Stat153 project

yuelinzhou

12/13/2020

1 Executive Summary

Gotham City is having Covid cases for the last few months, In 6/20/20, Gotham City started tracking the number of new cases until 8/18/20 (totally 60 days). We chose Differencing model $\nabla_7\nabla_1$ with ARMA(0,1)x(1,1)[7] in order to do the forecast. The model's forecast showed the outlook does look promising. Covid cases tend to decrease over the next 10 days after 8/18/20.

2 Exploratory Data Analysis

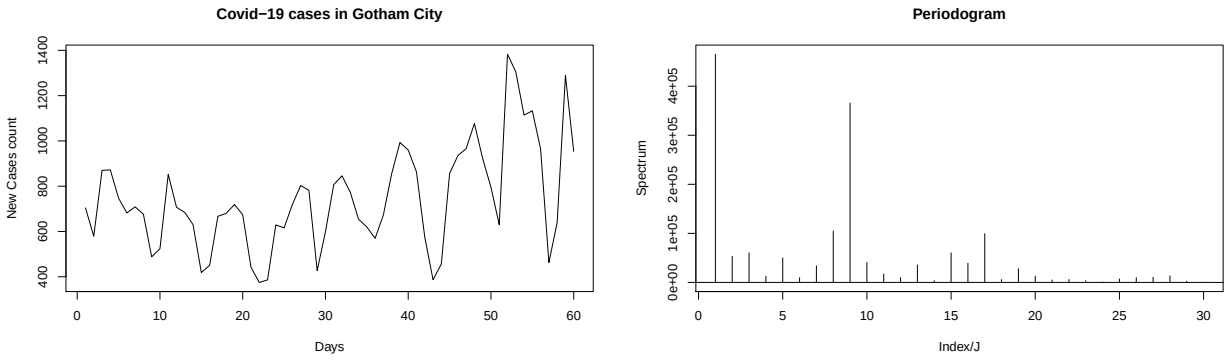

Figure 1: Covid new cases in Gotham City.

The data seems to follow an increasing linear trend. There's a strong seasonal pattern based on the periodogram in Figure 1. 2 largest spikes in the frequency of 1/60 and 9/60. the period of 9/60 is roughly 7 days. Heteroscedasticity occur in the 36th days which mean the varience new cases over time, and may imply that Covid-19 cases become serious than before.

3 Models Considered

To model the signal in this data, both a parametric model and a differencing approach are used. The remaining stationary "noise" will be addressed using ARMA models.

3.1 Parametric Signal Model

First, a parametric model is considered. A sinusoid which captures the increase in amplitude every week. A period are 3.53, 7 and 60 with time and indicator for the week of the Covid new cases. Below is the deterministic signal model written in Equation form, where X_t is the additive noise term.

$$Covid_t = \beta_0 + \beta_1 t + \beta_2 I_{\text{week1}} + \beta_3 I_{\text{week2}} + \beta_4 I_{\text{week3}} + \beta_5 I_{\text{week4}} + \beta_6 I_{\text{week5}} + \beta_7 I_{\text{week6}} + \beta_8 cos(\tfrac{2\pi t}{7}) + \beta_9 sin(\tfrac{2\pi t}{7}) + \beta_{10} cos(\tfrac{2\pi t}{60}) + \beta_{11} sin(\tfrac{2\pi t}{60}) + \beta_{12} cos(\tfrac{2\pi t}{3.53}) + \beta_{13} sin(\tfrac{2\pi t}{3.53}) + X_t$$

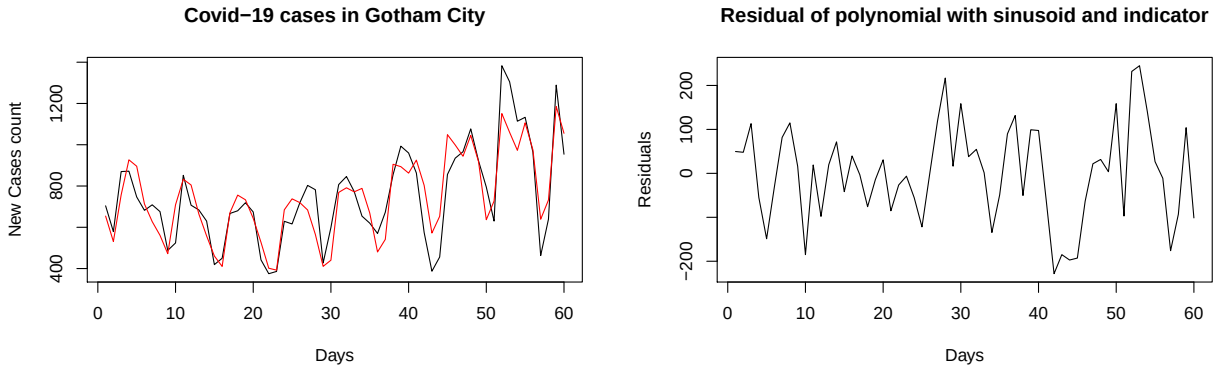

Figure 2: Those 2 plots are the fit and residuals, which looks reasonably stationary.

The left panel shows this model's fitted values in red, The right panel shows the residuals of this model. The right panel of Figure 2 shows that there could potentially be heteroscedasticity after the 36th days because that's where the Covid cases start to increase rapidly. when comparing the weekly cases, it could look very different after the 36th days. But my method of weekly indicator addressed those heteroscedasticity, now looks stationary.

3.1.1 Parametric signal with AR(1)

The ACF and PACF plots for these parametric models residuals are shown in Figure 3. There's 1 clear cutoff in PACF. ACF looks like bouncing between negative and postive which is a good sign for p = 1 as a potential fit. In addition, the P-value for ljung-box is pretty high for fitting AR(1). Therefore this model implies the theoretical ACF and PACF are indicated by the red circles in Figure 3.

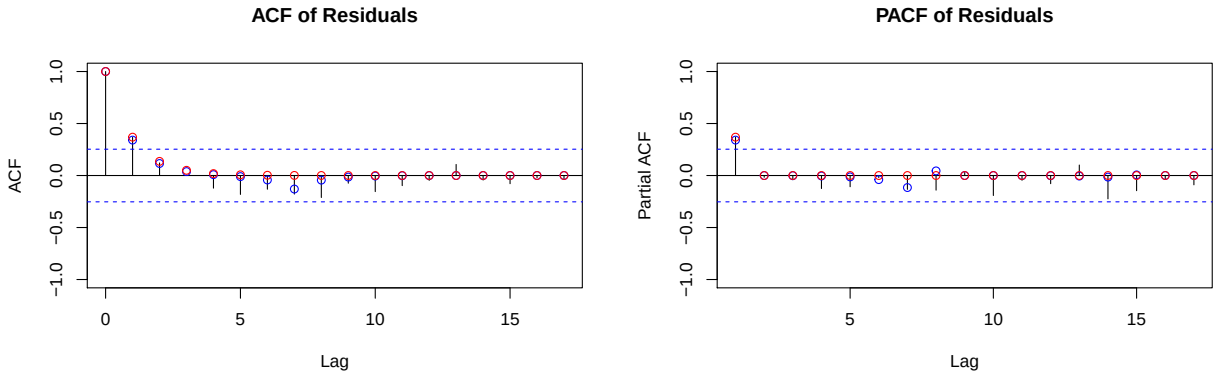

Figure 3: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for parametric signal model's residuals. Red circles is AR(1) model, blue circles is the ARMA(1,0)x(0,1)[7].

3.1.2 Parametric signal with ARMA(1,0)x(0,1)[7]

Based on the PACF, there's some large spikes in 7 and 14 which seems plausible as the seasonal MA (Q = 1) and S =7. The P-value for ljung-box is as high as AR(1). This model's theoretical ACF and PACF are

included as blue points on Figure 3, which look like an better fit to the sample ACF/PACF.

3.2 Differencing

Pursue stationarity with differencing. There's weekly effects pattern, so lag-7 differencing will also be beneficial. As I mentioned before, Covid cases looks like they are increase over time, so each week follows a increasing trend, differencing of lag-1 for each week will also be beneficial. and as this is now twice differenced, any linear or quadratic trend will be destroyed. Those fitted values of this differencing plot and the time series of the differencing plot are shown in Figure 4.

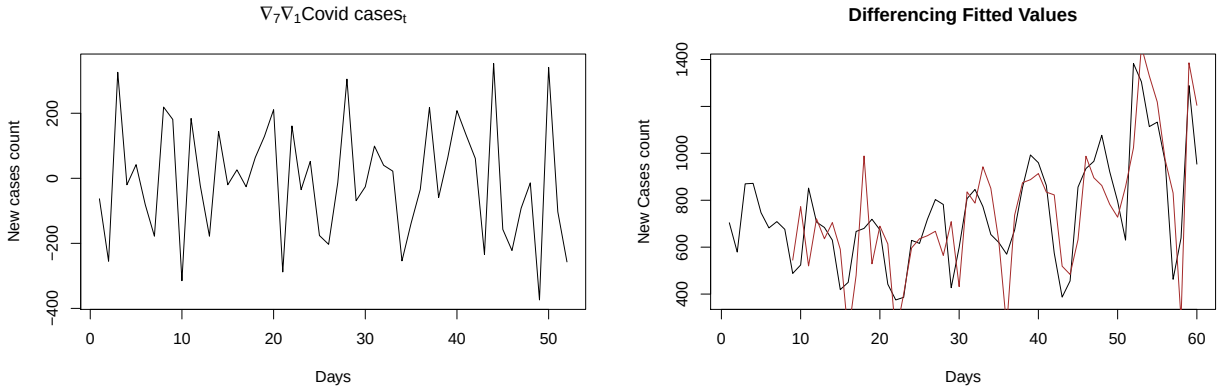

Figure 4: The left plot shows the differences themselves which looks pretty stationarity. The right plot shows data in black and the dfferencing fitted values in brown, they matches pretty well.

The dfferencing fitted values formula follows this equation:

$Covid_t = E(\nabla_7\nabla_1 Covid_t) + Covid_{t-1} + Covid_{t-7} - Covid_{t-8}$

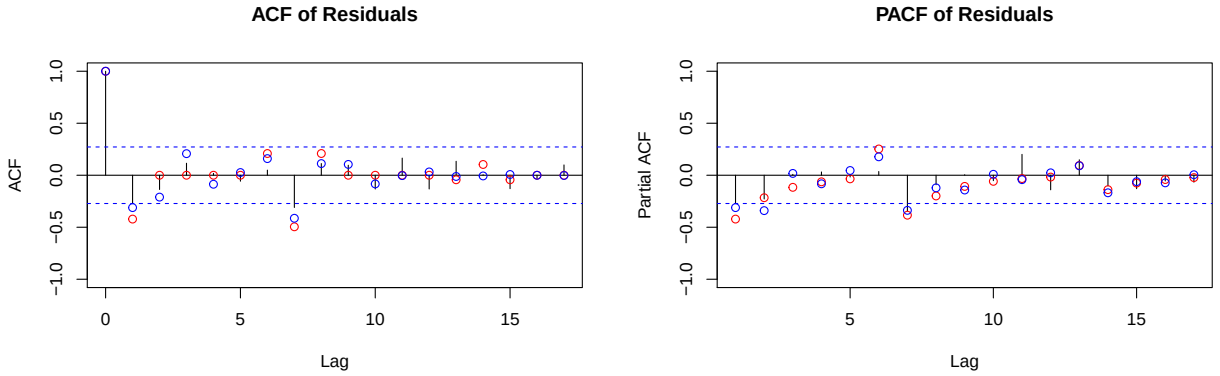

Figure 5: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the differencing model. Red circles is the ARMA(0,1)x(1,1)[7], the blue circles is ARMA(2,0)x(0,1)[7].

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

The sample ACF and PACF of residuals are shown in Figure 5. There's cutoff at lag 7 for both ACF and PACF, the plot suggest that P=1,Q=1 and S=7.The P-value for ljung-box increase rapidly for larger lags.

There also seems 1 cutoff in ACF, by adding q=1, the fit is much better. The fit of this choice is shown in Figure 5 with red circles.

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

Based on PACF, it looks like there's still 2 cutoff at lag 1 and 7, and setting up p=2 and S=7 will be beneficial. As mentioned in the previous model, the PACF suggest seasonal at lag 7, so set Q=1 will also be a good choice. The P-value for ljung-box is even higher than my previous differecing model. Therefore, the ACF and PACF of residuals in Figure 5 shows this ARMA(2,0)x(0,1)[7] model seems to fit better than ARMA(0,1)x(1,1)[7]. All blues circle in Figure 5 have smaller error than the red circles. Overall the shape looks pretty much the same.

4 Model Comparison and Selection

Table 1: Compare AIC,BIC and AICc those 4 ARMA models

	AR(1)	ARMA(1,0)x(0,1)[7]	ARMA(0,1)x(1,1)[7]	ARMA(2,0)x(0,1)[7]
AIC	12.17079	12.19235	13.04529	13.01168
BIC	12.27551	12.33197	13.23291	13.19930
AICc	12.17430	12.19949	13.06165	13.02805

Based on Table 1, AR(1) has the lowest AIC, BIC and AICc value than any other 3 models, which might be a best model for this Covid-19 dataset. But Further confirmation need to be addressed in Cross validation.

Cross validation

These four model options are compared through time series cross validation. cross validation is done on 10 nonoverlapping testing sets roll from the last 50 days (6/29/20 - 8/18/20) in 5 days segments. So there will be 50 forecasted points. The training sets consist of all data that occur before the testing set. Based on the performances of each models through root-mean-square prediction error (RMSPE). The model with the lowest RMSPE value will be chosen for prediction.

RMSE table

Table 2: Cross-validated out-of-sample RMSE for the four models

	RMSPE
Parametric Model + AR(1)	234.3643
Parametric Model + ARMA(1,0)x(0,1)[7]	202.1069
Weekly Differencing + daily Differencing + ARMA(0,1)x(1,1)[7]	189.5285
Weekly Differencing + daily Differencing + ARMA(2,0)x(0,1)[7]	211.8722

Based on the Table 2, the lowest RMSPE value is ARMA(0,1)x(1,1)[7] which is 189.5285, second lowest is ARMA(1,0)x(0,1)[7]. So ARMA(0,1)x(1,1)[7] is the chosen model for prediction.

5 Results

Let $Covid_t$ to be new Covid cases on day t, each Covid_t contain noise term X_t as the equation below. When taking Weekly Differencing + daily Differencing $\nabla_7\nabla_1$, approximately equal to X_t. Given result from Cross validation ARMA(0,1)x(1,1)[7] is the closest model to X_t. W_t is white noise with variance σ_W^2. The $E(\nabla_7\nabla_1 Covid_t)$ is the average different of Weekly Differencing + daily Differencing.

$Covid_t = E(\nabla_7\nabla_1 Covid_t) + Covid_{t-1} + Covid_{t-7} - Covid_{t-8}$

$\nabla_7\nabla_1 Covid_t = X_t$

$X_t = \Phi X_{t-7} + \theta W_{t-1} + \Theta W_{t-7} + \theta\Theta W_{t-8} + W_t$

$Covid_t = \beta_0 + \beta_1 t + \beta_2 I_{\text{week1}} + \beta_3 I_{\text{week2}} + \beta_4 I_{\text{week3}} + \beta_5 I_{\text{week4}} + \beta_6 I_{\text{week5}} + \beta_7 I_{\text{week6}} + \beta_8 cos(\frac{2\pi t}{7}) + \beta_9 sin(\frac{2\pi t}{7}) + \beta_{10} cos(\frac{2\pi t}{60}) + \beta_{11} sin(\frac{2\pi t}{60}) + \beta_{12} cos(\frac{2\pi t}{3.53}) + \beta_{13} sin(\frac{2\pi t}{3.53}) + X_t$

The second best model is ARMA(1,0)x(0,1)[7], which comes from Parametric Model, the longer equation above is the sinusoid with weekly indicator, $I_{\text{week i}}$ is the indicator if day t is in i th week, X_t is the noise term for ARMA(1,0)x(0,1)[7] which is not chosen for prediction.

5.1 Estimation of model parameters

All the estimates of the model parameters are given in Table 3 in Appendix 1. It's interesting to see that the $E(\nabla_7\nabla_1 Covid_t)$ is actually a negative number. This make sense because we expect the Covid cases to increase over time. And clearly the earlier time, the less cases that this city has.

5.2 Prediction

Figure 6 shows the forecasted values of Covid new cases for the next 10 days started from 8/18/20. The model predicts the next 10 days of Covid cases as we expected the Covid cases to increase. But supervisely the Covid cases tend to decrease over the next 10 days. Looking at the prediction intervals, even the light grey bands doesn't follows increasing trend, therefore we have a good news! Covid cases is decreasing after 8/18/20. In the future we hope Gotham City can beat the virus and keep everybody healthy.

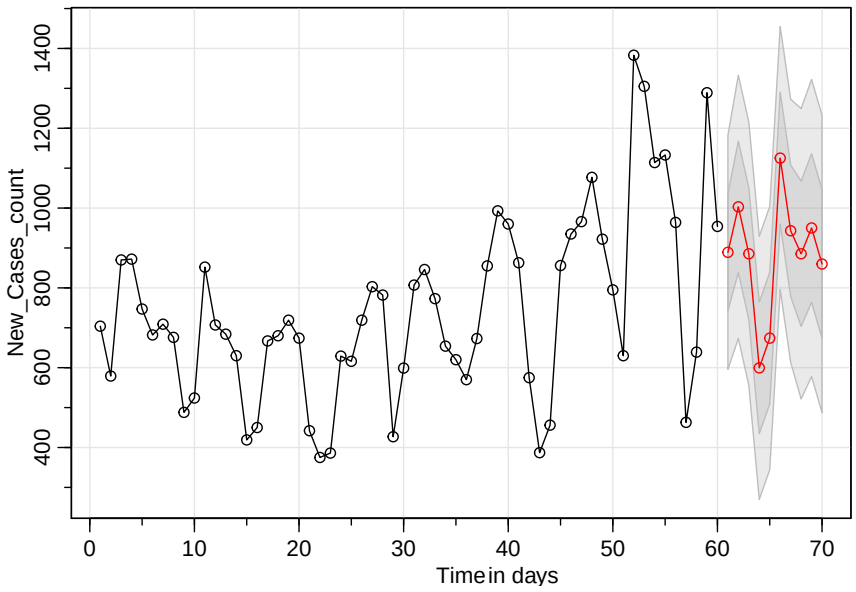

Figure 6: This is the forecast model for Covid new cases, x-axis is the time in days, y-axis is the number of count of Covid cases. The black points are the recent historical Covid data. The red points are the forecasts value after 60 days. The dark grey bands are the 68% prediction intervals. The light grey bands are the 95% prediction intervals.

6 Appendix 1 - Table of Parameter Estimates

Table 3: Estimates of the forecasting model parameters in ARMA(0,1)x(1,1)[7], with their standard errors (SE).

Parameter	Estimate	SE	Coefficient Description
$E(\nabla_7\nabla_1 Covid_t)$	-6.211538		Average differencing
θ	-0.5482681	0.1594	MA coefficient
Θ	-0.3970179	0.2581	Seasonal MR coefficient
Φ	-0.2087524	0.2803	Seasonal AR coefficient
σ_W^2	21106.84		Variance of White Noise