 doi:10.20944/preprints202102.0172.v1

Article

Spatio-Temporal Crime Predictions by Leveraging Artificial Intelligence for Citizens Security in Smart Cities

Umair Muneer Butt [1,†,‡], Sukumar Letchmunan [1,‡], Fadratul Hafinaz Hassan [1], Mubashir Ali [2], Anees Baqir [3], Tieng Wei Koh [4], and Hafiz Husnain Raza Sherazi [5,‡]

[1] School of Computer Sciences, Universiti Sains Malaysia, Malaysia; umair@student.usm.my
[2] Dept. of Management, Information and Production Engineering, University of Bergamo, Italy; mubbashircheema@gmail.com
[3] Department of Environmental Sciences, Informatics and Statistics, Ca'Foscari University of Venice; anees.baqir@unive.it
[4] Deparment of Software Engineering and Information System, Universiti Putra Malaysia, Malaysia; twkoh@upm.edu.my
[5] Tyndall National Institute, University College Cork, Cork, Ireland ;sherazi@tyndall.ie
* Correspondence: sherazi@tyndall.ie; umair@student.usm.my; sukumar@usm.my; fadratul@usm.my

Abstract: Smart city infrastructure has a significant impact on improving the quality of humans life. However, a substantial increase in the urban population from the last few years is posing challenges related to resource management, safety, and security. In order to ensure the safe mobility and security in the smart city environment, this paper proposes a novel Artificial Intelligence (AI) based approach empowering the authorities to better visualize the threats and to help them identifying the highly-reported crime zones yielding a greater predictability of crime hot-spots in a smart city. To this end, it first investigates the *Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)* to detect the hot-spots that have a higher risk of crimes to be committed. Second, for crime prediction, *Seasonal Auto-Regressive Integrated Moving Average (SARIMA)* exploited in each dense crime region to predict the number of crimes in future with spatial and temporal information. The proposed HDBSCAN and SARIMA based crime prediction model is evaluated on ten years of crime data (2008-2017) for *New York City (NYC)*. The accuracy of the model is measured by considering different time period scenarios i.e. (a) year-wise, i.e., for each year and (b) for whole period of ten years, using 80:20 ratio where 80% data was used for training and 20% data was used for testing. The proposed approach outperform with an average Mean Absolute Error (MAE) of 11.47.

Keywords: Citizen Security; Smart Cities; Crime Prediction; Artificial Intelligence; Safe City

1. Introduction

The primary objective of the smart city is to improve the quality of life in the city by efficient utilization of the city resources. The unprecedented transformation of urban areas has a significant effect on the social and economic development of the cities [1]. Because of technological advancements, smart cities infrastructure has been introduced that mainly focus on the quality of citizen life, better management of urban population issues, and sustainability in every aspect of their life [2,3]. Smart cities have empowered human life by exploiting technology to address socio-economic challenges such as education, health, transportation, economy, and public safety. However, the increased population in cities is posing challenges such as resource planning, public safety, and an enormous amount of data generated from sensors, cameras and, tracking devices [4].

To achieve smart city goals, collaboration is required among researchers, technology developers, government officials, industry, and citizen to present and develop ideas to cope with smart city challenges. One of the crucial challenges is to provide a secure and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

safe environment [5]. The availability of the enormous amount of data from the past few years has motivated the researchers to pursue research in the area of crime and criminal investigations by studying the crime trends and patterns and trying to make effective policies for better and peaceful communities [6,7].

Crime is a disorder in behavior, and it is a complex phenomenon of multiple dimensions closely associated with different diverse factors such as spatial, temporal, societal and ecological.Considering the crime trends and patterns of society is a critical issue while making decisions to relocate to a new city or avoid travelling to locations and places that have safety issues [8]. Based on historical data, forecasting crimes has been a subject of interest which gained much attention in research, hence resulted in proposing the significant number of different methods for the discovery of diverse aspects related to crime prediction [9].

For analyzing the data of crime to find the information regarding different perspectives of a crime event being occurred, data mining techniques have been proposed [10] e.g. association rule mining [11], classification [12], and clustering [13].Among other challenges being faced due to growth in urbanization, crime spikes with respect to seasonality is becoming a challenging and significant social problem [14]. Figure 1 shows the crime spikes over the years in the city of New York [15].

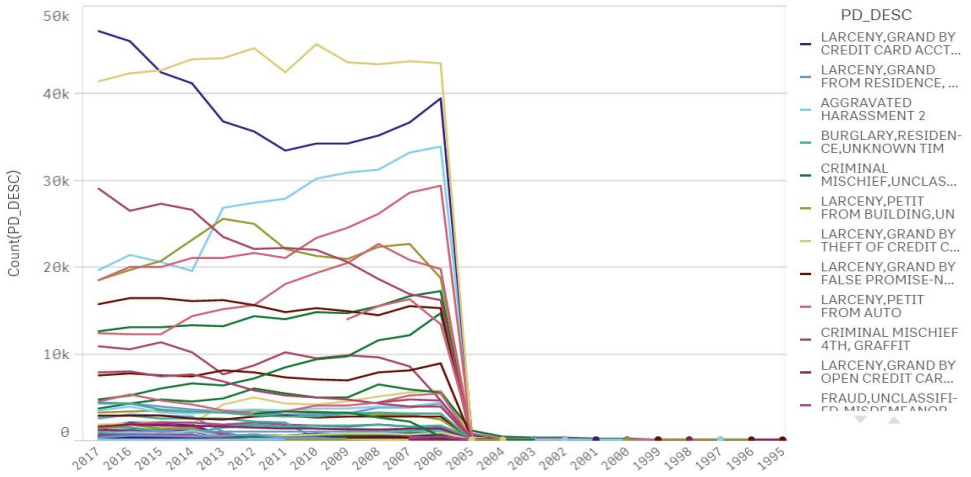

Figure 1. Crime Spikes Over the years in NYC.

In order to over come the crime spiking, effective policies needs to be in place. For this purpose, predictive policing is the one of the first which is based on using statistical predictions and techniques to analyze the likely future crime scenes or number of crimes [16].Higher crime rates increased complexities, which led to the new technologies helping police to analyze and understand crime trends and patterns [17]. Hence, there is a dire need to identify the areas which pose greater crime threats than other regions. In this regard, mapping of crime dense areas has emerged as a sublime analytical method for the identification of areas with greater crime risk for efficient and effective allocation of police resources [18].

Because of the exponential growth in data, it requires sufficient resources to process and get the required results. Therefore, among other goals, one of the main task was to utilize such clustering algorithm which requires less resources as compared to the baseline study [17] which used Density Based Spatial Clustering of Applications with Noise (DBSCAN). Specifically, our study aims to contribute towards achieving the following goals.

1. Given a raw dataset, perform pre-processing techniques to convert it into a form so that the models for hot-spot detection and crime prediction can be applied on it.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

2. From the processed dataset, detect a set of dense crime regions where each region is a spatial area where events of criminal activity have occurred with density higher than other locations.
3. Predict number of crimes in each dense crime region at a given timestamp.

The rest of paper is organized as follows: Section 2 discuss previously available literature on spatio-temporal crime hot-spot detection and crime prediction. Section 3 presents the proposed method and discuss in detail about the dataset being used in this paper and section 4.3 and 4.4 focuses on the detection of crime hot-spots and prediction of crimes respectively. While findings are concluded in section 6.

2. Related Work

Using different methods based on statistics for the purpose of identifying the likely locations or targets where crimes are likely to be occurred in order to prevent them is known as predictive policing [16]. On the basis of information extracted from those statistical methods, police may be enabled to design and monitor efficient patrolling strategies with the minimum resources available to them. Different *"state-of-the-art"* approaches have discovered different features which were static in nature,only including information which is historical [19], information related to geography and regarding demographics [20] in crime event prediction. This section focuses on the most significant research work done in crime hot-spot detection and prediction.

2.1. Spatio-Temporal Crime Hot-Spot Detection

Crime rates can vary according to geographic location, some locations can be identified as low and high risk areas. The trends of crimes can change with seasonal patterns and time of the year as well. Performing analysis with respect to their spatial properties has grown in the last decade. In this regard, analyzing crime hot-spots is an important and popular approach [21], [22], [23], [24]. It is evident from the studies that some locations have greater perception of crimes than the actual risk level [25]. Researches have been dedicated enough in the past years to provide suitable measures that could lead to the prevention and hence reduction of crimes. In this regard, Adelson et al. [26] proposed MLP, KNN, and Random forest based method for crime hotspot detection for achieving smart city goal of public safety in Natal city of Brazil. Sankar and Gopi [27] exploited deep learning for improving accuracy and time complexity in crime hotspot detection. To evaluate the algorithm crime data from 2010 to 2018 of Los Angeles, California was collected from police records. Experimental results show that deep learning based approach can significantly improves the crime hotspot detection accuracy.

Shino and Narushige [28] proposed a network-based approach to detect crime hotspots at street level. The algorithm was evaluated on robbery, burglary, and drug data of Chicago. Empirical analysis shows the effectiveness of the method in capturing detailed information of different types of crimes. Cheng et al. [29] applied FP-Growth algorithm to discover abnormalities in purchasing, buying, and travelling behaviour of criminals. Later, the DBSCAN algorithm is applied to detect dense crime regions using generated associated rules. The algorithm showed promising results with an accuracy of 73.9 %. Ravi and Bharti [30] present a detailed analysis on big data approaches for crime hotspot analysis and proposed an algorithm for crime prediction. They used Naive Bayes classifier for crime prediction and criminal identification on Cheltenham, United Kingdom crime data collected from police data portal. The algorithm can provide significant trends and patterns of crimes to police forces.

Yiqun and Shashi [31] highlighted theoretical limitations of existing crime hotspot detection approaches and proposed a robust Non-deterministic Normalization (NN) Scan Statistic algorithm for hotspot detection. It was achieved by presenting a novel Dynamic Linear Approximation algorithm which can significantly improve the computational complexity problem. The enhanced NN-Scan evaluated on crime data of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

Minneapolis, USA collected from City police. Experimental results showed the effectiveness of the algorithm as compared to state-of-the-art techniques.

[32] proposed a spatio-temporal ordinary kriging model which used not only minimal features like location of crime, its time and type, but also their correlation to predict future crime locations as well, which helped improve the accuracy . To apply this model, crime dataset from Philadelphia from January 2011 to December 2016 was used. The proposed method achieved 90.52% sensitivity and 88.63% specificity. The summary of the techniques previously proposed by different authors on hot-spot detection techniques is mentioned in table 1.

Table 1: Spatio-Temporal Crime Hot-Spot Detection Techniques.

Study	Method	Data	Finding
[33]	Hot Spots prediction model based on mixed spatial-temporal characteristics	Data of Main city zone of Nanchang ranging from 2014 to 2015	Optimal performance can be achieved by the prediction model if crime statistics are conducted on weekly basis
[34]	Spatio-Temporal Neural Network	Call for service data provided by Portland, Oregon Police Bureau for March 2012 through the end of December 2016	81.50% Accuracy
[35]	Kernel Density Estimation (KDE)	Crimes occurred in Manila, Philippines from the year 2012 to 2016	Criminal activities in Manila are at peak around 8:00 PM to 4:00 AM
[36]	spatio-temporal kernel density estimation (STKDE)	Data of residential burglaries in Baton Rouge, Louisiana in 2011	Southwest area of Baton Rogue is identified as the high-risk area
[32]	Spatio-temporal Ordinary Kriging	Crime dataset of Philadelphia from January 2011 to December 2016	90.52% Sensitivity
[17]	DBSCAN	Crimes Dataset of New York city and Chicago	Crime dense regions are discovered

2.2. *Spatio-Temporal Crime Prediction*

With the aim of preventing crimes to happen in future, number of methodologies have been proposed in the last few years, with the goal of providing efficient and effective law enforcement agencies' resources allocation [37–39]. Jason and Anthony [40] present a correlative analysis on the disastrous pandemics such as Spanish flu and Covid-19 and its effect on a country's economy and unemployment. They find a strong correlation between unemployment and crime. The authors proposed an ARIMA based crime forecasting model to predict the next six months of crime in Queensland, Australia. Violent crime data of March 2020 collected from the Queensland police department during the smart lock down. Violent crimes data consists of assaults, domestic violence, and sexual offense. Experimental results showed 95% confidence value in determining possible crimes across Queensland.

Sohrab et al. [45] proposed a supervised learning-based crime prediction algorithm using spatial and temporal information of the crime. Decision tree and K-Nearest neighbor algorithms have been used to train the crime prediction model. Moreover, Random Forest and AdaBoost algorithms are used to improve crime prediction accuracy. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

Table 2: Spatio-Temporal Crime Prediction Techniques.

Study	Method	Data	Finding
[41]	Probabilistic Model	Crime records ranging from June 2013 to May 2014 in Dhaka, Bangladesh	79.24% sensitivity
[19]	Cluster - Confidence - Rate - Boosting (CCR-Boost)	Ranging from January 2006 to December 2009, From a Police department in a city from northeastern, US	80% Accuracy
[42]	GA-BP neural network model	Crimes that occurred from 2008 to 2012, at city in South China	The accuracy results is based on the accuracy of input data
[35]	Naive Bayes	Gun shooting crimes incurred from the year 2012 to 2016 in Manila	77.78% Accuracy
[43]	Autoregressive Integrated Moving Average model (ARIMA)	Daily police data provided by the Public Security Bureau (PSB) of a city in China	Prediction results meet the expected requirements and are more accurate
[44]	Random Forest, Neural Network, SVM, Logistic Regression Model	The crime event records of Queensland, Australia from 01/2013 to 09/2013 and New York City from March, 2012 to February, 2013	With the inclusion of dynamic features across diverse types of criminal events, crime prediction performance can be significantly improved
[17]	Seasonal ARIMA	Crimes Dataset of New York city and Chicago	Average MAE is 27.03 for dense crime regions

algorithm was evaluated on 12 years of crime data of San Francisco and outperformed as compared to state-of-the-art techniques.

[43] implemented Auto Regressive Integrated Moving Average (ARIMA) model on daily police data provided by the Public Security Bureau (PSB) of a city in China. The prediction results met the expected requirements and were more accurate. [44] used Random Forests, Neural Network, SVM, Logistic Regression Model on the crime events recorded in Queensland, Australia from 01/2013 to 09/2013 and New York City from March, 2012 to February, 2013. It was concluded that with the inclusion of dynamic features across diverse types of criminal events, crime prediction performance can be significantly improved. The summary of the techniques previously proposed by different authors on hot-spot detection techniques is mentioned in table 2.

3. Proposed Crime Prediction Methodology

Studies show that incidents of criminal events is not equally distributed within a city [17]. Because of the unequal distribution of crimes, it can be considered as a location-oriented feature as some places can exhibit greater risk of crime to be committed than others [46]. Crime rates can change with respect to the geographic location of the area. Hence, resources allocation, by law enforcement agencies to counter crimes, proportional to the number of crimes being committed in certain areas, can be a challenging task if the locations with high risk areas are unknown. Therefore, an accurate model needs to be able to detect crime hot-spots and effectively forecast crimes with respect to time and location.

In this section, we have explained in detail the proposed crime hot-spot detection and prediction methodology and the number of steps required to perform to be able to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

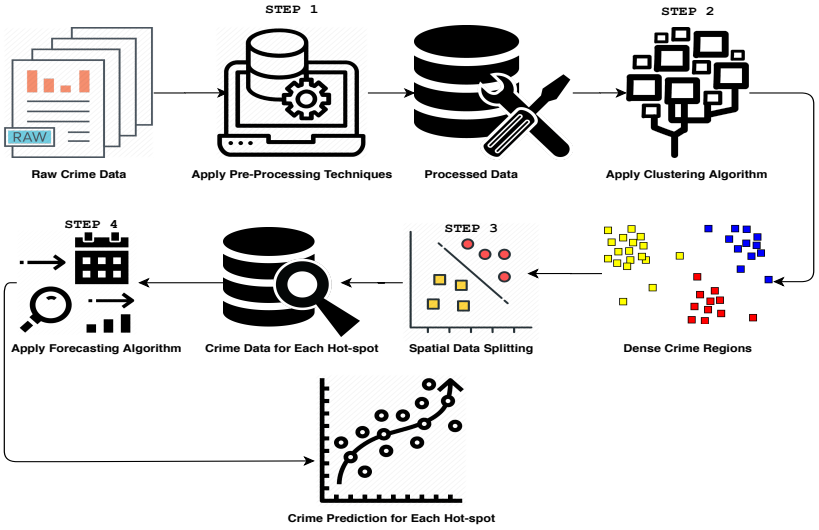

Figure 2. Steps of Spatio-Temporal Crime Prediction.

detect hot-spots and forecast number of crimes. The proposed model is based on the number of steps depicted in Figure 2.

The model consists of 4 steps as mentioned in figure 2. After the data is acquired from [15], as it is in raw form, **Step 1** is aimed towards converting it in processed form. The details of pre-processing techniques applied on the data are mentioned in section 4.2. In order to perform the required calculations, data needs to be in processed form. Details of pre-processing are discussed in section 4.2. To detect dense crime regions the goal of **Step 2** is to discover crime hot-spots by applying clustering algorithm. **Step 3** is used to perform splitting of crime data for each hot-spot. In **Step 4**, forecasting algorithm is applied to detect number of crimes for each hot-spot.

3.1. The Algorithm

As the work to be done in this study is based on two tasks, (1) Detect the crime dense regions (2) Forecast the number of crimes in each crime dense region. Therefore, to perform these two tasks, the proposed algorithm is a combination of two different algorithms (first to detect regions where crime density is greater, second to predict number of crimes in each dense region). The pseudo-code of the *"Spatio-Temporal Hot-Spot Detection and Crime Prediction algorithm (STHDCP)"* is depicted in Algorithm 1.

As discussed in the aforementioned steps of proposed model, crime hot-spots needs to be detected to perform the time series analysis. For that, clustering is performed and in each cluster time-series forecasting model is applied for prediction of crimes. Figure 2 depicts the processes of the proposed model of this study. For the *"detection of crime dense regions"* from the dataset the aim is to discover the areas where the occurrences or frequency of crime is greater than other adjacent areas, and they are to be discover automatically without a-priorily defining division in areas. This task can be modeled as a geo-spatial clustering instance, using clustering algorithm, which in this study is HDBSCAN, that processes spatial data after it has been filtered with respect to its temporal characteristics. The final output of these process is K number of clusters which corresponds to a dense crime region.

As this study is using HDBSCAN, which doesn't need a-priori information regarding number of clusters to be detected, rather, it is done automatically depending on the density of data points. The algorithm further consists the steps required to perform *"spatial data splitting"* of the original crime data, depending on the number of clusters found using the clustering model in the previous steps. In other words, data points which points to crime events are occurred belonging to the dataset, allocated to the i^{th}

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

Algorithm 1: Spatio-Temporal Hot-Spot Detection and Crime Prediction

Input: Raw Crime Dataset (RCD)

Output: $\mathcal{HS} = \{HS_1, ..., HS_K\}$: a set of K hot-spots;

$\mathcal{CP} = \{CP_1, ..., CP_K\}$: a set of K crime predictions for each hot-spot;

Method:

STEP 1: Execute *DataPreProcessing(RCD)* method to handle missing values and remove outliers from the targeted variables to get *Processed Data (PD)*;

STEP 2: Execute *DiscoverHotSpots(PD)* method to get $\mathcal{HS} = \{HS_1, ..., HS_K\}$;

STEP 3: Execute *SpatialDataSplitting(PD,HS)* to get $\mathcal{SDS}$ = {SDS$\frac{l}{1}$, . . . ,SDS$\frac{l}{K}$}: a set of K crime data for each hot-spot with spatial information;

STEP 4:

while *(for each k = 1, . . . , K)* **do**
    $CP_i \leftarrow$ DiscoverCrimePredictor(SDS$\frac{l}{i}$);
    $\mathcal{CP} \leftarrow \mathcal{CP} \cup CP_i$;
end

return $\{\mathcal{HS}, \mathcal{CP}\}$;

cluster are converted in a time series and gathered in the i^{th} output dataset, for i = 1, ...,K. The idea behind this step is to allocate the details of crimes belonging to each cluster, and partitioning them accordingly. The output of this step is K different time series datasets, each one containing the time series of crimes occurred in its associated dense region. Next step is aimed at *"extracting a specific crime prediction model for each crime dense region"*, analyzing the data of crime split in the previous step.

3.2. Detection of Crime Dense Region

The *"DiscoverHotSpots()"* method (Algorithm 1) performs clustering with respect to its spatial factor and each discovered cluster is a dense crime region. This step in this study is performed by applying HDBSCAN [47], which is an extended version of DBSCAN developed by [48]. It is extended by converting DBSCAN into a hierarchical clustering algorithm by extracting a flat clustering based in the stability of clusters.

Both algorithms have the minimum number of samples parameter which is the neighbor threshold for a record to become a core point. DBSCAN has the parameter epsilon, which is the radius those neighbors have to be in for the core to form. HDBSCAN has the parameter minimum cluster size, which is how big a cluster needs to be in order to form. This is more intuitive than epsilon because you probably have an idea of how big your clusters need to be to make actionable decisions on them. The absence of epsilon value in HDBSCAN gives it the freedom to discover the number of clusters of diverse densities and be more robust to parameter selection, as compared to DBSCAN which falls short of varying densities.

Considering the advantages that come with hierarchical clustering, another clustering algorithm was considered i.e. Hierarchical Agglomerative Clustering (HAC), but it was not suitable for this study because it has a time complexity of $O(n^3)$ and it requires $O(n^2)$ memory. However, to test the performance of algorithm, it was applied initially on 10k data with an interval of 10k. However, in comparison with HDBSCAN, clustering calculations by HAC were performed up to 55k data points only by consuming almost 108 seconds and utilizing maximum 23 GB of RAM. Whereas, HDBSCAN performed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

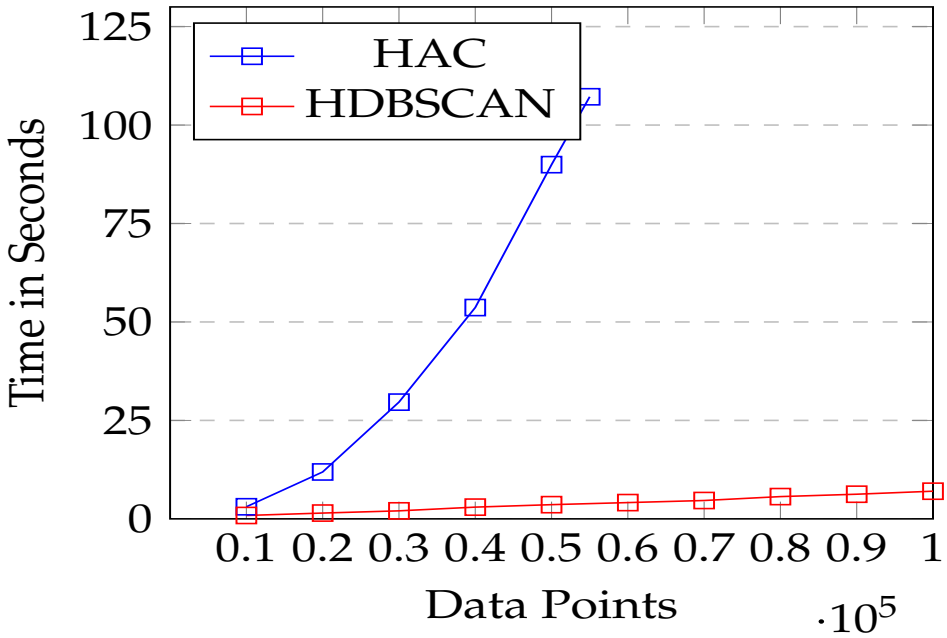

Figure 3. Time Taken by HAC and HDBSCAN to perform Clustering

calculations on 100k data points in less than 8 seconds. The results are shown in Figure 3.

Due to the intense memory and time requirements, and specifically, HAC is not a density based algorithm therefore it was not chosen for this study and instead HDBSCAN was used.

3.3. Extracting Spatio-Temporal Crime Predictors

Given a specific dense crime region, the *DiscoverCrimePredictor()* method which is mentioned in Algorithm 1 discovers a forecasting model to predict the number of crimes that will happen in its specific area. In our implementation, this has been performed by the *"Seasonal AutoRegressive Integrated Moving Average"* model (Seasonal ARIMA, or SARIMA) [49]. Which can be defined as a combination of auto-regression, moving average and difference modeling along with seasonality. Briefly, having the time series $\{y_t: t = 1...n\}$, where y_t is the the value of the time series at the timestamp t, an *ARIMA* p, d, q model can be written as mentioned in (1) below.

$$y_t^{(d)} = c + \phi_1 y_{t-1}^{(d)} + ... + \phi_p y_{t-p}^{(d)} + \theta_1 e_{t-1} + ... + \theta_q e_{t-q} + e_t \quad (1)$$

where c is a correcting factor, $\phi_1, \ldots, \phi_p$ are the regression coefficients of the auto-regressive part, $\theta_1, \ldots, \theta_q$ are the regression coefficient of the moving average part, $y_{t-1}, \ldots, y_{t-p}, e_{t-1}, \ldots, y_{t-q}$ are lagged values of y_t and lagged errors (p+q predictors), and e_t is white noise and takes into account the forecast error. In this study, ***Seasonal ARIMA*** model is exploited, which is an extension of classical **ARIMA**.

Considering the seasonal spikes in the data, i.e. the number of crimes can grow significantly in certain time of the year. Therefore, to cope with the seasonal element, **Seasonal ARIMA** model is build by including seasonal terms in **ARIMA** model. In the final formula of SARIMA, the additional seasonal terms are simply multiplied with the non-seasonal terms. A Seasonal ARIMA model is referred as *ARIMA(p,d,q)(P,D,Q)m*, where m is a periodicity factor, *(p,d,q)* and *(P,D,Q)* are the orders of the auto-regressive, differencing and the moving average part for the non-seasonal and seasonal model, respectively [49]. The problem with ARIMA is that if data is non-stationary, and possess clear trends then it requires a lot of differencing to make it stationary and that might

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

Table 3: New York City's Five Boroughs.

Jurisdiction	Land Area		Density	
	Square Miles	Square KM	Persons/Sq. Mi	Persons/km^2
The Bronx	42.10	109.04	34,653	13,231
Brooklyn	70.82	183.42	37,137	14,649
Manhattan	22.83	59.13	72,033	27,826
Queens	108.53	281.09	21,460	8,354
Staten Island	58.37	151.18	8,112	3,132
Total	**302.64**	**783.83**	**28,188**	**10,947**

be the reason that the seasonal spikes get ignored and it may gives undesired results. Hence, to deal with the seasonal element of the data, ARIMA is not suitable approach. Moreover, it is evident from the results extracted by [50], that to perform time series forecasting, ARIMA performs better than state of the art Machine Learning algorithms like SVM, RNN, KNN, and LSTM etc for time-series forecasting. And because the data has seasonal dips in it, that is why, SARIMA was applied on the dataset.

4. Experimental Evaluation

For the evaluation regarding efficiency and effectiveness of the proposed model, it was applied on the dataset acquired from [15]. The details of the dataset are discussed in the section .

4.1. Experimental Dataset

The data that we used to train the models and perform the experimental evaluation for New York City is housed on [15], a publicly available resource managed by the Mayors Office of Data Analytics (MODA) and the Department of Information Technology and Telecommunications (DoITT). The foucs of this study are all five boroughs of NYC i.e. *(1)* The Bronx *(2)* Brooklyn *(3)* Manhattan *(4)* Queens *(5)* Staten Island.

These regions are one of the most dense urban areas in the world, growing in terms of population, business and patterns of mobility. The total area of these Boroughs is 783.83 KM^2. Their details in terms of land area and density of the population as per land area are mentioned in table 3 below which are extracted from [51]. As per the information extracted from [51], Queens is the borough with the largest Land area with 281 square KM whereas Manhattan is the most densely populated borough with 72,033 persons per square miles. The dataset contains the information regarding all crimes that were reported in these borough starting from year 01-01-2008 to 31-12-2017.

The dataset[1] contains 4,952,699 rows representing the number of crimes occurred over the years in different locations. Whereas the average number of crimes recorded in a week are 10,318. The size of data size is approximately 2GB. Since Manhattan is the most densely populated borough among all five, hence it can be assumed and proved from Figure 4 that it is also the most dense region for crimes. The density of the crimes in all five boroughs is shown in Figure 4.

The density map in the Figure 4 indicates the crime density of regions, where some regions have very low density of crimes while some areas possess greater density because of the greater number of crime occurrences in the past. Starting from blue and going towards red, the colors depicts the density of crime from low to highest dense regions where number of crimes are greater than the other regions. And red, being the most dense regions, can be considered for efficient and effective police resource allocation region to control criminal activities there. The crimes frequency distribution

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

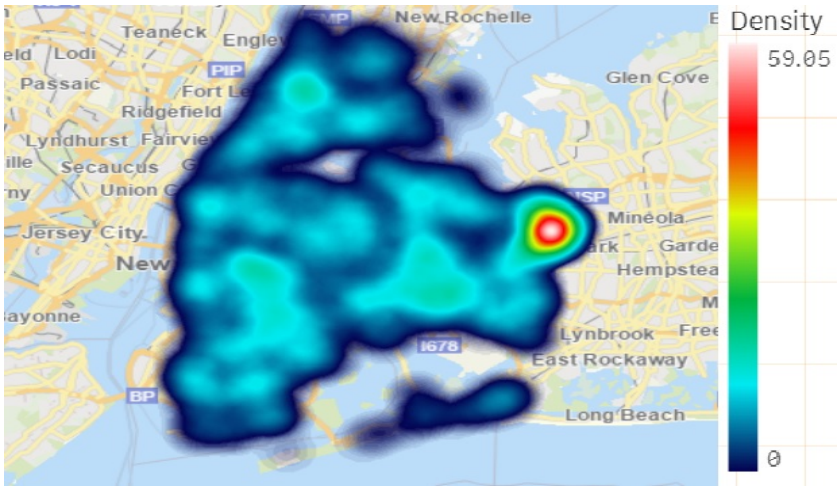

Figure 4. Crime Density in the Boroughs of NYC.

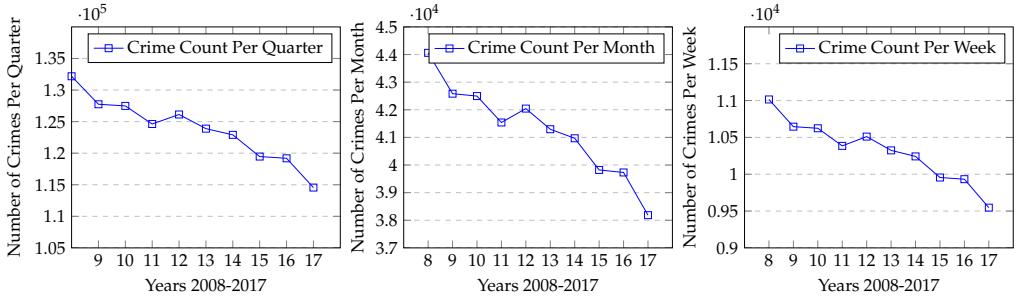

Figure 5. Number of Crimes Distribution by Quarter, Month and Week

per quarter, month and week for each year is shown in Figure 5 to better understand the trends of crimes every year.

4.2. Pre-Processing of Dataset

Pre-processing is a method of cleaning data to remove noise, handle missing values and deal with outliers i.e. the values that don't comply with normal distribution of values. These tasks are performed on the data to make it ready for further processing because to get accurate and better results, data must be cleaned before performing additional calculations. In this study, we have performed the following steps of pre-processing on our data to make it ready for further processing. (a) Data Formatting (b) Removing the outliers (c) Attribute Selection (d) Handling Missing Values. The data being used in this study was taken from [15], consisting of multiple attributes of diverse data types. The main focus of this study was to use the attributes i.e. CMPLNT_FR_DT (Date at which crime was reported), Latitude and Longitude to detect crime dense areas and to perform forecasting in those areas.

The date is ranged from 01-01-2008 to 31-12-2017 Latitude and Longitude coordinates for NYC are: 40.730610, -74.4. First of all, the datatype of CMPLNT_FR_DT was string in the dataset, which had to be converted into ***Date*** to perform time series calculations. From ***Pandas*** library of Python, ***"to_datetime"*** was used to convert the data from string to *Date* data type. Afterwards, the outliers in Latitude and Longitude attributes were detected and removed. For example, the Latitude values >= 41.0 and Longitude values <= -74.5 were removed to comply with the normal distribution of values, i.e. representing locations of NYC. Figure 6 depicts the distribution of data points as the Latitudes and Longitudes ***with*** outliers.

The Figure 6 depicts the uneven distribution of coordinates because of the outliers. Considering the actual range of coordinates, the data was sorted and the outliers were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

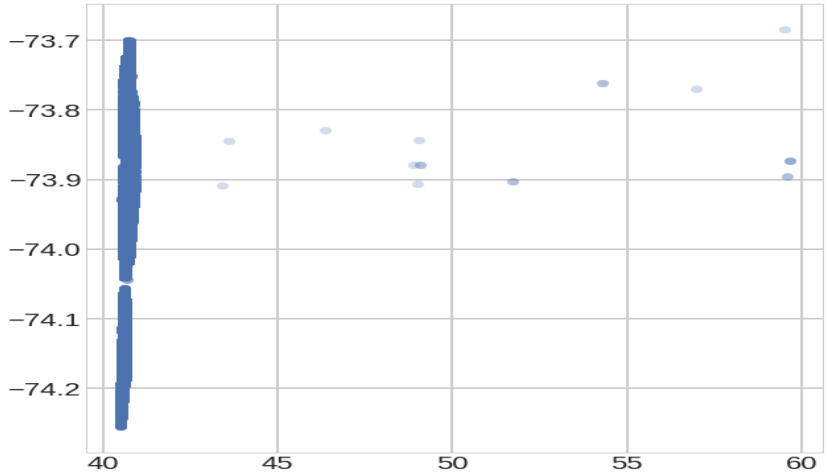

Figure 6. Latitude and Longitude ***with*** Outliers.

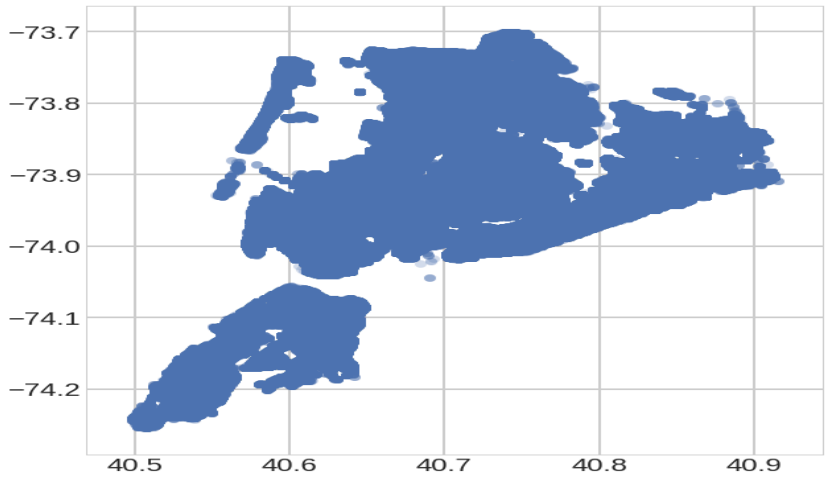

Figure 7. Latitude and Longitude ***without*** Outliers.

removed which did not belong to the mentioned range. Figure 7 depicts the distribution of data points as the coordinates ***without*** outliers.

To perform clustering, there must not be any missing values in the data, therefore, to handle the missing values in Latitude and Longitude, ***FillByMean*** method was used which computes the mean of all values and fill the missing values with that mean which is a reasonable approach considering that the mean will always belong from the range of those values, which were already normalized considering the removal of outliers.

4.3. Detecting Crime Hot-Spots from Dataset

As discussed in 3.2, the prior need to mention the number of clusters before the formation of clusters can be a major problem especially when you don't have prior knowledge how many number of clusters should you be looking for. Hence, there's a strong need of being able to calculate *K* number of clusters depending on the number of crimes or density of crimes in an area. As this study used HDBSCAN which automatically forms the number of cluster using the density of data points. The parameters which were used in HDBSCAN are **min_cluster_size** which is an intuitive parameter to select, it is the smallest size grouping that is wished to be consider to form a cluster, this study used 20000 as **min_cluster_size**. **min_samples** is another parameter which provides a measure of how conservative you want clustering to be. The larger the value, the more conservative the clustering is. The value of **min_samples** was 50 in this study and **distance metric** was *euclidean*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

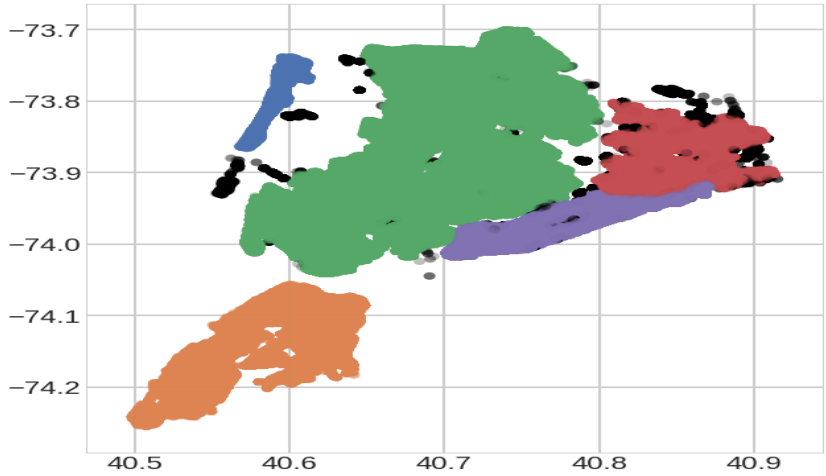

Figure 8. Crime Dense Regions in NYC Discovered using Euclidean Distance.

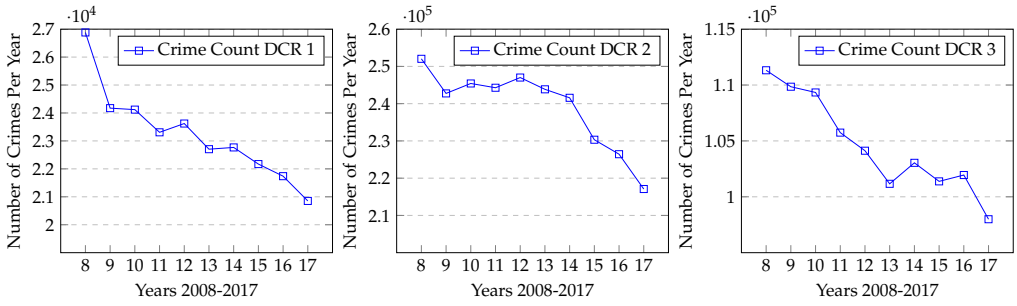

Figure 9. Crimes Frequency Distribution for DCR1, DCR2, and DCR3

Crime dense regions of years 2008 to 2017, discovered using HDSCAN algorithm are shown in Figure 8. Different colors are representing different crime dense regions. Algorithm discovered 5 different crime regions clearly visible through different colors. Black color represents noise which will be ignored.

From top three dense regions, following information in Figure 9 was extracted which shows the crime frequency distribution for each year. The Figure shows that crime frequency is gradually declining towards the end of 2017. Furthermore, it can be understood that DCR 3 is the most dense regions as the total number of crimes reported in this region were 2,390,751. Whereas, the total number of crimes reported in DCR1 and DCR2 were 232,328 and 1,045,887 respectively.

The information in Figure 9 shows the number of crimes reported in years 2008-2017. However, after detecting the dense crime regions as shown in Figure 8, their density is shown in figures 10 for DCR1, DCR2, and DCR3 for each year respectively. For each year, the density of all dense crime regions, is highlighted in the figures. The level of density is represented in the bar using the count of crime for each year. Based on the clusters formed, top three dense crime regions are shown in the figures with the location and density of crimes in those regions. One of the main contributions of this paper was to detect those regions, hence patrolling and other efficient strategies can be developed on the basis of this information leading towards possible significant drop in number of crimes. It is to be noted that as previously stated, not all locations possess same amount of danger or number of crimes, it is different for each location.

It is evident from the figures that DCR3 is the most dense crime region among those three.

4.4. Training and Evaluating the Regressive Crime Models

Among the number of crime dense regions extracted in section 4.3, top three dense crime regions are selected to predict the number of crimes in those regions. Crime

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

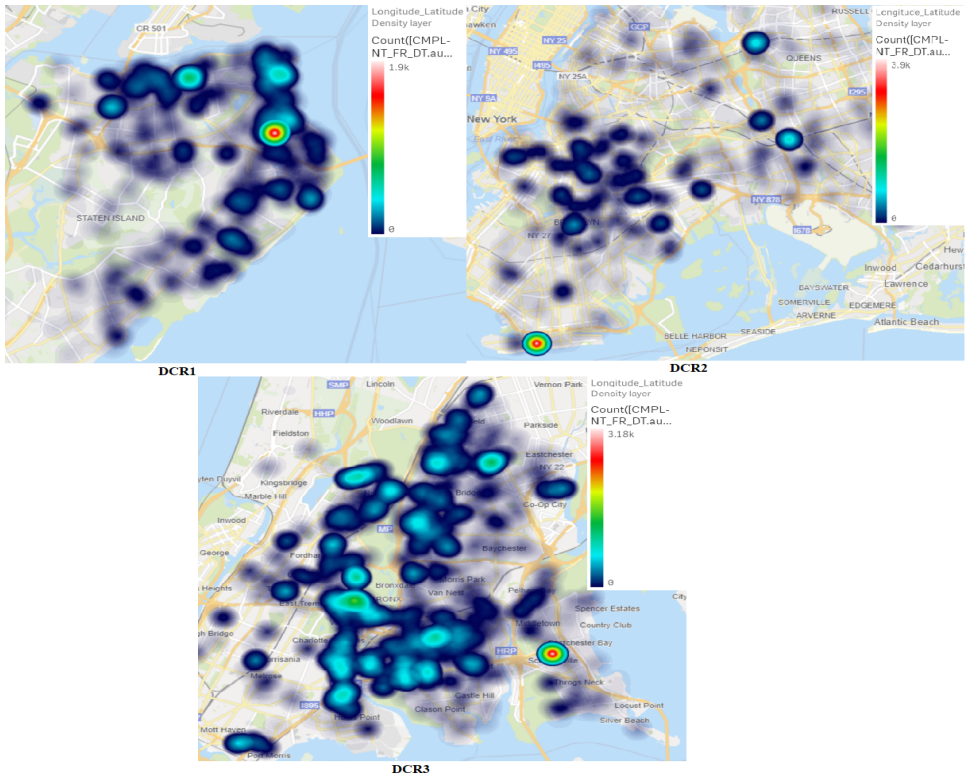

Figure 10. Crime Density of DCR1, DCR2, and DCR3 respectively.

regressive models have been extracted for years starting from 2008 till 2017 for each year, then for all years for all boroughs in the dataset.

The evaluation of the regressive functions trained and tested by 80:20 for the dataset of New York City. 80% data was used for training, and 20% data used for testing of the regressive functions for all years mentioned in the table. For those years, the significant trend in the crime count *DCR1, DCR2, and DCR3* is shown in Figure 11.

For *DCR1, DCR2, and DCR3* for years 2008-2017, following results were extracted from the data set which are shown in the Figure 12 for *DCR1, DCR2 and DCR3* respectively. The figures shows the forecasting values among the actual values.

The blue line represents the actual number of crimes in each year, whereas, the orange line in front of blue lines, represents the predicted number of crimes.

To better understand the forecasting results, *plot diagnostics* were applied on the model used for all three dense crime regions. There are four key parameters in the diagnostic which defines the quality of results.

1. **Standard Residuals:** There are no obvious patterns in the residuals
2. **Histogram Plus KDE Estimate:**The KDE curve should be very similar to the normal distribution
3. **Normal Q-Q:** Most of the data points should lie on the straight line
4. **Correlogram:** 95% of correlations for lag greater than one should not be significant

The diagnostics of the results depicted in Figure 12 are shown in Figure 13.

As per the four parameters to evaluate the diagnostics results and shown in the images above, it can be concluded from the results in figures for years 2008-2017, *DCR1* is the region that showed better forecasting performance and the difference between predicted and actual values is minimum compared to other crime dense regions. Whereas the KDE curve is very similar to the normal distribution, most of the data points are lying on the straight line and correlations for lag greater than are not significant. To evaluate the performance of forecasting on the dataset, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Error (ME), and Root Mean Square Error

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

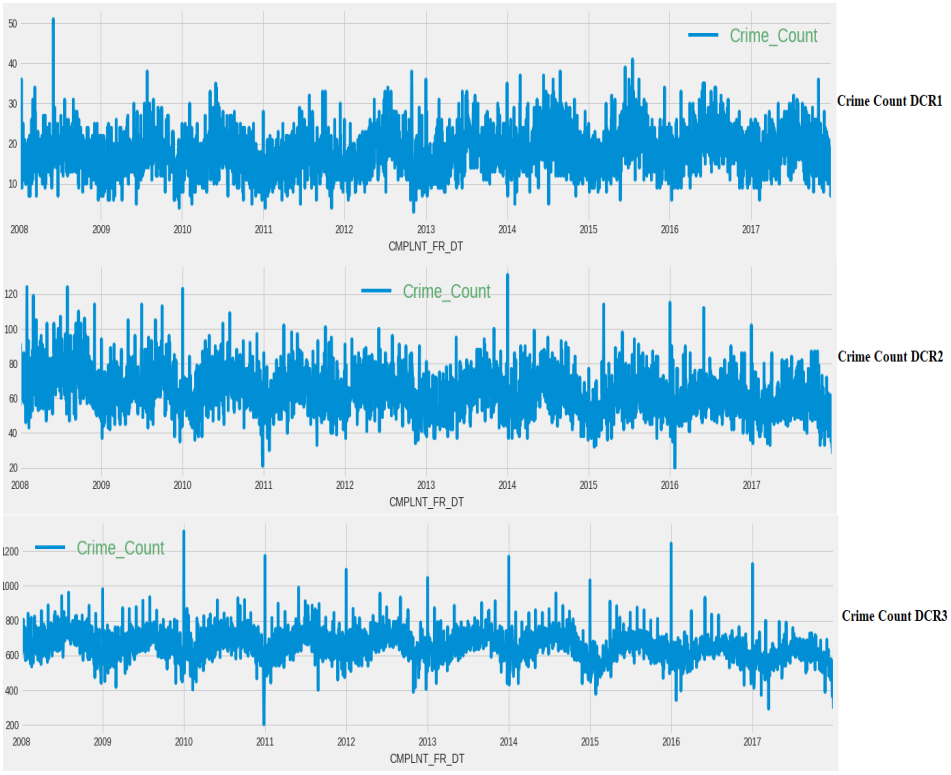

Figure 11. Crime Count over the Years for DCR1, DCR2, DCR3 respectively

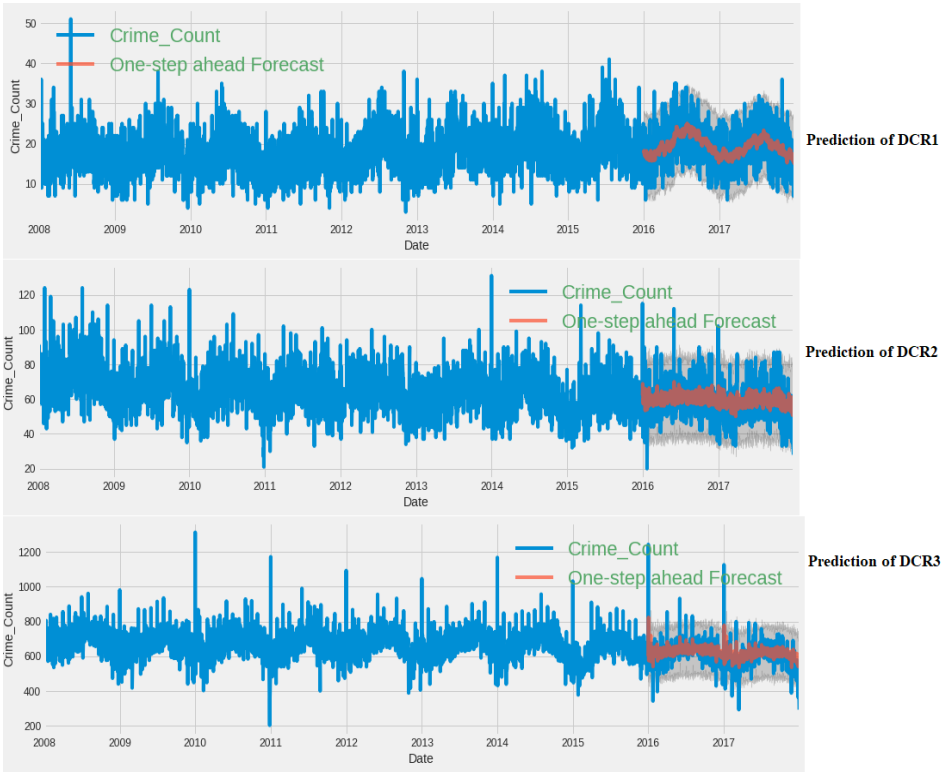

Figure 12. Crime Forecasting for DCR1, DCR2, DCR3 respectively

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

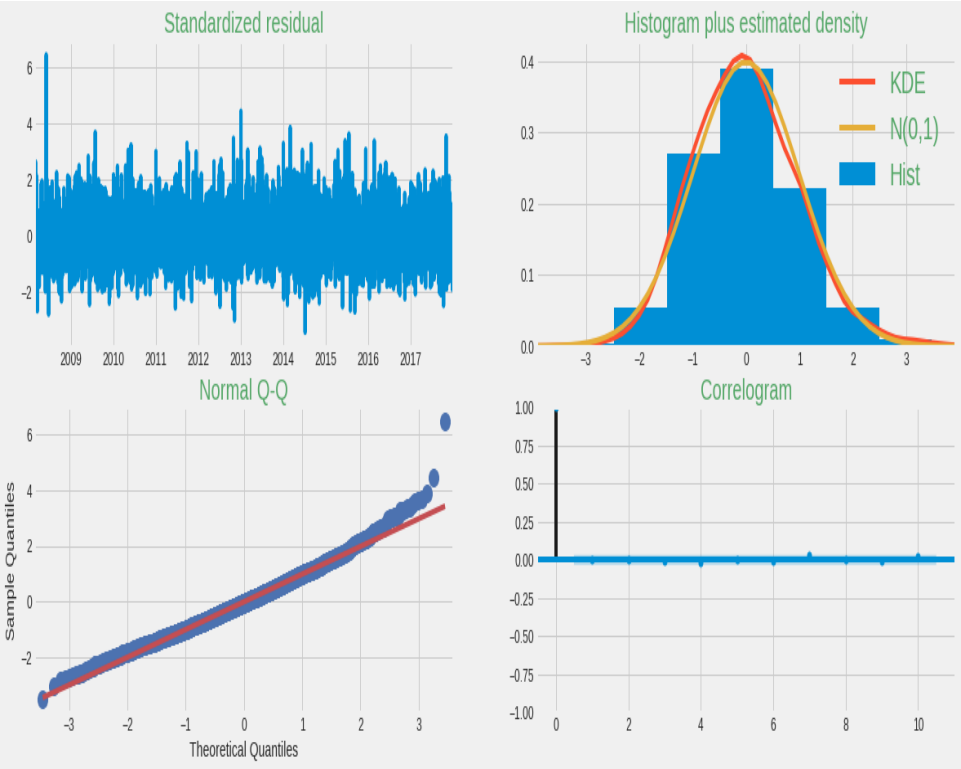

Figure 13. Results' Diagnostics of DCR1 Years 08-17.

(RMSE) have been used. The error values are calculated by comparing the predicted values with the actual values using testing and training data with 80-20 ratio i.e. 80% training and 20% testing data. The results achieved by applying the proposed model on the dataset, are shown in Figure 14.

It can be understood from the figures that DCR1 secured better results than other regions whereas DCR2 and DCR3 performed less better than DCR1. It can be due to the fact that among these three regions, DCR1 is the smallest with 232,328 number of crimes as mentioned in 4.3. Therefore, it can be understood that regions where number of crimes is lower than other, can give better results of forecasting in terms of error values used in this study to evaluate the results.

5. Comparative Analysis

To make our results more accurate and authentic, a comparative analysis is performed in this study with the results achieved by [17] by using DBSCAN for clustering. The comparison of the results of both studies is shown in Figure 15. The models which were used as an input for crime regressive models were different in both studies, moreover, their forecasting was based on weekly trends hence the value of m was 52. While in this study, as evident from Figure 5, the trend of crime was not non-stationary weekly, whereas a clear decline in number of crimes can be seen monthly as compared to weekly and quarterly distribution. Therefore, the value of m was 12 which represents the twelve months in a year.

Furthermore, the results achieved by [17] represents the results of year 2014-2016 only, whereas this study represents detailed analysis of years 2008 to 2017. In both studies, the results are depicted for each crime dense region for every year. However, for comparative analysis for both studies, the results are shown for common years only and for all four error measuring parameters, their average is used to compare the results.

It can be concluded from the results that the average performance of the model used in this study extracted better results than extracted by [17]. It is evident from the result that the result of three error evaluation metrics i.e. MAE, MAPE, and RMSE for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

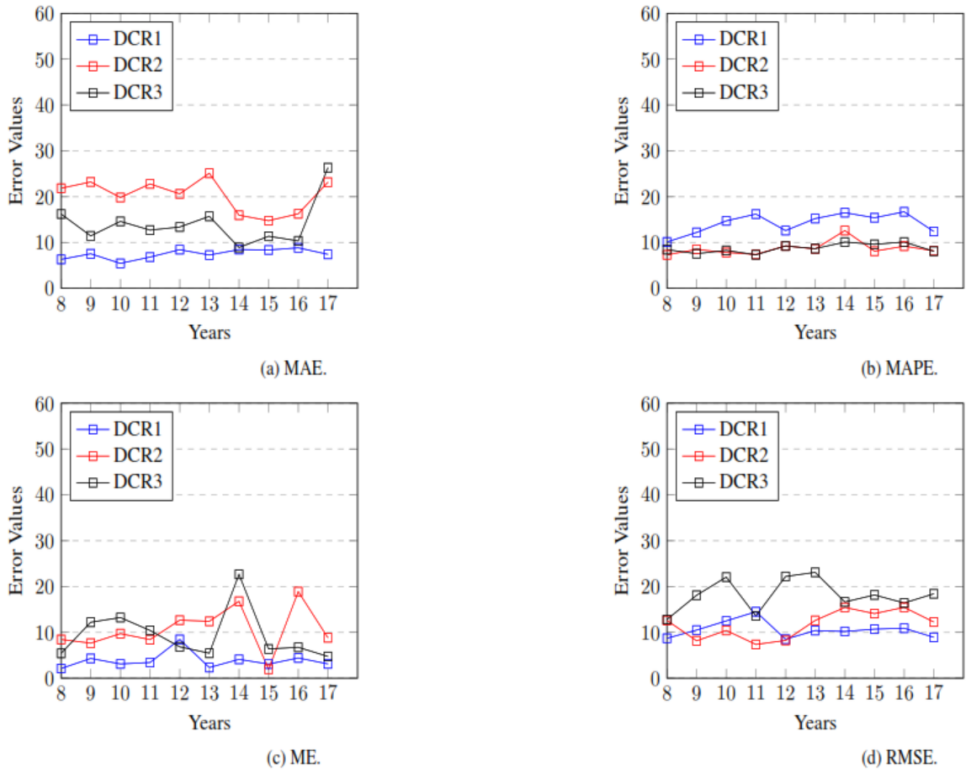

Figure 14. Forecasting Error Values

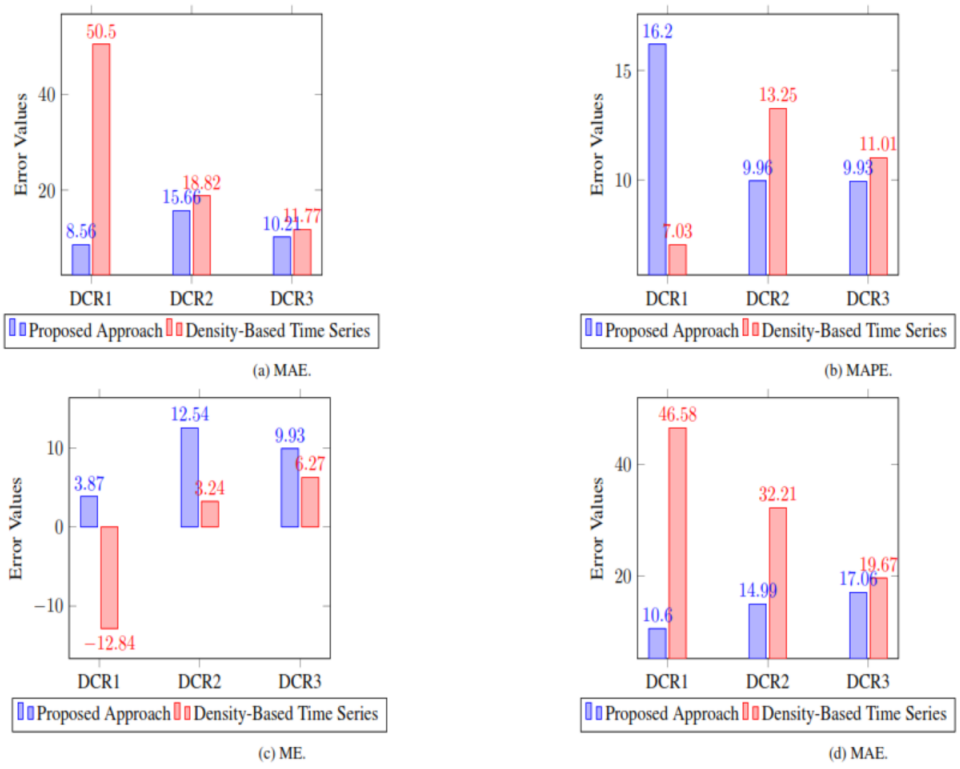

Figure 15. Comparison of Forecasting Error Values

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 February 2021 doi:10.20944/preprints202102.0172.v1

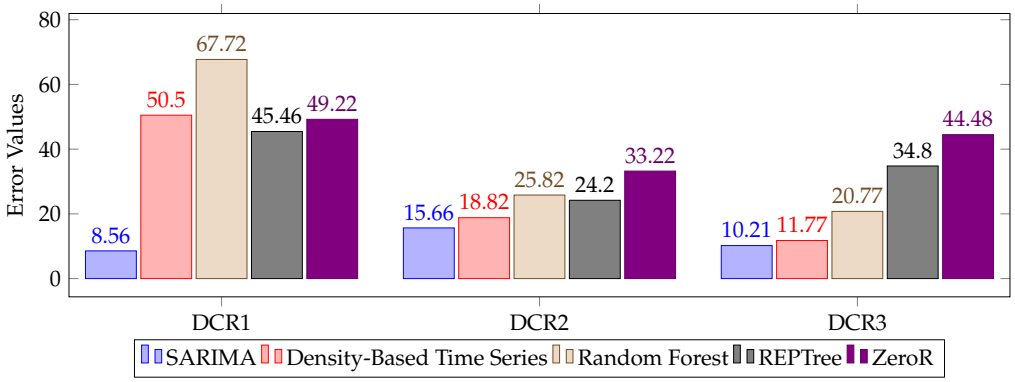

Figure 16. Comparison of Forecasting Error Values

this study is better than the comparative study. Only ME score for *DCR2 and DCR3* of comparative study is better than this study's result. This study extracted forecasting results for years 2008-2017 for top three crime dense regions which is more than the competitor's study.

Moreover, the results of this study are compared for extraction of crime predictors. Specifically, a comparative analysis of the results achieved using SARIMA in this study, is performed against the result extracted by classical regression algorithms such as Random Forest [52], REPTree [53], and ZeroR[54] which is shown in Figure 16.

To perform the comparison of the results between this study's and the aforementioned technologies, the performance of predicting number of crimes in top three crime dense regions was evaluated on the dataset used in this study. For each algorithm, results were extracted using suitable and accurate input parameters to get the best results the algorithm could offer. To compare the results, error evaluation metric MAE is used for two year ahead crime predictions using data of 8 years for training. Figure 16 summarizes the comparison' results and we can see that SARIMA results are generally better than others. This is to be noted that the comparison was performed on the same dataset and using the same timeline. Therefore, the window of comparison is same. These results confirm the appropriateness of the autoregressive model and its good performance in the crime prediction domain.

Hence, it can be concluded from the results depicted in the tables and figures above that the proposed model used in this study achieved better results than other studies mentioned in the literature.

6. Conclusion

One of the crucial challenges of smart city infrastructure is to provide a reliable and secure environment that is addressed by detecting crime hot-spots and predicting number of crimes in them. This information can be helpful for concerned stake-holders in providing a safe environment for a smart city' citizen. Because of the constant growth in data for smart cities, managing and utilizing the computational resources can be a challenging task. This paper proposed a cost-efficient and effective approach to perform the aforementioned tasks. The results were evaluated on a dataset of ten years of crimes reports.The experimental results show that the proposed system outperformed as compared to state-of-the-art systems with average Mean Absolute Error (MAE) of 11.47.

In the future, we aim to improve the proposed model by exploiting transfer learning. In this approach, the knowledge of an already learned crime prediction model is utilized to solve a related region crimes which can improve the performance and cost of learning. The clustering ensemble can also be used in the future for a more accurate and robust crime detection and prediction model.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Spencer, N.; Butler, D. Cities: the century of the city. *Nature* **2010**, *467*, 900–901.
2. Cicirelli, F.; Guerrieri, A.; Spezzano, G.; Vinci, A. An edge-based platform for dynamic smart city applications. *Future Generation Computer Systems* **2017**, *76*, 106–118.
3. Sherazi, H.H.R.; Iqbal, R.; Ahmad, F.; Khan, Z.A.; Chaudary, M.H. DDoS attack detection: A key enabler for sustainable communication in internet of vehicles. *Sustainable Computing: Informatics and Systems* **2019**, *23*, 13–20.
4. Iqbal, R.; Butt, T.A.; Afzaal, M.; Salah, K. Trust management in social Internet of vehicles: Factors, challenges, blockchain, and fog solutions. *International Journal of Distributed Sensor Networks* **2019**, *15*, 1550147719825820.
5. Dar, Z.; Ahmad, A.; Khan, F.A.; Zeshan, F.; Iqbal, R.; Sherazi, H.H.R.; Bashir, A.K. A context-aware encryption protocol suite for edge computing-based IoT devices. *The Journal of Supercomputing* **2019**, pp. 1–20.
6. Baqir, A.; ul Rehman, S.; Malik, S.; ul Mustafa, F.; Ahmad, U. Evaluating the Performance of Hierarchical Clustering algorithms to Detect Spatio-Temporal Crime Hot-Spots. 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). IEEE, 2020, pp. 1–5.
7. Butt, U.M.; Letchmunan, S.; Hassan, F.H.; Ali, M.; Baqir, A.; Sherazi, H.H.R. Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review. *IEEE Access* **2020**, *8*, 166553–166574.
8. Kaufmann, M.; Egbert, S.; Leese, M. Predictive policing and the politics of patterns. *The British Journal of Criminology* **2019**, *59*, 674–692.
9. Yi, F.; Yu, Z.; Zhuang, F.; Zhang, X.; Xiong, H. An Integrated Model for Crime Prediction Using Temporal and Spatial Factors. 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018, pp. 1386–1391.
10. Tapia-McClung, R. Exploring the Use of a Spatio-Temporal City Dashboard to Study Criminal Incidence: A Case Study for the Mexican State of Aguascalientes. *Sustainability* **2020**, *12*, 2199.
11. Altay, E.V.; Alatas, B. Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining. *Journal of Ambient Intelligence and Humanized Computing* **2019**, pp. 1–21.
12. Das, P.; Das, A.K. Application of classification techniques for prediction and analysis of crime in India. In *Computational Intelligence in Data Mining*; Springer, 2019; pp. 191–201.
13. Richmond-Rakerd, L.S.; D'Souza, S.; Andersen, S.H.; Hogan, S.; Houts, R.M.; Poulton, R.; Ramrakha, S.; Caspi, A.; Milne, B.J.; Moffitt, T.E. Clustering of health, crime and social-welfare inequality in 4 million citizens from two nations. *Nature Human Behaviour* **2020**, *4*, 255–264.
14. Brantingham, P.J.; Brantingham, P.L.; Song, J.; Spicer, V. Crime Hot Spots, Crime Corridors and the Journey to Crime: An Expanded Theoretical Model of the Generation of Crime Concentrations. In *Geographies of Behavioural Health, Crime, and Disorder*; Springer, 2020; pp. 61–86.
15. NYCOpenData. NYPD Complaint Data Historic | NYC Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/data, 2019. [Online; accessed 13-June-2019].
16. Meijer, A.; Wessels, M. Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration* **2019**, *42*, 1031–1039.
17. Catlett, C.; Cesario, E.; Talia, D.; Vinci, A. Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive and Mobile Computing* **2019**, *53*, 62–74.
18. Kawthalkar, I.; Jadhav, S.; Jain, D.; Nimkar, A.V. Predictive Crime Mapping for Smart City. In *Advances in Distributed Computing and Machine Learning*; Springer, 2020; pp. 359–368.
19. Yu, C.H.; Ding, W.; Morabito, M.; Chen, P. Hierarchical spatio-temporal pattern discovery and predictive modeling. *IEEE Transactions on Knowledge and Data Engineering* **2016**, *28*, 979–993.
20. Wang, X.; Brown, D.E. The spatio-temporal generalized additive model for criminal incidents. Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics. IEEE, 2011, pp. 42–47.
21. Hajela, G.; Chawla, M.; Rasool, A. A Clustering Based Hotspot Identification Approach For Crime Prediction. *Procedia Computer Science* **2020**, *167*, 1462–1470.
22. Xu, C.; Hu, X.; Yang, A.; Zhang, Y.; Zhang, C.; Xia, Y.; Cao, Y. Crime Hotspot Prediction Using Big Data in China. In *Handbook of Research on Managerial Practices and Disruptive Innovation in Asia*; IGI Global, 2020; pp. 351–371.
23. Hart, T.C. Hot Spots of Crime: Methods and Predictive Analytics. In *Geographies of Behavioural Health, Crime, and Disorder*; Springer, 2020; pp. 87–103.
24. Braga, A.A.; Turchan, B.S.; Papachristos, A.V.; Hureau, D.M. Hot spots policing and crime reduction: an update of an ongoing systematic review and meta-analysis. *Journal of experimental criminology* **2019**, *15*, 289–311.
25. Telep, C.W.; Hibdon, J. *Understanding and Responding to Crime and Disorder Hot Spots*; Department of Justice, 2019.
26. Araújo, A.; Cacho, N.; Bezerra, L.; Vieira, C.; Borges, J. Towards a crime hotspot detection framework for patrol planning. 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018, pp. 1256–1263.
27. Nair, S.N.; Gopi, E. Deep Learning Techniques for Crime Hotspot Detection. In *Optimization in Machine Learning and Applications*; Springer, 2020; pp. 13–29.

28. Shiode, S.; Shiode, N. A network-based scan statistic for detecting the exact location and extent of hotspots along urban streets. *Computers, Environment and Urban Systems* **2020**, *83*, 101500.
29. Cheng, B.; Li, W.; Tong, H. Prediction of Criminal Suspects Based on Association Rules and Tag Clustering. *Journal of Software Engineering and Applications* **2019**, *12*, 35–50.
30. Kumar, R.; Nagpal, B. Analysis and prediction of crime patterns using big data. *International Journal of Information Technology* **2019**, *11*, 799–805.
31. Xie, Y.; Shekhar, S. A nondeterministic normalization based scan statistic (NN-scan) towards robust hotspot detection: a summary of results. Proceedings of the 2019 SIAM International Conference on Data Mining. SIAM, 2019, pp. 82–90.
32. Deshmukh, S.S.; Annappa, B. Prediction of Crime Hot Spots Using Spatiotemporal Ordinary Kriging. In *Integrated Intelligent Computing, Communication and Security*; Springer, 2019; pp. 683–691.
33. Zhang, Q.; Yuan, P.; Zhou, Q.; Yang, Z. Mixed spatial-temporal characteristics based crime hot spots prediction. 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2016, pp. 97–101.
34. Zhuang, Y.; Almeida, M.; Morabito, M.; Ding, W. Crime hot spot forecasting: A recurrent model with spatial and temporal information. 2017 IEEE International Conference on Big Knowledge (ICBK). IEEE, 2017, pp. 143–150.
35. Baculo, M.J.C.; Marzan, C.S.; de Dios Bulos, R.; Ruiz, C. Geospatial-temporal analysis and classification of criminal data in Manila. 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). IEEE, 2017, pp. 6–11.
36. Hu, Y.; Wang, F.; Guin, C.; Zhu, H. A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied geography* **2018**, *99*, 89–97.
37. Brayne, S.; Christin, A. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems* **2020**.
38. Bhatti, M.H.; Khan, J.; Khan, M.U.G.; Iqbal, R.; Aloqaily, M.; Jararweh, Y.; Gupta, B. Soft computing-based EEG classification by optimal feature selection and neural networks. *IEEE Transactions on Industrial Informatics* **2019**, *15*, 5747–5754.
39. Mushtaq, H.; Siddique, I.; Malik, B.H.; Ahmed, M.; Butt, U.M.; Ghafoor, R.M.T.; Zubair, H.; Farooq, U. Educational Data Classification Framework for Community Pedagogical Content Management using Data Mining. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* **2019**, *10*, 329–338.
40. Payne, J.; Morgan, A. COVID-19 and Violent Crime: A comparison of recorded offence rates and dynamic forecasts (ARIMA) for March 2020 in Queensland, Australia **2020**.
41. Parvez, M.R.; Mosharraf, T.; Ali, M.E. A novel approach to identify spatio-temporal crime pattern in Dhaka City. Proceedings of the eighth international conference on information and communication technologies and development. ACM, 2016, p. 41.
42. Weihong, L.; Lei, W.; Yebin, C. Spatial- temporal forecast research of property crime under the driven of urban traffic factors. *Multimedia Tools and Applications* **2016**, *75*, 17669–17687.
43. Li, Z.; Zhang, T.; Yuan, Z.; Wu, Z.; Du, Z. Spatio-Temporal Pattern Analysis and Prediction for Urban Crime. 2018 Sixth International Conference on Advanced Cloud and Big Data (CBD). IEEE, 2018, pp. 177–182.
44. Rumi, S.K.; Deng, K.; Salim, F.D. Crime event prediction with dynamic features. *EPJ Data Science* **2018**, *7*, 43.
45. Hossain, S.; Abtahee, A.; Kashem, I.; Hoque, M.M.; Sarker, I.H. Crime Prediction Using Spatio-Temporal Data. *arXiv preprint arXiv:2003.09322* **2020**.
46. Wortley, R.K.; Mazerolle, L.A. *Environmental Criminology and Crime Analysis*; Vol. 6, 2016; pp. 1–294.
47. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* **2017**, *2*, 205.
48. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X.; others. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 1996, Vol. 96, pp. 226–231.
49. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: principles and practice*; OTexts, 2018.
50. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one* **2018**, *13*, e0194889.
51. Census.gov. United States Census Bureau. https://www.census.gov/, 2019. [Online; accessed 15-November-2019].
52. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
53. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical machine learning tools and techniques*; Morgan Kaufmann, 2016.
54. Nasa, C.; Suman, S. Evaluation of different classification techniques for web data. *International journal of computer applications* **2012**, *52*, 34–40.