Detecting Hot Spots on Crime Data Using Data Mining and Geographical Information System

Arbind Kumar Singh[1*], G. Manimannan[2]

[1]Senior Lecturer, Department of Statistics, S.N.S.R.K.S College Saharsa, B. N. Mandal University, Madhepura, Bihar, INDIA.

[2]Assistant Professor, Department of Statistics, Madras Christian College, Chennai, Tamil Nadu, INDIA.

[*]Corresponding Address:

manimannang@gmail.com

Research Article

Abstract: An attempt is made to introduce a new method of mapping the top level crime in various district/cities of Tamilnadu on the basis of crime parameters. Narrative information and crime records are stored digitally across individual police departments, enabling the collection of this data to compile a district wise database of crimes they committed in Tamilnadu. About 35 districts consist of twenty two important crime parameters from Police department database were considered for each year from 2009 to 2011 that could give different idea of the objectives and have important meaning in the literature. The unique feature of this study is the application of factor, k-mean clustering and Geographical Information System (GIS) analyses as data mining tools to develop the hidden structure present in the data for each year. Initially, factor analysis is used to uncover the patterns underlying crime parameters. The scores from extracted factors were used to find initial groups by k-mean clustering algorithm. The clusters thus obtained formed the basis for the further analyses as they inherent the structural patterns found by the factor analysis. Finally, the groups were identified as crimes belonging to High Crime Activity (HCA), Intermediate Crime Activity (ICA) and Low Crime Activity (LCA) in that order, which show the behavior of High Crime Activity cities, Intermediate Crime Activity cities and Low Crime Activity cities. From the present study it was observed that a little over 77 and more percentage of the total variation of the data was explained by the first for factors for each year. These four factors revealed the underlying structural patterns among the twenty two crimes parameter in the analysis. Also only three could be meaningfully formed clusters for each of the years. The results of this data mining and GIS could potentially be used to identify the hot spot and even prevent crime for the forth coming years.

Key Words: Data mining, Crime Parameters, Factor Analysis, k-means Clustering Techniques, Geographical Information System (GIS).

1.0 Introduction

Existing in this world, a lot of unemployment, hunger, starvation, together with the mind goes the wrong way. Terrorism, murder, robbery, theft and rape are growing every day. The amount of data being produced in modern society is growing at an accelerating pace. One of the areas where information plays an important role is that of law enforcement. Obviously, the amount of crime data gives rise to many problems in areas like data storage, data warehousing, data analysis and privacy. Already, numerous technological efforts are underway to gain insights into this information and to extract knowledge from it (Jerome *et.al.*2009*)*.

Intelligence agencies such as the CBI and NCRB (National Crime Record Bureau) are actively collecting and analyzing information to investigate terrorists' activities (Manish Gupta *et. al.)*. Local law enforcement agencies like, SCRB (State Crime Record Bureau) and DCRB (District Crime Record Bureau)/CCRB (City Crime Record Bureau) have also become more alert to criminal activities in their own jurisdictions. One challenge to law enforcement and intelligence agencies is the difficulty of analyzing large volumes of data involved in criminal and terrorist activities (Manish Gupta *et. al. 2008*). It has created its own executive tools to discharge assigned responsibilities. Indian constitution assigns task for maintaining law and order to the states and territories, and almost all routine policing, including apprehension of criminals, is carried out by state-level police forces. The police functioning have remained a constant area of governmental concern and efforts to improve it upon further and further (Chaudhary, 2003; Krishnamorthy, 2003).

Crime mapping and analysis have evolved significantly over the past 30 years. In the beginning, many agencies utilized city and precinct maps with colored pins to visualize individual crime events and crime plagued areas. Today, with the rapid advancement of technology, computer-based techniques for exploring, visualizing, and explaining the occurrences of criminal activity have been essential. One of the more influential tools facilitating exploration of the spatial distribution of crime has been GIS (Ratcliffe and McCullagh, 1999; Harries, 1999). As Murray et al. (2001) note, it is the ability to combine spatial information with other data that makes GIS so valuable. Furthermore, the sheer quantity of information available to most analysts necessitates an intelligent computational system, able to integrate a wide variety of data and facilitate the identification of patterns with minimal effort. Statistical Data mining holds the promise of making it easy, convenient, and practical to discover very large databases for organizations and users. In the present context the problem of crime has been studied, without making any assumptions with regard to the number of groups or any other structural patterns in advance, which reflected the classification of criminal based on certain crime parameters.

The study region of Tamil Nadu is the eleventh largest state in India by area and the seventh most populous state. It is the second largest state economy in India as of 2012. Raghuram Rajan panel report, Tamil Nadu was ranked as the third most developed state in India based on a MDS (Multidimensional Development Index). In the state with more than 72 million populations, and has a police force of 1, 20, 715. Police is a vital component of civil administration in Tamilnadu. The main objective of this paper is to investigate whether;

(i) Data mining paradigms together with well known *unsupervised* statistical model k-mean clustering and Factor Analysis can be used to exhibits the classification and to identify the hidden pattern of crime data.

(ii) A Geographical Information System (GIS) represents to identify the hot spot crime district/cities using final clusters center of k- mean algorithm.

The rest of the paper is organized as follows. Section 2 describes the methodology we have used, the database and the choice of crime parameters. Section 3 presents the proposed algorithm which is used as a benchmark to achieve the objective on applying one of the well known factor analysis, k-mean clustering model and Geographical Information System (GIS) and in Section 4 presents the empirical results. The conclusions of our study are presented in Section 5.

2.0 Methodologies and Database

This section brings out the discussion of the database; the crime parameters recorded various districts and cities of Tamilnadu. The district wise crime data were collected from secondary source of city crime data in police department during the period of 2009 to 2011 was considered as the database. The data mainly consists of five major categories, such as property, violent, crime against women, traffic violation and others crimes related parameters.

2.1 Selection of Variables

In this study, twenty two crime parameters were chosen among the many that had been used in crime records. These twenty parameters were chosen to assess property, violent, crime against women, traffic violation and others crimes. The crime parameters are given below.

Table 1: Crime Parameters during the study period

S. No.	Parameters
1.	MFG
2.	Dacoity
3.	Robbery
4.	Burglary
5.	Theft
6.	Murder
7.	Association of Town and city Management
8.	C.H.Notatm Crime
9.	Hurt
10.	Riots
11.	Rape
12.	Dowry death
13.	Molestation
14.	Sexual Harassment
15.	Cruelty by Husband
16.	Kidnapping Women
17.	Kidnapping Others
18.	Criminal Breach of Trust
19.	Arson
20.	Cheating
21.	Counterfeiting
22.	Other IPC crimes

3.0 Data Mining Techniques

Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously unknown and potentially useful information from the data in databases. In the present context data mining exhibits the patterns by applying few techniques namely, factor analysis **k**-means clustering and Geographical Information System (GIS).

As such KDD is an iterative process, which mainly consist of the following steps on the data collected;

Step 1: Data cleaning
Step 2**:** Data Integration
Step 3: Data selection and transformation
Step 4: Data Mining
Step 5: Knowledge representation

In general, a knowledge discovery process mainly consists of an iterative sequence of the following: Of these above iterative process Steps 4 and 5 are most important. If suitable techniques are applied in Step 5, it provides potentially useful information that explains the hidden structure. This structure discovers knowledge that is represented visually to the user, which is the final phase of data mining.

3.1 Factor Analysis

Factor analysis provides the tools for analyzing the structure of the interrelationships (correlations) among the large number of variables by defining sets of variables known as factors. In the present study, factor analysis is initiated to uncover the patterns underlying crime parameters. Orthogonal rotations such as Varimax and Quartimax rotations are used to measure the similarity of a variable with a factor by its factor loading.

3.2 k-Means Clustering Methods

McQueen (1967) suggests the term k-means for describing an algorithm of his that assigns item to the cluster having the nearest centroid (mean). Generally this technique uses Euclidean distances measures computed by variables. Since the group labels are unknown for the data set, k-means clustering is one such technique in applied statistics that discovers acceptable meaningful classes.

3.3 Geographical Information System (GIS)

A Geographic Information System (**GIS**) is a system designed to capture, store, manipulate, analyze, manage, and present all types of geographical data. The acronym *GIS* is sometimes used for Geographic Information System Science or Geospatial Information Studies to refer to the academic discipline or career of working with geographic information systems and is a large domain within the broader academic discipline of Geoinformatics (ESRI, 2011). In the simplest terms, GIS is the merging of cartography, statistical analysis, and computer science technology.

A GIS can be thought of as a system, it digitally makes and "manipulates" spatial areas that may be jurisdictional, purpose, or application-oriented. Generally, a GIS is custom-designed for an organization. Hence, a GIS developed for an application, jurisdiction, enterprise, or purpose may not be necessarily interoperable or compatible with a GIS that has been developed for some other application, jurisdiction, enterprise, or purpose. What goes beyond a GIS is a spatial data infrastructure, a concept that has no such restrictive boundaries.

In a general sense, the term describes any information system that integrates, stores, edits, analyzes, shares, and displays geographic information for informing decision making. GIS applications are tools that allow users to create interactive queries (user-created searches), analyze spatial information, edit data in maps, and present the results of all these operations (Maliene V, 2011). Geographic information science is the science underlying geographic concepts, applications, and systems.

The ability of a GIS to relate and synthesize data from a variety of sources enables analysts to examine various aspects of criminal activity, including the built environment, crime risk and opportunity measures, and offender search patterns. Several examples of the uses of a GIS for both strategic and tactical crime analyses have previously been cited. The utility of a GIS depends on (a) the accuracy of the data; (b) the data attributes associated with each incident; and (c) the database, mapping, and analytical capabilities of the GIS. Specifically, a GIS has two broad applications that can be used for tactical crime analysis: descriptive mapping and analytic mapping. In this paper we are using both type of analysis (Block, C. R, 1990).

3.4 Algorithms

A brief step-by-step algorithm to classify the crime parameter during the study period based on their overall crime is described below:

For the pruned data set the following algorithms is proposed to the crime parameters and visualize them on GIS map during each of the study period based on their overall crime parameters (*Table 1*).

Step 1: Factor analysis is initiated to find the structural pattern underlying the data set.

Step 2: **K** –means analysis is used to partition the data set into **k**-clusters using the factor scores obtained in **Step 1** as input.

Step 3: Construct a GIS Map using the final cluster centered (mean) values with appropriate hits of the crime district/cities in the data set that are assigned group labels in step 2.

4.0 Results and Discussion

Factor analysis is extended with the techniques of Varimax and Quartimax criterion for orthogonal rotation. Even though the results obtained by both the criterions were very similar, the varimax rotation provided relatively better clustering of crime data. Consequently, only the results of varimax rotation are reported here. We have decided to retain 76 and more percent of total variation in the data set, and thus accounted consistently four factors for crime parameters with eigen values little less than or equal to unity. *Table 2* shows variance accounted for each factors.

Table 2: Percentage of Variance explained by factors (Year-wise)

Factors	2009	2010	2011
1	42.02	37.65	36.18
2	15.83	23.64	24.43
3	13.20	10.85	09.80
4	12.89	06.19	06.49
Total	**83.94**	**78.33**	**76.60**

From the above table we observe that the total variances explained by the extracted factors are over 76 percent, which are relatively higher. After performing factor analysis, the next stage is to assign initial group labels to crime parameters. Step 1 of the algorithm is explored with factor score extracted by Step 2, by conventional **k**-means clustering analysis. Formations of clusters are explored by considering 2-clusters, 3-clusters, and 4-cluster and so on. Out of all the possible trials, 3-cluster exhibited meaningful interpretation than two, four and higher clusters. Having decided to consider only 3 clusters, it is possible to classify crime parameters as High Crime Activity (HCA), Intermediate Crime Activity (ICA) and Low Crime Activity (LCA) depending on whether the crimes belonged to Cluster 1, Cluster 2 or

Cluster 3 respectively. In all the years, Cluster 2 (**HCA**) is a group of crimes that have high values for the crime parameters, indicating that these district/city recorded high crime. The **ICA** with intermediate values for the crimes is grouped into Cluster 3 (**ICA**). This suggested that Cluster 3 is a group of crimes with ICA. Cluster 1 (**LCA**) are those crimes which perform low level well as compared to the Cluster 2 and Cluster 3. Most interesting to note that in all three years the top level crime was recorded in capital city of Tamilnadu and rest of the districts and cities switch over the level of crime in three years period. In spite of incorporating the results for crime records, only the summary statistics are reported in *Table 3*. The second, third and fourth columns in *Table 3* provide the groupings done by cluster analysis.

Table 3: Number of crime districts and cities with Cluster Centers

Factor Scores	**2009**		
	1	**2**	**3**
1	-0.211	5.153	0.466
2	0.065	1.204	-1.082
3	-0.210	-0.083	2.202
4	0.020	0.519	-0.380
Total	**31**	**01**	**03**

Factor Scores	**2010**		
	1	**2**	**3**
1	-0.129	4.959	-0.685
2	-0.148	0.768	4.128
3	-0.022	1.622	-0.886
4	-0.038	-0.344	1.626
Total	**33**	**01**	**01**
Factor Scores	**2009**		
	1	**2**	**3**
1	-0.150	5.275	-0.187
2	-0.312	0.255	2.276
3	0.020	-0.146	-0.113
4	-0.010	0.050	0.065
Total	**30**	**01**	**04**

1– Low Crime Activity (LCA), **2** – High Crime Activity (HCA) and **3** – Intermediate Crime Activity (ICA)

Figures 1 to 3 show the groupings of crime data into 3 clusters for the study period. We classify the district/cities in the form of Cluster (**HCA)**, the second as Cluster **SCA** and the third as Cluster **ICA** in terms of crime parameters.

The pruned data set is then subjected to the main algorithm as in *Section 3.4* to assign appropriate classes to the crime records. The classes of crime were identified in the hot spot of crime districts and cities. In addition, GIS is used efficiently in spatial data visualization due to its ability to represent the input data. Among the various visualization techniques the most widely used method for visualizing the cluster structure of the GIS is the distance measure technique. *Figures* 2 to 4 show the groupings of crime parameters into 3 clusters over the GIS map using the GIS visualization method. In the above Figures, each colour represents classes of crime records in various districts and cities in Tamilnadu.

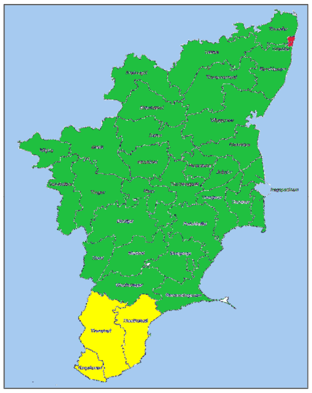

Figure 1: Crime mapping in the year 2009

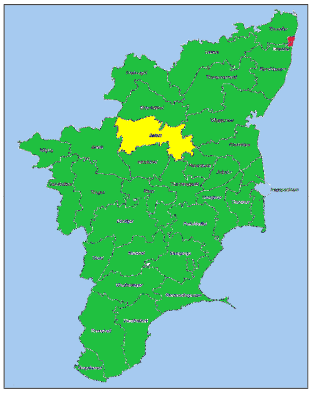

Figure 2: Crime mapping in the year 2010

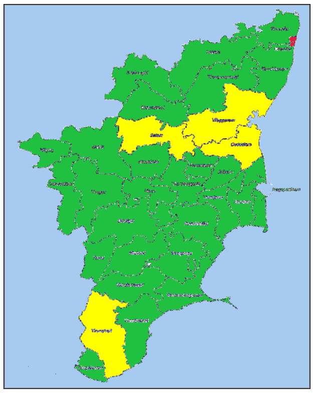

Figure 3: Crime mapping in the year 2011

5.0 Conclusions

The purpose of this paper has outlined several problematic aspects of optimization based cluster analysis for crime hot spot detection. Initially, factor analysis is used to identify the underlying structure based on twenty two crime parameters. The factor scores are used to partition the crime parameters into different clusters by using **k-**means clustering algorithm. The present analysis has shown that only 3 groups could be meaningfully formed for all the data. This indicates that only 3 types of crimes existed over a study period. Further, the crime parameters find themselves classified into *High Crime Activity* (HCA), *Intermediate Crime Activity* (ICA) and *Low Crime Activity* (LCA) categories depending on crime records. Rather than simply dismissing cluster analysis as being too complex for hot spot detection, additional research effort should be directed towards adopting existing statistical and GIS techniques make cluster detection more intuitive and useful for crime analysts. A generalization of the results is under investigation to obtain an incorporated class of 3 groups of crime parameters for any study period.

Reference

1. J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.
2. Block, C. R. (1990). Hot spots and isocrimes in law enforcement decision making. Paper presented at the conference on Police and Community Responses to Drugs: Frontline Strategis in the Drug War, University of Illinois at Chicago.
3. Geographic Information Systems as an Integrating Technology: Context, Concepts, and Definitions". ESRI. Retrieved 9 June 2011.
4. Maliene V, Grigonis V, Palevičius V, Griffiths S (2011). "Geographic information system: Old principles with new capabilities". Urban Design International 16 (1). pp. 1–6. doi:10.1057/udi.2010.25
5. Manish Gupta et. al. (2008), Crime Data Mining for Indian Police Information System, Proceeding of the Computer Society of India.
6. Murray, A.T., I. McGuffog, J.S. Western, and P. Mullins. 2001. “Exploratory spatial data analysis techniques for examining urban crime.” British Journal of Criminology. 41: 309-329.
7. Major Crime Trends – Tamil Nadu http://www.tnpolice.gov.in/crimeprofile.html
8. Major Crime Trends – Tamil Nadu http://www.tnpolice.gov.in/CAWChart.html
9. Richard A Johnson and Dean W Wichern (1992), Applied Multivariate Statistical Analysis, 3/ed, Prentice-Hall of India Private Limited, New Delhi.
10. Ratcliffe, J.H. and M.J. McCullagh. 1999. Hot beds of crime and the search for spatial accuracy. Journal of Geographical Systems. 1: 385-398.