Crime Hotspot Detection For Interesting Patterns

Sanjit Dash, Anubhav Panda

Abstract

This paper focuses on finding criminal hotspots of colocated spatio-temporal crimes. It analyses real-world crime datasets (NYC) and provides an exploratory data overview through a statistical analysis supported by several graphs. Then, it clarifies how we conducted the Colocation detection algorithm to produce interesting frequent patterns for criminal hotspots. In addition, the paper shows how we used SatScan and DBSCAN in order to find the hotspot of colocated crimes. The results of this solution could be used to raise people's awareness regarding the dangerous locations and to help agencies to take corresponding actions in a specific location within a particular time.

1. Introduction

Crimes are a common social problem affecting the quality of life and the economic growth of a society [1]. It is considered an essential factor that determines whether or not people move to a new city and what places should be avoided when they travel [2]. With the increase of crimes, law enforcement agencies are continuing to demand advanced geographic information systems and new data mining approaches to improve crime analytics and better protect their communities [3]. Although crimes could occur everywhere, it is common that criminals work on crime opportunities they face in most familiar areas for them [4]. Here the main Idea is to raise people's awareness by using a data mining approach which will determine the criminal hotspots by analyzing the type location and type of committed crimes . Our proposed solution can be used to save lives by advising people to stay away from the locations at a certain time of the day. In addition, having this kind of knowledge would help people to improve their living place choices. Police forces can get insights from this solution to increase the level of crime prediction

and prevention. This analysis can be used for police resources allocation. It can help in the distribution of police at most likely crime places for any given time, to grant an efficient usage of police resources [5].We hope to make our community a safer place by having all of this available information.

We are planning to perform Exploratory data analysis using the crime datasets (example NYC Crime Data sets) and we would like to compare different approaches to calculate Association among the crimes. Then we would like to try different approaches for Hotspot Detection, crime prediction using Spatial and Temporal criminal hotspots, crime classification, etc. The most commonly used methods are, entity extraction, clustering techniques, association rule mining, sequential pattern mining and classification[2].

Keywords

Data mining, crime predilection, crime classification, crime frequent patterns, NYC criminal hotspots, MEC, DBSCAN, SatScan

2. Problem Statement

2.1 Input

The input here is a set of crime locations, the date and time of their occurrence, type of crimes and their respective frequency

2.2 Output

The expected output is a hotspot pattern for associated crimes.

2.3 Goal

The major goal of this implementation is finding out the important features in the dataset, finding the interesting patterns in crime (co-located crimes) and finally computing the hotspots.

3. Significance of the Problem

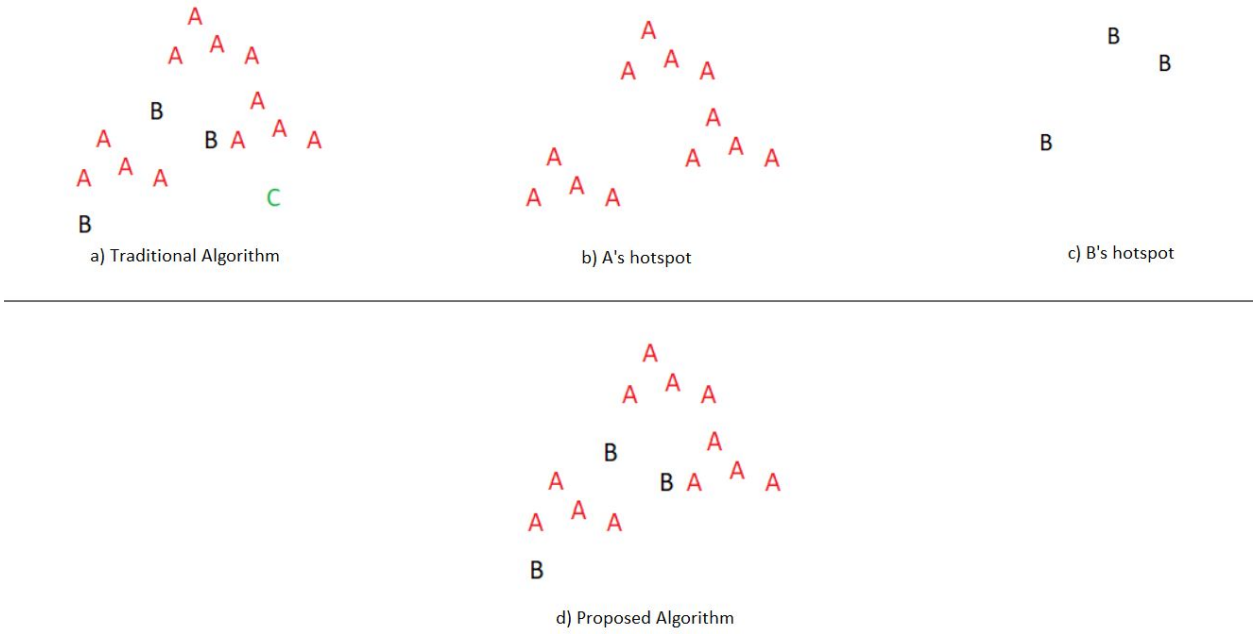

Figure 1

Let's consider a toy example denoting the spatial locations of three types of crimes (A, B and C). As we can see in Figure 1, we have a hotspot for A and a hotspot for B. But clearly we do not have a hotspot for C, as we have only one instance of C in the toy example. But the traditional hotspot detection algorithm does not differentiate between the types of crimes. It will run the hotspot detection algorithm on the whole dataset and will create hotspots of crimes densely located in location. So we will have C in the hotspot. But the proposed algorithm will first find which crimes can be associated together. Suppose crime type B is associated with crime type A. Then on running SatScan on the associated crimes we will get the output as shown in the above Figure 1 [d) Proposed Algorithm]. So as we can see C is eliminated from the final hotspot.
The various challenges of the problem statement can listed as :

- Difficulty of enumerating all potential hotspot areas. (computational complexity)
- Testing for statistical significance to reduce chance patterns.
- Find the patterns of crime occurrence within a certain time interval and radius
- Hyper-parameter tuning
- Non-existent addresses caused by typographical error.
- Address duplication problems that are caused by dozens of streets with the same name

4. Related Work

Most of the previous related works and their existing methods mainly identify crime hotspots based on the location of high crime density without considering either the crime type or the crime occurrence date and time. For example, related research work containing a dataset for the city of Philadelphia with crime information from the year 1991 - 1999 [1]. It was focusing on the existence of multi-scale complex relationships between both space and time . Another research titled “The utility of hotspot mapping for predicting spatial patterns of crime” looks at the different crime types to see if they differ in their prediction abilities[1] . Other existing works explore relationships between the criminal activity and the socio-economics variables such as education, ethnicity, income level, and unemployment [19].Despite all of the existing work, none of them consider the three elements (location, time, crime type) together. In addition, there is very little research that can accurately predict where crimes will happen in the future [7]. In our work we tried to provide a data-mining model which can be helpful for crime prediction based on crime types and using spatial and temporal criminal hotspots. So we can use the Co-location algorithm[15] for mining frequent patterns and here the goal is to come up with a list of all crime hotspots along with its related frequent time and then this result can be compared to Naïve Bayesian classifier, Decision tree classifier, etc. For the comparison precision, recall, F-1 Score, Support, etc. can be used.

5. Basic Concepts

5.1 Apriori algorithm

Apriori algorithm is a classic algorithm for generating association rules from datasets or databases[13]. The key idea of the algorithm is to begin by generating frequent itemsets with just one item and to recursively frequent itemsets with 2 items, then frequent 3-itemsets and so on till some stopping criterion is met. This is where computational complexity comes into the game. Apriori algorithm is based on very simple observation: subsets of frequent itemsets are also frequent itemsets. In other words, if some itemset is proven to be non-frequent, then it will not be considered by algorithm for further itemsets. To identify the k-itemsets that are not frequent, the algorithm needs to examine all subsets of size (k-1) of each candidate k-itemset. It generates candidate itemsets of length k from item sets of length k-1[13]. Then it does not consider the candidates which have an infrequent sub-pattern.

5.2 Colocation

Given a collection of boolean spatial features, the co-location pattern discovery process finds the subsets of features frequently located together[15]. Boolean spatial features is represented by the presence or absence of geographic object types at different locations in a two dimensional or three dimensional metric space, such as the surface of the Earth. Examples of Boolean spatial features are plant species, animal species, road types, cancers, crime, and business types. Co-location rules are models to infer the presence of spatial features in the neighborhood of instances of other spatial features. For example, “Nile Crocodiles Egyptian Plover” predicts the presence of Egyptian Plover birds in areas with Nile Crocodiles. The spatial co-location rule problem is different from the association rule problem, since there is no natural notion of transactions in spatial data sets which are embedded in continuous geographic space. Therefore a transaction-free approach should be used to mine co-location patterns by using the concept of proximity neighborhood. For spatial co-location patterns, a participation index is used. The participation index is used as the measure of prevalence of a co-location for two reasons[15]. First, this measure is closely related to the cross- function, which is often used as a statistical measure of interaction among pairs of spatial features. Second, it also possesses an anti-monotone property which can be exploited for computational efficiency[15].

5.3 Colocated Instances

Colocated instances are the clusters of the colocation patterns found. So once we run the co-location algorithm, we end up with a bunch of co-located crimes. Then we consider the data for only those co-located crimes. Then we cluster them by forming connected components. Each such connected component formed is a colocation instance. The main idea is to perform breadth first search based on the distance between two crimes. Crimes are considered to be neighbors if the distance between the two crimes are less than a certain threshold. So on connecting the neighbor crimes we end up with colocated instances.

5.4 Minimal Enclosing Circle (MEC)

A minimum enclosing circle is a circle which contains all the points (lie either inside the circle or on its boundaries). On getting the colocated instances, the Minimal Enclosing Circle of those instances are found. The task is to find the centre of the minimum enclosing circle (MEC). MECs are found using Welzl's Algorithm [18]. The idea of the algorithm is to randomly remove a point from the given input set to form a circle equation. Once the equation is formed, check if the point which was removed is enclosed by the equation or not [18]. The Input here is the set of points with X and Y coordinates and the output is the center and radius of MEC. It has a time and space complexity of O(n).

5.5 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) groups together points that are close to each other based on a distance measurement (based on a distance measure) and a minimum number of points. It also marks as outliers the points that are in low-density regions [17]. The DBSCAN algorithm basically requires 2 parameters:

eps: defines the number of points should be close to each other to be considered a part of a cluster. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbors[17].

minPoints: the minimum number of points to form a dense region. For example, if we set the minPoints parameter as 3, then we need at least 3 points to form a dense region.

5.6 Spatial Scan Statistics (SatScan)

The main goal of SatScan is to omit chance clusters. The steps in the algorithm can be represented as follows [20]:

Enumerate candidate zones & choose zone X with highest likelihood ratio (LR)

- LR(X) = p(H1|data) / p(H0|data)
- H0: points in zone X show complete spatial randomness (CSR)
- H1: points in zone X are clustered

If LR(Z) >> 1 then test statistical significance

- Check how often is LR(CSR) > LR(Z) using 1000 Monte Carlo simulations [20]

6. Dataset

In our study, we are planning to use two different datasets for real-word crimes in two cities of the US. We chose those cities from different states: NYC, Boston. To construct our data mining models, we mainly focused on the NYC dataset. After we have built the desired models, we plan to apply the same strategy to train the required models on the Boston dataset.

7. Proposed Algorithm

The first step in our project is to identify good quality datasets. We have identified some crime datasets like Boston, NYC and LA crime dataset. After that the dataset needs to be cleaned by removing redundant rows and rows having missing features. Then some pre-processing on the dataset needs to be done by merging some of the columns into one and removing the unnecessary columns. Then we do exploratory data analysis to find interesting features. The next step is to identify the associated patterns(crimes) by using Apriori algorithm and Colocation algorithm. Then we Identify the center of colocated instances by finding the center of the minimal enclosing circle. Then we applied SatScan/DBSCAN to detect the hotspot of the associated crimes. The proposed steps can be seen as a flowchart in Figure2.

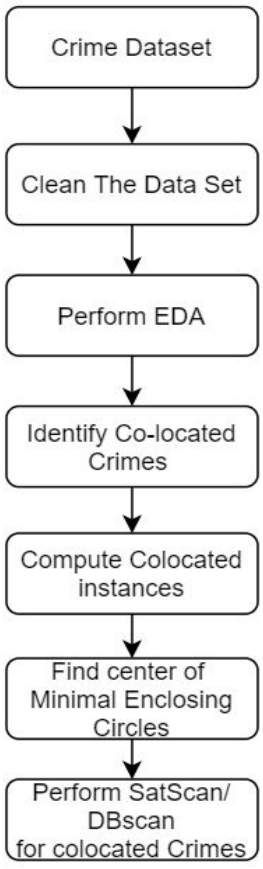

Figure 2

7.1 Toy Example

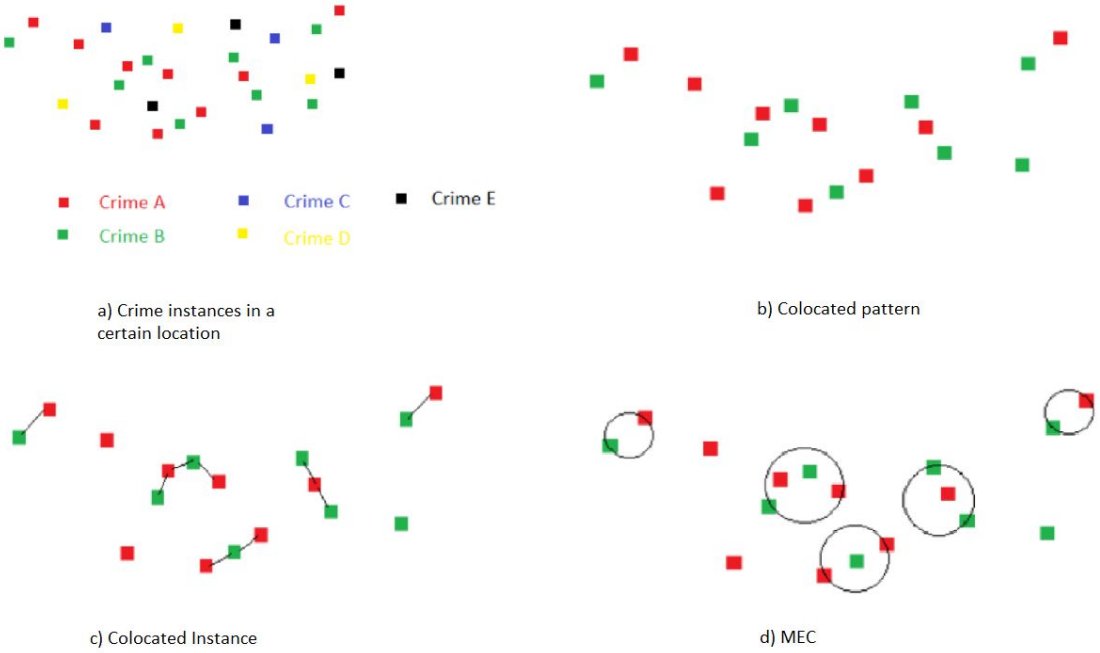

Figure 3

Given a set of crime data in a certain location we compute the center of the MEC's which is fed to DBScan and SatScan to find the hotspots of the co-located crimes. As we can see in the toy example in Figure 3a) we have five types of crime. Then it is found that crime A and crime B are co-located which can be seen in Figure 3b). Then for that pattern, the colocated instances are first found which can be seen in Figure 3c), and finally the minimum enclosing circles are found for those colocated instances (Figure 4d).

8. Contributions

We found the colocated crimes and we extended the point based clustering algorithms to circles (i.e. Colocation of points will be considered as a circle) and found the hotspots in the datasets.

9. Results and Validation of our contribution

In order to validate our method , we have performed our approach on the NYC dataset. Before going to the details of the result we would like to go through the details of the Dataset. The dataset contains the reports of NYC crimes from 2010-2015. But for the simplicity we have considered the crimes of 2015. There is a difference between this data and the original from Kaggle. We have applied some preprocessing of data by changing (hours and minutes to time...) and some are added (dayPart by the division of time) . Variables "hours" and "minutes" are joined into 1 continuous variable time - for instance: 15h 30min is now 15.5 (15 + 30/60). Variable "time" is divided into the categorical variable "dayPart" with 4 parts of the day for more simplicity. All NA's columns are removed from the dataset. The dataset contains the following features: date/time of crime, Latitude/Longitude, NewYork borough, location, crime description, offense level, premise description, etc. We have almost a million records. We have over 30 types of crime. We have categorized the offence level of the crime to Felony, Misdemeanor, Violation. We can see a sample plot between the crime count and offense level in Figure 4.

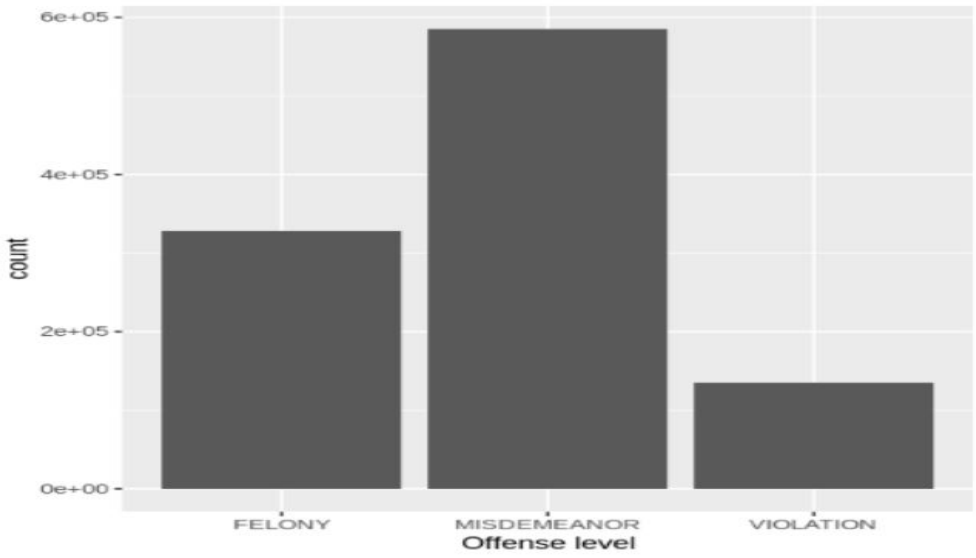

Figure 4

As shown in Figure 5, we can see how the crime is distributed in different boroughs.

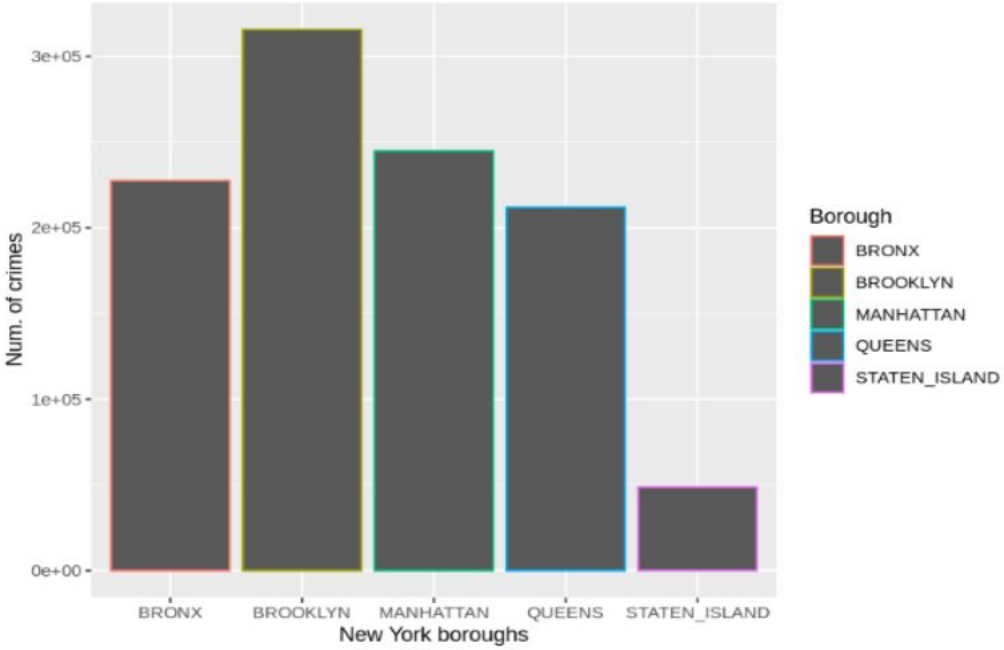

Figure 5

We can see a plot of offense level with time of the day overall through the new york in Figure 6.

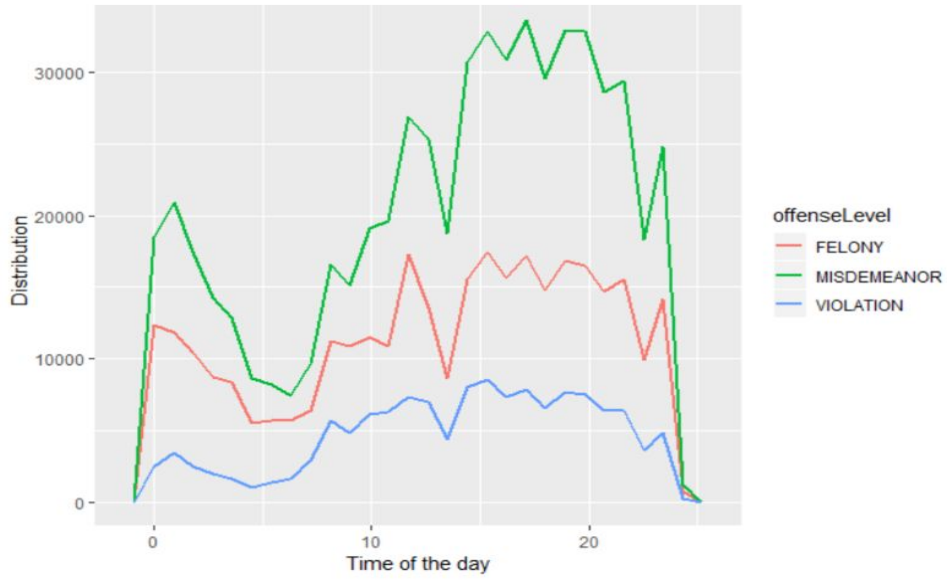

Figure 6

Figure 7 shows a plot of the crimes with respect to the borough offense level and the time of the day.

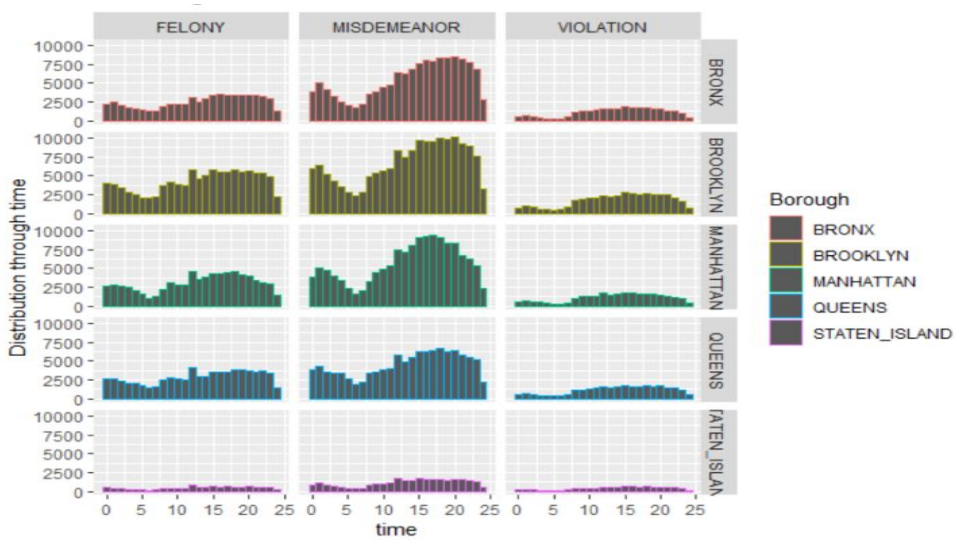

Figure 7

Once we computed the EDA of the data, we applied both colocation and Apriori Algorithm to find the interesting crime patterns. Since we used borough-transaction based association and there is a risk of Gerrymandering (results are sensitive to the choice of transaction boundaries for spatial partition), so we used colocation detection algorithm results. We found many interesting patterns like (FRAUDULENT_ACCOSTING,PROSTITUTION_AND_RELATED_OFFENSES), (GRAND_LARCENY,PETIT_LARCENY),(FELONY_ASSAULT,ROBBERY),(CRIMINAL_MISCHIEF_AND _RELATED_OF,HARRASSMENT_), etc. For simplicity we considered one of the patterns (GRAND_LARCENY, PETIT_LARCENY) and moved forward with it. It has a participation index value of 0.9875.

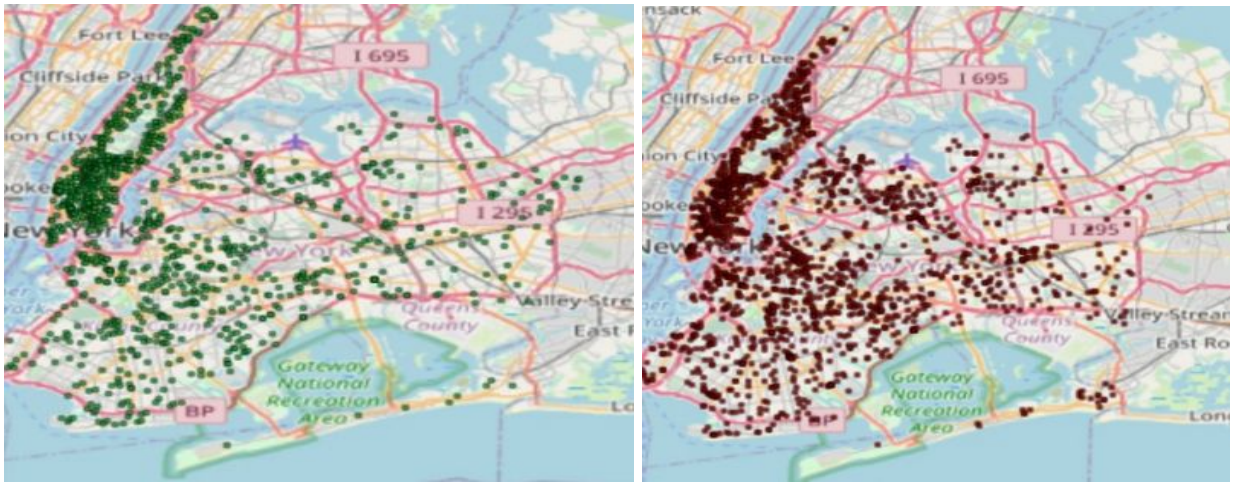

Figure 8 [Left : Grand Larceny | Right: Petit Larceny]

Then we identified all the crime locations having crime types GRAND_LARCENY, PETIT_LARCENY. Since we have a large number of records we considered the crimes that happened between 5pm-6:00am in the month of january 2015. The red points in Figure 8 shows the petit larceny whereas the green points in the Figure 8 shows grand larceny happened in NYC in the above mentioned time frame. Then we merged the colocation instances that are happening within 0.1 miles radius and computed the minimum enclosing circle that contains them. Then we passed the centre of the colocated instances to the clustering algorithm. We can see the result that we obtained by applying SatScan in Figure 9.

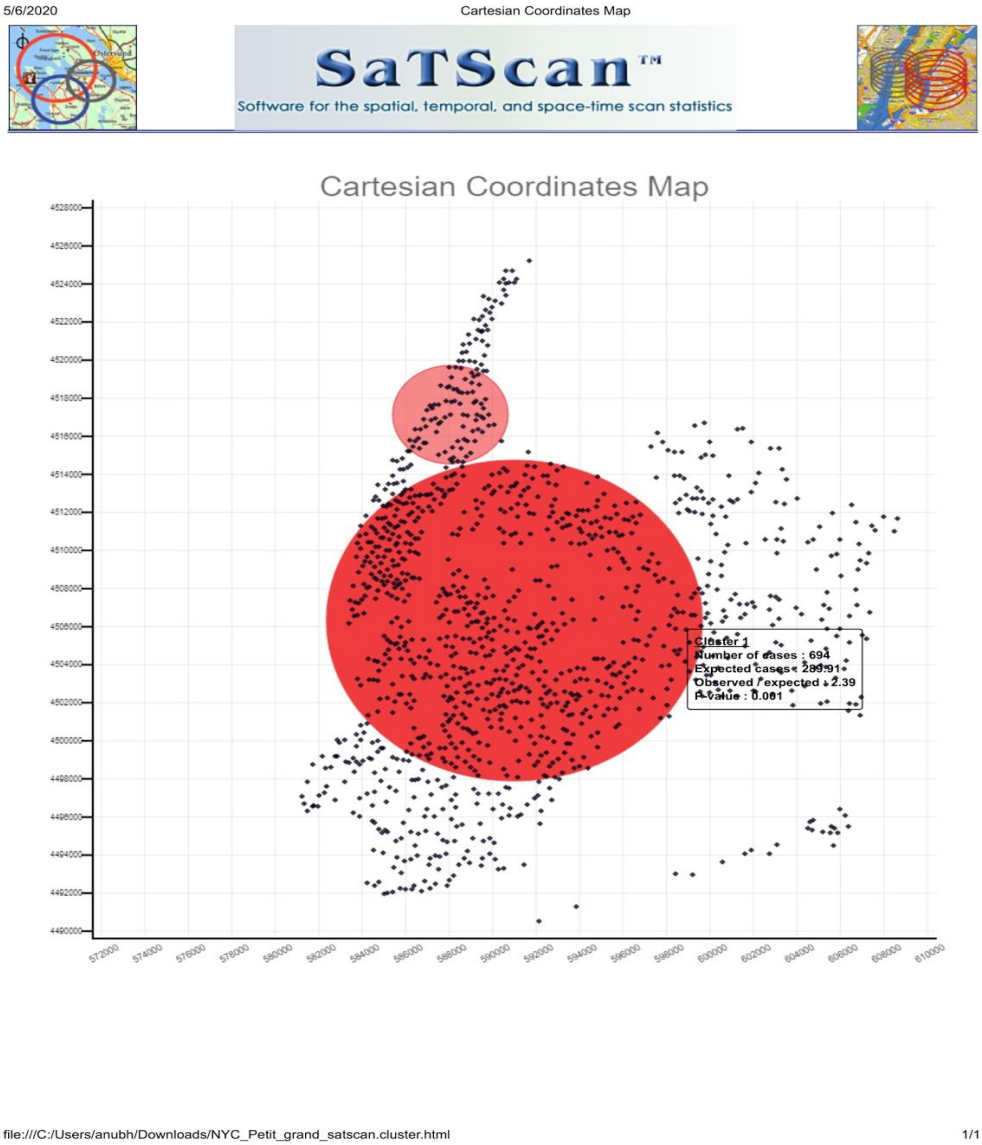

Figure 9

As we can see we got 2 circular shaped clusters by applying SATscan having p values 0.0001 respectively. Then we applied dbscan data and we got the following result.

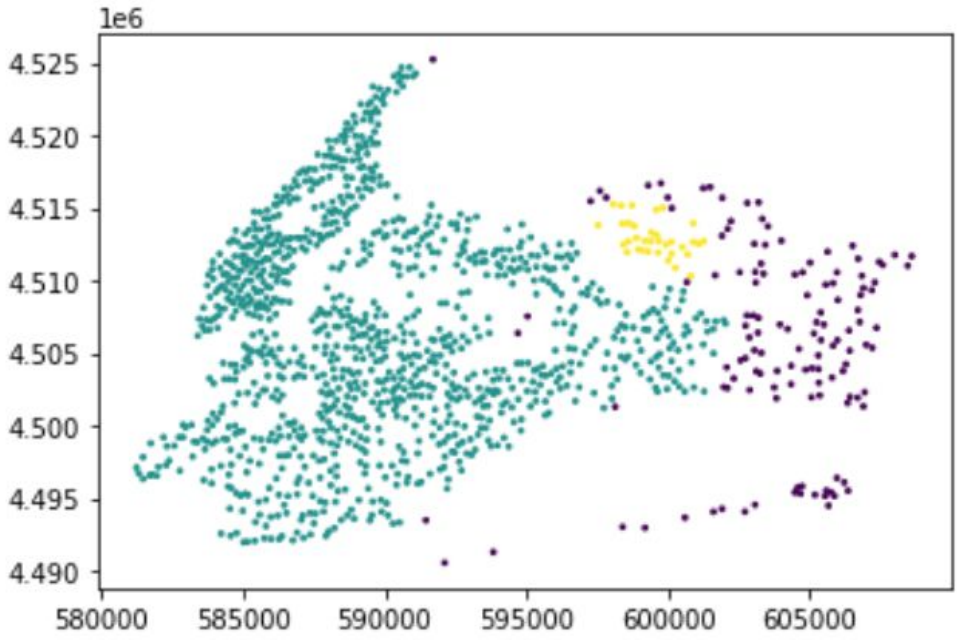

Figure 10

As can be seen from Figure 10, DBSCAN identified two clusters as well. The clusters are represented in the following colors (green, yellow), whereas the purple shows the outlier/noise points. As we know the SatScan computes likelihood ratio, density inside the circle, density outside the circle, log likelihood ratio for statistical significance of the identified clusters. But it has a limitation of identifying the clusters in the predefined shapes, whereas DBSCAN identifies the clustering in a density based manner. As we can see here, DBSCAN Is doing a better job compared to SatScan. The complete implementation can be found at https://github.com/anubhavpanda2/Crime-Hotspot-Detection-For-Interesting-Patterns

10. Conclusions and Future Work

In this project we used Colocation to find out the associated crimes and its hotspot by using SatScan and DBSCAN. Given a set of crimes our method follows the following steps: 1) Distribution of data by conducting Exploratory Data Analysis 2) Associated crimes having certain Participation ratio within a certain distance threshold. 3) Center of the circle containing the colocated Instances. 4) Hotspots by applying SatScan or DBSCAN. We have to choose between the above two algorithms based on the distribution of data. As we know, the SatScan cluster has statistical significance whereas DBSCAN provides clusters based on a density based approach. SatScan clusters are based on predefined shapes whereas DBSCAN clusters

do not have any predefined shapes. So based on the nature of the data we can choose SatScan or DBSCAN. Some of the future works include extending the SatScan from circles to other polygons. Also instead of approximating circles with the circles' centers, work can be done to feed the circles directly into SatScan and hence generalizing SatScan for polygons. The proposed method can also be extended to other crime datasets like the Boston and LA datasets. We are also planning to extend this approach of finding hotspots of colocation patterns to real world scenarios like finding the hotspots of the corona outbreak as well. This work can be extended with the use of Significant DBSCAN [16] which is a statistical, density based clustering algorithm. Further works can be done on the improving the computational aspect of the colocation detection algorithm.

Reference

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland, 'Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data', CoRR, vol. 14092983, 2014.

[2] R. Arulanandam, B. Savarimuthu and M. Purvis, 'Extracting Crime Information from Online Newspaper Articles', in Proceedings of the Second Australasian Web Conference - Volume 155, Auckland, New Zealand, 2014, pp. 31-38.

[3] A. Buczak and C. Gifford, 'Fuzzy association rule mining for community crime pattern discovery', in ACM SIGKDD Workshop on Intelligence and Security Informatics, Washington, D.C., 2010, pp. 1-10.

[4] M. Tayebi, F. Richard and G. Uwe, 'Understanding the Link Between Social and Spatial Distance in the Crime World', in Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12), Redondo Beach, California, 2012, pp. 550-553.

[5] S. Nath, 'Crime Pattern Detection Using Data Mining', in Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on, 2006, pp. 41,44.

[6] Crimereports.com, 2015. [Online]. Available: https://www.crimereports.com. [Accessed: 20-May2015].

[7] S. Chainey, L. Tompson and S. Uhlig, 'The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime', Security Journal, vol. 21, no. 1-2, pp. 4-28, 2008.

[8] Data.denvergov.org, 'Denver Open Data Catalog: Crime', 2015. [Online]. Available: http://data.denvergov.org/dataset/city-and-county-of-denver-crime. [Accessed: 20- May- 2015].

[9] Imgh.us, 2015. [Online]. Available: http://imgh.us/neighborhood_map.jpg. [Accessed: 20-May2015].

[10] O. Knowledge, 'Crime — Datasets - US City Open Data Census', Us-city.census.okfn.org, 2015. [Online]. Available: http://us-city.census.okfn.org/dataset/crime-stats. [Accessed: 20- May-2015].

[11] Laalmanac.com, 'City of Los Angeles Planning Areas Map', 2015. [Online]. Available: http://www.laalmanac.com/LA/lamap3.htm. [Accessed: 20- May- 2015]. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015

[12] Data.denvergov.org, 'Denver Open Data Catalog: Census Neighborhood Demographics (2010)',2015.[Online].Available:http://data.denvergov.org/dataset/city-and-county-of-denver-censusneighborhood-demographics-2010. [Accessed: 20- May- 2015].

[13] GitHub, 'asaini/Apriori', 2015. [Online]. Available: https://github.com/asaini/Apriori. [Accessed: 20- May- 2015].

[14] Scikit-learn.org, '3.3. Model evaluation: quantifying the quality of predictions — scikit-learn 0.17.dev0 documentation', 2015. [Online]. Available: http://scikit-learn.org/dev/modules/model_evaluation.html. [Accessed: 20- May- 2015].

[15] Huang, Y., Shekhar, S., & Xiong, H. (2004). Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and data engineering*, *16*(12), 1472-1485.

[16] Xie, Yiqun, and Shashi Shekhar. "Significant DBSCAN towards Statistically Robust Clustering." *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*. 2019.

[17] Kelvin Salton do Prado, Medium, “How DBSCAN works and why should we use it?” ”https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80”, 2017.

[18] Welzl, Emo. "Smallest enclosing disks (balls and ellipsoids)." *New results and new trends in computer science*. Springer, Berlin, Heidelberg, 1991. 359-370.

[19]Almanie, T., Mirza, R., & Lor, E. (2015). Crime prediction based on crime types and using spatial and temporal criminal hotspots. *arXiv preprint arXiv:1508.02050*.

[20] https://studyres.com/doc/3556519/part-1