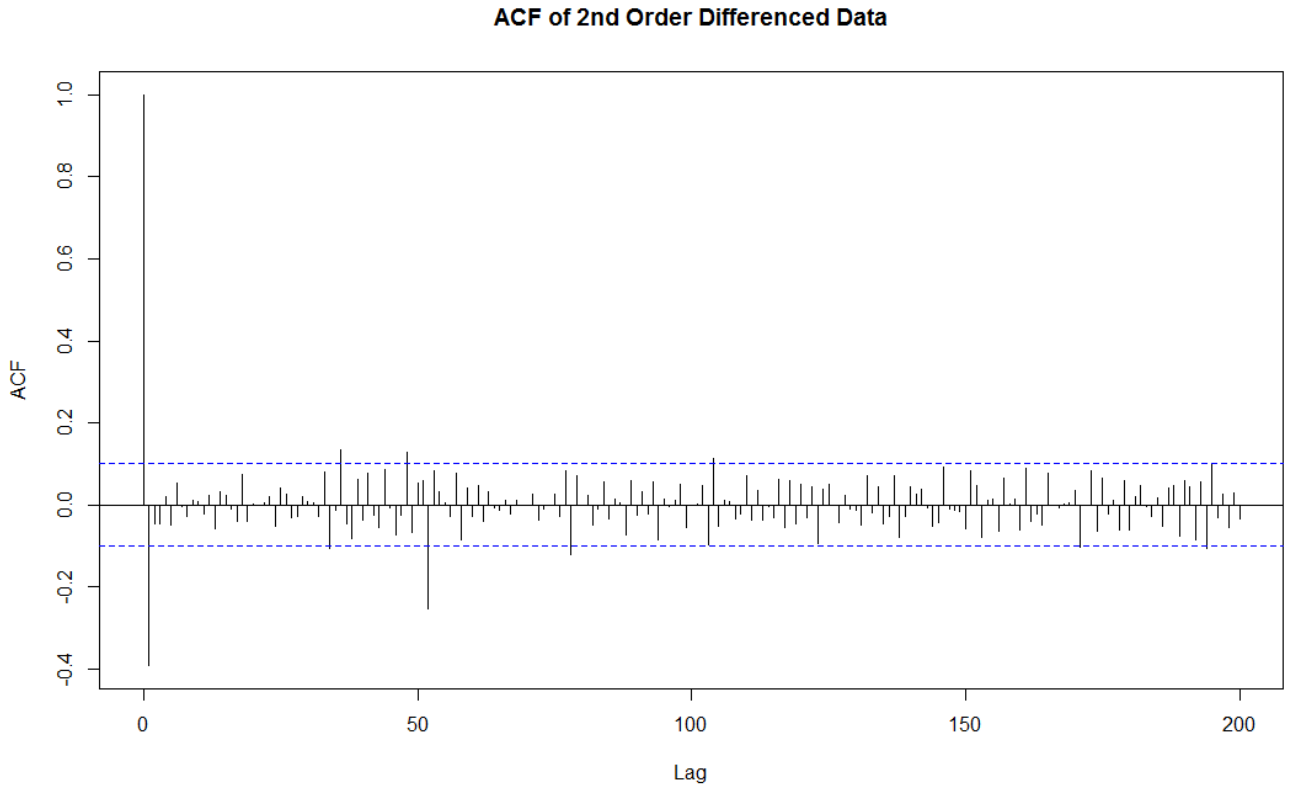


Figure 1.7b: PACF of the data after differencing again for seasonality

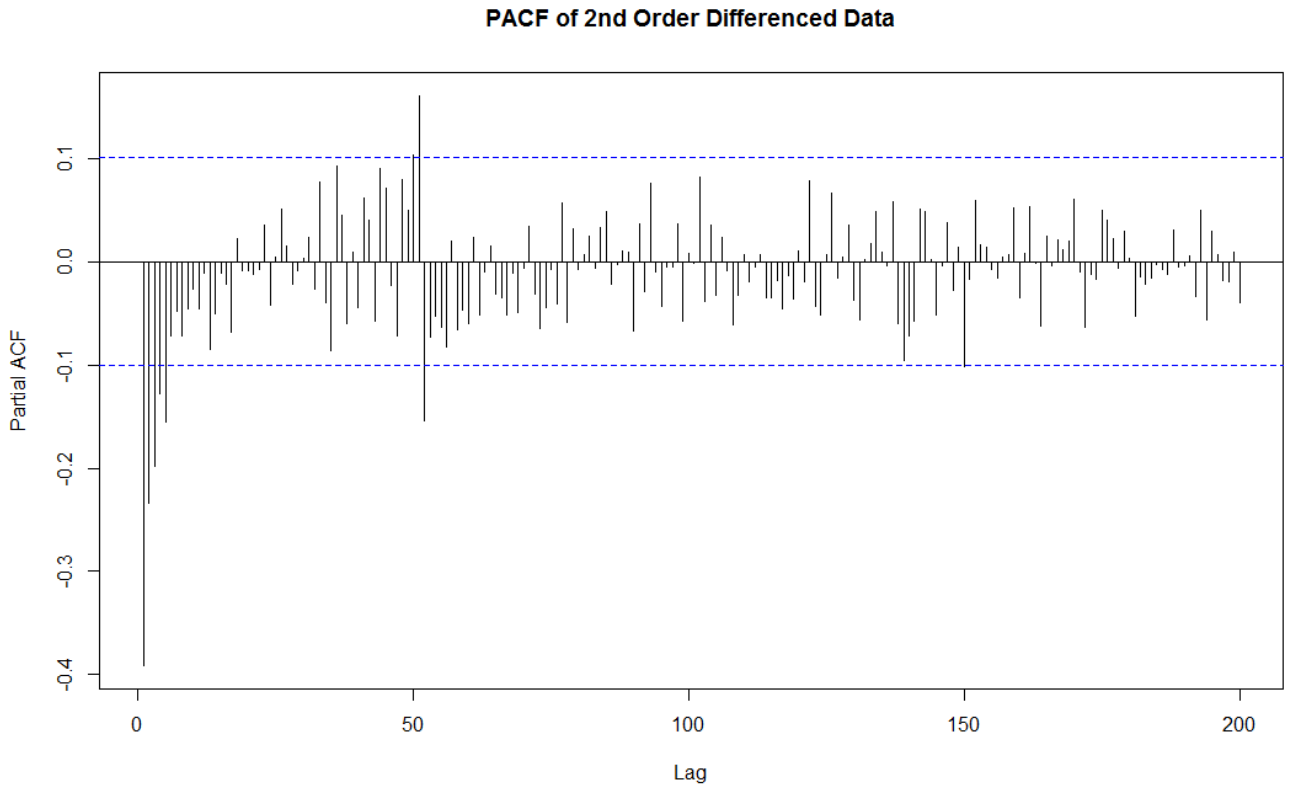


Figure 1.8a: Graph of the predicted data

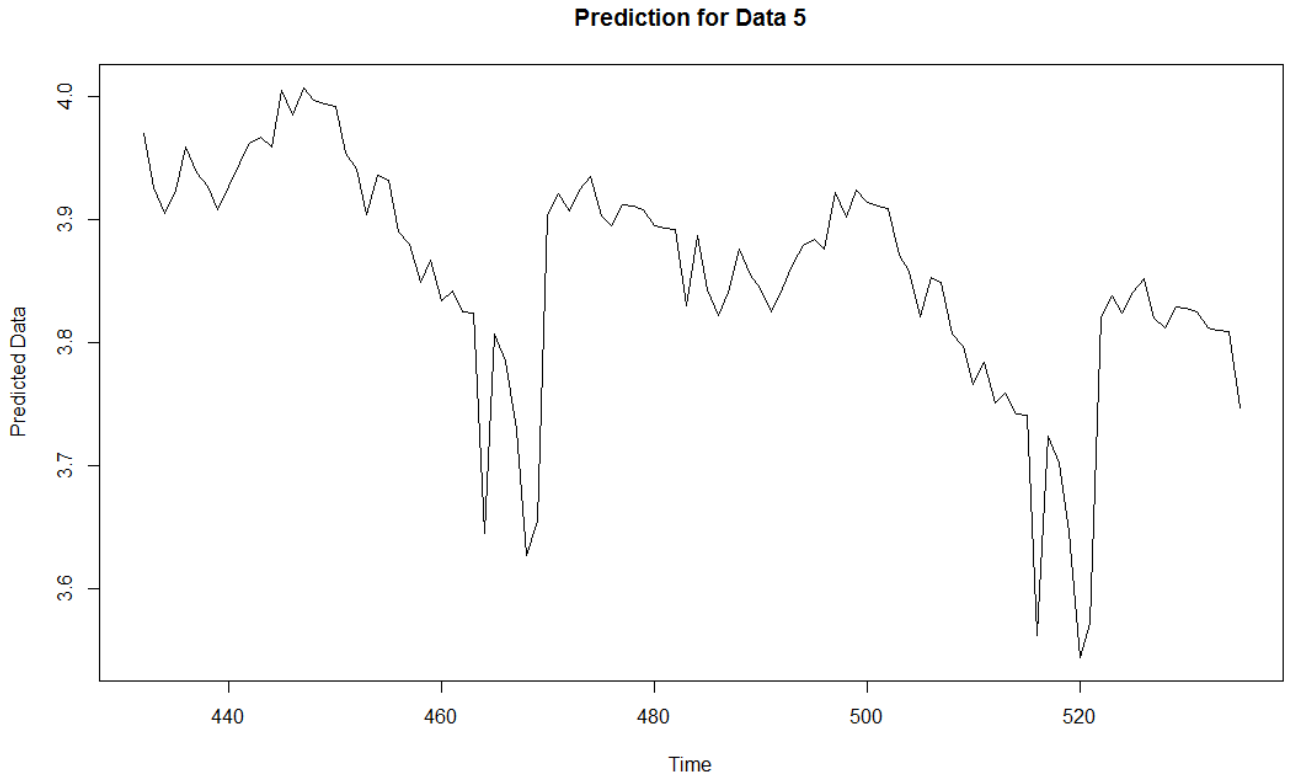


Figure 1.8b: Combined graph including predicted data and original data. Predicted data is in red.

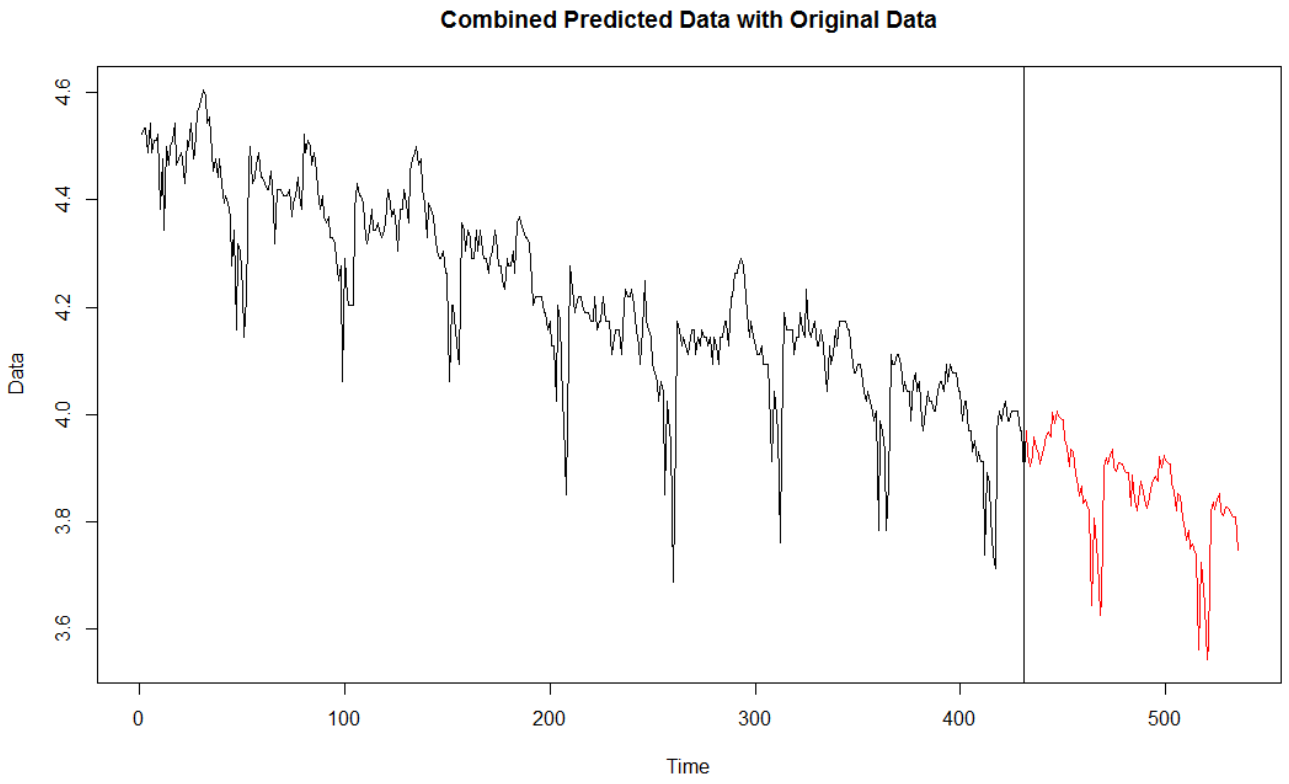


Figure 1.9a: Predicted data values without log transformation.

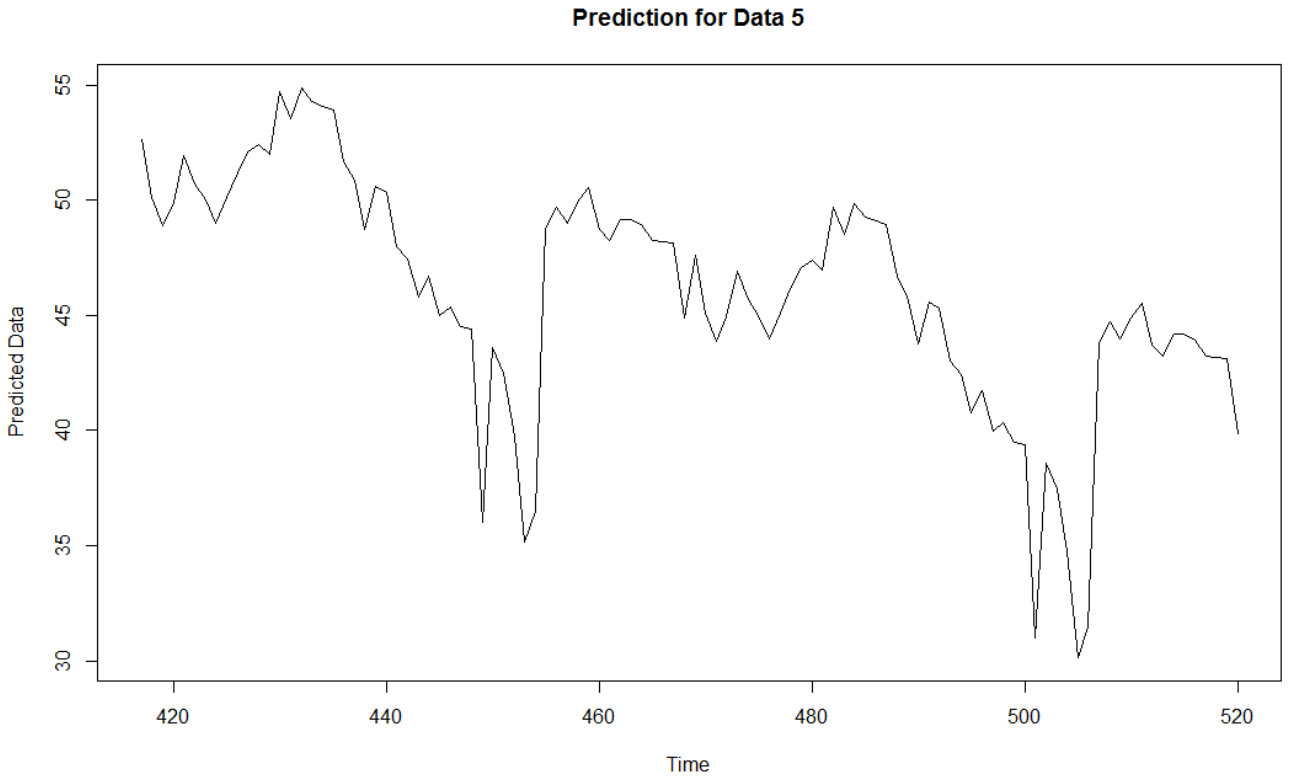
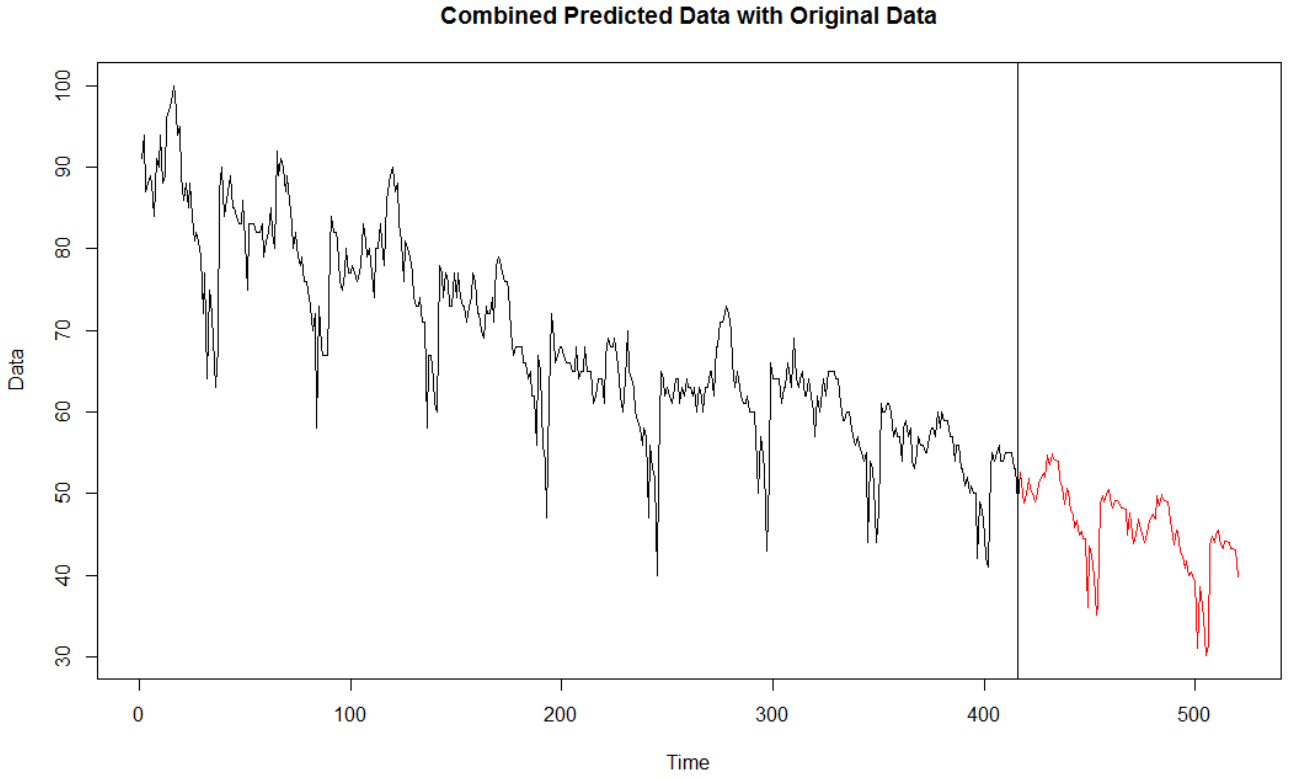


Figure 1.9b: Original data set with predictions included. Predictions are in red.


Combined Predicted Data with Original Data
Data
Time
30
40
50
60
70
80
90
100
0
100
200
300
400
500