# COVID-19 Forecast: Gotham City, Fifth Burrough

Jack Stehn

5/10/2021

## 1 Executive Summary

With resources running low, Gotham City needs predictions of new cases for the next ten days in each burrough so it can strategically allocate aid. COVID-19 cases in the fifth burrough of Gotham City are expected to remain high. A differencing model (Specified by a first difference and a lag 7 difference) with ARMA(2,1)x(0,2)[7] noise was used to forecast cases. It appears the recent wave has already peaked and flattened.

## 2 Exploratory Data Analysis

The data on COVID-19 cases started in April 2020. It shows a overall upward trend that has flattened out. It also has two major waves peaking in late July and late January as demonstrated in the left panel of Figure 1. There is a strong weekly seasonal pattern as demonstrated in the right panel of Figure 1. This is likely reflective of testing practices rather than the actual spread of the virus. Also of note is that the data is heteroscedastic with variance linked to mean.

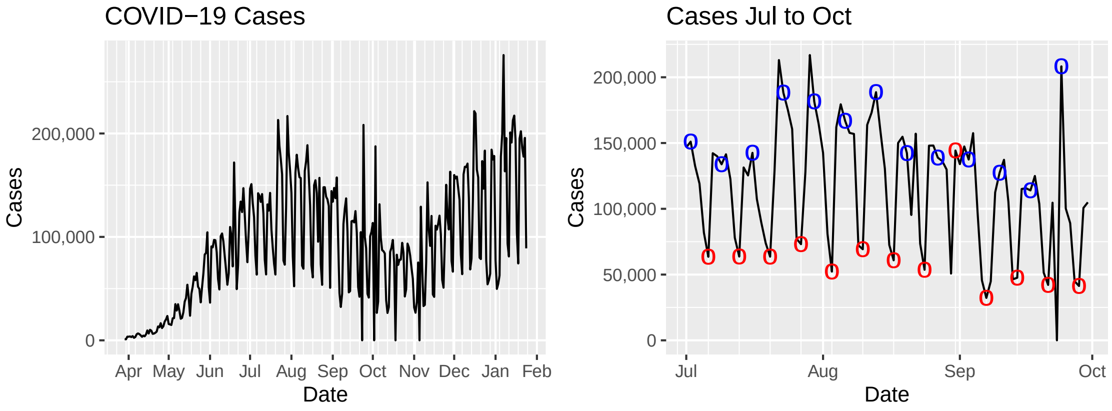

Figure 1: Daily COVID-19 Cases. In the right panel, red and blue circles denote Mondays and Thursdays respectively

When looking closer, there are also a few peculiar days in the dataset. Some days have 0 cases and the following day has a massive spike in cases. It can be assumed that this is due to a failure to report cases on those particular days and the number of cases is the result of both date's COVID-19 cases being combined. This requires imputation of those values for two reasons: To use a variance stabilizing transformation and to create a better model that more accurately reflects reality. The imputation on these dates was done as follows:

$$p := \frac{\left(\frac{cases_{t-7}}{cases_{t-7}+cases_{t-7+1}} + \frac{cases_{t+7}}{cases_{t+7}+cases_{t+7+1}}\right)}{2}$$

$$cases'_t = p \cdot cases_{t+1}, \;\; cases'_{t+1} = (1-p) \cdot cases_{t+1}$$

This imputation was made with a few assumptions. That 0 cases on day $t$ indicate there were no tests reported that day rather than having exactly 0 cases. Cases that would have appeared on day $t$ are counted on day $t+1$, so the total cases on day $t+1$ should be preserved and spread between the two days. The proportion of cases between $t$ and $t+1$ is similar to the proportions of nearby weeks.

A final observation to note is that there appears to be a drop in the number of cases reported on holidays and the day following a holiday relative to what we may normally expect on those days. Both the imputed values and holidays have been marked in Figure 2.

For the analysis, a log transform is used to stabilize the data. The variance seems much more stable on the log scale as shown in Figure 2.

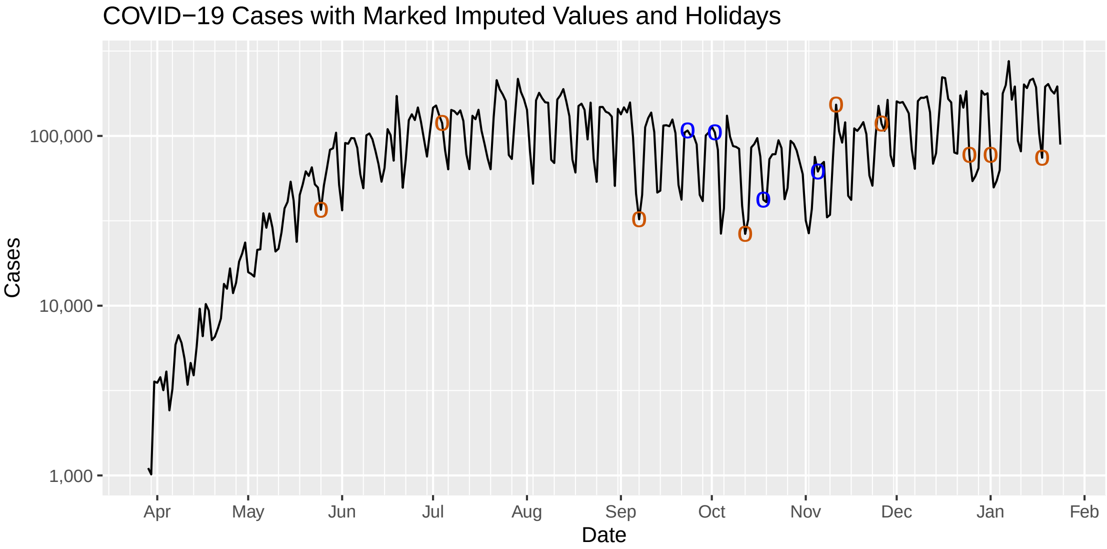

Figure 2: COVID-19 Case date on log scale with Imputed Values (Blue Circles) and Holidays (Orange Circles)

# 3 Models Considered

To model the natural signal in this data, both a parametric model and a differencing approach are used. Both of these models of the signal will be complimented with ARMA models for the remaining noise.

## 3.1 Parametric Signal Model

First, a parametric model is considered. For the base model, a degree 2 polynomial was used based on time. The waves that in the data are approximately represented by a 6 month period, so we created a sinusoid with that period and interacted that feature. To capture the weekly seasonality, indicators for each week day were used in this model. Finally indicator variables for whether the date was a holiday or a following day were added to the linear model. This is deterministic signal model is detailed in Equation (1) below, where $X_t$ is the additive noise term.

$$
\begin{aligned}
\log(cases_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \sum_{j=1}^{6} \beta_{2+j} t I_{\text{weekday}_{jt}} \\
+ \beta_9 t I_{\text{holiday}_t} + \beta_{10} t I_{\text{day after holiday}_t} \\
+ \beta_{11} t \cos\left(\frac{2\pi t}{6 * 30.5}\right) + \beta_{12} t \sin\left(\frac{2\pi t}{6 * 30.5}\right) + X_t
\end{aligned} \tag{1}
$$

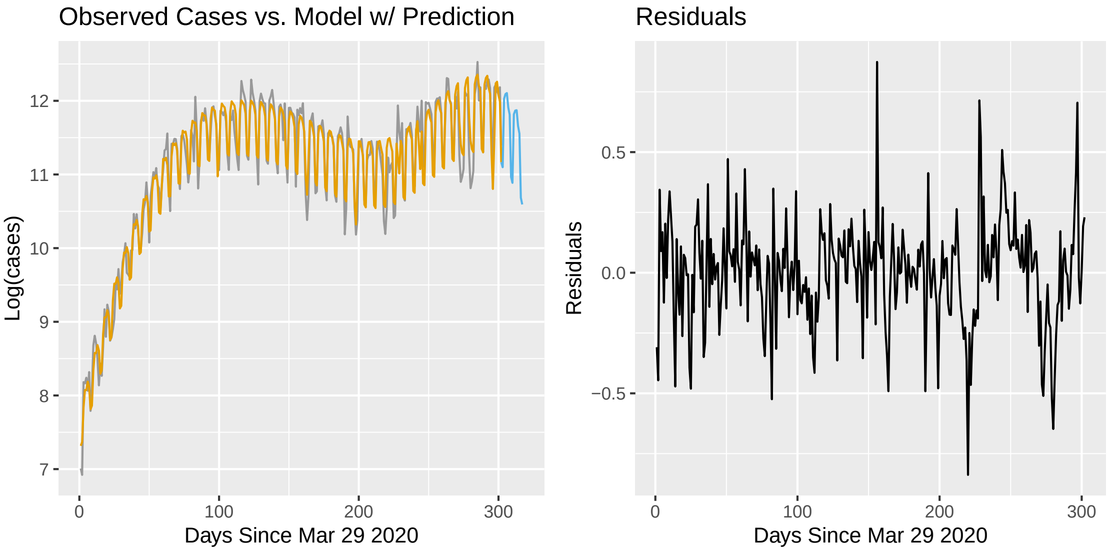

Figure 3: The parametric signal model. Left shows of fitted parametric model (orange), observed values (grey), and predicted values (blue) and the right panel shows the residuals of the model.

Figure 3 presents the fit as well as the residuals, which appear to be reasonably stationary. It also includes the predicted trend for 10 days to verify that the result is a reasonable trend.

### 3.1.1 Parametric Signal with AR(1)

The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for the parametric model residuals are shown in Figure 4. The PACF plot appears to have 1 significant value at lag 1. Furthermore, the ACF of appears to follow an exponential decrease with significant values only in the beginning. This immediately suggests an AR model. These two observations lead to proposing $p = 1$ as a potential fit. In Figure 4, marked with orange circles, we see the theoretical ACF of the fitted $AR(1)$ model. It decays very quickly, however it is not an unreasonable match.

### 3.1.2 Parametric Signal with AR(3)

In the PACF plot of 4, there is a significant value at lag 1 and the lags at 2 and 3 are possibly not significant. They are also close enough to significance that to warrant investigation. These observations suggest an AR model with $p = 3$ as a different potential fit. The theoretical ACF values of this model are marked by purple circles. It is a much better match with values closer to sample ACF and PACF. The tradeoff is that it is a more complicated model.

## 3.2 Differencing

As previously addressed, there is a locally linear trend of cases and weekly seasonality. While there are two waves of COVID-19 cases, it also appears to be locally linear. To address both of these things, a first

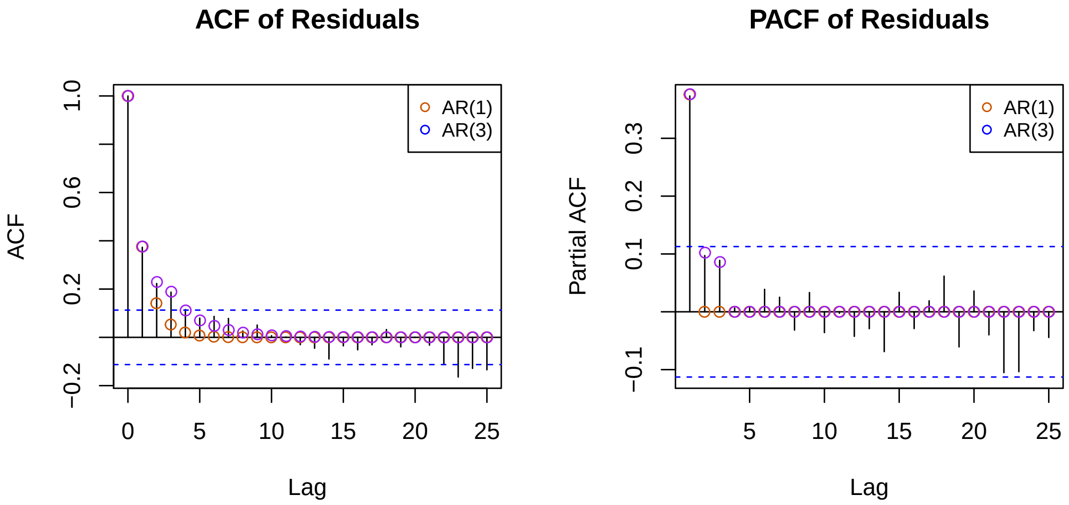

Figure 4: ACF/PACF of parametric model residuals with theoretical values of fitted AR(1) and AR(3) models. Sample ACF and PACF are shown with black lines.

difference and lag-7 difference are used. This is written as $\nabla_7 \nabla Cases$. The implied model is shown in the left panel of Figure 5. The right panel shows the time series of the differences, which appear stationary. Again, to address heteroscedasticity, a log-transform was preformed.

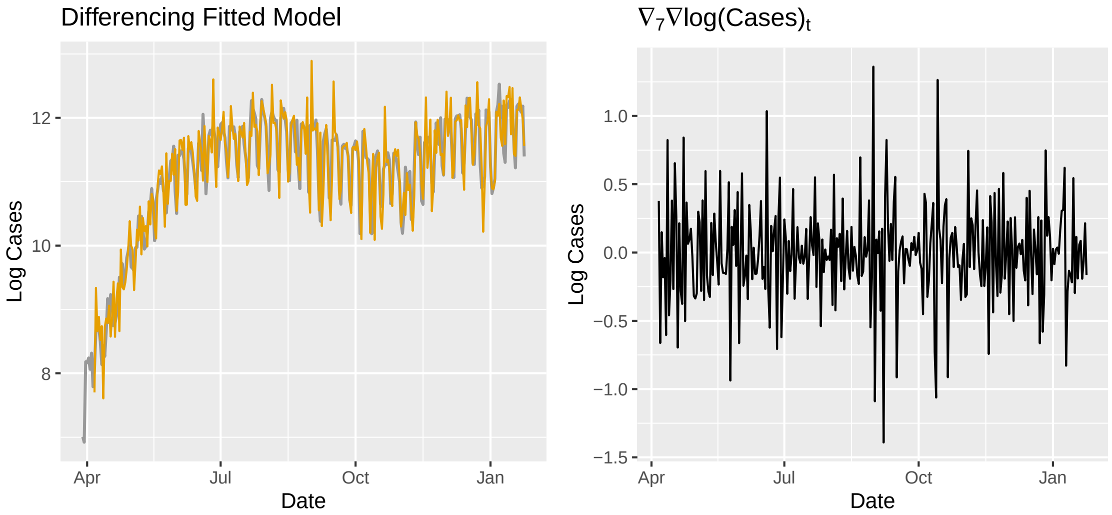

Figure 5: Diagnostics for differencing model. The left panel shows data in black and the fitted values in orange. The right plot shows the differenced time series.

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

The Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for the differenced model are shown in Figure 6. In the ACF plot, we can see significant values in multiples of 7, suggesting a seasonal ARMA model with S=7. There are no clear cutoffs values for these seasonal ACF and PACF values suggesting a multiplicative SARMA model is necessary and the cutoff rules for determining parameters cannot be used.

Significant values at every seventh lag of the PACF plot and 2 seasonal significant value in the ACF plot suggest that Q=2 and S=7. Following these peaks, there is a decaying PACF which suggests an positive value of q. There is also two significant values in the PACF, which suggests a p=2. With experimentation, a plausible model with p=2, q=1, Q=2 with S=7. The theoretical values of this fitted MSARMA(2,1)x(0,2)[7] the ACF and PACF are shown in figure 6 marked by orange circles.

### 3.2.2 Differencing with ARMA(2,2)x(3,1)[7]

In an alternative interpretation a significant value at 7 in the ACF plot and 3 possible seasonal significant values in the PACF plot suggest that P=3, Q=1, and S=7. Following these peaks, there is a decaying PACF which suggests an positive value of q. There is also two significant values in the PACF, which suggests a p=2. Together, with expirimentation, this gives us p=2, q=2, P=3, Q=1 with S=7. The theoretical ACF/PACF of the fitted model is demonstrated in figure 6. The purple circles fit the peaks significant values in the sample ACF/PACF.

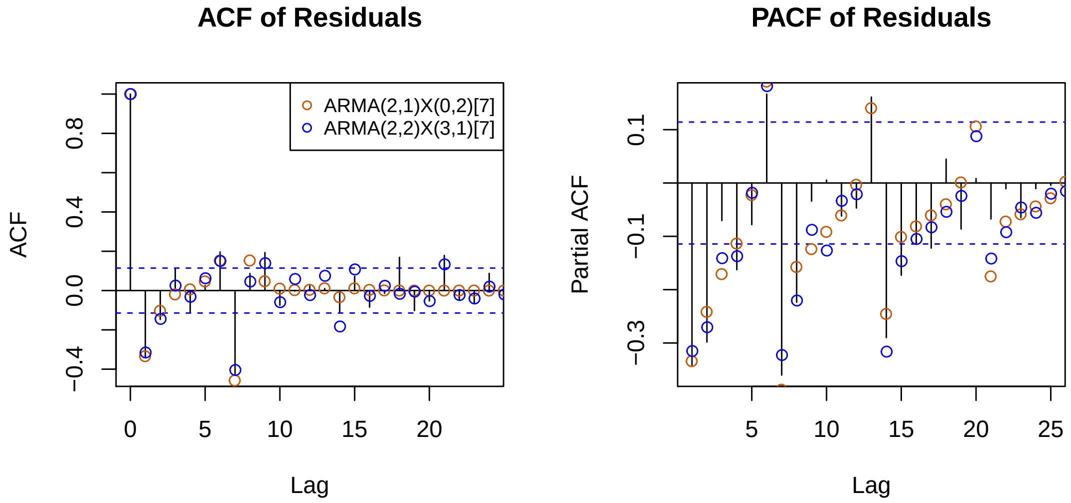

Figure 6: ACF/PACF of differencing model residuals with theoretical values of fitted ARMA models

## 4 Model Comparison and Selection

These four model options are compared through time series cross validation on the log-transformed data set. The nonoverlapping testing sets rolled through the last 100 days in the data, 10/17/2020 through 01/24/2021, in 10 day segments. Thus there will be 100 forecasted points over these 10 windows. The training sets consist of all data that occur before the appropriate testing set. The models' forecasting performances will be compared through root-mean-square prediction error (RMSPE). The model with the lowest RMSPE will be chosen as the model for predicting COVID-19 cases over the next 10 days.

Table 1 shows that the differenced model with ARMA(2,1)x(0,2)[7] has the lowest cross-validated prediction error with ARMA(2,2)x(3,1)[7] as a close second. Thus the differenced ARMA(2,1)x(0,2)[7] is chosen as the forecasting model.

## 5 Results

To forecast cases in the next 10 days (01/25/21 to 02/03/21), a model with differences at lag 7 and lag 1 will be used for the signal and augmented with an ARMA(2,1)x(0,2)[7] process for the noise. Let $Cases_t$

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

| | RMSPE |
|---|---|
| Parametric Model + AR(1) | 127836.06 |
| Parametric Model + AR(3) | 127836.06 |
| Daily Differencing + Weekly Differencing + ARMA(2,1)x(0,2)[7] | 38535.15 |
| Daily Differencing + Weekly Differencing + ARMA(2,2)x(3,1)[7] | 40549.25 |

be the number of cases at day $t$ with an noise term $X_t$. $X_t$ is a stationary process with 0 mean defined by ARMA(2,1)X(0,2)[7], $W_t$ is white noise with variance $\sigma_W^2$. This can be compactly written as ARIMA(2,1,1)x(0,1,2)[7]. The model can be represented as in Equation (2). $\phi_i, \Theta_i, \theta_i$ are all estimated in the next Appendix 1 Table 2. Note that $\Theta_1$ and $\theta_1$ have large magnitude with a negative with very tight bounds for the standard error. This suggests that $X_t$ is highly dependent on the white noise term $W_{t-7}$ and $W_{t-1}$.

$$\begin{aligned} log(Cases_t) =& \log(Cases_{t-1}) + \log(Cases_{t-7}) - \log(Cases_{t-8}) + X_t \\ X_t =& \phi X_{t-1} + \phi_2 X_{t-2} + W_t + \theta_1 W_{t-1} + \Theta_1 W_{t-7} + \theta_1 \Theta_1 W_{t-8} + \Theta_2 W_{t-14} + \theta_1 \Theta_2 W_{t-15} \end{aligned} \tag{2}$$

## 5.1 Prediction

Figure 7 shows the forecasted values of COVID-19 cases for the next ten days. It appears the recent wave has already peaked and flattened. Notably, the prediction for upcoming cases have high variance is cause for concern. This indicates that a rise in cases is not the expected outcome, but it is not probable.

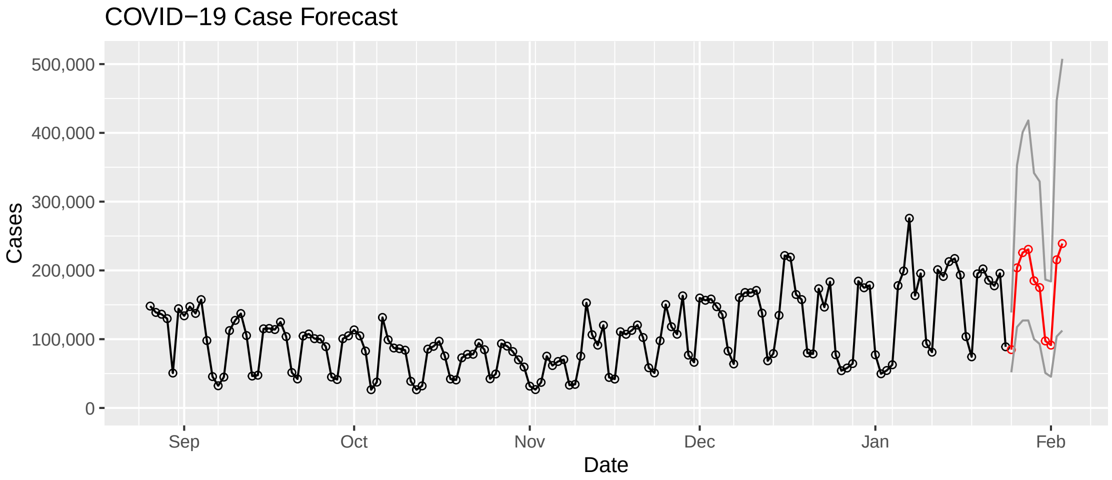

Figure 7: Forecasts of COVID-19 Cases in the fifth burrough of Gotham City from 01/25/21 to 02/03/21. The grey bands indicate the ±2 standard errors.

# 6 Appendix 1 - Table of Parameter Estimates for ARIMA(2,1,0)x(1,1,2)[7]

Table 2: Estimates of the forecasting model parameters in Equation (2) with their standard errors (SE). Note that this model includes a seasonal difference at lag 7 and a first difference.

| Parameter | Estimate | SE |
|---|---:|---:|
| $\phi_1$ | 0.2591 | 0.0874 |
| $\phi_2$ | -0.0167 | 0.0720 |
| $\theta_1$ | -0.7672 | 0.0695 |
| $\Theta_1$ | -0.7805 | 0.0672 |
| $\Theta_2$ | -0.0546 | 0.0695 |
| $\sigma_W^2$ | 0.060 | |