ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 11, May 2014

Spatial Patterns of Crimes in India using Data Mining Techniques

Ahamed Shafeeq B M[1], Dr. Binu V S[2]
Department of Computer Science & engg[1]. Manipal Institute of Technology[1],
Department of Statistics[2]
Manipal University[1 2], Manipal, India.

Abstract*— Applications of spatial data analytical techniques has increased in the last few years in almost all fields. Some of the applications are decision making in regional governance, to maintain law and order, disaster prediction and many more. The main objective of this paper is to study the influence of neighboring states crime-rate with the reference state using spatial data mining techniques. In our study, we take the GDP, literacy-rate, police-rate, Employment-rate and various crimes such as murder, dacoit and riots and the state as location data. The aim of the paper is to check the correlation between various crimes. The whole work is divided into two parts: 1) to check spatial autocorrelation between various crimes 2) to compare various attribute clusters and its relation. The spatial distribution of various crimes in the states of India and also the correlation between the above said attributes and crimes in 2012 analyzed using exploratory spatial analysis methods. The outcome of the study reveals that the crimes of Indian states' has positive spatial correlation among the states and also found spatial disparity in crime distribution between local states. The states with higher employment-rate are more affected by the crimes. The clustering is used to identify the patterns with different crime densities, Employment and Police force distribution. Finally, the thematic maps of clusters are used to compare its correlation. The crime clusters can be used for planning various security measures in the states.

***Index Terms*— Spatial Data Mining, Autocorrelation, ESDA, Clustering.**

I. INTRODUCTION

Spatial data is related to the particular geographic features with descriptions of each feature and differs from normal data with its volume and structure. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets [1 2]. The spatial datasets are very large in volume and most likely interrelated. Unlike business queries that may only return a few records, spatial queries for generating maps are routinely run against many thousands of features.

Crimes are a social nuisance and cost our society dearly in several ways. Any research that can help in avoiding crimes faster will be of very important for the safety of fellow citizens. According to ShyamVaran Nath[3], about 10% of criminals commit more than 50% of crimes. The increase in crime rate will bring down all economic progress activities. The police department can prevent crime through the knowledge benefits that derived from information acquired. The police department can modernize their services through information technology to turn police officers into better service providers and can use their power to pre-empt crime [4]. The major challenge to law enforcement and intelligence agencies is the difficulty in analyzing large volumes of data involved in criminal and terrorist activities. The model can be extended to smaller regions for effective enforcement of the police patrol. The crime incidents are the biggest worry factors for administrative departments. The most formidable difficulty in analyzing crime trends across a vast country like India is the gap between the incidence and reporting of crime. It is a well-known fact that not all crimes, or classes of crime, are reported to the police for various reasons. The central role of police department is to protect the lives and property of citizens against crimes. But the police force is usually relatively very small compared to the crime prone population they have to protect making them more of a reactive rather than preventive force.

The hot spot identification is the effective method of mapping the crime in high-density areas [5]. Crime hot spot is an area where the number of criminal events or disorder is higher than in any other places, or an area where people have a higher risk of victimization. According to the Routine Activity theory[6], crime occurs in condition of three elements intersecting in time and space. The three elements include potential victims, motivated offenders and formal or informal guardians [7]. Crime is an act that an offence against the public and the perpetrator of that act are liable to legal punishment. It is closely associated with geographical and demographic variables [8].

The economy of India is the tenth-largest in the world by nominal GDP and the third-largest by purchasing power parity[9]. The main challenge for the administration is to keep steady growth of any country. The Employment is one of the key factors in the growth of a country. The main factors that influence crime rates are economic adversity, rising unemployment, declining wages and GDP.

The objectives of the present study is to investigate i) the spatial pattern of various forms of crime among the states' of India. ii)Association between the clusters of various crimes and clusters of GDP, Population density, Police-rate and Employment-rate of states of India.

II. METHODOLOGY

A. Exploratory Spatial Data Analysis (ESDA)

Exploratory spatial data analysis method (ESDA) of spatial data mining is a set of techniques aimed at describing and visualizing spatial distributions, identifying atypical

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 11, May 2014

localizations or spatial outliers, detecting patterns of spatial association, clusters or hot spots [10 11]. Spatial auto correlation is a powerful technique for the analysis of spatial pattern. The spatial autocorrelation is an important index to test the coincidence of value similarity with location similarity [12]. In this study, with ESDA we analyze the spatial dependency and spatial heterogeneity of various forms of crimes and selected few attributes of 32 states in India using spatial autocorrelation. The various crimes studied are riots, dacoit, murder and overall crime rates. The other attributes studied from each of the 32 states are GDP, employment rate, population density and police rate (number of police per lakh population). The crime rates in various states for the year 2012 are used for the study and are obtained from National crime records bureau (http://ncrb.nic.in). The GDP of each state for the year 2012 is obtained from planning commission (http://planningcommission.nic.in). We carry out the global statistics by computing spatial autocorrelation index [13]. In this paper, we use GeoDa software designed by Anselin [5] to get spatial weight matrix and we measure global spatial autocorrelation using Moran's I. The spatial clusters of natural break-ups are used to cluster the states with various forms of crimes and other attributes such as GDP, police rate, population density and employment rate.

1) Spatial Weights Matrix: Spatial weights matrix W is the precondition of exploratory spatial data analysis. The appropriate choice of spatial weight matrix is the most difficult and a controversial methodological issue in exploratory spatial data analysis [14].The spatial weight matrix W is as shown below.

$$\begin{bmatrix} W_{11} & W_{12} \cdots W_{1n} \\ W_{21} & W_{22} \cdots W_{2n} \\ \cdots & \cdots \quad \cdots \\ W_{n1} & W_{n2} \cdots W_{nn} \end{bmatrix}$$

Where n is the region (state) numbers. If regions *i* and *j* are neighbours then *Wij*=1 otherwise *Wij*=0. The diagonal elements of the above matrix *Wii*=0 for *i* = 1,2,…,n.

2) Global Spatial Autocorrelation :It is believed that there exists autocorrelation between observed objects, when the same attribute of different objects in space present some regularity and not randomly distributed [14]. The measurement of global spatial autocorrelation is usually based on Moran's I statistic, this statistic which is given by[8]:

$$I = \frac{n \sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{(\sum_i \sum_j w_{ij}) \sum_i (x_i - \bar{x})^2} \qquad (1)$$

Where n is the observed regions; w_{ij} is ij^{th} element in the spatial weights matrix; x_i is the value of region i; $\bar{x}$ is the mean of all observation values. Moran's I varies between -1 and 1. On a given significant level, a value near 1 indicates that similar attributes are clustered (HH-Higher value region surrounded by Higher value region or LL-Lower value region surrounded by lower value region), and a value near -1 indicates that dissimilar attributes are clustered (HL-Higher value region surrounded by Lower value region or Lower value region surrounded by Higher value region). If a Moran's I is close to 0, it indicates a random pattern or absence of spatial autocorrelation.

3). Local spatial autocorrelation

Moran Scatter Plot: In this paper, we used the GeoDa software to obtain the Moran scatter plot of per capita GDP of Indian states in 2012, overall-crime and Employment rate. We considered states of India as the basic analysis unit and the attributes such as GDP, Crimes and Employment rate to study the spatial distribution in states of India.

B. Clustering

Clustering is an important task in spatial data mining and spatial analysis. Clustering is a process of grouping of similar items. The clustering is used in the second phase of analysis followed by global and local spatial autocorrelation. The plan is to divide the number of regions for each attribute into four clusters. The regions of clusters with different range of values can be identified as high, moderate, middle and low level clusters for all the attributes considered. The cluster output of the different variables compared against each other for their influence. The GeoDa software is used for clustering. The inputs are shape file (.shp file) of states of India and the related data of the corresponding locations (.dbf file). The technique used in the clustering is natural break-ups. The natural break-ups find "natural groupings" by minimizing the variance within each class using Jenks optimization. In Jenks optimization method, each region is assigned arbitrarily to different clusters. The optimization is done by minimizing each cluster's average deviation from its cluster mean, while maximizing each cluster's deviation from the means of the other clusters [16].

III. RESULTS

Table-1 gives the Moran's I statistic of the crime in 2012 for the 32 regions (28 states and 7 union territories). The result is based on the permutation approach with 1000 permutations. From the above table, we can deduce that there is a positive correlation on overall crime-rate of states.

TABLE 1: Moran's I Statistics for GDP and Crime-Rate(Indian states)

variables	Moran's I	Standard deviation	Expected Moran's I E(I)
Overall Crime-Rate	0.1734	0.1024	-0.0323

Figure-1 shows the Moran scatter plot for the data set (32 states) of India for overall crime-rate. It shows that 68.5% of the states are characterized by positive spatial association and the remaining 31.5% of the states are characterized by negative association. The association of similar values (High-High quadrant 31.3% and Low-Low quadrant 37.4%) is exhibited by 68.7% of Indian states.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 11, May 2014

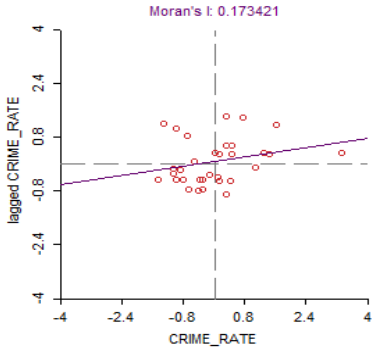

Fig 1: Moran's scatter plot (Overall crime-rate)

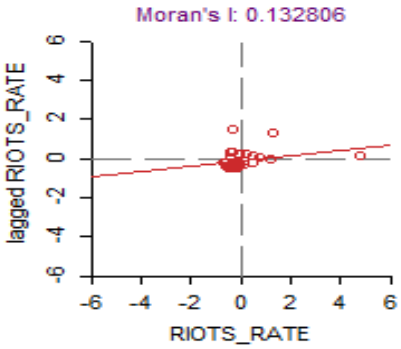

Fig 2: Moran's scatter plot (Riot-rate)

Figure-2 shows the Moran scatter plot for our data set(32 regions) of India for Riot-rate. It can be seen that the most Indian states characterized by positive spatial association. It shows that 81.25% of Indian states exhibited association of similar values (Low-Low quadrant 62.5% and in quadrant High-High quadrant 18.75%).

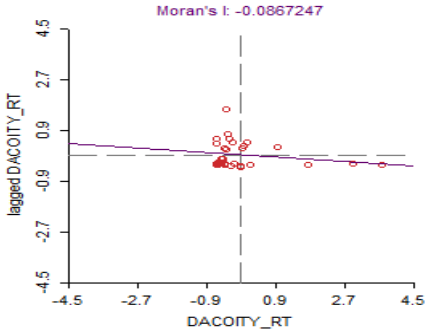

Fig 3: Moran's scatter plot (Dacoit-rate)

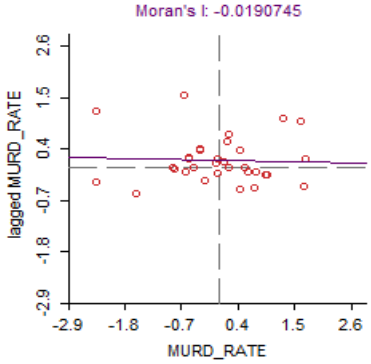

Fig 4: Moran's scatter plot(Murder-rate)

Figure-3 shows a small negative I value for the Dacoit-rate between the states of India. But it is negligible. Hence there is no correlation between the dacoit rate of the neighboring states. Similarly, murder rates also show a negligible negative autocorrelation among various states of India (Figure 4).

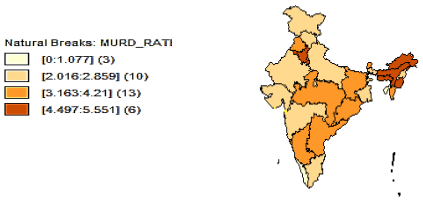

Fig 5: Murder-rate Cluster

Figure-5 shows the four clusters of murder-rate of 32 regions. The cluster shows a high murder- rate in Haryana and four north-eastern states.

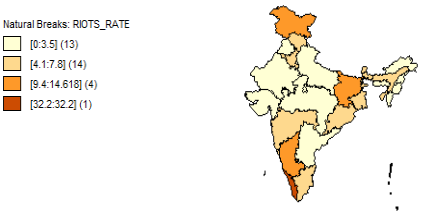

Fig 6: Riot-rate Cluster

Figure-6 shows the clusters of riot-rate in states of India. It shows highest riot-rate in Kerala state and second highest riots took place in Karnataka, West Bengal and Jammu &Kashmir

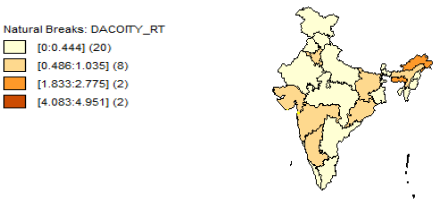

Fig 7: Dacoit-rate cluster

Figure-7 shows the highest dacoit-rate in Dadra and Nagar Haveli. It is found that the majority of the shows very low dacoit rate.

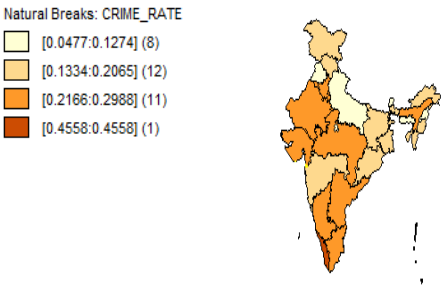

Fig 8: Overall Crime-rate Cluster

Figure-8 shows highest crime in Kerala. The cluster of second highest crime-rate includes eleven states.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 11, May 2014

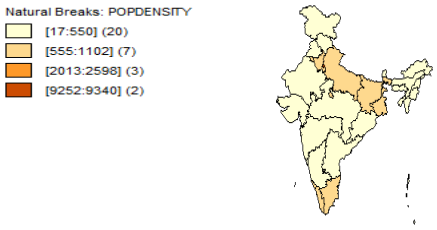

Fig 9: Population density Cluster

Chandigarh and Delhi has highest population per km2 area. The population is crowded in major cities.

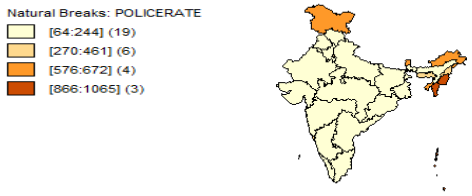

Fig 10: Police-rate Cluster

The maximum number of Police force is deployed in Andaman, Mizoram and Manipur. Majority of the states has less police /lakh population.

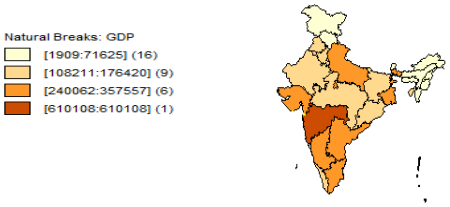

Fig 11: GDP Cluster

Figure-11 shows the clusters of regional economy distribution. The state with the highest GDP is Maharashtra. North-eastern states show poor regional economy.

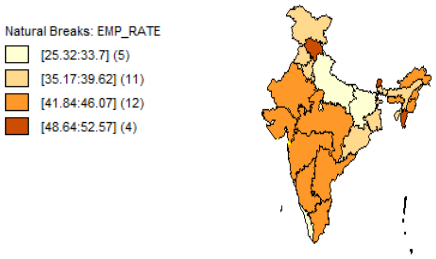

Fig 12: Employment-rate Cluster

Figure-12 shows the cluster with different employment rates. The states in the highest employment clusters are Himachal Pradesh, Mizoram, Sikkim and one union territory (Dadra & Nagar Haveli). Some of the states show correlation between Employment and GDP (Figure 11 & 12). It also shows that high murder-rate in under-developed states. The states with higher employment-rate is worst affected by the crimes (Figure 8 & 12). There is no much correlation of the various crimes with each other among the states' (Figure 5, 6 & 7).

IV. DISCUSSION

This study explores the spatial pattern of various forms of crime rates, GDP and employment rates of different states of India. The study focuses on autocorrelation of regional economy distribution, crime rates and employment rates in states of India and to construct clusters with various attributes for finding the relationship between the different clusters.

The crime clusters for different crimes in Uttar Pradesh have been studied and identified as safer zones [9]. This study is limited to cluster identification and to mark safer zones. Another study from China focused on regional economy distribution in different Chinese province [14]. This study is focused only on regional economy distribution. The investigation reveals that there is a global correlation among the regions in economy distribution and heterogeneity between local regions.

In our study, we used the GeoDa software designed by Anselin to obtain the Moran scatter plot of per capita GDP, overall crime rate and Employment rate of Indian states in 2012. Moran's scatter plot shows a positive autocorrelation for majority of the states' for GDP and Crime. There is a negative correlation for the employment rate of the states'. But few states show negative correlation for GDP and Crime. The murder rate & dacoit rate is found to be high (from murder-cluster & dacoit-cluster) in the north-eastern states compared to other parts of India. We do not find any relation between murder rate and riot rate in cluster comparison. There are no similar patterns for employment and GDP clusters. There is a correlation between overall-crime and GDP. The state Maharashtra is having highest GDP and comparatively less crime. The state Kerala has comparatively low employment and highest crime rate.

Our study is limited to check the correlation between the states for Crime, Police force, GDP and Employment. It will not find the reason (causal relation) for each outcome. The clusters are created for each attribute to compare with each other.

V. CONCLUSION

The Moran's scatter plot results show that there is a positive autocorrelation between states' overall crime-rate. There is no correlation of crimes such as dacoit-rate and murder-rate among the states. The spatial distribution cluster shows there is no spatial correlation between the various crimes in the states. By using the spatial distribution of regional(state-wise) per capita we can identify i) variations in income ii) potential market iii) state specific policies iv) industries that are driving economic growth and v) strategies based on the latest states' economy. Moran I index and Moran scatter plot is demonstrated to be a helpful tool in our study to reveal the characteristics of GDP and overall crime-rate of states of India. Exploratory spatial data analysis also tells us that there is a positive global spatial autocorrelation to distribution of economy of states of India. Moran scatter-plot is a very useful tool to identify positive and negative autocorrelation. We can also find heterogeneous regions in

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 11, May 2014

the local analysis and shows homogeneous relationship in the global level analysis.

REFERENCES

[1] Han, J., Kamber, M., Tung, A.K.H, "Spatial Clustering Methods in Data Ming: A Survey", Geographic Data Mining and Knowledge Discovery, pp. 1–29. Taylor & Francis, Abington, 2001.

[2] Haggett P, A.D.Cliff, A.Frey, "Location Analysis in Human Geography2: Locational Methods", John Wiley, New York, 1977.

[3] Shyam Varan Nath, "Crime Pattern Detection Using Data Mining", International IEEE/WIC/ACM Conference on Web Intelligence and Intelligent Agent Technology Workshops, 2006.

[4] Brown Mary Maureen & Brudney Jeffrey L., "Learning Organizations in the Public Sector? A Study of Police Agencies Employing Information and Technology to Advance Knowledge". Public Administration Review 63 (1), 30-43, 2003.

[5] Anselin L., Longley P.A, Good child M.F, Maguire D J, et al., "Interactive techniques and exploratory spatial data analysis, Geographical Information Systems, Principles, Technical Issues, Management Issues and Applications", John Wiley & Sons, Inc, pp. 253-266, 1999.

[6] Felson, Marcus and Rachel Boba. "Crime and Everyday Life. California", SAGE Publications, 2010.

[7] Kumar M. V.and Chandrasekhar C. "GIS Technologies in crime analysis and crime mapping", International Journal of Soft Computing and Engineering, vol. 1, November, 2011.

[8] Peng Chen, Tao Chen, Hongyong Yuan "GIS Based Crime Risk Analysis and Management in Cities", IEEE 2nd International Conference on Information Science and Engineering (ICISE), 2010.

[9] International Monetary Fund, World Economic Outlook Database, April 2013.

[10] Jitendra Kumar, Sripati Mishra, Neeraj Tiwari, "Identification of Hotspots and Safe Zones of Crime in Uttar Pradesh, India: Geo-spatial Analysis Approach", International Journal of Remote Sensing Applications, Vol.2 No.1 PP.15-19, 2012.

[11] Cliff A D, Ord J K, "Spatial Autocorrelation", Pion, London, 1973.

[12] Cliff A D, Ord J K. Spatial Processes: Models and Applications, Pion, London, 1981.

[13] Wang Yuanfei, He Honglin, "Spatial Data Analysis Method", Science Press, Beijing, 2007.

[14] Anselin L, "Spatial econometrics: methods and models", Kluwer Academic Publishers, Dordecht, 1998.

[15] Jian Lian, Xiaojuan Li, Huili Gong, Yonghua Sun and Lingling Li "Spatial Data Mining and Analysis of the Distribution of Regional Economy", IEEE International Workshop on Education Technology and Training & International Workshop on Geoscience and Remote Sensing, 2008.

[16] Robert McMaster and Susanna McMaster, "A History of Twentieth-Century American Academic Cartography", Cartography and Geographic Information Science, Vol. 29, No. 3, pp. 305-321, 2002.

AUTHOR BIOGRAPHY

Ahamed Shafeeq B M. has received the M.Tech. in computer science & engineering from Vishweshwaraya Technological University, India in 2006. He is presently working as an Assistant Professor in the department of Ccomputer Science and Engineering in Manipal Institute of Technology since 2009 and part time research scholar in Manipal University. His research interests are data mining, Algorithms and parallel processing. He has got sponsorship from Infosys Technologies in 2012 to present a research paper in an International conference held in Singapore. He is a life time member of the Indian Society for Technical Education (ISTE).

Binu V S received the PhD from the Manipal University, India in 2012. His thesis focused on estimation of uncertainty associated with intensity ratio in microarray data analysis. Currently, he is working as Associate Professor in Department of Statistics, Manipal University. His research focuses on multivariate methods, spatial epidemiology and error estimation in microarray data analysis.