Forecasting Chicago Crime Rates: A Time Series Analysis

Brenda Morales

brendamorales@ucsb.edu

University of California, Santa Barbara

June 2023

Abstract

Over the last few years, the city of Chicago has seen an alarming general increase in crime rates, raising significant risks for public safety and community well-being. This research investigation applies time series analysis to forecast Chicago's crime rates.. The dataset has been collected through the Citizen Law Enforcement Analysis and Reporting (CLEAR) system of the Chicago Police Department. This study will focus solely on the “Public Peace Violations” crime rates to see if it deviates from the overall increasing crime rates. In order to forecast this crime rate, two forecasting methods (i.e. SARIMA and GARCH models) are implemented. The results of the study contribute to us understanding how time series analysis may be employed when predicting crime rates. The SARIMA and GARCH models serve as the basis in this research to provide insights into the potential trends of crime in the city of Chicago, which can help law enforcement departments and legislators make well-informed choices about crime prevention and resource allocation.

Keywords: *time series, crime rates, forecasting, SARIMA, GARCH, Chicago Police Department*

1. Introduction

The project aims to draw insightful patterns and trends from the data and to predict “Public Peace Violation” crime rates in Chicago using methods of time series analysis. The purpose is to develop two accurate forecasting models that can assist in understanding and predicting crime trends. The chosen dataset is obtained from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR) system, which provides accurate and comprehensive crime data that is updated regularly for analysis purposes. The topic of crime rate prediction is of significant importance for law enforcement agencies and policymakers. By accurately forecasting crime rates, proactive measures can be taken to prevent criminal activities and allocate resources effectively. Chicago, being a major city with complex crime dynamics, provides a rich data set for studying crime patterns and testing predictive models.

Previous studies using the Chicago Police Department's crime data have explored various aspects of crime patterns, including spatial analysis, hotspots identification, and crime correlations with socioeconomic factors. These studies have provided valuable insights into understanding crime dynamics in Chicago and have laid the groundwork for further analysis and prediction.

Our project employs time series analysis techniques, specifically the Box-Jenkins approach to find the best Seasonal Autoregressive Integrated Moving Average (SARIMA) model and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model. SARIMA captures the temporal patterns and seasonality in the data, while GARCH addresses volatility and heteroskedasticity, which are important factors to consider in our crime rate analysis.

Through the analysis of the data set, significant discoveries related to crime trends and patterns in Chicago can be uncovered. This includes identifying seasonal variations in crime rates, detecting long-term trends, or understanding the impact of certain events or interventions on crime levels.

2. Data

The objective of this project is to investigate and analyze a dataset containing reported criminal incidents in the City of Chicago between January 1, 2001, and May 1, 2023. The dataset has a total of 7,799,141 observations and includes several explanatory variables such `ID`, `case_number`, `block`, `ICUR`, `primary_type`, `arrest`, `domestic`, `district`, `ward`, and `location`. The CLEAR system is updated on a monthly basis, so our dataset contains even the most recent observations. However, to dial in on the `primary_type` variable for the focus of this project, we will primarily examine the subgroup of the `primary_type` variable regarding "public peace violation," or civil disturbances.

Choosing to focus on this specific form of criminal violation was inspired by the important events of 2020, which were marked by widespread protests across the United States. As the public from all around the nation fought for structural changes to the criminal justice system and addressed ongoing racial injustice and police brutality, the Black Lives Matter movement developed widespread support. The protests attempted to bring attention to the disproportionate violence and abuse of Black people by law enforcement. One interest is in examining if the reported incidences of public peace violations show patterns and trends reflecting the elevated levels of civil disturbances experienced during the peak of the protests in 2020 given that the dataset covers a period that includes these significant events.

The dataset used for this study was taken from the Citizen Law Enforcement Analysis and Reporting (CLEAR) system of the Chicago Police Department. Additional information about the Chicago Police Department and their data can be found on their website (https://home.chicagopolice.org/). The original dataset can be accessed with the following link, https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.

With the help of this data as well as an analysis of the cases of public peace violations, we hope to learn more about the connection between civil disturbances and the more general social movements to predict future rates of public peace violation. This analysis has the potential to expand our comprehension of how these instances have impacted reported criminal violations and to help us better understand how to address law enforcement strategies and social justice initiatives.

3. Methodology

3.1 SARIMA (p, d, q) x (P, D, Q) Model

The SARIMA model is one of the many types of forecasting models we have covered for time series analysis. SARIMA, which is an acronym for Seasonal Autoregressive Integrated Moving Average, is an expansion of the ARIMA model by adding seasonal components to account for seasonal variations in the data. Autoregressive (AR), differencing (I), and moving average (MA) are the three basic elements that make up the SARIMA model. Additionally, Seasonal parameters P, D, Q, and S are also included in the SARIMA model. Hence, the SARIMA(p, d, q) x (P, D, Q) model is a powerful tool for accurate forecasting and time series data analysis as it effectively captures both short-term dynamic and seasonal changes through the combination of both elements. This includes: model identification and model selection, parameter estimation to find model coefficients, and statistical model checking by testing whether the estimated model conforms to the specifications of a stationary univariate process.

To evaluate the crime time series dataset using a seasonal (S)ARIMA model, we will apply the Box-Jenkins method. We will first plot the original time series data to gain understanding and determine if there are any apparent patterns and apply data transformations. To handle any non-stationary or non-linear behavior. Then, to ensure the model parameters' accuracy and dependability, we will estimate potential values for MA and AR. The goodness of fit of the model will be assessed, and any residual patterns or anomalies will be found using diagnostic tests. The best SARIMA model will then be chosen in accordance with the outcomes of the diagnostic, the standards of evaluation, and the predictive ability.

3.2 GARCH Model

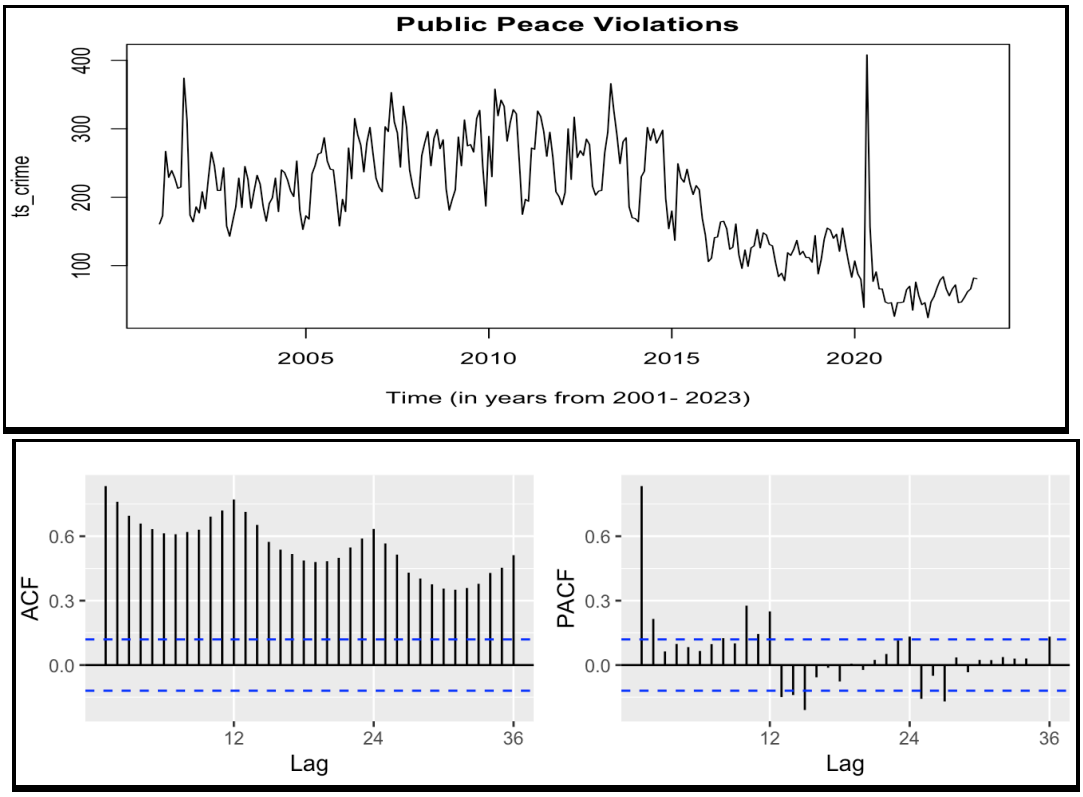
The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model is another common model for analyzing and forecasting volatility in time series data. Given the reliance on previous volatility and squared residuals, it represents the series' time-varying variance or volatility. The steps to compute a GARCH model typically include: specifying the model with the proper parameters, estimating the model to obtain the coefficients, evaluating the model's goodness of fit and diagnostic tests, predicting volatility based on the estimated model, and analyzing the results to gain insights into the time series' volatility behavior.

4. Results

4.1 Results from SARIMA (p, d, q) x (P, D, Q)

The first step of the Box-Jenkins method to find our SARIMA model begins by plotting the original time series data along with ACF and PACF plots. By doing so, it helps identify patterns and dependencies. The time series plot in **`Figure 1`** reveals a downward trend with apparent outliers which signifies that we cannot conclude the time series is stationary. There is also a sharp peak around 2020, as expected, but this is considered an outlier point so we will apply the `tsclean()` function to mutate the values accordingly (see **`Figure 1.1`** in Appendix). An Augmented Dickey Fuller Test was conducted to confirm that the time series is not stationary as our p-value of 0.2251 is larger than the 0.05 significance level, indicating we fail to reject the null hypothesis thus the time series is non-stationary and it has some time-dependent structure and does not exhibit constant variance over time. Additionally our ACF values lie outside the confidence interval and the PACF values are within this, so we may proceed with finding the ARIMA model.

`Figure 1: Original Time Series`



To make our crime data stationary, we will use differencing and log-differencing approaches which essentially subtracts each observation from its previous value, while log-differencing applies a logarithmic transformation before differencing. Prior to executing this task, it would be beneficial to investigate the decomposition of our time series, as seen in **`Figure 1.2`** (located in

the Appendix). It reveals the negative trend as the years increase as well as the seasonality. Next, the differencing and log-difference transformations remove trends, seasonality, stabilize variance, and improve modeling and forecasting. **Figure 2** reveals taking the logarithm of our cleaned crime data. From here, we can notice that it is a bit more stable with less variation, but it may be beneficial to transform it once more with the difference of our data. As a result, the bottom plot reveals our transformed stationary time series data.

Figure 2: Applying Transformations

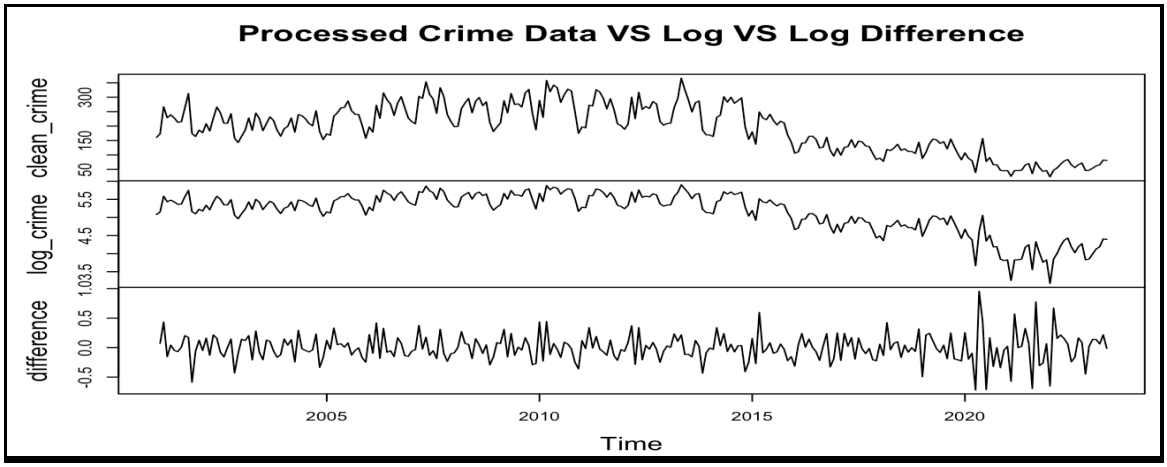


Figure 3: Logged Difference

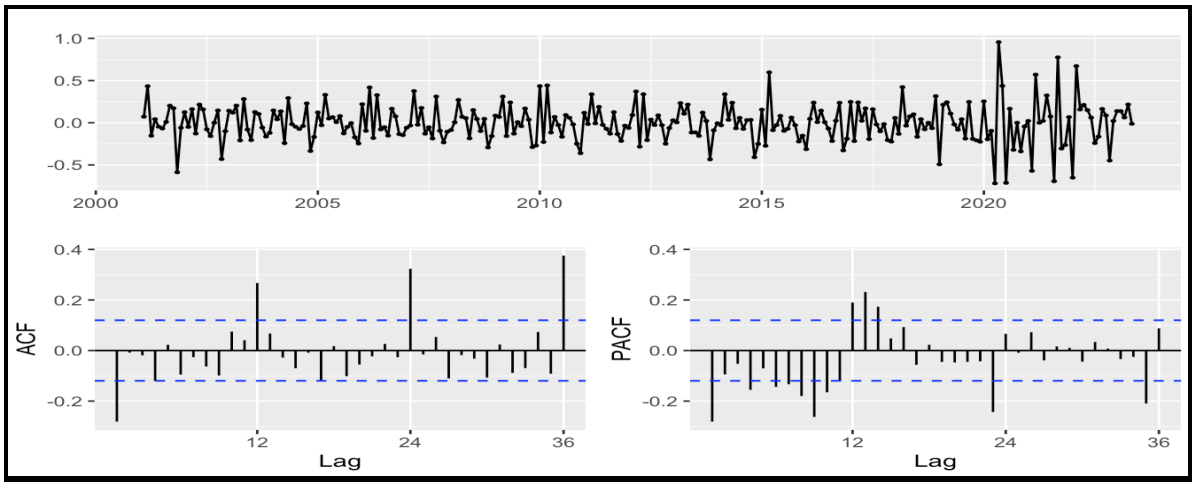
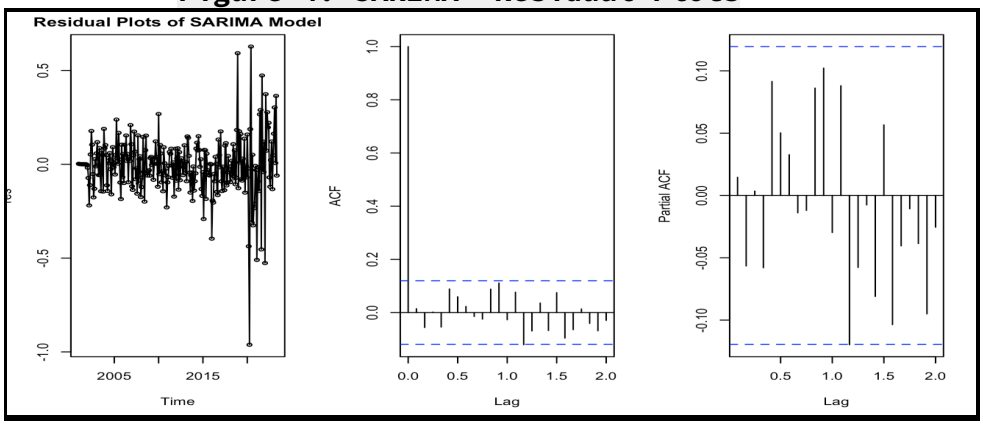


Figure 3 is a closer look at the logged difference from **Figure 2** along with ACF and PACF plots. These figures confirm that our model is stationary and we are not ready to move onto the next step: estimating parameters. Through a grid search (see code in the Appendix), we found the possible combinations of values for our SARIMA model parameters. Our search concluded the best fitting model parameters made up the SARIMA(0, 1, 1) x (2, 1, 2)[12] model for our values of (p, d, q) x (P, D, Q), respectively as that was the combination of parameters with the lowest AIC values. Additionally, this was confirmed using the `auto.arima()` function with the seasonal argument set to true (see **Figure 1.3)** in appendix.

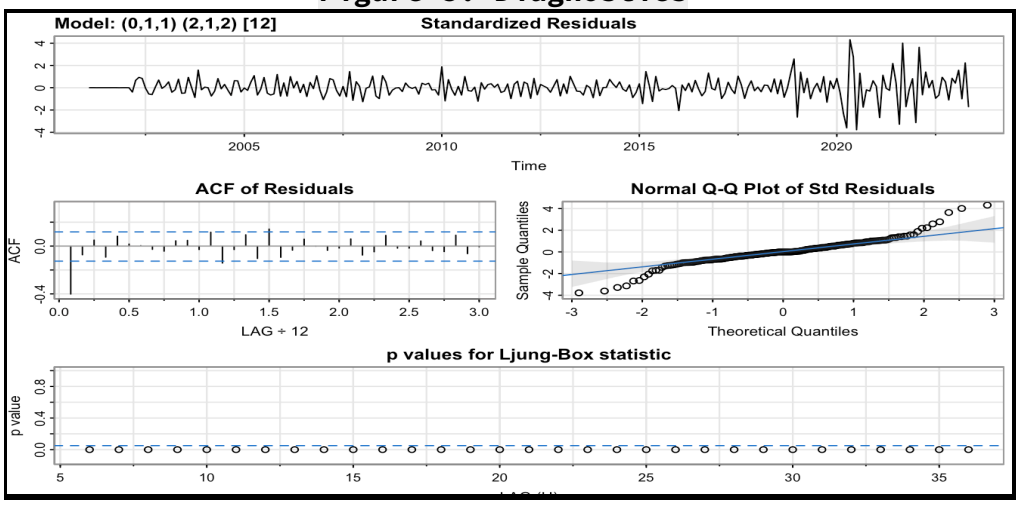
Below in **Figure 4**, you will find the resulting residual plots. To determine whether the SARIMA model successfully represents the underlying patterns in the time series, one can look at patterns such as randomness, absence of trends, and constant variance in the residuals. Our plots indicate no apparent trends as nearly all points (with the exception of 4) fall closely to the 0.0 value. This suggests that the model has effectively captured the overall trend present in the time series. This indicates that the model is adequately accounting for the systematic patterns and the residuals exhibit randomness around the model's predictions. The same can be said about the ACF and PACF models.

Figure 4: SARIMA Residual Plots



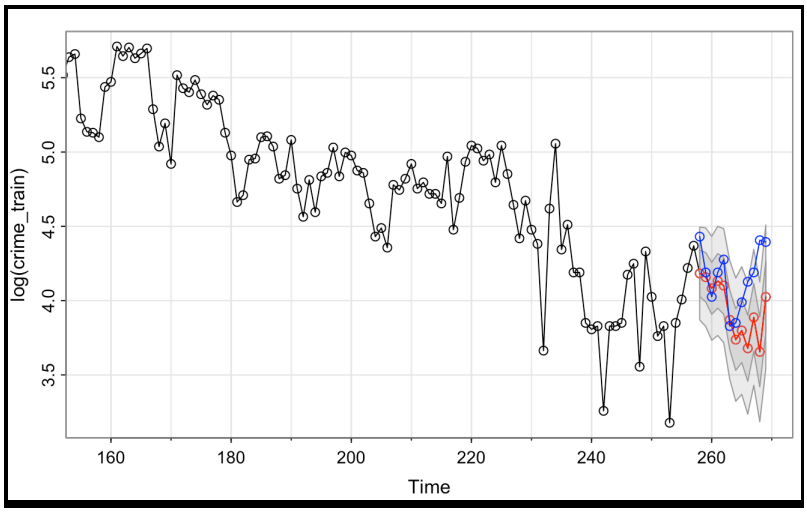
The standardized residual plot in Figure 5 helps assess the assumption of constant variance and identify any patterns or systematic deviations in the residuals.. Similarly, thes points in the QQ plot closely follow the diagonal line, meaning residuals are normally distributed. The autocorrelation values fall within the confidence bounds, it indicates that the residuals are uncorrelated. Lastly, the p-values of Ljung-Box statistics test the null hypothesis of no autocorrelation in the residuals. By running a Box-Pierce test, we determined our p-value was approximately 0.81, this p-value suggests the model is well at capturing the temporal dependencies in the data.

Figure 5: Diagnostics



Now that we have explored our model, we can test how accurate it is at actually forecasting future values. This was achieved by using the `sarima.for()` function on our training dataset, utilizing the best-fitting model parameters we identified. We then applied this model to the testing dataset. **`Figure 6`** shows the predicted values (in red) alongside the true data values (in blue), you can find the code for this in the Appendix. It is apparent that our model captures a downward trend in Public Peace Violations, as the predicted values align with the decreasing pattern observed in recent years. This alignment with the actual data points strengthens our confidence in the model's ability to forecast future trends in Public Peace Violations, enabling us to make informed decisions or predictions based on its output.

`Figure 6: SARIMA Forecasting`

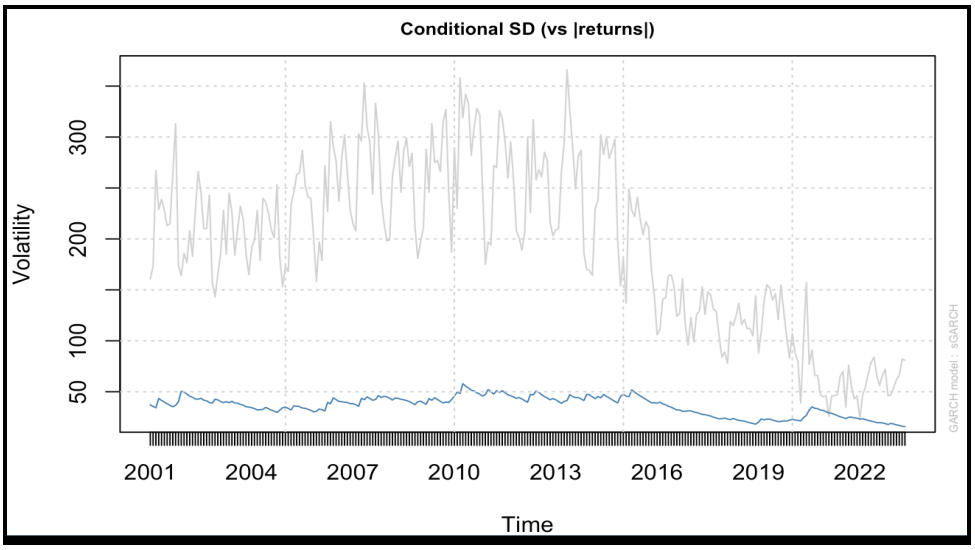


4.2 Results From GARCH Model

In order to capture and comprehend the volatility or variance clustering shown in the data, a GARCH model was used to analyze Chicago's crime rates. With the help of the GARCH model, we can simulate how the volatility patterns of crime rates change over time, showing how spikes in crime tend to be followed by a recurrence of these spikes and drops in crime by a recurrence of these drops. The GARCH model allows us to calculate the conditional volatility in crime rates by including lagged conditional variances, giving us insights into the dynamics of criminal risk and uncertainty across time. By setting and running our model, we can see the volatility of our data, as seen in Figure 8.

In **`Figure 7`**, it is clear that our conditional SD values are consistently low or close to zero, which points towards periods of relative stability or low volatility in the data. This suggests that the crime rates in Chicago are relatively predictable and less likely to experience sudden spikes or large fluctuations.

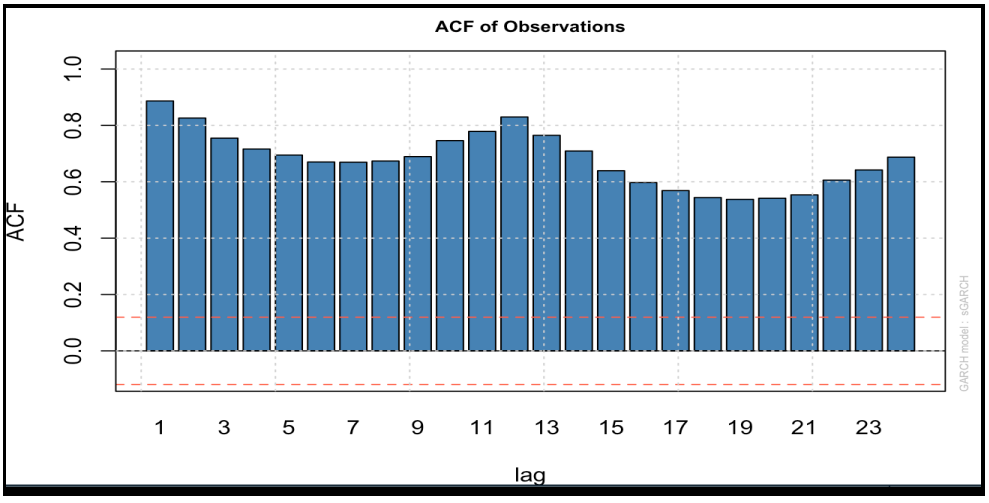
Figure 7: Volatility



In **Figure 8**, since all of the ACF values are over the threshold level when employing a GARCH model to analyze Chicago crime data, it suggests a strong and significant correlation between the crime rates at various time intervals. This implies that there is a considerable temporal relationship between Chicago's crime rates, with high crime rates in one time typically being followed by high crime rates in later periods.

The ACF plot's presence of autocorrelations over the threshold level suggests that the crime rate series has underlying patterns or trends that are not taken into account by the GARCH model. It implies that there might be more elements or variables affecting crime rates, and the model might need to be improved or expanded.

Figure 8: Autocorrelation of GARCH Model



5. Conclusion and Future Study

Overall, the primary objective of this project was to examine the trends in Chicago's "Public Peace Violation" crime rates. When the data was thoroughly evaluated, it became clear that the crime rates showed both a general decline and seasonal fluctuations, which meant that some months had greater crime rates than others. The Seasonal ARIMA model was employed, which turned out to be the best suitable for forecasting the time series, to accurately capture these characteristics. We had the opportunity to find patterns in the data and make significant predictions by using this model.

The volatility in the crime rate data was also visualized using the GARCH model. With the use of this model, we were able to assess the shifting nature of crime rates and comprehend the various degrees of uncertainty surrounding them. We learned more about the evolving volatility patterns over time by analyzing the conditional standard deviation plot.

Future research could delve more deeply into the underlying causes of the observed seasonal changes in crime rates. Examining additional variables like the weather, current affairs, or economic indicators could be beneficial to explain why certain months have higher crime rates than others. External factors may also improve forecasting accuracy and permit a more thorough investigation of crime trends by being incorporated into forecasting models. Additionally, legislators and law enforcement organizations may benefit from taking into account how particular policies or initiatives affect crime rates.

References

ARCH/GARCH Models | STAT 510. (n.d.). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat510/lesson/11/11.1

Box–Jenkins Method.” Wikipedia, 10 July 2022 en.wikipedia.org/wiki/Box%E2%80%93Jenkins_method.

Shumway, R. H., & Stoffer, D. S. (2011). Time Series Analysis and Its Applications. In *Springer eBooks*. https://doi.org/10.1007/978-1-4419-7865-3.

Appendix:

Figure 1.1: tsclean() identifies and removes all outliers, the resulting graph is shown below.

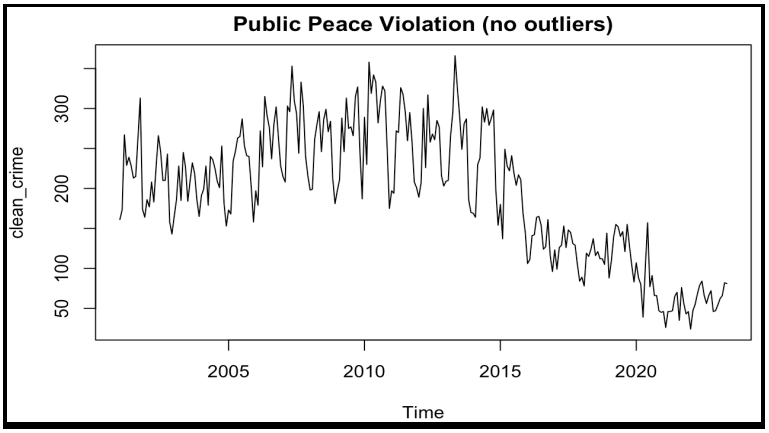


Figure 1.2: Decomposition

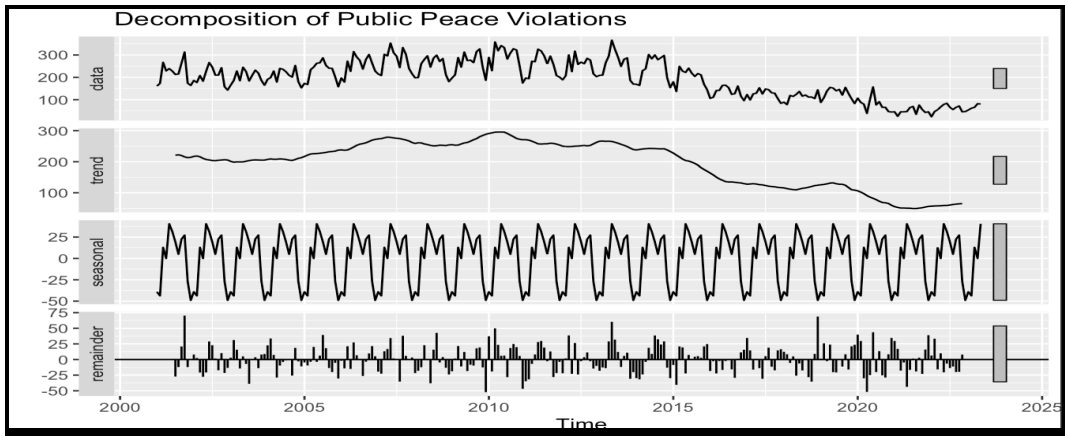


Figure 1.3: SARIMA Model

```
Series: clean_crime
ARIMA(0,1,1)(2,1,2)[12]

Coefficients:
          ma1     sar1     sar2     sma1     sma2
      -0.5730  -0.6123  -0.1220  -0.1270  -0.3226
s.e.   0.0631   0.4953   0.0982   0.4957   0.3649

sigma^2 = 662.7:  log likelihood = -1196.7
AIC=2405.41   AICc=2405.75   BIC=2426.68
Series: ts_crime
ARIMA(1,1,1)(1,1,2)[12]

Coefficients:
         ar1      ma1    sar1     sma1    sma2
      0.1831  -0.8008  0.5842  -1.4731  0.5405
s.e.  0.0846   0.0501  0.6474   0.6290  0.5378

sigma^2 = 1166:  log likelihood = -1272.88
AIC=2557.76   AICc=2558.1   BIC=2579.03
```

Figure 1.4: GARCH Results

```
Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  11.61869  0.00299939
 Shapiro-Wilk Test  R    W      0.9876262 0.02117352
 Ljung-Box Test     R    Q(10)  21.5856   0.01736063
 Ljung-Box Test     R    Q(15)  78.27356  1.442755e-10
 Ljung-Box Test     R    Q(20)  91.84894  3.515088e-11
 Ljung-Box Test     R^2  Q(10)  12.07332  0.2801794
 Ljung-Box Test     R^2  Q(15)  27.22433  0.02696084
 Ljung-Box Test     R^2  Q(20)  42.6265   0.002289708
 LM Arch Test       R    TR^2   17.77317  0.1227535

Information Criterion Statistics:
       AIC        BIC        SIC       HQIC
-0.2391030 -0.1855062 -0.2395399 -0.2175760
```

R Code:

```
# load the required packages
library(data.table)
library(forecast)
library(dplyr)
library(lubridate)
library(tseries)
library(stats)
library(astsa)
library(rugarch)
library(forecast)
library(fGarch)

# read in data
data<-fread('/Users/brendamorales/Desktop/final-project/Crimes_-_2001_
to_Present.csv')

# data processing & feature engineering
crime <- data %>%
  dplyr::filter(`Primary Type` == "PUBLIC PEACE VIOLATION") %>%
  dplyr::mutate(Date = mdy_hms(Date), # convert date time
         Date = date(Date), # subset only date component
         Month = month(Date)) %>% # new column for each month
  select(`Primary Type`, Month, Year) %>%
  group_by(Year, Month) %>%
  summarise(total = n()) %>%
  ungroup() %>%
  mutate(Date = parse_date_time(paste(Year,Month), orders = "Ym")) %>%
  select(Date, total)
```

```
# end in may 2023 since the month of June is not complete
crime <- crime %>%
  filter(Date < "2023-05-01") %>%
  select(Date, total)

#create the time series object
ts_crime <- ts(data=crime$total,
               start=2001,
               frequency=12)
ts_crime

# individual plots for original time series data, ACF and PACF
ts.plot(ts_crime, main = "Public Peace Violations", xlab='Time (in
years from 2001- 2023)')
acf(ts_crime)
pacf(ts_crime)

# alternative plot the time series along with the ACF and PACF graphs
ggtsdisplay(ts_crime, lag.max=20)

# identify outliers
tsoutliers(ts_crime)

clean_crime <-tsclean(ts_crime)
ts.plot(clean_crime, main='Public Peace Violation (no outliers)')

# decomposing out data
decompose(clean_crime) %>% autoplot(main="Decomposition of Public
Peace Violations")

# perform Augmented-Dickey Fuller test
adf.test(clean_crime)

# set training and testing datasets
n <- length(clean_crime)
crime_train <- clean_crime[1: (n-12)]
crime_test <- clean_crime[(n-11):n]

# .48641 is a low value indicating we should apply transformations
BoxCox.lambda(clean_crime)

# determine number of difference that is necessary
ndiffs(clean_crime)
nsdiffs(clean_crime)
```

```
# transforming the data  &  plotting it
log_crime <- log(clean_crime)
log_difference <- diff(log_crime)
plot.ts(cbind(clean_crime,  log_crime, difference), main="Processed
Crime Data VS Log VS Log Difference")

# SARIMA MODEL
# grid search
grid_parameters <- data.frame(expand.grid(P=1:2,
                                          Q=1:2,
                                          p=0:1,
                                          q=0:1),
                              AIC=NA, BIC=NA)

# loop that iterates between possible parameter combinations, returns
respective AIC and BIC values
for (i in 1:nrow(grid_parameters)) {
  val <- df[i, ]
  model <- arima(clean_crime,
                 order = c(val$p, 0, val$q),
                 seasonal=list(order=c(val$P, 1, val$Q),
                               period = 12), method='ML')

  grid_parameters[i, ]$AIC <- model$aic; grid_parameters[i, ]$BIC <-
BIC(model)

}

grid_parameters[order(grid_parameters$AIC)[1:3], ];
grid_parameters[order(grid_parameters$BIC)[1:3], ]


# getting same result with auto.arima()
fit <- auto.arima(clean_crime, seasonal=TRUE,)
fit

#  creating our model
model <- arima(x =log(clean_crime), order=c(0, 1, 1), seasonal =
list(order=c(2, 1, 2), period=12), method='ML')
model

#diagnostic check
res <- residuals(model)
par(mfrow=c(1,3))
plot.ts(res, type = 'o', main = "Residual Plots of SARIMA Model")
```

```
acf(res, main="ACF of Best SARIMA")
pacf(res, main='PACF of Best SARIMA')

#SARIMA model
sarima(log_difference, 0, 1, 1, P=2, D=1, Q=2, S=12, details = TRUE)

# running Box-Pierce test / Shapiro-Wilk normality test
Box.test(res)
shapiro.test(res)

# forecasting plot
n <-length(clean_crime)

forecasting_values <- sarima.for(log(crime_train),
                                 n.ahead=12,
                                 plot.all=F, p=0,
                                 d=1, q=1, P=, 2,
                                 D=1, Q=2, S=12)

# red indicates PREDICTED values
lines((n - 11):n, forecasting_values$pred, col = "red")

# blue indicates ACTUAL values
lines((n-11):n, log(crime_test), col="blue")

points((n - 11):n, log(crime_test), col = "blue")


# GARCH MODEL
summary(garchFit(~arma(1,0)+garch(1,0), diff(log(clean_crime))))

garch_order <- c(1, 1)  # adjust the GARCH order as desired
garch_model <- ugarchfit(data = clean_crime, spec =
ugarchspec(variance.model = list(model = "sGARCH", garchOrder =
garch_order)))

plot(garch_model)  # plot the conditional mean
```