Spatial Clustering with Web Processing Services (WPS)

HEMANTH KONGA
March, 2013

ITC SUPERVISOR
Dr. Ir. R.L.G. Lemmens

IIRS SUPERVISOR
Mr. Kapil Oberai

Spatial Clustering with Web Processing Services (WPS)

HEMANTH KONGA
Enschede, the Netherlands [March, 2013]

Thesis submitted to the Faculty of Geo-information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

THESIS ASSESSMENT BOARD:

Chairperson	: Prof. dr. ir. M.G. Vosselman
ITC Professor	: Dr. M. J. Kraak, ITC
External Examiner	: Dr. Surya S. Durbha, CSRE, IIT Mumbai
ITC Supervisor	: Dr. Ir. R.L.G. Lemmens
IIRS Supervisor	: Mr. Kapil Oberai

Indian Institute of Remote Sensing
Indian Space Research Organisation
Department of Space, Government of India,
ISO 9001-2008 Certified

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION,
UNIVERSITY OF TWENTE,
ENSCHEDE, THE NETHERLANDS

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, The Netherlands. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

DEDICATED *TO MY PARENTS…*

ABSTRACT

This research would facilitate the application of Geospatial Clustering Algorithms on spatial data as Web Processing Service. The crime-analysis section studies daily reports of serious crimes in order to determine the different facts of the crime like location, time, special characteristics, similarities to other criminal attacks, and various significant facts that might help to identify either a criminal or the existence of a pattern of criminal activity.

Data mining deals with the discovery of unexpected patterns and new rules that are "hidden" in large databases. Data mining is the process of extracting knowledge from these large databases. It serves as an automated tool that uses multiple advanced computational techniques, including artificial intelligence (the use of computers to perform logical functions), to fully explore and characterize large data sets involving one or more data sources, identifying significant, recognizable patterns, trends, and relationships not easily detected through traditional analytical techniques alone. Data mining can be defined as the identification of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data.

Spatial data mining is the process of extracting interesting knowledge from spatial databases. The spatial databases contain objects that represent space. The spatial data represents topological and distance information. This spatial object is organized by spatial indexing structures. Spatial data mining, or knowledge discovery in spatial database, refers to the extraction of implicit knowledge, spatial relocations, or other patterns not explicitly stored in spatial databases.

Clustering is an essential task in data mining to group data into meaningful subsets to retrieve information from a given dataset of Spatial Data Base Management System (SDBMS). The information thus retrieved from the SDBMS helps to detect urban activity centers for consumer applications. Clustering algorithms group the data objects into clusters wherein the objects within a cluster are more similar to each other and are more dissimilar to objects in other clusters. Spatial clustering is a major component of spatial data mining and is implemented as such to retrieve a pattern from the data objects distribution in a given data set. The clusters thus obtained would have to be interpreted to determine each one's significance in the context for which the clustering is carried out. Spatial clustering by itself is quite significant in that it is being implemented in a wide range of applications. Clustering process is a significant step towards the decision making process in such application areas as public safety measures, consumer related applications, ecological problems, public health measures and effective transportation facilities.

Keywords: *Web Processing Services (WPS), Crime Analysis, Geo Spatial Clustering, Clustering Algorithms, K-Means, DB Scan.*

ACKNOWLEDGEMENTS

.

I would like to express my gratitude to my ITC supervisor Dr. Ir .R.L.G. Lemmens for his support and constant encouragement. Without his patient guidance, this work could never have been a success.

I would like to thank my IIRS supervisor Mr. Kapil Oberai for his constant help throughout the project. Without his invaluable guidance, this work would never have been fruitful.

I am extremely thankful to Dr. Y.V.N. Krishna Murthy (Director IIRS), Mr.P.L.N.Raju (Group Head RSGG), and Dr.S.K.Srivastava (Head GID) for providing me infrastructural facilities to work in, without which this work would not have successfully completed.

Last but not least, I am greatly indebted to my family and friends, who have been a source of encouragement and inspiration throughout the duration of the project.

HEMANTH KONGA

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Crime Analysis

Crime analysis refers to the set of organized, analytical processes that provide timely, pertinent information about crime patterns and crime trend correlations. The main purpose of crime analysis is to get significant information from huge amount of data and circulate the final information to the officers in the field of investigation to help out in their efforts to capture criminals and curb criminal activity. (Santos, 2012)

Crime analysis deals with the study of crime and other police-related issues in order to help the police department in criminal investigation, crime and disorder reduction, crime prevention, and evaluation. Various elements of crime analysis can be demonstrated through this process. Usually when a study is taken up on a subject it means to inquire into the issue, investigate, examine closely, and/or scrutinize information. Crime analysis, then, is the focused and systematic examination of crime and disorder problems as well as other police related issues. Crime analysis involves the application of social science data collection procedures, analytic methods, and statistical techniques. (Ahmadi, 2003)

Crime is a one of the burning issues where the top priority is given by our government. Criminology is an area that focuses the scientific study of crime and criminal behaviour and law enforcement and is a process that aims to identify crime characteristics. The crime-analysis section studies daily reports of serious crimes in order to determine the location, time, special characteristics, similarities to other criminal attacks, and various significant facts that might help to identify either a criminal or the existence of a pattern of criminal activity. Such information is helpful in planning the operations of a division or district. (Wilson, 2002)

Crime analysis involves in finding out the truth of a given situation using different methods and the correct information to prove the truth so as to prepare an effective plan (Santos, 2012). The spatial nature of crime and other police-related issues is central to an understanding of the nature of a problem. In recent years, there has been a vast development in the field of Information technology. The improvements and development in computer technology and the availability of electronic data have created a larger role for spatial analysis in crime analysis. To understand the nature of crime and disorder visual displays of crime locations and their relationship to other events and geographic features are essential. Recent developments in criminological theory have encouraged crime analysts to focus on geographic patterns of crime, examining situations in which victims and offenders come together in time and space.

1.2. Crime Analysis using GIS

Chances of Crime are related to various factors like time, space, land etc. Geographic Information System (GIS) has emerged as a powerful analysis tool to support the decision-

making process involved in crime prevention. A Geographic Information System (GIS) is a special kind of database management system that supports analysis, organization and visualization of demographic and spatial (geographic) data as well as creation of new datasets. Initially GIS was mostly used for cartography and mapping. GIS takes raw data as the input and produces useful output information. GIS can also be referred to as a system of inter-related functions that can achieve several goals, such as data entry and storage, data-analysis and display. GIS tools allow integration of crime information systems with spatial data and assist in the production of accurate and high quality maps which makes GIS unique when compared to other database management systems.(Rogerson & Sun, 2001)

With the progress of information technology, several highly developed techniques are being invented to face and overcome the growing volume of crime. A computer based information system used for managing geographical data and also using these data to work out spatial problems is known as Geospatial Information(P. Canter, 2000)

1.3. Spatial Data Mining

Data Mining (also called Knowledge Discovery) is the process of analysing data from different perspectives and summarizing it into useful information or relationships. It is also characterized as the process of finding correlations or patterns among dozens of fields in large relational databases. Data Mining is the process of drawing out the information from huge databases while bringing collectively the techniques from machine learning, pattern recognition, figures, databases and visualization to address the issue of information extraction from large data bases. (Larose, 2005)

Data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis. This new field paves way to explore many future directions. Data mining can be defined as the identification of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data. (Adderley, Townsley, & Bond, 2007)

Spatial data mining is the process of discovering interesting, useful, non-trivial patterns from large spatial datasets. Spatial data mining (SDM) consists of extracting knowledge, spatial relationships and any other properties which are not explicitly stored in the database. SDM is used to find implicit regularities, relations between spatial data and/or non-spatial data. *"Data mining methods are not suited to spatial data because they do not support location data nor the implicit relationships between objects. Hence, it is necessary to develop new methods including spatial relationships and spatial data handling".* (Phillips & Lee, 2010)

1.4. Clustering Algorithms

Clustering is one of the important streams in data mining useful for discovering groups and identifying interesting distributions in the underlying data. Clustering is useful in several

investigative pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification. Financial data classification, spatial data processing, satellite photo analysis, and medical figure auto-detection etc. are the prime areas where clustering technique is widely adopted.(Hammouda & Karray, 2000) (Chandra & Anuradha, 2011)

1.5. Web Processing Services

A Web Processing Service (WPS) provides access to calculations or models which operate on spatially referenced data. A WPS can be configured to offer any sort of Geographic Information System (GIS) functionality to clients across a network, including access to pre-programmed calculations and/or computation models. The WPS standard provides a mechanism to identify the spatially-referenced data required by the calculation, to initiate the calculation, and to manage the output from the calculation so that it can be accessed by the client. WPS is a generic interface because it does not identify any specific processes that are supported. Instead, each implementation of WPS defines the processes that it supports, as well as their associated inputs and outputs. WPS is designed to standardize the way that GIS algorithms are made available through the Internet. It specifies a mean for a client to request the execution of a spatial calculation from a service (Schut, 2007).

1.6. Motivation and Problem Statement

Increase in Crime is one of the important concerns now-a-days. Most of the nations are facing the problems of increase in crime rate .This rate has increased twice or thrice since last 30 years (Ackerman & Murray, 2004). Countries worldwide are dealing with prevention of crime in order to fortify the public security. Efforts are being made by communities and government bodies to combat with crime. "*Crime analysis is a set of processes applied on relevant information about crime patterns. Administrative and operational personal can use the result of analysis to prevent and suppress of criminal activities and also for investigation aims*" (Ahmadi, 2003). Intellectual crime analysis enables the understanding of vibrant criminal activities, and also provides information about the likeliness of criminal act in particular place and time(Phillips & Lee, 2010). The purpose of crime analysis is to gain information from large data, and deliver the evidences to police and investigators in order to reduce crime. Recently, the innovative techniques such as data visualization, geographic information sciences etc. are being used to analyze the hot-spots of crime in urban areas (Tony H. Grubesic & Mack, 2008).

Geographic information science has reached to an extent where there is a huge amount of data but the information associated with the data is unorganized. From the large datasets, geospatial clustering forms a group of similar objects (observations, events) based on spatial and non-spatial attributes of the objects (Wang et al., 2011). It plays an important role in extracting useful patterns from the geo-referenced data. It has an application in quantifying various geographic patterns which is commonly used in disease surveillance, spatial epidemiology, population

genetics, landscape ecology, crime analysis and many more (Jacquez, 2008)(Wu & Grubesic, 2010). It integrates variety of large datasets with minimal efforts (T. H. Grubesic & Murray, 2001). With the help of spatial analysis, one can identify and examine the crimes patterns in geographic data. The spatial analysis explains the theory of co-relation of crime, with respect to poverty and its related conditions(Ackerman & Murray, 2004).

Now-a-days, the desktop GIS are being replaced by web based GIS in order to share the geographical data over web. It would be accessible to more people, instead of restricting the usage to a single user. "*A Web service is a software system designed to support interoperable machine-to-machine interaction over a network"* ("Web Services Architecture," 2012). These web services are adopted by geospatial web services to manage and process the geospatial data on the web. The interoperability technology of web services will help to share the geospatial data among various applications. It indorses the services to many users (Fenoy, Bozon, & Raghavan, 2012). Web Processing Service provides a standardized interface between the client and server for publishing the geospatial processes over the web. A WPS provides large number of GIS functionalities on spatially referenced data such as algorithms, calculations, models, etc., over the network (Schut, 2007). By incorporating the calculations, algorithms in WPS service, the users can simply request and process the geographical data over the web. *"WPS supports simultaneous processes via the HTTP GET and POST method, as well as the Simple Object Access Protocol (SOAP) and Web Services Description Language (WSDL)"* (Fenoy et al., 2012).

The potential users of this research work are the police department. They generally use proprietary software for crime analysis, which is a desktop based approach and restricted to a single system. The proprietary software is very costly and giving a license to every police station is very cost effective process and it requires the GIS knowledge to process the crime data. This data is sensitive, law bound and has constraints on its distribution over the web. The clustering service for crime analysis with WPS used by the police can be shared among the police stations. This could assist the police to identify the crime hotspot areas and take precautions in lowering down the crimes in such areas by patrolling etc. Data security could be ensured by restricting the access of WPS service to specific users who need authentication from the police. For this crime analysis, the police need not have any desktop GIS at the client side. By using web browser the police could be able to do crime analysis.

1.7. Research Identification

1.7.1. Research Objectives

The focus of the research objectives is as under:

- To Compare and evaluate various geospatial clustering algorithms to analyze crime data of Delhi.

- To evaluate possible criteria for comparing and evaluating geospatial clustering algorithms for crime data.
- To check the feasibility of Web Processing Service for crime analysis.

1.7.2. Research Questions

1. What are the algorithms available for geospatial clustering and what are the software and tools available to implement these algorithms?
2. How to evaluate the geospatial clustering algorithm for crime analysis?
3. Is the socio-economical be used for this?
4. How can the selected geospatial algorithm be served through WPS?
5. What are the input and output data formats/data structures suitable for WPS?
6. What is role of WPS in addressing interoperability?
7. Is Web Processing Service suitable for Crime Analysis?

1.8. Innovation aimed at

This research would facilitate the application of Geospatial Clustering Algorithms on spatial data which can provide particular patterns that will be used for the analysis and decision making. Implementation of these clustering algorithms as web processing service can provide online results by using internet that can be used for various applications. By using WPS service user can acquire clustered information about crime occurring in an area of interest. It is not necessary for a user to have any GIS software and GIS domain knowledge for data analysis. The clustering process would be conceded if the user has a precise crime data to input as Keyhole Markup Language (KML) file through an interface. The output generated by the web processing service will also be in form of KML file and can be directly used on Web based mapping services like Google Earth for analysis. This service would be very helpful especially for police to visualize the clusters of crime in several areas.

2. CRIME ANALYSIS AND GIS

2.1. Crime analysis in context of GIS

Crime is defined as an offence against a person, or his/her property (like theft and property damage) or the State regulation Crime, in the other sense breach or violation of public law, is defined in both legal and non-legal sense. Crime occurs in a variety of forms which police informally categorizes as being either major or minor. Major crime consists of the high profile crimes such as murder, armed robbery and non-date rape which may be referred as high profile crimes. The crime-analysis section studies daily reports of serious crimes in order to determine the different facts of the crime like location, time, special characteristics, similarities to other criminal attacks, and various significant facts that might help to identify either a criminal or the existence of a pattern of criminal activity.

Reviewing the past, the customary strategy of intelligence and criminal record maintenance has failed to meet the requirements of the present crime scenario. Manual processes neither provide accurate, reliable and comprehensive data round the clock nor does it help in trend prediction and decision support. The lack of public capability to guide and support the transition from "crime mapping in the police" to "mapping with the community" for local policies are increasing at alarming rate. The effective utilization of Information Technology helps to change the paradigm of public safety management and useful for community policing processes under this perspective. (Rogerson & Sun, 2001)

A GIS uses a digital map database to link spatial data to descriptive information. Also, GIS uses geographically referenced data as well as non-spatial data and includes operations which support spatial analysis. Due to its spatially distribution almost all human activity and natural phenomena can be studied using a GIS. GIS integrates various types of spatial data (databases, imagery, GPS coordinates, etc.). GIS plays an important role as it allows viewing and analysis of data based on geographical proximity and relationships. Looking at data geographically can often suggest new insights, scope and explanations. (Tony H. Grubesic & Mack, 2008)

GIS assisted crime mapping is often employed to understand the geographical distribution of crime, identify crime concentrated area, or hot spots, and facilitate deployment decisions regarding the duration and dosage of intervention programs. The crime data depicted on a spatial domain can be overlaid with education, sex and occupational data to obtain a correlation between each type of crime committed and the social conditions of the place. This helps us to identify the hotspots of different types of crime on a GIS platform. GISs can be used as a means to identify various factors contributing to crime, and thus allow police to proactively respond to the situations before they become tricky or more problematic. GIS now be viewed as a tool for which police analysts could obtain a better understanding of criminal activity from a geographic point of view. 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. The effectiveness of a GIS mainly depends on accuracy of the data, the features of the data connected with each incident and finally the database, mapping, and analytical capabilities of the GIS (P. Canter, 2000).

GIS has many applications in various fields today. This is very commonly used in the geographically related fields like cartography, crime analysis and urban planning. GIS used in mapping crime allows analysts to identify hot spots, along with other trends and patterns. It also helps analysts to overlay other datasets such as locations of shops, banks and schools, etc., to better understand the original causes of crime and help the investigators to invent strategies to deal with the problem. In the field of crime analysis GIS is also useful in assigning police officers and dispatching to emergencies. Police departments primarily employ GIS in various applications like, crime analysis, crime prevention, public information, and community policing. GIS helps in detecting the areas of frequent criminal activities, identifying the suspicious incidents for investigation. GIS has also proved to be helpful in enhancing the implementation of various policing methodologies to reduce overall crime (Ackerman & Murray, 2004)

2.2. Spatial Data Mining

Data mining deals with the discovery of unexpected patterns and new rules that are "hidden" in large databases. Data mining is the process of extracting knowledge from these large databases. It serves as an automated tool that uses multiple advanced computational techniques, including artificial intelligence (the use of computers to perform logical functions), to fully explore and characterize large data sets involving one or more data sources, identifying significant, recognizable patterns, trends, and relationships not easily detected through traditional analytical techniques alone (Xiong, 2001). Data mining can be defined as the identification of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data. Data mining is used in the fast retrieval of information from databases. Data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis. Data mining helps in dealing with the criminal cases more efficiently and also many future directions can be explored in this new field (Chen et al., 1996)

Spatial data mining is the process of extracting interesting knowledge from spatial databases. The spatial databases contain objects that represent space. The spatial data represents topological and distance information. This spatial object is organized by spatial indexing structures. Spatial data mining, or knowledge discovery in spatial database, refers to the extraction of implicit knowledge, spatial relocations, or other patterns not explicitly stored in spatial databases. Spatial data mining methods can he apply to extract interesting and regular knowledge from large spatial databases. This knowledge can be used for understanding spatial and non-spatial data and their relationships. This knowledge is very useful in Geographic Information Systems (GIS), image processing, remote sensing and so on. Knowledge discovered from spatial data can be of various

forms, like characteristic and discriminant rules, extraction and description of prominent structures or clusters, spatial associations, and others.(Jacquez, 2008)

Knowledge discovered from spatial data can be of various forms, like characteristic and discriminant rules, extraction and description of prominent structures or clusters, spatial associations, and others. Spatial data mining methods can apply to extract interesting and regular knowledge from large spatial databases.

Spatial Data Mining (SDM) plays an important role in:

a) Extracting interesting spatial patterns and features;

b) Capturing intrinsic relationships between spatial and non- spatial data;

c) Presenting data regularity concisely and at conceptual levels;

d) Helping to reorganize spatial databases to accommodate data semantics, as well as to achieve better performance.

SDM is used to find implicit regularities, relations between spatial data and/or non-spatial data. In effect, a geographical database constitutes a spatial-temporal continuum in which properties concerning a particular place are generally linked and explained in terms of the properties of its neighbourhood (Santhosh Kumar, 2012). Spatial data is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc. Spatial data mining tasks include: spatial classification, spatial association rule mining, spatial clustering, characteristic rules, discriminant rules, trend detection. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. All the members of the cluster have similar features. Members belong to different clusters has dissimilar features (Chandra & Anuradha, 2011).

2.3. Clustering Algorithms

Spatial cluster analysis plays an important role in quantifying geographic variation patterns. Cluster analysis is one of the popular approaches for detection of crime hot spots. Hotspot mapping is used by many policing and crime reduction practitioners to identify spatial patterns of crime. By using data from the past, hotspot mapping identifies where crime most densely concentrates, starting a decision-making process that considers where best to target enforcement and prevention resources.(Chakravorty, 1995)

Classification on large spatial databases is a difficult and computational heavy task. Using clustering algorithms is one way of doing it. Spatial clustering algorithms make use of the spatial

relationships among data objects to differentiate the groupings which are natural within the input data. Based on the clustering technique used the spatial clustering methods can be classified into different categories (Han, 2006). Data clustering is one of the popular methods of retrieving information from databases. There are different types of clustering methods available which of them having different significance(Wang et al., 2011). Clustering analysis in the field of data mining using algorithms has become very prominent due to its variety of applications. Data clustering algorithms are widely used in the variety of applications like image processing, communication fields, earthquake studies etc. and made the use of algorithms very popular. Two clustering techniques, K-means and DB Scan algorithms are considered for this research.

The requirements of a good clustering algorithm are:

1. Ability to work with high dimensional patterns
2. Scalability with large data sets
3. Ability to find clusters of irregular shape
4. Ability to detect noisy outliers
5. Should work in one scan or less of the data

2.3.1. K-means Clustering Algorithm:

The *k*-means algorithm is probably the most commonly used partition clustering algorithm
The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters, k. This algorithm has been discovered by several researchers across different disciplines, most notably Lloyd (1957, 1982), Forgey (1965), Friedman and Rubin (1967), and McQueen (1967).

The K-Means is a well-known iterative distance-based clustering algorithm, and also is one of the oldest, simplest and most widely used clustering algorithms. The basic algorithm is very simple. K-Means starts by randomly picking up k cluster centres (k has to be specified in advance). Then, it assigns each instance to the nearest cluster centre. Once this is done, the new mean (centroid) for each cluster is calculated and instances are assigned to the closest cluster centre. The centroid is (typically) the mean of the points in the cluster. By allocating each data object in a cluster to its nearest mean centre, a new set of clusters are identified. This whole process is repeated until the cluster centres stop changing (Faber, 1994). A squared-error function that is utilized as the Objective Function in the k-means algorithm is stated as below,

$$E = \sum_{i=1}^{k} \sum_{x \in Ci} | x - m_i |^2$$

Where x is the point in space representing the given object, and *mi* is the mean of cluster *Ci*

A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.(Xiang et al., 2005)

Figure 2.1 : Clustering Process

The K-Means is a well-known iterative distance-based clustering algori(Santhosh Kumar, 2012)thm, it also is one of the oldest, simplest and most widely used clustering algorithms.

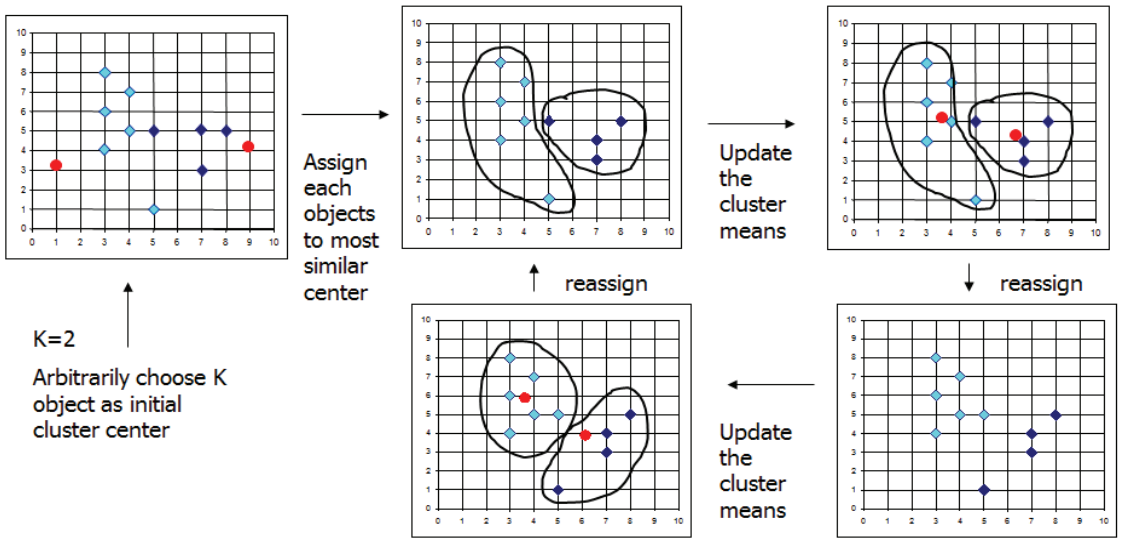


Figure 2.2 : K-means Clustering(Yang, 2010)

Researchers have shown that the k-means clustering in particular is rather easy to implement and apply even on large data sets. As such, it has been successfully used in various topics, ranging from market segmentation, computer vision, geostatistics and astronomy to agriculture Geosciences, etc. It often is used as a pre-processing step for other algorithms, for example to find a starting configuration. (Honarkhah & Caers, 2010). The continuous K-means algorithm being faster than the standard version can be used for datasets that are larger in size. (Faber, 1994)

The COPLINK national project developed by the University of Arizona Artificial Intelligence Lab was one of the prominent frame work for text mining, classification and clustering of crime data aiming to carry out difficult crime can be given as the best application of data mining in the field of crime analysis. The project deals with data pre-processing, and data gathering burdens handled by COPLINK CONNECT and extracting patterns out of large volumes of crime data. By using data mining and artificial intelligence handle by COPLINK DETECT, the two major components of the project(Chen et al., 2003).

2.3.2. The DBSCAN Algorithm

DBSCAN was the first density-based spatial clustering method proposed. The key idea is to define a new cluster, or extend an existing cluster, based on a neighbourhood. The neighbourhood around a point of a given radius (Eps) must contain at least a minimum number of points (MinPts). (Ester et al., 1996) The DBSCAN algorithm can identify clusters in large spatial data sets by looking at the local density of database elements, using only one input parameter. By this, the user gets a suggestion on which parameter value that would be suitable. Therefore, minimal knowledge of the domain is required. DB Scan can also determine what information should be classified as noise or outliers. The DB Scan algorithm can be used to find and classify the atoms in the data. Classification on large spatial databases is a difficult and computational heavy task which includes finding the right input parameters, localizing clusters of arbitrary shapes and, last but not least, doing the whole process in a reasonable time etc. The DBSCAN algorithm has a solution to all such problems as it can find those complicated cluster shapes in a very quickly, with given input parameter. A value for this parameter is also suggested to the user.

DBSCAN algorithm has the capability of Anomaly Detection in Temperature Data. A density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability.

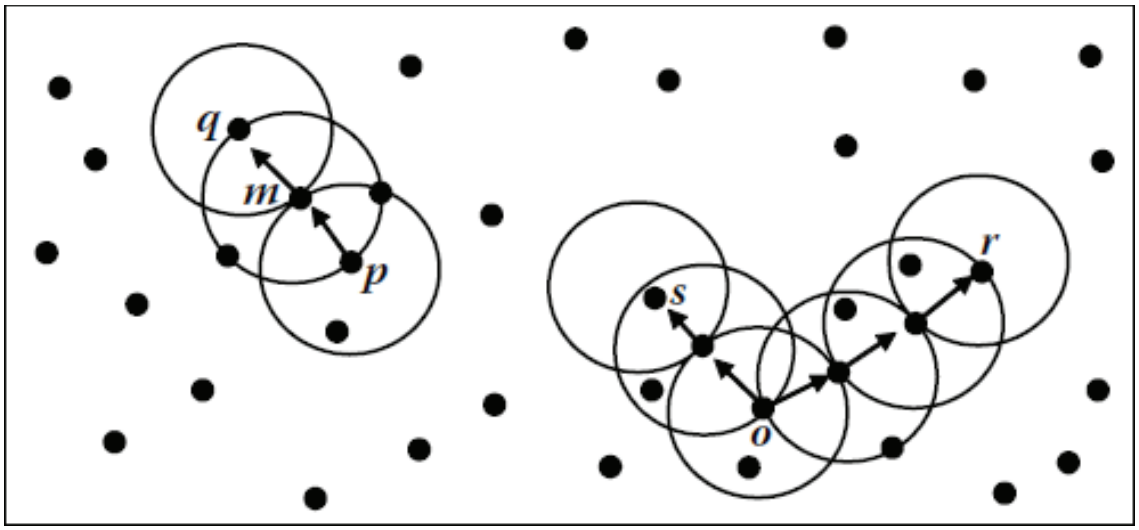


Figure 2.3 : DB Scan Clustering (Ester et al., 1996)

2.4. Web Processing Services

There has been an enormous advance in Web service technologies. The geographic information systems (GIS) field has been highly influenced by progress in Web services technologies. The field is now challenged with finding a way to incorporate multiple geographic services, each from a specific information community and region, in order to restrict the common practice of downloading and processing immense data sets using the conventional desktop GIS (Sun et al., 2011). A WPS provides client access across a network to pre-programmed calculations and/or computation models that operate on spatially referenced data. The calculation can be extremely simple or highly complex, with any number of data inputs and outputs. In the recent years, research on Web services has been growing interest aiming to provide solutions also for online generalization. (Bergenheim et al., 2009)

The progress in Web service technologies has influenced the geographic information systems (GIS) to a large extent field has been highly influenced by advances in Web services technologies. This has led to an effected in the increase of specialized geographic services that, for example, help in locating a map view, to visualize vector cartographic data etc. Traditionally the desktop GIS used to fetch even the simple analysis with complex manual data. (Granell et al., 2006)

The traditional desktop GIS is now challenged with a new type of GIS platform, which is easily accessible, easy-to-use, and easy-to-share known as Web-based GIS (also known as distributed GIS or Internet-based GIS). In today's world the web is becoming a combination of collaborative platforms. GIS along with web services has given out dynamic mapping services for environmental planning.(Sikder, 2008)

Web service is a software system designed to support Interoperable system-to-system interactions over a network (Booth, 2004). Service oriented architecture (SOA) and Service Oriented Architecture Protocol (SOAP) address these specifications and provide technical solutions for distributed systems. (Wehrmann, Gebhardt, Klingera, & Künzer, 2010)

A WPS is capable of handling more than a single process. There are three required operations performed by a WPS, namely GetCapabilities, DescribeProcess and Execute. The response of GetCapabilities, DescribeProcess both is in the form of a XML document. When an Execute request is send to the WPS, the final process implementation is carried out. The GetCapabilities returns service metadata, DescribeProcess returns a process description which includes resquests and responses and Execute returns the output or the results of the process. (Schaeffer, 2008)

WPS framework of the deegree2 project was used in the implementation of a bomb threat scenario. It is very much lengthy process to implement WPS from scratch. By using frameworks such as the deegree WPS the simple processes can be quickly implemented**.** (Stollberg & Zipf, 2007)

WSMT is an Integrated Development Environment (IDE) for Semantic Web Services, implemented in the Eclipse framework. This mainly helps the user in creating, editing and validating WSML descriptions and provides direct support for reasoning and discovery through specific views. The primary idea of this function is to allow the client to check that a specified Goal is equal to the estimated the set of the Web Services in the users' workspace. (Xu & Xinyan, 2009).

2.5. Literature Review and Summary

Crime, through its impact on victims and through its indirect effects on the wider community and their perception of crime, clearly has strong links on the society. There is endemic of criminal activities in the society today. The prevalence of crimes in the society depends on several factors. Geographic information science has reached to an extent where there is a huge amount of data, but the information associated with the data is unorganized. GIS is widely used application to analyse the hotspots of crime in different areas. A hotspot is a geographical area with higher than expected chances of incidences of certain events. Crime mapping using GIS is used in identification of hot spots of crime thus making the investigation easier and also time saving. The research employs geospatial clustering algorithms, K-means and DB Scan for crime analysis. The application and execution of these clustering algorithms as web processing service fetches online results which can be used in the process of crime analysis. The outcome of the application helps the police to overcome and reduce the chances of crime to a large extent.

3. STUDY AREA AND DATA USED

3.1. Study Area

The study area opted for the research is Capital city of India 'New Delhi'. *Delhi* has a total geographical area 1,483.01 sq. km. The city lies to the northern part of India with a centre latitude and longitude of 28.38° N and 77.13 E as its geographical locations. Delhi is one of the important metros of the country. It has a land frontier of about 15,200 km. The total population of Delhi as per the provisional figures is 16753235.

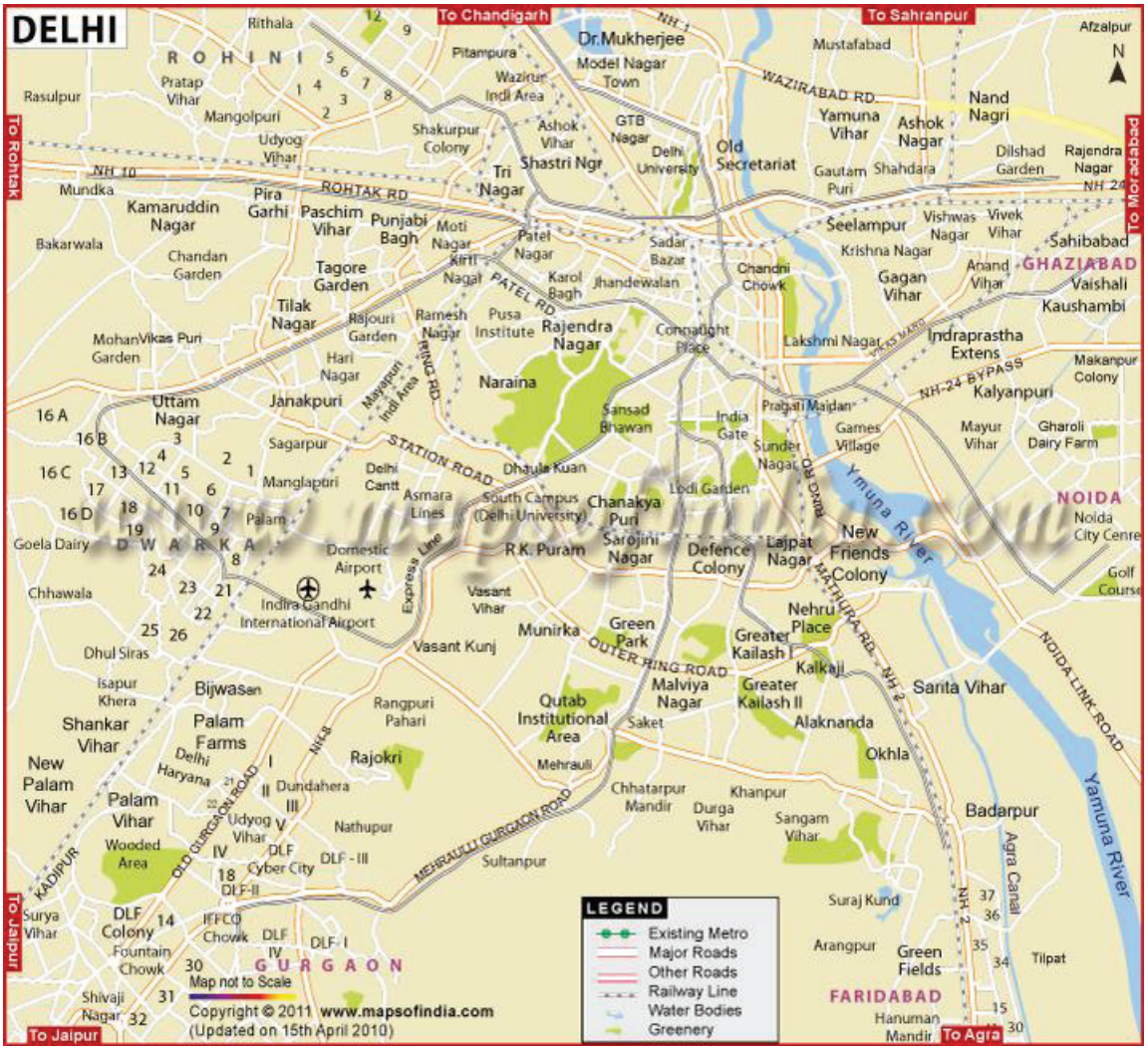


Figure 3.1 : Delhi Map

3.2. Scientific Significance of Study Area

Delhi has achieved a lot socio-economic progress during the last few decades. It is undoubtedly one of the most historic and aesthetic cities in the country drawing huge number of tourists every year and making the country grow socially and economically. It is not astonishing to say that the progress rate and crime rate are running parallel terming it as the most violent city. The national capital occupies the top slot for almost all violent crimes, including murder, rape, dowry death, molestation, kidnapping and abduction.

A report released by the National Crime Records Bureau stated that Delhi is one of the top four crime affected cities in India. Incidents like kidnapping for ransom, rape, molestation, abduction and attempted murders has seen sharp increase compared to the previous years. According to the report, about 44.6% of arrested criminals fall into the age group of 18 to 30 years. The situation has very much worsened in the recent years.

The parliament attack, series of bomb blasts etc are some of the important terrorist attacks. Criminals and terrorists flock to the capital city with an intension to destroy the harmony and peace of the city as well as the country.

The city being one of the most important metros of India should have been much secured for its citizens. The rate of crime in the city is increasing day by day and there is a need to put a break and slow down the crime rate. The growth in the crime rate is throwing a challenge to the police. The huge increase in the rate of crime in Delhi indirectly point outs the development of the country. The crime graph of Delhi clearly predicts that the city has increased rate of crime by 4-6% compared to the previous years prompting police to earmark special measures to combat street crime.

There has been an enormous increase in the crime in the recent years and Delhi is undoubtedly one of the major and top most victim metro cities of this crime world. The raise in the crime rate not only points out the national security but also the development of the country and the efficiency of the Government in controlling this.

The police people deal with the cases for months and years searching for the clues. For this, the data for the investigation must be accurate and assist the police in solving the crime cases efficiently. Crimes in Delhi during 2011 is as follows:- 543 Murders, 33 Dacoity , 562 Robbery, 1419 Burglary, 22899 Theft, 386 Attempt to Murder, 42 Arson, 568 Rape, 142 Dowry Death, 229 Eve teasing and 653 Molestation of Women.

3.3. Data Used

The data is vector point data where those points indicate location of crime with the relevant information (date, time, and type of crime) regarding that crime. LULC map and census information of the study area are also being used.

Crime data (raw data) of West Delhi has been used for the research. The date sets contain the records of crime occurred in West Delhi. The data has been obtained from the Delhi police department. The major highlight of the crime data lies on the auto theft in West Delhi. Motor vehicle thefts comprise of 27% of the total IPC crimes in Delhi.

The latest National Crime Records Bureau report says 'auto theft' in the country accounted for 44.4% (1,51,200 cases) of the total theft cases, which accounted for an increase of 2.5% in 2011 as compared to 2010 (1,47,475 cases). "The acute shortage of parking space and the general practice of parking vehicles on roadsides, coupled with the indifference of a majority of motor vehicle owners towards installing anti-auto theft equipment, is a major contributory factor to these thefts," said a senior police officer.("National Crime Records Bureau," 2012)

As per Times of India magazine, many of the stolen vehicles are used by criminals in the commission of other crimes, making this a problem area for the police. Delhi Police has been sending out advisories to the public through advertisements and leaflets requesting vehicle owners to buy security gadgets. Repeat offenders are also a big reason for the alarming increase in motor vehicle thefts. (The Times of India, 2012)

Table 3.1 : Crimes in Delhi("DelhiPolice," 2013)

Year	Miscellaneous I.P.C. Cases	Rape	Dowry Death	Eve Teasing	Molestation of Women	Snatch Theft	Motor Vehicle Theft
1980	12062	-	-	-	-	-	-
1985	13933	-	-	-	-	-	-
1990	16968	-	-	-	-	-	-
1995	26613	152	357	2730	502	-	-
1996	45972	132	484	2059	694	-	-
1997	3220	148	544	185	675	-	-
1998	30568	438	126	172	653	-	-
1999	27517	402	122	146	588	-	-
2000	28406	435	125	123	549	-	-
2001	29214	381	113	90	502	675	7894
2002	26010	403	135	976	446	605	7434
2003	26975	790	130	1599	489	532	7445
2004	29042	551	126	2132	601	775	8873
2005	30127	658	114	1714	762	1136	8862
2006	34036	623	137	556	718	1283	9366
2007	32065	598	138	414	868	1243	8874
2008	24097	466	129	328	611	1377	11020
2009	23881	469	141	238	552	NA	NA
2010	23776	507	143	126	601	NA	NA
2011	25877	568	142	229	653	NA	NA

YEAR 2011	Missing Persons			No. of persons traced		
	M	F	T	M	F	T
Below 18 years	2446	2665	5111	1879	1873	3752
Above 18 years	4587	4214	8801	3137	2774	5911

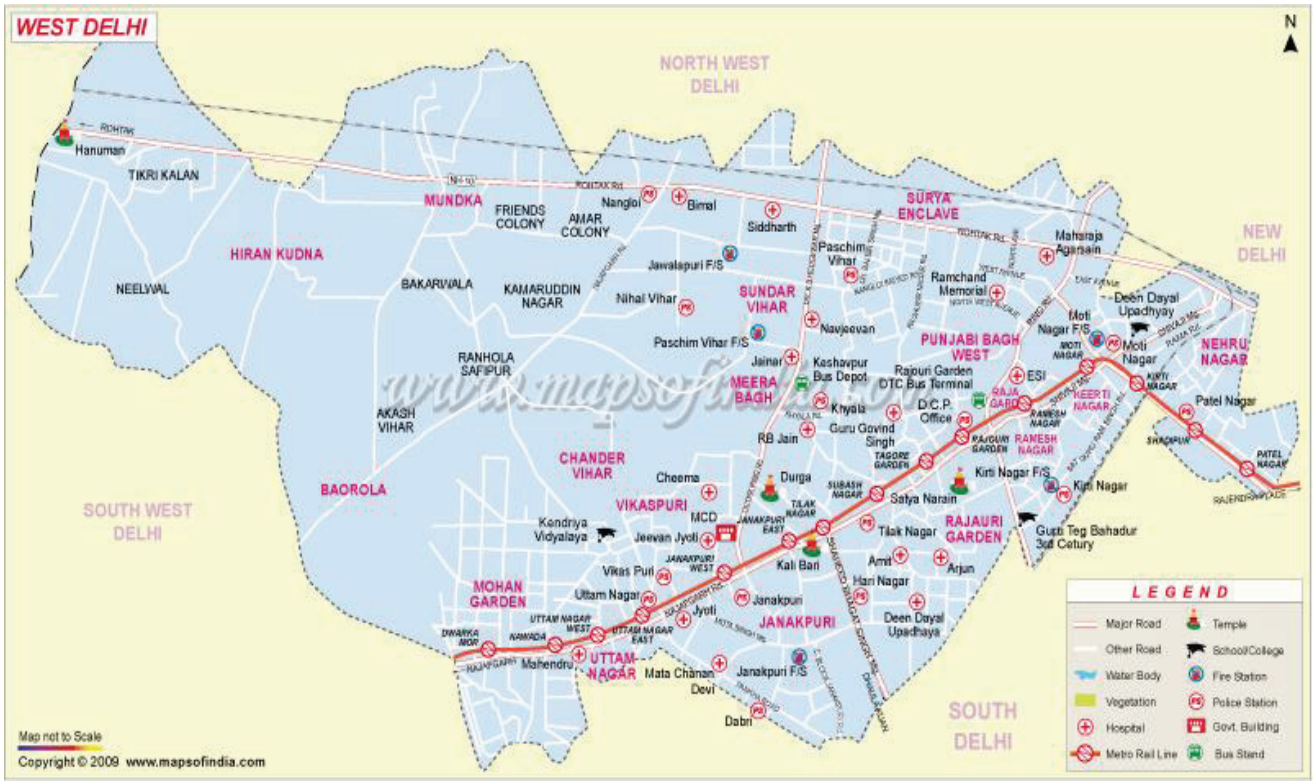


Figure 3.2 : West Delhi Map ("MapsofIndia," 2013)

4. PROTOTYPE IMPLEMENTATION

4.1. GeoCoding

Geocoding is a common technique used in crime mapping. This process employs two techniques. Firstly, addresses are searched for a database and secondly, algorithmic approaches to identify the closest locations with respect to the required address. The process of geocoding is generally used to achieve the approximate locations of street addresses. Address is an attribute data while locations of crimes on crime maps are mostly point data. Point data describes the location of crime most accurately, also making the identification of related addresses or other parameters easier to measure. (Ratcliffe, 2010)

In this study, address details of 17 taluks have been provided for Delhi West District. These addresses include a variety of addresses, ranging from apartments addresses, Schools, Industrial Sectors, Parks, etc. The process employed here is used mainly to identify the locations of the given addresses in terms of their spatial locations. A case that requires considerable stress here is the fact that not all the required addresses can be located accurately by proximity. Therefore, techniques of proximity or closeness of identifiable and valid locations are performed. For example, if the given address cannot be spatially located, the parks or schools nearby that have fixed locations are identified and locations of the required addresses are identified subsequently.

Various tools have been used in the process of establishing the addresses. Google Maps provides suggestions on the basis of the input query and the proximity of the address from other likely matching addresses. These suggestions were used to establish the locations of data in this study. OpenStreetMap and WikiMapia online tools were also used to generate the addresses of various locations in the study area.

4.2. Methodology

The present research work focuses on developing a crime analysis tool using different data mining techniques that can help police department to efficiently handle crime investigation and slow down the crime rate.

Over the past few years Geographic Information Systems (GIS) has become a standard tool for crime analysts in many police departments, regardless of their size. GIS assists one in seeing and understanding behaviour patterns, it also provides stakeholders with opportunities to join together in partnerships for the common good. A data pattern is an expression in some language describing a subset of the data or a model applicable to that subset. The application of data mining techniques in the field of crime can fetch important results. This is an area which includes exploring and detecting crimes and their relationships with criminals. (Chakravorty, 1995)

The research uses crime data of West Delhi which consists of the data sets comprising the records of crime occurred in West Delhi obtained from the Delhi Police department. There are several types of clustering algorithms like partitioning, hierarchical, model based and density

based etc. which exhibit their individual strengths and weaknesses. Hence, in order to determine the suitable clustering algorithm, K-means, and DB Scan techniques are implemented in the present paper. K-means is one of the most popular partitioning algorithms and DB Scan falls under density based category. The current research makes use of data mining algorithms K-means and DB Scan to analyse the trends in the data.

The applications of these data mining algorithms are endless and are determined by the requirement of the people and the community, time factor and other constraints.

The applications of data mining can be classified in the fields mentioned as under:

- In the field of marketing the targeted marketing programs are developed using this knowledge by discovering distinct groups in their customer bases.
- Helps in identification of areas of similar land use in the earth observation database.
- Successfully implemented in the field of Seismology.
- Medical and Scientific Research.
- Security agencies.

To a major extent the clustering algorithms have been widely used in the process of crime analysis only.

In the recent years a huge number of scientific researches and studies have been performed on crime data mining. These researches came up with new software applications for detecting and analysing crime data. In one of the research, crime analysis methods including neural networks, Bayesian networks, and genetic algorithms in forecasting and matching the crime incidents was introduced.(Oatley, Zelezniko, & Ewart, 2004)

K-means has been successfully used in health research too. The combination of Web GIS with k means has shown the most significant mortality rate for specific disease or cancers by creating maps K-means has produced a better mapping result. In this research K-means cluster algorithm allocates each point to one of a specified number of clusters and attempts to minimize the overall goodness of fit measures by using sum squared error. The process starts with, K-means randomly selecting the k representative as then cluster depending upon the distance between the representatives and the centre of the cluster it iteratively allocates the representatives to the closet' .Later it re-calculates the canter for each cluster and finally sees that the percentage of the points that change the membership is below the threshold by starting another round of membership task. (Zhang & Shi, 2010)

Successful investigations were carried out in the field of forensic sciences which is one of the most important areas of investigating crime by using the latest technology. K-means algorithm was also implemented in this field i.e., DNA and fingerprints and produced fruitful results. In the above research four cluster types existed within the datasets. Two different data sets were taken and K-Means algorithm was run on them where fingerprints and DNA were recovered which lead to an individual person being identified. (Adderley et al., 2007)

A research was conducted in United States in the medicine on the death cases caused by tumours in children every year. Mining the gene expression database for brain tumour was done which not only helped the doctors understand the tumour in a better way but also aided to provide more effective treatment for the tumour. This can be mentioned as one of the best applications of data mining in the field of medicine.(Larose, 2005)

Taking into view the related works carried out on the clustering algorithms, the research mainly use the data mining algorithms, K-means and DB Scan for its study. Due to their applications and implementation in various fields, the above two algorithms are made use of in the present research.

K-means is fast, robust and easier to understand and further gives best result when data set are distinct or well separated from each other. One problem that affects most police departments is the need to allocate throughout a city in a balanced and fair way. Too often, some police precincts or districts are overburdened with Calls for Service whereas others have more moderate demand. The issue of re-drawing or reassigning police boundaries in order to re-establish balance is a continual one for police departments. The K-means algorithm can help in defining this balance thought here are many other factors that will affect particular boundaries. The number of groupings, K, can be chosen based on the number of police districts that exist or that are desired. The locations of division or precinct stations can be entered in a secondary file in order to define the initial 'seed' locations. The K-means routine can then be run to assign all incidents to each of the K groups. The analyst can vary the location of the initial seeds or, even, the number of groups in order to explore different arrangements in space. Once an agreed upon solution is found, it is easy to then re-assign police beats to fit the new arrangement.
(Crime Stat, 2012)

DB Scan is one of the density based algorithms. The DB SCAN algorithm identifies the clusters of data objects based on the density-reachability and density-connectivity of the core and border points present in a cluster. An object not contained in any cluster is considered to be noise.
DB Scan algorithm was also successful in execution of the algorithm without input parameter.(Vijayalakshmi & Punithavalli, 2010)

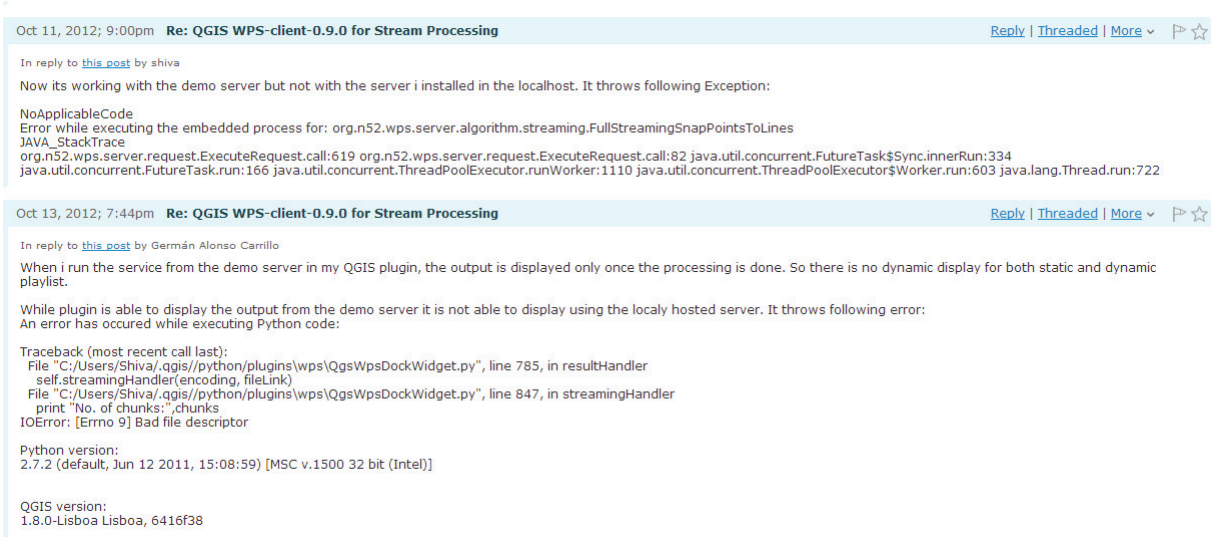
Studies have proved that DBSCAN algorithm a find arbitrarily shaped clusters. This algorithm has a notion of notion and requires just two parameters. The above algorithms have helped in developing crime analysis tool data mining techniques.

The methodology of this research is mainly divided into two main parts:

1. Comparison and selection of an algorithm
2. Design a framework for WPS service for crime analysis to incorporate geospatial clustering algorithm.

In the first phase, from different types of available clustering methods such as partition method (K-means) and density based methods (DB Scan) algorithms are selected. Selection is based on the literature review i.e., algorithms which have been used in most of the studies conducted on crime analysis and gave successful results. The text files are used as inputting the crime data. Here the data inputted by user (the police) are only considered and not retrieving from any other online resources. It is a good practice to store and index the crime data which would speed up the run time process of the algorithms which are executed on the web. After identification, these algorithms are used to generate various clusters i.e., hot spots of crime data with the help of standalone software. These algorithms are compared using the parameters such as, complexity, input parameters, the type of data that an algorithm supports, the shape of the clusters, ability to handle noise and outliers, clustering results and the clustering criterion etc.(Rehman & Mehdi, 2006). To correlate the crime data, satellite derived land use land cover map and also the other ancillary data such as, the census information to analyze the crime density are used. After formation of clusters of crime data, the clustered maps are overlaid on the land use map of the study area to know the crime density of that particular area (P. R. Canter, 1998). The other ancillary data as mentioned above would assist to extend crime analysis so as to analyze why that particular crime is happening in that area.

In the second phase, the geospatial clustering algorithm is incorporated in WPS. Based on algorithm, one of the WPS services 52 North WPS is used to incorporate the algorithm. For the cluster generation, standalone software R is used because the 52 North WPS is having backend R scripts. The writing algorithm code is uploaded to 52 North web admin console using upload R script option, then the algorithm is added to Local R Algorithm repository. After the adding algorithm, in the Get Capabilities request can show whether the uploaded algorithm is added or not. By using Describe Process request the user will be able to know the associate parameters required by that algorithm. But while executing the WPS request through the 52 North test client a java concurrent error occurred. Solution to this problem is not yet found as mentioned by many 52 North WPS users as shown below in figure 4.1.



Oct 11, 2012; 9:00pm **Re: QGIS WPS-client-0.9.0 for Stream Processing** Reply | Threaded | More

In reply to this post by shiva

Now its working with the demo server but not with the server i installed in the localhost. It throws following Exception:

NoApplicableCode
Error while executing the embedded process for: org.n52.wps.server.algorithm.streaming.FullStreamingSnapPointsToLines
JAVA_StackTrace
org.n52.wps.server.request.ExecuteRequest.call:619 org.n52.wps.server.request.ExecuteRequest.call:82 java.util.concurrent.FutureTask$Sync.innerRun:334
java.util.concurrent.FutureTask.run:166 java.util.concurrent.ThreadPoolExecutor.runWorker:1110 java.util.concurrent.ThreadPoolExecutor$Worker.run:603 java.lang.Thread.run:722

Oct 13, 2012; 7:44pm **Re: QGIS WPS-client-0.9.0 for Stream Processing** Reply | Threaded | More

In reply to this post by Germán Alonso Carrillo

When i run the service from the demo server in my QGIS plugin, the output is displayed only once the processing is done. So there is no dynamic display for both static and dynamic playlist.

While plugin is able to display the output from the demo server it is not able to display using the localy hosted server. It throws following error:
An error has occured while executing Python code:

```
Traceback (most recent call last):
  File "C:/Users/Shiva/.qgis//python/plugins\wps\QgsWpsDockWidget.py", line 785, in resultHandler
    self.streamingHandler(encoding, fileLink)
  File "C:/Users/Shiva/.qgis//python/plugins\wps\QgsWpsDockWidget.py", line 847, in streamingHandler
    print "No. of chunks:",chunks
IOError: [Errno 9] Bad file descriptor

Python version:
2.7.2 (default, Jun 12 2011, 15:08:59) [MSC v.1500 32 bit (Intel)]


QGIS version:
1.8.0-Lisboa Lisboa, 6416f38
```

Figure 4.1 : Error Screen Shot from 52 North geo processing community (52NorthWPS, 2012)

4.3. Research Flow

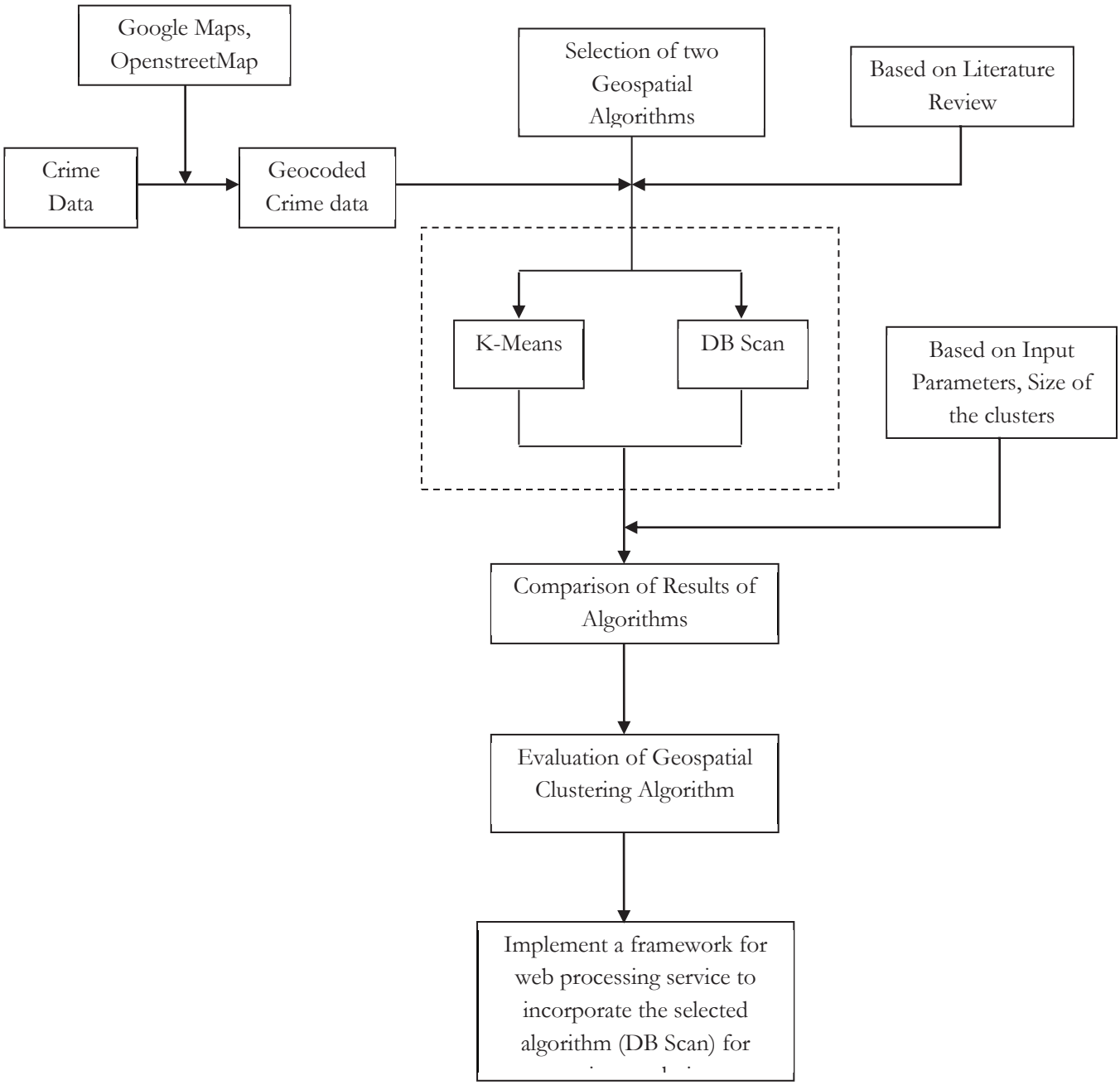


Figure 4.2 : Research Flow Diagram

4.4. System Flow

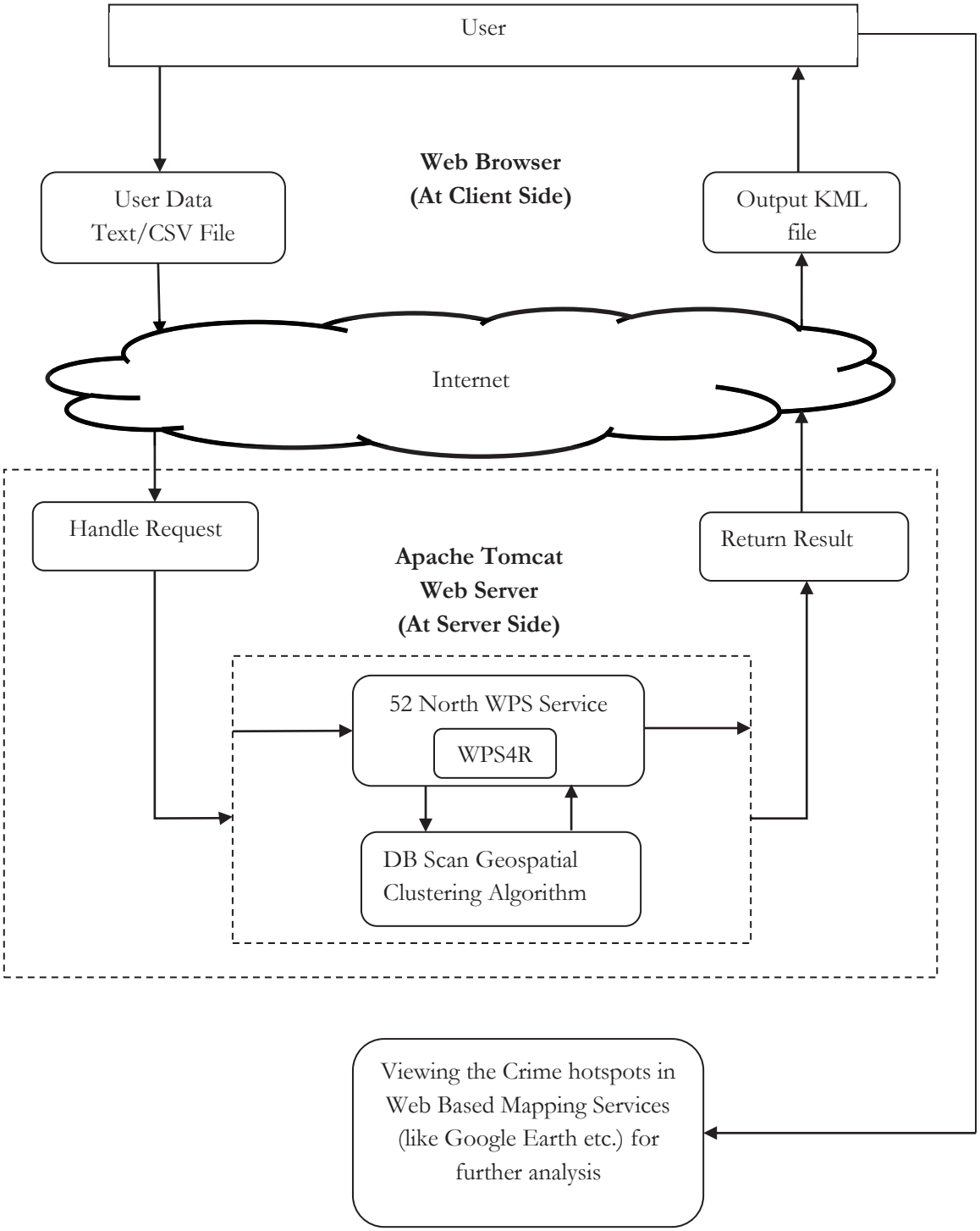


Figure 4.3 : System Flow Diagram

5. RESULTS AND DISCUSSION

The figure 5.1 shows the data of the study area. The plot shown depicts the data in terms of points on the image.

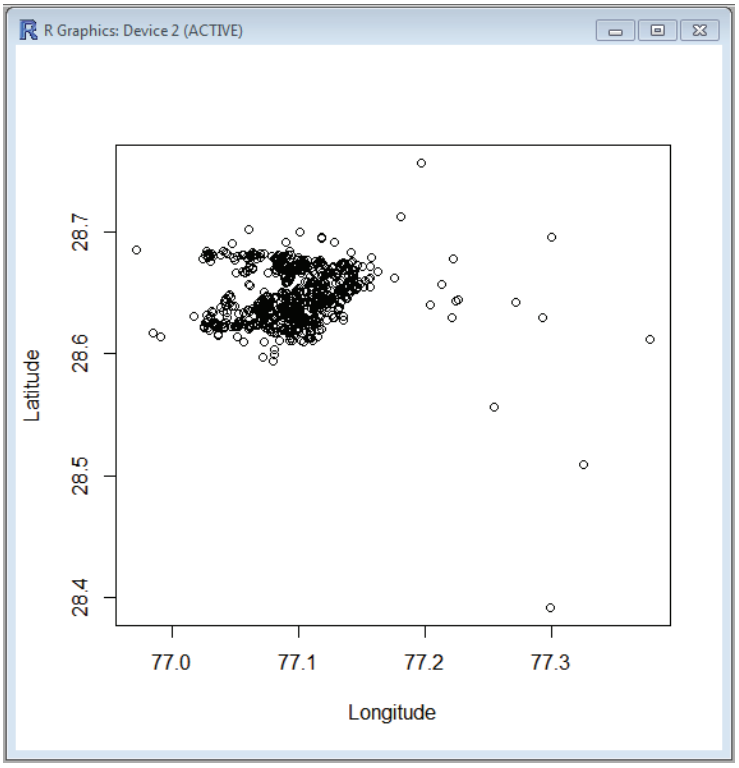


Figure 5.1 : Total Points of Crime in West Delhi

Below figure 5.2 shows the given point data which is used as input, using Delimited Text Layer. Coordinates of police stations in the study area are shown in red, in the same way as that of crime data. This is done to provide a visualization of the crime data, with respect to police stations locations.

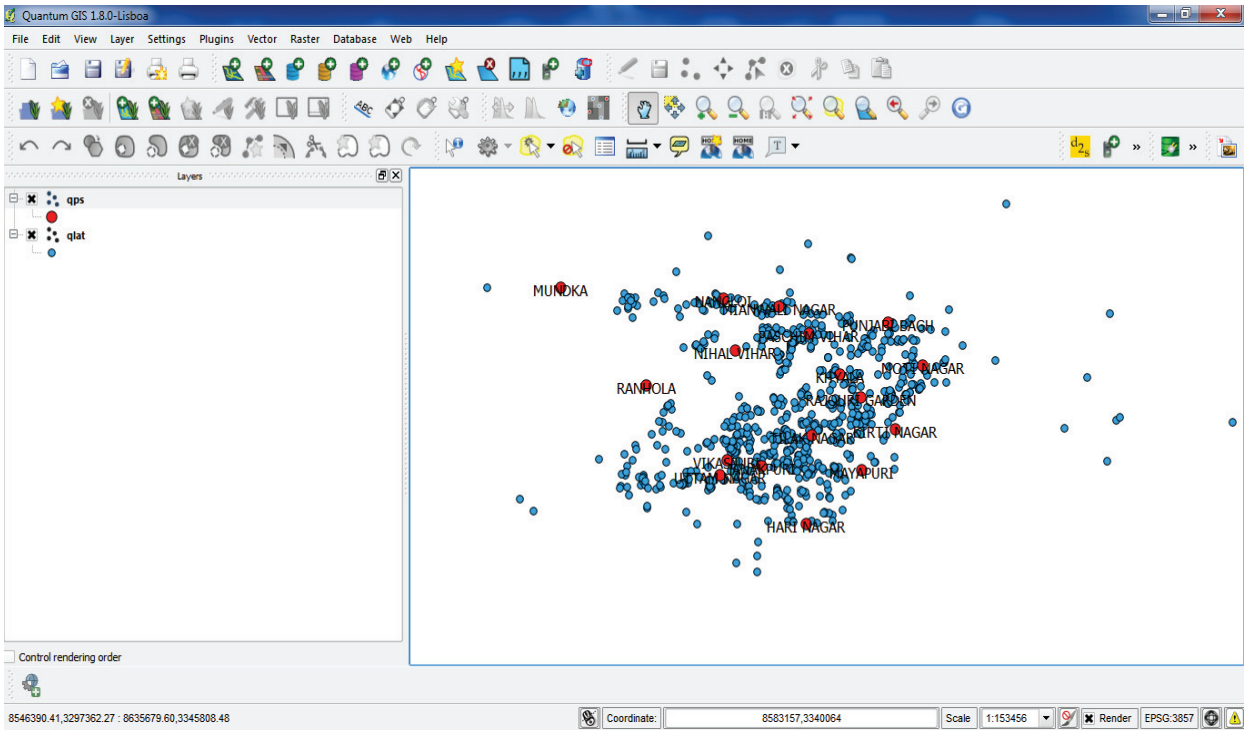


Figure 5.2 : Police Station Locations and Crime Incidents

The above data was converted into KML format and depicted in Google Maps to show the distribution of crime data in the study area.

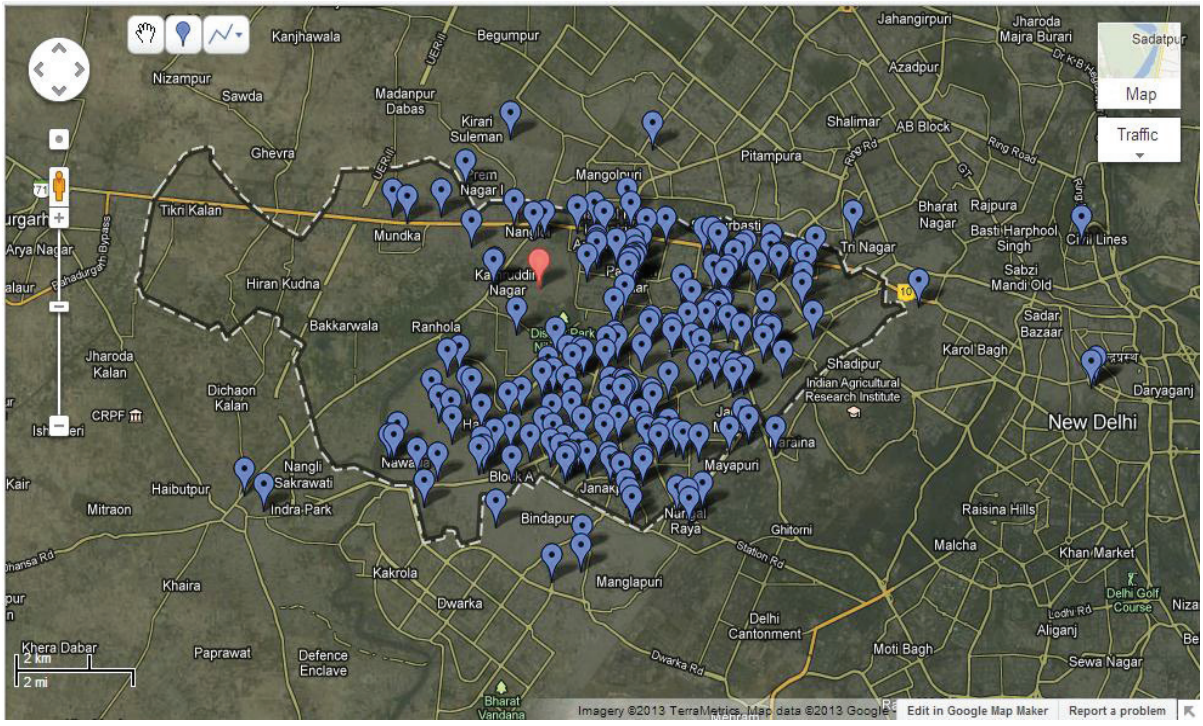


Figure 5.3 : Total Crime Incident locations in West Delhi with Google Maps

5.1. K-Means Algorithm Clusters:

In this algorithm, the value of K is considered to be 17, which is the number of police station in the study area. The crime density generated using K – means clustering algorithm defines the number of crimes falling in each of the police stations' jurisdiction. This helps in allocation of resources and better understanding of crimes and their occurrence. This algorithm is run using the R software, wherein the K means allows the generation of clusters with respect to the value of K, given as input. However, as the figures 5.4, 5.5, and 5.6 show, the K – means algorithm shows that random clusters are generated each time the code is executed. This is a disadvantage as the clusters form the main analysis parameter for the crime data. Also, the point allocated to which cluster is also obtained as information, as this defines which crime falls in the jurisdiction of which police station. The size of the cluster in each of the figures varies due to the presence of only one parameter. Also, the outlier data cannot be identified as it is also clustered in one of the classes.

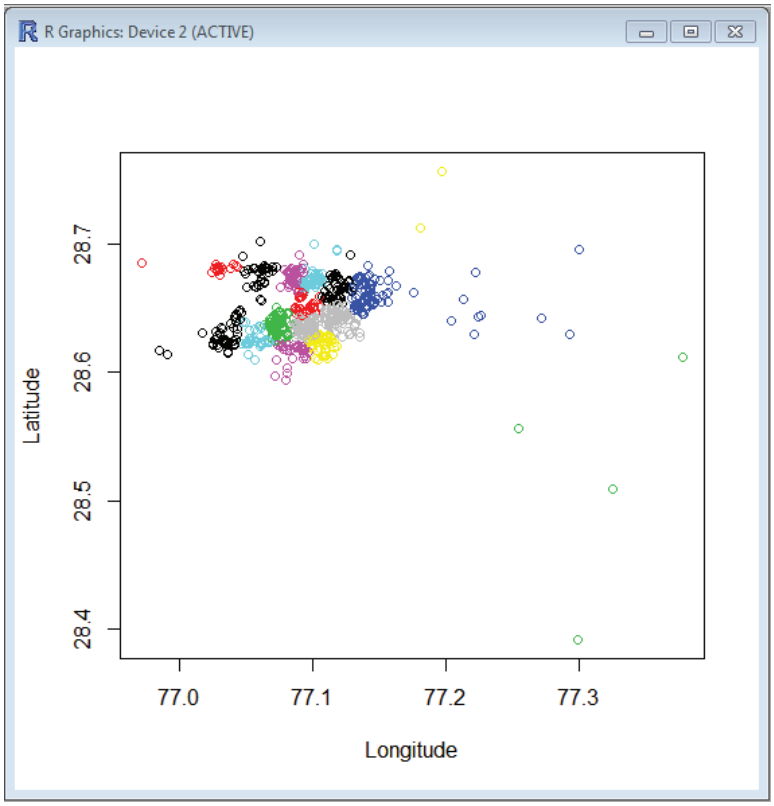


Figure 5.4 : K-means cluster with K=17

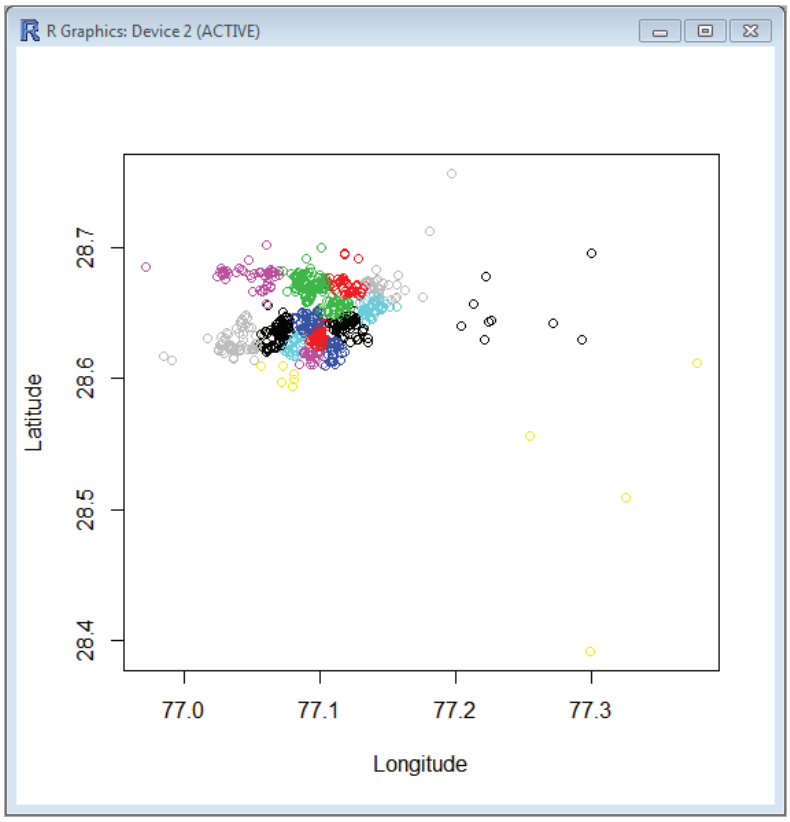


Figure 5.5 : K-means cluster with K=17

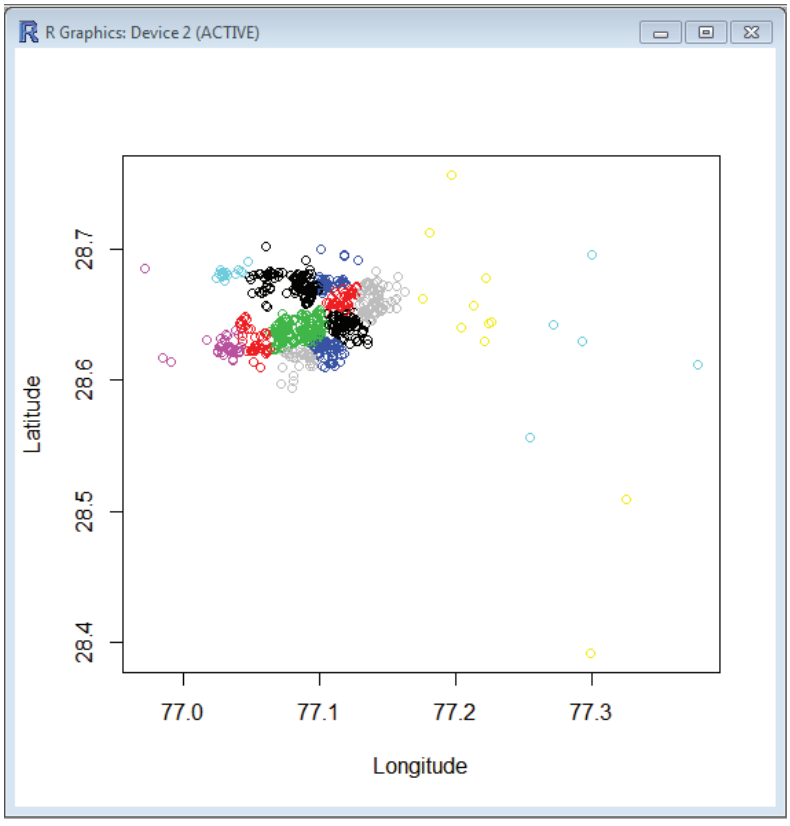


Figure 5.6 : K-means cluster with K=17

5.2. DB Scan Algorithm Clusters:

The parameters considered for DB Scan are EPS and minimum number of points for each cluster. These parameters are used to determine the radius of scanning while cluster formation and the number of points required in the formation of a cluster. Average distances between all the police stations, when considered, results in the formation of a single cluster. This is mainly due to the consideration of all the police stations, irrespective of their locations with respect to adjacent locations. The average of all the distances of separation of police stations, including close by and farther stations, results in a larger value, which when incorporated in DB Scan results in the allocation of all the crime data to one cluster. This has resulted in the consideration of DB Scan radius to be the minimum of half of the distance between the adjacent police station. This resolves the problem of clustering.

Also, the minimum number of points for clustering is determined by considering which police station has encountered the least number of crimes and considering the half of that value. This value is considered to be the minimum number of points for clustering.

Even consideration of half of the distance between adjacent police stations still gives a higher value. Therefore, assuming this minimum value of distances between all adjacent police stations is better suited for DB Scan algorithm.

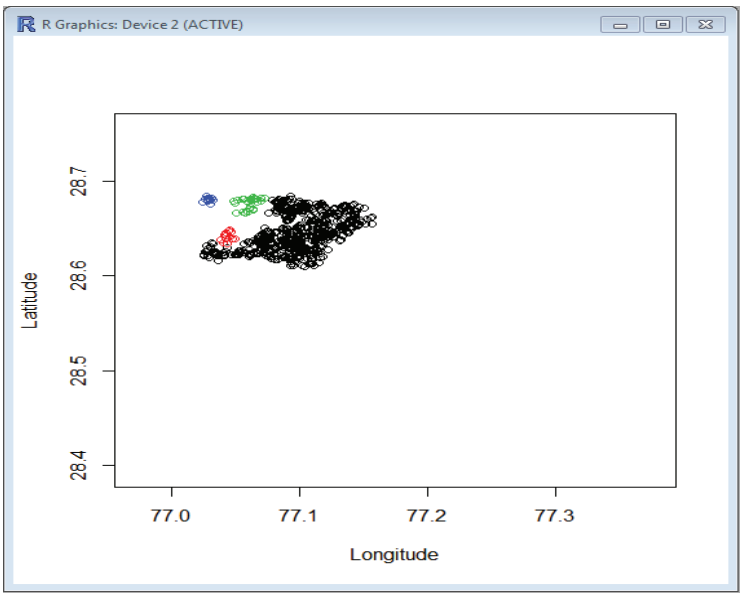


Figure 5.7 : DB Scan Clusters with eps=0.006, MinPts=6

EPS, when considered to 0.003, results in the formation of 28 clusters. The minimum number of points defining the cluster was set at 6, which is half of the minimum number of crimes in the police stations. This has resulted in the elimination of 249 points from the data as noise.

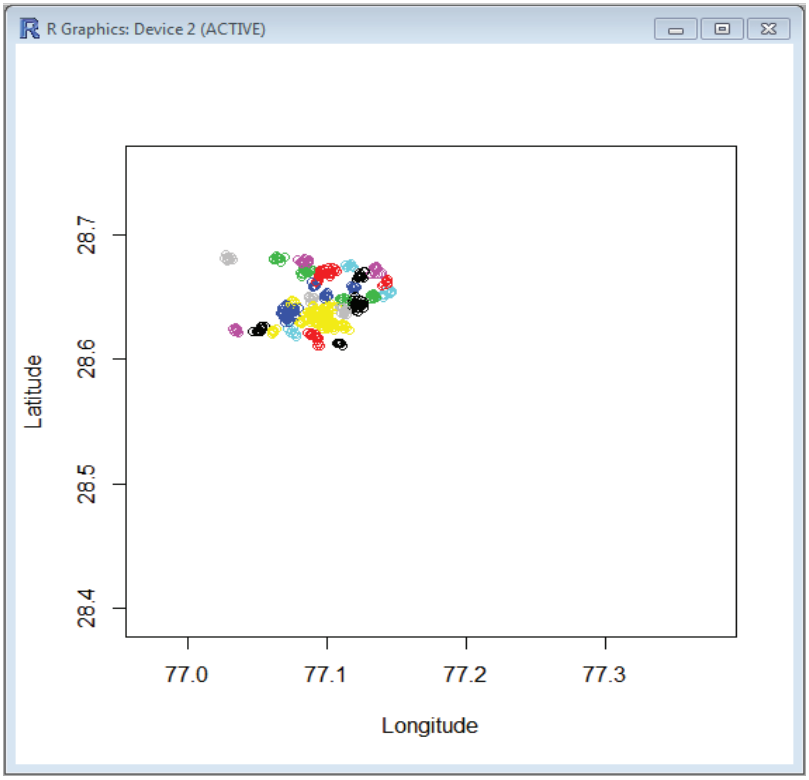


Figure 5.8 : DB Scan Clusters with eps=0.003, MinPts=6

EPS, when considered to 0.003, results in the formation of 34 clusters. The minimum number of points defining the cluster was set at 5, which is half of the minimum number of crimes in the police stations. This has resulted in the elimination of 200 points from the data as noise.

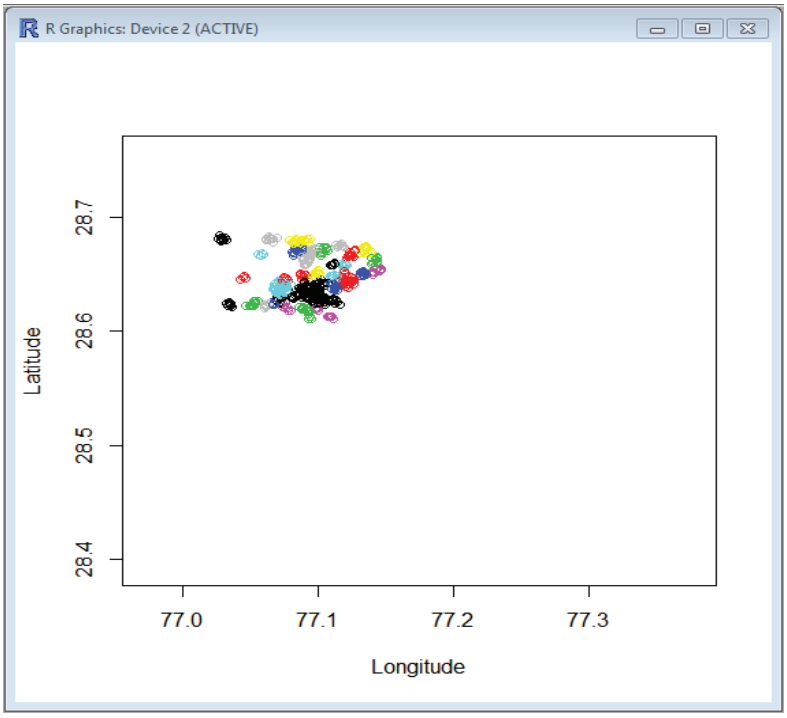


Figure 5.9 : DB Scan Clusters with eps=0.003, MinPts=5

From the above figures, it can be said that consideration of 5 points in a cluster to be minimum gives a better result. The allocation of 34 clusters with respect to 17 police stations and removal of 200 points as noise determines that this combination of 0.003 and 5 for EPS and minimum number of points respectively provides a better and more acceptable result.

The acceptance of results from DB Scan is better when compared to K – means clustering algorithm. The density of allocation of clustering is more acceptable for DB Scan, as it does not vary with the multiple executions of the R script. The DB Scan results are far more conclusive.

It is also to be noted that the above figures are not the results the user will encounter during the access and processing of data when using the WPS service. The figures show the process employed in this study, and not the end results. The user will deal with outputs of KML format, through the WPS service in the browser.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

6.1.1. Answers of Research Questions

1. What are the algorithms available for geospatial clustering and what are the software, tools and frameworks available to implement these algorithms?

   The various types of geospatial clustering algorithms are available, namely, partition methods, hierarchical methods, density based methods, grid based methods, model based methods. Out of these, for this study, Partition method and density method are considered. K-means from partition method and DB Scan from density based are the main algorithms for selected for this study based on the literature review. The choice has previously been justified in the methodology.

   Clusters of crime data have been formed using these algorithms, utilising the packages and frameworks of R software. The main motive behind choosing R software is due to its compatibility with 52North WPS service and presence of predefined functions of K-Means and DB Scan algorithms. The advantage of using 52North service is due to its compatibility with backend R scripts, to provide a Web Processing Service.

2. How to evaluate the geospatial clustering algorithm for crime analysis?

K-means algorithm deals basically with the clusters of point data. K value was considered to be equivalent to the number of police stations to form the same number of clusters of crime, in order to calculate the number of crimes falling under the jurisdiction of each of the stations. The disadvantage of this process is the resultant generation of random clusters, of varying sizes, for the same parameter of number of police stations, i.e. 17.
DB Scan algorithm, however deals with two parameters, requires a radius of scanning and a minimum number of points required to define a cluster. Optimum radius was considered to be the half of the distance between police stations, due to the fixed nature of locations of police stations. Minimum of half of the distance is taken as an attempt to reduce the number of crimes calculated by algorithm to lie in a conflicted jurisdiction. The added constraint of minimum distance ensures that police stations at the far ends of the study area are not considered, and only the adjacent stations are taken. The other parameter of minimum points is estimated after an analysis of the number of crimes occurring in the jurisdictions of police stations, and the minimum is considered, and the half of those number of cases were considered as the values for input into DB Scan algorithm. This algorithm gives a constant set of clusters, irrespective of number of times the algorithm is executed. The outlier data is excluded as noise in the cluster formation procedure, unlike that of K-means which gave random outputs. So DB Scan algorithm is evaluated to serve as a WPS service.

3. Is the socio economic layer may be used to extend the crime analysis?

   The socio economic layers like LULC, census data, etc. can be used to extend crime analysis. Linking crime analysis data with other socio economic layers of data helps in extending the understanding of the occurrence of crime, locations, factors influencing crime, etc.

4. How can the selected geospatial algorithm be served through WPS?

   The algorithm written in R is uploaded into the 52North Web Admin Console. It is added into the local R algorithm repository. To check the process request, WPS Get Capabilities request is used. The user can know the associated input and output parameters through the Describe Process request. Test Client is used for running the required process and obtaining the clustering result.

5. What are the input and output data formats/data structures suitable for WPS?

   KML and GeoRSS are the most widely used as input for WPS services. Police inputs are mostly in text formats, wherein KML and GeoRSS inputs are difficult. Therefore, this

study deals with accepting input files through Text or Csv formats, however, providing them with KML outputs, which are Google Earth compatible, which makes it easier to analyse the crime data.

6. What is role of WPS in addressing interoperability?

   This service of WPS is platform independent. Irrespective of number of police stations and their equipment, the data from all these sources can be input and processed.

7. Is Web Processing Service suitable for Crime Analysis?

   WPS is suitable for crime analysis. Instead of desktop based software for processing crime data, on fly WPS processing enhances the ability of the user to access and analyse the data through the internet. Service interoperability is also an advantage for this service to do Crime analysis process. The service helps in allocating resources for border line cases or jurisdiction conflicts. Any number of users (police personnel) can access the data and subsequently process it for required outputs.

6.2. Recommendations

The process of cluster formation can be enhanced by using the concept of indexing. If the database connected to the server is previously indexed, the formation of clusters and searching of addresses is greatly enhanced. Also, the speed of finding the address or the location increases because it is not necessary to scan the whole database for the required data.

In this research only DB Scan was used to provide WPS service. But by providing multiple algorithms as WPS service will help the user to access multiple algorithms on fly.

REFERENCES

52NorthWPS. (2012). 52 North GeoProcessing Community Forum. Retrieved from http://geoprocessing.forum.52north.org/QGIS-WPS-client-0-9-0-for-Stream-Processing-td4025029.html

Ackerman, W. V., & Murray, A. T. (2004). Assessing spatial patterns of crime in Lima, Ohio. *Cities*, *21*(5), 423–437. doi:10.1016/j.cities.2004.07.008

Adderley, R., Townsley, M., & Bond, J. (2007). Use of data mining techniques to model crime scene investigator performance. *Knowledge-Based Systems*, *20*(2), 170–176. doi:10.1016/j.knosys.2006.11.007

Ahmadi, M. (2003). *Crime Mapping and Spatial Analysis*.

Bergenheim, W., Sarjakoski, L. T., & Sarjakoski, T. (2009). A web processing service for GRASS GIS to provide on-line generalisation. In *Proceedings of the 12th AGILE International Conference on Geographic Information Science* (pp. 1–10). Retrieved from http://agile.gis.geo.tu-dresden.de/web/Conference_Paper/CDs/AGILE%202009/AGILE_CD/pdfs/126.pdf

Canter, P. (2000). Using a geographic information system for tactical crime analysis. *Analyzing crime patterns: Frontiers of practice*, 3–10.

Canter, P. R. (1998). Geographic information systems and crime analysis in Baltimore County, Maryland. *Crime mapping and crime prevention*, *8*, 157–190.

Chakravorty, S. (1995). Identifying crime clusters: The spatial principles. *Middle States Geographer*, *28*, 53–58.

Chandra.E, (first), & Anuradha.V.P. (2011). A Survey on Clustering Algorithms for Data in Spatial Database Management Systems. *International Journal of Computer Applications*, *24*(9), 19–26. doi:10.5120/2969-3975

Chen, H., Atabakhsh, H., Petersen, T., Schroeder, J., Buetow, T., Chaboya, L., … Casey, D. (2003). COPLINK: Visualization for crime analysis. In *Proceedings of the 2003 annual national conference on Digital government research* (pp. 1–6). Retrieved from http://dl.acm.org/citation.cfm?id=1123230

Crime Stat. (2012). Hotspot Analysis II. Retrieved from http://www.icpsr.umich.edu/CrimeStat/files/CrimeStatChapter.7.pdf

DelhiPolice. (2013). Retrieved from http://www.delhipolice.nic.in/

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining* (Vol. 1996, pp. 226–231). Retrieved from http://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science*, *22*, 138–144.

Fenoy, G., Bozon, N., & Raghavan, V. (2012). ZOO-Project: the open WPS platform. *Applied Geomatics*, *5*(1), 19–24. doi:10.1007/s12518-011-0070-0

Granell, C., Lemmens, R., Gould, M., Wytzisk, A., De By, R., & Van Oosterom, P. (2006). Integrating semantic and syntactic descriptions to chain geographic services. *Internet Computing, IEEE*, *10*(5), 42–52.

Grubesic, T. H., & Murray, A. T. (2001). Detecting hotspots using cluster analysis and GIS. In *Fifth Annual International Crime Mapping Research Conference*. Presented at the Annual International Crime Mapping Research Conference, Dallas, TX.

Grubesic, Tony H., & Mack, E. A. (2008). Spatio-Temporal Interaction of Urban Crime. *Journal of Quantitative Criminology*, *24*(3), 285–306. doi:10.1007/s10940-008-9047-5

Hammouda, K., & Karray, F. (2000). A comparative study of data clustering techniques. *Tools of intelligent systems design. In Course Project, SYDE*, *625*. Retrieved from http://www.pami.uwaterloo.ca/pub/hammouda/sde625-paper.pdf

Honarkhah, M., & Caers, J. (2010). Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling. *Mathematical Geosciences*, *42*(5), 487–517. doi:10.1007/s11004-010-9276-7

Jacquez, G. M. (n.d.). Spatial Cluster Analysis. In J. P. Wilson & A. S. Fotheringham (Eds.), *The Handbook of Geographic Information Science* (pp. 395–416). Oxford, UK: Blackwell Publishing Ltd. Retrieved from http://doi.wiley.com/10.1002/9780470690819.ch22

Larose, D. T. (2005). *Discovering knowledge in data: an introduction to data mining*. Hoboken, N.J: Wiley-Interscience.

MapsofIndia. (2013). Retrieved from http://www.mapsofindia.com/delhi/districts/west-delhi.html

Ming-Syan Chen, Jiawei Han, & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866–883. doi:10.1109/69.553155

National Crime Records Bureau. (2012). Retrieved from http://ncrb.nic.in/

Oatley, G. ., . Zeleznikо, J., & Ewart, B. W. (2004). Matching and Predicting Crimes. In *Applications and Innovations in Intelligent Systems XII in Proceedings of AI2004*. Presented at the The Twenty-fourth SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence.

Phillips, P., & Lee, I. (2010). Crime analysis through spatial areal aggregated density patterns. *GeoInformatica*, *15*(1), 49–74. doi:10.1007/s10707-010-0116-1

Ratcliffe, J. (2010). Crime Mapping: Spatial and temporal challenges. *Handbook of quantitative criminology*, 5–24.

Rehman, M., & Mehdi, S. A. (2006). Comparison of density-based clustering algorithms. Lahore College for Women University, Lahore, Pakistan, University of Management and Technology, Lahore, Pakistan. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download? &rep=rep1&type=pdf

Rogerson, P., & Sun, Y. (2001). Spatial monitoring of geographic patterns: an application to crime analysis. *Computers, Environment and Urban Systems*, *25*(6), 539–556.

Santhosh Kumar, C. . (2012). Spatial Data Mining using Cluster Analysis. *International Journal of Computer Science and Information Technology*, *4*(4), 71–77. doi:10.5121/ijcsit.2012.4407

Santos, R. B. (2012). *Crime Analysis With Crime Mapping*. Sage Publications, Incorporated.

Schaeffer, B. (2008). Towards a Transactional Web Processing Service. *Proceedings of the GI-Days, Münster*. Retrieved from http://www.gi-tage.de/archive/2008/downloads/acceptedPapers/Papers/Schaeffer.pdf

Schut, P. (2007). Open Geospatial Consortium Inc. OpenGIS® Web Processing Service. Open Geospatial Consortium.

Sikder, I. U. (2008). Geospatial Web Services in environmental planning. In *Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on* (pp. 424–429). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4803101

Stollberg, B., & Zipf, A. (2007). OGC Web Processing Service Interface for Web Service Orchestration Aggregating Geo-processing Services in a Bomb Threat Scenario. *Web and Wireless Geographical Information Systems*, 239–251.

Sun, C., Wang, G., Mu, B., Liu, H., Wang, Z., & Chen, T. Y. (2011). Metamorphic Testing for Web Services: Framework and a Case Study (pp. 283–290). IEEE. doi:10.1109/ICWS.2011.65

The Times of India. (2012). Articles Times of India. Retrieved from http://articles.timesofindia.indiatimes.com/2012-07-03/delhi/32522760_1_vehicle-owners-theft-cases-motor-vehicle-thefts

Vijayalakshmi, S., & Punithavalli, M. (2010). Improved varied density based spatial clustering algorithm with noise. In *Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on* (pp. 1–4). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5705763

Wang, J., Wang, X., & Liang, S. H. L. (n.d.). GeoClustering: A Web Service for Geospatial Clustering. Retrieved from http://www.ucalgary.ca/wangx/files/wangx/chapter_draft.pdf

Web Services Architecture. (2012). Retrieved March 10, 2013, from http://www.w3.org/TR/ws-arch/

Wehrmann, T., Gebhardt, S., Klingera, V., & Künzer, C. (2010). Web services enabled architecture coupling data and functional resources. *ISPRS Commission IV–Geospatial Data and Geovisualization: Environment, Security and Society*, 15–19.

Wilson, R. E. (2002). *Mapping Crime Understanding Hot Spots*. Presented at the Mapping & Analysis for Public Safety Program.

Wu, X., & Grubesic, T. H. (2010). Identifying irregularly shaped crime hot-spots using a multiobjective evolutionary algorithm. *Journal of Geographical Systems*, *12*(4), 409–433. doi:10.1007/s10109-010-0107-7

Xiang, Y., Chau, M., Atabakhsh, H., & Chen, H. (2005). Visualizing criminal relationships: comparison of a hyperbolic tree and a hierarchical list. *Decision Support Systems*, *41*(1), 69–83. doi:10.1016/j.dss.2004.02.006

Xiong, L. (2001). Data Mining: Concepts and Techniques. Retrieved from http://www.mathcs.emory.edu/~lxiong/teaching/cs570/share/slides/07.pdf

Xu, C., & Xinyan, Z. (2009). Semantic discovery of OGC WPS-based remote sensing image processing. In *Geoinformatics, 2009 17th International Conference on* (pp. 1–5). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5293510

Yang, Q. (2010). *Clustering*. Retrieved from www.cse.ust.hk/~qyang/337/slides/cluster.ppt

Zhang, J., & Shi, H. (2010). Geo-visualization and Clustering to Support Epidemiology Surveillance Exploration (pp. 381–386). IEEE. doi:10.1109/DICTA.2010.71