STA6857 - Applied Time Series Analysis - 4/17/2020

Chicago Crime Rate: Univariate and Multivariate Time Series Analysis

Daniel Gluck

Melody Green

Daniel Hyun

Hoiyin Leung

Salih Tuzen

Table of Contents

Introduction

Background Information

Crime is a major component of managing any metropolitan area, and Chicago is no exception. Cook County, the home of Chicagoland, is the residence of over five million people, which means that proper law enforcement management and policy are essential to its daily city operations. Despite being a beautiful example of modern architecture and a cornerstone of the Midwest, Chicago retains a bad reputation for its excessive crime rate. Though crime in Chicagoland has been in decline for many years, like with any city, crime data remains an important part of Chicago's governmental and infrastructure considerations.

Crime data is harnessed by multiple types of government and private institutions for a variety of goals. In the obvious case, that of law enforcement, such data is harnessed for the management and optimization of the agencies used to combat crime. Seeing where and how crime rates are going to peak allows for police and other law officials to delegate manpower and resources properly. Similarly, crime data is used by politicians, bureaucrats, and other policymakers and regulators to assess the effectiveness of their initiatives. Without having proper access to crime data, they do not know if a given law or policy meant to dissuade crime or alleviate recidivism in a given area is having a proper impact. Finally, private organizations and advocacy groups use crime data to better learn the needs of those they serve. By knowing the crime statistics in a given region, they are better able to target their initiatives to have a larger positive impact.

Crime data and analysis is important for all of these groups, and the analysis of criminal happenings is so important that it is integrated into the hierarchies of the city, state, and federal governments in the United States. Almost every state in the United States has its own Statistical Analysis Center for crime data. For Illinois, that is the Research & Analysis Unit of the Illinois Criminal Justice Information Authority (ICJIA). Agencies such as them take in information from sources across the state and work directly with the FBI's Uniform Crime Reporting (UCR) Program to form databases on all crime-related happenings in the United States.

One of the primary causes of Chicago's high crime rate is its issues with unemployment and poverty. Chicago is one of the most segregated cities in the United States. Within the city proper, areas of extreme poverty, where the highest murder rates in the city are concentrated, can often be observed next to areas of extreme wealth. Studies such as those by the Metropolitan Planning Council advocate addressing rampant unemployment issues as a means of combating Chicago's relatively high crime rates. To say the least, there is likely to be a link shown in any time series analysis that is carried out between Chicago's crime and unemployment rates.

Objectives

The primary goal for this analysis is to use univariate and multivariate time series models to predict the crime rate in Chicago and examine the relationship between the crime rate and the unemployment rate. For this, SARIMA will be used to look at both subclasses of crime and the aggregate crime rate, as well as the unemployment rate. Then, Granger Causality will be used to establish a link between these different crime categories and unemployment. Afterward, a basic VAR will be carried out, followed up by a Two-Stage Model and sVARMA, in order to portray the relationship between crime and unemployment. The outcomes of these methods will be compared against basic Linear Regression in order to highlight the shortcomings of conventional regression and to demonstrate the importance of understanding how to analyze data using a temporal perspective.

Literature Review

Janko and Gurleen (2015) use national and regional Canadian data to analyze the relationship between economic activity as reflected by unemployment rate and crime rates. They use the following model to describe the link between the long-run and short-run effects of unemployment and crime rate.

$$\Delta C_t = \beta_0 + \beta_1 \Delta U_{t-1} + \beta_2 Z_{t-1} + \beta_3 \Delta C_{t-1} + \varepsilon_t$$

Where ΔC_t represents the change of crime rate at time t

β_1 captures the short-run relationship between change in the unemployment rate and change in crime. $Z_{t-1} = C_{t-1} - \gamma U_{t-1}$ captures the long-run relationship between the unemployment rate and crime rate. If γ equals to 0, there is no long-run relationship.

This study finds no significant long-run relationship between crime rate and unemployment rate, but there is a significant short-run relationship between property crime (fraud and theft) and the unemployment rate. Janko and Gurleen use linear regression to explore the relationship between the crime rate and the unemployment rate. Our project aims to use more sophisticated time series models like SARIMA and VAR to explore the link between the crime rate and the unemployment rate.

Data

Source

Our data comes from 2 sources. The crime data is obtained from the Chicago Data Portal supported by the City of Chicago. It has an open-source dataset of Chicago crimes from 2001 to the present. It contains variables such as the date, type, and location of each crime reported on a daily basis.

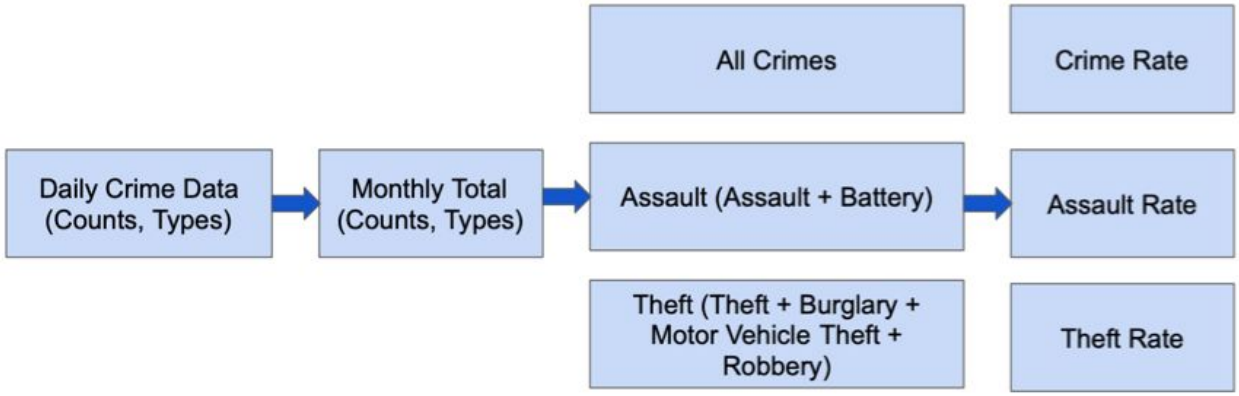
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g

The second data source is the Illinois Department of Statistics. It has monthly unemployment rates in each county. We use the Cook county Chicago Metropolitan Area data.

http://www.ides.illinois.gov/xxlmi/Pages/Local_Area_Unemployment_Statistics.aspx

Data Processing

The time frame of this project is from Jan 2001 to July 2019. Raw data is aggregated to the monthly level. Then all crimes are added as All Crimes. The two most common types of crimes Theft (Theft, Burglary, Motor Vehicle Theft, Robbery) and Assault (Assault and Battery) are extracted and aggregated. Then different crime rates are calculated according to the equations below. These are crime rates per 100,000 people.



$$Crime\ Rate = \frac{Monthly\ Count\ of\ All\ Crimes}{Population} \times 100,000$$

$$Theft\ Rate = \frac{Monthly\ Count\ of\ Theft}{Population} \times 100,000$$

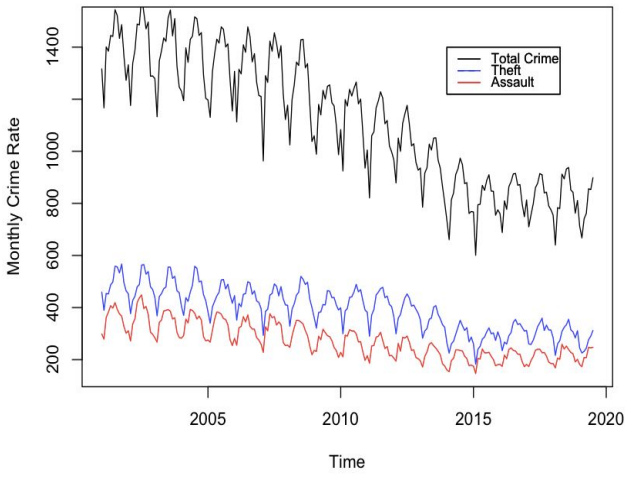
$$Assault\ Rate = \frac{Monthly\ Count\ of\ Assault}{Population} \times 100,000$$

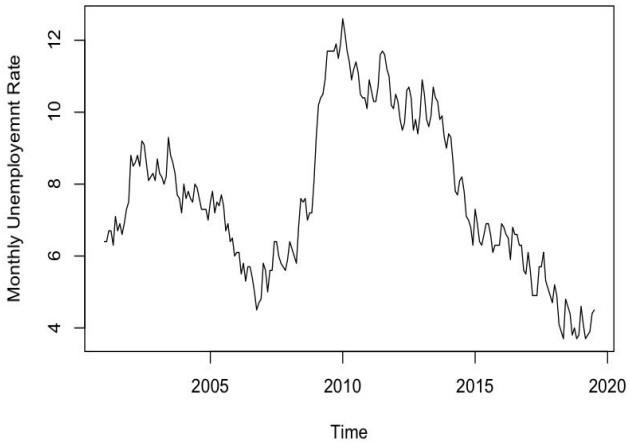
The table below is a brief summary of the 4 time series in the project. Among all crimes, about 1/3 are theft, less than 1/3 are assault. There are 223 data points in each time series. We use the data from Jan 2001 to Dec 2017 as a train set, those from Jan 2018 onwards as the test set.

	Mean (%)	**S.D.**
Crime Rate	1113.77	250.86
Theft Rate	391.90	85.90
Assault Rate	272.98	67.82
Unemployment Rate	7.66	2.2

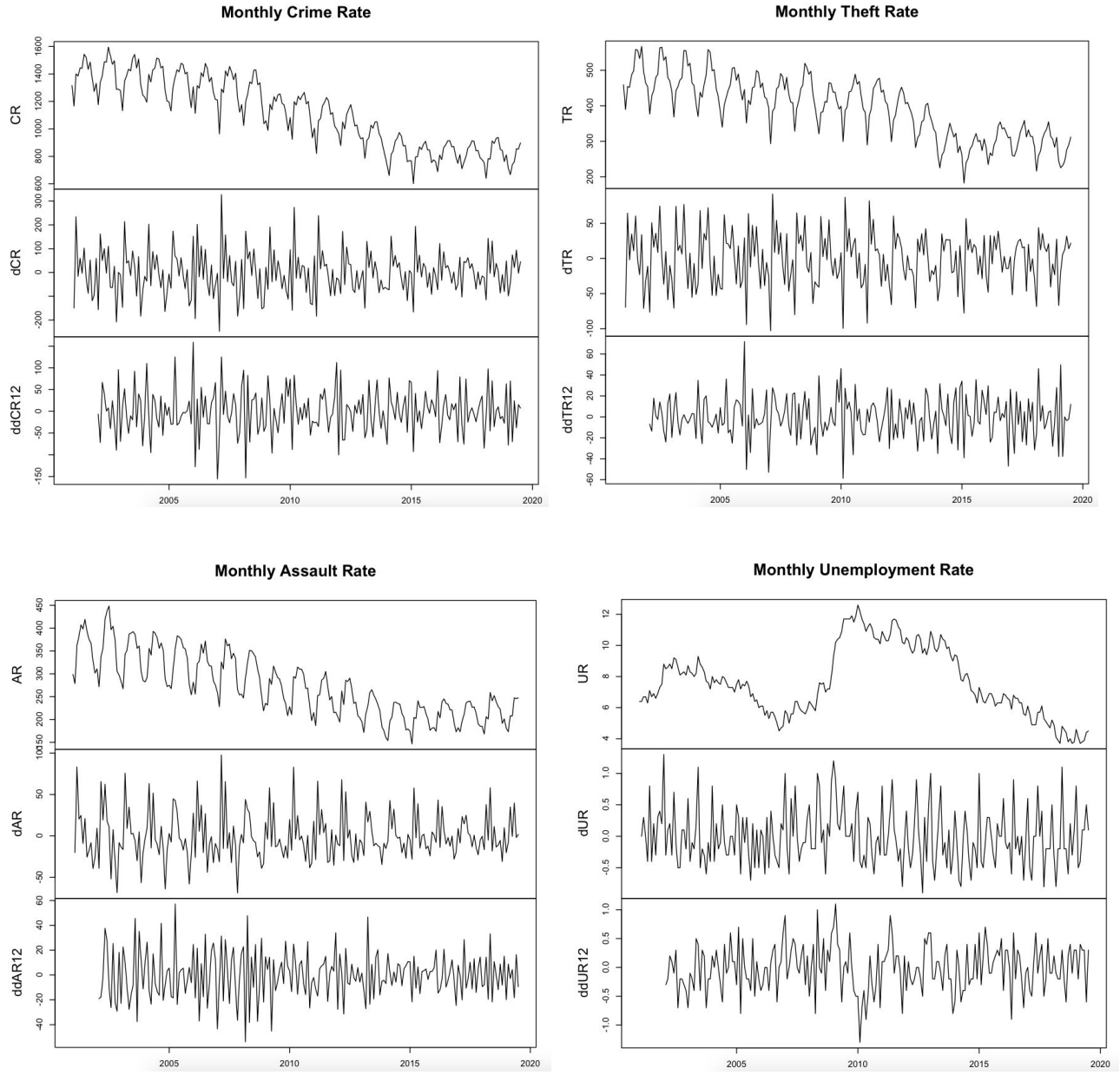
Descriptive Analysis

The following plots depict the monthly total crime, theft, and assault rates, followed by the monthly unemployment rates for the City of Chicago from January 2001 until July 2019.





The three crime rate series have a distinct 12-month seasonality; while this same feature is not quite evident in the unemployment rate, in later sections we show that this series also has a 12-month seasonality. Additionally, all of the above exhibit fluctuations in their variance. Because of these two characteristics, differencing and seasonal differencing is required to make our data stationary. As such, the following plots exhibit how each of the differencing steps we take makes each series stationary.



The top panels show the undifferenced data, the middle shows the once-differenced data, and the bottom panels show the data after differencing and 12-month seasonal differencing, for each of the time series. We further show that these steps are necessary in the following section Unit Root Testing.

Unit Root Testing

One important step in SARIMA modeling is to check if the time series is stationary or not. In other words, we check if the mean and variance of a time series are constant over time. This step is needed to detect the correct AR (auto-regressive), MA (moving average), and I (integrated) terms for both seasonal and non-seasonal components. A formal test to check for seasonality is the Augmented Dickey-Fuller test. The ADF has the following null and research hypothesis;

Null Hypothesis (H0): time series has a unit root, meaning it is non-stationary.

Alternate Hypothesis (H1): time series does not have a unit root, meaning, it is stationary.

We have seen that all the variables have either some kind of trend or seasonality. As a first step, we applied first-order differencing and conducted ADF tests. The p-values in the below table show that all series do not have a unit root, meaning they are stationary. Thus, we are ready to move to the SARIMA process knowing that the integration term is 1.

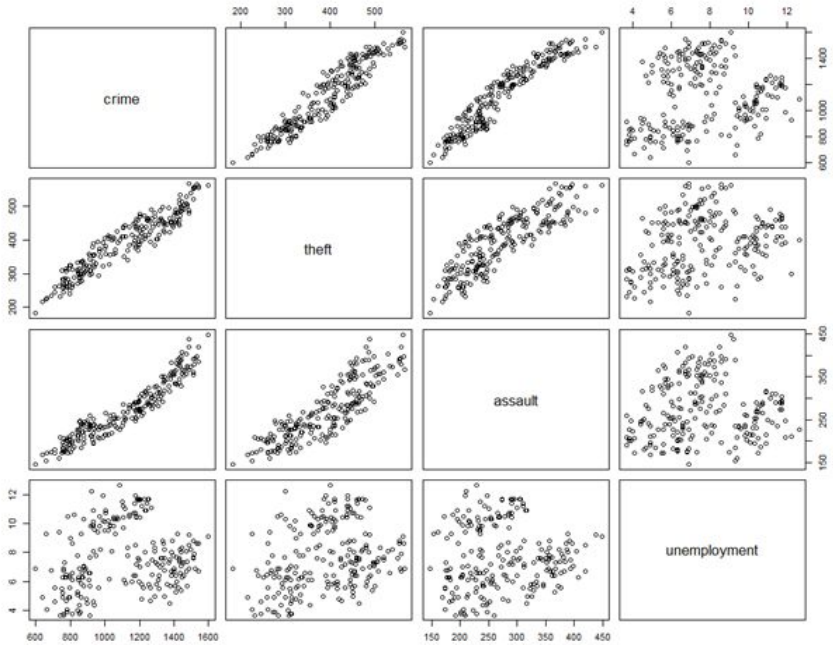
Variable	P-value	Conclusion
Crime Rate	0.0039	Stationary
Assault Rate	0.0029	Stationary
Theft Rate	0.0042	Stationary
Unemployment Rate	0.0071	Stationary

Modeling

Linear Regression

To start off the modeling process, we wanted to start with classical linear regression in the context of time series. Our main objective with fitting a linear regression model was to not only give us a baseline model for which we could compare other models, but to also give us insight into its limitations in predicting time series data. We also wanted to see how much our knowledge and understanding of time series modeling that we've learned over the semester could be applied to real data, and why it can be much more efficient over the classical linear regression model in this context.

When modeling with multiple linear regression, one assumption that needs to be met is the independence of predictors from other predictors in the model - or in other words, multicollinearity. From the descriptive analysis section earlier, we saw that the crime rate had high collinearity with theft rate and assault, which makes sense given the nature of the data. This can be further seen in a scatterplot matrix of our variables, below. We see that there is a strong correlation between crime, theft, and assault rate. Since our objective with this project is to explore the relationship between crime rate and unemployment rate with other variables, we decided to remove the theft rate and assault rate from our linear regression process.



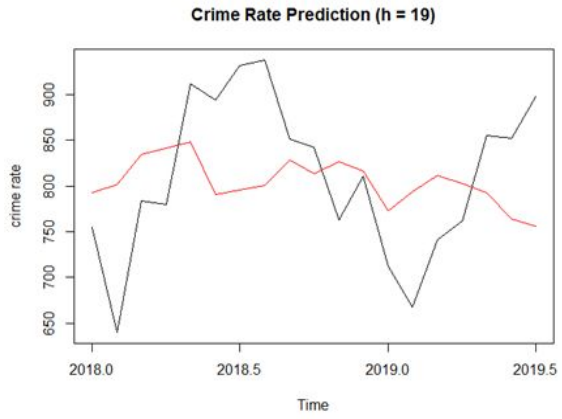
With only the crime rate and unemployment rate left, we shift our focus on the scatterplot matrix between crime rate and the unemployment rate to explore their relationship. We notice that as the unemployment rate increases, crime increases as well, indicating a possible linear relationship.

There is also the possibility that as the unemployment rate continues to increase, crime rate peaks but then falls, indicating a negative quadratic relationship. With this in mind, we decided to try a few tentative models and find which one performs the best.

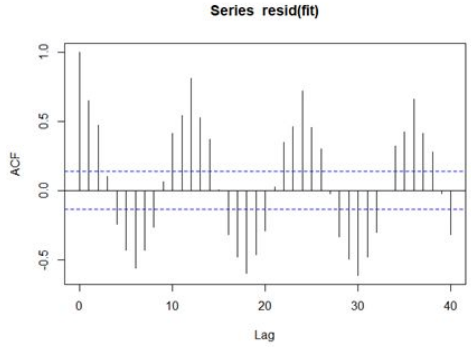
In the following table, we have the results of three different tentative models for predicting the crime rate. The first model is a trend only model. The second model has trend and linear unemployment rate. Finally, the third model has a trend, linear unemployment rate, and negative quadratic unemployment rate. Based on the coefficient of determination (), AIC, and BIC, we can see that the third tentative model performed the best in predicting the crime rate. Roughly 80% of the variance in the crime rate can be explained by the model.

Tentative Model	R^2	AIC	BIC
$CR_t = \beta_0 + \beta_1 t + w_t$	.786	10.52588	10.57172
$CR_t = \beta_0 + \beta_1 t + \beta_2 UR_t + w_t$	.785	10.53292	10.59405
$CR_t = \beta_0 + \beta_1 t + \beta_2 UR_t - \beta_3 (UR_t)^2 w_t$	.796	10.48602	10.56242

Fitting the model to the training data and predicting on the testing set, we can see the results in the plot below. The black line indicates the actual data, and the red lines are the predicted values. As we can see, the predicted values do not perform very well, with an RMSE of ~89.



As expected, it predicted the points in a linear manner, thereby not capturing the ups and downs of seasonality. Furthermore, if we look at the autocorrelation of residuals of the fitted model below, we can see a high correlation at every 12 lags, indicating that there is a high presence of seasonality in the data that the linear model is not able to capture. The autocorrelations are also exceeding the threshold (indicated by the dotted blue line) showing significance in the residuals. In addition to the model not taking into account the seasonality, another weakness is that it only performs predictions based on concurrent values of the unemployment rate, without using potentially useful information based on its past values. Performing linear regression on the other three variables as dependent variables performed similarly, so without showing all the outputs, we decided to move on to SARIMA models.



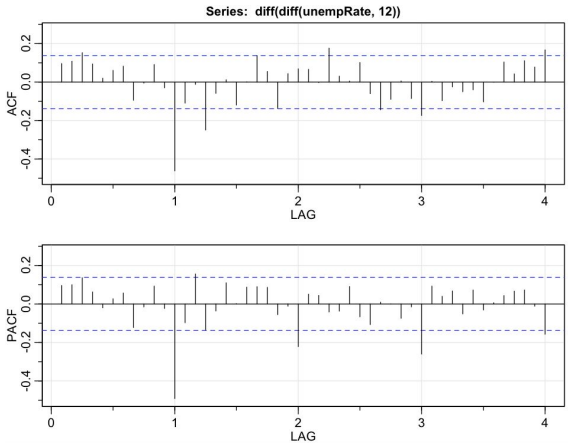
SARIMA

In the section above, we saw that there were some severe limitations when performing linear regression on time series data. Linear regression does not take into account seasonality or patterns in the data, and it also does not use historical data, which can potentially contain significant information in predicting the future. In this section, we will use SARIMA models, which will utilize autoregression, moving averages, and seasonality.

Unemployment Rate

Identifying Orders:

Our primary interest in this project is to detect the relationship between the crime rate and other variables in the data set. Before that, however, we wanted to see how the unemployment rate can be predicted from its own past values. Initially, we checked the ACF, PACF plots of the first differenced time series and noticed persistent seasonal peaks. That could potentially obscure the orders, therefore ACF and PACF of first differenced and seasonally differenced time series is plotted.



We see above that ACF cuts off at lag 1s, while the PACF tails off every ks (k=1, 2, 3, ...) implying a season MA (1) model. For the non-seasonal component, we look at the lower lag values. Both plots

tail off after 3 lags, implying an ARMA (3, 3) model. Our proposed model for this data set is thus: ARIMA (3, 1, 3)x(0, 1, 1)$_{12}$

Fitting and Prediction:

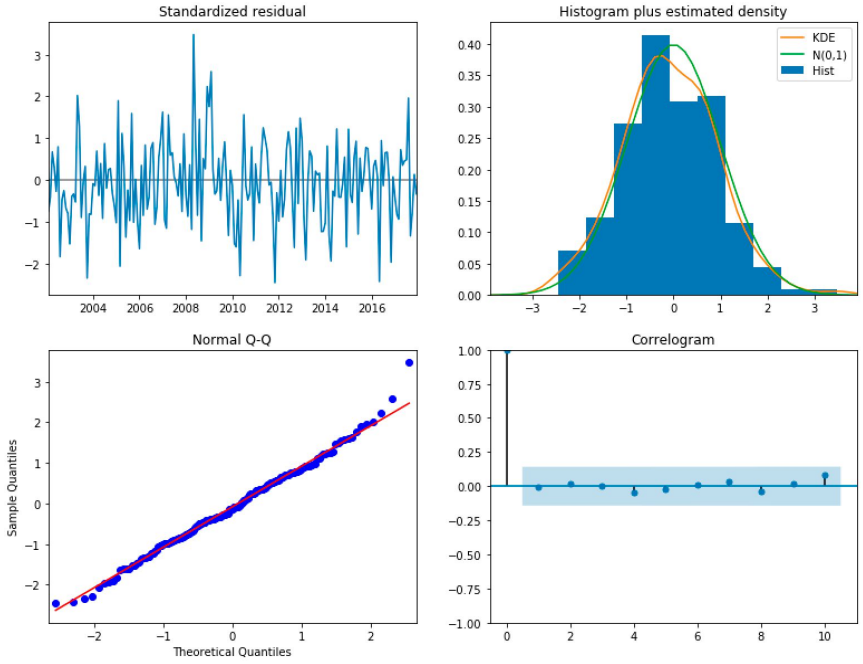
We fit the SARIMA model with the orders identified above. The p-values are significant for all coefficients except MA lag2. The AIC is 118, the lowest we could obtain. Next, we will validate our model by looking at diagnostics. The Q-Q plot and histogram show the residuals are approximately normally distributed. We don't see an apparent pattern in the plot of the standardized residuals. The correlogram plot also shows the residuals are not correlated. The diagnostic plots verify the model's reliability.

Statespace Model Results

Dep. Variable:	UnemRate	No. Observations:	204
Model:	SARIMAX(3, 1, 3)x(0, 1, 1, 12)	Log Likelihood	-51.424
Date:	Thu, 16 Apr 2020	AIC	118.847
Time:	11:54:52	BIC	144.866
Sample:	01-01-2001	HQIC	129.386
	- 12-01-2017		
Covariance Type:	opg		

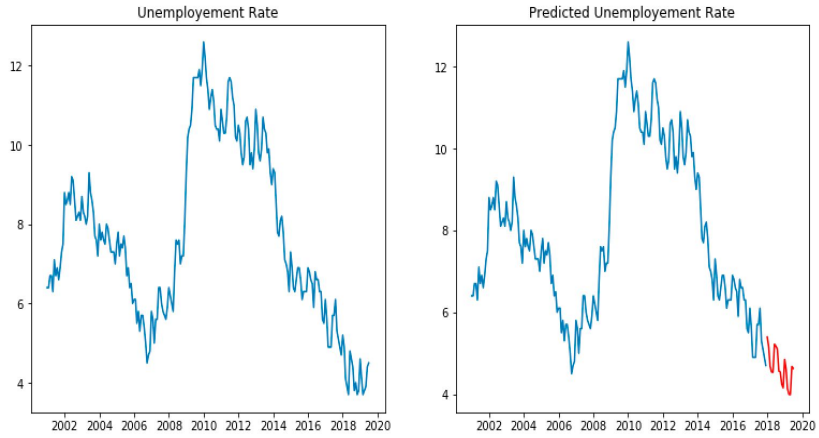
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7862	0.099	-7.962	0.000	-0.980	-0.593
ar.L2	0.5144	0.160	3.217	0.001	0.201	0.828
ar.L3	0.8696	0.092	9.403	0.000	0.688	1.051
ma.L1	0.8505	0.131	6.477	0.000	0.593	1.108
ma.L2	-0.2935	0.203	-1.447	0.148	-0.691	0.104
ma.L3	-0.6942	0.123	-5.650	0.000	-0.935	-0.453
ma.S.L12	-0.8800	0.081	-10.911	0.000	-1.038	-0.722
sigma2	0.0920	0.010	9.161	0.000	0.072	0.112

Ljung-Box (Q):	35.55	Jarque-Bera (JB):	2.17
Prob(Q):	0.67	Prob(JB):	0.34
Heteroskedasticity (H):	0.93	Skew:	0.18
Prob(H) (two-sided):	0.79	Kurtosis:	3.37



Upon verifying the model, we used it to predict unemployment rates from January 2018 to 2019 July. We see a good matchup between the actual values on the left and the predicted ones on the right

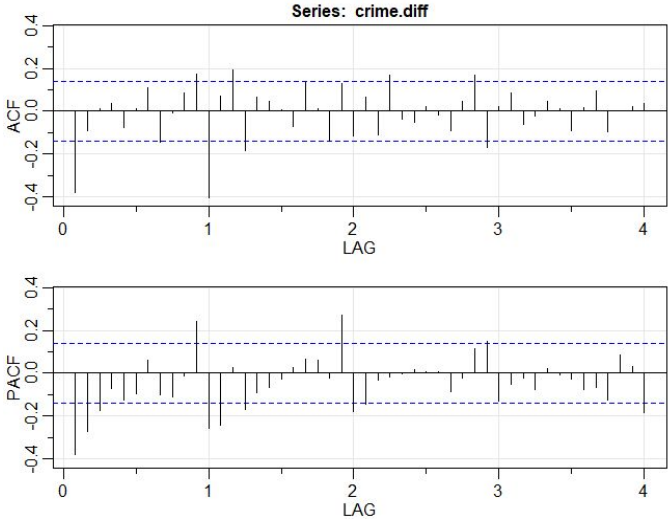
with red color. The accuracy of the prediction is also quantified with an RMSE value of 0.36. We will compare models with RMSE measures in our final comparison.



Crime Rate

Identifying Orders:

Similar to the unemployment rate, we look at the ACF and PACF plot of the first differencing of seasonal differenced crime rate series to identify the orders for the SARIMA model. In the plot below, we can see that for the seasonal component (lags ks for k=1, 2, 3, …), the ACF cuts off after lag 1, while the PACF tails off, indicating an MA(1) behavior. For the non-seasonal component, we look at the lower lags. The ACF plot shows a sharp cutoff after lag 1, while the PACF plot again tails off, indicating an MA(1) behavior. Thus, our SARIMA model for crime rate is ARIMA (0, 1, 1)x(0, 1, 1)$_{12}$

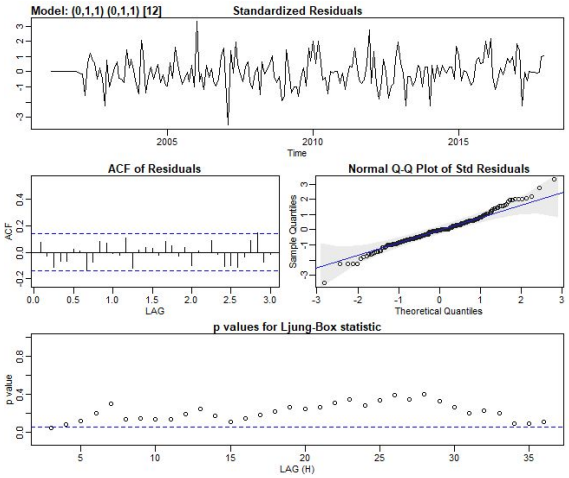


Fitting and Prediction:

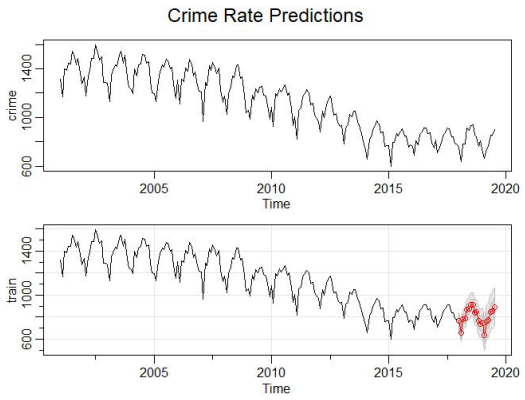
Using the same partitioning as in the previous sections, with the testing set consisting of crime rate data points from January 2018 to July 2019, we fit the SARIMA model to the training data and observe the results below. We see that both the seasonal and non-seasonal components are significant in predicting the crime rate. Our lowest AIC and BIC for SARIMA models on crime rate were 9.61 and 9.65, respectively.

Furthermore, looking at the diagnostics of the fitted model below, we see that the standardized residuals are approximately a white noise. This is also confirmed by the normal Q-Q plot, which indicates no significant departure from normality in the residuals. Also, looking at the ACF of residuals, there does not seem to be any serial correlation, which would otherwise be shown with ACFs exceeding the blue dotted line. This is also verified by looking at the Ljung-Box test statistics. The high p-values show that the autocorrelation in the residuals is random.

```
Coefficients:
          ma1     sma1
      -0.6173  -0.6782
s.e.   0.0668   0.0508

sigma^2 estimated as 1350:  log likelihood = -1058.76

$degrees_of_freedom
[1] 208

$ttable
     Estimate     SE  t.value p.value
ma1   -0.6173 0.0668  -9.2467       0
sma1  -0.6782 0.0508 -13.3518       0

$AIC
[1] 9.608705

$AICc
[1] 9.608954

$BIC
[1] 9.654141
```



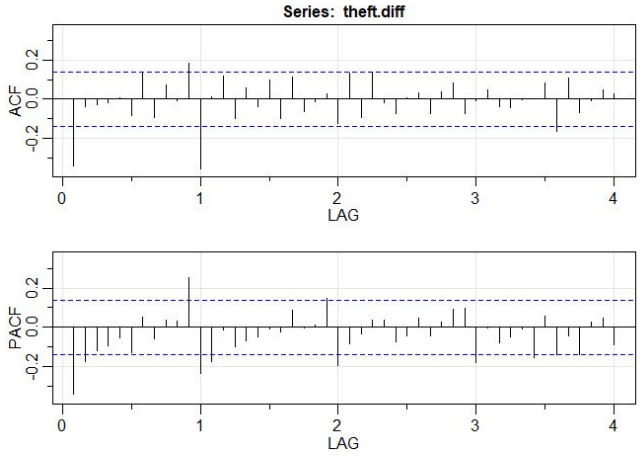
The next step is to perform a prediction using the model on the testing set. Looking at the 19-steps ahead prediction results below, we see that the SARIMA model performs really well in predicting crime rate. The upper plot shows the actual values of the crime rate series, and in the bottom plot, we can the predicted value in the red line. Overall, the SARIMA model was able to predict the crime rate with an RMSE of 25.13.



Theft Rate

Identifying Orders:

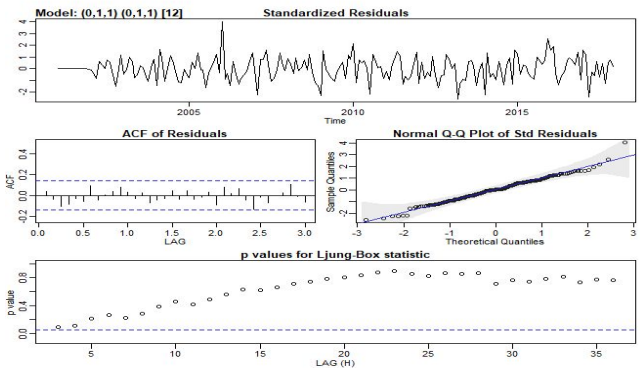
To identify the orders of the SARIMA model for the theft rate, we will again look at the ACF and PACF plot. Looking at the plots below, we see that it behaves similarly to the crime rate series. Looking at the seasonal lags ks (k = 1, 2, 3, …), we see that the ACF cuts off after lag 1, and the PACF tails off, indicating an MA(1) model. For the lower lags, again, the ACF cuts off after lag 1, and the PACF tails off, typical of a MA(1) model. Therefore, our SARIMA model for predicting theft rate is ARIMA (0, 1, 1)x(0, 1, 1)$_{12}$



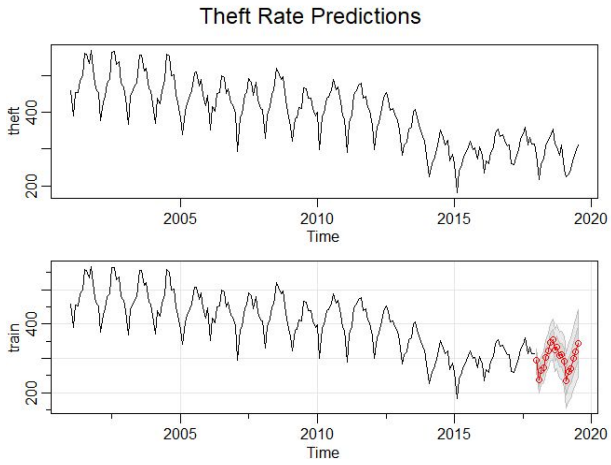
Fitting and Prediction:

We use the SARIMA model from above and fit it on our training data. Below, we show the results of fitting the model, and we can see that the p-values of the seasonal and non-seasonal components are significant. Overall, the lowest AIC and BIC we got for predicting the theft rate was 7.99 and 8.04, respectively. We also see from the model diagnostics that the standardized residuals follow a white noise, with the Q-Q plot showing normality of the residuals. The ACF of the residuals plot shows no indication of serial correlation, and the Ljung-Box test further confirms that the correlation of residuals follows a random pattern.

```
Coefficients:
          ma1     sma1
      -0.4569  -0.6547
s.e.   0.0712   0.0552

sigma^2 estimated as 249.3:  log likelihood = -880.9

$degrees_of_freedom
[1] 208

$ttable
     Estimate     SE  t.value p.value
ma1   -0.4569 0.0712  -6.4181       0
sma1  -0.6547 0.0552 -11.8635       0

$AIC
[1] 7.999072

$AICc
[1] 7.999322

$BIC
[1] 8.044508
```



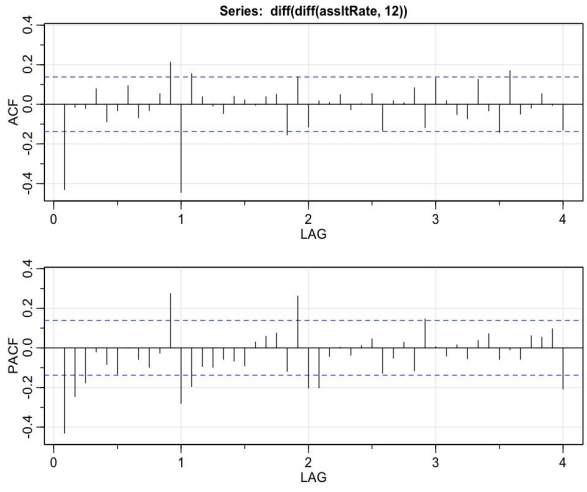
Based on these results, we know that this SARIMA model is valid for predicting the theft rate. Next, using this model to predict the theft rate in the testing data, we look at the results of the 19-steps ahead prediction below. The upper plot shows the actual values, while the lower plot shows the actual values + predicted values (in red). We can see that predicted values closely resemble the actual values, along with the 95% confidence intervals, which seem reasonable. Our forecasting results give an RMSE of 20.83. Considering the range of theft rate, this SARIMA model performs really well with predictions.



Assault Rate

Identifying Orders:

The last variable to fit the SARIMA model is the assault rate. We repeat the steps in previous sections. This model will show us how well we can predict the assault rate if we only use its own lag and shock values as predictors. The ACF and PACF plots of the first differenced time series showed persistent seasonal peaks. Therefore, we apply a seasonal difference over the first differenced data. The ACF and PACF of the resulting series are below. The ACF and PACF behave the same way as the crime rate and the theft rate, so our proposed model is once again ARIMA (0, 1, 1) x $(0, 1, 1)_{12}$



Fitting and Prediction:

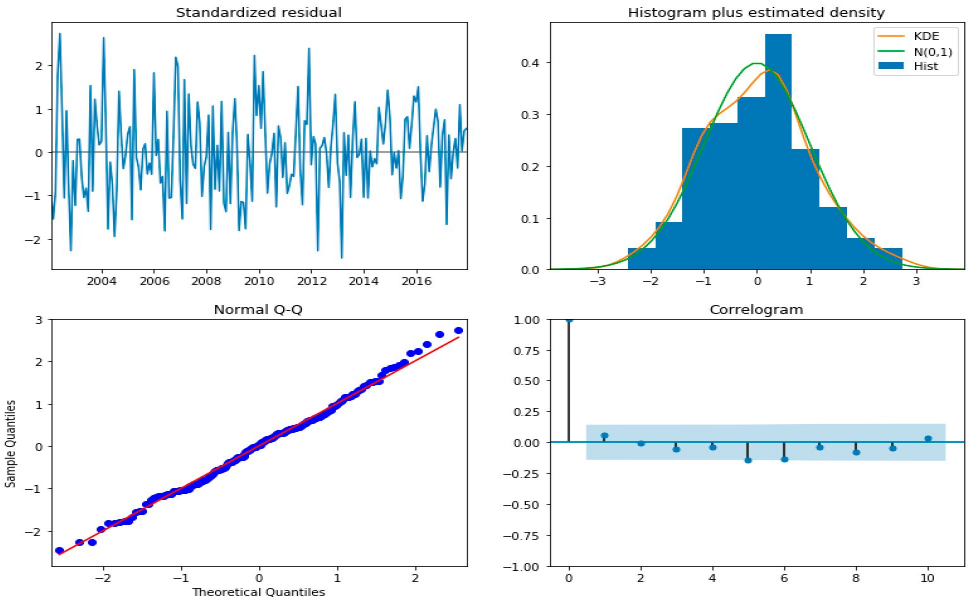
We fit the SARIMA model with seasonal and non-seasonal MA1 terms and with seasonal/non-seasonal differences. The p-values are significant for all coefficients. The AIC is 1545. We fit a few other models and noticed they didn't help to lower the AIC. Since AIC is calculated based on residuals, it is affected by the range of the variable that is being predicted. So, we only used AICs to compare models for the same outcome variable. We also see that the Ljung-Box test p-value of 0.38 is not significant. This means the residuals overall are random and independent. Next, we will look at the diagnostics. The Q-Q plot and histogram show the residuals are approximately normally distributed. We don't see an apparent pattern in the plot of the standardized residuals. The correlogram plot also shows the residuals are not correlated. The diagnostic plots again verify the model's reliability.

Statespace Model Results

Dep. Variable:	AssaultRate	No. Observations:	204
Model:	SARIMAX(0, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-769.980
Date:	Tue, 14 Apr 2020	AIC	1545.959
Time:	17:18:01	BIC	1555.716
Sample:	01-01-2001 - 12-01-2017	HQIC	1549.911
Covariance Type:	opg		

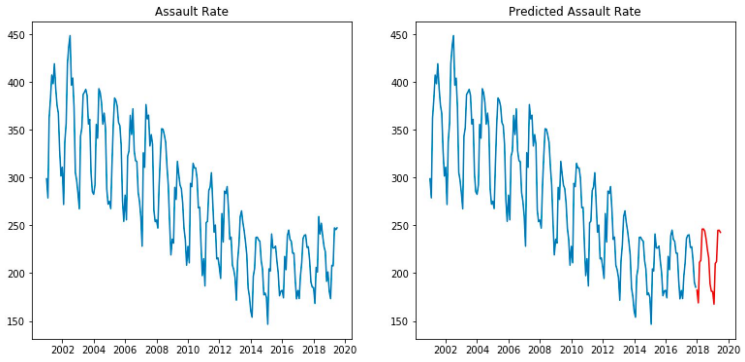
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.5836	0.060	-9.739	0.000	-0.701	-0.466
ma.S.L12	-0.6796	0.076	-8.965	0.000	-0.828	-0.531
sigma2	178.3254	19.388	9.198	0.000	140.325	216.325

Ljung-Box (Q):	42.12	Jarque-Bera (JB):	0.81
Prob(Q):	0.38	Prob(JB):	0.67
Heteroskedasticity (H):	0.44	Skew:	0.14
Prob(H) (two-sided):	0.00	Kurtosis:	2.86



Finally, we will check how the identified model will predict the test set which has assault rates from January 2018 to 2019 July. The actual data points and the predicted values are shown in the below plot. The RMSE value based on these 19 predictions is 7.35. This is a pretty low error rate when we

consider the range and standard deviation of the assault rate. The standard deviation of the assault rate is 67.82.



Multivariate Analysis: Granger Causality

SARIMA is a sophisticated univariate time series model; however, it does not consider the influence of other time series. In this project, we want to explore whether the unemployment rate, crime rate, theft rate, or assault rate are interrelated to one another. We use the Granger causality test to check whether these relationships are significant, if so, in what way they are affecting each other.

Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another. It does not test a true cause-and-effect relationship, but it can test if a variable comes before another in time series. The null hypothesis for the test is that lagged x-values do not explain the variation in y. In other words, it assumes that x(t) doesn't Granger-cause y(t).

It uses F-test to check if the unrestricted model is significantly different from the restricted model. The unrestricted model includes the AR terms of another time series but the restricted model does not.

$$Restricted\ Model \rightarrow y(t) = \sum_{i=1}^{\infty} \alpha_i y(t-i) + c_1 + v_1(t)$$

$$Unrestricted\ Model \rightarrow y(t) = \sum_{i=1}^{\infty} \alpha_i y(t-i) + \sum_{j=1}^{\infty} \beta_j x(t-j) + c_2 + v_2(t)$$

$$F = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/(n-k)}$$

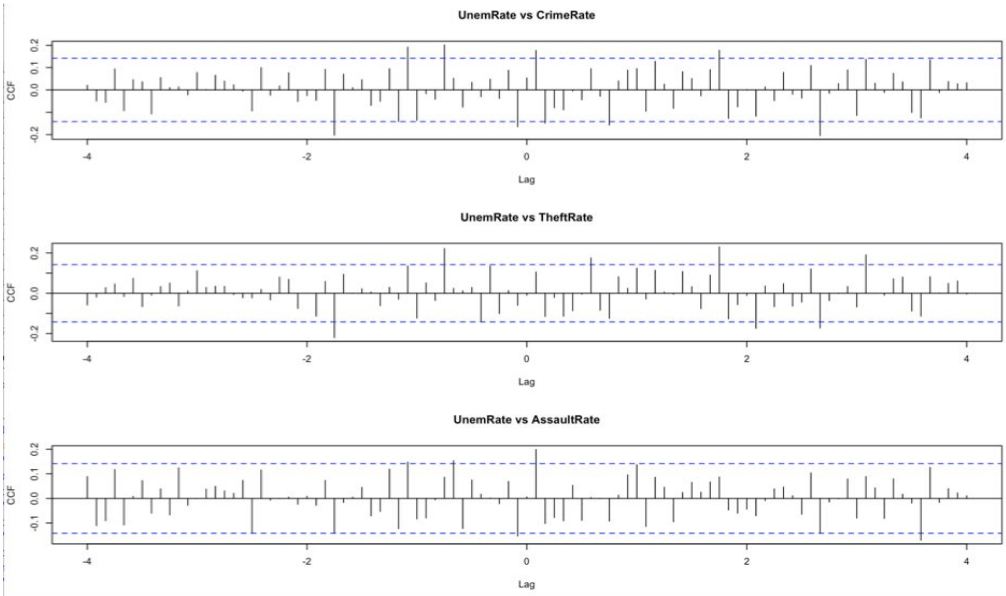
n: number of observations

k: number of parameters from unrestricted model

p: the difference in numbers of parameters from the restricted model

The results of Granger causality test are shown in the following table. If we look at the past 3 lags, only the crime rate and assault rate show evidence of leading the unemployment rate. But if we look at the past 12 lags, all except one are interrelated to one another. The unemployment rate leads assault rate but the assault rate does not lead unemployment rate. However, despite their similar patterns shown in time plots, it is interesting to see the crime rate, theft rate, and assault rate do not Granger cause one another.

	P-Values (Order = 3)	P-Values (Order = 12)
CrimeRate ~ UnemRate	0.272	0.015
UnEmRate ~ CrimeRate	0.008	0.002
TheftRate ~ UnemRate	0.300	0.018
UnemRate ~ TheftRate	0.225	0.002
AssaultRate ~ UnemRate	0.394	0.040
UnemRate ~ AssaultRate	0.007	0.070
Insignificant p-values among CrimeRate, TheftRate, and AssaultRate		



The above CCF plots further illustrate their interrelatedness. In the first plot, the unemployment rate is correlated with crime rate both in the positive and negative lags, indicating unemployment rate both leads and lags crime rate. The same is true for the unemployment rate and the theft rate. However, in the third plot, the CCF is only significant on the positive side, not the negative side, indicating the unemployment rate leads the assault rate but not in the other direction. The result justifies the use of VAR modeling for this system of time-series.

VAR

Similar to the Granger Causality analysis above, we employ a Vector AutoregRessive model (VAR) to investigate the interdependencies of the four times series. The general form of a VAR process of order p is:

$$x_t = \alpha + \sum_{j=1}^{p} \phi_j x_{t-p} + w_t$$

Here, x_t is a vector containing all of the sub-models, or variables, being fit by the model. So for our analysis, this would be the Crime, Theft, Assault, and Unemployment Rate data. The alpha parameter is used to fit a constant value to each of these data sets, while w_t is their corresponding white noise estimates. For each lag value from 1 to p, we fit a transition matrix, Φ_j, which reveals how each variable contributes to itself and the others, from 1 to p months in the past.

For our VAR analysis, we start by identifying the most appropriate order, p, for the VAR model, then use these results to find the equation that best describes each time series. We analyze the residuals for model accuracy, and finish by analyzing the predicted values of the last 19 months to their actual recorded values.

Identifying Orders

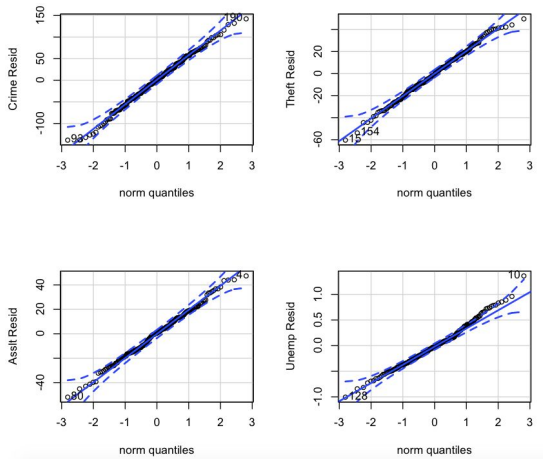
Using the R package MTS, we run the command VARorder on our vector of time series, x_t. Since our data is clearly seasonal, we test p values up to a maximum of 12. The following results outline the best orders according to the AIC, BIC, and HQ selection criteria.

```
selected order: aic =  12
selected order: bic =  3
selected order: hq =  12
```

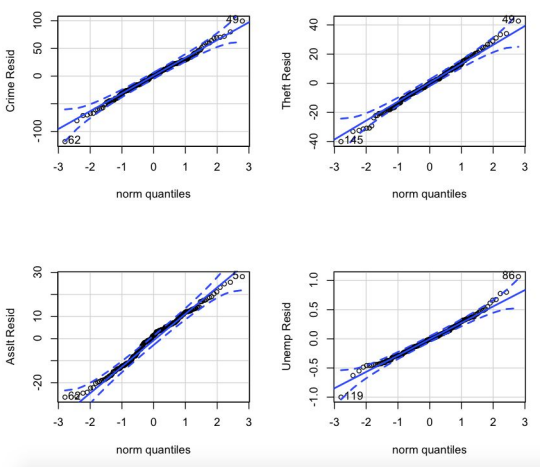
From this output, we see that AIC and HQ indicate that the most appropriate order for our model is p=12, while the BIC, as usual, selected a smaller model of order p=3. To determine which model to proceed with, we first fit both to our data and analyze the resulting residuals. We already know that the AIC and HQ values will favor the larger model, so further comparisons using the selection criteria are unnecessary.

Except for a few deviations in the unemployment rate for the smaller model, and in the assault rate for the larger model, the residuals from both are normally distributed, indicating both models are valid. Since the larger model is additionally backed by more selection criteria than the smaller, this is the model with which we proceed.

Normal Plot of Residuals for VAR(3) Model



Normal Plot of Residuals for VAR(12) Model



Fitting the Model

Our next step is to fit the VAR(12) model to x_t using the MTS function VAR. Preliminary results show that this model is dominated by insignificant predictors, which will lead to inaccurate estimations and predictions. Below are two examples of this, as seen in the Φ estimations for the Crime Rate and Unemployment rate time series. The stars indicate which estimates are actually significant; both images show how few significant estimates there are.

Crime Rate

	Estimate	Std. Error	t value	Pr(>|t|)	
crime.l1	-0.38576	0.21567	-1.789	0.07578	.
theft.l1	0.82717	0.36057	2.294	0.02324	*
asslt.l1	1.38864	0.44423	3.126	0.00215	**
unemp.l1	-29.18826	9.47567	-3.080	0.00248	**
crime.l2	0.73163	0.26293	2.783	0.00612	**
theft.l2	-0.94195	0.46919	-2.008	0.04657	*
asslt.l2	-0.35793	0.52074	-0.687	0.49298	
unemp.l2	26.55050	13.19036	2.013	0.04601	*
crime.l3	-0.41493	0.25206	-1.646	0.10192	
theft.l3	0.80203	0.45525	1.762	0.08025	.
asslt.l3	0.44222	0.51001	0.867	0.38735	
unemp.l3	7.35773	13.46116	0.547	0.58551	
crime.l4	0.23573	0.25275	0.933	0.35256	
theft.l4	-0.40673	0.46211	-0.880	0.38026	
asslt.l4	-0.32314	0.51402	-0.629	0.53059	
unemp.l4	-1.70796	13.54555	-0.126	0.89984	
crime.l5	0.45386	0.25118	1.807	0.07288	.
theft.l5	-0.80337	0.47445	-1.693	0.09258	.
asslt.l5	-0.72521	0.50609	-1.433	0.15405	
unemp.l5	-16.46802	13.61397	-1.210	0.22841	
crime.l6	-0.14666	0.25681	-0.571	0.56883	
theft.l6	0.39111	0.48414	0.808	0.42053	
asslt.l6	0.20877	0.51382	0.406	0.68512	
unemp.l6	18.85090	13.76125	1.370	0.17288	
crime.l7	-0.28182	0.25849	-1.090	0.27744	
theft.l7	0.74764	0.49133	1.522	0.13030	
asslt.l7	0.39371	0.51377	0.766	0.44475	
unemp.l7	-27.95740	13.67307	-2.045	0.04272	*
crime.l8	-0.10849	0.25649	-0.423	0.67293	
theft.l8	0.21613	0.49160	0.440	0.66086	
asslt.l8	-0.18095	0.51070	-0.354	0.72362	
unemp.l8	7.13636	13.80364	0.517	0.60596	
crime.l9	-0.06921	0.25046	-0.276	0.78270	
theft.l9	0.18355	0.48352	0.380	0.70479	

Unemployment Rate

	Estimate	Std. Error	t value	Pr(>|t|)	
crime.l1	-2.950e-03	1.877e-03	-1.572	0.1183	
theft.l1	6.639e-03	3.138e-03	2.116	0.0361	*
asslt.l1	5.648e-03	3.867e-03	1.461	0.1463	
unemp.l1	9.234e-01	8.248e-02	11.196	<2e-16	***
crime.l2	-1.980e-03	2.289e-03	-0.865	0.3885	
theft.l2	3.368e-03	4.084e-03	0.825	0.4109	
asslt.l2	8.265e-04	4.533e-03	0.182	0.8556	
unemp.l2	1.127e-01	1.148e-01	0.981	0.3281	
crime.l3	3.389e-03	2.194e-03	1.545	0.1246	
theft.l3	-5.271e-03	3.963e-03	-1.330	0.1855	
asslt.l3	-7.361e-03	4.439e-03	-1.658	0.0995	.
unemp.l3	-3.407e-02	1.172e-01	-0.291	0.7717	
crime.l4	7.944e-04	2.200e-03	0.361	0.7186	
theft.l4	-5.395e-03	4.022e-03	-1.341	0.1820	
asslt.l4	-3.891e-03	4.474e-03	-0.870	0.3859	
unemp.l4	-5.595e-02	1.179e-01	-0.475	0.6359	
crime.l5	1.001e-03	2.186e-03	0.458	0.6478	
theft.l5	-9.315e-04	4.130e-03	-0.226	0.8219	
asslt.l5	1.859e-04	4.405e-03	0.042	0.9664	
unemp.l5	1.345e-01	1.185e-01	1.135	0.2583	
crime.l6	-3.124e-03	2.235e-03	-1.398	0.1644	
theft.l6	1.061e-02	4.214e-03	2.518	0.0129	*
asslt.l6	2.821e-03	4.472e-03	0.631	0.5291	
unemp.l6	-9.975e-02	1.198e-01	-0.833	0.4063	
crime.l7	-4.258e-04	2.250e-03	-0.189	0.8502	
theft.l7	-1.392e-05	4.277e-03	-0.003	0.9974	
asslt.l7	2.135e-03	4.472e-03	0.477	0.6338	
unemp.l7	1.353e-01	1.190e-01	1.137	0.2575	
crime.l8	7.458e-04	2.233e-03	0.334	0.7388	
theft.l8	-5.191e-03	4.279e-03	-1.213	0.2271	
asslt.l8	1.563e-03	4.445e-03	0.352	0.7257	
unemp.l8	-1.848e-01	1.201e-01	-1.538	0.1262	
crime.l9	-8.368e-04	2.180e-03	-0.384	0.7017	
theft.l9	2.041e-03	4.209e-03	0.485	0.6285	

To account for these superfluous parameters, we run the MTS function refVAR, which refines the fitted VAR model by removing the parameters that fall outside of a specified threshold, which we set to 1.96. Thus we remove the extraneous parameters to get the following models:

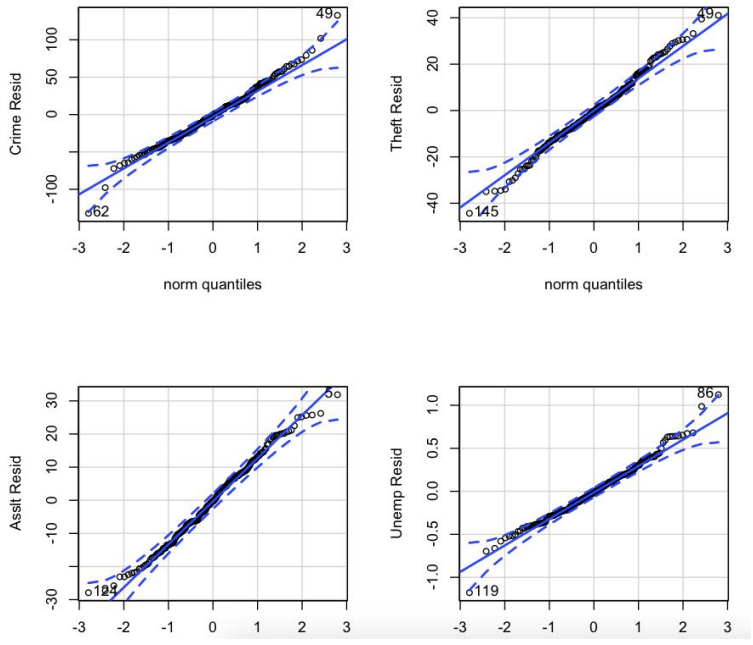
$$\widehat{CR} = 0.27CR_{t-2} - 0.34CR_{t-8} + 0.14CR_{t-10} + 0.16CR_{t-11} + 0.54CR_{t-12} + 0.85TR_{t-8} - 0.90TR_{t-11} + 1.05AR_{t-1} - 21.2UR_{t-7} + 31.8UR_{t-9} - 12.83UR_{t-12}$$

$$\widehat{TR} = -0.46CR_{t-1} - 0.19CR_{t-2} - 0.04CR_{t-6} - .12CR_{t-7} + 0.18CR_{t-11} + 0.18CR_{t-12} + 1.05TR_{t-1} - 0.26TR_{t-2} + .38TR_{t-7} - 0.28TR_{t-11} + 0.79AR_{t-1} - 0.33AR_{t-11} - 10.2UR_{t-7} - 8.17UR_{t-1} + 9.5UR_{t-2} + 16.0UR_{t-9} - 7.58UR_{t-12}$$

$$\widehat{AR} = -0.18CR_{t-1} + 0.14CR_{t-2} - 0.07CR_{t-3} - 0.05CR_{t-9} + 0.13CR_{t-12} + 0.20TR_{t-9} - 0.20TR_{t-12} + 0.90AR_{t-1} + 0.23AR_{t-10} - 11.32UR_{t-1} + 10.6UR_{t-2}$$

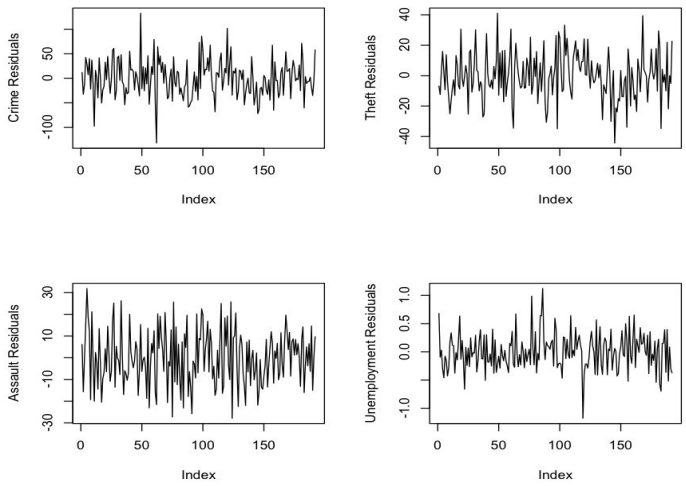
$$\widehat{UR} = 0.001CR_{t-5} - 0.002CR_{t-6} + 0.003TR_{t-1} - 0.007TR_{t-4} + 0.01TR_{t-6} - 0.003AR_{t-11} + 0.98UR_{t-1} + 0.12UR_{t-5} - 0.16UR_{t-8}$$

Residual Analysis

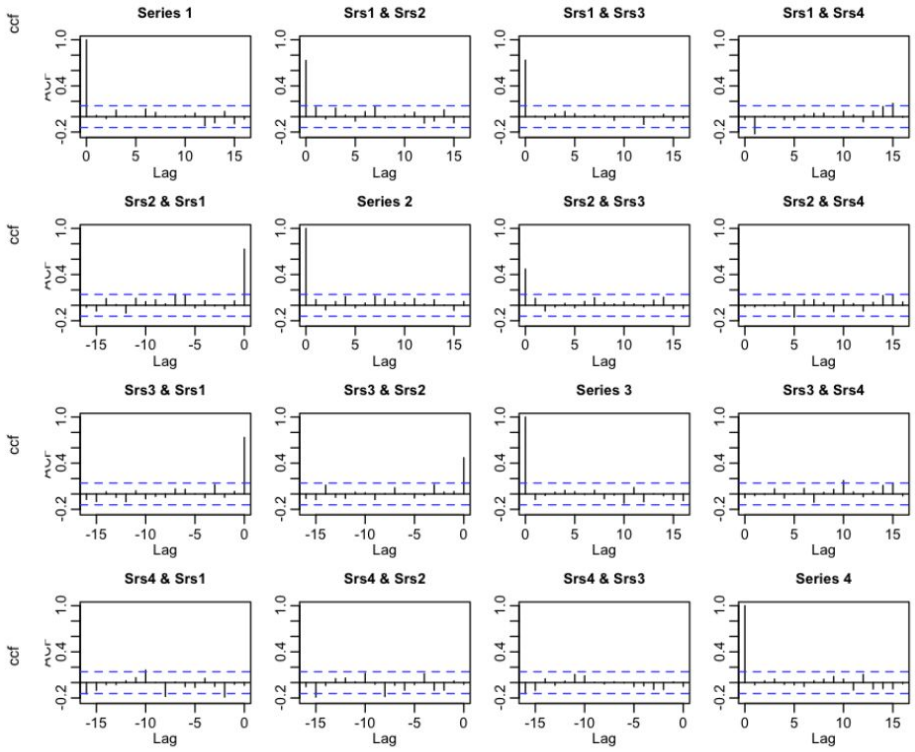


Next, we analyze the residuals of the final, refined VAR(12) model. The above figure shows that the residuals for each of the sub-models are sufficiently normal, though there are a few deviations outside of 95% confidence intervals, particularly for the Unemployment Rate residuals. The next plot

shows how these residuals behave as expected, i.e. like white noise residuals, with a zero mean and no clear trends in the variance.



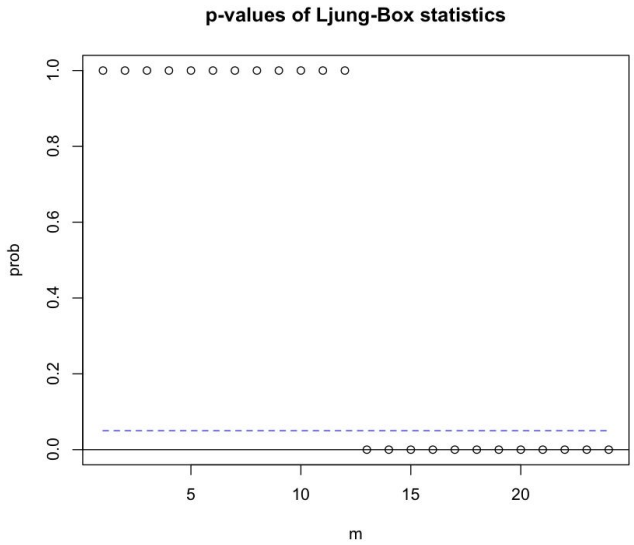
Next, we analyze the autocorrelation and cross-correlation functions of each set of residuals. Here, the series follows the same order as before, with Crime Rate being the first, followed by Theft, Assault, and Unemployment Rates. While most of the ACF and CCFs stay within the 95% confidence intervals at all lags, we see that this does not hold true for the CCF of the residuals between series 1 & 4, 2 & 4, and 3 & 4. That is, estimations between Assault Rate and the other 3 variables are not quite up to par. Indeed, we will see further evidence in the next section (Predictions).



In addition, this table of selection criteria depicts how the final, refined model is an improvement over both the VAR(3) and the unrefined VAR(12) models:

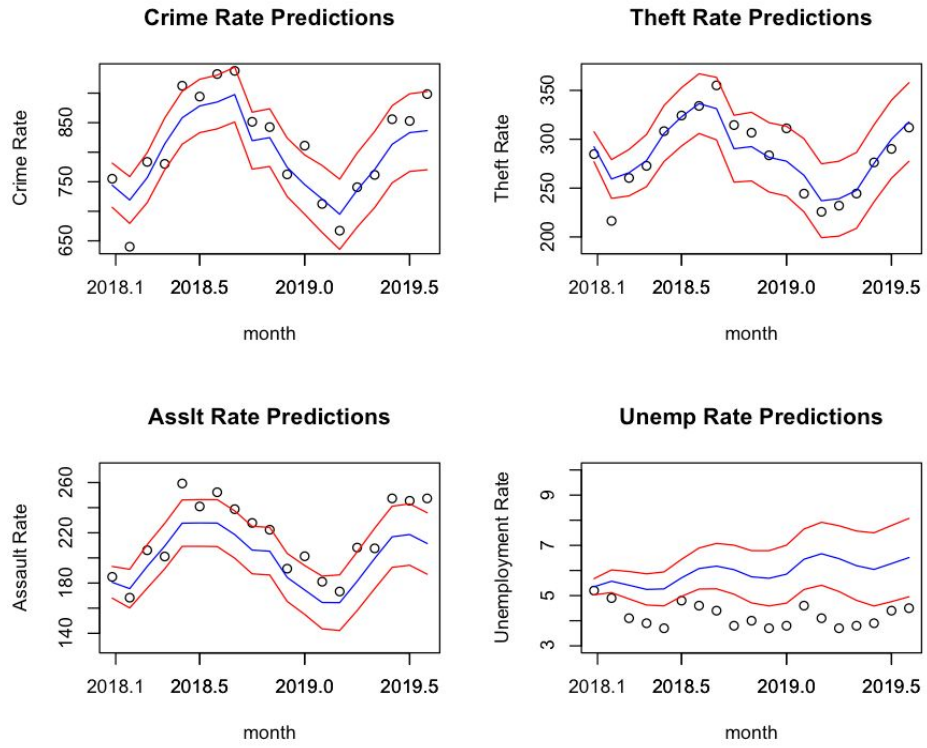
	VAR(3)	VAR(12)	Refined VAR(12)
AIC	16.04	14.24	14.40
BIC	16.82	17.36	15.18
HQ	16.36	15.50	14.72

The last plot to investigate is that of the Ljung-Box statistics, seen below. It seems the residuals are sufficiently uncorrelated for lags 1-12 but are very much correlated for any of the subsequent lag values. This indicates that there is a clear 12-month seasonality in the data that was not taken into account in our VAR analysis. This will be further analyzed and discussed in the section titled Two-Stage Model and sVARMA Model.



Predictions

The final step in our VAR analysis is to predict the final 19-month values for each variable, as well as their corresponding 95% confidence ranges, and compare these predictions to their real-world realizations. In the four plots below, the blue lines indicate these predictions as calculated using the refined VAR(12) model, with the red lines denoting their upper and lower confidence limits. The black circles are the real crime and unemployment rates as reported by the City of Chicago for the months of January 2018 to July 2019.



The Crime Rate and Theft Rate models seem to accurately predict the expected values, while the Assault Rate model becomes a bit questionable since we notice how many of the values start to drift above the upper prediction limit. Unfortunately, the model seems unable to accurately predict future unemployment rates. This is potentially due to the seasonality issues mentioned above, as well as our omission of a linear trend parameter in our VAR model. Our efforts to correct these issues are outlined in the following section.

Multivariate Modeling: Two-Stage Model and sVARMA Model

This system of time-series shows a strong seasonal pattern, the previous VAR model does not remove seasonality, as a result, the prediction falls short of precision and the residuals are not random as reflected in the Ljung-Box statistics.

Our group proposes two ways to address this issue: a two-stage model and an sVARMA (Seasonal Vector Autoregressive Moving Average) model. The two-stage model applies the previous SARIMA models to the training data. The residuals are then used to feed a VAR model. The prediction from the VAR model is used to add or subtract the SARIMA model prediction in the test set. The intuition behind this method is to use SARIMA model to remove seasonality pattern in the first stage and the second stage uses VAR to do multivariate modeling. This combined method should give a more precise prediction than using VAR alone.

The second method is the sVARMA model, which is a VARMA model with seasonal adjustment. The k-dimensional time series z_t is a VARMA(p,q) process if

$$\phi(B)z_t = \phi_0 + \theta(B)\alpha_t$$

Where ϕ_0 is a constant vector, $\phi(B) = I_k - \sum_{i=1}^{p} \phi_i B^i$ and $\theta(B) = I_k - \sum_{i=1}^{q} \theta_i B^i$ are two matrix polynomials and α_t is a sequence of iid random vectors with mean zero and positive-definite covariance matrix α_t.

The general form of a sVARMA model is:

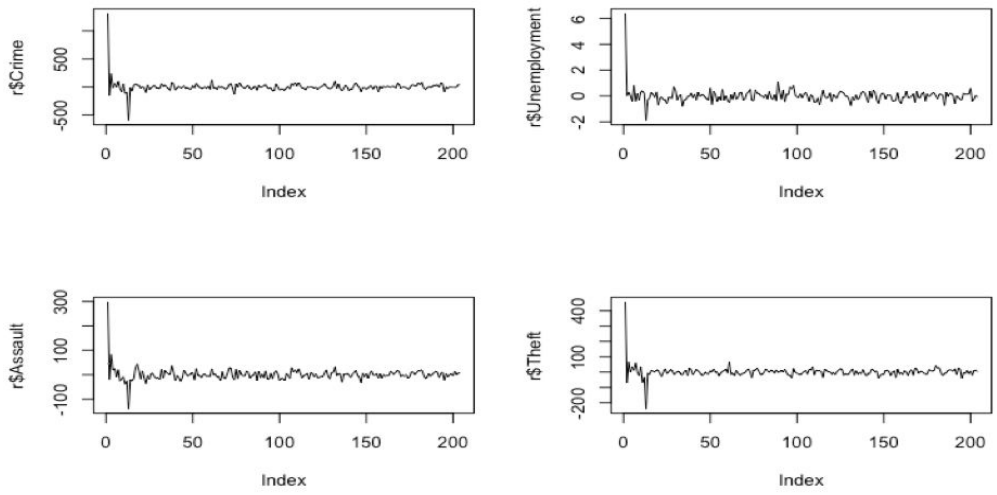
$$(1-B)(1-B^s)z_t = \theta(B)(\Theta B)\alpha_t$$

Where both θ(B) and Θ(B) are matrix polynomials of order q and Q, respectively, with q < s. s > 1 denotes the number of seasons within a year, and $\{\alpha_t\}$ is a sequence of iid random vectors with mean zero and positive-definite covariance matrix ∑α. $(1 - B^s)$ is referred to as the seasonal difference, and (1 – B) the regular difference.

The modeling process of the two-stage model and sVARMA model is described in the next section.

Two-Stage Model

The first stage of this model is to develop a customized SARIMA model for each time series as described in the earlier part of this report. These four SARIMA models are then applied to the training data to obtain four sets of residuals. These training residuals are plotted below. They all have a similar pattern, the first 13 observations are highly unstable. Therefore, they are removed from the dataset.

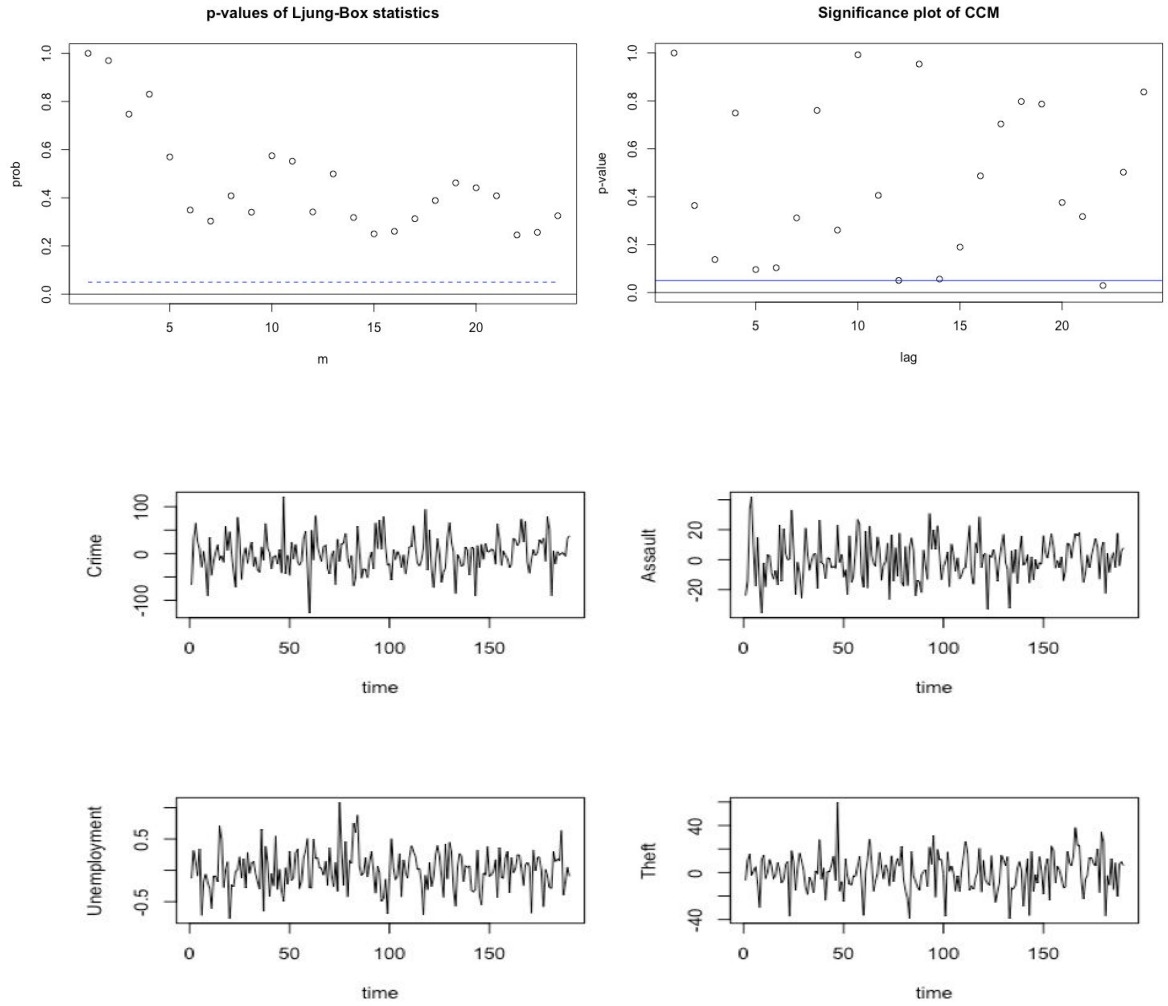


These SARIMA residuals are then fed to a VAR(1) model. The original VAR(1) model has 16 parameters which are not all significant. This model is then simplified to remove any parameters

with t-values under 1 simultaneously. As a result, only 12 parameters remain, the AIC, BIC, and HQ values are improved. The result is summarized in the following table:

	No. of Parameters	AIC	BIC	HQ
VAR(1)	16	13.90	14.17	14.01
Refined VAR(1)	12	13.86	14.05	13.94

Residual diagnosis is then conducted on the refined VAR(1) model. In the CCM plot, most points pass the threshold indicating there is no significant cross-correlation detected. All points pass the Ljung-Box statistics indicating the residuals are random and independent. The residual plots show the residuals are similar to white noise.



The refined VAR(1) model is:

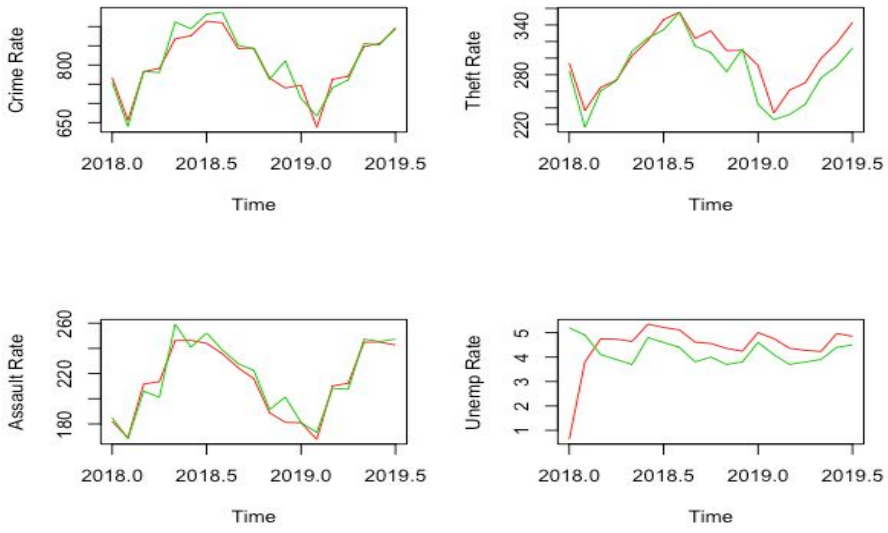
$$\hat{CR}_t = -1.69CR_{t-1} - 8.87UR_{t-1} + 0.51AR_{t-1} + 0.34TR_{t-1}$$

$$\hat{UR}_t = -0.03 + 0.0015TR_{t-1}$$

$$\hat{AR}_t = -0.046CR_{t-1} - 3.74UR_{t-1} + 0.181AR_{t-1}$$

$$\hat{TR}_t = -0.245CR_{t-1} + 0.48AR_{t-1} + 0.31TR_{t-1}$$

This model is used to make a 19-steps prediction. The result is then added or subtracted from the SARIMA prediction. The result of this two-stage model is shown below. The red line indicates the predicted values whereas the green line indicates the actual values. Generally speaking, the prediction is close to the actual values, except for the first few predictions from the unemployment rate. The RMSE of each time series is listed in the following table.



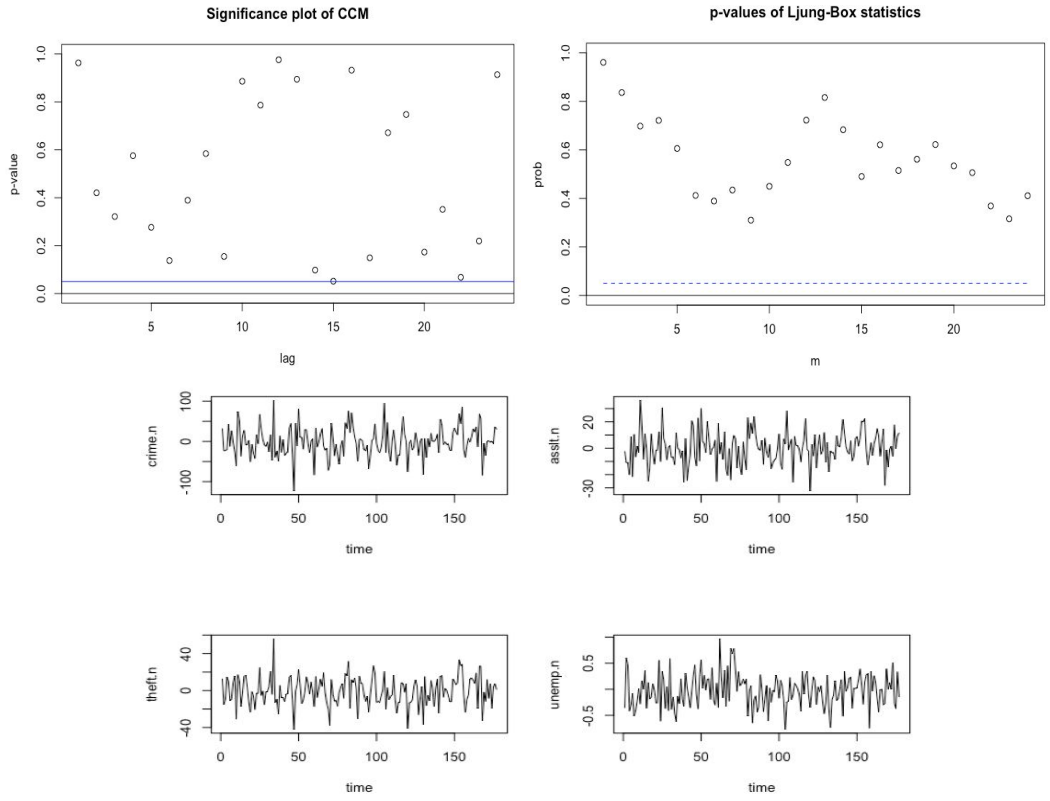
	RMSE
Crime Rate	24.98
Unemployment Rate	1.23
Assault Rate	31.92
Theft Rate	20.80

sVARMA Model

The next method we use to address the seasonality issue is to employ the sVARMA model. sVARMA model has the same notation as SARIMA model, the orders are specified as sVARMA(p,d,q) x (P, D,Q)s. A number of low order models are tried and sVARMA(0,1,2) x (0,1,1)12 model is chosen because it has the lowest AIC and BIC values. This model has 48 parameters and most of them are not significant. It is then simplified, removing all parameters with t-values under 1.2 simultaneously. As a result, 25 parameters remain and the AIC and BIC values slightly improved.

	No. of Parameters	AIC	BIC
sVARMA(0,1,2) x (0,1,1)12	48	13.75	14.56
Refined sVARMA	25	13.53	13.96

Residual analysis of the refined sVARMA model does not show any major issue. Most points on the CCM plot past the threshold level and all the p-values in Ljung-Box statistics are not significant. The residual plots do not show any specific trend.



The refined model has the following matrices:

MA(1)-matrix

	[CR]	[AR]	[TR]	[UR]
[CR]	0.60	0	0	7.35
[AR]	0.133	0.36	-0.34	0
[TR]	0	0	0.64	3.03
[UR]	0	-0.002	0	0

MA(2)-matrix

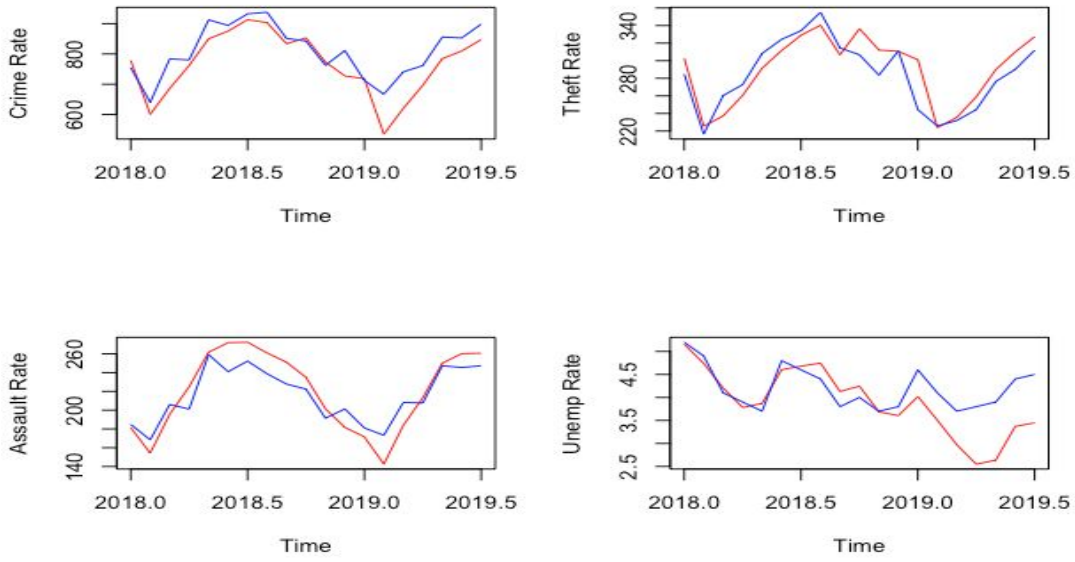
	[CR]	[AR]	[TR]	[UR]
[CR]	0.26	-0.52	0	0
[AR]	0	0	0	0
[TR]	0	-0.11	0.19	0
[UR]	0.004	-0.01	0	-0.17

Seasonal MA coefficient matrix

MA(12)-matrix

	[CR]	[AR]	[TR]	[UR]
[CR]	0.56	0.33	0	14.66
[AR]	0	0.71	-0.13	4.11
[TR]	0.07	0	0.61	4.72
[UR]	0	0	0	0.73

Finally, the result of the refined sVARMA model is shown below. The red line shows the predicted values whereas the blue line shows the actual values. The predicted values are quite close to the actual values except for the end of unemployment rate. The gap is increasing as prediction moves further from the origin.



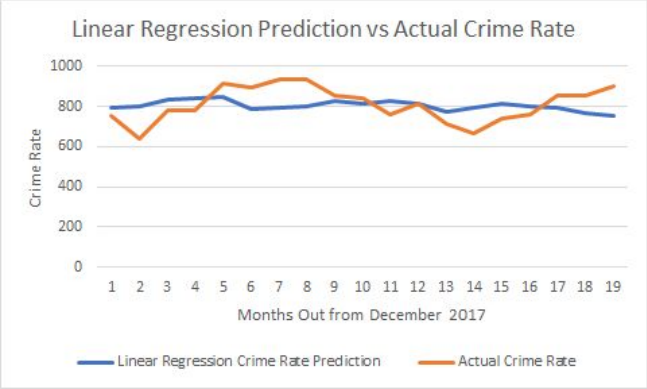
The RMSE of sVARMA model is shown in the following table.

	Crime Rate	Theft Rate	Assault Rate	Unemployment Rate
RMSE	61.22	20.25	17.79	0.61

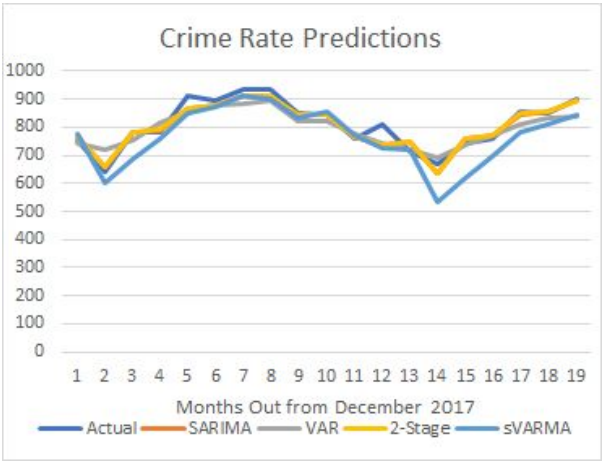
Conclusion

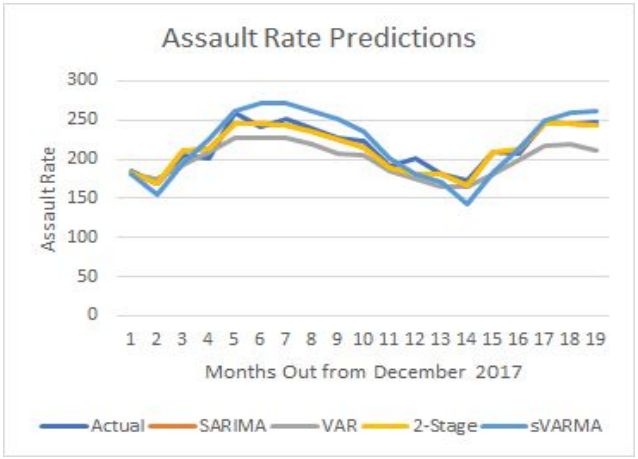
Model Comparison

Each model has a different method of tackling the data, and so produces a different prediction for the nineteen months following the end of 2017. While some of the time series models are better than others, they all remain a better fit to the actual data than the Linear Regression, an example of which for crime is seen below, with an RMSE of 88.85.

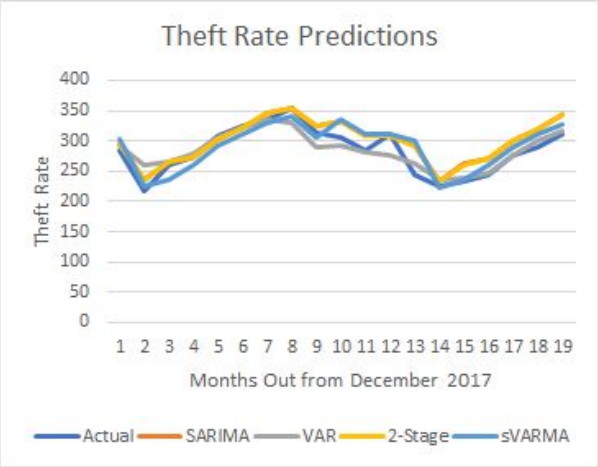


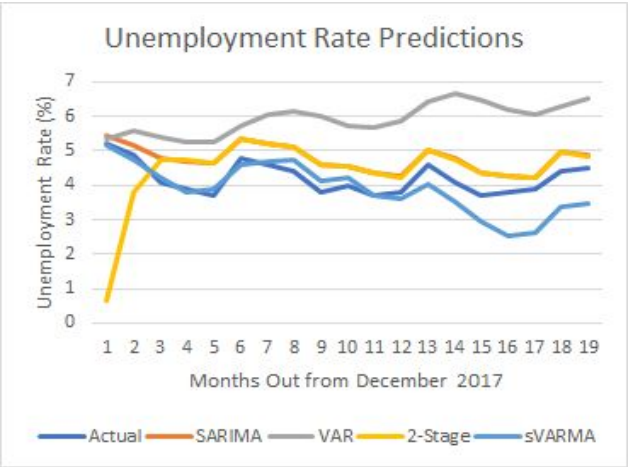
Looking at this graph, we can tell that simply regressing the crime rate vs the unemployment rate misses nuances in the data that analyzing with respect to time series does not. In this instance, the most noticeable difference between the predicted regression values and the actual data is there is no consideration for seasonality. Indeed, the natural undulation of the data over time is missed.





Next, we compare the four time series models, two of which are on the top of the next page. That is, the SARIMA, the basic VAR, the Two-Stage, and the sVARMA. Just looking at the plots, the time series data fit and undulate with the actual data in a way the Linear Regression does not. However, to confirm this improved fit, simply compare the RMSE of the Linear Regression for Crime, 88.85, to the highest RMSE for the time series models, the VAR with an RMSE of 38.42. A table of the RMSE for the predictive models is viewable on the next page, page 32.





For the most part, all of the trends for the predictions are similar, with tighter predictions at the beginning of the predictive series, with further deviation as we go further out from the time of prediction, which is expected. The exception to this is the Two-Stage, with its one almost uncharacteristic first prediction point.

Comparing the time series models themselves, especially comparing the RMSE, the clear winner appears to be the SARIMA, with the Two-Stage and sVARMA improving upon the basic VAR in different ways. The SARIMA was the best in Assault and Unemployment predictions, while virtually tied with the Two-Stage for Crime. The SARIMA, Two-Stage, and sVARMA are also virtually tied for Theft.

Theft is also, oddly, where the basic VAR outperforms all the other models. However, due to its poor performance on all the other variables, especially its overestimation for Unemployment, VAR is hardly the most optimal model.

RMSE Comparison

	SARIMA	VAR	2-Stage	sVARMA
Crime:	25.13	38.42	24.98	61.22
Theft:	20.83	16.44	20.81	20.26
Assault:	7.35	20.63	31.92	17.79
Unemp:	0.60	3.39	1.23	0.61

Indeed, the Two-Stage and sVARMA are both successful improvements upon the basic VAR. However, both the Two-Stage and the sVARMA have their issues. For the sVARMA, this is Crime, while, for the Two-Stage, it is Unemployment. However, while the sVARMA appears to be off due to overestimating the contributions of seasonality, for the Two-Stage, it appears to be one poor prediction mistake.

Were it not for the odd first prediction for Unemployment, the Two-Stage would likely have comparable RMSE to both the SARIMA and the sVARMA and be the clear better improvement of the two over the basic VAR. Despite investigating reasons for this, no current explanation has been

found. This leaves the SARIMA the current best model. However, one can certainly conclude that, given the complexity of this data, that it is a boon to have multiple models from which to analyze the data.

Problems and Future Thoughts

Though the best model has been found in the SARIMA, there are still some issues during analysis that remain unanswered. As mentioned previously, there were different issues with the Two-Stage and the sVARMA, and the Two-Stage, in general, had a single odd prediction that remains unexplained.

Additionally, there are a variety of issues that still remain in the data. For instance, we are not sure if high correlation might have affected the results, and the Granger causality established that there are different connections between the aggregate and subclasses of crime. For instance, though crime lags unemployment in general, assault does not.

Going forward, there are various possible improvements that could be done and alternate models that could be pursued. First, that odd first prediction in the Two-Stage can be re-explored. Secondly, though it was desired to use Spectral Analysis, time was not conducive to allowing its use. Thus, in the future, Spectral analysis is likely one of the new methods to be pursued for prediction. Finally, taking to heart the concept of improving the various forms of VAR, more variables could always improve the model. Variables such as Homelessness, Inflation, and various socioeconomic indicators could provide insights that simply are not present within the current dataset. Like the Time Series course in general, this exercise has taught us that considering new and novel aspects of a problem can provide fruitful and enlightening results.

References

B.V., Elsevier (2020). “Granger Causality Test.” ScienceDirect. https://www.sciencedirect.com/topics/social-sciences/granger-causality-test

Janko Z, Popli G (2015). “Examining the link between crime and unemployment: a time-series analysis for Canada.” Applied Economics Vol.47, No.37, 4007-4019

Lauritsen, Janet L., Cork, Daniel L., et al (2016). “MODERNIZING CRIME STATISTICS.” National Academies Press. https://www.nap.edu/read/23492/chapter/5#89

Novara, Marisa and Khare, Amy. (2017) “Two Extremes of Residential Segregation: Chicago’s Separate Worlds & Policy Strategies for Integration.” Metropolitan Planning Council. https://www.jchs.harvard.edu/sites/default/files/a_shared_future_two_extremes_residential_segregation.pdf

Tsay R.S. (2014). Multivariate Time Series Analysis With R and Financial Applications, Wiley.

Zhu, Wei (2018). “VAR MODELS & GRANGER CAUSALITY.” STAT 497 LECTURE NOTES http://www.ams.sunysb.edu/~zhu/ams586/VAR_Lecture3.pdf

Shumway, Robert H. *Time Series Analysis and Its Applications: With r Examples*. Springer Science+Business Media, 2017.

Appendix

Descriptive Analysis

```
crimes = read.csv('/Users/rae/Desktop/sorted_crime.csv')
crimeRate = crimes[,2]
theftRate = crimes[,3]
asslt Rate = crimes[,4]
unempRate = crimes[,5]
CR = crimeRate
dCR = diff(crimeRate)
ddCR12 = diff(dCR, 12)
TR = theftRate
dTR = diff(theftRate)
ddTR12 = diff(dTR, 12)
AR = asslt Rate
dAR = diff(AR)
ddAR12 = diff(dAR, 12)
UR = unempRate
dUR = diff(unempRate)
ddUR12 = diff(dUR, 12)
```

Linear Regression

```
#univariate time series#############################################################
crime <- ts(data$CrimeRate, start = c(2001, 1), frequency = 12)
theft <- ts(data$TheftRate, start = c(2001, 1), frequency = 12)
assault <- ts(data$AssaultRate, start = c(2001, 1), frequency = 12)
unemployment <- ts(data$UnemRate, start = c(2001, 1), frequency = 12)

#multivariate time series###########################################################
crime.mts <- cbind(crime, theft, assault, unemployment, trend = time(crime),
                   unemployment2 = -1*(unemployment^2))

#crime rate#########################################################################
train <- window(crime.mts, end = c(2017, 12))
test <- window(crime.mts, start = 2018)

#tentative model 3
unemployment2 <- -1*(unemployment^2)
fit <- lm(crime ~ trend + unemployment + unemployment2, data = crime.mts)
summary(fit)
summary(aov(lm(crime ~ cbind(trend, unemployment, unemployment2), data =
crime.mts)))

num <- length(crime)
AIC(fit)/num - log(2*pi)
BIC(fit)/num - log(2*pi)

#fit and prediction
fit <- lm(crime ~ trend + unemployment + unemployment2, data = train)
pred <- predict(fit, test)
```

```
plot(test[,1],
     main = "Crime Rate Prediction (h = 19)",
     ylab = "crime rate")
lines(as.numeric(time(test)), pred, col = 'red')

forecast::accuracy(pred, test[,1])

plot(fit)
acf(resid(fit), lag.max = 40)
```

SARIMA Models

UNEMPLOYMENT RATE:

```
#prediction
train <- window(unemployment, end = c(2017, 12))
test <- window(unemployment, start = 2018)

unemployment.fit <- sarima(train, 3, 1, 3, 0, 1, 1, 12)

par(mfrow = c(2, 1), oma = c(0, 0, 2, 0))
plot(unemployment)
unemployment.fore <- sarima.for(train, 19, 3, 1, 3, 0, 1, 1, 12, plot.all =
TRUE)
mtext("Unemployment Rate Predictions", side = 3, outer = TRUE, cex = 1.5)
forecast::accuracy(unemployment.fore$pred, test)
```

CRIME RATE:

```
#partitioning data
train <- window(crime, end = c(2017, 12))
test <- window(crime, start = 2018)

#fitting the model to the training data
crime.fit <- sarima(train, 0, 1, 1, 0, 1, 1, 12)

#forecasting crime data using testing data
crime.fore <- sarima.for(train, 19, 0, 1, 1, 0, 1, 1, 12, plot.all = TRUE)
```

THEFT RATE:

```
#prediction performance
accuracy(crime.fore$pred, test)
train <- window(theft, end = c(2017, 12))
test <- window(theft, start = 2018)

theft.fit <- sarima(train, 0, 1, 1, 0, 1, 1, 12)
theft.fore <- sarima.for(train, 19, 0, 1, 1, 0, 1, 1, 12, plot.all = TRUE)

accuracy(theft.fore$pred, test)
```

ASSAULT RATE:

```
#assault rate#################################################################
#SARIMA: (0, 1, 1) x (0, 1, 1)12

sarima(assault, 0, 1, 1, 0, 1, 1, 12)

#prediction
train <- window(assault, end = c(2017, 12))
test <- window(assault, start = 2018)

assault.fit <- sarima(train, 0, 1, 1, 0, 1, 1, 12)

par(mfrow = c(2, 1), oma = c(0, 0, 2, 0))
plot(assault)
assault.fore <- sarima.for(train, 19, 0, 1, 1, 0, 1, 1, 12, plot.all = TRUE)
mtext("Assault Rate Predictions", side = 3, outer = TRUE, cex = 1.5)

forecast::accuracy(assault.fore$pred, test)
```

Granger Causality Test

```
library(lmtest)
grangertest(CrimeRateD2 ~ UnemRateD2 , order = 3)  #P = 0.2722
grangertest(CrimeRateD2 ~ UnemRateD2 , order = 12) #P = 0.015
grangertest(UnemRateD2 ~ CrimeRateD2 , order = 3)   #P = 0.008
grangertest(UnemRateD2 ~ CrimeRateD2 , order = 12)   #P = 0.00177 **

grangertest(TheftRateD2 ~ UnemRateD2 , order = 3) #P=0.2995
grangertest(TheftRateD2 ~ UnemRateD2 , order = 12) #P=0.01754
grangertest(UnemRateD2 ~ TheftRateD2 , order = 3) #P=0.2252
grangertest(UnemRateD2 ~ TheftRateD2 , order = 12) #P=0.0017

grangertest(AssaultRateD2 ~ UnemRateD2 , order = 3) #P=0.394
grangertest(AssaultRateD2 ~ UnemRateD2 , order = 12) #P=0.04
grangertest(UnemRateD2 ~ AssaultRateD2 , order = 3) #P=0.0065
grangertest(UnemRateD2 ~ AssaultRateD2 , order = 12) #P=0.0702

grangertest(AssaultRateD2 ~ CrimeRateD2 , order = 3) #P=0.687
grangertest(AssaultRateD2 ~ CrimeRateD2 , order = 12) #P=0.4237
grangertest(CrimeRateD2 ~ AssaultRateD2 , order = 3) #P=0.7554
grangertest(CrimeRateD2 ~ AssaultRateD2 , order = 12) #P=0.94

grangertest(TheftRateD2 ~ CrimeRateD2 , order = 3) #P=0.1529
grangertest(TheftRateD2 ~ CrimeRateD2 , order = 12) #P=0.77
grangertest(CrimeRateD2 ~ TheftRateD2 , order = 3) #P=0.2794
grangertest(CrimeRateD2 ~ TheftRateD2 , order = 12) #P=0.4834

grangertest(TheftRateD2 ~ AssaultRateD2 , order = 3) #P=0.534
grangertest(TheftRateD2 ~ AssaultRateD2 , order = 12) #P=0.869
grangertest(AssaultRateD2 ~ TheftRateD2 , order = 3) #P=0.654
grangertest(AssaultRateD2 ~ TheftRateD2 , order = 12) #P=0.7964
```

VAR Model:

```
all = cbind(crime, theft, asslt, unemp)
library(MTS)
VARorder(all, maxp=12)
fit1 = VAR(all, p=3)
fit2 = VAR(all, p=12)
library(car)
par(mfrow=c(2,2))
qqPlot(fit1$residuals[,1], ylab='Crime Resid')
qqPlot(fit1$residuals[,2], ylab='Theft Resid')
qqPlot(fit1$residuals[,3], ylab='Asslt Resid')
qqPlot(fit1$residuals[,4], ylab='Unemp Resid')
par(mfrow=c(2,2))
qqPlot(fit2$residuals[,1], ylab='Crime Resid')
qqPlot(fit2$residuals[,2], ylab='Theft Resid')
qqPlot(fit2$residuals[,3], ylab='Asslt Resid')
qqPlot(fit2$residuals[,4], ylab='Unemp Resid')
bet1 = refVAR(fit1, thres=1.96)
bet2 = refVAR(fit2, thres=1.96)
MTSdiag(bet2, adj=188)
pred1 = VARpred(bet1, h=19)
pred2 = VARpred(bet2, h=19)
ind = 1:length(ctest)
month = ind/12+2018
plot(month,ctest, main='Crime Rate Predictions', ylab='Crime Rate')
lines(month, pred2$pred[,1], col='blue')
up1 = pred2$pred[,1]+pred2$se.err[,1]
lo1 = pred2$pred[,1]-pred2$se.err[,1]
lines(month, up1, col='red')
lines(month, lo1, col='red')
axis(1, at = c(2018.1, 2018.5, 2019, 2019.5))
plot(month,ttest, main='Theft Rate Predictions', ylab='Theft
Rate',ylim=c(190,370))
lines(month, pred2$pred[,2], col='blue')
up2 = pred2$pred[,2]+pred2$se.err[,2]
lo2 = pred2$pred[,2]-pred2$se.err[,2]
lines(month, up2, col='red')
lines(month, lo2, col='red')
axis(1, at = c(2018.1, 2018.5, 2019, 2019.5))
plot(month,atest, main='Asslt Rate Predictions', ylab='Assault Rate',
ylim=c(130 ,270))
lines(month, pred2$pred[,3], col='blue')
up3 = pred2$pred[,3]+pred2$se.err[,3]
lo3 = pred2$pred[,3]-pred2$se.err[,3]
lines(month, up3, col='red')
lines(month, lo3, col='red')
axis(1, at = c(2018.1, 2018.5, 2019, 2019.5))
plot(month,utest, main='Unemp Rate Predictions', ylab='Unemployment Rate',
ylim=c(3, 10))
lines(month, pred2$pred[,4], col='blue')
up4 = pred2$pred[,4]+pred2$se.err[,4]
```

```
lo4 = pred2$pred[,4]-pred2$se.err[,4]
lines(month, up4, col='red')
lines(month, lo4, col='red')
axis(1, at = c(2018.1, 2018.5, 2019, 2019.5))
```

Two-Stage Model:

```
library(MTS)
Remove extreme values 1:13 rows
r1 = r[-c(1:13),c(2:5)] #remove extreme values and dates in the SARIMA
residuals

#Order Selection
m1 = VARorder(r1)
selected order: aic =  1
selected order: bic =  0
selected order: hq =  1

m2 = VAR(r1,1) #fitting a VAR(1) model
AIC =  13.90402
BIC =  14.17646
HQ  =  14.01437

m2a=refVAR(m2,thres=1) #remove simultaneously insignificant parameters,
threshold of t-test is 1, 12/16 remains
AIC =  13.86321
BIC =  14.05052
HQ  =  13.93908

MTSdiag(m2a) #residual diagnosis
VARpred(m2a,19) #Predict the next 19 lags
```

sVARMA Model:

```
library(MTS)
m14=sVARMA(D4,order=c(0,1,2),sorder=c(0,1,1),s=12)
m14a=refsVARMA(m14,thres=0.8)#refining model
m14b=refsVARMA(m14a,thres=1)
m14c=refsVARMA(m14b,thres=1.2)
MTSdiag(m14c) #Residual diagnosis
D5=sVARMApred(m14c, orig=205, h=19) #prediction
```