Forecasting the number of Covid Cases

Nathan Lam

1 Executive Summary

The current events of the global pandemic continues to affect Gotham City. During the months between March 2020 and January 2021, Covid-19 was extremely infectious. There are different waves and peaks of the number of cases and we want to be able to model the number of cases so we can have a reasonable estimate to account for future cases. The forecast is made using Weekly Differencing + Daily Differencing + ARMA(1,2)x(1,1)[7] and it predicts that the number of new cases will continue to show seasonality but the local trend will flatten in the next 10 days.

2 Exploratory Data Analysis

The number of covid cases shown on the left in Figure 1 looks to be composed of a big wave made out of smaller waves, there is at least two dominant frequencies in this process. The effects of seasonality also appears to increase over time which implies the variance of the process is changing. The process of the data in the beginning is very different to the process at the end, and you can see this between May and June where the size of the waves changes from from an exponential growth.

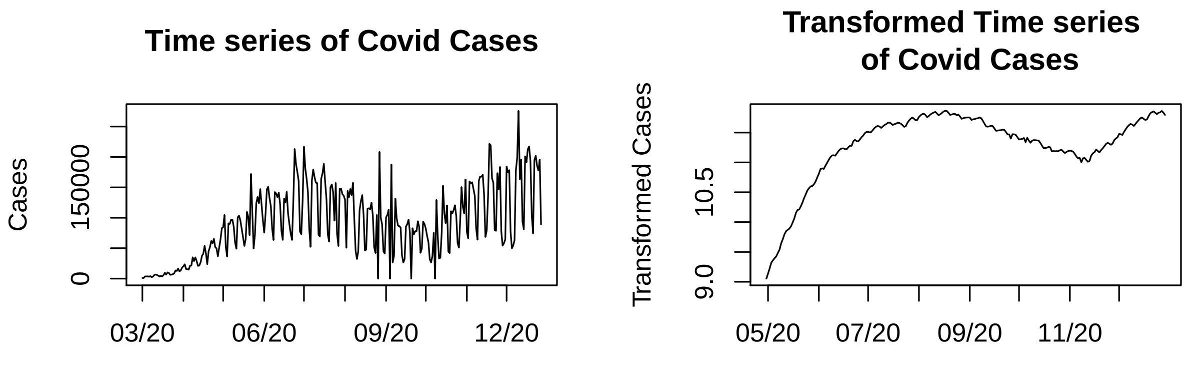


Figure 1: On the left, plot of raw time series for cases. On the right, plot of time series for cases after an exponential filter and log transformation. The time axis on both plots are marked by month and year.

In addition to the seasonality, there is also an trend line, possibly a polynomial trend, but it is apparent that cases is increasing over time. An ideal trend would have their slope approach 0, signaling the rate of spread is decreasing.

To help remedy the erratic behavior, we will transform and smooth the data. Using an exponential smoother and log transformation, we produce the right plot in Figure 1. The general shape is still preserved, which is what we want for out predictions. The trend made by the data is not as obstructed by the high variance before filtering, and after filtering, it appears maybe a cubic model could possibly fit. The process of filtering also removes some data points and the filter used truncates the data by 30 data points, which is why the start of the data will be different from plots without filtering. When the number of new covid cases is mentioned as a variable, it will be referred to as ‘Cases’ and will be in reference to after a filter has been applied.

3 Models Considered

The two models being considered is a parametric model and a differencing model. After fitting both models, a SARIMA model will be fitted to each to achieve white noise.

3.1 Parametric Model

The parametric model utilizes polynomials, an indicator, and sinusoids variables. The polynomial is meant to capture the almost cubic shape of the time series, an indicator is added to capture some seasonality in the weekly spikes, and five sinusoids are added to capture the remaining frequencies not captured by the indicator variable. The parametric signal model is mathematically described in Equation (1).

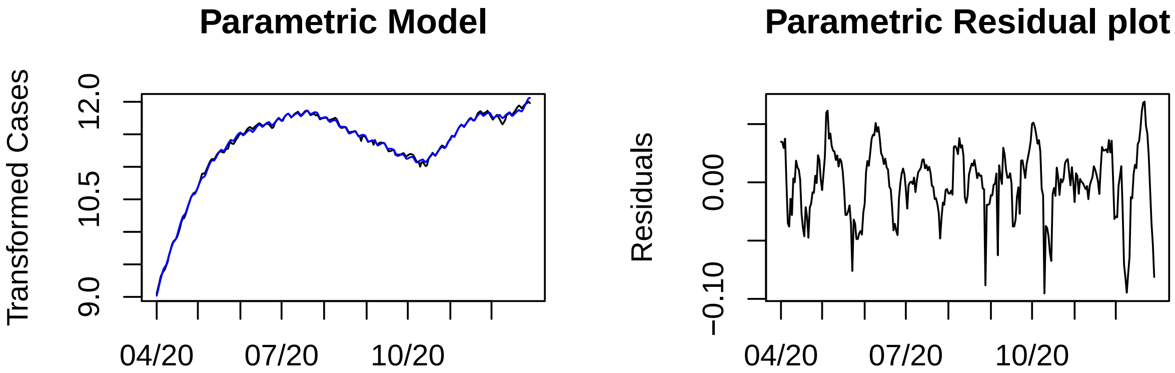


Figure 2: On the left, parametric model fitted in blue on the transformed data in black. On the right, plot of residual of the parametric model. The time axis on both plots are marked by month and year.

$$\begin{aligned}log(\text{cases}_t) = &\sum_{k=0}^{3}\beta_k t^k + \beta_4 I_{\text{day of week}_t} + \beta_5 cos(2\pi\frac{t}{18}) + \beta_6 sin(2\pi\frac{t}{18}) + \beta_7 cos(2\pi\frac{t}{58}) + \beta_8 sin(2\pi\frac{t}{58})\\ &+ \beta_9 cos(2\pi\frac{t}{72}) + \beta_{10} sin(2\pi\frac{t}{72}) + \beta_{11} cos(2\pi\frac{t}{96}) + \beta_{12} sin(2\pi\frac{t}{96})\\ &+ \beta_{13} cos(2\pi\frac{t}{144}) + \beta_{14} sin(2\pi\frac{t}{144})\end{aligned} \tag{1}$$

The fit of the parametric model, left plot in Figure 2, has some minute discrepancies, but it overall resembles the transformed data very closely. Observing the residual graph, right plot in Figure 2, the residual time series looks about stationary.

3.1.1 Parametric Signal Model with ARMA(2,2)

To diagnose an ARMA model, the Autocorrelation Function (ACF) plot and the Partial Autocorrelation Function (PACF) plot, shown in Figure 3. The PACF has a significant point at lag = 1 and the ACF resembles a sinusoidal decay, not necessarily an exponential decay. The significant PACF lag implies at least an AR(1) will fit, but the sinusoidal decay in the ACF plot also implies an AR model with at least two coefficients with opposite signs. From these observations, it was trial and error to find a p and q that roughly reproduces the same ACF and PACF plots.

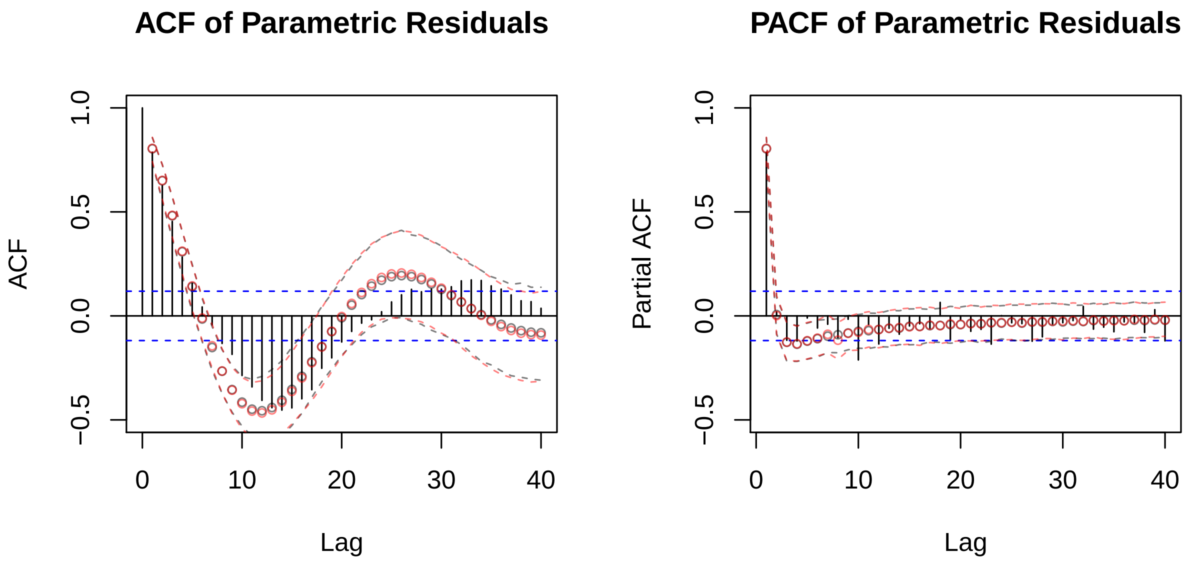


Figure 3: On the left, ACF of parametric residuals. On the right, PACF plot of the parametric residuals. The black circles and dotted lines represent the confidence interval following an ARMA(2,2) model. The red circles and dotted lines represent the confidence interval following a ARMA(2,2)(1,0)[7]

One chosen model is an ARMA(2,2) and to verify the fit of the model, ACF and PACF values are bootstrapped by generating data 1000 times following the parameters of the chosen model to create a 95% confidence interval using the 25th and 975th percentiles as the bands. The confidence bands are the curving dotted lines and the mean fits are the circles, like what is shown in Figure 3 and these will be referred to as 'CI'. Bootstrapping these plots will be used as a comparison between an approximated theoretical model and the data as a measure of fit.

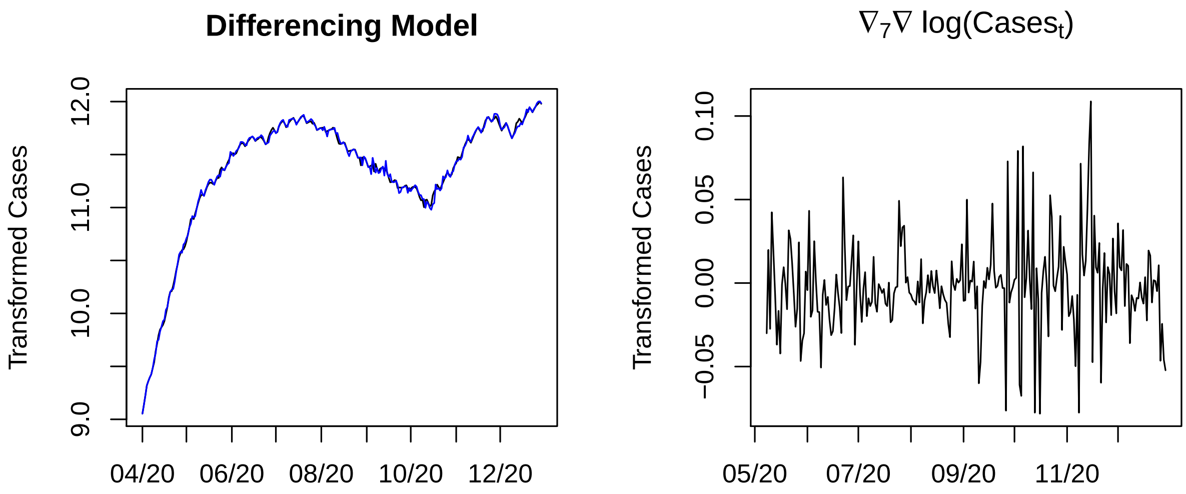
The confidence interval for ARMA(2,2) is shown in black in Figure 3, and in observing the fit, most of the data stay inside the bands with the few bars outside. Since a majority of the bars are stay inside the bands, the fit seems reasonable.

3.1.2 Parametric Signal Model with ARMA(2,2)(1,0)[7]

Another model to consider is adding a minor seasonal component to ARMA(2,2), this leads to considering an ARMA(2,2)(1,0)[7]. The fit of the model can be seen as the red bootstrap CI in Figure 3. The coefficient for the seasonal ARMA is considerably smaller than the non seasonal coefficients, so it doesnt have that big of an impact on the model. Observing the ARMA(2,2) CI and ARMA(2,2)(1,0)[7] CI agrees with this as both models are very similar. ARMA(2,2)(1,0)[7] looks to be a reasonable fit.

3.2 Differencing Model

The PACF plot from the parametric model implies Cases has some kind of autoregression, so a lag 1 difference might be useful. After applying this difference, Cases still shows weekly seasonality, which implies a lag 7 differencing is needed to remove that. Explicitly, the differencing model is a difference of lag 7 on a difference of lag 1 or weekly differencing + daily differencing. The fit of the differencing model can be observed on the left plot of Figure 4. The differencing model captures a lot of the data, variations between the model and the data is more apparent here than in the parametric model, but overall it seems to fit the data rather closely.


Differencing Model

$\nabla_7\nabla$ log(Cases$_t$)

Figure 4: On the left, differencing model fitted in blue on the transformed data in black. On the right, plot of transformed data after differencing

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

The SARIMA models being considered are diagnosed using the ACF and PACF plots in Figure 5. Ignoring lag 0 in the ACF plot, there is a significant bar with a clear cut off at lag 7, this implies there is at least a seasonal MA(1) component. The PACF plot shows seasonal decay, which agrees with the ACF, but this model does not capture this pattern. Through trial and error, the ARMA(1,1)(0,1)[7] model was obtained and fits the model rather closely. The fit of this model can be seen in Figure 5 as the black CI in the ACF and PACF plots. A majority of the data is within the CI bands, so this looks to be a reasonable fit.

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

An alternative is to consider a slightly more complex model. The ARMA(1,2)(1,1)[7] model was obtained similarly through trial and error based on using seasonal MA(1) as a starting point. While the model is different from the previous model considered, the SAR and one of the MA parts of the model have coefficients close to zero, so the effects of the model are similar. The fit of this model can be seen in the red CI on the ACF and PACF plots in Figure 5. Like the CI for ARMA(1,1)(0,1)[7], the CI here also captures a majority of the data, so this appears to be a good fit.

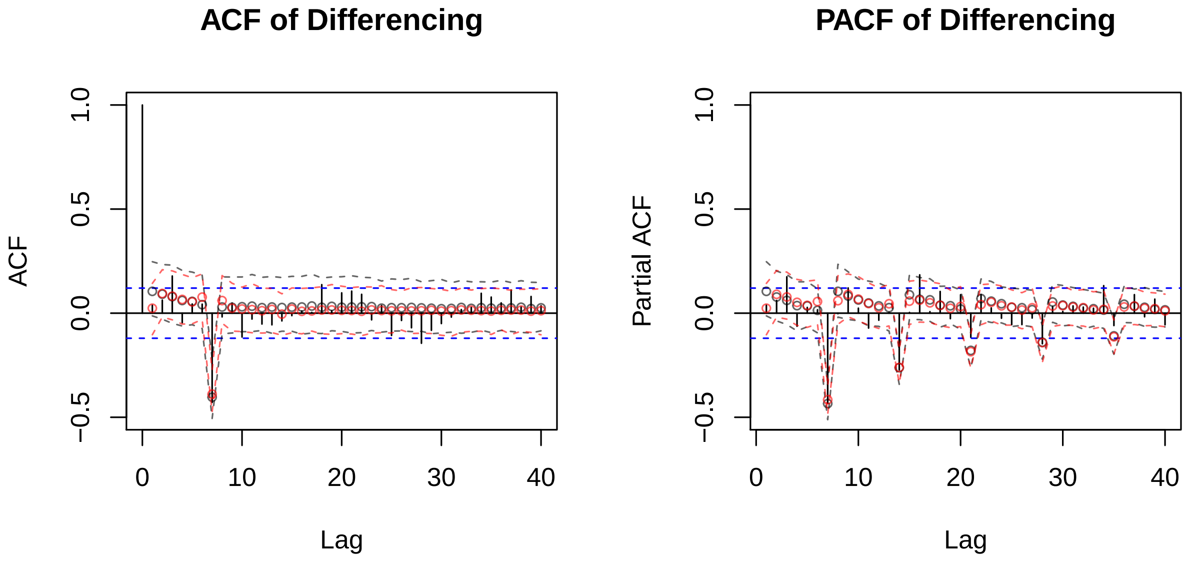


Figure 5: On the left, ACF of differencing. On the right, PACF plot of the differencing. The black circles and dotted lines represent the confidence interval following a ARMA(1,1)(0,1)[7] model. The red and dotted lines represent the confidence interval following a ARMA(1,2)(1,1)[7] model.

Table 1: Cross Validated RMSPE per considered model.

	RMSPE
Parametric Model + ARMA(2,2)	1.6589870
Parametric Model + ARMA(2,2)x(1,0)[7]	1.6589227
Weekly Differencing + Daily Differencing + ARMA(1,1)x(0,1)[7]	0.1157175
Weekly Differencing + Daily Differencing + ARMA(1,2)x(1,1)[7]	0.1153835

4 Model Comparison

To pick a final model for forecasting, each model will be measured using the Root Mean Square Predictive Error (RMSPE) through cross validation. From June 29, 2020 to January 24, 2021, Cases cases will be partitioned into 19 intervals of 10 days. One iteration of cross validation will fit the models and then calculate the Mean Square Root Predictive Error (MSPE) divided by 19 × 10 per model, and once every MSPE has been calculated and summed up, a square root is applied to get RMSPE. The best model will be the model with the smallest RMSPE.

The aggregated measured RMSPE per model can be seen in Table 1. Weekly differencing + daily differencing + ARMA(1,2)x(1,1)[7] performed overall the best with weekly Differencing + Daily Differencing + ARMA(1,1)x(0,1)[7] as a very close second.

5 Results

The forecasting model selected is weekly differencing + daily differencing + ARMA(1,2)x(1,1)[7]. Let $log(\text{cases}_t) = Y_t$ be the log transformed number of cases at day t after applying an exponential filter, X_t be a stationary process following ARMA(1,2)x(1,1)[7], and W_t be white noise with variance σ_W^2. The model can be mathematically described like in Equation (2).

$$\begin{aligned}Y_t &= \nabla_7 \nabla Y_t + X_t \\ &= Y_{t-1} + Y_{t-7} - Y_{t-8} + \phi_1 X_{t-1} + \Phi_1 X_{t-7} - \phi_1 \Phi_1 X_{t-8} \\ &+ W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \Theta_1 W_{t-7} + \theta_1 \Theta_1 W_{t-8} + \theta_2 \Theta_1 W_{t-9}\end{aligned} \tag{2}$$

The ARMA components in Equation (2) can be derived based on the form $(I - \phi_1 B)(I - \Phi_1 B^7)X_t = (I + \theta_1 B + \theta_2 B)(I + \Theta_1 B^7)W_t$, this describes ARMA(1,2)x(1,1)[7] using back-shift notation, where I is the identity operator and B is the back-shift operator. Expanding this form, solving for X_t, and substituting X_t is what is given in Equation (2). The varables ϕ, Φ, θ, and Θ are coefficients for the ARMA model.

5.1 Estimation of model parameters

The estimates of the coefficients can be seen in Table 2. The strongest three coefficients imply a noticeable ARMA(1,1)x(0,1)[7] process, which is the first differencing ARMA model considered, and this makes sense as that model was barely behind ARMA(1,2)x(1,1)[7] when cross validating RMSPE.

5.2 Prediction

Figure 6 plots the Cases_t from August 24, 2020 to January 23, 2021 in black and is appended with the forecasts for the next 10 days in red. The forecast predicts that the number of new cases in the next 10 days will have no trend while still showing some seasonality. This implies maintaining the same amount of resources from the last few days to support the new cases before the rate of new cases changes again.

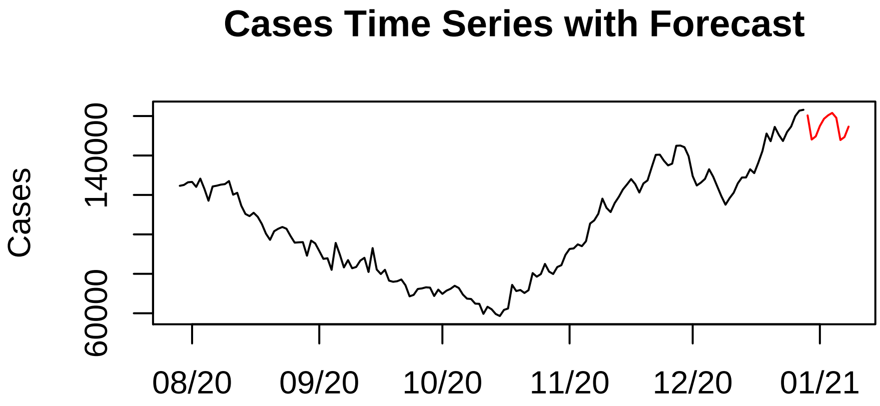


Figure 6: Time Series of number of covid cases and the predictions for the next 10 days. In black, the number of covid cases. In red, the forecast for the next 10 days

6 Appendix 1 - Table of Parameter Estimates

Table 2: Estimated coefficients of model as seen in Equation (2), with their standard errors (SE)

Parameter	Estimate	SE	Coefficient Description
ϕ_1	0.9908	0.0158	Non-seasonal AR coefficient 1
Φ_1	0.0433	0.0651	Seasonal AR coefficient 1
θ_1	-0.9211	0.0597	Non-seasonal MA coefficient 1
θ_2	0.0882	0.0614	Non-seasonal MA coefficient 2
Θ_1	-1.0000	0.0472	Seasonal MA coefficient 1
σ_W^2	0.0004058		Variance of White Noise