Projection of Covid-19 Cases in Gotham City

Annie Dang

5/9/2021

1 Executive Summary

Covid-19 cases remain high in the fifth borough of Gotham City and resources are running low. Our goal is to predict the number of Covid-19 cases in the next couple of days so that our city leadership can better allocate aid and resources. Historically, there have been less cases in the spring compared to other seasons, and more cases during the weekdays compared to weekends. According to our parametric model with ARMA(1,0)x(1,0)[10] noise, the amount of cases in the next following day will remain high and we will see some of our highest number of cases. These cases, however, reach its peak and then drop by the thousands after a couple of days, as we have seen historically.

2 Exploratory Data Analysis

There are two seasonal trends for Covid-19 cases in Gotham City: one with a shorter period and another with a longer period, as seen in the left panel of Figure 1. Cases also seem to have an increasing trend over time and the variance of the cases look to be changing with time.

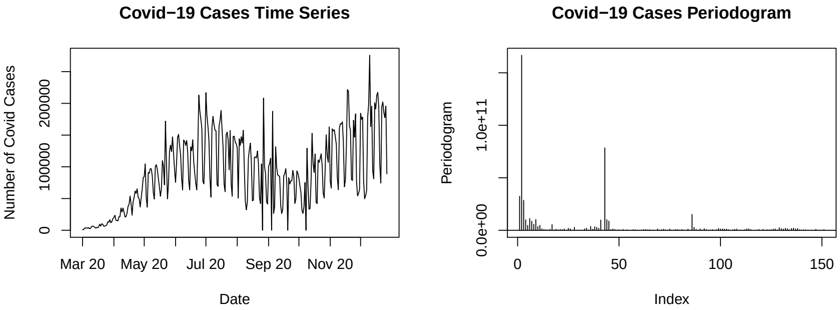

Figure 1: Daily counts of Covid-19 cases in the fifth borough of Gotham City, starting from 3/29/20. In the right panel is the periodogram of the cases.

Upon closer inspection using the periodogram in the right panel of Figure 1, we can see large values for index j = 2 and j = 43. This implies a period of 151 and 7.02 respectively. This may imply 151-days and weekly seasonal patterns of Covid-19 cases.

3 Models Considered

To model the data, a parametric model and a differencing model was used.

3.1 Signal Model 1: Parametric Model

The data seems like it may have some quadratic and cubic trends, thus to account for this, time-squared and time-cubed variables are added to this model. An indicator variable that indicate whether the season is Spring and an indicator that indicates whether the day is in the weekend are also added to this model. There are also interactions between sinusoids with periods 151 and 7, time, the Spring indicator variable, and the weekend indicator variable. Based on Figure 1, the variance seems to be increasing with time, in a manner that is not quite linear. Thus, a log variance stabilizing transformation is applied to the data. The signal model is detailed in Equation (1) below, where X_t a noise term.

$$\begin{aligned} log(Cases_t) = {} & \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 I_{\text{spring}_t} + \beta_5 I_{\text{weekend}_t} + \beta_6 cos\left(\frac{2\pi t}{151}\right) + \beta_7 sin\left(\frac{2\pi t}{7}\right) \\ & + \beta_8 I_{\text{spring}_t} I_{\text{weekend}_t} + \beta_9 t I_{\text{weekend}_t} + \beta_{10} t I_{\text{spring}_t} + \beta_{11} t I_{\text{spring}_t} I_{\text{weekend}_t} \\ & + (I_{\text{weekend}_t} + I_{\text{spring}_t} + t + I_{\text{weekend}_t} I_{\text{spring}_t} + t I_{\text{weekend}_t} + t I_{\text{spring}_t} + t I_{\text{spring}_t} I_{\text{weekend}_t}) \times \\ & \sum_{j=1}^{7} (\beta_{2j+10} cos\left(\frac{2\pi t}{151}\right) + \beta_{2j+11} sin\left(\frac{2\pi t}{7}\right)) + X_t \end{aligned} \tag{1}$$

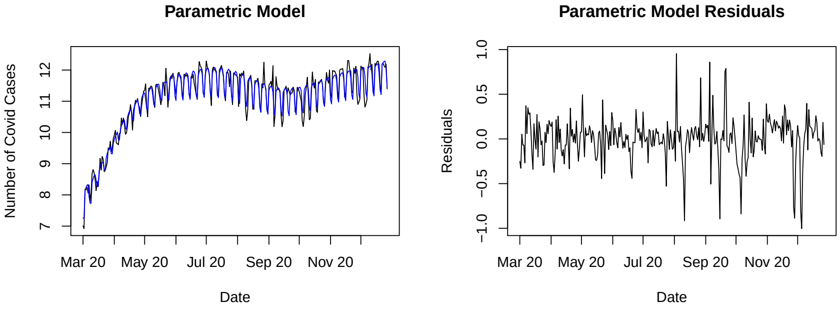

Figure 2: The parametric signal model. The left panel shows the model's fitted values in blue and the actual logged transformed data in black. The right panel shows the residuals of this model.

3.1.1 Parametric Signal with ARMA(1,0)x(1,0)[10]

The ACF and PACF plots for the parametric model residuals are shown in Figure 3. For the autocorrelation function (ACF), there are large values at lag 1. For the partial autocorrelation function (PACF), there are large values at lag 10 and lag 1. The large PACF and ACF values at lag 1 may imply q=1 or p=1. The large PACF value at lag 10 suggest that there is also a seasonal auto-regressive component (P=1) with period 10 (S=10). Thus, the final model proposed is ARMA(1,0)x(1,0)[10]. As shown by Figure 4, this ARMA model fits relatively well. The ACF values of the residuals are white noise for the most part, with one or two values just slightly surpassing the 95% confidence bands. Almost all of the p-values are insignificant in the Ljung-Boxplot, meaning there isn't enough evidence to reject the hypothesis that the data was generated from this ARMA model. All things considered, this is a relatively good fit.

3.1.2 Parametric Signaling with ARMA(1,1)x(1,0)[10]

As seen by the original ACF and PACF plots (Figure 3), the PACF and ACF values seem to taper off as lag values increase, showing a diminishing effect. This suggest that p>0 and q>0. Thus, for this model, p=1 and q=1 is used. To account for the large PACF value at lag 10, a seasonal auto-regressive element of period 10 or S=10 is also added to this model. As such, the final model is ARMA(1,1)x(1,0)[10]. The

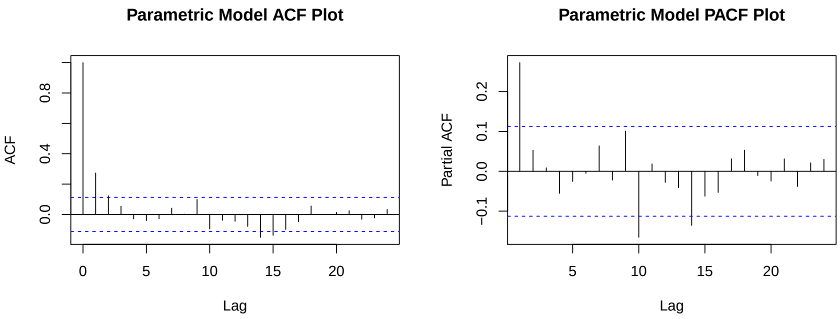

Figure 3: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the parametric signal model's residuals on the left and and right panel respectively.

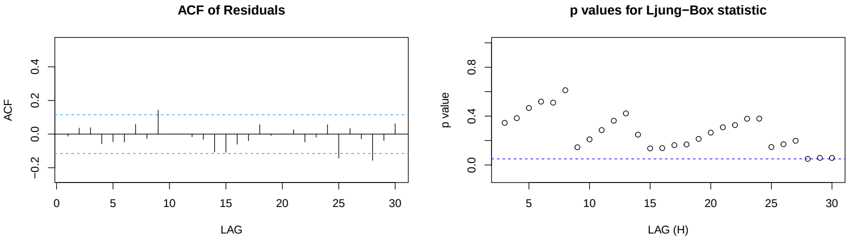

Figure 4: Diagnostic plots for ARMA(1,0)x(1,0)[10] on the parametric signal model. On the left panel is the ACF of residuals and on the right panel is a plot of the p-values for the Ljung-Box statistic for this ARMA model.

Ljung-Boxplot in Figure 5 shows that most of the p-values are insignificant, however some do fall slightly below the significance threshold. There are also one or two slightly significant residual ACF values. Taken all together, this model fits relatively well but most likely isn't the best fit for this data.

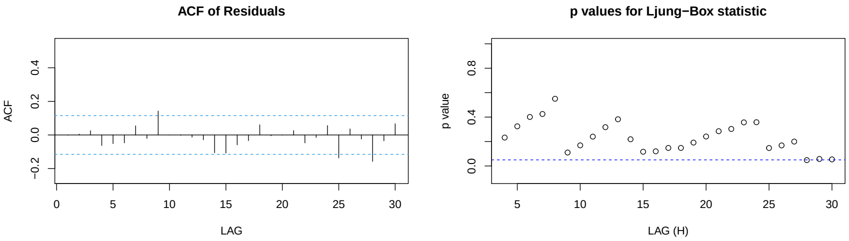

Figure 5: ACF and Ljung-Box plots for ARMA(1,1)x(1,0)[10] on the parametric signal model in the left and right panels respectively.

3.2 Signal Model 2: Differencing

As shown by the periodogram of the data, there are weekly effects and effects every 151 days. Because of such, lag-151 and lag-7 differencing is used. This is now twice differenced data that will remove any linear and quadratic trends. The differenced data, as shown by the right panel of Figure 6, looks relatively stationary. This differenced model is detailed in Equation (2) below.

$$\nabla_7\nabla_{151}\mathrm{Y}_t = \mathrm{Y}_t - \mathrm{Y}_{t-7} - \mathrm{Y}_{t-151} + \mathrm{Y}_{t-144} \tag{2}$$

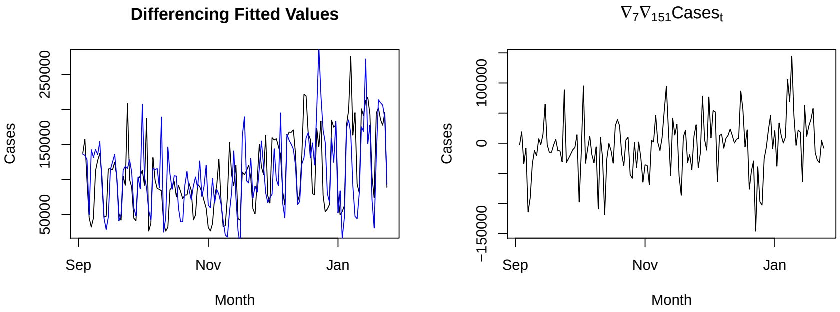

Figure 6: The differenced signal model. The left panel shows the data in black and the fitted values in blue. The right panel is the differences, used to assess stationarity.

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

The ACF and PACF for these differences are shown in Figure 7. The ACF and PACF both show a diminishing effect, which suggest p>0 and q>0; p=q=1 might be reasonable. After assessing the ACF of residuals for ARMA(1,1) of this differencing model, there was a large ACF value at lag 7, suggesting S=7 and Q=1 or SMA(1)[7]. After implementing SMA(1)[7] with ARMA(1,1), our final model becomes ARMA(1,1)x(0,1)[7]. As shown by Figure 8, this ARMA model fits well. The ACF values are white noise and fall in between the confidence bands and all of the p-values are insignificant in the Ljung-Boxplot.

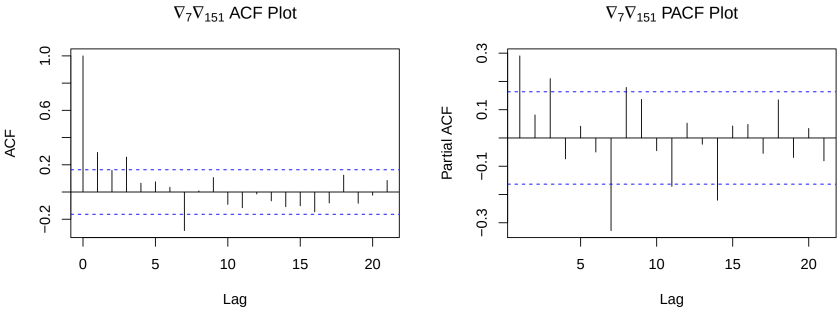

Figure 7: ACF and PACF values for the differencing model on the left and and right panel respectively.

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

The PACF value of the differencing model displayed values of large magnitude at lag 7 and 14, which could mean S=7 and P=2 or SAR(2)[7]. After implementing this with ARMA(1,1), we get ARMA(1,1)x(2,0)[7]. As shown by Figure 9, this ARMA model fits well. The ACF values are white noise and all of the p-values are insignificant in the Ljung-Boxplot.

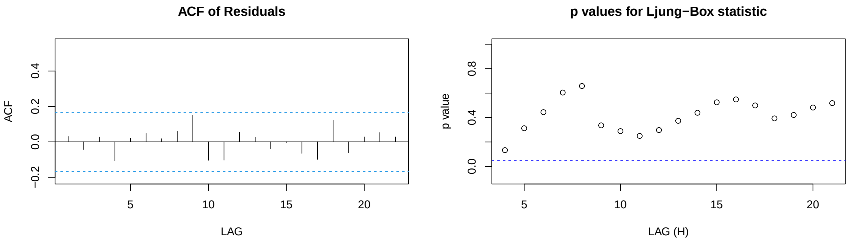

Figure 8: ACF and Ljung-Box plots for ARMA(1,1)x(0,1)[7] on the differencing model in the left and right panels respectively.

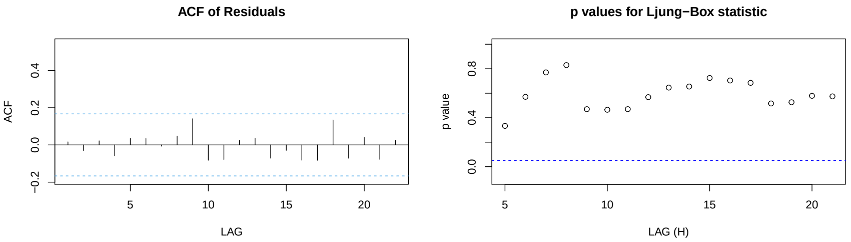

Figure 9: ACF and Ljung-Box plots for ARMA(1,1)x(2,0)[7] on the differencing model in the left and right panels respectively.

4 Model Comparison and Selection

The two parametric and two differencing models are compared using time series cross validation. The non-overlapping testing sets roll through the last 120 days in intervals of 10 days. That is, there will be forecasts, provided in windows of 10 days segments, for dates 9/25/20 through 1/24/21. The training set will contain data points from all dates prior to the respective testing set. To assess the model's forecasting performances, root mean squared error (RMSE) is used and the model with the lowest RMSE value will be selected for predicting Covid-19 cases for the next 10 days. Table 1 shows that the parametric model with ARMA(1,0)x(1,0)[10] has the lowest cross-validated RMSE value and thus will be used as the final model for forecasting.

Table 1: Cross-validated out-of-sample root mean squared error for the four models under consideration.

	RMSE
Parametric Model + ARMA(1,0)x(1,0)[10]	45389.99
Parametric Model + ARMA(1,1)x(1,0)[10]	45440.38
151-Day Differencing + Weekly Differencing + ARMA(1,1)x(0,1)[7]	48694.24
151-Day Differencing + Weekly Differencing + ARMA(1,1)x(2,0)[7]	48955.91

5 Results

To forecast sales for the next 10 days after 1/24/21, a parametric model of time will be used. Let $log(Cases_t)$ be the log-transformed number of cases on day t with additive noise term X_t as previously shown in Equation (1). Let X_t be a stationary process defined by ARMA(1,0)x(1,0)[10], where W_t is the white noise term with variance σ_W^2. The final model is stated below as Equation (3).

$$\begin{aligned}
\log(Cases_t) = {} & \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 I_{\text{spring}_t} + \beta_5 I_{\text{weekend}_t} + \beta_6 cos\left(\frac{2\pi t}{151}\right) + \beta_7 sin\left(\frac{2\pi t}{7}\right) \\
& + \beta_8 I_{\text{spring}_t} I_{\text{weekend}_t} + \beta_9 t I_{\text{weekend}_t} + \beta_{10} t I_{\text{spring}_t} + \beta_{11} t I_{\text{spring}_t} I_{\text{weekend}_t} \\
& + (I_{\text{weekend}_t} + I_{\text{spring}_t} + t + I_{\text{weekend}_t} I_{\text{spring}_t} + t I_{\text{weekend}_t} + t I_{\text{spring}_t} + t I_{\text{spring}_t} I_{\text{weekend}_t}) \times \\
& \sum_{j=1}^{7} (\beta_{2j+10} cos\left(\frac{2\pi t}{151}\right) + \beta_{2j+11} sin\left(\frac{2\pi t}{7}\right)) + X_t \\
X_t = {} & \phi X_{t-1} + \Phi X_{t-10} - \phi \Phi X_{t-11} + W_t
\end{aligned} \tag{3}$$

There are two binary indicators in this model. I_{weekend_t} indicates if day t is a weekend day (Sunday, Saturday, and also includes Monday). I_{spring_t} indicates if day t is in the Spring season. ϕ, Φ, and all of the β's are coefficients of the model that will estimated and provided in the next subsection.

5.1 Estimation of model parameters

Estimates of the model parameters are given in Table 2 in Appendix 1. It's interesting to note that the Spring season averages -4 times more cases than it would any other season.

5.2 Prediction

Figure 10 shows the forecasted values of Covid-19 cases on the ten days following 1/25/21. The model predicts that a peak of Covid-19 cases is near, with the highest cases on Tuesdays, Wednesdays, Thursdays, and Fridays, which reflects our initial intuition. Although this is contributing to the increasing trend in Covid-19 cases, the maximum predicted value is not the highest of this dataset.

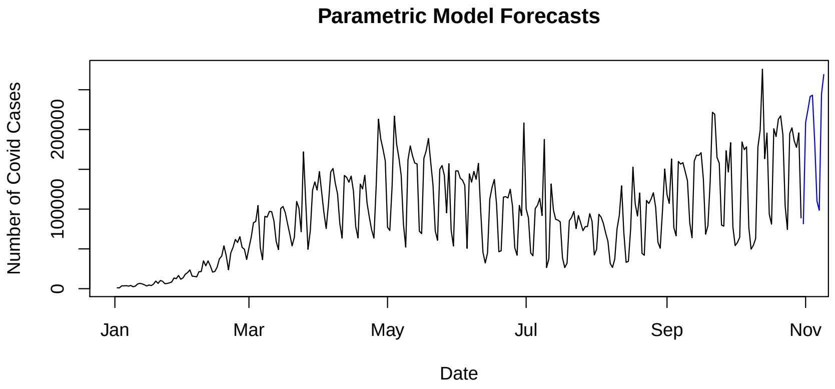

Figure 10: Forecasts of Covid-19 cases for Gotham City. The x-axis is time in months. In black are the recent historical Covid-19 data and in blue are the forecasts for the 10 days following 1/24/2021.

Appendix 1 - Table of Parameter Estimates

Table 2: Estimates of the forecasting model parameters in Equation (3), with their standard errors (SE), and a description of the coefficients.

Parameter	Estimate	SE	Coefficient Description
β_0	14.53e+01	1.198e+00	Intercept
β_1	-3.401e-02	2.158e-02	Time
β_2	9.188e-05	1.207e-04	Time Squared
β_3	-2.845e-09	2.118e-07	Time Cubed
β_4	-4.984e+00	1.053e+00	Spring Season
β_5	1.127e-01	1.793e-01	Weekend
β_6	1.018e+00	1.630e-01	Cosine
β_7	1.018e+00	1.381e-01	Sine
β_8	-9.587e-01	5.847e-01	Spring× weekend interaction
β_9	-1.039e-03	8.775e-04	time× weekend interaction
β_{10}	4.524e-02	1.541e-02	time × spring interaction
β_{11}	-9.631e-02	1.697e-01	Cosine × weekend interaction
β_{12}	-6.075e-01	2.478e-01	Sine × weekend interaction
β_{13}	-2.483e+00	3.904e-01	Cosine× spring interaction
β_{14}	1.179e-01	1.955e-01	Sine × spring interaction
β_{15}	-4.323e-03	8.178e-04	Cosine × time interaction
β_{16}	4.988e-04	6.813e-04	Sine × time interaction
β_{17}	2.537e-02	1.477e-02	time × spring × weekend interaction
β_{18}	3.407e-01	4.661e-01	Cosine × spring × weekend interaction
β_{19}	3.262e-01	3.543e-01	Sine × spring × weekend interaction
β_{20}	1.217e-04	8.186e-04	time × cosine × weekend interaction
β_{21}	-1.798e-04	1.215e-03	time × sine × weekend interaction
β_{22}	1.614e-02	4.696e-03	time × cosine × spring interaction
β_{23}	-2.332e-04	2.923e-03	time × sine × spring interaction
β_{24}	1.189e-02	4.771e-03	time × cosine × weekend × spring interaction
β_{25}	-8.102e-04	5.270e-03	time × sine × weekend × spring interaction
ϕ	0.2882	0.0554	AR coefficient
Φ	-0.1326	0.0577	Seasonal AR coefficient
σ_W^2	0.05458		Variance of White Noise