# COVID-19 Prediction Project

Danny Wu

5/10/2021

## 1. Executive Summary

Using a SARIMA(2,1,1)x(2,1,1)[7] model, we predict COVID-19 cases in the fifth borough of Gotham City to exhibit the same weekly seasonality as before, however we also expect cases to be higher than the what we have observed over past few weeks (with the exception of a massive spike in cases on January 7, 2021). City leadership should therefore increase the aid to provide greater support, preventing the system from being potentially overwhelmed as cases rise.

## 2. Exploratory Data Analysis

Daily COVID-19 cases have dramatically increased since March of last year. As seen in Figure 1, there is very strong trend in the data. Cases grew from March to August then slightly declined until November, where it rapidly grew and exceeded the levels before the decline. There is also weekly seasonality, as seen in the fluctuations which occur every seven days in Figure 1. It is also clear that the data exhibits heretoscedasticity as the variance of daily COVID-19 cases has been increasing over time.

There are a few anomalies in which entries will have zero counts with the following entry having an abnormally large number of cases (marked by the red and orange points respectively in Figure 1). These anomalies are likely due to cases being miscounted and accidentally moved to the next day. There are four of these instances which we correct for by dividing the number of counts on the second date in half and setting that as the number of counts for both dates.

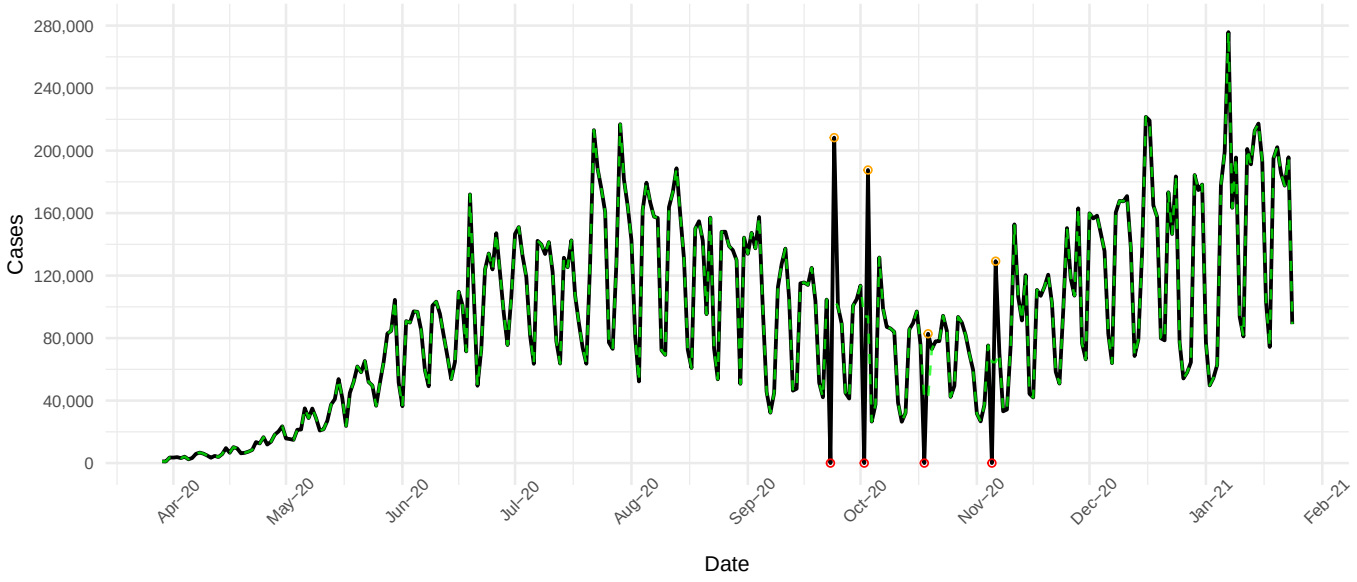


Figure 1: Daily COVID-19 cases from March 29, 2020 to January 24, 2021. Black solid line indicates original time series. Dashed green line represents corrected dataset with red and orange points indicating erroneous data points.

# 3. Models Considered

## 3.1 Parametric Signal Model

First we consider a parametric signal model. From a periodogram we see that there are two dominant fourier frequencies at 2/302 and 43/302 corresponding roughly to periods of period of 151 and 7 days. Thus we create a sinusoid with frequency 1/151 to model the larger trend, and then use indicators for each day of week to model the weekly seasonality. We also include terms for time and interaction between time and the day of the week in order to capture to the upwards trend of the data. To deal with the heteroskedasticity, we first log the data, and then also include an interaction term between the sinusoid, time, and the day of week indicators to capture the fluctuations of the sinusoud's magnitude over time. Our parametric model is as follows:

$$
\begin{aligned}
log(\text{Cases}_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \sum_{i=0}^{5} \Bigg[ & \beta_{3+6i} I_{\text{weekday}_{it}} + \beta_{4+6i} t I_{\text{weekday}_{it}} + \beta_{5+6i} I_{\text{weekday}_{it}} \cos\left(\frac{2\pi t}{151}\right) \\
& + \beta_{6+6i} I_{\text{weekday}_{it}} \sin\left(\frac{2\pi t}{151}\right) + \beta_{7+6i} t I_{\text{weekday}_{it}} \cos\left(\frac{2\pi t}{151}\right) + \beta_{8+6i} t I_{\text{weekday}_{it}} \sin\left(\frac{2\pi t}{151}\right) \Bigg] \\
& + \sum_{j=0}^{2} \left[ \beta_{39+2j} \cos\left(\frac{2\pi t}{151}\right) + \beta_{40+2j} \sin\left(\frac{2\pi t}{151}\right) \right] \qquad (1)
\end{aligned}
$$

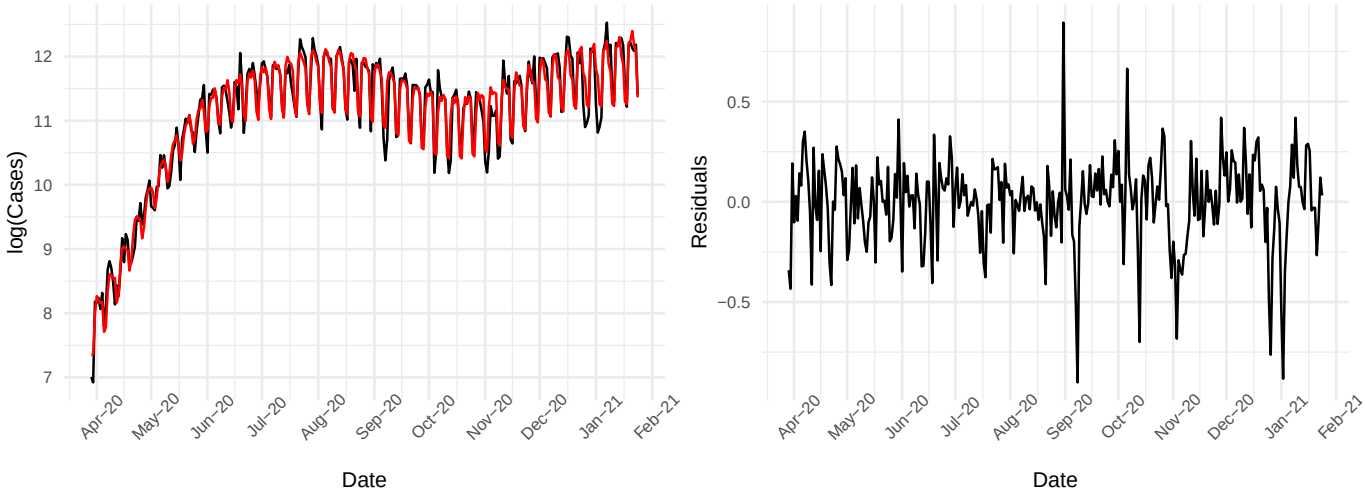


Figure 2: Left panel: Our parametric signal model with the fitted values plotted in red and the logged data plotted in black. Right panel: Residuals of the aforementioend parametric signal model. The residuals look relatively stationary.

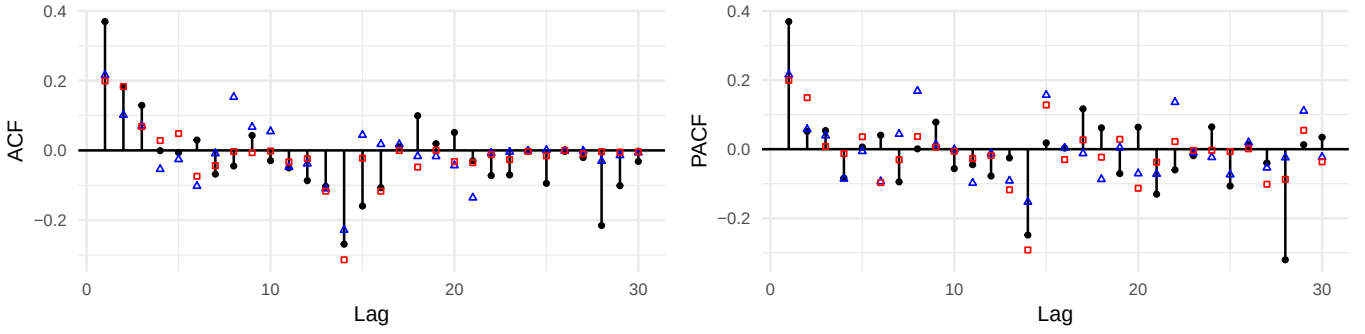


Figure 3: Sample ACF and PACF values for our logged parametric model (in black). Theoretical distributions for ARMA(0,3)x(2,1)[7] in blue triangles, and ARMA(2,2)x(1,2)[7] in red squares.

### 3.1.1 Parametric Signal Model with MSARMA(0,3)x(2,1)[7]

Looking at the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) values of our residuals in Figure 3, we observe large magnitudes at lags 1, 2, 3, 14 and 28 on the ACF plot and at lags 1, 14, and 28 in our PACF plot. The significant values in lags 1, 2, and 3 in the ACF plot the and the significant value at lag 1 in the PACF suggest an ARMA(1,3) fit. In addition, the 2 large magnitudes at lags 14 and 28 in both ACF and PACF plots suggest SARMA(2,2). Though the seasonal autocorrelation appears to begin at lag 14, through trial and error, we found that a period of 7 works much better, and through additional tweaking, we arrive at an MSARMA(0,3)x(2,1)[7]. Looking at the plot in Figure 4 we see that the p-values for the Ljung-Box statistic are all very insignificant, indicating that we cannot reject the hypothesis that the stationary process we observe was generated from this model.

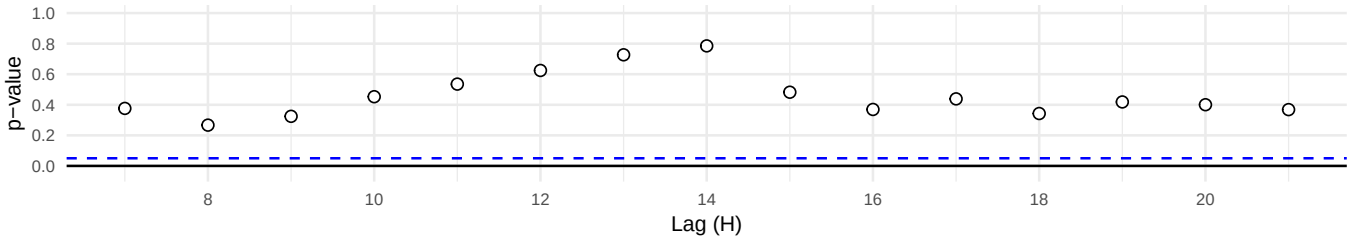


Figure 4: Plot of p-values for Ljung-Box statistic for Parametric Signal Model with MSARMA(0,3)x(2,1)[7]

### 3.1.2 Parametric Signal Model with MSARMA(2,2)x(1,2)[7]

The R function auto.arima() suggests an MSARMA(2,2)x(1,2)[7] model. This model is plausible, as the Ljung-Box statistics are all almost all of large magnitude as seen in Figure 5. In addition, in Figure 3, the theoretical ACF and PACF points from the MSARMA(2,2)x(1,2)[7] model, shown in red, appear to fit the sample ACF and sample PACF slightly better than the prior MSARMA(0,3)x(2,1)[7] model, further signaling good fit.

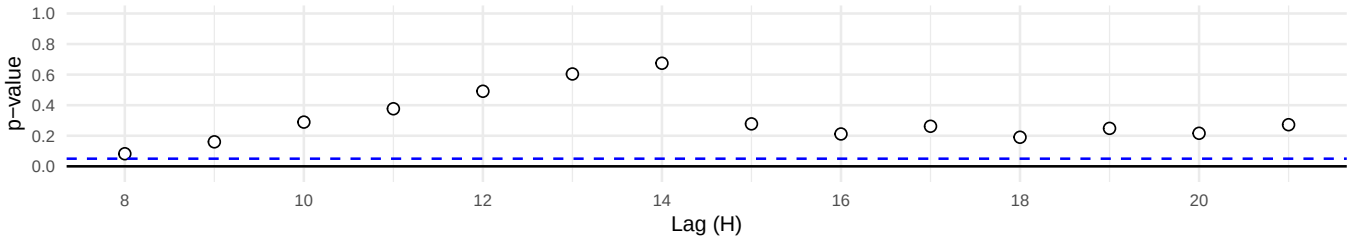


Figure 5: Plot of p-values for Ljung-Box statistic for Parametric Signal Model with MSARMA(2,2)x(1,2)[7]

## 3.2 Differencing Model

We now try a differencing approach. Because there exists heteroskedasticity, we apply a variance stabilizing transformation (VST) by logging all of the values. Since there is weekly seasonality, we applying differencing with a lag of 7. Looking at the residuals, there is still a slight downward trend so we apply differencing once again. Our resulting differenced data can be represented by the equation $\nabla_1\nabla_7 Log(\text{Cases}_t) = Log(\text{Cases}_t) - Log(\text{Cases}_{t-1}) - Log(\text{Cases}_{t-7}) + Log(\text{Cases}_{t-8})$ and is shown in Figure 6.

### 3.2.1 SARIMA(2,1,1)x(2,1,1)[7]

Looking at the ACF and PACF values for our differenced data in Figure 7, there are large magnitudes at lags 1 and 7 for the ACF plot. This suggests q = 1 and Q = 1 with a seasonal period of 7. In the PACF plot we see large magnitudes at lags 1, 2, 7 and 14, which suggests p = 2 and P = 2. Through some trial and error, we arrive at an MSARMA(2,1)x(2,1)[7], and as seen through the the Ljung-Box statistics in Figure 8, almost all p-values are

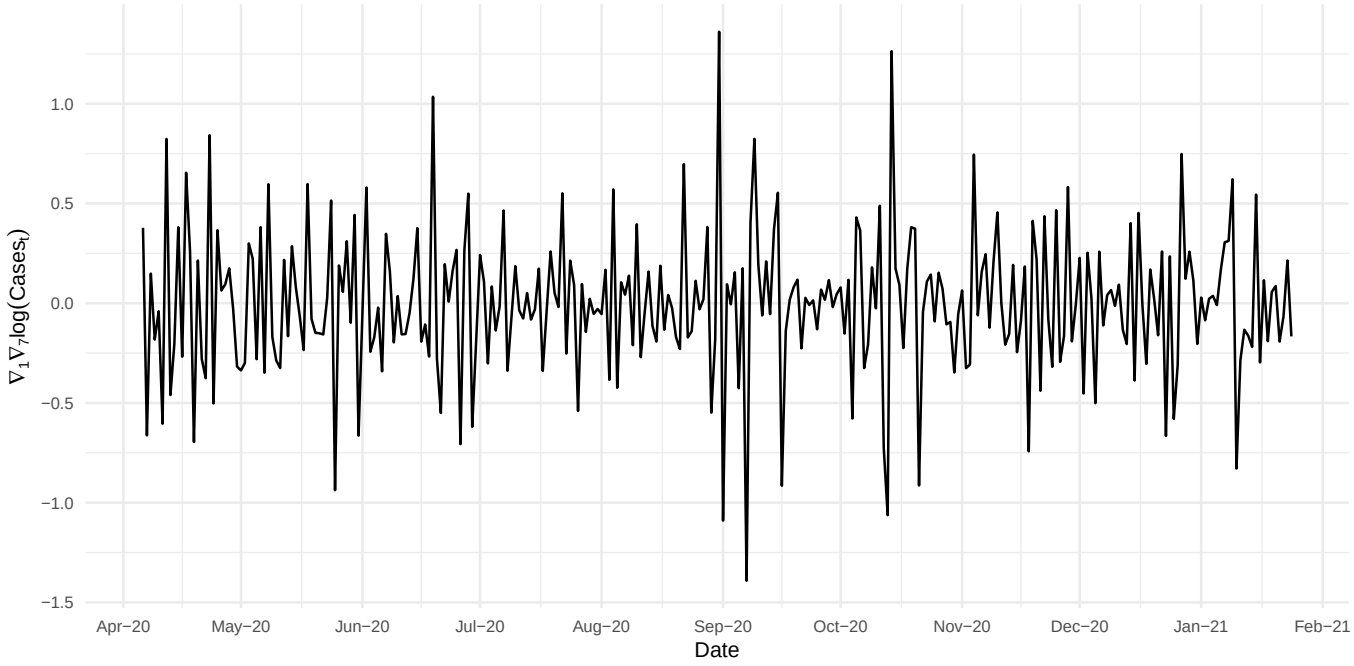


Figure 6: Plot of VST transformed data with lag 7 and lag 1 differencing applied. The result looks relatively stationary.

insignificant indicating this model is a good fit for the differenced data. We can represent our MSARMA noise model combined with differencing concisely as a SARIMA(2,1,1)x(2,1,1)[7] model for the entire dataset

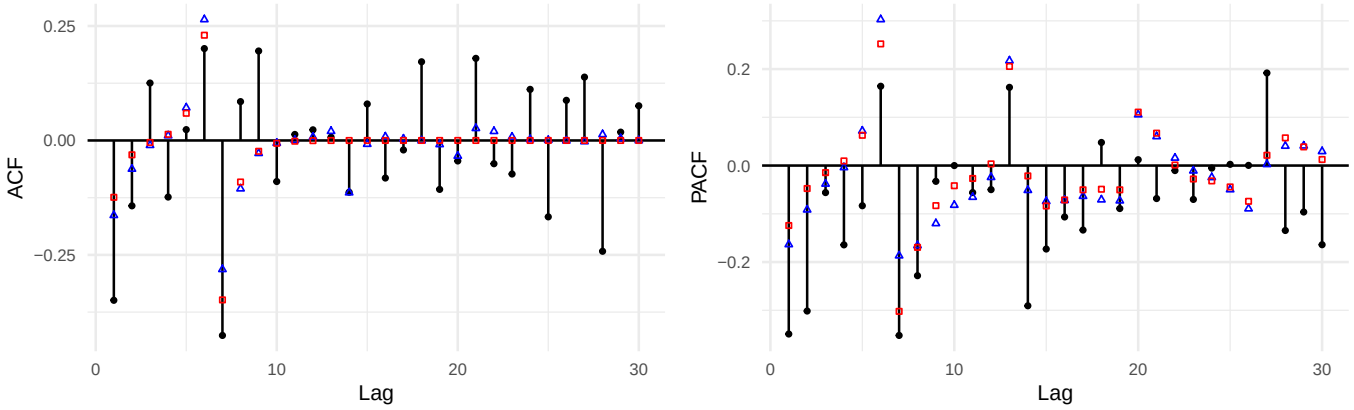


Figure 7: Sample ACF and PACF values from our differencing model. Theoretical distributions for MSARMA(2,1)x(2,1)[7] in blue triangles, and MSARMA(1,1)x(0,1)[7] in the red squares

### 3.2.2 SARIMA(1,1,1)x(0,1,1)[7]

For our second noise model on the differenced data, we try to simplify the prior MSARMA(2,1)x(2,1)[7] model while still satisfying the necessary diagnostics tests. After some trial and error, we arrive at an MSARMA(1,1)x(0,1)[7]. As seen in Figure 9, this model's p-values for the Ljung-Box statistic are all insignificant, indicating a good fit. The benefit of this model is that it is relatively less complex than the prior model by a magnitude of 3. We can represent our MSARMA noise model combined with differencing concisely as aSARIMA(1,1,1)x(0,1,1)[7] model for the entire process.

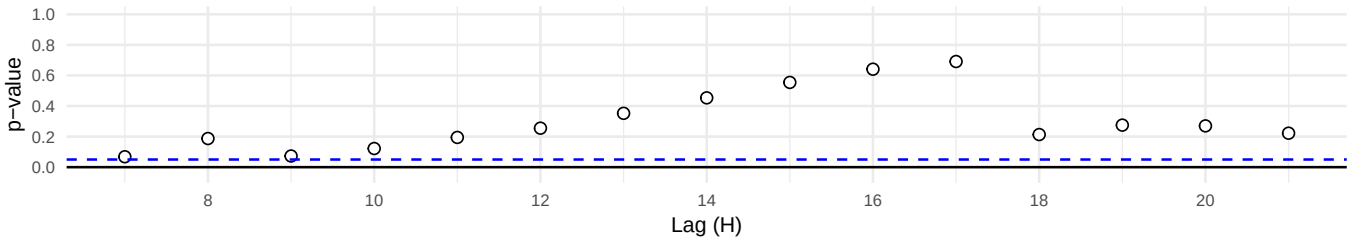


Figure 8: Plot of p-values for Ljung-Box statistic for SARIMA(2,1,1)x(2,1,1)[7]

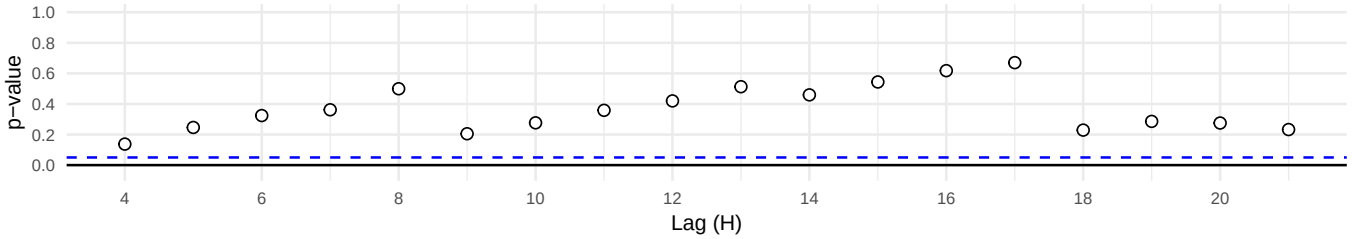


Figure 9: Plot of p-values for Ljung-Box statistic for SARIMA(1,1,1)x(0,1,1)[7]

# 4 Model Comparison and Selection

The four proposed models are compared through time-series cross validation. We roll through the last 20 weeks of data, from 9/6/20 to 1/24/21, in 7 day increments and forecast a week forward using all past data. For each of these 20 sets of 7 forecasts, we calculate the summed squared error (SSE) of that group of 7 forecasts. These values are then aggregated and used to calculate the root mean squared prediction error (RMSPE) for each of the models, listed in Table 1. We see that our SARIMA(2,1,1)x(2,1,1)[7] has the lowest RMSPE, and thus will be chosen as the model for predicting cases over the next 10 days.

Table 1: Cross-validated out-of-sample RMSPE for the four models under consideration.

| | RMSPE |
|---|---|
| Logged Parametric Model + ARMA(0,3)x(2,1)[7] | 48020.77 |
| Logged Parametric Model + ARMA(2,2)x(1,2)[7] | 47280.17 |
| VST + SARIMA(2,1,1)x(2,1,1)[7] | 34483.58 |
| VST + SARIMA(1,1,1)x(0,1,1)[7] | 34509.06 |

# 5 Results

The following SARIMA(2,1,1)x(2,1,1)[7] model will be applied to the logged data in order to forecast the next 10 days of COVID-19 cases. As this gives predictions in the logged scale, we will then exponentiate to produce our final forecasts.

$$Log(\text{Cases}_t) = Log(\text{Cases}_{t-1}) + Log(\text{Cases}_{t-7}) - Log(\text{Cases}_{t-8}) + X_t \tag{2}$$

$$\begin{aligned} X_t = {} & \phi_1 X_{t-1} + \phi_2 X_{t-2} + \Phi_1 X_{t-7} - \Phi_1\phi_1 X_{t-8} - \Phi_1\phi_2 X_{t-9} + \Phi_2 X_{t-14} - \Phi_2\phi_1 X_{t-15} \\ & - \Phi_2\Phi_2 X_{t-16} + W_t + \theta W_{t-1} + \Theta W_{t-7} + \Theta\theta X_{t-8} \end{aligned} \tag{3}$$

$$\hat{\mu}_X = -0.02716285 \tag{4}$$

## 5.1 Estimation of Model Parameters

The estimates of the model parameters are given in Table 2. Interestingly, we see that the coefficients on the MA and seasonal MA terms are very large in magnitude, indicating that the stationary process after differencing is very dependent on past values of the white noise. The AR and seasonal AR terms are not as large in magnitude, perhaps implying that past values of the stationary process itself are not as important.

Table 2: Estimates of the MSARMA(2,1)x(2,1)[7] model parameters in Equation (3)

| Parameter | Estimate | SE | Coefficient Description |
|---|---|---|---|
| $\phi_1$ | 0.2353 | 0.0885 | AR coefficient (1) |
| $\phi_2$ | -0.0248 | 0.0732 | AR coefficient (2) |
| $\theta$ | -0.7609 | 0.0702 | MA coefficient |
| $\Phi_1$ | -0.0065 | 0.0833 | Seasonal AR coefficient (1) |
| $\Phi_2$ | -0.1282 | 0.0761 | Seasonal AR coefficient (2) |
| $\Theta$ | -0.7789 | 0.0677 | Seasonal MA coefficient |
| $\sigma^2_W$ | 0.06029 | N/A | Variance of White Noise |

## 5.2 Prediction

Figure 10 shows the forecasts of COVID-19 cases for the next ten days from January 25, 2021 to February 3, 2021. The model predicts that cases will naturally exhibit weekly seasonality with lower cases on Sundays and Mondays, however we see that the cases during the middle of the week (Tuesday, Wednesday, Thursday) will be higher than almost all past levels of COVID cases (with exception of the peak on January 7, 2021). This upwards trend in the mid-week cases will continue during the following week with even higher numbers of cases.

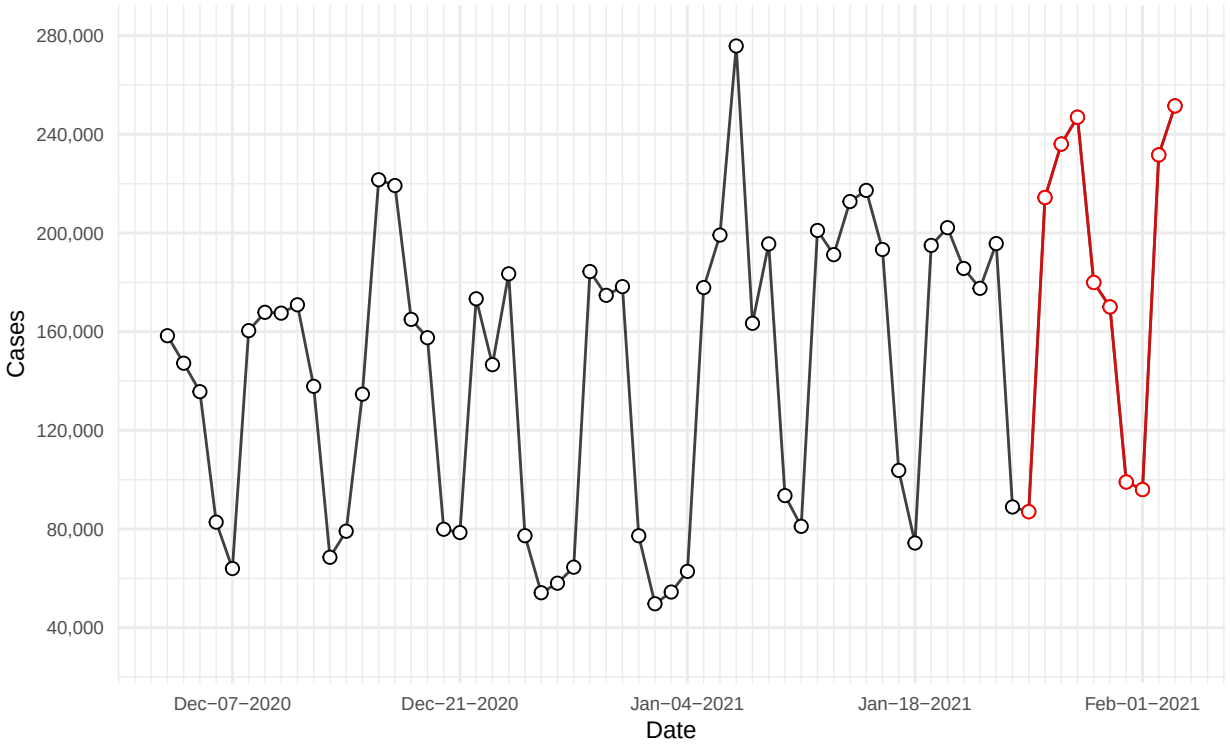


Figure 10: Forecasts of COVID-19 cases from January 25, 2021 to Feburary 3, 2021. The black points are obversed values, the red points are the forecasted values. Though truncated,