Time Series Forecasting Methods for Analyzing Crime Data

Kyle Arbide

60172270026632916

CS6444 - Introduction to Big Data and Analytics

Spring 2022

1. Introduction

The collection and analysis of crime data is of significance for many police departments and federal agencies, due to its application towards improving public safety and quality of life. As the size of data grows across all industries, the scale and depth of this type of analysis is growing as well. The goal of some of the more recent crime data studies is to apply the newest and most effective predictive modeling strategies from Big Data, including deep learning and neural networks. The target for these Big Data models is to outperform the historical models that have been applied effectively in the past. Few studies have managed to reach this target, and thus the excitement towards discovering a potential breakthrough will fuel future researchers.

The two most important elements of crime data are time and space. Spatial analysis of crime data involves the differentiation of regions, to understand crime density and hotspots. The regions can vary in size from city level, to neighborhoods, all the way down to point level analysis. Understanding the spatial element to the data is crucial to crime analysis and forecasting, as different regions may contain different characteristics in terms of seasonality or trends. This paper will discuss how the spatial analysis has been used in some of the studies covered but will not go into detail on the intricacies of that type of analysis. Instead, the research from this study will surround the temporal element of crime analysis. First, research will be reviewed surrounding the statistics of time series analysis (seasonality, trend, and stationarity). Each of these plays a large role in any effective crime forecasting paper, and some interesting insights and patterns have been studied. Next, the discussion will cover the use of ARIMA forecasting models, their application towards crime forecasting, and potential benefits and value added for police departments and agencies. Finally, the results of the ARIMA forecast will be compared to those from a study that utilized an LSTM deep learning model, to understand how crime forecasting is evolving.

Research has shown that crime patterns not only vary across space but also across different crime types. Many analyses in crime forecasting perform studies using aggregated data across all crime types, which could be potentially misleading or uninformative. Different crime types have the potential to follow different trends and seasonal patterns. This paper will explore both violent and property crimes, to understand how their patterns may differ and to understand where it is and isn't appropriate to use aggregated data across the two. Excluded from the scope of this paper will be large scale crimes, such as cybercrimes and terrorism, and financial crimes, such as fraud and extortion.

The overall goal for this research is to establish the theoretical basis for crime forecasting. There is still a large amount of skepticism surrounding the ability to detect patterns and hotspots from crime data. This paper will look to determine if crime forecasting models have shown the ability to add value towards policing, and where they can continue to grow.

2. Background

The interest in researching crime data forecasting stems from its prevalence, potential benefits, and complexities. At first, people were skeptical at the application of forecasting models in crime, due to the uniqueness and randomness associated with most crimes. However, a few key breakthroughs around the 1980s established the opportunity associated with detecting crime patterns. One of the first effective crime forecasting studies was published in 1999 in the UK, the result of which led to better budget requests and resource deployment for the police department. In modern day, the purpose of crime forecasts is to support crime. There are also potential applications to predicting offenders and potential victims, but those topics won't be covered in this research.[1]

Presently, crime analysis and forecasting has cemented itself as a key component of policing. In 2017, a study found that 89% of police departments in the US employed a full or part time data analyst. The main area of focus for these analysts was criminal apprehension and hot spot identification. This was especially true for larger agencies, who are statistically more likely to adopt and maintain crime analysis. Even with this large scale application, many studies found that police departments are not using their analysts to develop advanced models. A repetitive cycle was found where managers would request basic, low effort work from their analysis. The results of this analysis would typically be unhelpful towards the prevention of crime, causing managers to lose faith in the data and creating a self-fulfilling prophecy. Issues like this are due partially to the lack of knowledge of the capabilities of advanced crime forecasting.[2]

---

[1] Wilpen Gorr and Richard Harries, "Introduction to crime forecasting," *International Journal of Forecasting* 19, no. 4 (October 2003): 551-553.
[2] Rachel Boba Santos, *Crime Analysis with Crime* Mapping (Los Angeles: SAGE Publications, 2017), https://books.google.com/books?id=RLmcDAAAQBAJ, 16-18.

Academic research plays a crucial role in the advancement of crime forecasting methodologies. Apart from the work of a select few analysts' at large agencies, most of the development of crime forecasting occurs in the research space. The lack of adoption and relative newness of the field means it is difficult to stay updated on the newest findings and that many issues remain unresolved. This paper aims to develop the readers understanding of the methods, benefits, and pitfalls of the most recent discoveries in the field of crime forecasting.

3 Time Series Analysis

3.1 Seasonality

Understanding seasonality of time series data is critical to any forecasting analysis. In some cases, the detection of seasonality in the data can be as simple as visualizing time series plots and noticing quarterly or monthly patterns. Autocorrelation plots can also be used as effective method of determining the seasonal lag in time series data. Occasionally more advance methods are needed, such as Chi-Square Goodness-of-Fit testing.[3]

Research dating back to the 1970s has theorized that changes in routine activities have an impact on crime rates. For example, if people are spending more time outside, there is more opportunity for burglary on empty homes, as well as more people at risk of becoming a victim of a violent crime. This suggests that during colder months, when people are more likely to stay indoors, both property and violent crime rates will drop. However, this has been found to be a gross generalization, as these trends have not been found universally. Other factors, such as children being out of school in the summer, seem to have just as much if not greater impact on crime rates. To get a better understanding of seasonality, and the different patterns that appear as a result of it, this section will review the seasonality of crime data across multiple cities.[4]

Andresen and Malleson researched the effects of seasonality in Vancouver CMA. At the time of the analysis, Vancouver had a population of nearly 550,000. The research was performed both with individual crime types as well as across all crime types. The seasonality in the dataset

[3] Eleazar Chukwunenye Nwogu, Iheanyi Sylvester Iwueze, and Valentine Uchenna Nlebedim, "Some Tests for Seasonality in Time Series Data," *Journal of Modern Applied Statistical Methods* 15, no.2, (November 2016): 382-385.

[4] Martin A. Andresen and Nicolas Malleson, "Crime seasonality and its variations across space," *Applied Geography* 43, (September 2013): 25-35.

is calculated by measuring the seasonal variation across dissemination areas and census tracts to understand if seasonal spatial patterns are unique to what is seen in the yearly aggregate data. The results for the assault and burglary crime types can be seen below in *Tables 1 & 2*.

	All Year	**Spring**	**Summer**	**Fall**	**Winter**
All Year		0.546	0.491	0.518	0.391
Spring	0.566		0.418	0.382	0.382
Summer	0.543	0.674		0.355	0.355
Fall	0.554	0.656	0.662		0.327
Winter	0.507	0.661	0.648	0.673	

Table 1: Indices of similarity, all year and seasons, census tracts and dissemination areas, Vancouver, 2001, assault

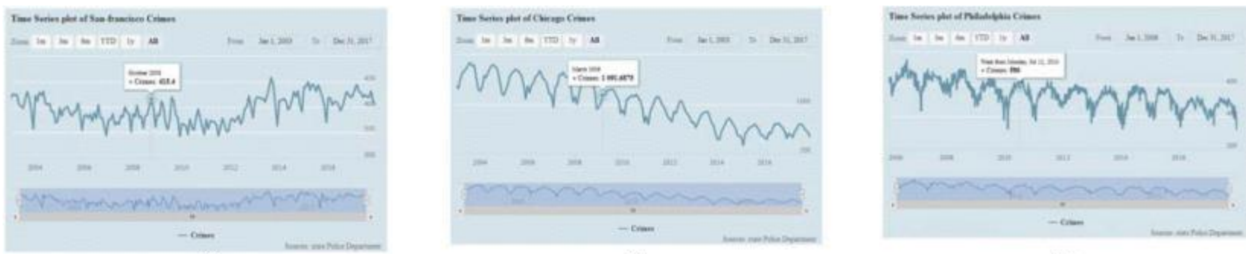
	All Year	**Spring**	**Summer**	**Fall**	**Winter**
All Year		0.355	0.436	0.382	0.391
Spring	0.425		0.336	0.264	0.264
Summer	0.458	0.373		0.327	0.327
Fall	0.415	0.335	0.363		0.336
Winter	0.427	0.380	0.369	0.360	

Table 2: Indices of similarity, all year and seasons, census tracts and dissemination areas, Vancouver, 2001, burglary

Note: Adapted from Andresen and Malleson, *Crime seasonality and its variations across space*, 30-31 Table 3 and 4.

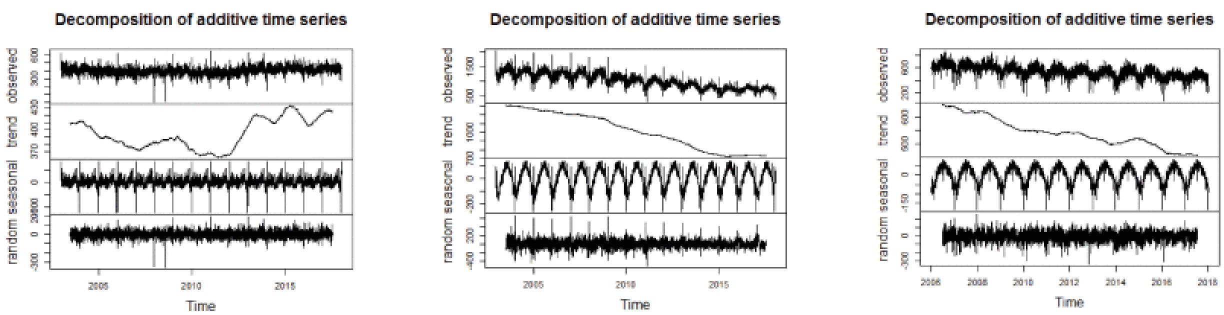
The authors set a threshold of 0.8 to determine a high degree of similarity. Looking at the tables above, none of the relations between season for both burglary and assault reach that threshold for similarity. The results suggest that yearly aggregated data is not a valid threshold due to the lack of similarity between the seasons. This conclusion challenges the use of aggregated yearly data in crime analysis as it does not constitute a representative sample of how crime rates vary over seasons.

Another study by Feng et al. performed seasonal decomposition before fitting an LSTM model on data from San Francisco, Chicago, and Philadelphia. Their analysis was performed across all crime types and with data from 2003 to 2017. The time series plots and decomposed time series plots in *Figures 1 and 2* show clear seasonality in the Chicago and Philadelphia. However, the San Francisco plot shows little to no seasonality. The authors link this difference to local climates, as San Francisco is the only one of these cities that maintains similar temperatures throughout the year.[5]



(a) (b) (c)

Figure 1: Time series plot of crime incidents for cities of (a) San-Francisco, (b) Chicago, and (c) Philadelphia



(a) (b) (c)

Figure 2: Decomposed time series to show how crime evolved over time in three cities of San Francisco (a), Chicago (b), and Philadelphia (c). For each original time series (top), we show the estimated trend component (2nd top), the estimated seasonal component (3rd top), and the estimated irregular component (bottom)

Note: Adapted from Feng et al. *Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data*, 106114-106117 Figures 2 & 8

[5] Mingchen Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data", *IEEE Access* 7, (July 2019): 106111-106123.

3.2 Trend

Along with the seasonal component, decomposing time series helps identify the trend related to the data. Identifying trend is a key component of fitting any time series model. For example, in the study above, Meng et al. identify the trend for all crime types in San Francisco, Chicago, and Philadelphia. The trend for each city can be seen in *Figure 2***.** Again, the data shows that crime trends are not consistent across cities, with San Francisco showing an upwards trend in recent years, and Chicago and Philadelphia showing a consistent downwards trend.

However, it is not clear that aggregating across all crime types is an accurate method of measuring trend. A 2010 study by Tseloni et al. sought out to analyze and compare trends across different crime type. This study is particularly interesting because the authors use an aggregation of data generated by sampling crime from countries all around the world. One of the goals of this analysis was to understand the relationship between the trends of different crime types, across both time and space. The study presents *Figure 3*, a time series plot of the internationally sampled data from 1988 – 2004. The authors observe from this figure that theft from car rates have dropped much more dramatically in recent years than assault, at a rate of 77.1% to 20.6% respectively. Another observation that was made is that crime rates have dropped uniformly across each of the displayed crime types after 1995.[6]

---

[6] Andromachi Tseloni et al., "Exploring the international decline in crime rates," *European Journal of Criminology* 7, no.5 (September 2010): 375–394.

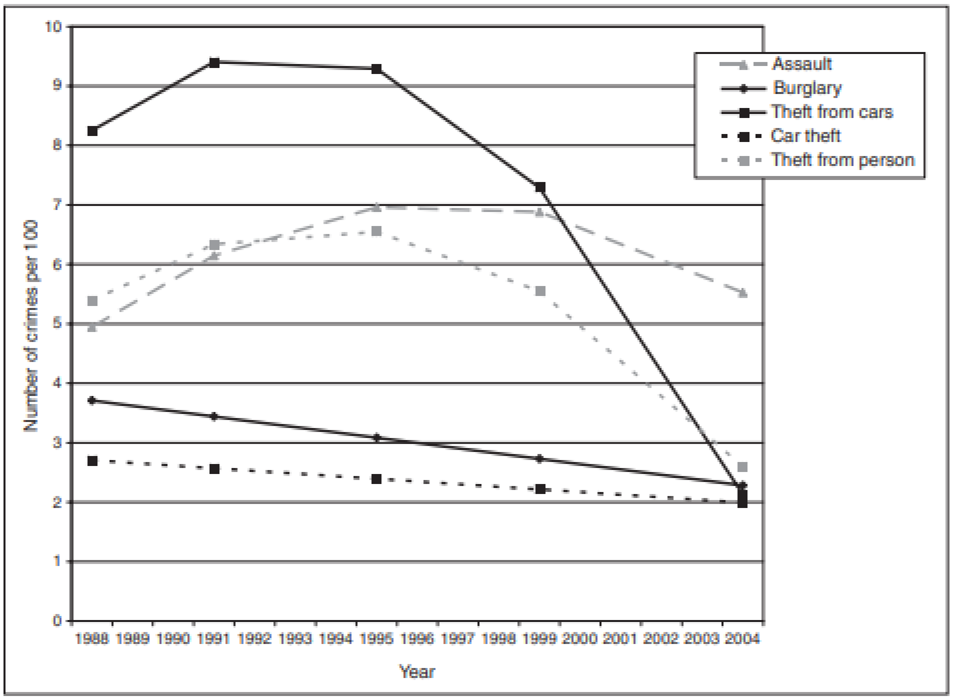


Figure 3: International trends in crime incidence, 1988-2004 (model estimates)

	Burglary	Theft from car	Car theft	Assault	Theft from person
Between-countries estimated correlation ($\hat{\rho}_{si}$) i,s = 1,2,…,5					
Burglary	1.00				
Theft from car	0.40 (0.53)▼	1.00			
Car theft	0.64 (0.70)▼	0.44 (0.59)▼	1.00		
Assault	0.58 (0.55)§	0.32 (0.13)*	0.85 (0.72)▼	1.00	
Theft from person	0.92 (0.87)▼	0.51 (0.74)▼	0.51 (0.71)§	0.38 (0.28)▼	1.00
Between-years estimated correlation ($\hat{\rho}_{si}$) i,s = 1,2,…,5					
Burglary	1.00				
Theft from car	0.43 (0.55)▼	1.00			
Car theft	0.30 (0.36)	0.23 (0.31)§	1.00		
Assault	0.44 (0.40)*	0.58 (0.50)▼	0.03 (0.07)	1.00	
Theft from person	0.48 (0.56)▼	0.60 (0.69)▼	0.09 (0.15)	0.45 (0.46)▼	1.00

Notes: Non-weighted data. Estimated correlation from multivariate multilevel models of crime incidence with linear trend for burglary and theft of car and non-linear trend for the other three crime types.
*.05 < *p*-value ≤ .10; § .005 < *p*-value ≤ .05; ▼ *p*-value ≤ .005

Table 3: Estimated correlations of crime incidence rates based on multivariate multilevel trend (and baseline) mode

Note: Adapted from Tseloni et al., *Exploring the international decline in crime rates*, 383 Figure 1 and Table 2

3.3 Crime Types

For both seasonality and trend, studies have shown that there are identifiable differences across crime types. The research by Tseloni et al. mentioned above goes on to measure the correlation of crime rates across countries and across years. The results in *Table 3* show that most crime types show at least some correlation. Some crime types, such as Burglary and Theft from person, show very strong correlation between countries, meaning countries with higher levels of one of these crimes is very likely to show higher levels of the other. Low correlation between years, which is seen between car theft and assault, suggests that the assault rate one year is independent to the car theft rate for that year. These conclusions raise the questions, how should crime types be treated during modeling and is the aggregation of data across crime types poor practice?

Research by Andresen and Linning focused on answering these questions exactly by analyzing the similarity of disaggregated crime types across space. They obtained data from Vancouver and Ottawa from 2001 and 2006 respectively. They transformed the data by obtaining a confidence interval for the percentage of crime that occurred on each street in the datasets across crime types. Then, the correlations across crime types are measured for both property and violent crimes. The results showed low correlation across violent and property crimes in Vancouver, and low correlation across violent crimes in Ottawa. The authors discovered high correlation between property crimes in Ottawa but attributed this to the fact that these crimes were condensed to only a few blocks. The results of the analysis confirm the speculation from earlier over the aggregating of crime types and its ineffectiveness in practice.[7]

Unfortunately, not every model referenced going forward will be using disaggregated data. Although it is best to have high granularity across crime types, this type of data is not always used. Lack of crime types in a study is typically an issue of data availability and confidentiality. As technology improves, higher quality crime data is becoming more readily available, and future studies should focus on performing their analysis across crime types.

[7] Martin A. Andresen and Shannon J. Linning, "The (in)appropriateness of aggregating across crime types," *Applied Geography* 35, no. 1-2 (November 2012): 275-282.

3.4 Stationarity

The last characteristic of time series analysis to mention is stationarity. The stationarity of time series data plays a large role in model selection and fitting. If time series data is determined to be non-stationary, it places constraints of which forecasting models can be properly applied. Stationarity is measured by taking the auto-correlation of the data using an ADF test. The auto-correlation of time series data is the level to which each value is correlated with each of its lag observations.

For crime data, most studies immediately assume that the data is non-stationary. Most authors don't take the extra time to prove non-stationarity before model fitting, but instead reference other literature that has discussed this. One study by Janko and Popli, however, does consider this before beginning their analysis. They perform the ADF test on each of the crime types in their dataset, the results of which are displayed in *Table 4*. They perform the test both with and without trend and concluded non-stationarity across all crime types. This confirms the assumption of non-stationarity that will be made in much of the research that is to be discussed.[8]

	No trend		With trend	
Variable	ADF statistic	*p*-Value	ADF statistic	*p*-Value
LTC	−2.188	0.2108	−2.220	0.4789
LVIO	−2.482	0.119	−1.600	0.7924
LPC	−0.744	0.8349	−2.837	0.1838
LBE	−0.337	0.9201	−2.329	0.4178
LFR	−0.699	0.8469	−2.841	0.1825
LATH	−1.446	0.5598	−1.694	0.7534
LROB	−3.227	0.0185	−3.224	0.0797
UEM	−2.558	0.1020	−3.030	0.1239

Table 4: Results of the ADF test: national level

Note: Adapted from Janko and Popli, *Examining the link between crime and unemployment*, 4015 Table 4

[8] Zuzana Janko and Gurleen Popli, "Examining the link between crime and unemployment: a time series analysis for Canada," *Applied Economics* 47, no. 37 (March 2015): 4007-4019.

3.5 Data Preprocessing

Each study describes their own approach to data preprocessing, and the most common steps taken are described in this section. First, nearly every study dealt with missing values in their data. Different strategies were applied to deal with the missing values, from imputation techniques to just removing the rows completely. Imputation strategies ranged from mean imputation to anticipating the missing values using k-nearest-neighbors. Date values in the data were also handled during preprocessing. One study claimed to have issues with incorrectly formatted dates, requiring them to drop those data points. Another study chose to parse the dates, creating 5 columns for year, month, day, hour, and minute for each crime. A more unique approach to preprocessing came from a study by Kim and Phillips, who were trying to understand the impact of COVID-19 on shootings. They chose to implement dummy variables for key events that could have impacted shooting rates. These events included the COVID-19 pandemic and Black Lives Matter protests.[9] Safat et al. used the bootstrap random sampling method to select features with low bias. Bootstrap random sampling requires dividing the data into three random subsets. From there, the random subsets are used to estimate the population parameters of the data.[10] What most studies failed to do during preprocessing is apply transformations to the data. We have already established that both trend and seasonality can be present in crime data. Trend and seasonal adjustments can be useful in transforming the nonstationary data into a more stationary form.[11] This can prove useful during time series modeling and forecasting.

[9] Dae-Young Kim and Scott W. Phillips, "When COVID-19 and guns meet: A rise in shootings," *Journal of Criminal Justice* 73, (March 2021): 4.
[10] Wajiha Safat et al, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," *IEEE ACCESS* 9, (May 2021): 70080-70094.
[11]Douglas C. Montgomery, Cheryl L. Jennings and Murat Kulahci, *Introduction to Time Series Analysis and Forecasting* (Hoboken: Wiley, 2015), 48.

3.6 Time Series Models for Forecasting

Once the data has been preprocessed and the statistical background of the time series data has been established, the next step of forecasting is model selection. A wide range of model types have been applied to crime data with the goal of predicting future trends in a given city. Different potential models include regression, exponential smoothing, support vector machines, and neural networks. However, the most popular method for crime data is ARIMA forecasting. The next section goes in depth on multiple studies than have fit ARIMA models. These results will then be compared to those achieved using LSTM models, to see if deep learning is the most effective technique for crime forecasting.

3.6.1 ARIMA

Autoregressive Integrated Moving Average (ARIMA) is a standard method of time series forecasting. It combines Autoregressive (AR) and Moving Average (MA) processes to build a more composite model. The parameters for an ARIMA model are (p, d, q) where p is number of lag observations for the AR, d is time difference between observations, and q is lagged observations of the MA. ARIMA models can also account for seasonality in seasonal ARIMA models.[12]

In 2020, Yim et al. set out to forecast homicide rates across the US. Using one year ahead prediction intervals, they split the data into training and test sets, with 1990 being the cut off. They specified prediction intervals of 50, 80, 90, and 98 percent, which would give an idea towards the efficacy of the model. As a comparison, they also performed the same ARIMA forecasting with other violent crime types, to provide more evidence that the model can generate effective forecasts.[13]

The model was run with a variety of parameters, seemingly on a trial and error basis, and an ARIMA(1,1,0) model was selected for having the lowest corrected AIC value. The autocorrelation function for the residuals returned between the desired thresholds and the authors

---

[12] Sima Siami-Namini, Neda Tavakoli and Akbar Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," *2018 17th IEEE International Conference on Machine Learning and Applications* (January 2019): 1394-1401.

[13] Ha-Neul Yi, Jordan R. Riddell and Andrew P. Wheeler, "Is the recent increase in national homicide abnormal? Testing the application of fan charts in monitoring national homicide trends over time," *Journal of Criminal Justice* 66, (January 2020): 5-10.

concluded that the model was adequate at forecasting crime over time. The test set of crime rates between 1991 and 2018 was then used to determine the performance of the model, and the results are shown in *Table 5*. The 50% prediction interval captured 43% of the true values, slightly less than it was expected to. The interesting part of this result is the model mostly overestimated the homicide rates, which the authors attributed to the steady decline starting in the 1990s.

The 80, 90, and 98 percent prediction intervals performed at about the level they were expected to, with the 80% interval still showing some overprediction. The other crime types that were used to endorse the use of the ARIMA model performed about the same, with the exception of aggravated assault, which performed subpar across all prediction intervals. The authors also experimented with changing the cut off for training and testing split. Interestingly, the model underperforms when the train cut off is between 1980 and 1994, but overperforms between 1995 and 2000. This is potentially because the crime rates appear to change less drastically in the 2000s but could also be explained by the fact the model no longer has to predict many years in the future.

Out of Sample, 1991–2018								
	Homicide		**Aggravated Assault**		**Robbery**		**Rape**	
	N	%	N	%	N	%	N	%
Coverage of 50% Intervals								
Low(25th)	12	43	23	82	17	61	17	61
Covered	12	43	4	14	11	39	11	39
High(75th)	4	14	1	4	0	0	0	0
Coverage of 80% Intervals								
Low(10th)	3	11	14	50	5	18	5	18
Covered	24	86	14	50	23	82	23	82
High(90th)	1	4	0	0	0	0	0	0
Coverage of 90% Intervals								
Low(5th)	2	7	9	32	3	11	3	11
Covered	25	89	19	68	25	89	25	89
High(95th)	1	4	0	0	0	0	0	0
Coverage of 98% Intervals								
Low(1th)	0	0	3	11	0	0	0	0
Covered	27	96	25	89	28	100	28	100
High(99th)	1	4	0	0	0	0	0	0
Total	28		28		28		28	

Table 5: Coverage Rates for the 50%, 80%, 90%, and 98% prediction intervals

Note: Adapted from Yi, Riddell and Wheeler, *Is the recent increase in national homicide abnormal?*, 6 Table 2

The main takeaway from this study can be found when looking at the point prediction for the 2015 homicide increase. This event occurs outside the 90% prediction interval from the ARIMA model. This result along with results of prior studies suggests that the spike in homicide rates from 2015 is an anomaly, and that the trend is likely to be downwards in the future. Models run on the other violent crimes in the dataset confirmed the same result, with the exception of rape which continues to have an upwards trend after a surge in 2015. The work from this study is a prime example of how value can be added from analysis and forecasting of Big Data in crime. Percent change statistics would imply that violent crime rates have been up in the last 5 years, which is true. However, it takes a more sophisticated statistical model to see that this is not representative of the way crime is trending and will likely continue to trend.

The next study was conducted by Nitta et al in 2018, with the goal of forecasting Chicago crimes. The motivation of the paper is to create an accurate forecast that can help police spend and distribute resources more efficiently. They us a wide variety of machine learning strategies throughout their research, including k-nn for imputation, LASSO estimation for feature selection, and Naïve Bayes for performance measurement. This discussion, however, will only go in detail over their forecasting methods, which includes fitting an ARIMA model.[14]

This study does not provide as much detail into the process of fitting their model as the last that we discussed. In fact, the authors do not even provide the parameters they applied to their final ARIMA model. Instead, they display a chart that compares the performance of their model to other time series forecasting models. It is unclear where they received the performance metrics for the other models in the figure. *Figure 4* is the visual representation of this comparison between models. The results claimed that the ARIMA model fit in this study outperformed support vector machines, kernel density estimation, and deep neural network at crime forecasting.

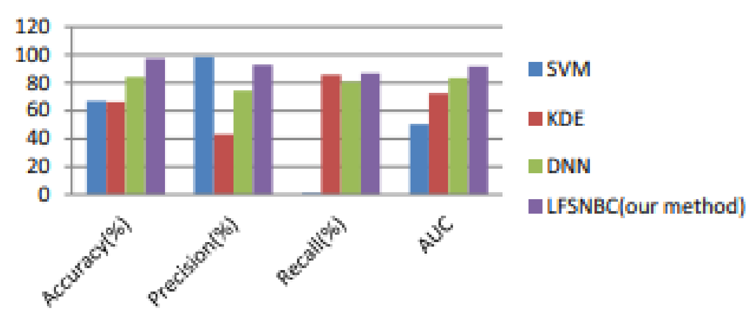


Figure 4: Comparisons performance with different algorithms

Note: Adapted from Nitta et al., *LASSO-based feature selection and naïve Bayes classifier for crime prediction and its type*, 194 Figure 4

There is very little detail used in this study throughout the model selecting and fitting section. The authors shared few insights about the process of fitting their ARIMA model and

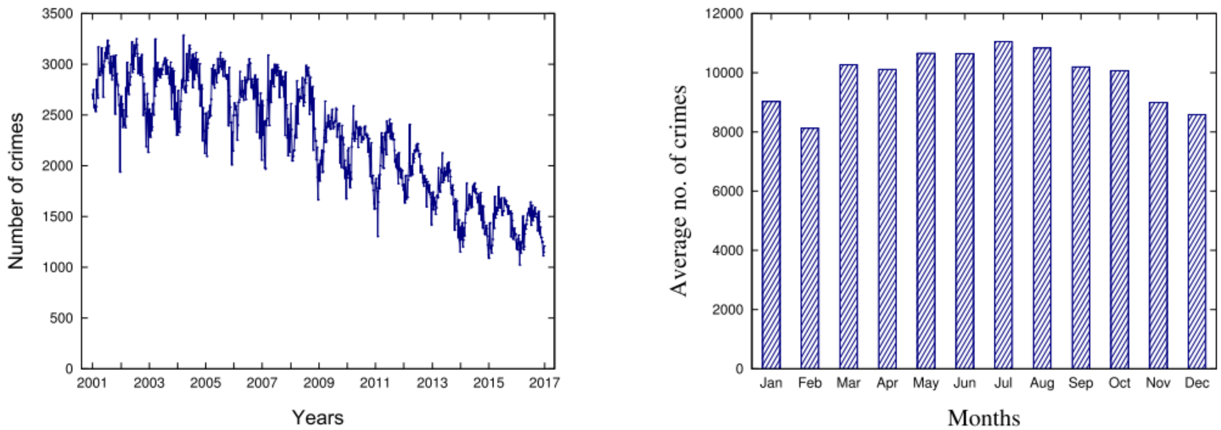
[14] Gnaneswara Rao Nitta et al., "LASSO-based feature selection and naïve Bayes classifier for crime prediction and its type," *Service Oriented Computing and Applications* 13, (July 2019): 187-197.

share even less about the algorithms with which their model was compared. It is not even clear what period they were using during forecasting. Although some of the strategies used by the authors in this research are intriguing, the lack of clarity into the model fitting and results raises skepticism over the claims of developing a highly effective crime forecasting model.

In 2019, Catlett et al. took a spatio-temporal approach to crime prediction in New York City and Chicago. The data for Chicago ranged from 2001 to 2017, and for New York from 2006 to 2017. The spatial element of their approach included clustering the regions in each data set based on their crime density. For this step, they applied K-means clustering to the regions. Each cluster was then split into its own dataset, and the timeseries forecasting was applied across them.[15]

The authors chose to perform their forecasting by fitting a seasonal ARIMA, or SARIMA, model. A seasonal ARIMA model is similar to the normal ARIMA model, except with the inclusion of additional seasonal terms. This decision came after analyzing the time series plots which are shown in *Figures 5 and 6.* The time series plots appear to have clear seasonality. They observed that the Autumn and Winter times of year typically would see the lowest number of crimes, which aligns with what was discussed before. For fitting the model, the authors chose 2013 for their train-test split for both cities.

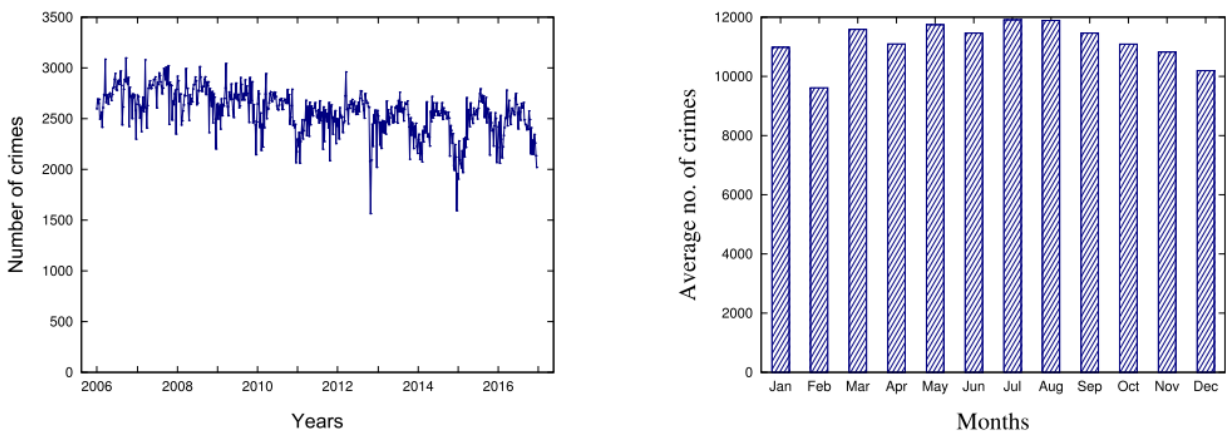
[15] Charlie Catlett et al., "Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments," *Pervasive and Mobile Computing* 53, (February 2019): 62-74.



(a) N. of crimes vs time

(b) Distribution by month

Figure 5: CHI crime data: number of crimes vs time and their distribution by month



(a) N. of crimes vs time

(b) Distribution by month

Figure 6: NYC crime data: number of crimes vs time and their distribution by month

Note: Adapted from Catlett et al., *Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments*, 67-71 Figures 4 and 8

The ARIMA approach was then applied to both cities, starting with Chicago. The Chicago dataset had been divided into three crime regions with varying densities. For each region, different parameters of the ARIMA model were set. Each model used a periodicity of 52 weeks and was tested against a one, two, and three years ahead prediction horizion. The results are shown in *Table 6* for each crime dense region. Comparing MAE measurements to those from a model which included the entire area of the dataset, they note that the spatially split model preformed significantly better across every region. The MAPE measurement shows a maximum forecasting error of 11.90%, 9.63%, and 18.66% for years one, two, and three respectively. Overall, the model seems the generate strong forecasts.

Time	MAE				MAPE			
	Area	CDR1	CDR2	CDR3	Area	CDR1	CDR2	CDR3
2014	88.86	30.20	14.47	11.15	6.19	8.68	10.86	11.90
2015	74.54	28.24	12.13	9.24	5.42	7.60	8.94	9.62
2016	81.47	31.04	12.86	13.83	6.29	10.14	9.99	18.66
Time	ME				RMSE			
	Area	CDR1	CDR2	CDR3	Area	CDR1	CDR2	CDR3
hline 2014	−62.96	30.19	−8.36	−2.67	108.34	35.98	19.01	13.57
2015	−27.05	4.75	−1.80	−1.57	97.77	34.61	14.98	11.43
2016	−48.77	−24.94	−5.36	−11.02	115.16	38.31	15.95	16.66

Table 6: Forecast error measures vs years, for the whole area and the top three largest crime dense regions in Chicago

The next step is to see if these results can be replicated on the New York City dataset. Again, the data is through K-means clustering into 3 regions of varying crime densities. Each region is fit with its own parameters, but a consistent 52 week periodicity. The data is tested against a one, two, and three year forecasting horizion, and the results are shown in *Table 7*. The MAE measurements confirm the results seen from the Chicago dataset, as the forecast for each crime dense region performs better than for the entire dataset. The authors also noted that the ARIMA forecasting method performed the best of the models they had attempted, which included random forest, REPTree, and zero rule.

Time	MAE				MAPE			
	Area	CDR1	CDR2	CDR3	Area	CDR1	CDR2	CDR3
2014	135.30	52.15	10.56	12.46	6.17	7.26	8.15	12.59
2015	141.68	51.51	20.86	11.06	6.04	7.42	14.95	9.87
2016	117.49	47.85	25.05	11.80	4.79	6.41	16.64	10.58
Time	ME				RMSE			
	Area	CDR1	CDR2	CDR3	Area	CDR1	CDR2	CDR3
2014	−80.05	−41.17	2.11	−6.55	184.81	70.99	14.23	14.73
2015	−0.6	−28.57	15.63	1.51	177.08	66.08	24.60	14.45
2016	31.21	−8.00	23.84	2.68	143.73	57.80	29.85	16.19

Table 7: MAE, MAPE, ME, and RMSE prediction errors vs years, for the whole area and the top three largest crime dense regions in New York City

Note: Adapted from Catlett et al.**,** *Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments*, 69-73 Tables 3 and 5

The knowledge to be gained from this study is the importance of spatial granularity in the time series analysis of crime data. The model showed much stronger performance when the data had been preprocessed into regions with similar crime densities. The study also confirms strength of the ARIMA model in time series forecasting, especially in crime. One shortcoming of the study, however, is their failure to account for crime types in the data. This could have been influential in both the clustering and forecasting steps, where regions could be clustered based on crime levels across crime types, and forecasting could be performed for different crime types. Still, this resource can be used to generate a strong argument for the inclusion of spatial clustering in crime forecasting.

The final study for this section was performed by Payne and Morgan in 2020. Payne and Morgan were interested in studying and forecasting the affect of the COVID-19 outbreak on property crime rates. They performed their analysis over multiple types of property crimes, including property damage, shop theft, burglary, and motor vehicle theft. They compared their forecasted results to the true values for the beginning months of the epidemic to understand the impact on different crime types. Their data captured monthly crime rates in Queensland, Australia, between February 2014 and March 2020.[16]

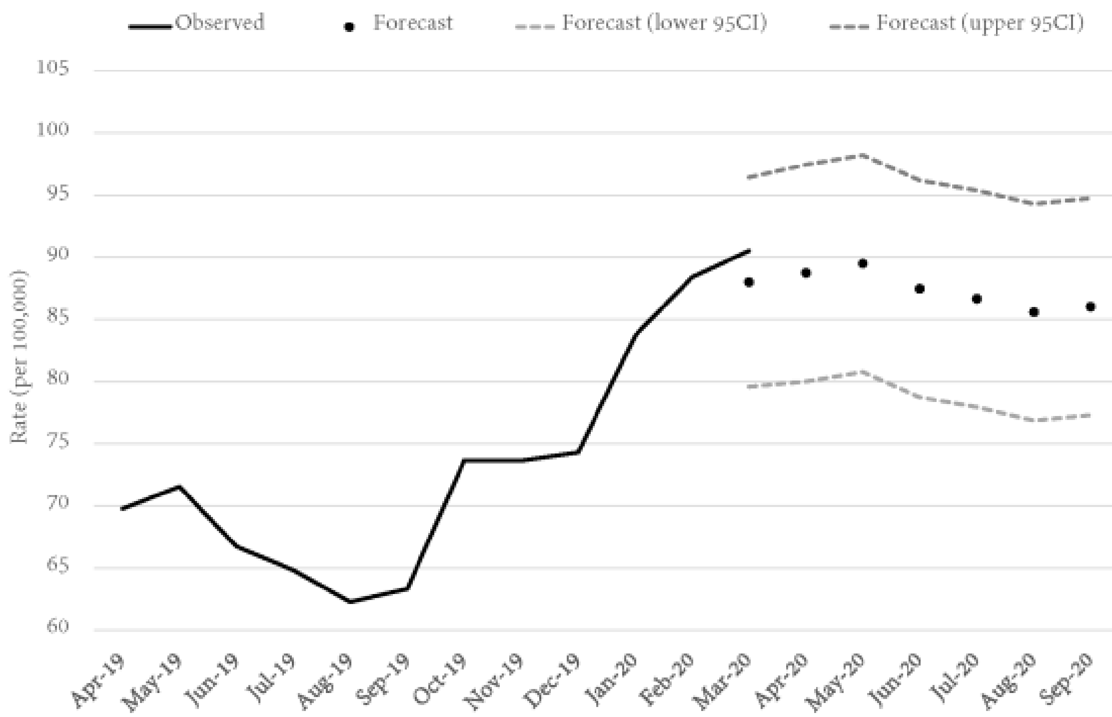
Payne and Morgan are very thorough in performing the required time series statistical tests before selecting a model. For each crime type, they apply the ADF test to check for non-stationarity in the data. From here, they use auto correlation, partial auto correlation, moving average, and auto regression to determine the ARIMA parameters that are appropriate for that crime type. After performing those steps, they can confidently calculate the forecasting confidence interval, and compare the point prediction for March 2020 to the true value. They repeated this analysis for 6 different crime types, but this discussion will only describe those results from burglary, and motor vehicle theft.

Burglary rates in Queensland have fluctuated between 50 and 90 incidents per 100,000 population since 2014. In March of 2020, this was up to a rate of 90.5 incidents per 100,000. The AC and PAC tests indicated the need for both seasonal and trend parameters, so a seasonal ARIMA model was fit. The results of this seasonal ARIMA model can be seen in *Figure 7*. The forecasted value for burglary rates in March 2020 fell only slightly below the true value,

---

[16] Jason L. Payne and Anthony Morgan, "Property Crime During the COVID-19 Pandemic: A Comparison of Recorded Offense Rater and Dynamic Forecasts (ARIMA) for March 2020 in Queensland, Australia", (May 2020): 5-10.

suggesting there may have been little impact as a result of the pandemic restrictions. The model also predicts that the burglary rates will continue going up into the later months of the pandemic. This seems to follow the upwards trend of burglary rates that appears to have started in August 2019, so it would be incorrect to assume that the pandemic had caused an increase in burglaries. The true value also did not fall outside the 95% confidence interval for March 2020.



Source: Queensland offence rates, Open Data Portal

Figure 7: Short-term BURGLARY rate forecasts (actual rates and ARIMA forecasts)

Motor vehicle theft historically occurred at a much lower rate, with between 14 and 32 instances per 100,000 population. Again, the true value for motor vehicle theft rate in March 2020 was up from this historical range, with a rate of 38.8 incidents per 100,000. The AC and PAC plots suggested the need for seasonal and trend considerations once again, and ultimately the same parameters were specified as were for the burglary model. This time however, the model forecasted a higher rate of motor vehicle theft than what had actually occurred, but the true value still fell within the confidence interval. The trend of motor vehicle theft rates, seen clearly in *Figure 8*, had already been rising before the pandemic, so again it is impossible to attribute this rise or future rises to pandemic restrictions.

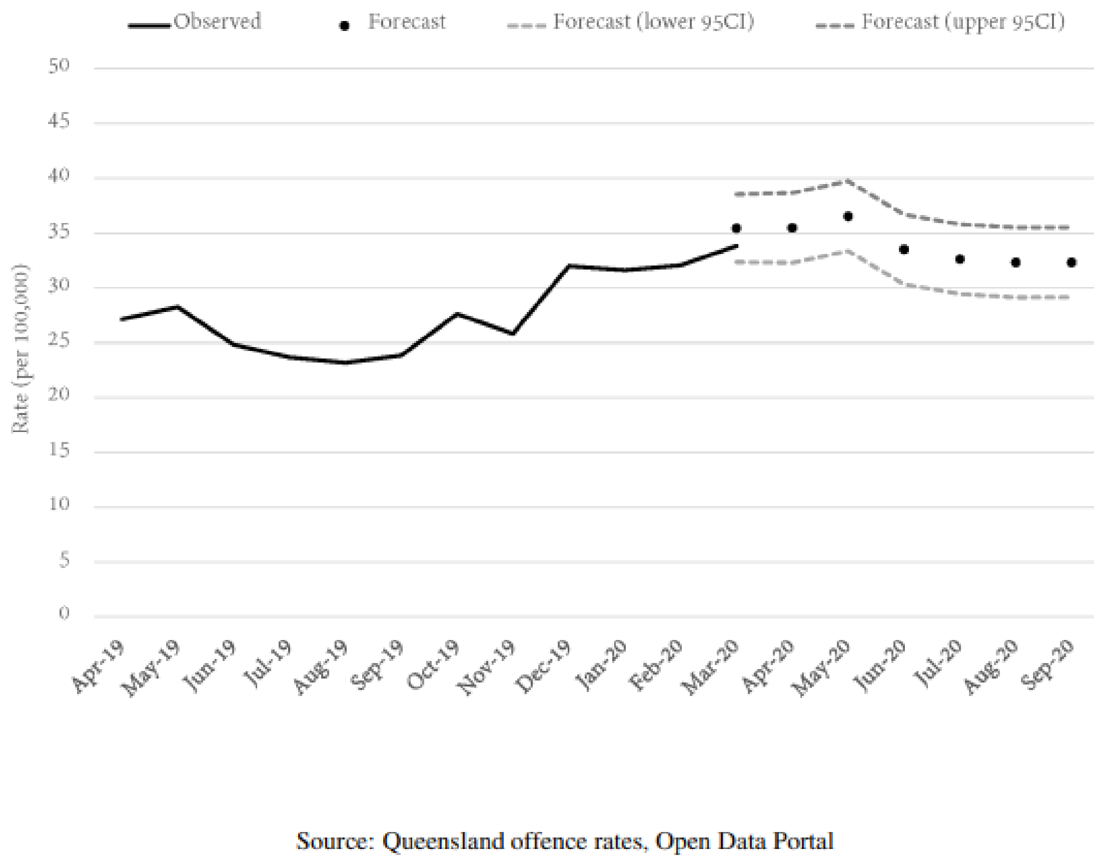


Figure 8: Short-term MOTOR VEHICLE THEFT rate forecasts (actual rates and ARIMA forecasts)

Note: Adapted from Payne and Morgan, *Property Crime During the COVID-19 Pandemic: A Comparison of Recorded Offense Rater and Dynamic Forecasts (ARIMA) for March 2020 in Queensland, Australia*, 9-10 Figure 11 and 13

The study was able to find statistically significant drops in shop stealing, fraud, and other theft rates using this method. The crime types mentioned above did not see significant change but applying this methodology to more recent data may reveal different information. This case provides excellent insight about the importance of calculating time series statistics before setting ARIMA parameters, as well as the effectiveness of performing forecasting analysis across crime types.

The studies in this section that performed the best at generating accurate forecasting results, drawing valid conclusions, and contributing to domain knowledge were those that were granular will their model fitting process and critical in the analysis of the forecasting results. The study by Catlett et al. showed how understanding spatial crime data can lead to more effective modeling within a city. Payne and Morgan showed that modeling can be applied to different crime types within a city and doing so can help achieve more detailed results into the crime trends.

3.6.2 LSTM

Long short term memory is a type of RNN. It is applicable to time series because it has the ability to remember values from earlier stages. It is capable of memorizing sequences in the data and leveraging that information for forecasting. This section will look at the application of an LSTM model in crime forecasting with the goal of concluding if deep learning models can generate valuable insights, like those seen in the ARIMA model.[17]

Feng et al. performed a study in 2019 that looked at crime data in San Francisco, Chicago, and Philadelphia. The goal of their analysis was to generate insight that could prove valuable to police in their distribution of resources. The data was aggregated monthly and contained records from between 2003 and 2017 for both San Francisco and Chicago, and between 2006 and 2017 for Philadelphia. The authors performed some spatial analysis, but none was included in their forecasting models. They also conducted some exploratory analysis on crime types, but still performed their prediction with aggregated data across all crime.[18]

Without going into too much detail on the specifics of each model that was fit during the study, the LSTM model generated results that were comparable to other models in Chicago, and better than other models in Philadelphia. These results are shown in *Table 8*. The model performed best with 3 years of training data across every city. Optimization of the parameters for the LSTM model was also performed and were computed to be 50 for layers and 60 for cell state.

The results from this study show that not only can LSTM be applied to crime data and generate useful forecasts, but also in some cases it can perform better than traditional methods. This example used aggregated spatial and crime type data for the cities that were analyzed,

---

[17] Siami-Namini, Tavakoli and Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," *2018 17th IEEE International Conference on Machine Learning and Applications*, 1398.
[18] Mingchen Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," *IEEE Access* 7, (July 2019): 106118-106121.

which has been proven not to be best practices. This suggests that there is still opportunity for research into deep learning forecasts across higher levels of crime type and spatial granularity.

TABLE 1. Comparison of different algorithms/models in terms of RMSE and spearman correlation under different sizes of training samples.

city	Years for training	RMSE-Prophet	Correlation-Prophet	RMSE-LSTM	Correlation-LSTM	RMSE-Neural Network	Correlation-Neural Network
San-Francisco	10	38.21	0.384	55.21	0.354	54.79	0.097
San-Francisco	5	35.70	0.402	54.18	0.365	48.04	0.232
San-Francisco	4	36.18	**0.415**	53.96	0.411	48.04	0.236
San-Francisco	3	**35.65**	0.398	**45.65**	**0.423**	41.62	**0.291**
San-Francisco	2	91.93	0.087	160.2	0.098	**41.17**	0.128
San-Francisco	1	100.56	0.182	95.96	0.122	-	-
Chicago	10	76.89	0.560	77.01	0.532	77.19	0.367
Chicago	5	68.21	0.652	69.12	0.549	92.74	0.551
Chicago	4	66.75	0.654	67.45	0.612	88.04	0.492
Chicago	3	**66.68**	**0.658**	**67.15**	**0.625**	**75.06**	0.505
Chicago	2	67.42	0.632	68.14	0.576	75.90	**0.552**
Chicago	1	100.51	0.02	78.98	0.459	-	-
Philadelphia	10	51.83	0.716	50.65	0.709	82.21	0.422
Philadelphia	5	56.07	0.728	51.23	0.698	71.19	0.486
Philadelphia	4	55.37	0.728	49.22	0.714	67.74	**0.588**
Philadelphia	3	**48.73**	**0.729**	**48.15**	**0.725**	**63.68**	0.537
Philadelphia	2	50.35	0.718	57.16	0.705	170.12	0.128
Philadelphia	1	100.71	0.098	140.63	0.562	-	-

Table 8: Comparison of different algorithms/models in terms of RMSE and spearman correlation under different sizes of training samples

Note: Adapted from Feng et al. *Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data*, 106119 Table 1

4 Conclusion & Future Work

Big Data analysis is becoming an important tool towards the understanding of crime trends. It provides the potential to add value towards police departments and local governments due to its application in resource distribution and policy. Time series analysis is the most relevant of the big data tools for crime, especially forecasting methods. The key characteristics of a proper time series analysis were discussed in this paper, with examples being provided from research in the crime industry. Seasonality was seen throughout most of the studies presented in this report. One study even concluded that the time of year has a measurable impact on certain crime rates. Its importance towards model selection and time series decomposition was also established. Trend is potentially the most important time series characteristic in crime analysis. Distinguishing between anomalous changes and long term trend is important to understanding how crime is evolving in a given area. Stationarity was found to be often overlooked in crime related time series studies. It is an important element to establish before selecting a forecasting method. Future work in crime analysis should focus on establishing these characteristics as they apply to the crime data they are referencing, as those studies who did seemed to arrive at the most accurate results.

Multiple forecasting methods were also described, with the main focus being on ARIMA, seasonal ARIMA, and LSTM models. These models are the most popular for the time series forecasting of crime data and have been referenced in many more studies that were not mentioned here. Few studies went into detail with their logic behind setting the parameters for their ARIMA model. Still, a majority of the time ARIMA models proved to be effective and assist the researcher in arriving at their goal. As for the LSTM model, the study referenced here believes it has application in crime forecasting. A wider variety of LSTM models should be studied to determine its true value, and if it should be preferred to the ARIMA method.

Spatial characteristics were a recurring quality of the crime data that typically influenced the nature of an analysis. Most of these studies were performed with crime data from different cities, which is why patterns for trend and seasonality could vary drastically from study to study. When performing a crime analysis or forecast, it is important to align the work with studies applied to the same city's crime data, as patterns and qualities of the time series data is not necessarily consistent across cities. Even within cities, different crime patterns and densities were noticed. When different regions within a city were treated separately, it was found to have a positive influence on the ability to forecast crime rates. Further research into this subject should look into the different methods applied towards spatial separation and clustering of crime data.

The separation of crime types also proved to be influential. Some of the studies referenced used aggregated data across all crime types, although other research has suggested that this may be incorrect. Those studies that did separate their crime types were able to arrive at multiple conclusions and achieve a much more granular understanding of how crime is evolving in the described area. Future studies have the potential of deepening the understanding of the relationship between crime rates across crime types. The independence or dependence between property and violent crimes should be analyzed before aggregating the types together for a forecasting model.

Overall, crime forecasting methods have already shown their ability to provide value, but additional focus on a few key time series and crime elements could lead to more thorough and accurate studies. These methods combined with the potential benefit provided from deep learning model outlines the exciting future of crime forecasting.

References

Andresen, Martin A. and Linning, Shannon J. "The (in)appropriateness of aggregating across crime types." *Applied Geography* 35, no. 1-2 (November 2012): 275-282.

Andresen, Martin A. and Malleson, Nicolas. "Crime seasonality and its variations across space." *Applied Geography* 43, (September 2013): 25-35.

Catlett, Charlie et al. "Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments." *Pervasive and Mobile Computing* 53, (February 2019): 62-74.

Feng, Mingchen et al. "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data." *IEEE Access* 7, (July 2019): 106111-106123.

Gorr, Wilpen and Harries, Richard. "Introduction to crime forecasting." *International Journal of Forecasting* 19, no. 4 (October 2003): 551-555.

Janko, Zuzana and Popli, Gurleen. "Examining the link between crime and unemployment: a time series analysis for Canada." *Applied Economics* 47, no. 37 (March 2015): 4007-4019.

Kim, Dae-Young and Phillips, Scott W. "When COVID-19 and guns meet: A rise in shootings." *Journal of Criminal Justice* 73, (March 2021): 1-10.

Montgomery, Douglas C, Jennings Cheryl L, and Kulahci, Murat. *Introduction to Time Series Analysis and Forecasting.* Hoboken: Wiley, 2015.

Nitta, Gnaneswara Rao et al. "LASSO-based feature selection and naïve Bayes classifier for crime prediction and its type." *Service Oriented Computing and Applications* 13, (July 2019): 187-197.

Nwogu, Eleazar Chukwunenye, Iwueze, Iheanyi Sylvester, and Nlebedim, Valentine Uchenna. "Some Tests for Seasonality in Time Series Data." *Journal of Modern Applied Statistical Methods* 15, no.2, (November 2016): 381-399.

Payne, Jason L. and Morgan, Anthony. "Property Crime During the COVID-19 Pandemic: A Comparison of Recorded Offense Rater and Dynamic Forecasts (ARIMA) for March 2020 in Queensland, Australia." (May 2020): 1-32.

Safat, Wajiha et al. "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques." *IEEE ACCESS* 9, (May 2021): 70080-70094.

Santos, Rachel Boba. *Crime Analysis with Crime* Mapping. Los Angeles: SAGE Publications, 2017. https://books.google.com/books?id=RLmcDAAAQBAJ.

Siami-Namini, Sima, Tavakoli, Neda, and Namin, Akbar Siami. "A Comparison of ARIMA and LSTM in Forecasting Time Series." *2018 17th IEEE International Conference on Machine Learning and Applications* (January 2019): 1394-1401.

Tseloni, Andromachi et al. "Exploring the international decline in crime rates." *European Journal of Criminology* 7, no.5 (September 2010): 375–394.

Yi, Ha-Neul, Riddell, Jordan R, and Wheeler, Andrew P. "Is the recent increase in national homicide abnormal? Testing the application of fan charts in monitoring national homicide trends over time." *Journal of Criminal Justice* 66, (January 2020): 1-11.