Final Project: Covid Cases

Suchir Joshi

May 10, 2021

1 Executive Summary

This dataset contains data of the daily number of new cases in Gotham City's fifth bureau. The stated goal is to forecast the number of new cases detected on each day for the next ten days, and two different modelling approaches are discussed below. Our chosen approach is a non-parametric differencing model, with lag-1 and lag-7 differencing to model the signal, and ARMA(1,1)x(1,1)[7] noise. This yields the best cross-validation score and will be used for prediction.

2 Explanatory Data Analysis

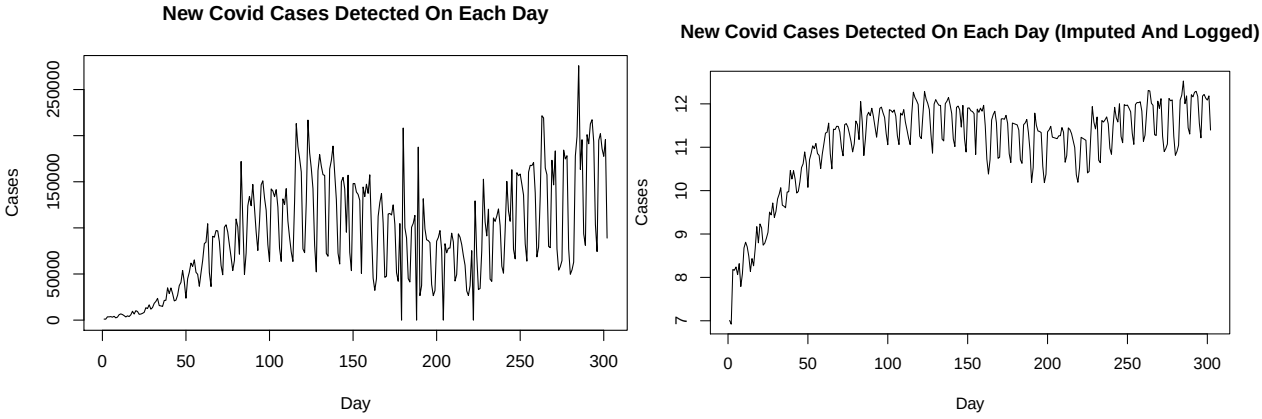

Figure 1: The new Covid-19 cases detected on each day in Gotham City. In the x-axis, Day 1 refers to March 29, 2020, as that is when the data collection began; Day 2 is March 30, 2020, and so on. The case numbers are shown on the y-axis. On the left is the raw data, and on the right is the imputed and logged data.

In Figure 1, there exists a plausibly cubic (degree-4) trend, as well as apparent weekly seasonality. This is possibly because some days of the week could be processing more cases than other days, so there could be an inherent weekly pattern in the specifics of data collection. Furthermore, the variance of cases seems to increase over time, which means a Variance Stabilizing Transform such as taking the natural log of the data could be useful.

Lastly, some of the days after Day 150 show 0 new cases, which seems to be an anomaly in the way the data was processed. To rectify this, the missing data has been imputed via the following scheme: if Day_T shows 0 new cases, the interval X of cases from Day_{T-1} to Day_{T+2} is considered. From this interval, we create the following table:

Percentile of X	Variable
0	A
33.33	B
66.67	C
100	D

We then assign Day_T to B, and Day_{T+1} to C. This imputation technique was chosen because it not only smooths out the overall trend, but also takes the anomously high next value into account. After performing this imputation, and the previously mentioned Log VST, the resulting plot appears more regular.

3 Models Considered

There are two components to each considered model: signal and noise. For modelling the signal, a parametric degree-4 polynomial trend with lag-7 differencing is considered, as is a non-parametric approach with lag-1 and lag-7 differencing. Both of these signal approaches will have two accompanying ARMA models each to model the remaining noise.

3.1 Quartic Parametric Signal Model With Seasonal Differencing

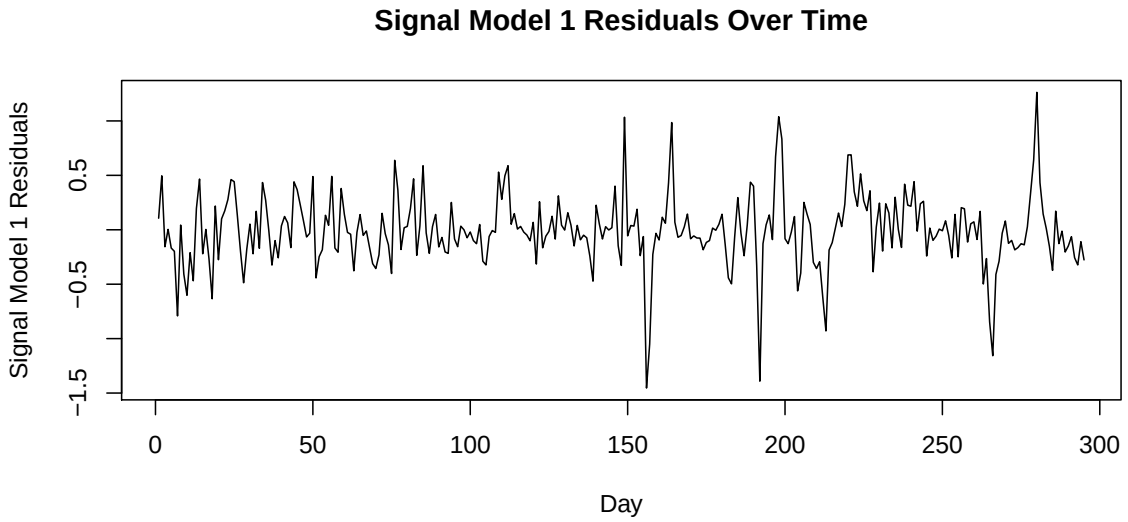

Figure 2: The residuals of the quartic polynomial model.

In Figure 2, the residuals visually appear quite stationary. Now, we examine the ACF and PACF plots of the residuals.

The ACF plot in Figure 3 exhibits 2 significant lags, after which it tapers off to zero in the form of a damped sin wave, reducing in amplitude. The PACF plot in Figure 3 more gradually tapers off towards 0; however, early on, there are significant values at every 7th lag. This suggests atleast one MA term with Q=1, S=7 should be used. Using the auto.sarima results to guide us, combined with empirical examination of different values, we find that additionally adding two AR terms yields a better fit. Below are two appropriate noise models.

(Note: for ease of forecasting, the non-differenced residuals will be used as input to the ARIMA models, which will handle the seasonality.)

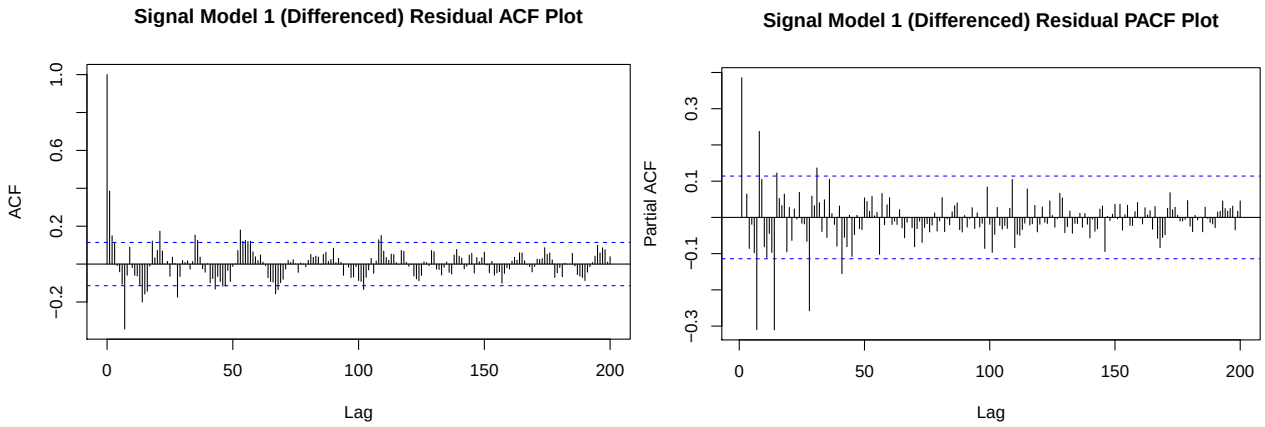

Figure 3: The ACF and PACF plot of the Model 1 Residuals.

3.2 Quartic Parametric Signal Model with ARMA(2,0,0)x(0,1,1)[7]

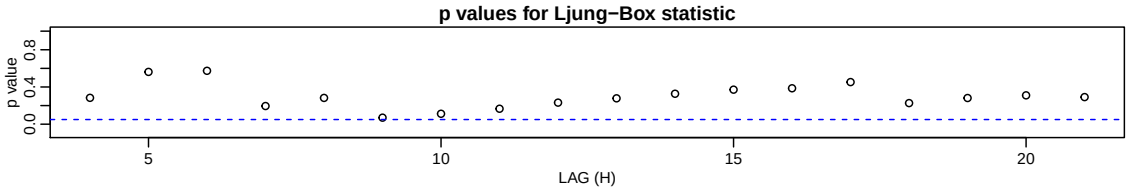

Figure 4: The SARIMA diagnostics for the Signal Model 1's Noise Model 1.

The diagnostics of this noise model in Figure 4 appear quite stationary, as the p-values for the Ljung-Box statistic are all non-significant, and virtually all of the ACF and PACF lags exhibit values within the confidence intervals of significance (figure not included).

3.3 Quartic Parametric Signal Model with ARMA(1,0,0)x(1,1,1)[7]

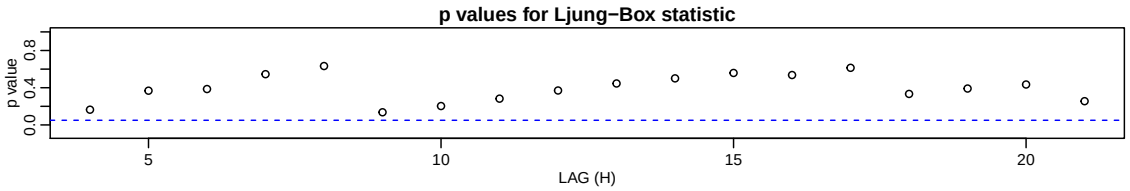

Figure 5: The SARIMA diagnostics for the Signal Model 1's Noise Model 2.

The diagnostics of this model in Figure 5 appear quite stationary as well. The p-values for the Ljung-Box statistic are all non-significant, and virtually all of the ACF and PACF lags exhibit values within the confidence intervals (figures not included). Therefore, we conclude that both noise models are good fits.

4 Differencing Signal Model

In Figure 6, the residuals of this model visually appear decently stationary. Now, we inspect the ACF and PACF plots of the residuals.

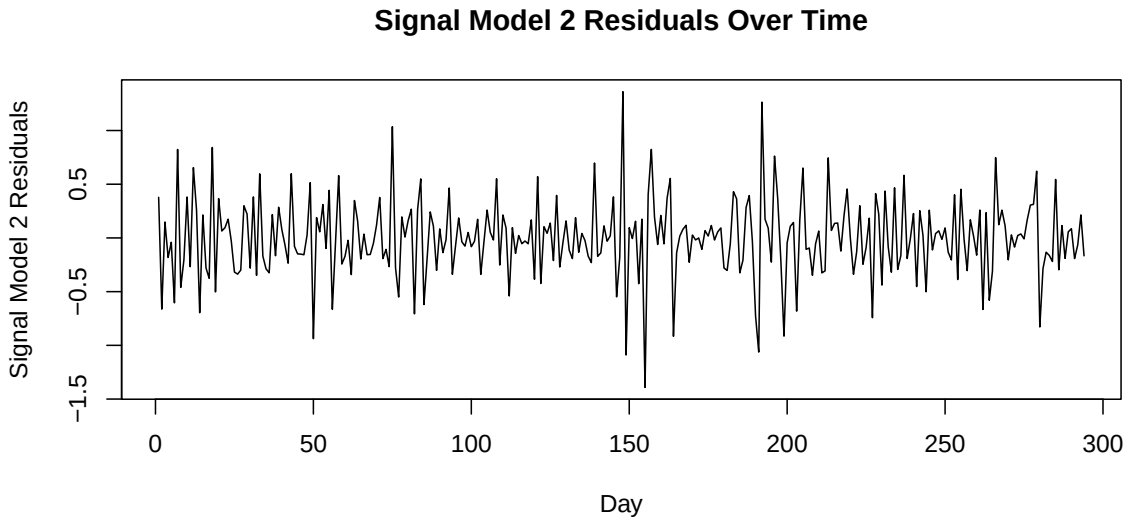

Figure 6: The residuals after differencing

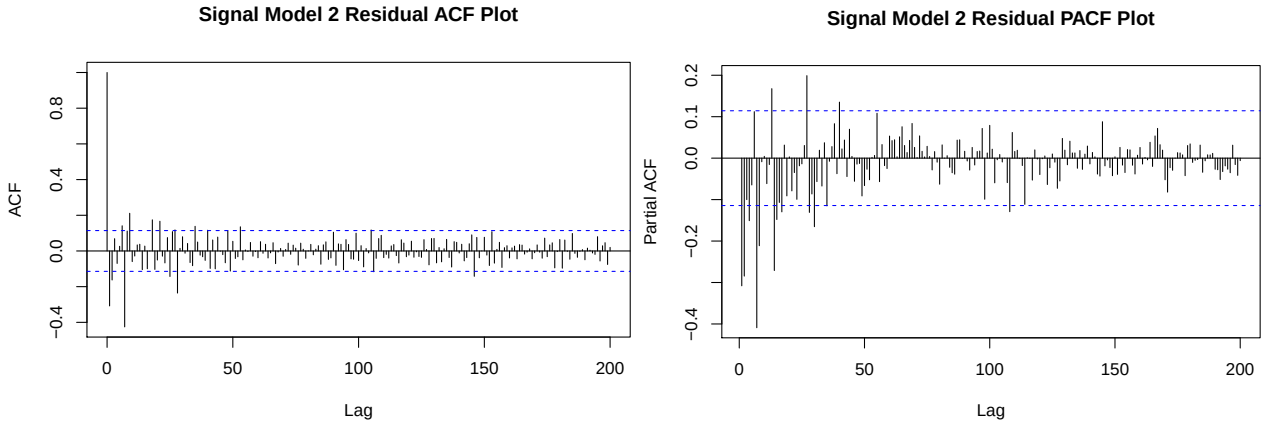

Figure 7: The ACF and PACF plot of the Model 2 Residuals.

The ACF plot in Figure 7 exhibits a few significant lags, especially at early multiples a=of 7, but thereafter the lags are largely all within the confidence bands. The PACF plot in Figure 7, on the other hand, more gradually tapers off towards 0; it resembles a damped sin wave decreasing in amplitude. This suggests that atleast one AR term with P=1, S=7 should be used. In addition, the auto.sarima results guide our empirical investigation, where we find that adding more MA terms yields a better fit. Below are two chosen noise models.

4.1 Differencing Model With ARMA(1,0,1)x(2,0,1)[7]

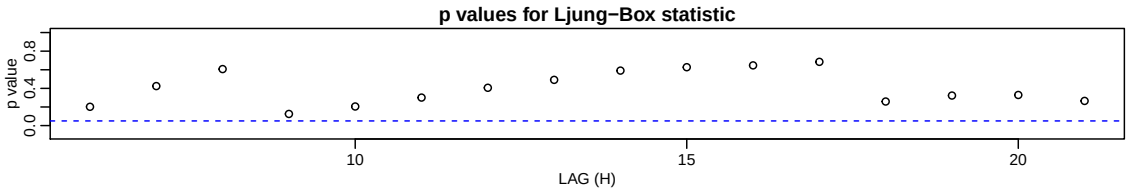

Figure 8: The SARIMA diagnostics for the Signal Model 2's Noise Model 1.

After examining the diagnostics of this noise model in Figure 8, the residuals appear quite stationary. The p-values for the Ljung-Box statistic are all non-significant, and virtually all of the ACF and PACF lags are contained within the confidence intervals of significance (figure not included).

4.2 Differencing Model With ARMA(0,0,2)x(2,0,1)[7]

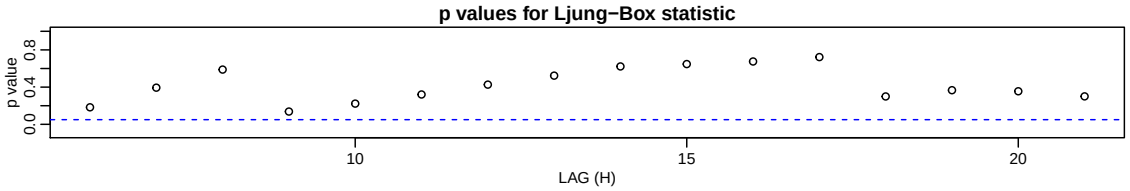

Figure 9: The SARIMA diagnostics for the Signal Model 2's Noise Model 2.

After examining the diagnostics of this noise model in Figure 9, the residuals appear quite stationary. The p-values for the Ljung-Box statistic are all non-significant, and virtually all of the ACF and PACF lags are contained within the confidence intervals of significance (figure not included).

5 Model Comparison and Selection

Time series cross validation is used to compare the four candidate models. The testing sets are non-overlapping and roll through the last 140 days in the data, from day 162 to day 302, or 9/6/20 to 1/24/21. This is done in 10 day intervals, and the training sets comprise of all data occuring prior to the start of the respective testing set. To estimate and compare model performance in forecasting, we use RMSE, or root-mean-square error; the model with the lowest total RMSE will be selected to forecast future cases.

As seen in Table 1, the candidate model with the lowest cross-validated RMSE is the differencing model with ARMA(1,0,1)x(2,0,1)[7]. However, all of them have quite similar RMSE. Therefore, this differencing model with ARMA(1,0,1)x(2,0,1)[7] is selected to forecast future cases.

Table 2: Out-of-sample root mean squared error during cross-validation for the four candidate models.

	RMSE
Quartic Parametric Model + ARMA(2,0,0)x(0,1,1)[7]	51255.87
Quartic Parametric Model + ARMA(1,0,0)x(1,1,1)[7]	50603.40
Differencing Model + ARMA(1,0,1)x(2,0,1)[7]	50395.66
Differencing Model + ARMA(0,0,2)x(2,0,1)[7]	50410.92

6 Results

We propose the following non-parametric model for forecasting future cases. Let Cases_t be the number of new cases detected on day t, and let X_t be a noise term defined by ARMA(1,0,1)x(2,0,1)[7]. W_t is a white noise term, with variance σ_W^2. Lastly, let log denote the natural logarithm.

$\log(\text{Cases}_t) = \log(\text{Cases}_{t-1}) + \log(\text{Cases}_{t-7})$ - $\log(\text{Cases}_{t-8}) + X_t$, where

$$X_t = \phi X_{t-1} + \Phi X_{t-7} - \phi\Phi X_{t-8} + W_t + (\theta + \Theta_1)W_{t-1} + (\theta\Theta_1 + \Theta_2)W_{t-2} + (\Theta_2\ \theta)W_{t-1}$$

6.1 Estimation of Model Parameters

We provide estimates of the model parameters in Appendix 1, Table 2.

6.2 Prediction

Here are the model predictions for the number of new cases detected on each of the next ten days. Overall, the model's predictions seem to be in line with the general weekly seasonality observed, but it doesn't predict this recent upward trend in cases to continue. In fact, it predicts a plateau, if not slight decrease, in the number of new cases going forward. This bodes cautiously well for the residents of Gotham City's fifth bureau.

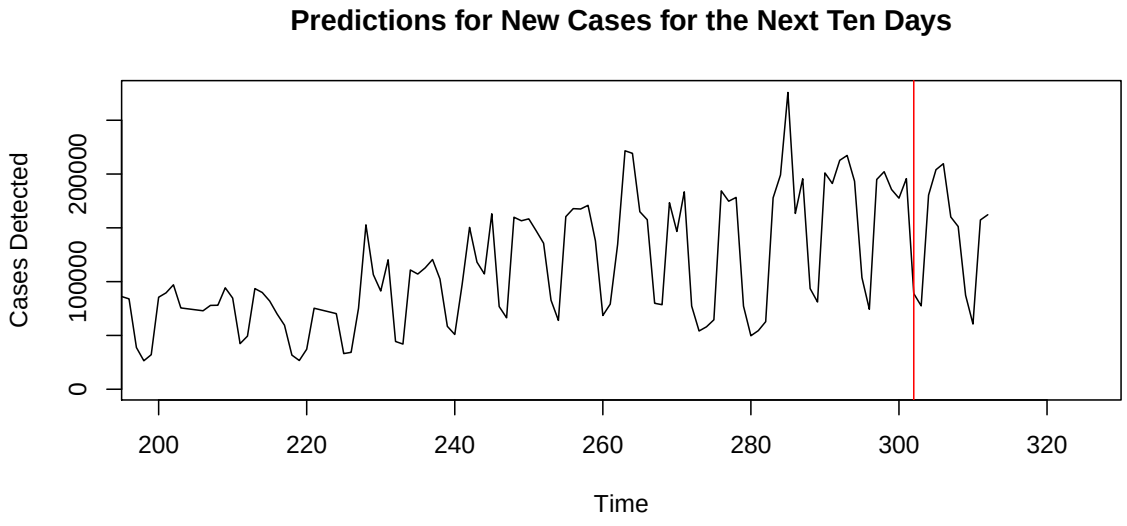

Figure 10: The model's prediction of new cases detected for the next ten days. The last date for which data is available is depicted by the vertical red line. The x-axis is time in days, and the y-axis is cases.

7 Appendix 1 - Parameter Estimates Table

Table 2: Parameter estimates of the model that was used to forecast.

Parameter	Estimate
ϕ	0.2978
θ	-0.8024
Φ	-0.8357
Θ_1	0.0462
Θ_2	-0.0981
σ_W^2	0.0625