[redacted]

Predicting Crime:

Time Series Analysis of Crime Rate in the City of Chicago

[redacted]

[redacted]

Date: 14/06/2024
Pages: 10
Characters: 21.443

Table of Contents

Introduction

Crime is defined as an act against the law because socially harmful and punishable by the state authorities, which influences a nation's economy, reputation and quality of life (Sivapriya, Vijay Ganesh, Pradeeshwar&al, 2023).

As presented by Mark Shaw, Jan van Dijk and Wolfgang Rhomberg in their 2003 paper "*Determining trends in global crime and justice: an overview of results from the United Nations surveys of crime trends and operations of criminal justice systems*", the general trend of crime has been decreasing recent years in some areas of the world such as North America, remained stable in others, while some parts of the world still face a high proportion of violent crimes. In this paper the geographical focus will be on the United States of America, in particular on the city of Chicago.

Both the Federal Bureau of Investigation (FBI) and the Bureau of Investigation Statistics (BJS) have highlighted significant reductions in violent and property crime rates in the country since the early 1990s, but the population is still keen to believe crime is up and some areas are more dangerous than others (Gramlich, 2024). In Chicago for example the number of homicides in 2023 was 50% more than 10 years ago, with younger victims and not many criminals caught (Gowins&Josko, 2024).

Because of this, innovative technologies and predictive techniques are starting to be implemented on large scale in crime analytics, with the goal of better understand the phenomenon and prevent crime. Predictive policing is the name of one of these new implementations, where law enforcement use statistical data to predict which areas have higher probability of crime and guide them in decision making. These predictions come with many benefits such as better resource allocation and better identification of the people involved, but also drawbacks such as lack of transparency and stigmatization of some groups (Meijer&Wessels, 2019).

Nevertheless, this study will not focus on the different results and ethics of the application but only on the technical fundamental part of forecasting. Starting from a dataset extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system, recording from 2001 to the present, the paper will attempt to forecast the total crime count in the years to come. This forecasting aims to determine future crime trends based on historical analysis. The study will take into consideration major exogenous events' impact and the nature of crime itself, which is neither systematic nor entirely random (Yu, Ward, Morabito& Ding, 2011).

To do the aimed forecasting two models will be used: AutoRegressive Integrated Moving Average (ARIMA), in particular Seasonal ARIMA, and Exponential Smoothing. These two models are the most widely used approaches in time series forecasting because they can provide complementary insights. Exponential Smoothing accounts for error, trend, and seasonality, while ARIMA can capture a huge range of patterns and describe the autocorrelation in the data (Hyndman&Athanasopoulos, 2018). These models apply greatly to crime data considering its seasonality pattern for different crimes and trends due to socio-economic changes, law enforcements practices and other factors.

Related Work Literature

Studies before have analyzed crime in different areas of the world using models like SARIMA and Exponential Smoothing, for example in the paper "*SARIMA: A Seasonal Autoregressive Integrated Moving Average Model for Crime Analysis in Saudi Arabia*", Talal&al (2022) implement a SARIMA (0,0,0)(2,0,2) to predict crime patterns in Saudi Arabia and compared it with a random forest model and a XGB model. SARIMA model together with Holt-Winters has been used also on the "*Chicago Crime Time Series Analysis*" available on RPubs by RStudio, which is using the same dataset this work is based upon but targeting for forecasting one crime (theft) during hours of the day. The validity of Exponential Smoothing has been proven by works such as "*Hybrid of deep learning and exponential smoothing for enhancing crime forecasting accuracy*" by Butt&al (2022) and

"*Crime Data Forecasting using Exponential Smoothing*" by Everly Chua&al (2020) where the method is utilized to determine the most accurate forecasts for different types of crime incidents. Taking previous works under consideration, the decision to work with ARIMA and Exponential Smoothing was reinforced.

Dataset Description

The Chicago Crimes Dataset was extrapolated from the Chicago Data Portal and it considers data from 2001 until seven days prior the export date, so in the case of this paper until the second week of May 2024. The dataset initially was extensive, with numerous columns and more than 8 million observations. The *.cvs* file was imported on RStudio and analyze, starting from the *head()* function that highlighted columns and initial observations, as seen in Figure A1 in the Appendix. The data was subsequently cleaned, missing values accounting for 1% of total were removed,

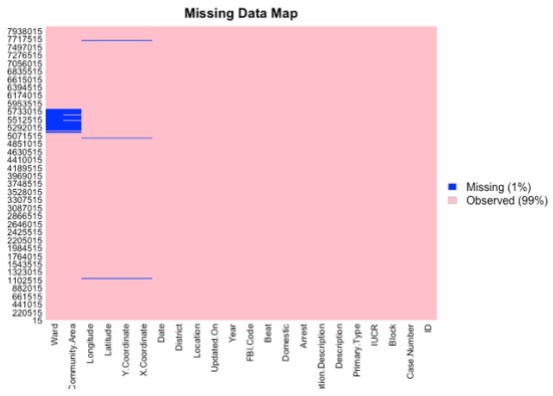

Figure 1: Missing Data Map

formats of columns like Date were fixed and duplicates were checked. Before moving into the focus of the forecast, two plots to better understand how levels of crime change during the hours of the day and what are the 5 most committed crime in Chicago were created (Figure A3 and A4). Then before proceeding selecting the necessary columns for the prediction, it was decided to exclude all observation from before 2003 since in the first two years the program was still being set up and so the numbers of crime recorded are not accurate for Total Number of Crimes and for the sake of analysis also the year 2024 was dropped considering only the first 5 months are present. At this point only the columns of Date and Primary.Type, which contains the crime committed, were selected. Date changed in the format yearmonth, discarding Hours-Day, while Primary.Type was grouped by Date, summarize and renamed *Total_Crime_Count*, so without distinguishing by single crime types but summing them up all together. Lastly the dataset was transformed into a tsibble for future purposes.

Figure 2: Total Crime Counts in Chicago in the Years

In Figure 2 it's possible to observe the overall trend of crime, which had a significant decline from the early 2000s until 2015, followed by a stabilized period with some fluctuations.

Seasonality is visible and consistent through times annually, as showed by the peaks appearing at regular intervals. Further analysis is done in the next session with regard to seasonality, stationarity and structural breaks.

Methodology

Mathematical Transformation

Since the data is showing variation, a transformation was considered probably useful. To stabilize variance and improve normalization of the data, the Box-Cox logarithmic transformation was chosen, since it includes both logarithmic and power transformation. After the transformation, the lambda value was observed to be equal to 1. If lambda is equal to 1, then the data is shift downwards, but the shape of the time series doesn't change (Hyndman&Athanasopoulos, 2018). By plotting the graph (Figure B1), that is confirm.

Seasonality

The seasonal plot Figure 3 shows a clear seasonal pattern where crime tends to spike during warmer months (May to August) and lower during colder periods (January to February). These seasonal peaks indicate that external factors such as the weather, influence crime rates. To further investigate this, the seasonal subseries plot (Figure 4) was plotted, with the horizontal lines indicating the means for each month, confirming seasonal peaks during the summer months.

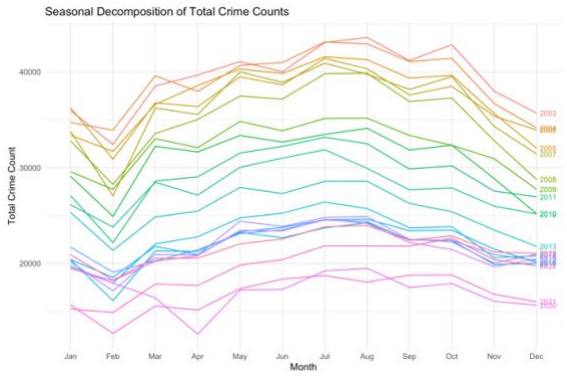

Figure 3: Seasonal plot of Total Crime Counts

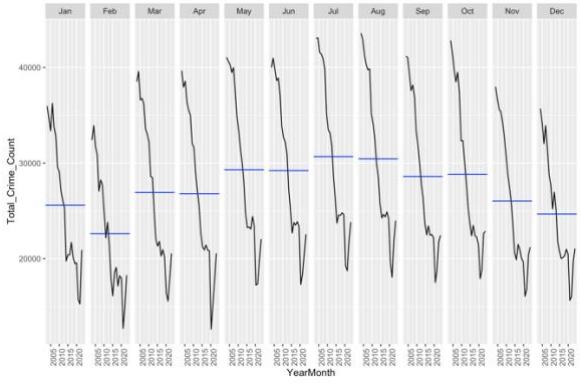

Figure 4: Seasonal Subseries Plot of Crime in Chicago

Lagged scatterplots are also available in the Appendix (Figure B2), exhibiting a stable seasonal pattern in the data. The Autocorrelation Function (Figure 5) and Partial Autocorrelation Function (Figure B3) were plotted too. From ACF it's possible to observe the presence of a trend, given by the slow decrease as lags increase, and seasonality, given by the scallop shape (Hyndman&Athanasopoulos, 2018). The periodic spikes around lags 12 and 24 could suggest yearly seasonality, which is in line with the monthly data. Considering the bounds created by the blue dashed lines, the series is not white noise, since all the lags spike outside of the bound.

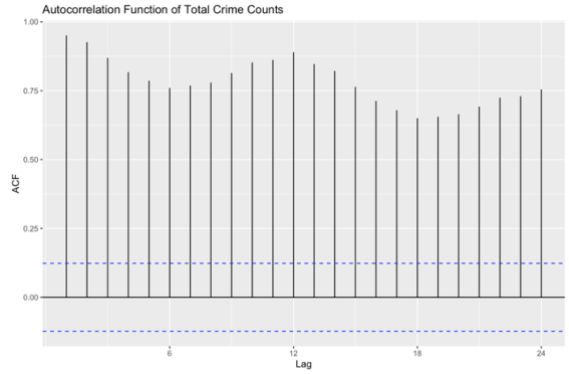

Figure 5: Correlogram of crime rate

Lastly the Seasonal Trend decomposition using Loess was carried out, considering it's a robust and versatile method for decomposing a time series into three components: Seasonal, Trend, and Remainder.

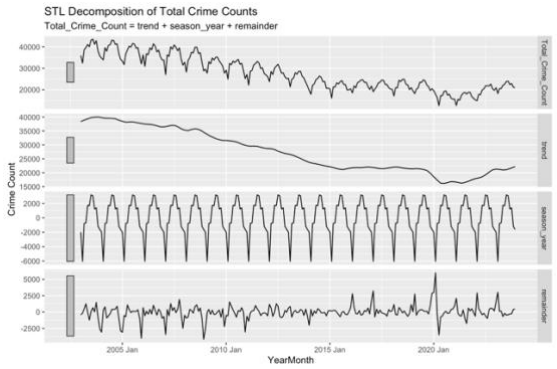

Figure 6: STL Decomposition of the data

The decomposition shows the long term decline in trend, while the seasonal component is confirmed as a clear, recurring yearly pattern, like supposed and proved by all the previous plots. In the remainder some irregularities are visible, which might be caused by anomalies not captured by trend or seasonal components.

Stationarity

Before proceeding to forecasting, it's important to take under consideration the requirements of the chosen models, and ARIMA needs data to be stationarity to achieve accuracy in prediction. The crime data presented has trend and seasonality, so it's not stationary. This is shown also in the tests executed to determine stationarity. There are in fact different unit roots tests that can be used, in particular the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test and the Augmented Dickey–Fuller test (ADF). In the KPSS test the null hypothesis is that the data are stationary, so if we reject the null hypothesis there's a unit root present. In the ADF instead the null hypothesis is that there's a unit root so data is not stationary and if the null hypothesis is rejected then there is stationarity (it's a left-tailed test) (Hyndman&Athanasopoulos, 2018). Both tests have been done on the transformed Box-Cox data and all results are visible in the Appendix. For KPSS *mu* (constant mean) the value of the statistic is 3.886, greater than the p-value at all significance levels and KPSS *tau* (constant trend) test significance is 0.3777, again greater than the p-value at all significance levels, so we reject the null hypothesis of stationarity.

For the ADF test both *drift* and *trend* were tested and in one case the test statistic is -1.8536 with p-values of -3.44 (1% significance), -2.87 (5% significance) and -2.57 (10% significance), so the value is not negative enough and the null hypothesis of non-stationarity cannot be rejected.

Because of this, differencing is needed to reach stationarity, which is the computation of the differences between consecutive observations

(Hyndman&Athanasopoulos, 2018). In the case of the chosen dataset both a seasonal difference and a first difference were needed to obtain stationary data. In fact, only with seasonal differentiation ADF test would reject the null hypothesis with -4.44 < -3.46 (drift) / -4.78 < -3.99 (trend) and so reaching stationarity, but the KPSS test was still resulting in non-stationarity rejecting the null hypothesis with a test statistic of 0.71.

Therefore first differencing was applied as a second step, transforming the data to what can be seen in Figure 7.

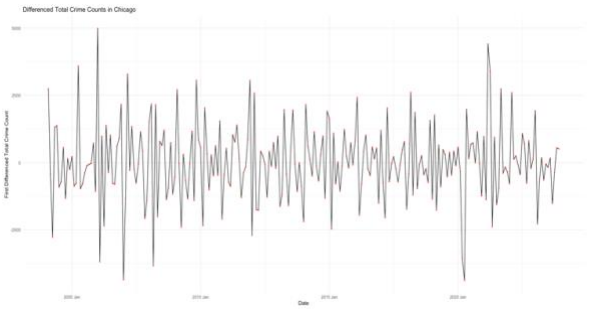

Figure 7: Plot of Differenced data

Testing again this time both KPSS and ADF results prove stationarity, in fact with KPSS we do not reject the null hypothesis with a test statistic of 0.0249 (mu) and critical values at different significance levels of 10%: 0.347, 5%: 0.463, 2.5%: 0.574, 1%: 0.739. Same for KPSS tau with a test statistic of 0.0181 and critical values at different significance levels of 10%: 0.119, 5%: 0.146, 2.5%: 0.176, 1%: 0.216. With ADF it's instead possible to reject the null hypothesis, highlighting once again that stationarity has been reached.

This can be noticed also looking again at previous plotted graphs.

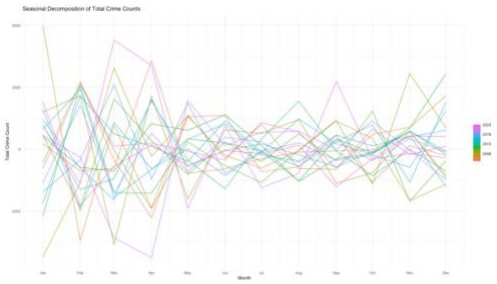

Figure 8: Seasonal Decomposition of Differenced Data

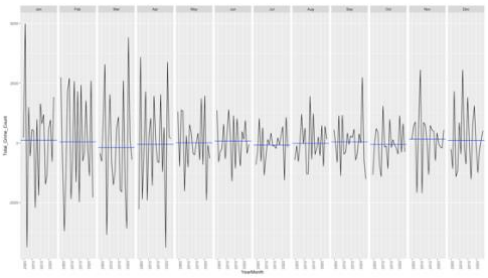

Figure 9: Seasonal Subseries of Differenced Data

For example from the seasonal subseries it's visible how seasonal peaks are no longer present and looking at the ACF (Figure 10) and PCAF (Figure B5) plots we can see the differencing has partially removed trend and seasonality, but some correlations remain, confirming Seasonal ARIMA as the model to work with.

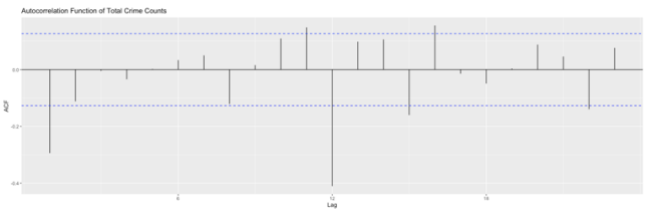

Figure 10: ACF of Differenced Data

Structural Breaks

A structural break is when a time series suddenly changes at a certain point in time. Different economic and political events can cause it in socioeconomic data (Antoch, Husková, Prášková&Veraverbeke, 2018).

Different approaches have been used to inquire structural breaks in the data: the Quandt likelihood ratio (QLR) Test, the Step Indicator Saturation (SIS) test, the Conventional Cumulative Sum (CUSUM) and the Moving Sum (MOSUM) test of residuals. All the plotted results are available in the Appendix and they all indicate proof of structural changes, the major ones around year 2010 (aftermath of the financial crisis) and around year 2015 (implementation of new policies such as the Police Data Initiative by President Barack Obama)(The White House, n.d.). The breakpoint around 2020 particularly emphasized by Figure C16 is probably correlated with Covid19.

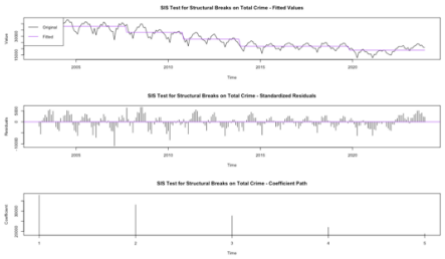

Figure 11: SIS Test for Structural Breaks Plot

Figure 11 shows the SIS test, which examines three aspects: fitted values, standardized residuals and coefficient path. All three graphs further validate the presence of structural breaks aligning with known events and policy changes in Chicago.

Considering how intervening on all these structural breaks could leave us with insufficient data for the forecasting, the final decision was to keep data as it is currently, while being aware of the risk of bias and error in the results.

Forecasting

Models: SARIMA and ETS

For forecasting as said before two models were selected: Seasonal ARIMA and Exponential Smoothing.

Seasonal ARIMA was chosen because it's capable of modelling seasonal data, autocorrelation and patterns, making it a strong choice for forecasting something as volatile as crime.

Exponential Smoothing was also chosen because it can capture both trends and seasonal patterns and can handle well volatility. While SARIMA models are powerful in handling moving average components, Exponential Smoothing provides a different perspective by focusing on the weighted average of past observations. This complementary approach allows for a more comprehensive evaluation of the forecast.

Seasonal ARIMA is the union between the non-seasonal part of the model (basic ARIMA represented by (p, d, q)) with the seasonal part of the model (P, D, Q)m. ARIMA is the combination of the order of the autoregressive part p (AR), the degree of first differencing involved d (I), and the order of the moving average part q (MA). The seasonal differencing is recorded in the seasonal component of the model (Hyndman& Athanasopoulos, 2018). To decide how to construct the model, ACF and PACF of the differenced data are plotted to determine the components. *auto.arima* can also be used on

the non-differentiated data to find the best model through R suggestion.

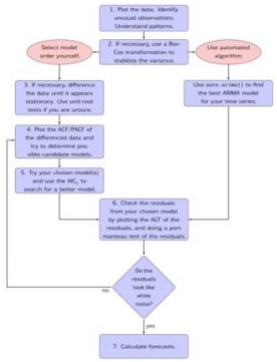

Figure 12: General process for forecasting taken from Hyndman&Athanasopoulos, 2018

For ETS, Box-Cox transformed data will be used since stationarity is not required and both the simpler Exponential Smoothing State Space Model *ets()* and the more complex Exponential Smoothing Holt-Winters *hw()* will be tested and then compared with SARIMA for forecasting.

Results

Starting from SARIMA with the use of *auto.arima* the model given was ARIMA(0,1,2)(0,1,1)[12]. Then *auto.arima* with *stepwise* set to FALSE resulted in model ARIMA(1,1,1)(2,1,1)[12]. The presence of different autoregressive and moving average in the two models indicates the different approaches implemented to analyze the patterns of the data, with the second model taking into account more complexity.

Then considering the ACF and the PACF plots, it was noticed a significant spike at lag 1 and lag 12 in the PACF, one differencing order was done and so was one seasonal differencing, and different spikes were present in ACF, models ARIMA(1,1,1)(1,1,1)[12] and ARIMA(1,1,2)(1,1,1)[12] were both evaluated.

Model	AIC	AICc	BIC
ARIMA(0,1,2)(0,1,1)[12]	4035.59	4035.76	4049.5
ARIMA(1,1,1)(2,1,1)[12]	**4031.57**	**4031.93**	4052.43
ARIMA(1,1,1)(1,1,1)[12]	4036.64	4036.9	4054.02
ARIMA(1,1,2)(1,1,1)[12]	4038.59	4038.95	4059.45

Table 1: SARIMA models criterions

The models were compared on the base of three criterions: AIC (Akaike Information Criterion), AICc (Corrected Akaike Information Criterion) and BIC (Bayesian Information Criterion). The best model is the one with smaller values, because it indicates better fitting. In this case, the ARIMA(1,1,1)(2,1,1)[12] model has the lowest AIC and AICc, defining it as the best fit among the models.

Residuals plots were also observed to determine which model could be the best fit. ARIMA(0,1,2)(0,1,1)[12] showed some autocorrelation in the ACF of residuals and some potential outliers in the histogram. The same goes for ARIMA(1,1,1)(1,1,1)[12]. The best model is confirmed to be ARIMA(1,1,1)(2,1,1)[12], presenting a better fit, a normal distribution with fewer outliers and the residuals are centered around zero with no apparent trend.

To confirm the ARIMA model selected, it's important also to operate the *Ljung-Box Test* to check for autocorrelation in the residuals.

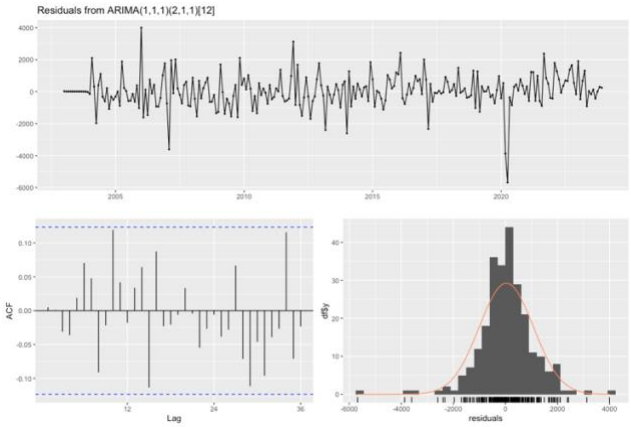

Figure 13: Residuals from ARIMA(1,1,1)(2,1,1)[12]

In this test the null hypothesis is of no autocorrelation, and with lag 5 p-value=0.9836, lag 10: p-value=0.5648, lag 15: p-value = 0.5173, and lag 20: p-value = 0.6679 we fail to reject the null hypothesis, suggesting the model is appropriate.

Following that, ETS was investigated. Starting from the STL Decomposition of the Box-Cox transformed data, additive error, additive trend and additive seasonality were assumed given the consistent trend and seasonal patterns in the decomposition and therefore model ETS(A,A,A) was taken under consideration. The decision was to test it together with ETS(M, Ad, A) for multiplicative error, damped trend and additive seasonality to explore another possible option. Finally Holt-Winters, which is a widely used approach for time series that exhibit both trend and seasonality. The decision was to implement it additive (there is also an option for multiplicative), considering how STL shows the seasonal pattern of consistent amplitude over time.

Model	AIC	AICc	BIC
Auto ETS(A,N,A)	4973.09	4975.13	5026.03
ETS (A, A, A)	4975.90	4978.52	5035.90
ETS (M, Ad, A)	5047.26	5050.20	5110.79
Holt-Winters	**4087.42**	**4090.18**	**4146.59**

Table 2: ETS models criterions

In this case the lowest AIC, AICc and BIC indicating the better fitting model to be the Holt-Winters, since it captures trend and seasonality more effectively.

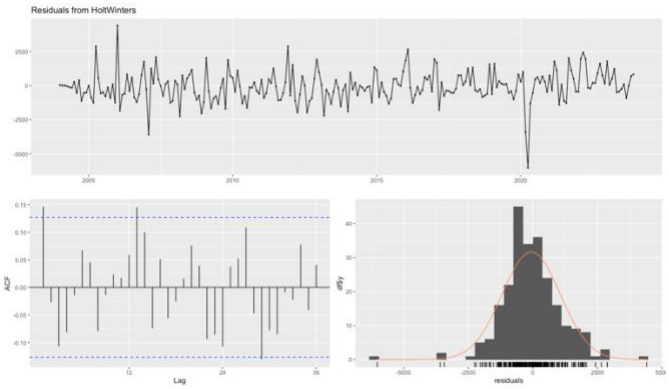

Figure 14: Residuals from Holt-Winters

Checking also the plots of the residuals, it's possible to notice they oscillate around zero, so the model has captured the main patterns. The residuals are approximately normally distributed and the ACF plot shows no significant autocorrelation and. Final confirmation of this was given by the Ljung-Box test, where the Holt-Winters model resulted the only one with a p-value big enough (0.07335) to not reject the null hypothesis.

Forecast

Based on all the consideration made in the section above and comparing all relevant criterions, the SARIMA model appears to be the best model for the forecast. In this final

section we will consider forecast and errors to achieve a final decision.

After diving the data in train size of 80% and test size of 20%, the two best models were trained and the accuracy metrics were evaluated: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE).

Model	RMSE	MAE	MAPE	MASE
ARIMA(1,1,1)(2,1,1)[12]	**956.34**	**682.65**	**2.486**	**0.463**
Holt-Winters	1030.49	766.84	2.803	0.520

Table 3: Models Error Analysis

The RMSE relatively low for SARIMA suggests the model's predictions are close to the actual values and MAE confirms the forecast to be quite accurate. Also SARIMA's MAPE is lower than the one of Holt-Winters, indicating smaller percentage errors. The Holt-Winters model is still performing adequately well, but the SARIMA model has demonstrated a strong performance across all metrics of the paper.

Observing the forecasting for the next 2 years in Figure 15 and 16, the plots look similar for the two models, they have overlapping confidence intervals (shaded areas) indicating quite the same level of uncertainty. However the visualizations suggest a slight upward trend in the SARIMA forecast, probably because of the model's complexity and the recent upward movements, while the Holt-Winters model focuses more on long-term trend, but again the difference is extremely minimal.

Eventually, the final decision is reached considering all the metrics, residual analysis and forecast evaluations, and the SARIMA model is chosen as the better performing one.

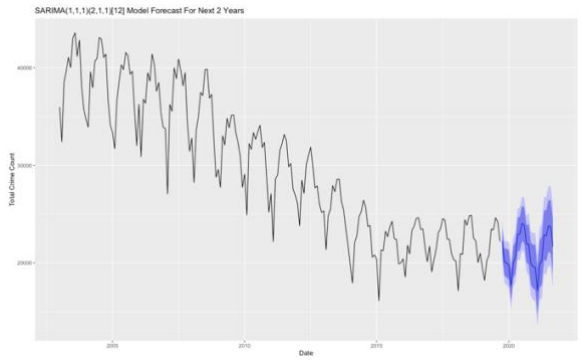

Figure 15: Forecast from SARIMA

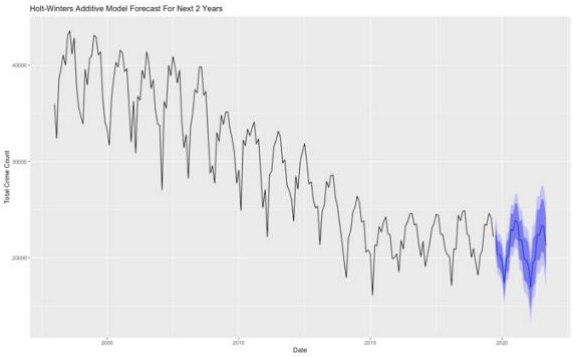

Figure 16: Forecast from Holt-Winters

Conclusions

In this paper, the aim was to choose a forecasting model for the crime rate in Chicago, selecting between Seasonal ARIMA and Exponential Smoothing. The analysis required extensive evaluation of the models, preceded by thorough data manipulation. The final decision was for the SARIMA model, considering the metrics and the acceptable forecast errors relative to the range of crime counts. Further improvements are possible, with the hope that crime forecasting will pave the way for a safer Chicago.

References

Antoch, J., Husková, M., Prášková, Z., & Veraverbeke, N. (2018). *Change-point analysis in panel data*. CentAUR: Central Archive at the University of Reading. doi:10.1080/07474938.2018.1454378

Butt, U. M., Letchmunan, S., Hassan, F. H., & Koh, T. W. (2022). *Hybrid of deep learning and exponential smoothing for enhancing crime forecasting accuracy*. PLOS ONE, 17(9), e0274172. https://doi.org/10.1371/journal.pone.0274172

Chua, E., & Tumibay, G. (2020). *Crime data forecasting using exponential smoothing.* International Journal of Advanced Trends in Computer Science and Engineering, 9(1.1), 69-75. https://doi.org/10.30534/ijatcse/2020/1391.12020

Gowins, H., & Josko, J. (2024). *Report: Chicago violent crime spikes 11%, arrests in just 11%.* Illinois Policy. Retrieved June 10, 2024, from https://www.illinoispolicy.org/report-chicago-violent-crime-spikes-11-arrests-in-just-11/

Gramlich, J. (2024). *What the data says about crime in the U.S*. Pew Research Center. Retrieved June 10, 2024, from https://www.pewresearch.org/short-reads/2024/04/24/what-the-data-says-about-crime-in-the-us/

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Meijer, A., & Wessels, M. (2019). *Predictive Policing: Review of Benefits and Drawbacks*. International Journal of Public Administration, 42(12), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

Noor, T. H., Almars, A. M., Alwateer, M., Almaliki, M., Gad, I., & Atlam, E.-S. (2022). *SARIMA: A seasonal autoregressive integrated moving average model for crime analysis in Saudi Arabia*. Electronics, 11(23), 3986. https://doi.org/10.3390/electronics11233986

Sacharinawati, S. (2022). *Chicago Crime Time Series Analysis*. RPubs. Retrieved June 10, 2024, from https://rpubs.com/sacharinawati/TS_Analysis

Shaw, M., Dijk, J., & Rhomberg, W. (2003). *Determining trends in global crime and justice: An overview of results from the United Nations surveys of crime trends and operations of criminal justice systems*. Forum on Crime and Society, 3(1), 35-63.

Sivapriya, G., Vijay Ganesh, B., Pradeeshwar, U. G., Dharshini, V., & Al-Amin, M. (2023). *Crime prediction and analysis using data mining and machine learning: An approach that helps predictive policing*. FMDB Transactions on Sustainable Computer Letters, 1(2), 64-75.

The White House. (n.d.). Retrieved June 10, 2024, from https://obamawhitehouse.archives.gov/

Yu, C.-H., Ward, M. W., Morabito, M., & Ding, W. (2011). *Crime forecasting using data mining techniques*. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 779-786). IEEE. https://doi.org/10.1109/ICDMW.2011.56

Appendix

Dataset Description

```
> head(crime)
        ID Case.Number                  Date                Block IUCR       Primary.Type
1 11037294    JA371270 03/18/2015 12:00:00 PM     0000X W WACKER DR 1153 DECEPTIVE PRACTICE
2 11646293    JC213749 12/20/2018 03:00:00 PM  023XX N LOCKWOOD AVE 1154 DECEPTIVE PRACTICE
3 11645836    JC212333 05/01/2016 12:25:00 AM   055XX S ROCKWELL ST 1153 DECEPTIVE PRACTICE
4 11645959    JC211511 12/20/2018 04:00:00 PM    045XX N ALBANY AVE 2820      OTHER OFFENSE
5 11645601    JC212935 06/01/2014 12:01:00 AM   087XX S SANGAMON ST 1153 DECEPTIVE PRACTICE
6 11646166    JC213529 09/01/2018 12:01:00 AM 082XX S INGLESIDE AVE 0810            THEFT
                        Description Location.Description Arrest Domestic Beat District Ward
1     FINANCIAL IDENTITY THEFT OVER $ 300                BANK  false    false  111        1   42
2 FINANCIAL IDENTITY THEFT $300 AND UNDER           APARTMENT  false    false 2515       25   36
3     FINANCIAL IDENTITY THEFT OVER $ 300                      false    false  824        8   15
4                     TELEPHONE THREAT           RESIDENCE  false    false 1724       17   33
5     FINANCIAL IDENTITY THEFT OVER $ 300           RESIDENCE  false    false 2222       22   21
6                            OVER $500           RESIDENCE  false     true  631        6    8
  Community.Area FBI.Code X.Coordinate Y.Coordinate Year           Updated.On Latitude
1             32       11           NA           NA 2015 08/01/2017 03:52:26 PM       NA
2             19       11           NA           NA 2018 04/06/2019 04:04:43 PM       NA
3             63       11           NA           NA 2016 04/06/2019 04:04:43 PM       NA
4             14      08A           NA           NA 2018 04/06/2019 04:04:43 PM       NA
5             71       11           NA           NA 2014 04/06/2019 04:04:43 PM       NA
6             44       06           NA           NA 2018 04/06/2019 04:04:43 PM       NA
  Longitude Location
1        NA
2        NA
3        NA
4        NA
5        NA
6        NA
```

Figure A1: Data First Observations

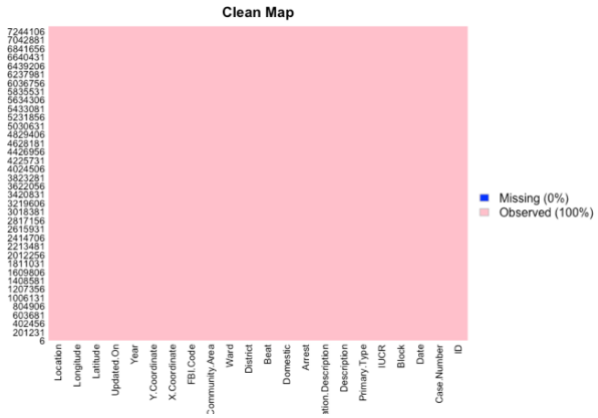

Figure A2: Clean Map no missing values present

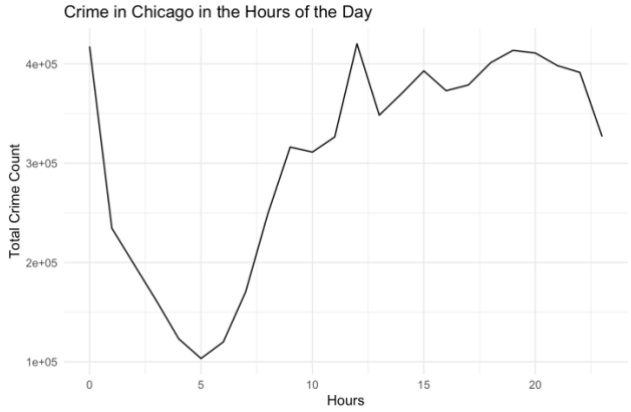

Figure A3: Crime in the city throughout the hours of the day

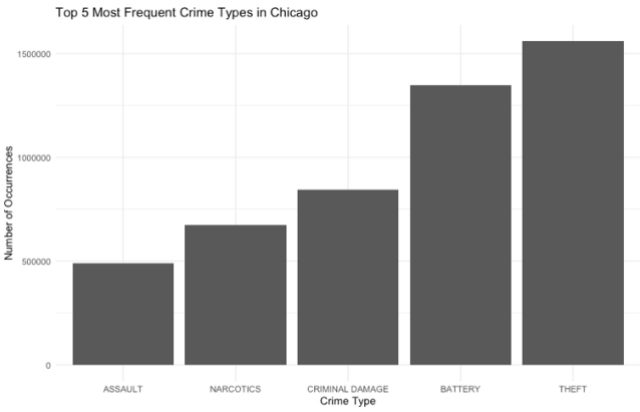

Figure A4: Top 5 Crimes in Chicago

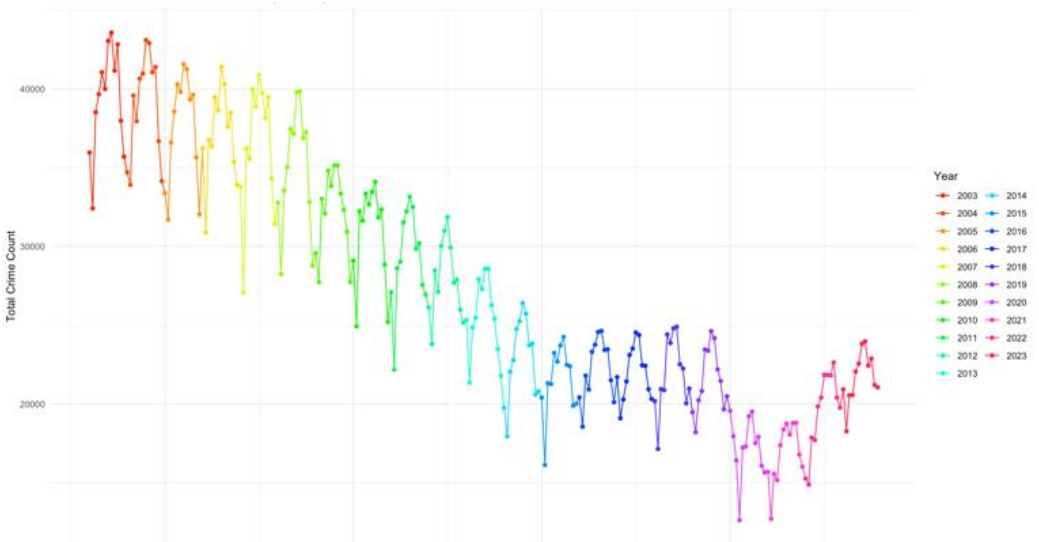

Figure A5: Total Crime Counts in Chicago through the Years with years color marked

Methodology

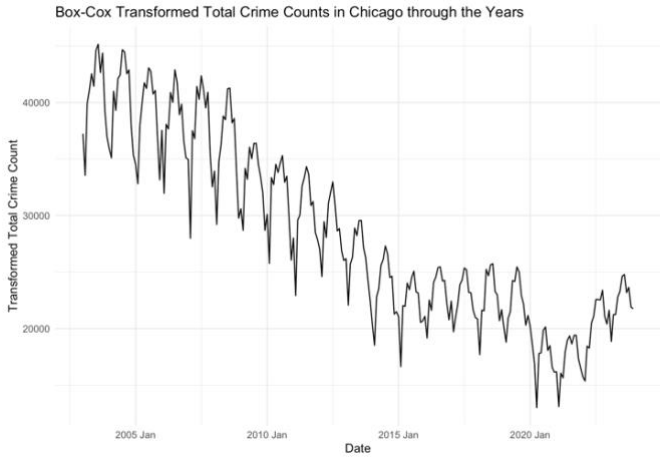

Figure B1: Box-Cox transformation on dataset

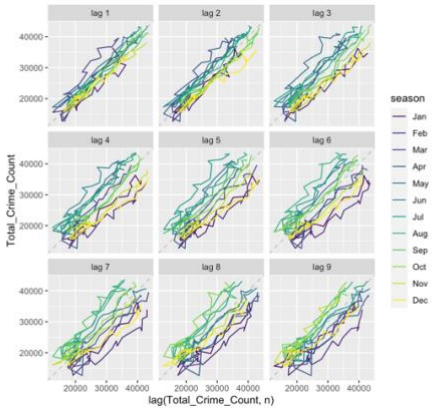

Figure B2: Lagged scatterplots for monthly crime

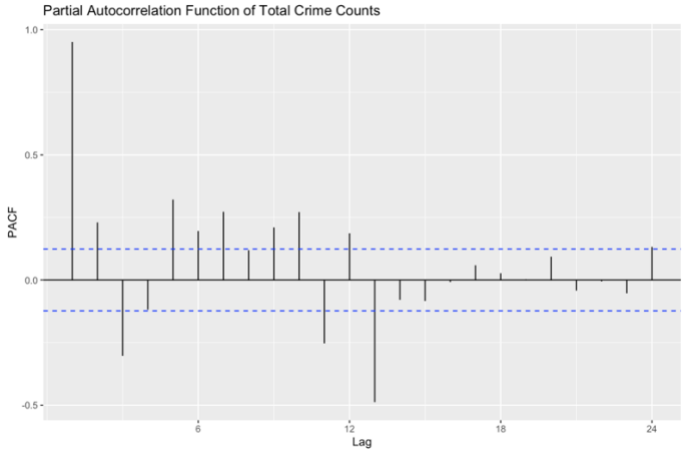

Figure B3: Partial Autocorrelation Function of crime rate

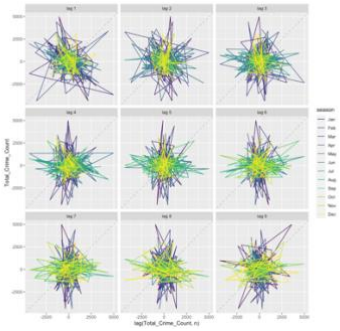

Figure B4: Lagged scatterplots of differenced data

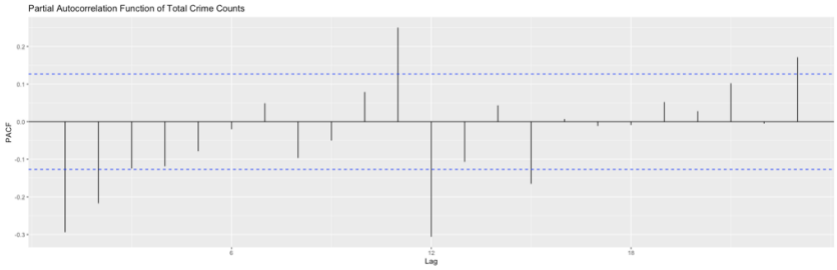

Figure B5: PACF of Differenced Data

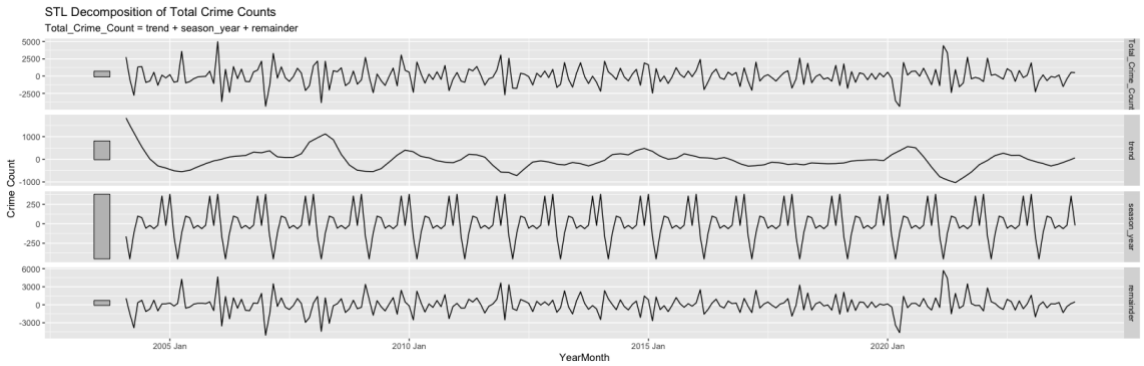

Figure B6: STL Decomposition of Differenced Data

Unit Root Tests

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 5 lags.

Value of test-statistic is: 3.886

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Figure C1: KPSS mu on transformed data

```
######################
# KPSS Unit Root Test #
######################

Test is of type: tau with 5 lags.

Value of test-statistic is: 0.3777

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.119 0.146  0.176 0.216
```

Figure C2: KPSS tau on transformed data

```
###########################################
# Augmented Dickey-Fuller Test Unit Root Test #
###########################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-6466.7 -1583.7   126.2  1337.8  7656.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 890.71443  532.82196   1.672 0.095851 .
z.lag.1      -0.03458    0.01866  -1.854 0.064992 .
z.diff.lag   -0.23066    0.06162  -3.744 0.000226 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2279 on 247 degrees of freedom
Multiple R-squared:  0.07469,   Adjusted R-squared:  0.0672
F-statistic: 9.969 on 2 and 247 DF,  p-value: 6.858e-05


Value of test-statistic is: -1.8536 1.8047

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

Figure C3: ADF drift on transformed data

```
###########################################
# Augmented Dickey-Fuller Test Unit Root Test #
###########################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-6939.4 -1449.3   352.6  1527.7  6629.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8143.03150 1755.28017   4.639 5.69e-06 ***
z.lag.1       -0.20574    0.04351  -4.728 3.82e-06 ***
tt           -20.12730    4.65726  -4.322 2.25e-05 ***
z.diff.lag    -0.14315    0.06287  -2.277   0.0237 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2201 on 246 degrees of freedom
Multiple R-squared:   0.14,     Adjusted R-squared:  0.1295
F-statistic: 13.35 on 3 and 246 DF,  p-value: 4.228e-08


Value of test-statistic is: -4.7283 7.515 11.1794

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Figure C4: ADF trend on transformed data

```
###########################################
# Augmented Dickey-Fuller Test Unit Root Test #
###########################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-5469.6  -685.3    45.7   609.7  3954.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  319.49734  307.39990   1.039 0.299766
z.lag.1       -0.01803    0.01108  -1.628 0.104977
z.diff.lag1   -0.37961    0.06624  -5.730 3.23e-08 ***
z.diff.lag2   -0.08508    0.06072  -1.401 0.162563
z.diff.lag3   -0.12588    0.06062  -2.076 0.038997 *
z.diff.lag4   -0.20729    0.06119  -3.387 0.000834 ***
z.diff.lag5   -0.19104    0.06121  -3.121 0.002039 **
z.diff.lag6   -0.18657    0.05987  -3.117 0.002071 **
z.diff.lag7   -0.19752    0.05989  -3.298 0.001133 **
z.diff.lag8   -0.26550    0.05980  -4.440 1.42e-05 ***
z.diff.lag9   -0.18930    0.06105  -3.100 0.002181 **
z.diff.lag10  -0.01292    0.06077  -0.213 0.831808
z.diff.lag11  -0.10989    0.05982  -1.837 0.067522 .
z.diff.lag12   0.52679    0.05897   8.933  < 2e-16 ***
z.diff.lag13   0.10378    0.06381   1.626 0.105271
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1208 on 223 degrees of freedom
Multiple R-squared:  0.7489,    Adjusted R-squared:  0.7332
F-statistic: 47.51 on 14 and 223 DF,  p-value: < 2.2e-16


Value of test-statistic is: -1.6278 3.029

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

Figure C5: ADF drift on lags

```
###########################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-5490.3  -682.0    48.7   610.3  3956.9

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -90.50984 1912.89656  -0.047 0.962304
z.lag.1        -0.00836    0.04591  -0.182 0.855656
tt              1.07296    4.94058   0.217 0.828273
z.diff.lag1    -0.38964    0.08088  -4.818 2.69e-06 ***
z.diff.lag2    -0.09704    0.08207  -1.182 0.238321
z.diff.lag3    -0.13742    0.08072  -1.702 0.090069 .
z.diff.lag4    -0.21871    0.08080  -2.707 0.007318 **
z.diff.lag5    -0.20161    0.07830  -2.575 0.010682 *
z.diff.lag6    -0.19594    0.07391  -2.651 0.008600 **
z.diff.lag7    -0.20609    0.07184  -2.869 0.004519 **
z.diff.lag8    -0.27323    0.06971  -3.920 0.000118 ***
z.diff.lag9    -0.19619    0.06893  -2.846 0.004840 **
z.diff.lag10   -0.01896    0.06695  -0.283 0.777281
z.diff.lag11   -0.11541    0.06511  -1.773 0.077668 .
z.diff.lag12    0.52195    0.06316   8.264 1.28e-14 ***
z.diff.lag13    0.10037    0.06584   1.524 0.128842
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1210 on 222 degrees of freedom
Multiple R-squared:  0.749,     Adjusted R-squared:  0.732
F-statistic: 44.16 on 15 and 222 DF,  p-value: < 2.2e-16


Value of test-statistic is: -0.1821 2.0264 1.3428

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Figure C6: ADF trend on lags

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-5660.1  -808.3   -39.4   690.2  4531.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -178.31142   93.90212  -1.899   0.0588 .
z.lag.1       -0.20138    0.04534  -4.441 1.38e-05 ***
z.diff.lag    -0.19292    0.06362  -3.033   0.0027 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1307 on 235 degrees of freedom
Multiple R-squared:  0.1585,    Adjusted R-squared:  0.1513
F-statistic: 22.13 on 2 and 235 DF,  p-value: 1.565e-09


Value of test-statistic is: -4.4414 9.8631

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.46 -2.88 -2.57
phi1  6.52  4.63  3.81
```

Figure C7: ADF drift on seasonally differenced data

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-5860.4  -757.2   -43.9   710.8  4734.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -475.45640  190.91741  -2.490  0.01346 *
z.lag.1       -0.22554    0.04712  -4.787 3.01e-06 ***
tt             2.28677    1.28114   1.785  0.07556 .
z.diff.lag    -0.18152    0.06364  -2.852  0.00473 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1301 on 234 degrees of freedom
Multiple R-squared:  0.1698,    Adjusted R-squared:  0.1592
F-statistic: 15.95 on 3 and 234 DF,  p-value: 1.804e-09


Value of test-statistic is: -4.7866 7.6986 11.5477

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.99 -3.43 -3.13
phi2  6.22  4.75  4.07
phi3  8.43  6.49  5.47
```

Figure C8: ADF trend on seasonally differenced data

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 4 lags.

Value of test-statistic is: 0.7101

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Figure C9: KPSS mu on seasonally differenced data

```
######################
# KPSS Unit Root Test #
######################

Test is of type: tau with 4 lags.

Value of test-statistic is: 0.2652

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.119 0.146  0.176 0.216
```

Figure C10: KPSS tau on seasonally differenced data

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 4 lags.

Value of test-statistic is: 0.0249

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Figure C11: KPSS mu on differenced data

```
######################
# KPSS Unit Root Test #
######################

Test is of type: tau with 4 lags.

Value of test-statistic is: 0.0181

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.119 0.146  0.176 0.216
```

Figure C12: KPSS tau on differenced data

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-5745.0  -824.6   -34.2   774.3  4752.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.33386   86.36965   0.015 0.987691
z.lag.1     -1.58049    0.10236 -15.441  < 2e-16 ***
z.diff.lag   0.21868    0.06331   3.454 0.000655 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1330 on 234 degrees of freedom
Multiple R-squared:  0.6652,    Adjusted R-squared:  0.6623
F-statistic: 232.5 on 2 and 234 DF,  p-value: < 2.2e-16


Value of test-statistic is: -15.4406 119.2062

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.46 -2.88 -2.57
phi1  6.52  4.63  3.81
```

Figure C13: ADF drift on differenced data

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-5793.6  -807.5   -43.7   787.6  4811.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -73.13706  174.70209  -0.419 0.675866
z.lag.1      -1.58155    0.10255 -15.422  < 2e-16 ***
tt            0.62059    1.26482   0.491 0.624133
z.diff.lag    0.21902    0.06342   3.454 0.000657 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1332 on 233 degrees of freedom
Multiple R-squared:  0.6655,    Adjusted R-squared:  0.6612
F-statistic: 154.6 on 3 and 233 DF,  p-value: < 2.2e-16


Value of test-statistic is: -15.4225 79.2932 118.9397

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.99 -3.43 -3.13
phi2  6.22  4.75  4.07
phi3  8.43  6.49  5.47
```

Figure C14: ADF trend on differenced data

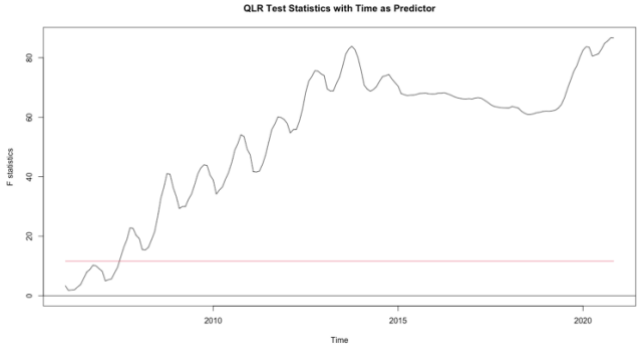

Figure C15: QLR test on crime data

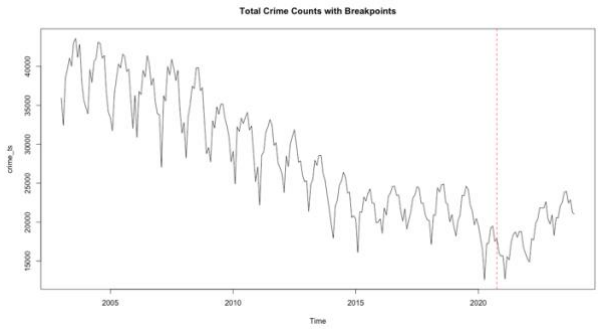

Figure C16: Breakpoints

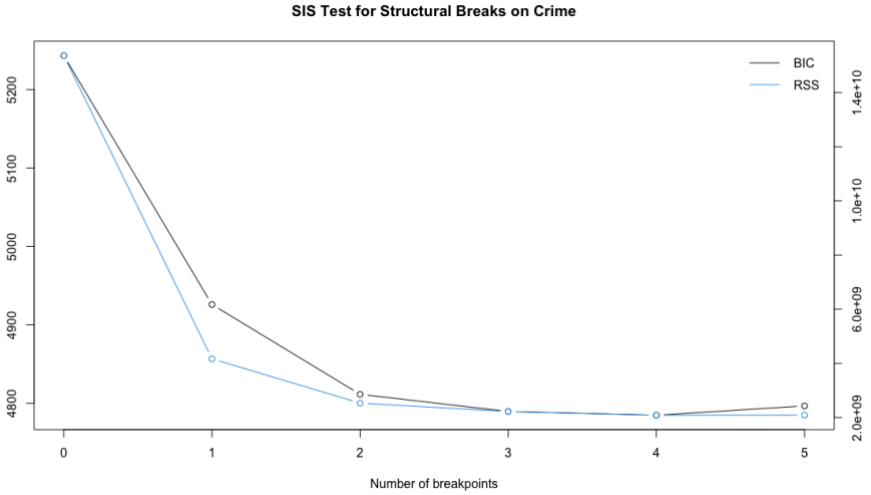

Figure C17: SIS test on crime data

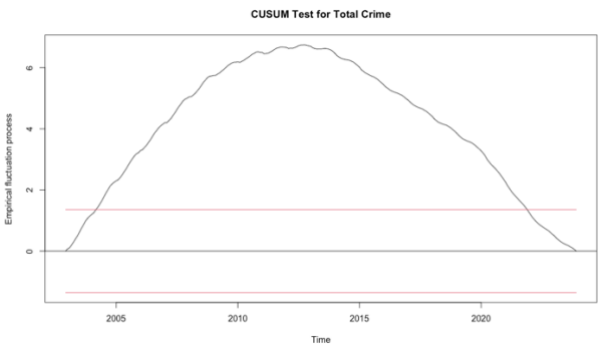

Figure C18: CUSUM test

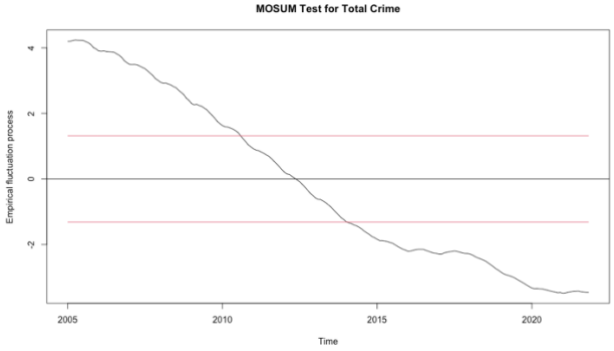

Figure C19: MOSUM test