Analysis of Violent Crime in Los Angeles County

Xiaohong Huang
UID: 004693375

March 20, 2017

Abstract

Violent crime can have a negative impact to the victims and the neighborhoods. It can affect people's lives physically and emotionally. It could also impose huge cost to the taxpayers. Therefore, it seems meaningful if we could identify patterns in the crime data, so that it could be used to reduce the violent crimes. In this paper, the violent crime data in Los Angeles County from year 2014 to 2016 is analyzed, with the goal of finding an ARIMA model to forecast the number of violent crimes. Nonparametric and parametric spectral estimation are also performed to estimate the possible cycles for the violent crime.

1 Introduction

Exposure to violent crime can have a negative impact on people's lives and the neighborhoods. The effect seems to be more obvious on children. According to the paper from US Department of Housing and Urban Development [1], "In general, exposure to violence puts youth at significant risk for psychological, social, academic, and physical challenges and also makes them more likely to commit violence themselves. Exposure to gun violence can desensitize children, increasing the likelihood that they act violently in the future." The consequences sound horrifying and violent crime is obviously something that could influence the whole community.

In addition to threatening people's lives and safety, violent crime imposes huge cost to the community in various ways. A large number of violent crimes could lower the value of people's properties in the community. Having more crimes means it will cost more taxpayers' money to maintain the police system and the prisons. According to the paper from Professor Shapiro [2], "The costs borne by the American public for this level of criminal activity are significant. Medical care for assault victims, for example, costs an estimated $4.3 billion per year. We spend $74 billion per year on incarcerating 2.3 million criminals, including some 930,000 violent criminals." In short, analysis for the violent crime data seems to be important as it is closely related to everyone in the neighborhood. Finding patterns in the crime data could be useful for the law enforcement in terms of making strategies to reduce violent crime and protect the public.

The goal of this project is analyze the trend and periodicity of the violent crime in Los Angeles county. In addition, we are interested in fitting an ARIMA model and make a 10-day ahead prediction.

2 Dataset

From the definition of the FBI's Uniform Crime Reporting (UCR) Program, "violent crime is composed of four offenses: murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault." Therefore, the violent crime data in this project is the aggregated sum of these four offenses.

The dataset is downloaded from the Los Angeles County Sheriff's Department. The raw data contains all the crimes reported every day and only the crimes that falls into the category of violent crime is counted for this project. The consolidated dataset is daily data and each observation represents the total number of violent crimes happened in one day. The dataset used in this project is from 01/01/2014 to 12/31/2016 with 1096 observations. The last 10 days' observation(12/21/2016 - 12/31/2016) is used as the testing data and the remaining data is used as the training data.

3 Data Exploration

Figure 1 shows the original data and data looks very choppy. The average of the data is 31.90. According to Business Insider [3], "The average US violent-crime rate was roughly 36.6 offenses per 10,000 people in 2014." So the Los Angeles County is doing slightly better than the nationwide average.

To see if there is trend in the data, the violent crime data is regressed on time t. Figure 1 shows the regression line. Table 1 shows the intercept and slope of the linear regression are both significant. The estimated regression line is $x_t = -4731.0014 + 2.3632t + w_t$. Figure 2 shows the decomposition of the dataset into seasonal, trend and irregular components, and the graph for the trend component also shows an upward trend. Therefore, it seems there is an increasing trend in the dataset.

Looking at Figure 1 again, it looks like there is some cycle in the original data. If the data is ordered by the number of violent crime in decreasing order each year, it is interesting to see that the highest number of violent crime each year usually occurs roughly between August and October. The seasonal component in Figure 2 also shows the number will rise to a peak in the second half of the year and then decrease to around 0 at the end of the year. Figure 3 shows the ACF and PACF for the original data. Here each lag is the multiple of 1/365. The ACF plot at certain lags shows relatively slow decay and the ACF is still significant after lag 45. It looks like there is some weekly cycle in the original data. Therefore, it seems there are multiple seasonality in the dataset.

To make the data stationary, two methods are considered: detrending and taking first difference. Figure 4 shows the residuals after detrending the data and figure 5 shows the data after taking the first difference. The ACF and PACF for the detrended data seem to have some seasonality. Also, the plot for the detrended data still shows cycle similar to

the original data. For example, the number usually goes up at the second half of the year. On the other hand, the ACF for the first differenced data seems to be cut off at lag 1 and the PACF seems to be tail-off, so it looks like the differenced data follows the pattern for an ARIMA model. However, if the differenced data is used to fit the model, it might be difficult to interpret the estimated model. In general, the model without difference order is preferred when there is no good explanation for the differenced data. Therefore, we will consider fitting models using the first differenced data and the detrended data.

4 Modeling Fitting and Diagnostics

4.1 ARIMA Model

Table 2 shows the candidate models and the corresponding AIC. For ARIMA models that have zero difference order, the detrended data is used to fit the model. It shows ARIMA(1,0,2) has the lowest AIC, so ARIMA(1,0,2) is chosen to be the preferred model. The estimated model for the detrended data is $x_t = -0.041 + 0.976x_{t-1} - 0.850\hat{w}_{t-1} - 0.080\hat{w}_{t-2} + \hat{w}_t$. The $\hat{\sigma}_w = 6.4$ with 1,082 degrees of freedom. The regression coefficients are significant except for the constant term. The coefficient for the AR term is positive, which might suggest the number of violent crime tomorrow is positive related to the current number of violent crime. On the other hand, the coefficients for the MA term is negative, which might suggest the next observation is negative related to the shock of the previous two values in the series. If the previous shock is positive and large, then the next observation is more likely to be smaller. Such negative relationship might account for the choppiness shown in the original dataset. It seems interesting to see that the AR term and the MA term have opposite relationship to the current observation. If the coefficients for the AR term and the MA term are both positive, then the data might be smoother. If the goal is to reduce the number of future crimes, then the model suggests reducing the current number of crimes help lower the number of violent crimes tomorrow.

To see whether the ARIMA(1,0,2) is a good fit for the data, figure 6 shows the residual diagnostic for the model. The residuals seem to have constant mean around 0 and are roughly normally distributed. However, some ACF at certain lags are significant and it looks like it follows some cycle. Also, the Ljung-Box statistics shows significance after lag 8. This might suggest the residuals are not independent and they are are not white noise. Therefore, ARIMA(1,0,2) does not seem to be a good fit for the data.

Since the ARIMA(1,0,2) model is not a very good fit for the data , it is not surprising to see that the predictions are not very accurate. Figure 7 shows the 10-day ahead prediction using the ARIMA(1,0,2) with the trend added back to the prediction. The predictions are almost like a flat line, which fail to capture the variations in the data. Even though ARIMA model is easier to interpret, if the goal is to make a prediction for the data, then ARIMA model might not be the best model to use. Therefore, it seems more practical to try the seasonal ARIMA model.

4.2 Seasonal ARIMA Model

Looking back to figure 3, the characteristics of the ACF of the original data shows a strong peaks at lag = 1s,2s.... in the autocorrelation function with s = 7. Similarly, the PACF plot shows minor peaks at lag = 1s,2s.... Therefore, it is very likely that there is weekly seasonality in here. Figure 8 shows the ACF and PACF for the original data after taking the weekly difference. It seems the ACF is cutting off after lag 1s and the PACF is tailing off in the seasonal lags with s = 7. This looks like a pattern for seasonal ARIMA. Figure 9 shows the original data after weekly differencing, and it seems the periodicity shown in figure 1 is removed. Even though the variance seems to be relative smaller in the first half of the year in 2014 and 2015, the differenced data looks roughly stationary.

Table 3 shows the seasonal ARIMA models and the corresponding AIC. Here seasonal ARIMA(1,0,1)x$(1,1,1)_7$ has the lowest AIC, so it is the preferred model. Table 4 shows the coefficients for the seasonal ARIMA(1,0,1)x$(1,1,1)_7$ model and $\hat{\sigma_w^2} = 36.88$. It is interesting to see that the coefficient for the nonseasonal AR is positive while the coefficient for the seasonal AR is negative. Since coefficient for the nonseasonal AR is larger than the coefficient for the seasonal AR in absolute value, this might suggest the weekly difference for tomorrow is positive related to the weekly difference today, while the weekly difference one week ago will have a smaller negative effect for the weekly difference tomorrow. Similar to the ARIMA model, the coefficient for the nonseasonal AR has the opposite sign to the nonseasonal MA term. The coefficient for the nonseasonal MA and seasonal MA are both negative. Since the seasonal ARIMA(1,0,1)x$(1,1,1)_7$ has a seasonal difference order, it seems difficult to interpret the MA and SMA terms with this model. This could be a drawback if we are interested in finding a model that is easy to interpret.

Figure 10 shows the residual diagnostic for the seasonal ARIMA(1,0,1)x$(1,1,1)_7$ model. The residuals seem to have constant mean around 0. There seems to be some outliers in the normal Q-Q plot. Since the number of outliers is small, the residuals still seems to be roughly normally distributed. There are still some autocorrelations that are close to be significant, but the cyclical patterns are not as obvious as shown in the ARIMA model. The Ljung-Box statistics also shows the residuals are independent. Therefore, it looks like the residuals are white noise and the seasonal ARIMA(1,0,1)x$(1,1,1)_7$ model might be a good fit for the data.

Figure 11 shows the 10-day ahead prediction. Most of the testing data falls inside the 95% prediction interval. Another thing to notice is that the 95% prediction interval for seasonal ARIMA(1,0,1)x $(1,1,1)_7$ is wilder than the prediction interval for ARIMA (1,0,2). However, the prediction still fails to capture the sudden drop on 12/24/16. Looking at the data right before the prediction, there seems to be a weekly cycle from 12/12/16 to 12/18/16, which decreases first and then increases. In the seasonal ARIMA model with weekly seasonality, the period from 12/22/16 to 12/25/16 is supposed to be increasing in the weekly cycle. This might explain why the prediction goes up to 37.71 on 12/24/16 while the actual data drops to 15 on that day. Looking at figure 1 again, the week near New Year usually corresponds to a below average number of violent crimes. However, the seasonal ARIMA model might not be able to capture this kind of pattern. One possible reason for the inaccuracy is that the data might not be strictly in a weekly cycle. This is also reflected in the ACF plot of figure 10. The significance for the autocorrelation might be from the changing cycle. Since the violent crime can be affected by many other factors, it is reasonable to have changing

cycles in the dataset. However, since the 95% prediction interval is able to contain most of the testing data, the seasonal ARIMA model seems to be a better model than ARIMA model if the goal is to forecast the number of violent crimes.

5 Spectral Analysis

The periodogram in figure 12 shows the time series contains three peaks at frequencies of about 0, 52 and 110. Here the frequency axis is labeled in multiples of $\Delta = 1/365$. It is also obvious that the periodogram is choppy. Since the original data in figure 1 is very choppy, it is not surprising to see that the raw periodogram is choppy as well. There are many small spikes in the periodogram, which might be caused by the noise. In order to identify the predominant period, the modified Daniell kernel is used to smooth the periodogram. As shown in figure 13, There are three major peaks and the periods are $1/0.649 = 1.54$ year, $1/52.24 = 0.019$ year or 7 days and $1/12 = 0.083$ year or 30 days. It is worth noting that the smoothed periodogram still has many small spikes at the high frequencies, which might be corresponding to the changing cycles as discuss above. It might also suggest the data is not completely sinusoidal and the minor peaks capture the non-sinusoidal behavior of the signal.

An approximate 95% confidence interval for the spectrum $f_S(1/0.649)$ is [0.38,0.93] and an approximate 95% confidence interval for the spectrum $f_S(1/52.24)$ is [0.27, 0.66]. Since the lower bound of the confidence interval is not higher than any other periodogram ordinate, it seems difficult to establish significance of the peak.

Another way of estimating the spectral density is to fit an AR model to the data and use the spectral density of the AR model as the approximation. Figure 14 shows the spectral density of an AR(28) model. Here the AR(28) model is chosen by the AIC criterion. There are two major peaks and the periods are $1/0.73 = 1.37$ years and $1/52.3 = 0.019$ year or 6.98 days, which looks similar to the smoothed periodogram. It is interesting to see that the peak corresponding to the 30-day cycle in smoothed periodogram is almost smoothed out in the parametric spectral estimation. Therefore, only the periods of 1.4 $\sim$1.5 years and 7 days are included in the final conclusion.

6 Conclusions and Future Improvement

In this project, the seasonal ARIMA(1,0,1)x $(1,1,1)_7$ is chosen to be the best model. The residual analysis shows the seasonal ARIMA model seems to fit well. The seasonal ARIMA model seems to produce a slightly better prediction than the ARIMA(1,0,2) model, with the cost that the seasonal ARIMA model is harder to interpret. The spectral analysis suggests there seems to be a predominant period of around 1.4$\sim$1.5 years and 7 days in the dataset. The weekly cycle found in the spectral analysis seems to match the seasonal period for seasonal ARIMA(1,0,1)x $(1,1,1)_7$ model that is chosen.

For the future work, some improvements can be done for the analysis. One disadvantage of the current model is that it cannot explain what drives the number of violent crime to a relatively higher number during summer and fall and why the data will drop to a relatively

lower number at the end of the year. It will be interesting if we can incorporate other information such as the demographics of the area and the employment rate. Maybe the extra information will help explain such pattern. In addition, for this project only three years' data is used, and an upward trend is observed. However, it might not be the case when more data is included. The long term effect for the violent crime might be different. For example, the previous 20 years' data could be used to run the analysis again and it will be interesting to see whether the trend is increasing or decreasing.

7 Figures

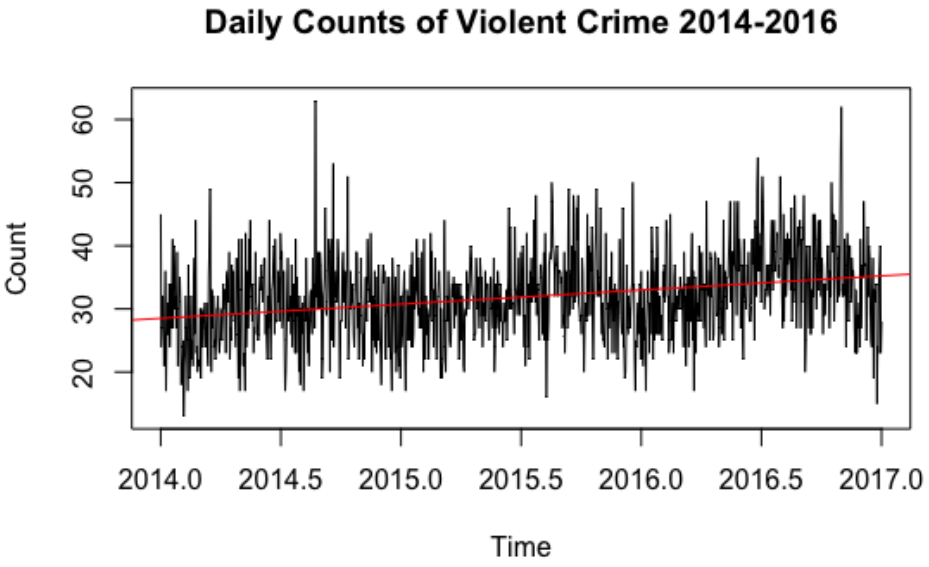

Figure 1: Original Data with Regression Line

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-4731.0014	467.8698	-10.11	$< 2e-16$***
Time	2.3632	0.2321	10.18	$< 2e-16$ ***

Table 1: Summary of Linear Regression

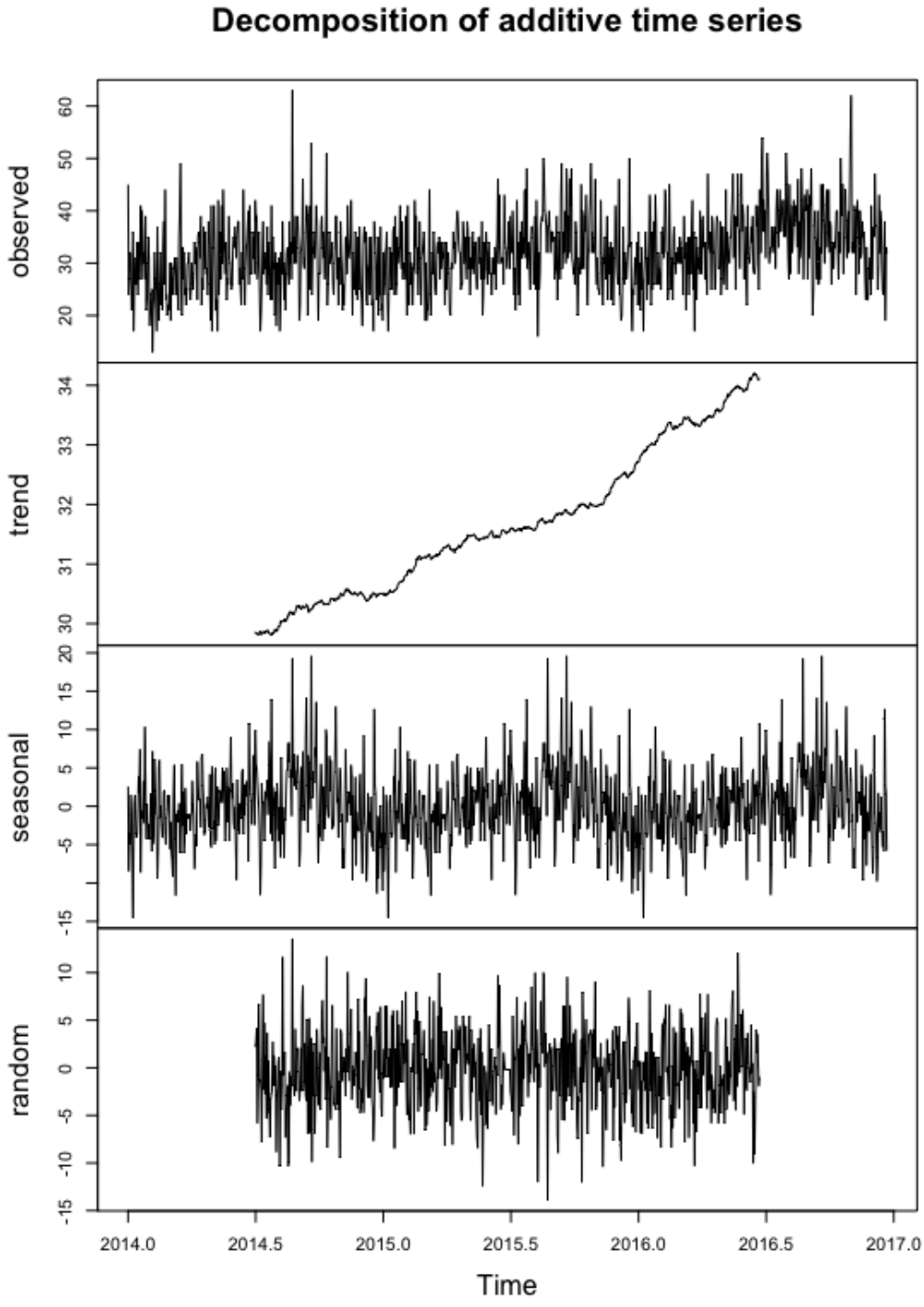

Figure 2: Decomposition of Time Series

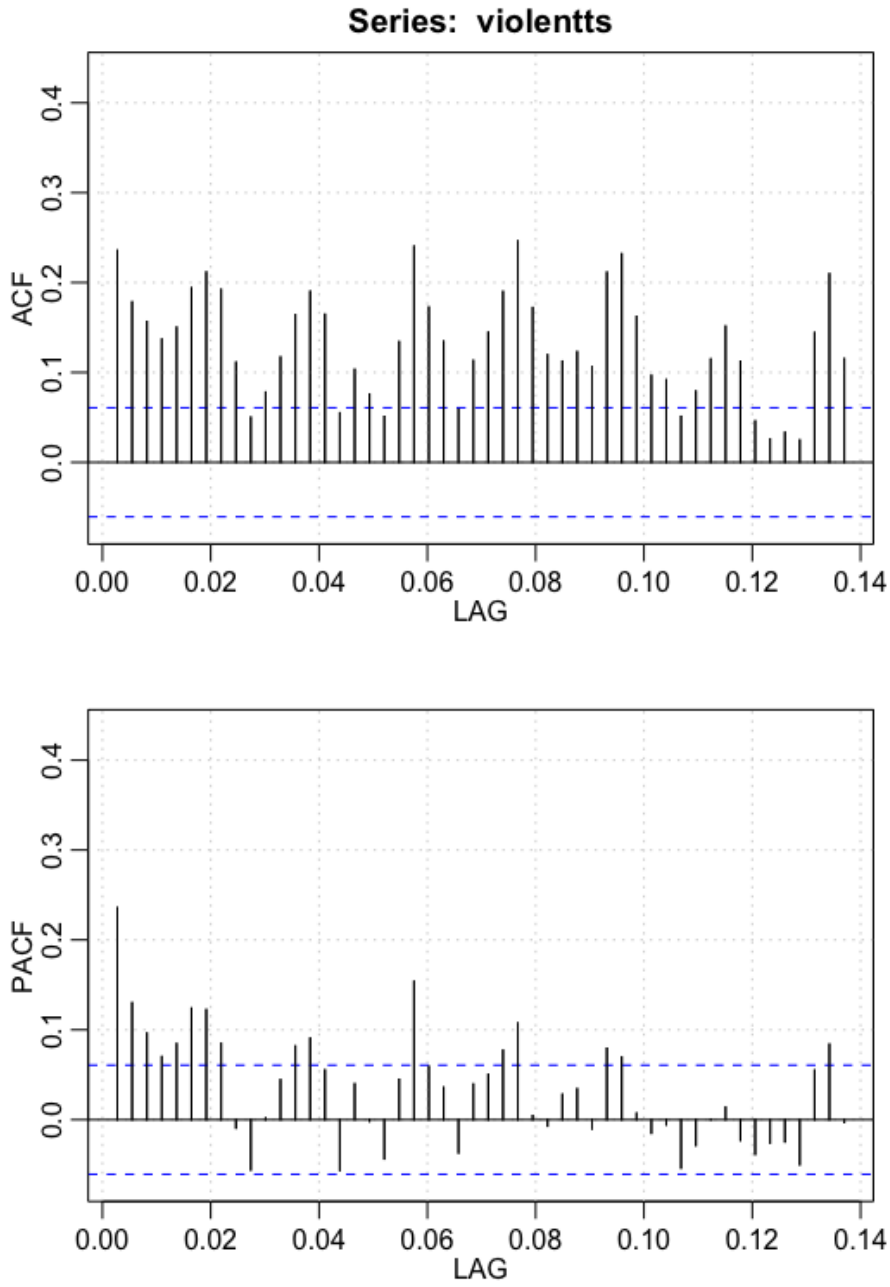

Figure 3: ACF and PACF for Original Data

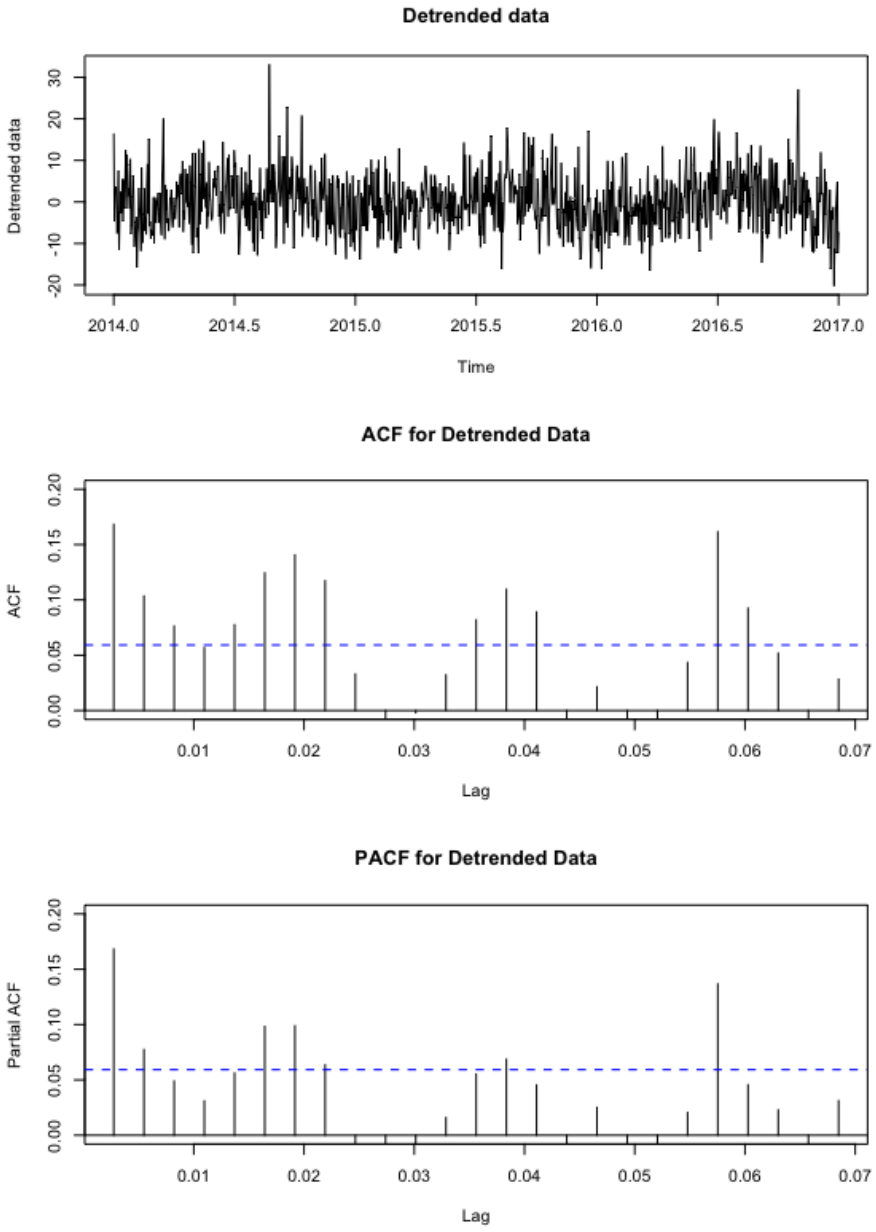

Figure 4: Detrended Data and the ACF and PACF Plot

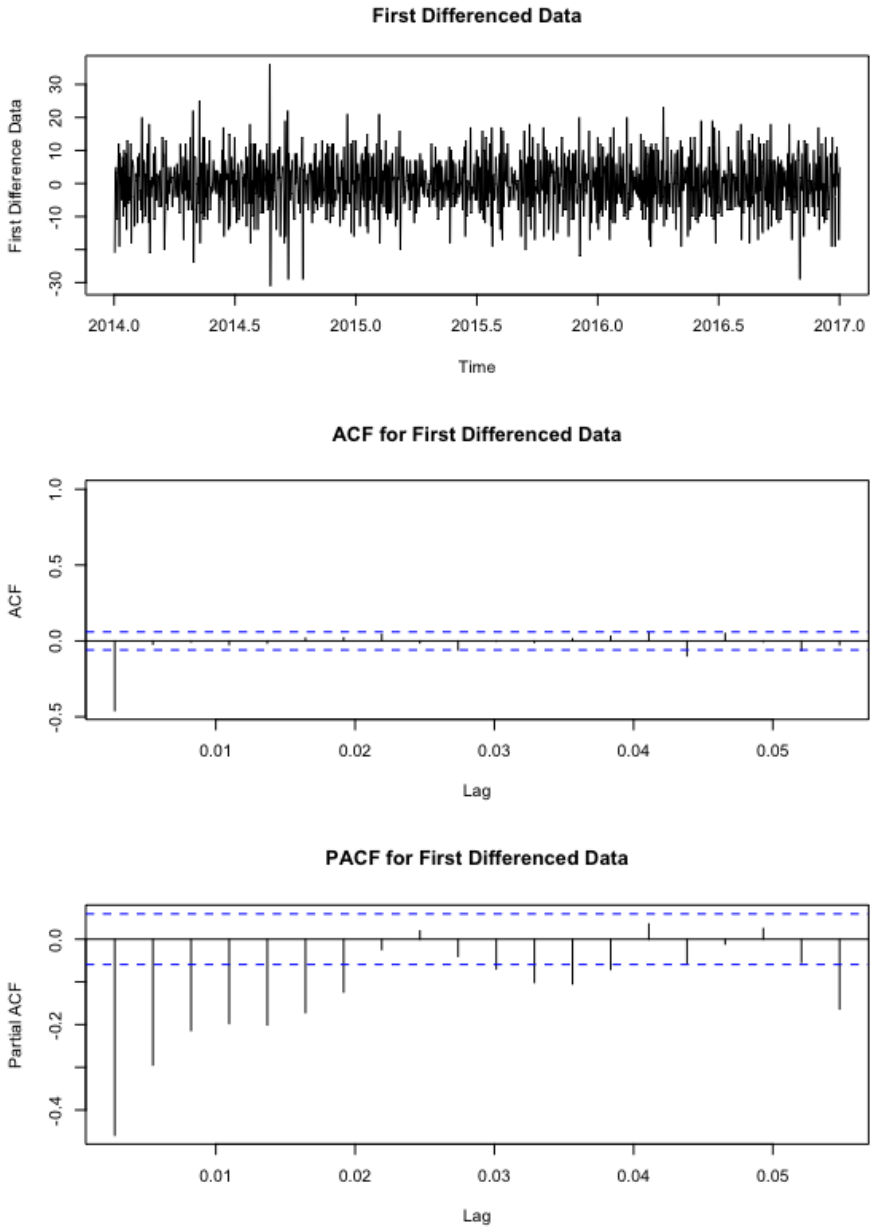

Figure 5: First Differenced Data and the ACF and PACF Plot

Model	AIC
ARIMA(1,1,2)	4.7272
ARIMA(1,1,1)	4.7259
ARIMA(1,1,3)	4.729
ARIMA(0,1,3)	4.7272
ARIMA(0,1,2)	4.7264
ARIMA(0,1,1)	4.7327
ARIMA(1,1,0)	5.043
ARIMA(1,0,1)	4.7231
ARIMA(1,0,2)	**4.7189**
ARIMA(1,0,3)	4.7202
ARIMA(0,0,3)	4.738
ARIMA(0,0,2)	4.7395
ARIMA(0,0,1)	4.7441
ARIMA(1,0,0)	4.7405

Table 2: AIC for Different Models

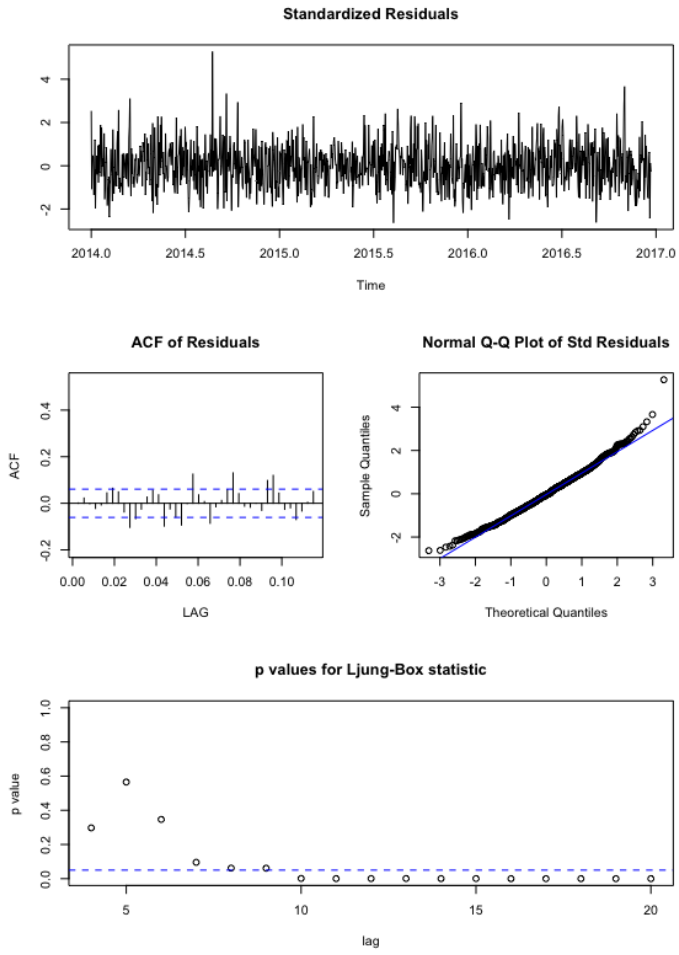

Figure 6: Residual Analysis for ARIMA(1,0,2)

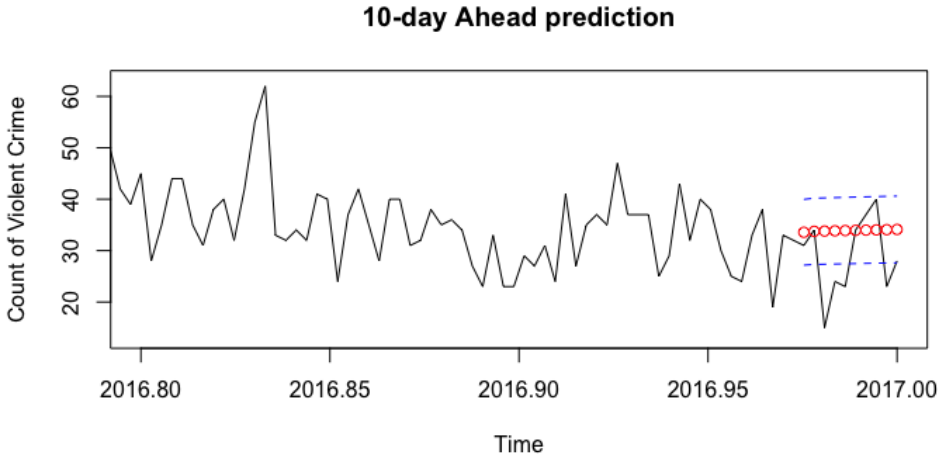

Figure 7: 10-day Ahead Prediction for ARIMA(1,0,2)

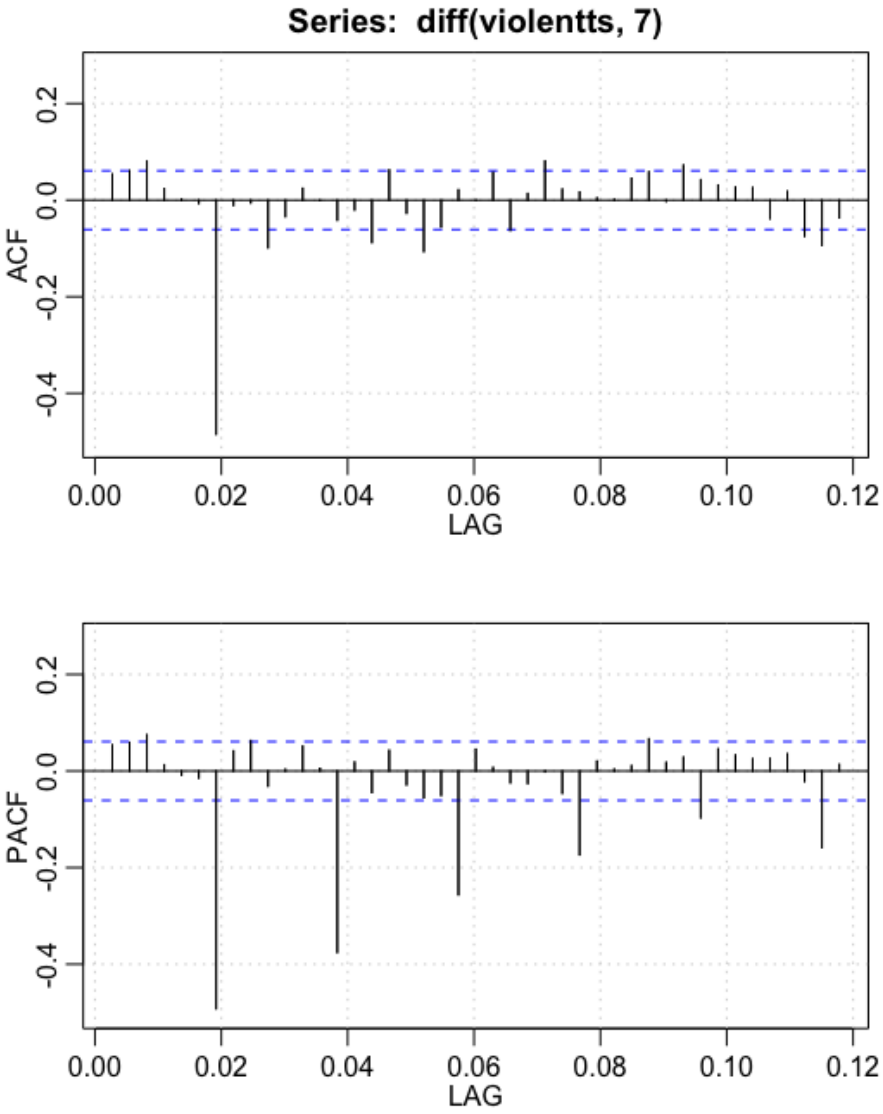

Figure 8: ACF and PACF for Weekly Differenced Data

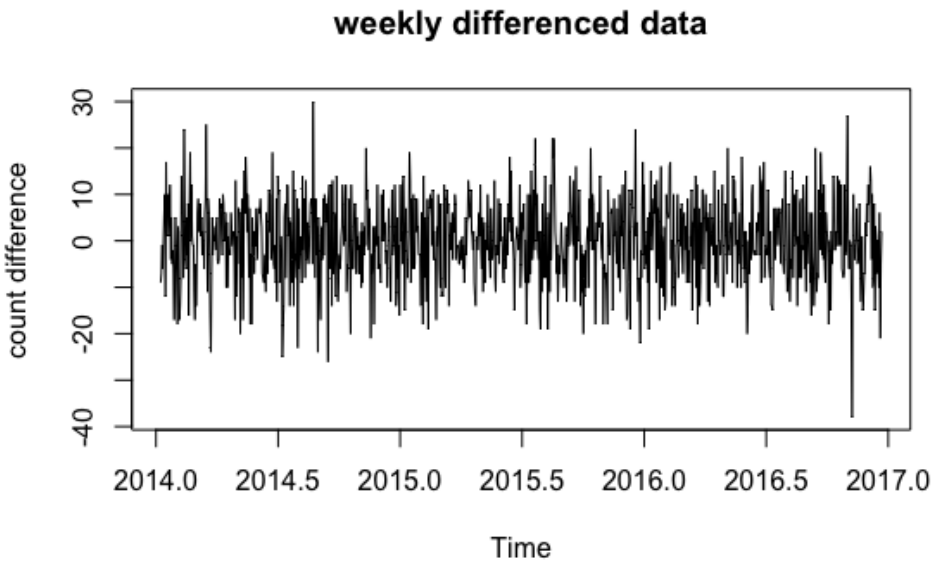

Figure 9: Weekly Differenced Data

Model	AIC
ARIMA(1,0,1) x $(0,1,1)_7$	4.6177
ARIMA(1,0,1) x $(0,0,1)_7$	4.7268
ARIMA(1,0,1) x $(1,0,1)_7$	4.6286
ARIMA(1,0,2) x $(0,1,1)_7$	4.6195
ARIMA(1,0,2) x $(1,1,1)_7$	4.6186
ARIMA(1,0,1) x $(1,1,1)_7$	**4.6167**
ARIMA(1,0,1) x $(1,0,0)_7$	4.7264
ARIMA(0,0,0) x $(1,1,1)_7$	4.6746

Table 3: AIC for Seasonal ARIMA Models

	ar1	ma1	sar1	sma1	constant
	0.9332	-0.8340	-0.0507	-1.0000	0.0062
s.e.	0.0358	0.0565	0.0319	0.0085	0.0014

Table 4: Coefficients for Seasonal ARIMA(1,0,1) x $(1,1,1)_7$

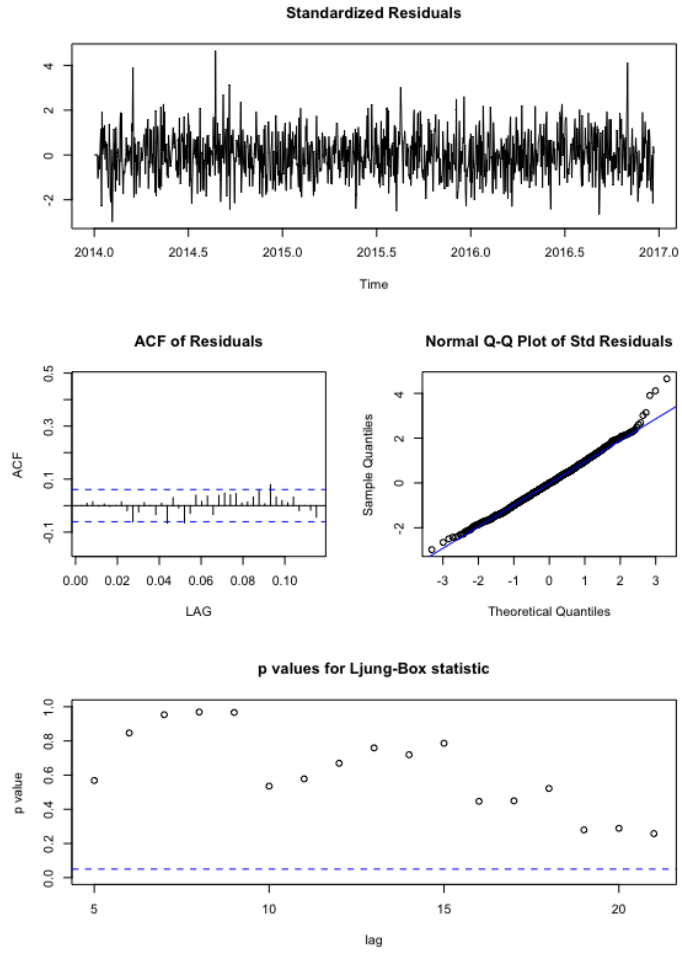

Figure 10: Residual Diagnostic for Seasonal ARIMA(1,0,1) x $(1,1,1)_7$

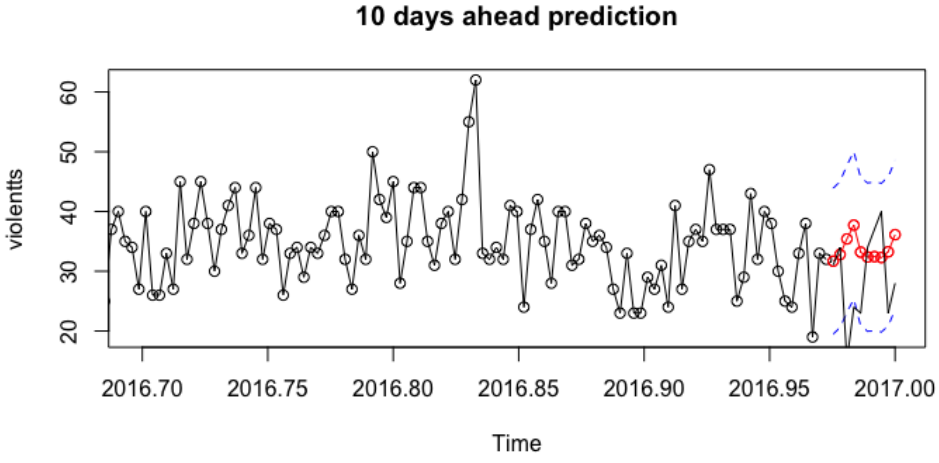

Figure 11: 10-day Ahead Prediction for Seasonal ARIMA(1,0,1) x $(1,1,1)_7$

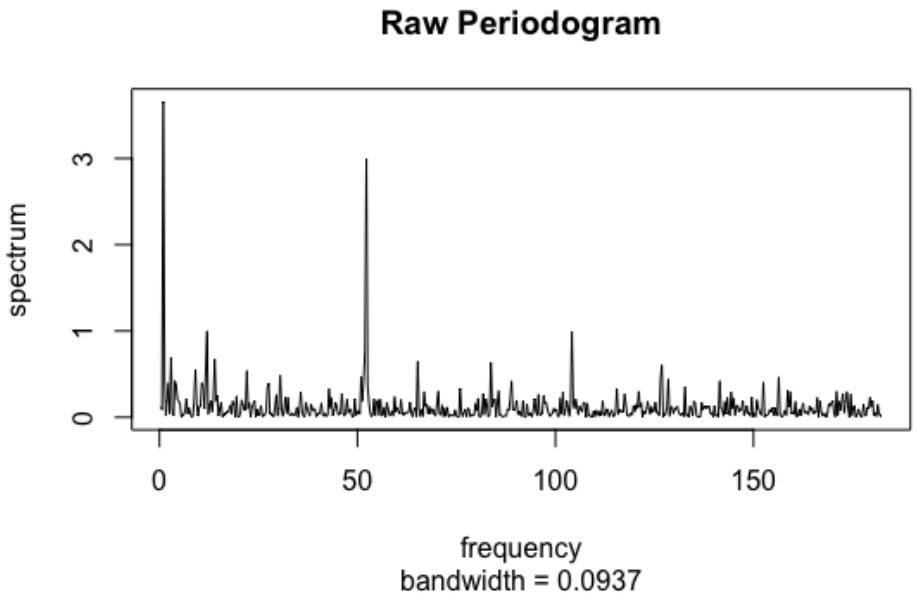

Figure 12: Periodogram

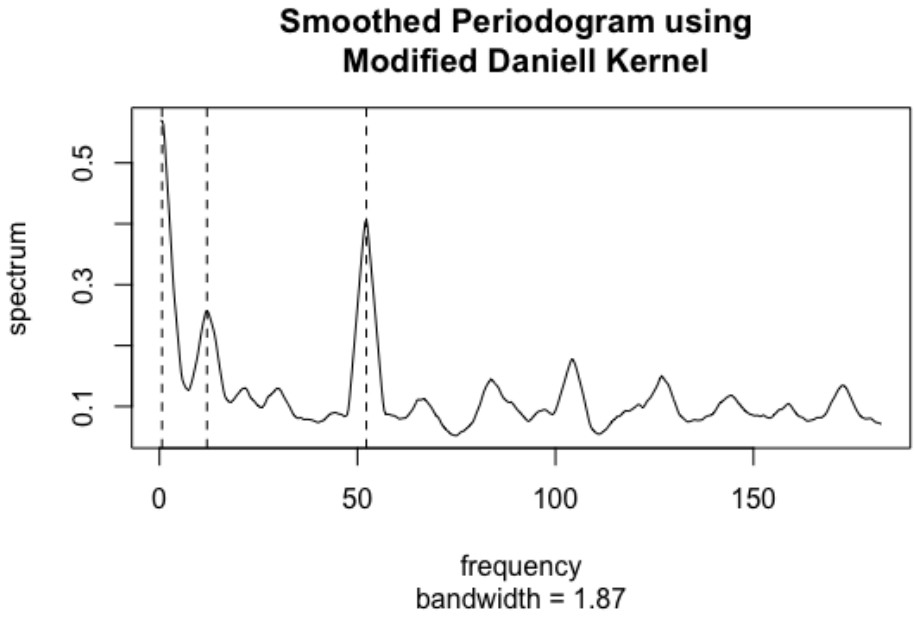

Figure 13: Smoothed Periodogram

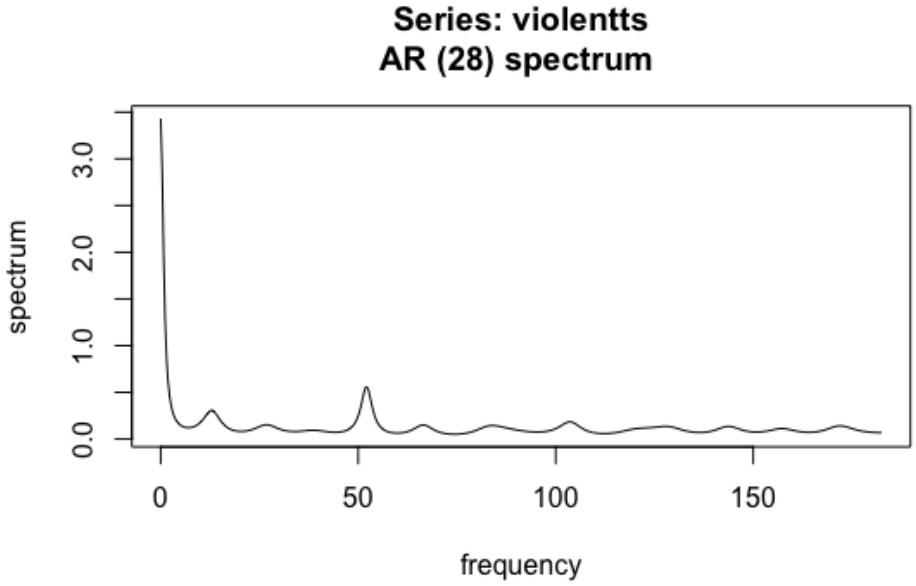

Figure 14: AR Model Approximation

References

[1] Sackett, Chase. Summer 2016, *Neighborhoods and Violent Crime*, https://www.huduser.gov/portal/periodicals/em/summer16/highlight2.html.

[2] Shapiro, R. J., & Hassett, K. A. (2012). *The economic benefits of reducing violent crime: A case study of 8 American cities.* https://www.americanprogress.org/issues/economy/reports/2012/06/19/11755/the-economic-benefits-of-reducing-violent-crime/

[3] Bender, Jeremy, & Kiersz, Andy. *The FBI's most violent cities in each state.* http://www.businessinsider.com/most-violent-cities-in-each-state-2016-1