RESEARCH REVIEW JOURNALS
Volume-04
Issue-05
May-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[UGC Listed Journal]

Crime Hot Spot Identification Using Kernel Density Estimation

[1]Pranay Ghosh & [2]Dr. Jitendra Sheetlani

[1]Research Scholar, Department of Computer Science, Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P.
[2]Research Guide, Department of Computer Science, Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P.

ARTICLE DETAILS

Article History
Published Online: 25 May 2019

Keywords
Crime, Hot, Spot, Kernel, Density

ABSTRACT

A crime detection approach known as hot-spot mapping, which aids police officers in identifying high-crime areas and responding more effectively. The Modified Cut Clustering method (MCC) was used to discover crime locations with less manpower, and the findings were discussed in the preceding chapter. However, when dealing with huge amounts of crime data, the MCC does not enable efficient grouping of crime site facts. This problem is addressed in the proposed study by introducing a methodology known as the social crime data Aware Kernel Density Estimation based Serial crime Detection approach (SAKDESD), which is used to classify comparable crimes based on their degree of similarity. The serial crime data set, as well as the social data collection, are used to group serial offences in this study.

1. Introduction

Hotspot Analysis

With the rapid advancement of technology, a computer-based approach for discovering, visualising, and investigating illegal activities has emerged. Suspect is a term used in criminology to describe the person who committed the crime. The person who is the target of the crime is known as the victim. A crime hot spot is described as a region having an extremely high rate of criminal activity [1].

A geographical cluster of crime, or a large number of crimes in a certain location at a specific time, is referred to as a cluster of crime features. Only by examining and looking for a large number of evidences related to the crime scene can a thorough investigation be conducted. A geo-spatial plot of the crime on the map of the police jurisdiction is frequently used to visually show such clusters [2].

Serial crimes are repeated threats made by the same person in a similar manner in various locations. To ensure public safety, these offences must be discovered and addressed. Investigation departments are in charge of ensuring social safety by locating criminals who are accountable for a variety of hazards that occur around the world [3].

Detecting serial crime hotspots and the people involved would be a time-consuming procedure that would need to be analysed and processed quickly. The investigation procedure is less efficient due to a lack of manpower and a large amount of data concerning crimes that occur in numerous areas. Several attempts have been made by various researchers to resolve these challenges by employing data mining technologies to make investigative processes easier [4].

Using the kernel density estimation methodology, comparable types of crimes are recognised and aggregated at various places in terms of criminal attributes, overcoming obstacles such as grouping and overlapping problems encountered by the modified graph cut clustering (MGCC) technique. Data about social crime would be in an unstructured format that would need to be pre-processed in order to be handled efficiently. To find the most comparable subjects in the crime data set, the latent semantic technique is used. As a result of this effort, the investigation methods involved in the detection of serial crimes can be more flexible [5].

Social Crime Data Aware Kernel Density Estimation

Crimes are a type of behaviour problem that has a number of negative consequences for our society. Those offences must be detected, and the perpetrator must be recognised, in order to prevent future criminal acts by the same individuals. Depending on the structure and features of the crime, it may take several forms. “Personal trait crime” and “serial crime” are two of the most common types of crimes that occur in various regions. Individuals that commit personal trait crimes do so for their own personal motivations. Serial crimes are crimes that are committed repeatedly by an individual or a group of persons in different locations. Serial crimes are even more harmful than personal trait crimes, and they must be discovered in order to ensure security [6].

Serial crimes can be identified by looking for similarities in the characteristics of crimes that are occurring in multiple areas. If the crime locations are divided into different types and represented in distinct regions in the visualization, the crime mapping will be successful. This method works well with structured serial crime data, but not so well with unstructured data such as social media data. The serial crime data set acquired from the police department will include criteria such as crime details and behaviours, which are insufficient for accurately forecasting repeated offences [7].

The suggested study introduces social crime data aware kernel density estimation based serial crime detection algorithms, which would find serial crimes by identifying similar features that exist in crimes of type T that occurred at different geographic coordinates p. Social media data concerning crime behaviour is gathered from different social media sites and combined with serial crime data to accurately forecast serial crime. The unstructured format of social media data is pre-processed to remove extraneous words and display the dataset in a structured format. Following pre-processing, Latent dirichlet allocation (LDA) is used to identify the most essential elements in the document based on their importance. The kernel density for the retrieved features would then be

determined using LDA. This kernel density is derived using the serial crime features that can be found in the serial crime data set. Finally, the spatial point 'p' with the highest kernel density is chosen as the serial crime hotspot [8].

The flow diagram of the proposed work is given in Figure 1 as follows:

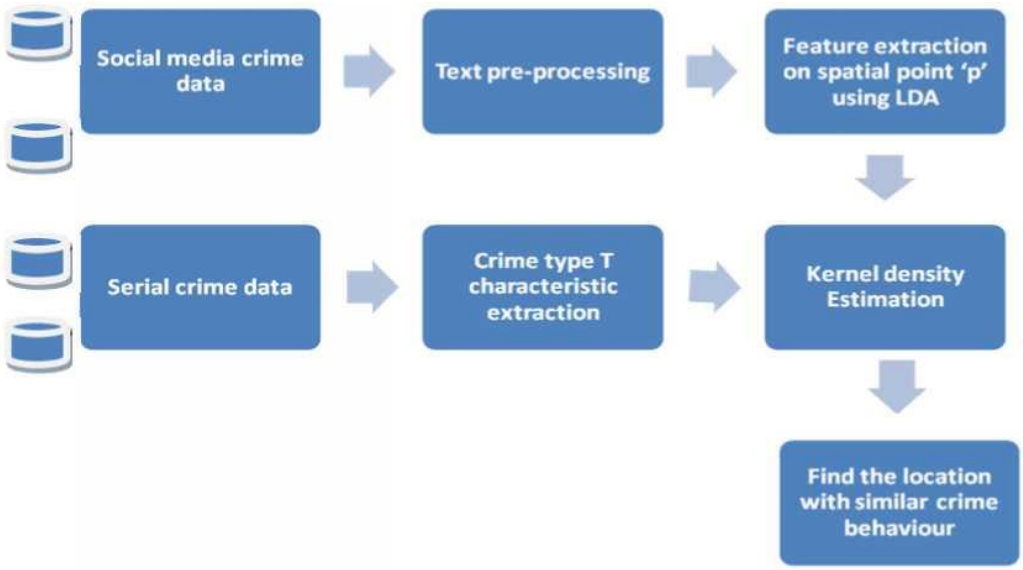

Figure 1. Flow of kernel density estimation based serial crime detection

The above flow diagram provides the steps in serial crime detection based on public comments on social media data in different locations. This process flow is explained in the following sub sections.

2. Methods:

Data Collection

Two types of data sets, namely the Crime Data Set and the Social Media Crime Data Set, are studied in this proposed research study for locating serial crimes that are occurring in various crime sites and sharing comparable features.

The crime data set studied in this study includes qualities such as murder, dacoity, robbery, theft, and breaking and entering, as well as latitude and longitude values of crime, which indicate the actual location where the crimes occurred. The different sorts of crimes are examined using this data, and then related sorts of crimes are grouped together to forecast serial crimes.

The social media crime data sets are acquired from various public social media web sites such as Facebook and Twitter, and they consist of numerous user comments about crimes occurring in various locations, which can provide in-depth information about repeated crimes. It has the potential to expose the semantic significance of crime data from various crime locales. It would be in an unstructured manner with a slew of irrelevant tags and terms. By using the latent semantic approach behaviour for detecting the similarity topics contained in the social crime data set, the unstructured social crime data set is pre-processed to get meaningful structured format, which can lead to accurate prediction and efficient grouping of serial offences.

After obtaining the data, pre-processing will be performed on the social media data set in order to represent it in a structured fashion, allowing semantic meaning to be extracted.

Text Pre-Processing

The most important procedure for converting unstructured data into structured format is text pre-processing. It will also remove any redundant terms from the data set that are unrelated to the notion. Noises, html tags, and advertisement messages would make up the social crime data obtained from online social media web sites. This could result in a large dimensionality difficulty when processing the social crime data set, since each tag would be treated as a single feature. The kernel density would be calculated for those features that aren't needed, increasing the calculation overhead. The following steps would be included in the text pre-processing:

Text cleaning on the internet: The process of removing online text characters such as html tags, scripting tags, and advertisement contents is known as online text cleaning. This stage will get rid of any content that isn't related to the concepts.

White space removal: After doing online text cleaning, white space removal is performed to ensure that the documents are fully represented. The removal of white space is utilized to generate entire phrases from which semantic content can be retrieved in a flexible manner.

Expanding abbreviations: This stage involves processing the sentences from beginning to end in order to expand abbreviations. The whole meaning of the document can be retrieved by abbreviating the sentences.

Stemming: Stemming is the process of avoiding the use of derivative terms in order to avoid repetition. The Porter stemming method is utilized in this study to accomplish the stemming process, which reduces inflected words to their basic words.

The procedure of removing the starting and finishing words from sentences from which the meaning of the material cannot be derived is known as start and stop word removal. 'A, an, the, and so on' are some of the first words. The stop words evaluated in this study are 'the, is, at, which,' which are

eliminated before the social media crime data is processed.

Feature Detection using Latent Drichlet allocation

Latent drichlet allocation is a generative model based on natural language processing that is used to observe the more similar parts of the documents that are fed as input. In two ways, LDA improves the document similarity discoveries. These are the distribution of words and the distribution of topics. This would be determined by comparing the documents to the details of crime type T from the crime data set. The documents in this section contain public opinions concerning crime statistics acquired from social media platforms. The likelihood of corresponding crime type T belonging to the words of document D is defined as word distribution. The likelihood of corresponding crime type T belonging to document D is defined as the topic distribution. This is accomplished by submitting and analysing various document subjects. The following is the functioning procedure:

Algorithm 1:

Input : Social crime data set gathered in particular time period, Crime types

Output : Features

Repeat

Collect the input documents and initialize weight values of all documents as null

Find the log likelihood to find the words that are most related

$$-2\ln\lambda = 2^{\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)} \quad \text{.....(1)}$$

Assign the temporary crime types for every word that has more likelihood present in the user review

a. If any is repeated multiple times assign with different crime types

Find the similarity of the word and crime type

8. Update the weight values of reviews-based similarity

9. until final solution obtained

Where

$-2\ln\lambda \rightarrow$ Log likelihood ratio

$O_i \rightarrow$ Observed value

$E_i \rightarrow$ Expected value

Expected Value

The aforesaid algorithm generates a pseudo code for the latent Dirichlet allocation technique, which is used to locate the most related and relevant phrases in the social crime data set in terms of the various sorts of crimes represented in the serial crime data set. This procedure will assign a crime category label to each and every word in the document. This technique would be repeated for each sort of crime evaluated in this study's methodology. Finally, all words labelled with associated crime types T will be considered characteristics f (p) of the crime types at a specific spatial point p. These traits would be used to describe serial crimes that occur in various parts of the world in a comparable way. The place where these crimes occur more frequently would be found based on these features by computing the kernel density of those specific traits.

Kernel Density Estimation of Features

In terms of statistics, kernel density estimation is defined as a non-parametric method for determining the probability density distribution of a related feature in a certain region. The density level of the features f (p) at the relevant geographic location 'p' in terms of crime type T is determined using KDE in this study. KDE is a data smoothening strategy that can discover the density estimation of a related feature on a certain spatial location while ignoring the noise present in the region. This smoothing is accomplished in this study by adjusting the bandwidth parameter 'h' to its ideal value. This 'h' value denotes the amount of data that needs to be processed on the surface.

KDE estimation would be done for all features values in terms of every crime types T within a particular time period. The equation is used to estimate the kernel density of features at the spatial point 'p' for the crime types that are considered and calculated by using the following formulae:

$$f(p) = k(p,h) = \frac{1}{Ph}\sum_{j=1}^{p}\left(\frac{\| p - p_j \|}{j}\right)$$

.....(2)

Where

p→ spatial point in which density to be calculated

h→ bandwidth of KDE used to smoothen the data processing problem

P → Total number of crime types T that are considered

j→ single crime location during the time period

K → density function

||.|| → Euclidean distance

The probability of possibility of occurrence of a given feature in the geographic point 'p' in terms of different crime types T is calculated in this research study utilizing the probability density function approach.

The following algorithm depicts the complete work flow of this kernel density based serial crime detection in social crime data.

Algorithm 2:

Gather the serial crime data from the different social web sites

Pre-process the data gathered from the social media web sites

2. a. Remove the html tags, advertisement

2. b. Remove the white spaces present in the data set

2. c. Abbreviate the acronyms

2. d. Remove the start and stop words from the data sets

2. e. Apply porter stemmer algorithm to remove the stemming words

3. Find the most relevant features that are related to the crime types present in the data set by using LDA

4. For every features fi∈F

5. Find the kernel density function k of features f in spatial point p for every crime type T

$$f(p) = k(p,h) = \frac{1}{ph}\sum_{j=1}^{p} K \quad \text{.....(3)}$$

6. End for

7. Return spatial point p with more kernel density

The procedure above calculates kernel density estimate

for features at distinct spatial points p in a given time window. This method yields more accurate results in determining the places where serial crimes occur most frequently. This proposed study study is executed in the matlab simulation environment in terms of performance measure values, which are evaluated and compared to the existing technique, which is detailed mentioned below.

3. Result and Discussion:

The social crime data collection, which is collected from several social media web sites, was used to conduct the experimental test for this work. The social crime data set contains information about the nature and types of crimes that can occur in a variety of locales and at various times in the public eye. The proposed research for this project aims to identify the areas where serial crimes occur more frequently. This is discovered by studying and predicting a variety of attributes from a social crime data set that is most comparable to crime type T. This process is tested in a Matlab simulation environment and compared to the previous method. This performance test is done to show that the proposed work in this research, dubbed Social crime data aware kernel density estimation based serial crime detection technique (SAKDESD), will outperform the existing technique, modified cut clustering algorithm (MCC).

The mantel index and the jaccard index are performance indicators that are used to evaluate performance. This performance analysis is presented in a graphical format, which is explained in depth in the parts that follow.

Data Set

In this study, the social crime data set and the acquired dataset from police stations are used to predict where serial crimes are most likely to occur in the real world. This data collection is compiled from a variety of social media websites in terms of various criminal acts taking place in various crime hotspots. This data set would include information such as the type of crime, the number of crimes that occurred within a given time period, and people's reactions to specific crimes that occurred in various locations. The collected crime data set from several areas in Coimbatore, India, was used to analyse and predict a series of burglaries. The many sorts of crimes are assessed using this information, and then comparable sorts of crimes are grouped together to predict the various crimes.

Mantel Index calculation for SAKDESD and MCC

The mantel index is a statistic for calculating the association between several features found in crime types that are comparable. The proposed research technique's Mantel index should be higher than the existing research strategy, which reflects a high level of data correlation. The mantel index is calculated as follows:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(x_{ij} - \bar{x})}{S_x} \cdot \frac{(y_{ij} - \bar{y})}{s_y} \quad \text{.....(4)}$$

Where

x, y =variables measured at locations i, j

n = number of elements in the distance matrices

S_x, S_y = standard deviation of variable x and y

$\bar{x}, \bar{y}$ = mean value of variables x and y.

The actual values obtained for the mantel index is given in table 1.

Table 1. Mantel Index values attained by SAKDESD and MCC

Number of Centre Data Points	Mantel Index	
	SAKDESD	MCC
2	0.75	0.48
4	0.79	0.63
6	0.84	0.76
8	0.87	0.78
10	0.91	0.82
12	0.94	0.85
14	0.94	0.88
16	0.94	0.90
18	0.98	0.92
20	1	0.94

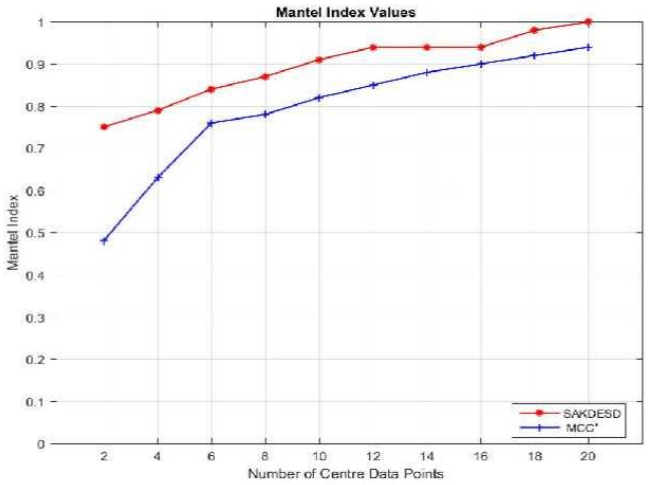

Figure 2. Comparison on Mantel Index Measure(SAKDESD and MCC

Figure 2 compares and evaluates mantel index values as well as existing and suggested research scenarios. The number of data points is counted on the x axis, whereas mantel index values are counted on the y axis. The planned MWMO-MCC has an average Mantel index of 0.90, while the present GCC has an average Mantel index of 0.80. As a result, the total Mantel index values of the suggested MWMO-MCC technique are 11.11 percent higher than the GCC technique. As a result, it can be demonstrated that the proposed research method, SAKDESD, produces superior results than MCC.

Jaccard Index calculation for SAKDESD and MCC

The jaccard index is used to represent the similarity between the data points. The jaccard index value is used to measure how much the crimes happened in different locations are matched with each other. The jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \text{.....(5)}$$

Where

A, B= Data points

The actual values of Jaccard index obtained for both existing and proposed approach is indicated in table 2.

Table 2 .Jaccard Index values attained by SAKDESD and MCC

Number of Centre Data Points	Jaccard Index	
	SAKDESD	MCC
2	0.77	0.55
4	0.84	0.57
6	0.86	0.61
8	0.93	0.61
10	0.95	0.78
12	0.99	0.79
14	1	0.84
16	1	0.89
18	1	0.94
20	1	0.96

The next picture 4.3 shows a graphical depiction of the comparison of the proposed study work with the current research work for the above-mentioned actual values.

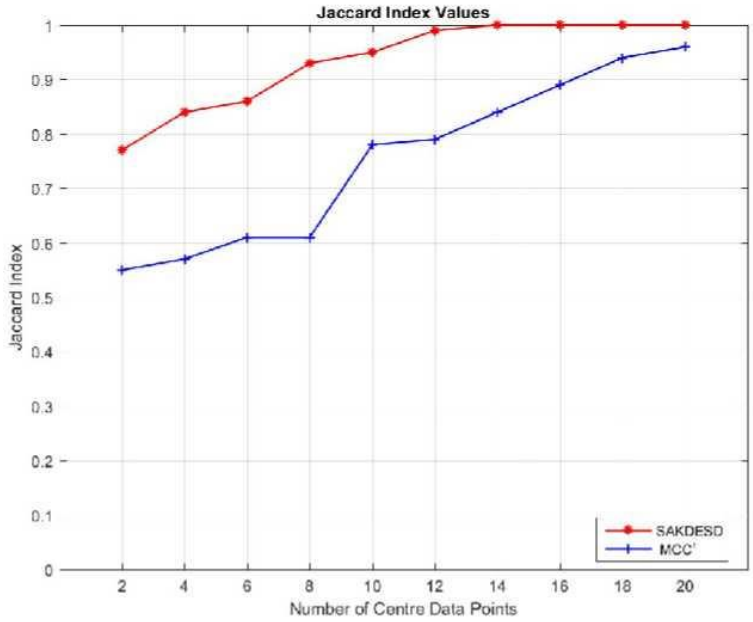

Figure 3. Comparison on Jaccard index Measure(SAKDESD and MCC)

The jaccard index values are analysed, and the existing and suggested research scenarios are compared, as shown in Figure 4.3. The number of centre data points is represented on the x axis, while the jaccard index values are represented on the y axis. The suggested SAKDESD method has a Jaccard Index of 0.93, while the present MCC has a Jaccard Index of 0.75. It is known that the suggested SAKDESD's overall jaccard index value is 19.35 percent higher than the MCC approach. As a result, it can be demonstrated that the proposed research method, SAKDESD, produces superior results than MCC.

Figure 4 shows the GIS representation of clustered results of crimes which were happened in the various crime locations are depicted using MCC as follows:

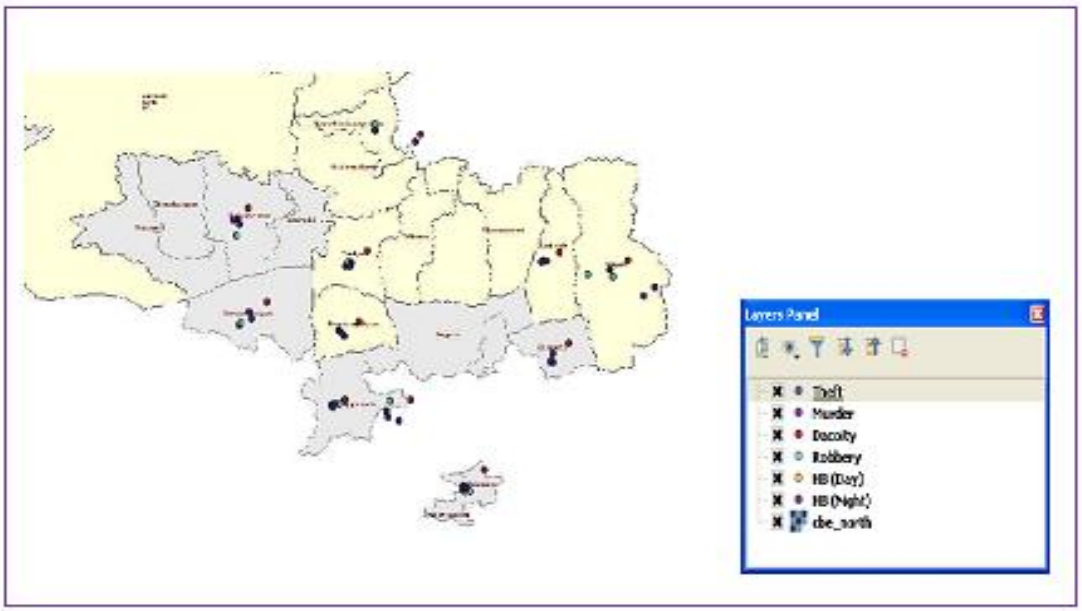

Figure 4. Clustering crime spots using MCC

MCC appears to cluster crime hotspots effectively, however the overlapping problem lowers overall performance.

The GIS representation of clustered results of crimes that occurred in various crime sites is presented as follows in Figure 5 using SAKDESD.:

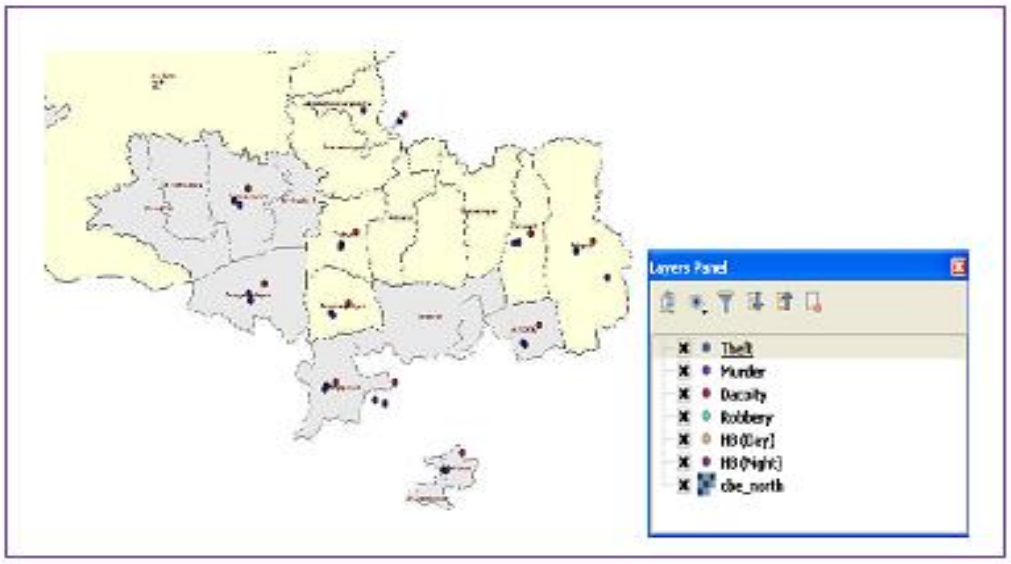

Figure 5. Clustering crime spots using SAKDESD

It may be deduced that the SAKDESD clusters crime spots more effectively and performs better than the MCC, as seen by the graphical representation of performance metrics.

4. Conclusion

The Social crime data Aware Kernel Density Estimation based Serial crime Detection (SAKDESD) approach is examined in this chapter. Serial crime detection is vital in the police investigation department since it is difficult to conduct manually [9]. It has been noted that the proposed study aims to determine the location where serial crimes occur most frequently in terms of crime kinds. This is accomplished by employing a novel method known as social crime data aware kernel density estimation based serial crime detection [10]. The experimental results suggest that the proposed research outperforms the existing strategy in terms of improved performance measures such as the mantel and jaccard index values.

References

1. Soliman, T. H., Al Ommar, K., andMahdy, Y. B. (2015). Developing Spatio-Temporal Dynamic Clustering Algorithms For Identifying Crime Hot Spots In Kuwait. Journal of Engineering Sciences Assiut University Faculty of Engineering, 43(1), 1-15.
2. Bajpai, D. (2012). Emerging Trends in Utilization of Data Mining in Criminal *Investigation: An Overview. Journal of Environmental Science, Computer Science and Engineering and Technology, 1(2), 124-131.*
3. Zhou, G., Lin, J., and Ma, X. (2014).A Web-Based GIS for Crime Mapping and Decision Support. In *Forensic GIS* (pp. 221-243).Springer Netherlands.
4. Shafeeq, A., and Binu, V. S. (2014). Spatial Patterns of Crimes in India using Data Mining Techniques. International Journal of Engineering and Innovative Technology (IJEIT), 3(11), 291-295.
5. Mukaka, M., White, S. A., Mwapasa, V., Kalilani-Phiri, L., Terlouw, D. J., and Faragher, E. B. (2016). Model choices to obtain adjusted risk difference estimates from a binomial regression model with convergence problems: An assessment of methods of adjusted risk difference estimation. *Journal of Medical Statistics and Informatics,* 4(1), 5.
6. Singh, A. K., and Manimannan, G. (2013). Detecting Hot Spots on Crime Data Using Data Mining and Geographical Information System.Int J of Statistika and Mathematika, ISSN, 2277-2790.
7. Law, J., Quick, M., and Chan, P. (2014). Bayesian spatio-temporal modeling for analyzing local patterns of crime over time at the small-area level. *Journal of quantitative criminology,* 30(1), 57-78.
8. Kamiran, F., Karim, A., Verwer, S., and Goudriaan, H. (2012). Classifying sociallysensitive data without discrimination: an analysis of a crime suspect dataset. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on* IEEE., 370-377.
9. Ghorpade, A. S., Sumbe, Y. B. and and Nalavade, J. E. (2014). Identity Crime *Detection for Multiple Applications. International Journal of Current Engineering and Technology, 4(3), 1892-1896.*
10. Bowers, K. J., Johnson, S. D., and Pease, K. (2004). Prospective hot-spotting the future of crime mapping. *British Journal of Criminology,* 44(5), 641-658.

