Stat 153 Project

Kathleen Nie

12/14/2020

Executive Summary

Stuff City Inc. is a (made-up) publicly traded company who specialize in home-improvement products and services. This data set includes the closing price of Stuff City from 2016 to 2020, which is primarily affected by the closing prices of the past 8 days. The end goal is to predict the closing prices for the next two weeks, or the next 10 data points. According to the selected differencing model with ARMA(0, 1)x(0,2)[7] noise, the closing prices of the next two weeks will not largely deviate from the closing prices of the past week and a half.

Exploratory Data Analysis

Upon graphing the time series, there doesn't seem to be any strong seasonality or trend. After grouping the series by the year and month they're in and taking the mean of the respective closing prices, we observe both a yearly and monthly trend in the respective left and right panels of Figure 1. After an initial plotting of the time series data, there seems to be a sharp price decrease in the beginning of 2020. This outlier may later cause heteroscedastic behavior within fitted residual plots. Observing the acf plot in right panel of Figure 2, there is large autocorrelation within our lag values and a geometric decay, which suggests we may have to difference our data. If we take a closer look and decompose the time series and graph the decomposition, it's clearer that there is some type of trend, seasonality, and randomness component (Figure 3). This step is achieved using decompose(data) function.

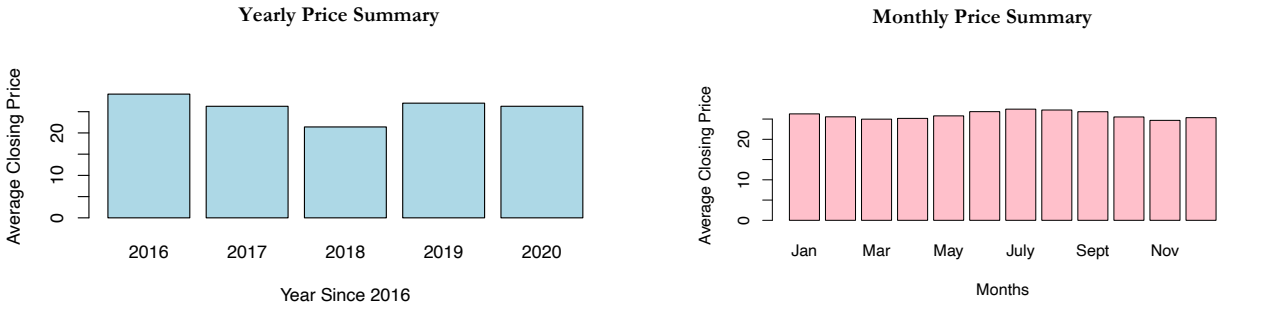


Figure 1: On the left, the average closing price of each year. The second bar graph visualizes average closing price of each individual month.

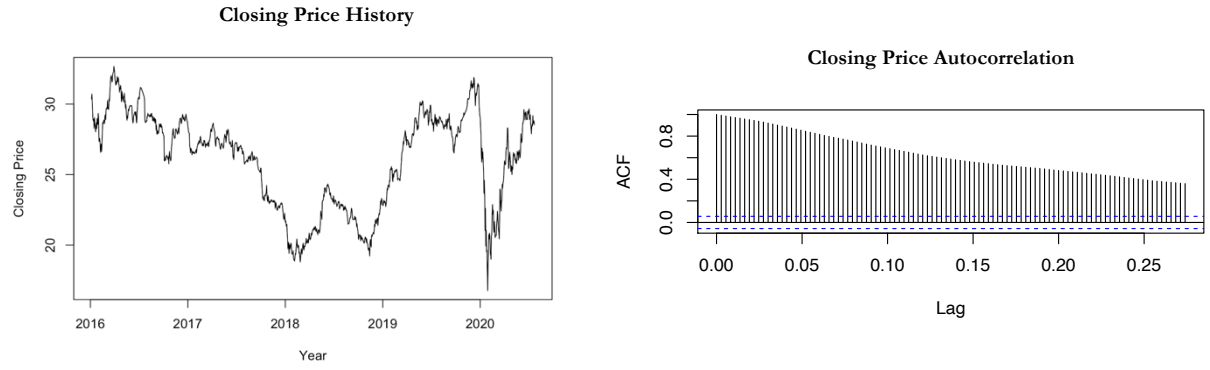


Figure 2: First graph shows a time series plot of Stuff City Inc. closing prices from 2016 to September of 2020. The acf plot showcases the autocorrelation values of closing prices with no clear visible cutoff.

Looking at the decomposition graph in Figure 3, there seems to be an observed trend over the 4 years that looks kind of parametric. While the seasonality component does not seem to affect a huge amount of the data, it does have an affect of around plus or minus $1 on the closing price. Further observing the seasonality component, the seasonal periodogram plot seems to have three significant frequencies around 0.006, 0.011, and 0.027. These frequencies correspond to about 7, 13, and 31 days, suggesting there's some sort of weekly and monthly seasonality component.

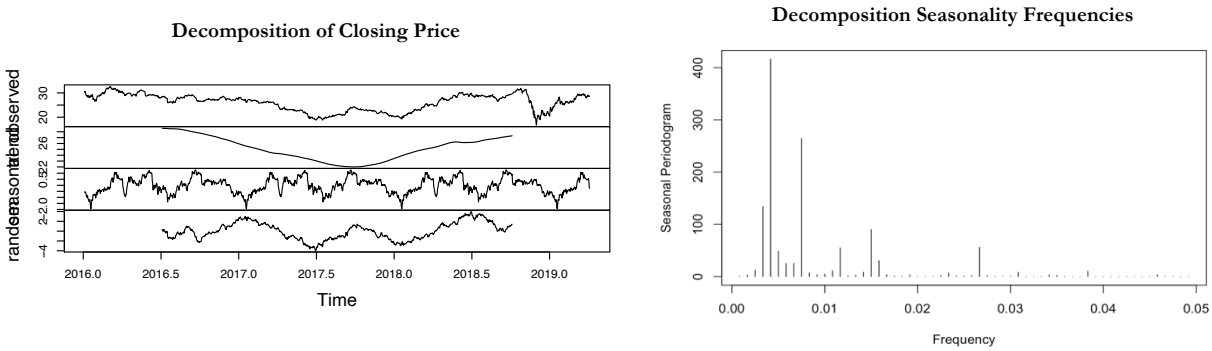


Figure 3: Left panel contains the decomposed components of Stuff City Inc's historical closing price. The following periodogram illustrates significant frequencies found in the seasonality component of the previously decomposed closing prices. Large spikes roughly correspond to periods of 1 week, 2 weeks, and 1 month

Models Considered

To model the yearly trend within the data and its relatively random behavior, both a parametric and differencing approach will be explored. In the parametric model, the trend (which is likely attributed by the closing price average per year and month), seasonality, and randomness seen in the decomposition will be taken into account. In the differencing model, various order differencing and lags will be considered. Both of these signal models will be complimented with ARMA models for the remaining noise

Parametric Signal Model

Looking at the periodogram in Figure 4 below, the frequency is strongest around 0.002. As a result, a sinusoid with period 365.25 is interacted with the year and month of the data. Accounting for a possible quadratic trend that was seen in the decomposition earlier, time squared is added. Additionally, indicators for each month and year are also included. To account for heteroscedasticity towards the end of the data set, the log of the original closing price is taken. This deterministic signal model is detailed in Equation(1) below, where X_t is the additive noise term.

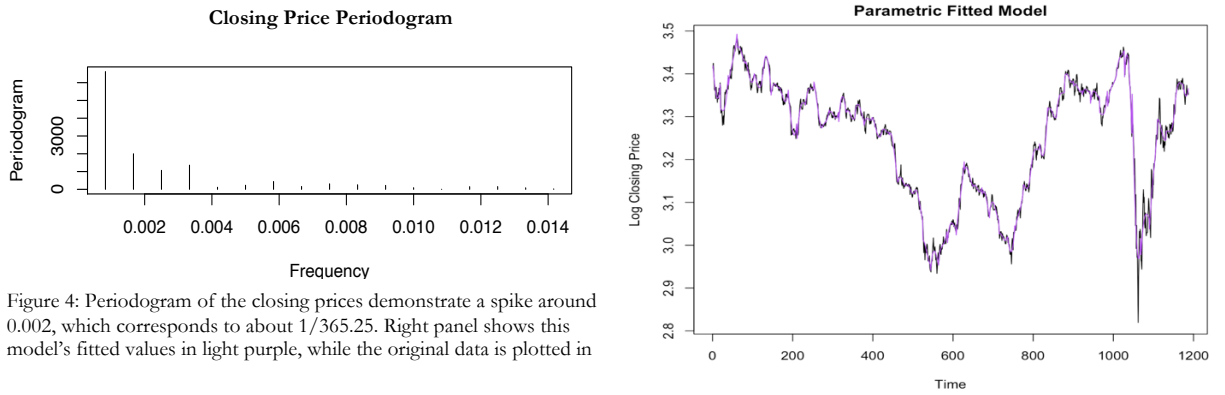


Figure 4: Periodogram of the closing prices demonstrate a spike around 0.002, which corresponds to about 1/365.25. Right panel shows this model's fitted values in light purple, while the original data is plotted in

$$log(close)_t = \beta_0 + \beta_1 t + \sum_{i=1}^{12}\sum_{i=1}^{4} \beta_{1+i+j} I_{\text{month}_{it}} I_{\text{year}_{jt}} sin\left(\frac{2\pi t}{365.25}\right) + \sum_{i=1}^{12}\sum_{i=1}^{4} +\beta_{50+i+j} I_{\text{month}_{it}} I_{\text{year}_{jt}} cos\left(\frac{2\pi t}{365.25}\right) + X_t \quad (1)$$

Observing the residuals in Figure 5 on the right, this plot looks mostly stationary except towards the end. In the corresponding position of the original graph, there is a huge unexpected closing price drop at that time index. Since such a drop only occurs once in this data set, it will be treated as an outlier, so it does not largely affect our prediction model.

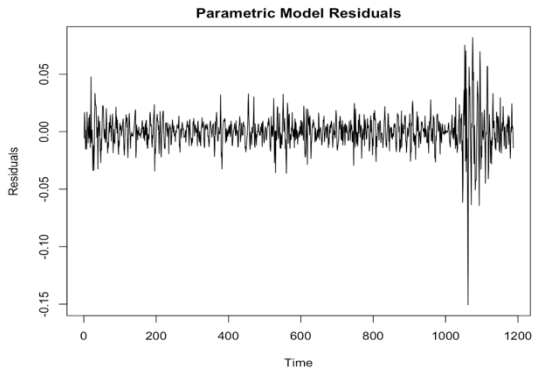


Figure 5: Residual plot of log(original data points) minus parametric model values

Parametric Signal with ARMA(1, 2)

To start off, auto.arima function is used to create a rough starting point. Taking the suggestion from auto.arima, the left panel of Figure 6 below shows the results of fitting an ARMA(1, 2). The ACF and PACF plot shows that improvements can still be made as there are still a good number of significant values in both.

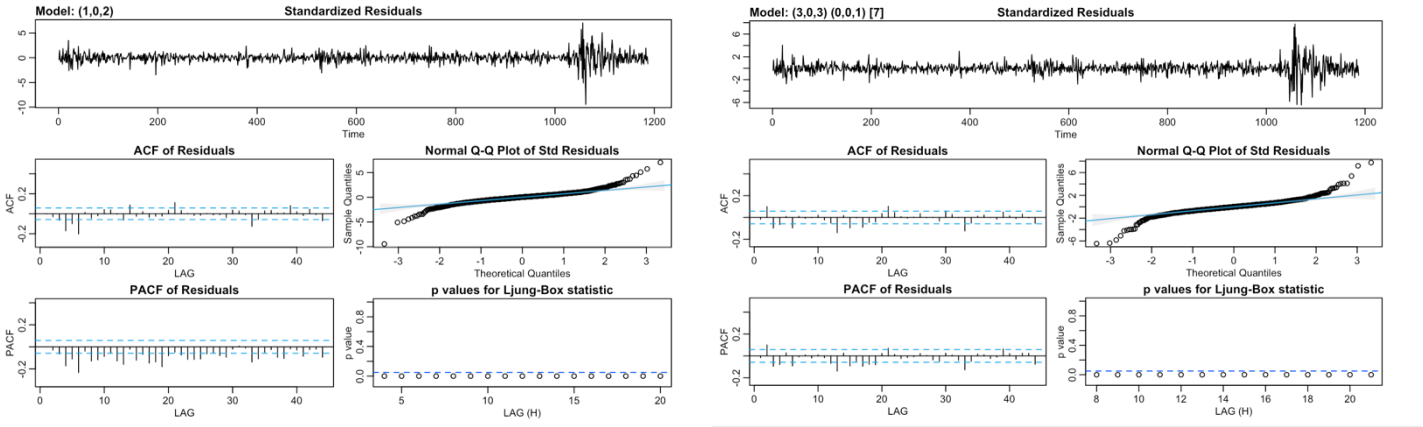


Figure 6: Results of sarima_wPACF( ) function for ARMA(1, 2) (left) and ARMA(3,3)x(0,1)[7] (right)

Parametric Signal with ARMA(3,3)x(0,1)[7]

For a better idea on how to determine p, q, P, Q values, an acf and pacf plot of the parametric model residuals was created. Although not shown, the acf plot cut off after around lag 7 while the pacf plot did not have a clear cut off. Using this information, it suggests that a SARIMA model with P = 1, S = 7 and Q > 0. After several experimentations, the ARMA(3,3)x(0,1)[7] had the best IC values out the several possible models. The acf and pacf plots in as seen in the right panel of Figure 6 above suggests that this is a better value than the previous model given by auto.arima() (compare with acf/pacf of left panel).

Differencing Model

To account for possible monthly seasonal trends, second order difference at lag 1 and lag 7 will be taken. Differencing by lag 1 removes deterministic trend and differencing by lag 7 removes any weekly seasonality trend. The reason these lags were chosen is because they showed significant values in the periodogram in Figure 3. In the left side of Figure 7 below, the black lines show the values of the original closing price data while pink lines show the fitted values of the forementioned differencing model. On the right panel, the residual plot, which is also the second order difference at lag 1 and 7 is plotted. Overall, residuals look relatively stationary without heteroscedasticity. However, the variance towards the end increases as the original time series data hold more outliers/drastic changes in 2020. Since the differencing model fits a majority of the data pretty well, this increased variance at the end can be set aside.

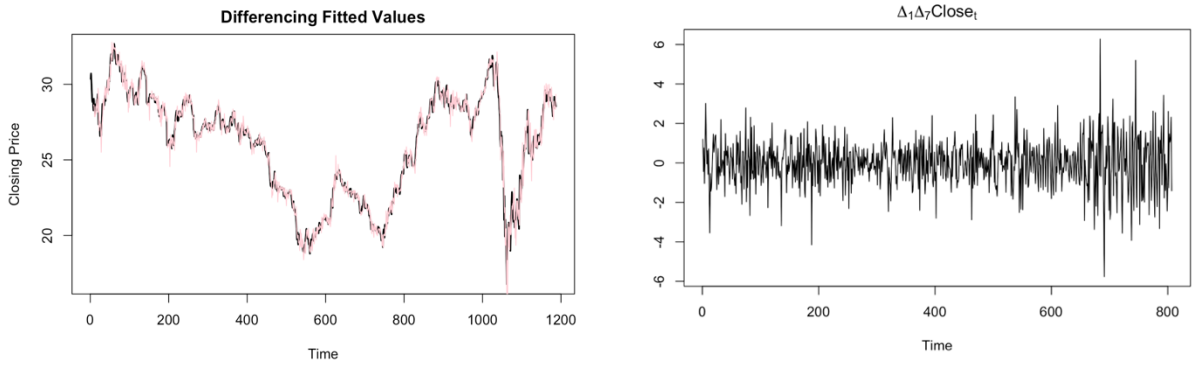


Figure 7: First time series plot shows differencing model fitted values in pink, and the original data in black. The second plot graphs differencing model residuals, which are also the values of at second order differencing. Heteroscedasticity does not seem to be prevalent in the residuals.

Differencing with ARMA(2, 1)x(0,2)[7]

The acf and pacf plot in the left panel of Figure 8, a significant value in the acf graph at lag 7 and possibly at lag 14 is observed, suggesting Q = 2. There are also some smaller consistent spikes possibly suggesting that p > 0. Directly below in the pacf graph, there seems t be a tapering off pattern also occurring at multiples of lag, which indicates that Q > 0. To start off, the right panel of Figure 8 showcases the results of fitting the sarima_wPACF( ) for model ARMA(2, 1)x(0,2)[7]. Although the acf and pacf residual plots look ok, the Ljung-Box statistics seem to have room for improvement.

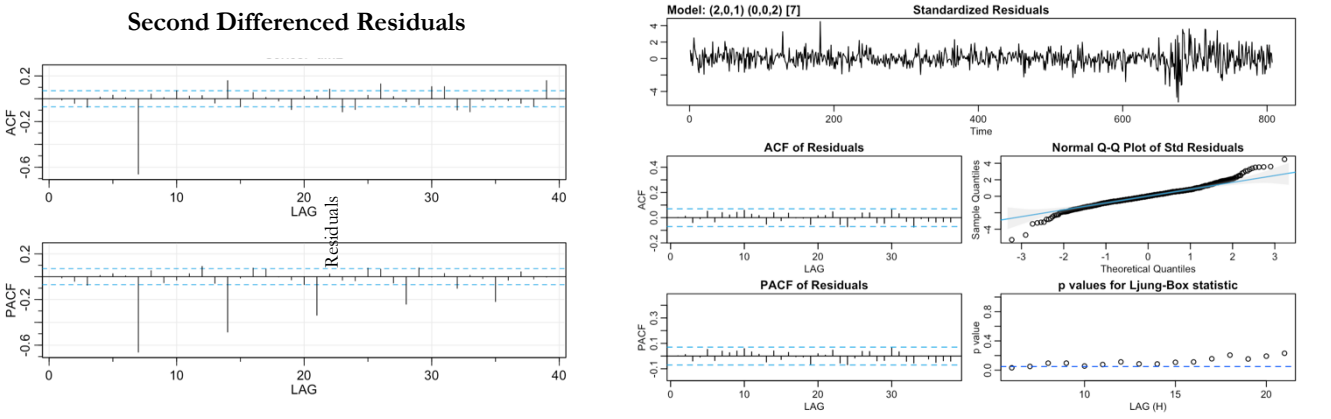


Figure 8: Left side graphs the acf and pacf plots of the residual plot in Figure 7. The right side shows results for sarima_wPACF( ) function for ARMA(2,1)x(0,2)[7]

Differencing with ARMA(0, 1)x(0,2)[7]

To help improve the previous model, perhaps the smaller spiked values in the acf plot aren't significant so in this model, p = 0 rather than 2. As illustrated in the AIC in Table 1, this model fits slightly better than the one above.

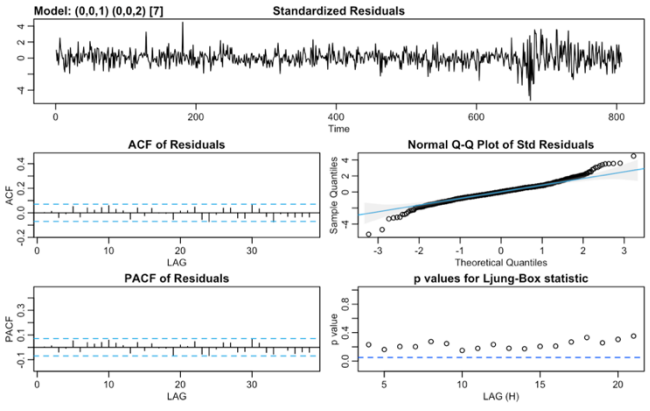


Figure 8: The plots to the right show results for sarima_wPACF( ) function for ARMA(0,1)x(0,2)[7]. The observed Ljung-Box statistics indicate a better fit than ARMA(2,1)x(0,2)[7]

Model Comparison and Selection

Comparing the AIC values of the two parametric models, the parametric model + ARMA(3,3)x(0,1)[7] shows the lowest AIC value. Comparing AIC values for the two differencing models, the differencing + ARAMA(0,1)x(0,2)[7] has the lowest AIC value and the best Ljung Box statistics p-values.

AIC values for the four models under consideration.

	AIC Values
Parametric Model + ARMA(1,2)	-5.865840
Parametric Model + ARMA(3,3)x(0,1)[7]	-6.267744
Second Order Differencing + ARMA(2, 1)x(0,2)[7]	0.912000
Second Order Differencing + ARMA(0, 1)x(0,2)[7]	0.911100

Table 1: AIC values for the four models under consideration. AIC values however cannot be compared across the differencing and parametric models.

The four models are also compared through time series cross validation. This cross validation rolls through the last 190 days in the data in 10 day windows. Thus, there will be 190 forecasted points that are tested against the last 190 points of original data. The root-mean-square prediction error, RMSPE, is used to determine which model produces the best fit. Table 2 below shows that the differencing with ARMA(0,1)x(0,2)[7] noise term is the best overall fit according to this cross-validation exercise, and therefore this model will be used for forecasting.

Cross-validated out-of-sample root mean squared prediction error for the four models under consideration.

	RMSPE
Parametric Model + ARMA(1,2)	6.674419
Parametric Model + ARMA(3,3)x(0,1)[7]	6.674237
Second Order Differencing + ARMA(2, 1)x(0,2)[7]	2.243087
Second Order Differencing + ARMA(0, 1)x(0,2)[7]	2.231680

Table 2: Cross validated out-of-sample root mean square prediction error for the four models under consideration. RMSPE can be compared across the parametric and differencing model.

Results

To forecast the next ten closing prices starting from Monday 9/21/2020, a differencing model will be used. The differencing mathematical equation model is shown in Equation (2). Let Y_t be the stock's closing price on day t with additive noise term X_t. As seen in the residual plot in Figure 7, X_t is a stationary process defined by ARMA(0,1)x(0,2)[7], where W_t is white noise with variance σ_W^2. Since the ARMA(0,1)x(0,2)[7] is applied, the ARMA model equation is shown in Equation(3) with one MA (θ) coefficient and two SMA (Θ_1, Θ_2) coefficients.

$$Y_t = Y_{t-1} + Y_{t-7} - Y_{t-1-7} + X_t + E(X_t) \tag{2}$$

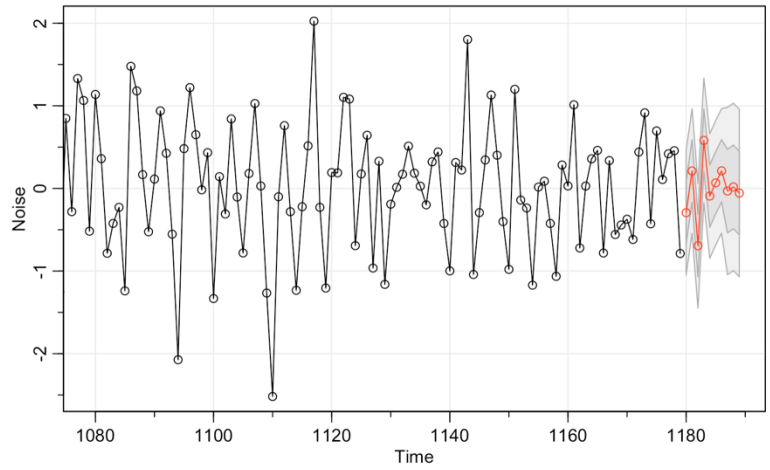
$$X_t = W_t + \theta W_{t-1} + \Theta_1 W_{t-7} + \Theta_2 W_{t-14} \tag{3}$$

Estimation of model parameters

Estimation of the model parameters are given in Table 3 of Appendix 1 along with the residual noise variance estimate. Sigma squared is relatively small, indicating the range of white noise terms isn't very large. It is interesting to note that the MA(1) and SMA(2) coefficients are all negative, with Θ_1, or the white noise term from 7 data points before seeming to have the most significant effect on the current noise term.

Forecasting

The first plot in Figure 9 below illustrates the sarima.for() function's next ten prediction values for the selected ARMA(0,1)x(0,2)[7] and their respective confidence intervals. As the confidence intervals show, the noise terms affect minimally affects the next ten closing price points. The following plot projects the next ten forecasted closing prices starting from Monday 9/21/2020. Predicted points land around the closing prices of the previous few points as the selected differencing model does not take into consideration values greater than lag 8.



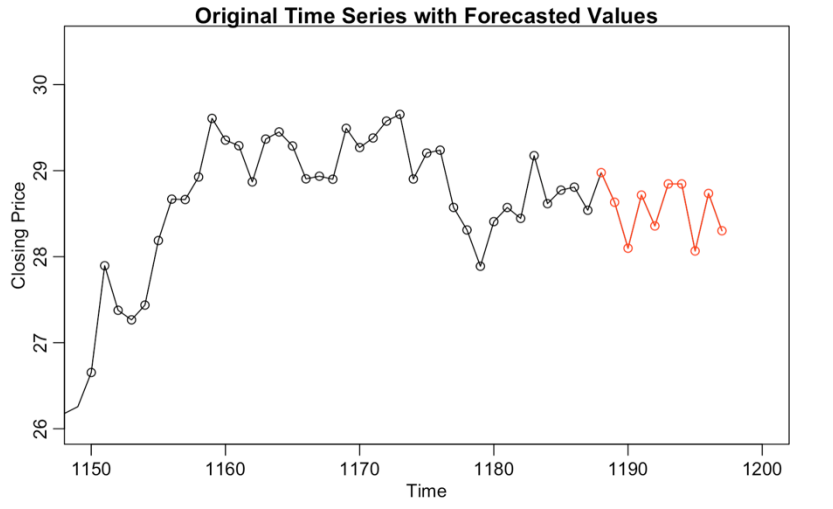


Figure 9: The first plot forecasts the next 10 X_t terms using the sarima.for( ) function as graphed in red. The black points are past second differenced residual values. The second plot forecasts the next ten closing price points for Stuff City Inc. starting from Monday, 9/21/2020. Historical closing prices are detailed in black while prediction values are detailed in red.

Appendix 1 – Table of Parameter Estimates

Table 3: Estimates of the forecasting model parameters in Equation (3) with their standard errors (SE).

Parameter	Estimate	SE	Description
θ	-0.0168	0.0279	MA coefficient
Θ_1	-0.8994	0.0288	Seasonal MA coefficient 1
Θ_2	-0.0727	0.0289	Seasonal MA coefficient 2
σ_W^2	0.0142		Variance of white noise