Cluster Detection and Analysis with Geo-spatial Datasets Using a Hybrid Statistical and Neural Networks Hierarchical Approach

Salar Mustafa Majeed

Faculty of Advanced Technology

University of Glamorgan

Pontypridd

March 2010

A thesis submitted in partial fulfilment of the requirements of the University of Glamorgan for the degree of Doctor of Philosophy

TABLE OF CONTENTS

Appendix E

LIST OF FIGURES

LIST OF TABLE

Acknowledgements

I would like to express my sincere thanks to my supervisors, Dr Ian D Wilson, Dr Jamal RM Ameen and Prof. Andrew J Ware for their support, guidance and constructive criticism they have provided throughout this research project.

I would like to thank Ministry of Higher Education and Scientific Research of Kurdistan, University of Salahaddin, College of Eduction and Department of computing, in particular, Prof. Dr Karim S Abdul, Dr Mohamad S Mohamad and Dr Ahmed A Dezaye for allowing me to carry out this research on A full- time at the University of Glamorgan.

Thanks are due to my fellow researchers for their support throughout my time at the University of Glamorgan, in particular, Robert, Christian and Richard.

My heartfelt appreciation goes to my parents, husband, sisters, brothers, Prof. Dr Nabil A Fakhre and my son Ranj for their continued love, support and patience.

Finally, thanks all the members of Computing and Mathematic Department at the University of Salahaddin.

Declarations

Certificate of Research

This is to certify that, except where specific reference is made, the work presented in this thesis is the result of the investigation undertaken by the candidate

Candidate .....

Director of studies .....

Certificate of Research

This is to certify that neither this thesis nor any part of it has been presented or is being currently submitted in candidature for any other degree other than the degree of Doctor of Philosophy of the University of Glamorgan.

Candidate .....

ABSTRACT

Spatial datasets contain information relating to the locations of incidents of phenomena for example, crime and disease. Areas that contain a higher than expected incidence of the phenomena, given background population and census datasets, are of particular interest. By analysing the locations of potential influence, it may be possible to establish where a cause and effect relationship is present in the observed process.

Cluster detection techniques can be applied to such datasets in order to reveal information relating to the spatial distribution of the cases. Research in these areas has mainly concentrated on either computational or statistical aspects of cluster detection. Each clustering algorithm has its own strengths and weakness. Their main weaknesses causing their unreliability can be estimating the number of clusters, testing the number of components, selecting initial seeds (centroids), running time and memory requirements. Consequently, a new cluster detection methodology has been developed in this thesis based on knowledge drawn from both statistical and computing domains. This methodology is based on a hybrid of statistical methods using properties of probability rather than distance to associate data with clusters. No previous knowledge of the dataset is required and the number of clusters is not predetermined. It performs efficiently in terms of memory requirements, running time and cluster quality. The algorithm for determining both the centre of clusters and the existence of the clusters themselves was applied and tested on simulated and real datasets. The results which were obtained from identification of hotspots were compared with results of other available algorithms such as CLAP (Cluster Location Analysis Procedure), Satscan and GAM (Geographical Analysis Machine). The outputs are very similar.

GIS presented in this thesis encompasses the SCS algorithm, statistics and neural networks for developing a hybrid predictive crime model, mapping, visualizing crime data and the corresponding population in the study region, visualizing the location of obtained clusters and burglary incidence concentration 'hotspots' which was specified by clustering algorithm SCS. Naturally the quality of results is subject to the accuracy of the used data. GIS is used in this thesis for developing a methodology for modelling data containing multiple functions. The census data used throughout this construction provided a useful source of geo-demographic information. The obtained datasets were used for predictive crime modelling.

This thesis has benefited from several existing methodologies to develop a hybrid modelling approach. The methodology was applied to real data on burglary incidence distribution in the study region. Relevant principles of statistics, Geographical Information System, Neural Networks and SCS algorithm were utilized for the analysis of observed data. Regression analysis was used for building a predictive crime model and combined with Neural Networks with the aim of developing a new hierarchical neural Network approaches to generate a more reliable prediction. The promising results were compared with the non-hierarchical neural Network back-propagation network and multiple regression analysis. The average percentage accuracy achieved by the new methodology at testing stage increase 13% compared with the non-hierarchical BP performance. In general the analysis reveals a number of predictors that increase the risk of burglary in the study region. Specifically living in a household in which there is 'one person', 'lone parent', household where occupations are in elementary or intermediate and unemployed. For the influence of Household space, the results indicate that the risk of burglary rate increases within the household living in shared houses.

1 Introduction

This Chapter presents the thesis outline together with the objectives of the research. The content of subsequent chapters is described.

1.1 Background

The following section provides the background for the diverse topics addressed in the objectives.

1.1.1 Cluster detection

Cluster detection is mainly an exploratory data analysis process which aims to sort different objects into groups or clusters; clusters are sets of observed data with a significant degree of 'similarity' within each set and a significant degree of dissimilarity between unrelated sets. The popular measure to assess the similarity between pairs of observations or clusters is Euclidean distance. Distance measures are problematic if the elements to be clustered have many categorical attributes. The thesis presents a new cluster detection methodology, called Salar's Clustering with Significance (SCS) section 3.3. A new methodology is based on a hybrid of statistical method using properties of probability rather than distance to associate data with clusters. Details on this can be found in section 3.3.2. Clusters in this thesis are regions of high density separated by regions of lower density. The algorithm was tested using both artificial and real (crime) spatial datasets

The principal functions of clustering are to name, display, summarise, predict and explain. Thus, all objects in the same cluster will be given the same name 'looking for characteristic properties'. Objects are displayed, in order that subtle differences

may become more apparent by physically adjoining all objects in the same cluster. Data are summarized by referring to properties of clusters rather than to properties of individual objects'. Such a summary makes the data easier to understand and to manipulate (Hartigan, 1975:6).

Hotspot

Hotspots are specific locations or small areas that suffer a large number of incidents (Rachel, 2005: 273). There are several ways to identify hotspots such as Graduated colour mapping, grid mapping and density mapping. Hotspots generally fall into one of the following three categories:

Dispersed hotspot: The incident locations within the hotspot are spread throughout the hotspot area but are more concentrated than incidents in other areas.

Clustered hotspot: The incident locations within the hotspot group together in one or more smaller clusters.

Hot point: The incidents occur at one particular place. Unlike the incidents in a clustered hotspot, which form one or more clusters, the incidents in a hot point are centered at one address or place (Rachel, 2005:273).

A new cluster detection methodology was utilized for performing the procedure of identification hotspots. The clusters with high levels of crime (hotspots) are those with a crime rate greater than or equal to crime rate of the study region. The details of this procedure can be found in section 4.6.1. The results which were obtained from identification of hotspots were compared with results of other available algorithms such as CLAP, Satscan and GAM. The outputs are very similar (Section 4.8).

Possible Applications

Clustering algorithms can be applied in many fields, for instance:

- Marketing: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records (Egeli, 2003);

- Biology: classification of plants and animals given their features (Yona, 1999);
- Libraries: book ordering;
- Healthcare: Health status related to areas of deprivation (Richards and Ameen,2005);
- City-planning: identifying groups of houses according to their house type, value and geographical location (Jarvis and Wilson, 2005);
- Earth quake studies: clustering observed earth quake epicenters to identify dangerous Zones (Teanby and Kendall, 2004);
- Medicine: clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies (Neill, 2006);
- Social network analysis: clustering may be used to recognize communities within large groups of people (Yu, 2007);
- Crime zoning: determining area exhibiting elevated concentration of crime (Hartigan, 1975 and Corcoran, 2003).

Over the recent past and with the increasing power of computers, research on cluster detection and analysis has grown quickly in many disciplines simultaneously and often independently of each other. Amongst these disciplines, a few stand out as being especially important for the development of cluster analysis.

Limitations in the study of spatial systems

The spatial analyst's tool box includes techniques for quantifying spatial patterns, modelling risk surfaces, and assessing relationships between the outcomes and potential exposures. These techniques allow researchers to determine whether observed spatial patterns are statistically significant, to identify the locations of clusters, hotspots and cool spots, to construct maps showing excesses and deficits relative to a risk model, and to quantify association between two spatial variables (such as cancer incidence and putative environmental exposures). Although these techniques can be quantitatively powerful, the inferences that can be drawn from them have attendant limitations. These can be faced through the analyses of spatial

patterns, spatial associations and/or the use of randomization (Monte Carlo)-based techniques.

All methods have attendant limitations while those related to spatial methods include the amount of knowledge required, the selection and specification of spatial weights and the subjectivity of the methods themselves.

The spatial data used in many geographic studies have inherent limitations attributable to granularity, spatial and temporal mismatch, under-reporting, misdiagnosis, the use of location as an exposure surrogate, human mobility, location and attribute uncertainty, static representation, as well as topological errors that result in erroneous spatial weights (Geoffrey, 2004).

1.1.2 Geographical Information Systems

Geographical Information Systems (GIS) are computer assisted systems for the storage, integration, analysis and display of geographic data. GIS has become more evident to researchers in an expanding array of disciplines, such as the demand for spatial analysis, spatial modelling and spatial statistics (Stillwell, 2004). GIS are used for handling maps of different kinds, represented as several different layers, where each layer represents a unique phenomenon. The data share a common location which allows the integration of data from all sources and types under a single platform. Researchers integrate data to reveal trends and relationships that bring new perspectives to previously held beliefs about places and events. For example possible layers that are used to analysis and present the spatial distribution of disease are: parasite drug resistance, average age per census tract, hospital and health stations, patients' home locations, census boundaries and zip code. GIS can reveal and display spatial patterns hidden in tables and databases. Display of geographic data can be adjusted by changing the symbols, colors and legend classifications and analysis information in historical records, images and maps.

GIS are used for a variety of purposes and functions. GIS provides means for managing business information of many kinds a according to its location. Bringing together data with a shared spatial component reveal trends and patterns that are not apparent with tabular databases. Businesses have used GIS to:

- Analyse markets;
- Modelling spending patterns;
- Analyse parcels of land;
- Optimizing media campaigns;
- Creating sales territories;
- Selecting future business sites.

Marine Biology Research use GIS to:

- Store, map and analyse data from seafloor mapping expeditions
- Monitor species distribution, abundance and migration patterns;
- Map sources and paths of pollutants in marine environments;
- Manage Coastal zones.

Humanities Research use GIS to:

- Place historical analysis in geographic context;
- Determine and illustrate changes through time;
- Interpret texts in relation to historical maps;
- Analyse and present the spatial distribution of literature, art or material culture;
- Map linguistic, ethnic and cultural traits.

GIS has become one of the most important developments in crime analysis. By combining geographic principles and geo coded location data with crime data and criminological theories, GIS allows the analysis of crime incidents across time and space.

The new cluster detection technique (SCS) accommodates GIS in terms of a prediction crime model; by mapping and displaying distribution of crime in a study region, counting number of points within polygons and integration of different data layer. These processes are illustrated in chapter five.

1.1.3 Artificial Neural Networks

Artificial Neural Networks (ANN) "are computational modelling tools that have recently re-emerged and found extensive acceptance in many disciplines for modelling complex real-world problems" (Basheer and Hajmeer 2000). ANNs are typically organised in layers, with each layer conected to the next. Layers are made up of a number of nodes which contain an activation function. Patterns are presented to the network via the input layer, which communicates to one or more 'hidden' layers where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to on output layer where the result is output.

The attractiveness of ANNs comes from "their remarkable information processing characteristics pertinent mainly to non linearity, high parallelism, fault and noise tolerance, learning and generalization capabilities" (Basheer and Hajmeer 2000).

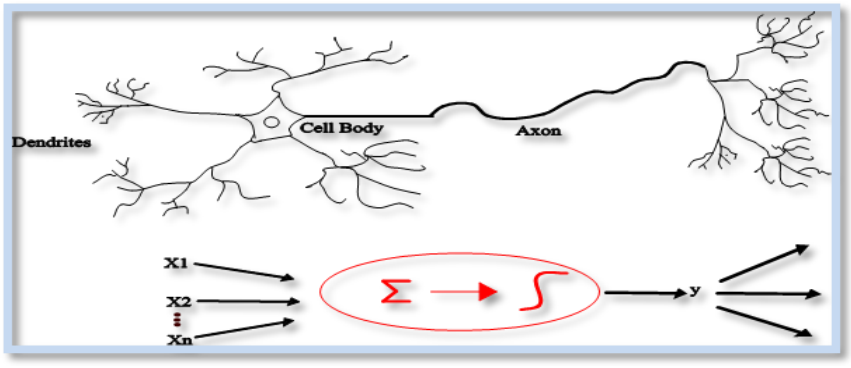


Figure 1.1 Real neuron and artificial neuron model (Daniel, 2005)

Transfer function

The transfer function of a unit sums the weighted input (net) from all connected units and squashes it into a finite range of values (Swingler, 1996: 62). The purpose of this transformation is to modify the output levels to a reasonable value, 0 to 1 or -1 to 1. The chosen function is typically differentiable, non-linear and monotonic to provide a smooth mapping between continuous variables. Several most frequently used transfer functions are: step function, logistic sigmoid function, hyperbolic tangent function and linear function. The logistic sigmoid function and hyperbolic tangent (tanh) function are most commonly used. The logistic sigmoid (s- shaped) whose shape is shown in Figure 1.2 is a real function F: R → (0,1), defined by the expression

$$f(net_i) = \frac{1}{1 + \exp(-net_i)} \qquad 0 < f < 1 \qquad (1.1)$$

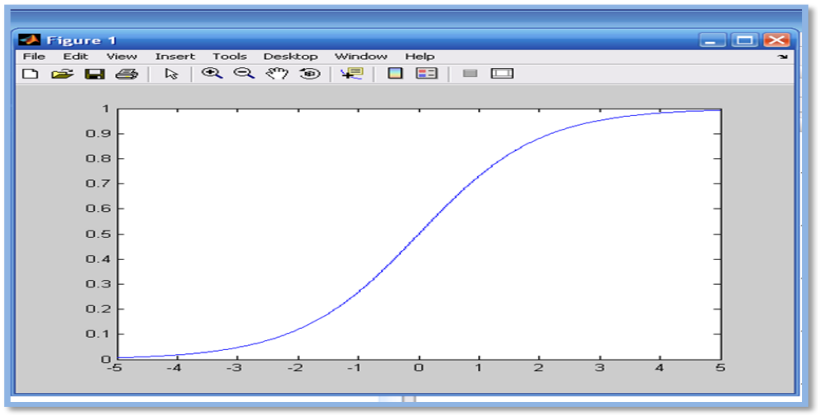


Figure 1.2 Graph of logistic sigmoid function(s- shaped)

The hyperbolic tangent function is a sigmoid curve, like the logistic function except that output lies in the range (-1, 1). An alternative to the sigmoid is the symmetrical sigmoid. Whose shape is shown in Figure 1.3 and is defined by the expression

$$f(net_i) = \frac{e^{net_i} - e^{-net_i}}{e^{net_i} + e^{-net_i}} \qquad -1 < f < 1 \qquad (1.2)$$

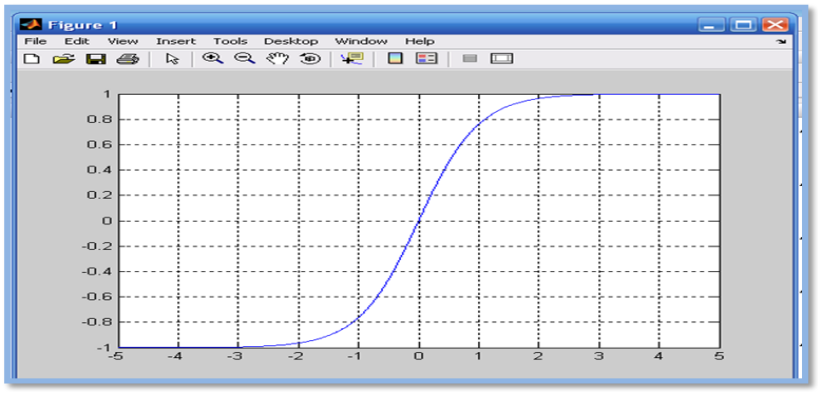


Figure 1.3 Graph of hyperbolic tangent function (symmetrical sigmoid curve).

Selecting a transfer function is determined by the nature of the data and what the network is trying to learn. The experiments which presented in this thesis the logistic sigmoid function suggested for back propagation network and tanh for hierarchical network. The evidence is purely empirical (Section 6.9).

Number of hidden layers and nodes

The size of networks depends on the number of layers and the number of hidden-units per layer. In a feed-forward multilayer neural net, there are one or more layers of hidden neuron units between the input and output neuron layers (Kung, 1993:31). The number of hidden layer and neuron are not predetermined, but are solved in practice by trial and error. Feng (2006) suggested that a network start with a one –hidden layer. If the one hidden layer dose not train well, then the number of neurons or the training and testing tolerances or both can be changed. The accuracy of the resulting model is affected by the number of hidden neurons. Since the number of hidden neurons directly affects the number of parameters in the model, a neural net needs a sufficient number of hidden neurons to enable it to properly model the underlying behavior. Neural learning is considered successful if the system can perform well on test data on which the system has not been trained (Mehrotra, 1997: 85). Some researchers have proposed for choosing the number of

hidden node, for example, Lawrence (1998) suggested that a best estimation for the number of hidden neurons is half of the sum of inputs and outputs. Related to training data size Lawrence's proposed as follows:

$$\frac{N}{10} - i - o \leq h \leq \frac{N}{2} - i - o \qquad (1.3)$$

Where *N* is the number of training data, *i* is the number of input neurons, *o* is the number of output neurons and *h* is the number of hidden neurons.

Type of learning

"Learning" is the process of calculating the weights of neurons in a network (Fausett, 1994:15). There are two main types of learning in a network: "supervised'' and "unsupervised''. In supervised learning both the inputs and the outputs are provided. The neural network system receives the output, computes the error, which are the difference between computed and actual output. The weights are adjusts according to the error. Supervised learning includes: back-propagation (BP), radial basic function (RBF), probabilistic neural network (PNN), generalized regression neural network (GRNN). On the other hand, the actual outputs are not known in unsupervised learning. Inputs are available to the network, and the weights cannot be adjusted based on the actual output. This type of learning is commonly used for pattern recognition problems and clustering. Kohonen's self-organizing network is based on unsupervised learning.

The first step in the development of a neural network model is to select an appropriate neural network paradigm by matching the application requirements with the paradigm capabilities. The application in this thesis required a powerful feature of both supervised and unsupervised learning. Supervised learning was used since the network would be trained that included the result. For this application, a back-propagation neural network was used to predictive crime rate in the study

region (Section 6. 9). Unsupervised SOM learning is used for developing hierarchal neural network for reducing the dimension of the input data set.

Learning Rules

Procedures for modifying the weights on the connection link in a neural net. The learning rule is the mathematical equation that determines the increment or decrement by which weights of a processing element change during the learning phase. There are four more commonly used: Delta rule, Delta- Bar- Delta, Extended Delta- Bar- Delta rules and Kohonen's rule (Fausett, 1994: 429).

Delta rule

The error in the output layer is computed as the difference between the computed and the actual output of a neuron. This error is transformed by the derivative of the transfer function, and is back- propagated to prior layers. This process of back-propagating the error continues until the first layer is reached.

Delta – Bar – Delta (DBD)

This learning rule was developed in order to improve the convergence speed of Delta rule. This is each connection in the network has its own learning rate and change those rates continuously as the learning progresses.

Extended- Delta- Bar- Delta (EDBD)

Extended Delta- Bar- Delta, developed by Minal and Williams, is an extension of DBD which introduces a momentum term for each connection. This varies with time (Fausett, 1994: 427).

Neural network experiments in this thesis used the learning rule extended- Delta-Bar. Experimentally work best (Section 6.9).

Kohonen's rule

Since Kohonen's network does not depend on known outputs, the weights are adjusted using the input into the neuron *i:*

$$\Delta W_{ji} = \eta (input_i - w_{ji}) \quad (1.4)$$

Where $input_i$ is the input that neuron *i* receives from the external environment. η is learning coefficient and Δw_{ji} is the adjustment of the connection weight from neuron *j* to neuron *i* .

Possible Applications

Artificial Neural Networks (ANNs) "provide a range of powerful techniques for solving problems in pattern recognition, data analysis and control. They have several notable features including high processing speeds and an ability to learn the solution to a problem from a set of examples"(Bishop, 1994:1803). The following examples represent only a sampling of areas in which ANN have been successful. ANNs were found to be more efficient in solving complex and non linear optimization problems than statistical techniques (Hopgood, 1993), having been successfully applied in clustering and visualization of high dimensional data (Vesanto 2000, Hollman 1999 and Zhang 1993).

ANNs are being used by many technical analysts to make predictions about stock prices based upon a large number of factors such as past performance of other stocks and various economic indicators (Egli, 2003). An ANN has been shown to learn to predict future values of the time series from its past values, which is highly relevant in crime prediction (Corcoran & Wilson 2003, Olligschaleger, 1997). ANN research in medicine includes modelling parts of the human body and recognising diseases from various scans (Thomas, 2002 and Joseph, 2001).

The utility of ANNs, such as Back-Propagation (BP) and Self-Organizing Map (SOM) are used in this thesis for developing hierarchical neural networks, with the aim of developing predictive crime models.

1.1.4 Regression Analysis

Regression analysis is "a statistical methodology used for explaining or modelling the relationship between a dependent variable and one or more predictors (explanatory variables)" (David, 1998). It can be of different types, e.g. linear regression, non linear, multiple linear regression and multivariate multiple regression, when there is more than one dependent variable. Regression analysis has several possible objectives including; prediction of future observations, assessment the relationship between dependent and explanatory variable and general description of the structure of data.

A multiple linear regression model (MLR) implemented in the thesis to predict crime model in the study area. A general mathematical form of MLR shown in the equation below;

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \quad ; \qquad \varepsilon \sim N[0, \sigma^2] \tag{1.5}$$

The regression parameters β_i are estimated using the least squares method which uses the criterion that the solution must give the smallest possible sum of squared deviations of the observed Y_i from the estimates of their true means provided by the solution. The error term ε is assumed to be normally distributed with a mean of zero and a variance of σ^2. In particular, the estimated regression equation is

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_i X_i \tag{1.6}$$

and the observed residual ε is $Y - \hat{Y}$.

This methodology is widely used in business, the social and behavioural sciences, the biological science and many other disciplines. MLR is a common and useful tool for model building and has wide applicability in prediction within a variety of areas. For example Bolzan et al. (2008) used MLR method to predict hatchability in an artificial incubation process. The cost models developed by Margaret et al. (2002) used MLR in order to identify those variables that demonstrated a strong linear relationship with the cost. It has also been used by Ameen et al.(2003) to quantify a claim for increased costs in construction engineering. Most published medical research uses regression analysis in predicting the outcome of patients with a variety of diseases. This methodology developed rapidly with the increasing power of computers. There are many computer software packages that can be used to perform regression analysis. For instance SAS, SPSS, Answer Tree and Minitab.

The statistical technique, multiple linear regression (MLR) analysis is used in this thesis first for building a predictive crime model, identify potentially significant predictive variables among characteristics of burgled households and the level of their contribution in the performance of the model and predicting of future crime rate in the study region. The burglary rates in this analysis are expressed as the burglary incidence per number of households in each polygon (census wards). The models are applied across a number of geographical high burglary incidence concentration 'hotspots', a number of geographical space (parcel) and across the clusters. Which were identified by a new methodology and GIS was used to visualize their locations in the study region. Clustering leads to increase the predictive accuracy of a crime model by identify the problem associated to the characteristic of the people within their location (section 6.8.2). Secondly, MLR combined with Neural Networks with the aim of developing a new hierarchical neural Network approaches to generate a more reliable prediction.

1.1.5 Ecology of crime

Crimes are a human phenomena and the assumption is that their occurrence in a spatial and spatio-temporal framework. Crime is not constant: it varies from person to person, and it varies for each individual person a cross time and space. It is varies as the awareness of opportunities to commit a crime vary. Targets are not constant. The distributions of targets vary in time and space (Ronald, 1993: 266). The early ecology of crime studies started with Clifford Shaw's seminal study of delinquency in Chicago (1929). This dealt with concentrations of crime in central business districts. Concentration of delinquents' residences varied inversely in proportion to distance from the city center. Ecological theories attempt to explain individual actions in general features of the social structure in which an individual is embedded. The social ecology is focused to "place- based theories" of crime and routine activities (Gorr and Anselin, 2000: 218). Place-based theories where the objective is to derive an understanding of mechanisms upon individual actions. Crime is a social construct, and therefore some understanding of criminological theory, for example, Routine activities, Pattern theory, rational choice perspective theory and Awareness theory is important in their interpretation and to understand patterns of crime.

Routine activities theory

Routine Activity Theory (RAT) is one of the main theories of "environmental criminology". It was developed by criminologists Cohen and Felson (1979). The theory states that criminal events results from motivated offenders, suitable targets, an absence of guardian capable of preventing the criminal act. A converging of offenders and victims occurs non-randomly in time and space. The three elements must be present at the same time and in the same space when a crime occurs. RAT introduces an important tool in crime analysis, the crime triangle. The crime triangle diagram is illustrated in Figure 1.4. Cohen and Olligschlaeger (1993) applied routine activity theory to illicit drug markets, recognizing the tacit coordination required of dealers and buyers for drug transactions (John, 1995: 6).

Figure 1.4 Diagrammatic representation of routine activity theory
(http://www.crimereduction.homeoffice.gov.uk/skills/skills08.htm)

Pattern theory

This theory integrates crime within a geographic context that demonstrates how the environments people live in and pass through influence criminality. It describes the distribution of offenders, targets, handlers, guardians, and managers over time and place. This theory has three main concepts: nodes, paths, and edges. Nodes refer to where people travel to and from such as home, school and entertainment area and the paths among them. Edges, refers to the boundaries of areas where people live, work or shop. The theory specifically focuses on places and the lack of social control or other measures of guardianship that are informal needed to control crime (Ronald, 1993: 284).

Rational choice theory

Developed by Ronald (1979). This theory focuses of primarily on properties of the offender, the rational choice perspective explains the conditions that are needed for specific crimes to occur, and thus emphasizes the role of crime opportunities in crime causation. The theory helped criminologists to focus on the particulars of criminal acts (Alex, 2002: 87).

Awareness Theory

Brantingham and Brantingham (1991) suggested that crime has four dimensions: victim, offender, geo- temporal and legal. Concentrating on the spatial element of crime is significant to understand the behavior of offenders. A crime's space can be chosen either on purpose or accidentally by either the victim or the offender according to their life styles. Several things have an effect on the crime rate of an area. For example, what type of people live in particular space and what type of security is available (Ronald, 1993: 269).

This thesis presents building of hybrid predictive models for crimes based on real data. The spatial distributions of residential burglaries are chosen as the foci of this analysis. The model construct with 28 potential explanatory variables among characteristics of burgled households, for the purpose of estimating the relationship between burglary rate and characteristics of burgled households. These include Resident Population, Occupation, Qualifications, Socio-Economic, Household composition and Household spaces.

1.2 Research Objectives

The aim of developing a hybrid modelling approach utilizing some relevant principles of Statistics, Geographical Information Systems (GIS), Neural Networks and in general, Information Technology for the analysis of observed data. The methodology is applied to real data on crime. The objectives of this research were to:

- Develop a new algorithm based on statistical theory for identifying clusters within spatial data;
- Generate artificial datasets, based on established practice, for use as a proof of concept for a general purpose algorithm for detecting clusters within spatial datasets, which have been used to evaluate the effectiveness of the developed cluster determination algorithm;

- Acquire real world spatial datasets for:
    - testing the developed algorithm and identification of ''hotspots'' in the study region;
    - generating predictive models.

- Utilize GIS to accommodate the new cluster detection technique (SCS) in terms of a predictive crime model for:
    - mapping;
    - display distribution of crime and the corresponding population in a study region;
    - visualize the location of obtained clusters which was specified by clustering algorithm SCS;
    - display distribution of burglary incidence concentration 'hotspot' which was identified by clustering algorithm SCS;
    - identify the total number of cases within polygons (census wards);
    - integrating information from a variety of sources such as crime data, population data and census data associated with the observed data.

- Building of hybrid predictive models for crimes based on real data (crime).
    - Develop a predictive crime model based on statistical methodology;
    - Develop a new a hierarchical neural network methodology based on statistical methodology and two proposed ANN learning algorithms, unsupervised Self-Organizing Map and supervised back-propagation.

1.3 Thesis outline

The remainder of the thesis is divided into five chapters. Chapter two discusses the relevant issues to the topics under investigation. Chapter three describes the development of a new cluster detection methodology. Chapter four presents the new algorithm applied to real crime dataset and its rotation. Chapter five introduces

construction and interpretation of maps using crime data in the area under study and data modelling. Chapter six describes the application of the two methodologies: regression analysis and artificial neural networks (ANNs) for the purpose of predictive crime modelling. Chapter seven presents conclusions and some potential for future research.

1.4 Software

The Matlab language is utilized for implementing the clustering Algorithm with significance. Matlab is a high level programming language for technical computing.

The utility of ArcGIS Desktop 9.3 software is employed to analysis and display of geographic data and integrate information from a variety of sources associated with the observed data.

Finally, Neuralware, 2001 software and statistics package Minitab version 15 are utilized to develop a predictive crime model.

2 LITERATURE REVIEW

This chapter discusses the relevant literature pertaining to the field of clustering, statistical theory in clustering and crime prediction. Blends science from three distinct research domains is provided.

2.1 Introduction

This section provides a review of literature relevant when producing a general purpose, hybrid modelling methodology. In addition, the reader is provided with an overview of literature relevant to the case study that will be used to test the proposed methodology. As such, the following discusses the relevant literature pertaining to the field of clustering, statistical theory in clustering and crime prediction. The significance of this and the following related wide-ranging study, which blends science from three distinct research domains, should not be lost on the reader.

2.1.1 Cluster detection

Over the recent past and with the increasing power of computers, research on cluster detection and analysis has grown quickly in many disciplines simultaneously and often independently of each other. Amongst these disciplines, a few stand out as being especially important for the development of cluster analysis.

Several studies utilize information from K-means to derive a new representation. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. Other studies employ specific properties of statistical distribution. Determining clusters depend more on the researcher's goals than on a theory. Selected studies are outlined below:

K-Means

K-Means (Mac Queen, 1967) is one of the simplest unsupervised learning algorithms. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The algorithm is based on specifying the desired number of clusters, *k*. Randomly assign *k* points to be initial cluster centers. The clustering is done by minimizing the sum of squares of distances between data and the corresponding cluster centroids. The center of each cluster is then re-estimated. This process is repeated until convergence is achieved, such that the convergence criterion has been met. Its disadvantages are that it works on the assumption that the initial centers are provided. Thus the results depend on the suggested value *k*; applicable only when the mean is defined (numerical data); it does not do well with overlapping clusters (Daniel, 2005: 90). There are a lot of applications of the K-means clustering, such as in classification analysis, Artificial intelligent, image processing and unsupervised learning of neural network. Clustering algorithms have been applied in molecular biology for gene expression data analysis. Lus et al.(2004) was used K-means to identify genes function, by partitioned genes into groups based on the similarity between their expression profiles. They found that the enrichment of genes of similar function increased within the cluster. Clustering algorithm has been widely used in computer vision such as image segmentation and database organization. The purpose of clustering is to group images whose feature vectors are similar by similarity judgment standard; meanwhile to separate the dissimilar images. Liu and Yu (2009) used K-means clustering algorithm to image retrieval system. Image retrieval algorithms, retrieval is according to feature similarities with respect to the query, ignoring the similarities among images in database. They addressed this problem by introducing a graph-theoretic approach for image retrieval post-processing step by finding image similarity clustering to reduce the images retrieving space. The results of experiments on the testing images showed that the efficiency and effectiveness of K-means algorithm in analysing image clustering.

GAM

The Geographical Analysis Machine (GAM) is an early attempt to automate cluster detection. This was created and subsequently refined by Openshaw (1987, 1988 & 1991). GAM identifies clusters based on the spatial distribution of incidents combined with the background population. A point on a map randomly defines and a circle with a random radius draw a round that point. The numbers of point are count with each random circle and compare the result with an expected value based on an assumption a random generating process. Then the circle draws on the map if it contains a significantly higher number of observations than expected. The procedure is repeated by select a new point on the map. GAM addressed the problem of purely statistical analyses used previously which did not handle the special characteristics of spatial data. One of the greatest assets of GAM is the exhaustive nature of its search, which required no previous knowledge of the data or the study region. It returns information of significance for mapping. The major disadvantage of GAM is that on any large dataset, the time required for the exhaustive search makes the use of GAM nearly impossible. This is because the time complexity of GAM is: O (l* r*n), where l is the number of locations to be tested, which is proportional to the real extent of the dataset and the distance between neighboring points on the grid, r is the number of radii to be tested at each location, and n is the number of points in the dataset. GAM spends much of its time examining regions in which there may be no relevant cases or not even any background population at all. As such, GAM was criticised for this incredibly computer intensive process, which was particularly significant if the study area was large (Fotheringham and Zhan, 1996). GAM is widely applicable to many different types of data; for example, disease data, crime data and traffic accidents. The GAM method was developed to analyse child leukaemia data for Northern England.

Satscan

Satscan (Kulldorff, 2005) used circle to scan whole area to find hotspot. A moving circle is centered on each point and then the radius of the circle is expanded based

on a user assumption. It only returns circles such that no circle's centriod was contained in another circle, thus limiting the amount of overlap between adjacent circles. The circle with the maximum likelihood is the most likely cluster. The process of moving circle in Satscan lets the running time grow quadratically because the time complexity of Satscan is: O (n*l*m), where m is the number of Monte carol tests and n=l, the default circle locations are the data points. The choice of region size will impact upon the success of the technique. If the regions chosen are large then the precision of case and population positioning becomes generalised and may lead to reduced accuracy of results. In contrast to GAM, Satscan is less intensive, testing fewer points. Satscan has been applied in a very wide range of application areas, such as medical, criminology and demography. Examine geographic variation in prostate cancer grade and stage was the subject of a study by Klasssen et al. (2005). The data (20928 Maryland men) was derived from cancer registry during 1992- 1997. Satscan was used to test for significant local clusters. Four statistically significant clusters identified of high and low rates of stage at diagnosis and higher histological grade of tumour. Chiehwen et al. (2004) were employed Satscan to examine the geographic variations in breast cancer mortality in Texas females according to three predominant racial groups (non- Hispanic white, Black, and Hispanic females) over a twelve- year period. Spatiotemporal variations in breast cancer mortality affected racial groups at varying levels. There was neither evidence of hotspot clusters nor persistent spatiotemporal trends of excess mortality into the present decade. Non- Hispanic whites in the Gulf Coast and Hispanics in West Texas carried the highest burden of mortality, as evidenced by spatial concentration and temporal persistence.

CLAP

The cluster Location Analysis Procedure (CLAP) (Jarvis, 2006) technique for cluster detection is capable of analysing a point dataset and determining the significance of the quantity of cases found, when compared with the quantity expected. CLAP analyses each case in turn by counting the number of other incidents within a

certain radius, which itself is progressively increased while a high ratio of incidence is returned. The algorithm gives a reliable output, but as length, running times most especially on large datasets and large memory is needed. This is also the case for GAM and Satscan algorithm. Jarvis (2006) was employed this technique to analyse Leukaemia dataset and identified clusters of high crime incidence in Cardiff by Corcon (2003).

SOM

Artificial Neural Networks (ANNs) are extensively used for both classification and clustering. Jain and Mao (1996) stated that competitive neural networks are often used to cluster input data. In competitive learning, similar patterns are grouped by the network and represented by a single unit (neuron). This grouping is performed automatically based on data correlations. Well-know examples of ANN used for clustering include learning vector quantization (LVQ) and Self-Organizing Map (SOM). The goal of SOM is to convert a complex high-dimensional input signal into a simpler low–dimensional discrete map. Thus, SOM are appropriate for cluster analysis where underlying hidden patterns among records and fields are sought (Vesanto & Alhoniemi, 2000). SOMs structure the output nodes into clusters of nodes, where nodes in closer proximity are generally more similar to each other than to other nodes that are farther apart. SOM are based on competitive learning, where the output nodes compete among themselves to be the winning node, the only node to be activated by a particular input observation. However, the number of clusters that can be produced by this methodology is limited within a fixed number of output nodes. SOM for clustering has been established for a number of researchers such as Corcoran (2003) used SOM for clustering Temporal, spatial crime datasets. Andrew (2004) used SOM for clustering components of the din cyst class.

2.1.2 Statistical theory in clustering

Many methods have been proposed for estimating the number of clusters, testing the number of components and also alternative methods for selecting initial seeds (centroids). Selected studies are outlined below:

One of the most widely used clustering techniques is the K-means algorithm. Solutions obtained from this technique are dependent on the initialization of cluster centers, which in turn highly influence final results. There are many algorithms which have been proposed to compute initial cluster centers for K-means clustering (Baradley, 1998; Likas, 2003; Deelers, 2007). These study the whole feature space to select *k* initial samples. Deelers et al. (2007) proposed the algorithm based on the data partitioning along the data axis, either the x-axis or y-axis. Centroid distance of cluster *C* computed by:

$$centroidDi \quad st = \frac{1}{n}\sum_{i=1}^{n} dsum_{i} \tag{2.1}$$

Where $dsum_{i}$ is the summation of distances between the adjacent data. The partition boundary is the plane perpendicular to the principal axis and passes through a point *m* whose $dsum_i$ approximately equals to centroidDist. Cluster C divided into two clusters by this perpendicular plane. Clusters are partitioned one at a time until the number of clusters equals the predefined number of clusters, *K*. The centers of the *K* clusters become the initial cluster centers for *K*-means. However, new algorithm, SCS has employed specific properties of statistical distributions to estimate the position of the centers. This approach is much more logical and efficient. The experimental results show that the new algorithm performs better than random initialization and can reduce running time of the algorithm.

Available methods for estimation of the number of mixture components (*k*) include bootstrapping[1] the likelihood ratio test statistic and optimizing a variety of validity

[1] Bootstrapping is a statistical re-sampling method employed to estimate a population parameter.

function. In the case of finite mixtures of normal distributions Miloslavsky and Mark (2003) estimate the number of components *K,* by minimizing the distance $D(\hat{f}_k, f)$ between true density f and its projection, with respect to this distance, on the space spanned by the mixture model with *K* components, $\hat{f}_k$ over *k=1, 2, …, K*, $D(\hat{f}_k, f)$ =0 for $k \geq k^*$, k^* components of a true model. The distances considered are Kullback-Leibler and use cross validation to estimate these distances. This can be cumbersome and complicated in applications, while the new approach SCS is much simpler and it is easier to predict the cluster sizes. Cluster sizes are determined according to the requirements of confidence intervals.

Tibshirani et al. (2001) proposed a method for estimating the number of clusters in a dataset called 'gap statistic'. The idea of this method supposes that the data $\{x_{ij}\}$ clustered into k clusters $C_1, C_2, \ldots, C_k$ *with i=1,2,…,n, j=1,2,…,p*, consists of p features measured on n independent observations, with C_r denoting the indices of observations in cluster *r*, and $n_r = |C_r|$. Let $d_{ii'}$ denote the distance between observations *i* and i', $D_r = \sum_{i,i' \in c_r} d_{ii'}$. The estimated value *k* has been the optimal number of clusters for which log (w_k) falls the farthest below the reference curve of the data distribution, where

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \qquad (2.2)$$

Hence, Tibshirani et al. (2001) defined

$$Gap_n(k) = E_n^* \{\log(w_k)\} - \log(w_k) \qquad (2.3)$$

Where E_n^* denotes expectation under a sample of size *n* from the reference distribution, the estimate $\hat{k}$ will be the value maximizing $Gap_n(k)$ after taking the sampling distribution into account. This approach used K-means, as the first clustering algorithm. Therefore, they increased the complexity of their algorithm. In

the new algorithm (SCS) the multi-modality of the projected data is used as an indicator for the number of distributions in the mixture. Thus, no cluster numbers are predetermined.

Their main weaknesses (Deelers 2007; Miloslavsky and Mark 2003; Tibshirani 2001) causing their unreliability can be summarised in the following:

- Estimating the number of clusters;
- Testing the number of components; and
- Selecting initial seeds (centroids).

Consequently, a new cluster detection methodology has been developed taking advantage of new advances in Information Technology, Geographic Information Systems, computer technology and the principles of basic descriptive statistics. The approach is a natural combination between advancements in data mining and classical statistics. It is effective in terms of cluster detection, using a hybrid of statistical methods and properties of probability rather than distance to associate data with clusters. The approach relies on a mixture of a finite number of marginal probability distributions projected on principle axes from which bivariate (multivariate) probability distributions are reconstructed to best represent clustered data. The multi-modality of the projected data is used as an indicator for the number of distributions in the mixture. The parameters of the mixture model are estimated by the method of maximum likelihood. Radii of clusters are computed according to the requirements of confidence intervals.

2.1.3 Crime prediction

The purpose of police crime prediction is to directly support crime prevention and law enforcement. It is an exciting new area, which brings together the disciplines of statistics, machine learning, artificial intelligence, criminology, and psychology and database technology.

The prediction of crime has seen an increase of research attention over the last decades and is more widely practiced by police. This is due to fact that Geographical Information systems (GIS) had become an important tool for police agencies. The mapping of crimes and the identification of hot spots has become regular practice. The 'criminality profile' of places was established based on theories like routine activities, the ecology of crime and hot spots.

There have been numerous studies using varity of related methodology such as statistic, neural networks for the spatial and temporal analysis of crime; the following outlines some of the recent ones. Law enforcement agencies have a continuing need to predict time and locations of crimes. Liu and Brown (2003) showed that predicted crime locations for next week based on data from the previous week. They suggest that the likelihood of a criminal incident at a specified location is based on past incidents of the same type and independent spatial attributes or features. They compare two prediction models for hot spots that relate the features in an area to the predicted occurrence of crime through the preference structure of criminals, and conclude that the model performs significantly better with the extra features.

Gorr and Olligschlaeger (2003) study monthly crime data over the period 1991-1998. Using holdout samples the subset of actual time and rolling horizon experimental design, to compare the forecasting accuracy of model fit to past data. They contrast the forecast accuracy of univariate time series models with naïve methods. This method used time series data points themselves as forecasts. They find the classical seasonal patterns of increased property crime levels late in the year and increased aggression crimes in summer due to increased social interactions, cold weather might increase burglary, and robbery crimes due to seasonal economic pressures or unemployment. Felson and Poulsen (2003) studies show that crime varies greatly by hour of day more than by any other variable and criminal activity

between 5: AM and 4:49 AM the next morning. This suggests that a prediction for specific time periods might be valuable for police planning.

Several studies show risk factors of crime. Roncek and Maier (1991) found relationship between levels of crime and the number of taverns and lounges located. Drug hotspots tended to be in areas with poverty and low family cohesion Gorr and Olligschlaeger (1993).

Bowers et al. (2004) investigated the relationship between area type, housing type, level of victimization and repeat victimization. The results have demonstrated that the influences of area and housing type being burgled interact. For instance a detached house located in deprived area is at over seven times the risk of a detached house in affluent areas. Analysis of the relationships between the spatial patterns of residential burglaries and the socio economic characteristics of neighborhoods in London has been examined by Malczewski, et al. (2005) using geographically weighted regression. The result shows that there were significant spatial variations in the relationships between the relative risks of residential burglaries and the average value of dwelling and the percentage of multifamily housing. Edmark (2005) studied the effects of unemployment on property crime rates. According to the theory of economics of crime, increased unemployment rates lead to higher property crime rates. A high crime rate leads to unemployment because new firms do not want to settle in a criminal area and existing businesses leave. It might also be the case that people who have once been in prison have difficulties finding a job and for this reason contribute to a high unemployment rate.

ANN's are presented as one technique that offers minimal user interaction in addition to dynamic adaptability, and thus a potential operational forecasting solution. One of the earliest was that of Olligschlaeger (1997), who employed back-propagation to predict areas where future drug markets will emerge. In more recent work, Olligschlaege and Gorr (2001) found that ANNs outperform multiple

regression leading indicator models when the set of leading indicators is rich and numerous.

Corcoran, et al. (2003) used hotspots (spatial clusters of crime) for forecasting. The Gamma Test (GT) was applied to each cluster to assess suitability for predictive modelling. ANN and comparative linear regression forecasting models were constructed using the GT, and compared to a "random walk" model. For crime analysis software, Oatley and Ewart (2003) used a Kohonen neural network for matching crimes against the offender list and the Bayesian belief network for prediction of re-victimisation.

Craglia, et al. (2001) reported the strengths of GIS based spatial analysis with census (socioeconomic) data for modelling high-intensity urban crime area. Three police force areas in England and Wales were used to develop the model. These areas that raise special policing problems, such as sometimes found violent forms of crime within them, resident population defect to co-operate with the police. The model suggests that high-intensity crime areas are characterized by populations that are deprived and live at high density and have higher levels of population turnover. This is done through a statistical analysis (regression) which uses data from the census. The spatial datasets within the GIS was used to integrate data on the boundaries of the high-intensity area with socio-economic data. Ratcliffe (2001) derived his study over 14,000 burglaries over two years for separately examining the spatial and temporal patterns of residential and non residential burglary. The study showed that the highest probability for residential burglaries was between 8am and about 6pm, the period that most people were at work. The residential burglary levels were lower over the weekend and overnight. For the spatial analysis of residential burglary the researchers examined the Canberra region. 'Hotspots' include the more established suburbs of the inner-north of the city and the inner south-east. The housing characteristics of the residential burglary hotspots vary considerably across the city.

This thesis presented the developing a hybrid predictive models for crime based on real data (burglary incidence). Both regression methodology and neural networks have been used for predictive crime modelling. Historical data with background population and census datasets are used for predictive crime modelling. The obtained models based on the observed data in the study region and which presented in chapter six are reasonable.

2.2 Summary

This chapter describes the relevant literature pertaining to the field of clustering, statistical theory in clustering and crime prediction. Cluster detection has led to the development of several techniques; the results indicate that different techniques often have different aims. In view of many studies that are described, some researchers prefer the K-means idea and others employ specific properties of statistical distribution; determining clusters depend more on the researcher's goals than on a theory. Their main weaknesses causing their unreliability can be estimating the number of clusters, testing the number of components, selecting initial seeds (centroids), running time and memory requirements. Consequently, a new cluster detection methodology has been developed in this thesis based on knowledge drawn from both statistical and computing domains.

Law enforcement agencies have a continuing need to predict time and locations of crimes. Predictive crime models in this thesis are created using several existing methodology, such as regression analysis, Geographical Information System, Neural Networks and a new clustering algorithm. Spatial distributions of residential burglaries are chosen as the foci of this analysis.

The next chapter introduces a new cluster detection methodology by taking advantage of the new advances in Information Technology, Geographic Information Systems, computer technology and the principles of basic descriptive statistics.

3 Spatial Clustering with Significance

This chapter describes the development of a new cluster detection methodology called Salar's Clustering with Significance (SCS). It relies on knowledge drawn from both a hybrid of statistical methods and computing domains. The algorithm for determining both cluster centers and the existence of the clusters themselves is given. The methodology has been tested on simulated datasets with promising results. In addition the chapter introduces some of popular distance measures to compute distances (similarities) between two clusters.

3.1 Introduction

Cluster detection is mainly an exploratory process which aims to sort different objects into groups or clusters. Clusters in this thesis are regions of high density separated by region of lower density.

Over the recent past and with the increasing power of computers, research on cluster detection and analysis has grown quickly in many disciplines, simultaneously and often independently of each other. Amongst these disciplines, a few stand out as being especially important for the development of cluster analysis.

Cluster detection techniques can be applied to datasets in order to reveal information relating to the spatial distribution of cases. As was pointed out in Section 2.1, there are several studies for clustering that have concentrated on computational or statistical aspects of cluster detection. Each clustering algorithm has its own strengths and weaknesses. Their weaknesses related to the following cases: selecting initial seeds; estimating the number of clusters; testing the number of components; running time; and memory requirements. For example, Deelers (2007), Miloslavsky and Mark (2003), and Tibshirani (2001). Consequently, during the research documented in this thesis, a new cluster detection methodology has been developed that uses a hybrid of statistical methods and properties of probability distributions rather than distance to associate data with clusters.

3.2 Background

The following section provides some definitions and background relevant to development of a new cluster detection methodology SCS. For determining the center of clusters, the existence of the clusters themselves and computing the distance between clusters.

3.2.1 Normal Distribution

Normal distribution (Gaussian) is one of the most commonly used distributions in statistics. It can be used to model many real-world phenomenon. The normal distribution is completely determined by its parameters, which are its expected value μ and its variance σ^2. It has bell-shaped curve; symmetrical and unimodal; the mean, the median and mode all coincide and correspond to the highest point on the curve. The tails of the curve extend to infinity into the right and left. That it is possible theoretically to obtain values at any distance from the mean. The standard deviation determines how flat and wide the curve is (see Figure 3.1). In the normal distribution for instance, 95 percent of the data values will be within two standard deviation of the mean (see Figure 3.2).

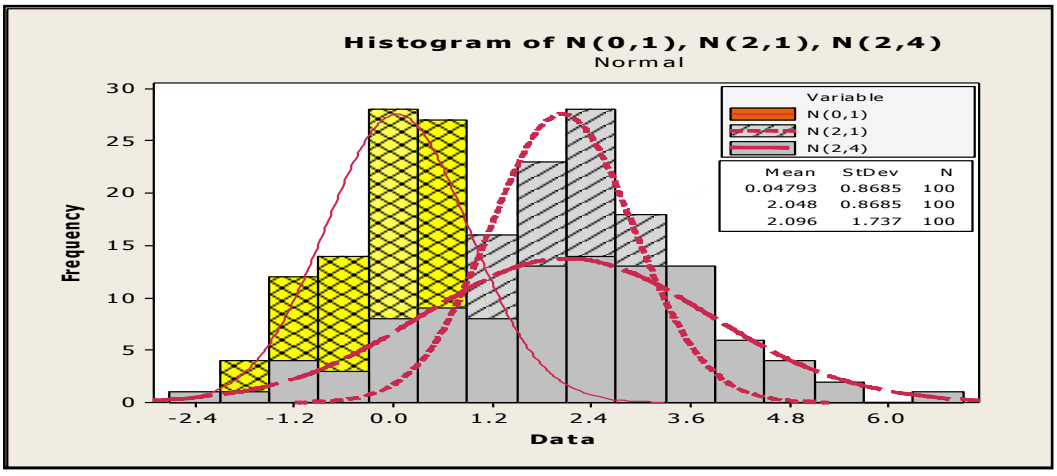


Figure 3.1: Illustrates the property of the normal distribution; the mean of the distribution determines the location of the center of the graph and the standard deviation determines the spread of the graph.

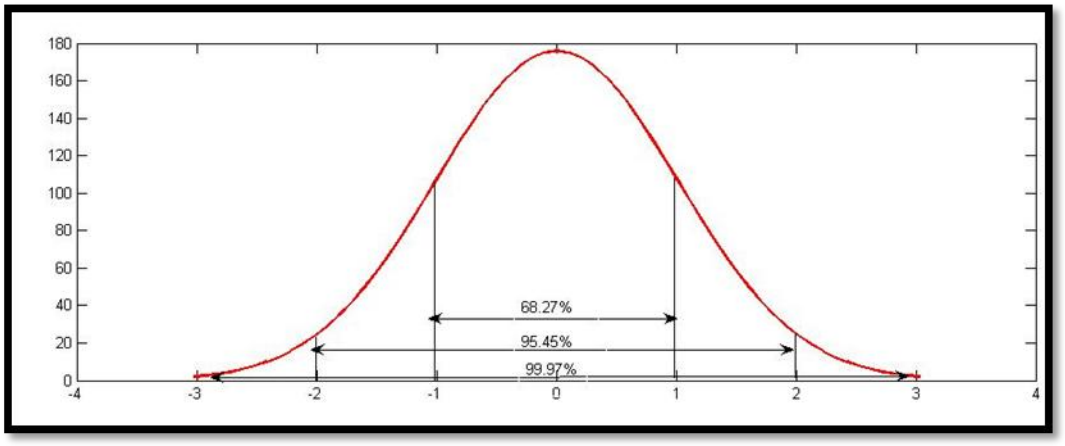


Figure 3.2: Areas under the curve for any Normal distribution

3.2.2 Generating Random Variables

Many of the methods in computational statistics require the ability to generate random variables from known probability distributions. The main MATLAB program has a function called randn that will generate numbers from the standard normal distribution, and can obtain a normal random variable *X* with mean μ and variance σ^2 by means of a transformation. Letting *Z* represent a standard normal random variable (possibly generated from randn), the desired *X* can be found from the relationship (3.1) (Martinez, 2002).

$$X = Z * \sigma + \mu \qquad (3.1)$$

Table 3.1 A Sample of the artificial data

N(0,1)	N(2,1)	N(2,4)	N(0,5)	N(5,7)
0.69	2.69	3.38	-2.1628	8.5744
0.8156	2.8156	3.6312	-8.3279	6.4734
0.7119	2.7119	3.4238	0.6267	-4.2007
1.2902	3.2902	4.5805	1.4384	5.4178
0.6686	2.6686	3.3372	5.9546	-1.745
1.1908	3.1908	4.3817	6.3489	-2.4829

3.2.3 Confidence interval

A confidence interval is a range of values within which the true value of the parameter lies with some specified probability. Confidence intervals are determined based on the stated hypotheses. In the case of equality of the mean to a specific number, a two-sided confidence interval will be optimal. For example, a 95% confidence interval containing the true mean when the variance of the random variable is assumed unknown would be:

$$(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}) \quad (3.2)$$

Where α represents the stated level of significance, *s* is the sample standard deviation, *n* is the sample size, $\bar{x}$ is the sample mean and t stands for a point on the t-distribution scale. This is the same to say that wrong decisions would be accepted as far as their probability of occurrence does not exceed α (=0.05 or 5% level of significance).

3.2.4 Histograms

Histograms are plots of sampling frequency distributions attempting to graphically represent estimates of the frequency distributions of the populations that sampled datasets (sample) have been drawn from. Histograms are used to:

1. Summarize data to understand general characteristics of distributions such as shape, spread or location;
2. Suggest possible probabilistic models;
3. Determine unusual behavior.

3.2.5 Maximum likelihood estimator

Maximum likelihood is a popular statistical method for estimation of model parameters. Let $x_1, \ldots x_n$ be a random sample from any probability density function f (x_i,θ) that depends on an unknown parameter θ. The likelihood function

L (θ) = f(x_1, . . . x_n, θ) = TT f (x_i, θ) is the joint probability density function. The function $\hat{\theta}$ = g (x_1, . . . x_n) that maximize L (θ) with respect to θ is the maximum likelihood estimator of θ . The maximum likelihood estimator (mle) can often be found by setting the first derivative equal to zero: $\frac{dL(\theta)}{d\theta} = 0$ (Krzanowski, 1998).

In the study presented in this thesis maximum likelihood was used to estimate the position of the center of the clusters. For this purpose the function mle, provide by MATLAB statistics toolbox was utilized.

3.2.6 Similarity Measures

Any clustering algorithm requires some type of measure to assess the similarity between pairs of observations or clusters. Computing distances (similarities) between two clusters can be performed in different ways, the popular methods are:

Single Linkage (Nearest Neighbor)

In single linkage, the distance between clusters is defined as the distance between the two closest objects in two clusters. The distance d(*r,s*) between cluster *r* and *s* is computed as d(*r,s*)=Min{ d(*i,j*): Where object *i* is in cluster ***r*** and object *j* is in cluster *s*}**.** Here the distance between every possible object pair (*i,j*) is computed. The minimum value of these distances is said to be the distance between cluster ***r*** and ***s*****.** At each stage of hierarchical clustering, the clusters ***r*** and ***s***, for which d(*r, s*) is minimum, are merged. This measure of inter-group distance is illustrated in Figure 3.3.

Complete linkage (Farthest Neighbor)

Complete linkage is the opposite of single linkage. The distance between clusters is defined as the maximum distance between pairs of objects in two clusters. The distance d(*r,s*)between cluster *r* and *s* is computed as d(*r,s*)=Max { d(*i,j*): Where

object *i* is in cluster ***r*** and object *j* is in cluster ***s*** }. Here the distance between every possible object pair (*i, j*) is computed. The maximum value of these distances is said to be the distance between cluster ***r*** and ***s*****.** This measure is illustrated in Figure3.3.

Average linkage

In average linkage, the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each cluster. In the average linkage method, d(*r, s*) is computed as d(*r,s*)= $\frac{1}{n_r . n_s} \sum_{j \in s} \sum_{i \in r} d(i, j)$ where n_r and n_s are the sizes of the clusters ***r*** and ***s*** respectively**.**

This measure is illustrated in Figure 3.3.

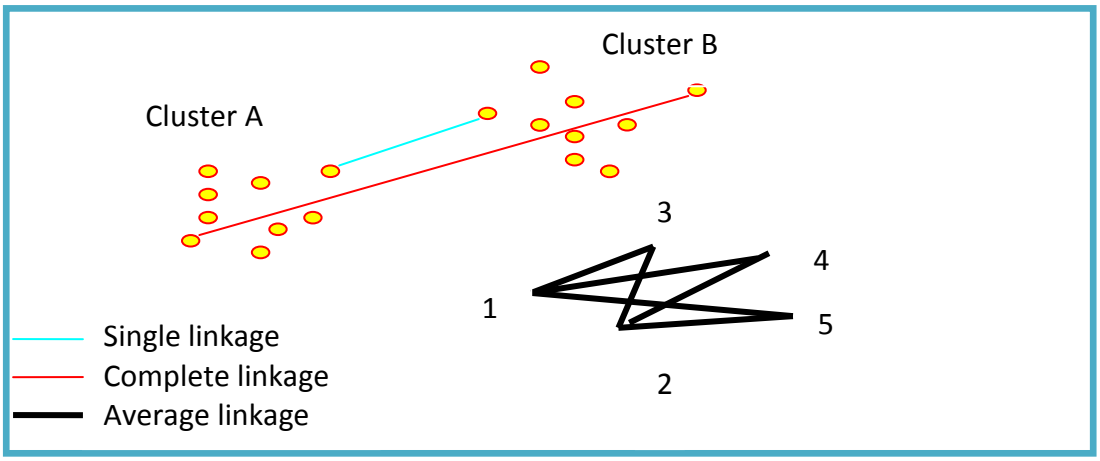


Figure 3.3: Examples of three inter-cluster distance measures: single, complete and average (Everitt, 2001).

In the study presented in this thesis Kullback-Leibler divergence (KL) was used to compute distances between two clusters rather than Euclidean distance.

The Kullback-Leibler Distance

Kullback and Leibler(1951) studied a measure of information from the statistical viewpoint; this measure involved two probability distributions associated with the

same experiment. The Kulback-Leibler divergence is a measure of the difference between two probability distributions (over the same event space). The Kullback-Leibler (KL) divergence between two probability distributions *p* and *q* on a finite set X is defined as

$$KL(p||q) = \sum_{x \in X} P(x).\log \frac{P(x)}{q(x)}, \quad \text{if X is discrete} \tag{3.3}$$

and as

$$KL(p||q) = \int_{-\infty}^{\infty} P(x).\log \frac{P(x)}{q(x)}\,dx, \quad \text{if X is continuous} \tag{3.4}$$

Since the Kullback-Leibler divergence is not a distance metric, it is not symmetrical and does not satisfy the triangle inequality. Therefore, various measures have been introduced, such as DKLD1 (Bigi, 2003) and DKLD2 (Bennet, 1998). The original asymmetrical definition of the KL distance is changed into a symmetrical version:

$$DKLD1 = KL(p||q) + KL(q||p) \tag{3.5}$$

$$DKLD2 = \frac{1}{2}\left[KL\left(p||\frac{p+q}{2}\right) + KL\left(q||\frac{p+q}{2}\right)\right] \tag{3.6}$$

Myrvoll and Soong (2003) demonstrated that the divergence between two multivariate normal distributions as the following form:

If $X \sim N(\mu_x, \Sigma_x)$ and $Y \sim N(\mu_y, \Sigma_y)$ are independent random variables, then

$$D(X, Y) = \frac{1}{2}\text{Trace}\left[(\Sigma_x^{-1} + \Sigma_y^{-1})(\mu_x - \mu_y)(\mu_x - \mu_y)^T + \Sigma_x \Sigma_y^{-1} + \Sigma_y \Sigma_x^{-1} - 2I\right] \tag{3.7}$$

Where μ and Σ are the corresponding mean vectors and covariance matrices, respectively, the superscript T to a vector means 'transposed' and -1 to a matrix means its inverse, 'trace' means the sum diag of a matrix ($\sum a_{ii}$).

In this thesis formula 3.7 was employed to compute the distances between the output clustering in Figure 4.4, Section 4.4. MATLAB was used to implement the formula 3.7. The program can be found in Appendix A and is included on the attached CD.

3.2.7 Clustering by density Estimation

Clusters are viewed as regions of the observation space in which the observations are with high density (mode), separated by regions of low observation density. Each mode is associated with a cluster center and each observation is assigned to the cluster with the closest center. For an unknown density distribution, the probability *P* that an observation *x* falls in region R with radius r_n and volume V_n. Suppose K_n of n observations falls in R, and then the density at *x* can be estimated by

$$\hat{P}_n(x) = K_n / (n \cdot V_n) \quad (3.8)$$

The performance of the estimation influence with the choices of the parameter K_n and V_n. A small K_n or V_n leads to smaller bias while a greater bias with larger choice of K_n or V_n. Two approaches of Parzen window and k-nearest neighbour have been existed for choice of V_n. The volume V_n in the Parzen window approach is specified as a function of n, such as $V_n = \sqrt{n}$ but let K_n be a random variable. In the nearest-neighbour approach, K_n is specified as a function of *n*, such as $K_n=\sqrt{n}$, but let V_n be a random variable (study by Duda (1973) cited Cios 2007: 166).

Wong and Lane (1983) estimated the underlying density of observations by the K_n-nearest neighbour method. Two observations x_i and x_j are said to be neighbours if x_i is one of the K_n nearest neighbours of x_j and if x_j is among the K_n observations closest to x_i. The dissimilarity between neighboring patterns x_i *and* x_j is given by:

$$d(x_i, x_j) = \frac{1}{2\hat{P}_n(x_i)} + \frac{1}{2\hat{P}_n(x_j)} \quad (3.9)$$

Pairs of observation that are not neighbours are assigned arbitrarily large dissimilarities. Definition (3.9) in this study is redefined by taking the weighted average of the probability density functions as illustrated in Section 3.2.8.

A histogram is the simplest way to represent the frequency distribution of a dataset graphically. Histograms are easy to create and are computationally feasible. Histograms can be used to identify modes in the datasets. Regions with relatively high frequency counts are the potential modes or cluster center and the boundaries between clusters fall in the valleys of the histogram. Thus, it is suited for summarizing the information about relative frequencies of observations in large datasets. Therefore a histogram is used in this study for the identification of characteristics of distributions of datasets such as shape, spread. That is to delimit the concentrated location.

3.2.8 Weighting

To weight a variable means to give it greater or lesser importance than another variable. When the variable is continuous the most commonly employed weights for a variable are the reciprocal of its standard deviation or the reciprocal of its range (Everitt 2001).

The dissimilarity definition (3.9) is redefined by taking the weighted average of the probability density functions; let σ_r and σ_s be variance of clusters *r* and *s* respectively, and then the weights could be defined as follows:

$$W_r = \frac{\sigma_s}{\sigma_r + \sigma_s} \tag{3.10}$$

and

$$W_s = \frac{\sigma_r}{\sigma_r + \sigma_s}$$

This is the strength of each cluster to pull dissimilarity to its self.

3.2.9 Finite mixture distribution

Finite mixtures of distributions have provided a mathematical based approach to the statistical modelling of a wide variety of random phenomena, such as modelling the distribution of diseases (crime) from mixed population. Finite mixtures provide suitable models for cluster analysis ''if each group of observations in a data set suspected to contain clusters, comes from a population with a different probability distribution"(Everitt 2001: 118). A mixture distribution is a compounding of statistical distributions. That is, when sampling from mixed populations. Each component is with a different probability density function. For a random variable *X*, finite mixture models decompose a probability density function f(x) into sum of *K* weighted densities. A model with *K* components is written in the form:

$$f(x;\theta_i)=\sum_{i=1}^{K} p_i \; g(X;\theta_i) \qquad (3.11)$$

When P_i denotes the proportion of the *ith* weight (mixing coefficient for the *ith* term) and $\sum_{i=1}^{K} p_i = 1.\; g(X;\theta_i)$ denotes a probability density with parameters represented by the vector θ_i. There is variety of estimation methods that have applied to estimate the parameter of the vector θ, such as maximum likelihood, least squares and moments. Finite mixtures with multivariate normal components have been widely used to model multivariate data of a continuous nature, while, multivariate Bernoulli densities suitable for categorical data. The *t* distribution provides a longer-tailed alternative to the normal distribution and thus provides a more robust approach to the fitting of mixture models (McLachlan 2000: 6).

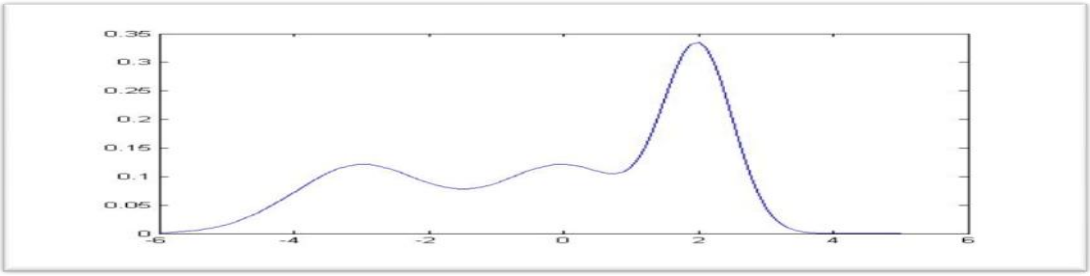


Figure 3.4: Density of a mixture of three Normal N (-3, 1); N (0, 1); N (2, 0.5) with weights 0.3, 0.3 and 0.4.

3.3 Clustering Algorithm with Significance

This algorithm is based on a hybrid of statistical methods and employs specific properties of statistical distributions to determine centers with boundaries of each cluster drawn by suggesting a possible significance level. This algorithm is significantly less time consuming, efficient and objective for finding clusters within the datasets.

The algorithm is composed of the following steps:

Step 1 Examine data outside their region:

- draw lines parallel to x-axis and y-axis (use the notation L_x and L_y to denote it respectively).
- Project each data point first on the L_x-axis and then on the L_y-axis.
- draw the histograms on the L_x-axis and L_y-axis.
- specify the lower density in the histograms that are created in the L_x-axis and L_y-axis.

Step 2 Employ specific properties of statistical distributions to predict possible cluster centers.

Compute maximum likelihood for each distribution on the L_x-axis and L_y-axis (μ_{xi} and μ_{yi} denoted coordinates of center cluster i).

Step 3 Find the optimal size of any clusters.

For each distribution as appropriate assume a confidence interval to determine the radius of each (cluster) ellipse.

The algorithm iterates step 2 and 3.

MATLAB was used to implement the algorithm. The program can be found in Appendix *A and is included on the attached CD*.

3.3.1 Data Representation

In this chapter results derived from the application of artificial data to the developed algorithm are presented. The advantage of using the artificial data is that it can be used as a validation tool. More specifically, it relates to the possibility of knowing in advance what to look for, what is important in the data and what kind of classification can be obtained from a clustering method. The most direct visualisation is a two-dimensional plot showing the objects to be clustered as points. A random sample of size 100 and 6000 was generated from a normal distribution with mean μ and standard deviation σ and this is illustrated in Table 3. 1.

3.3.2 Experimental results on an artificial dataset

By projecting the data into their marginal dimensions separately and formulating single histograms, acceptable forms of the marginal distributions with their general characteristics could be obtained (see Figure 3.7). Following this, different approaches could be used in order to estimate the position of the center of the clusters. For example, using the means of the produced marginal distributions or their modes, depending on the objective of the study, or using maximum likelihood as it has been followed here (Figure 3.8).

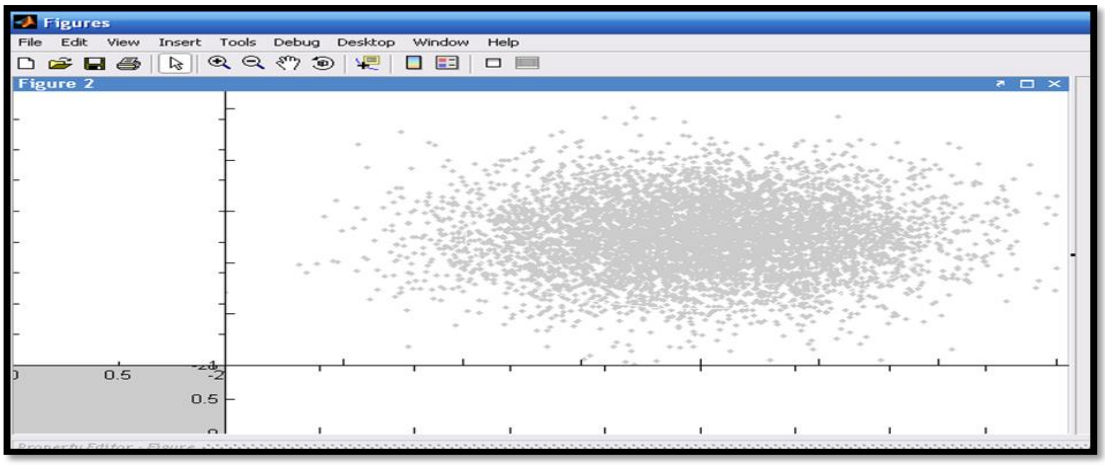


Figure 3.5: Two dimensional plots of 6000 data points illustrated in Table 3.1, N (0, 5) and N (5, 7).

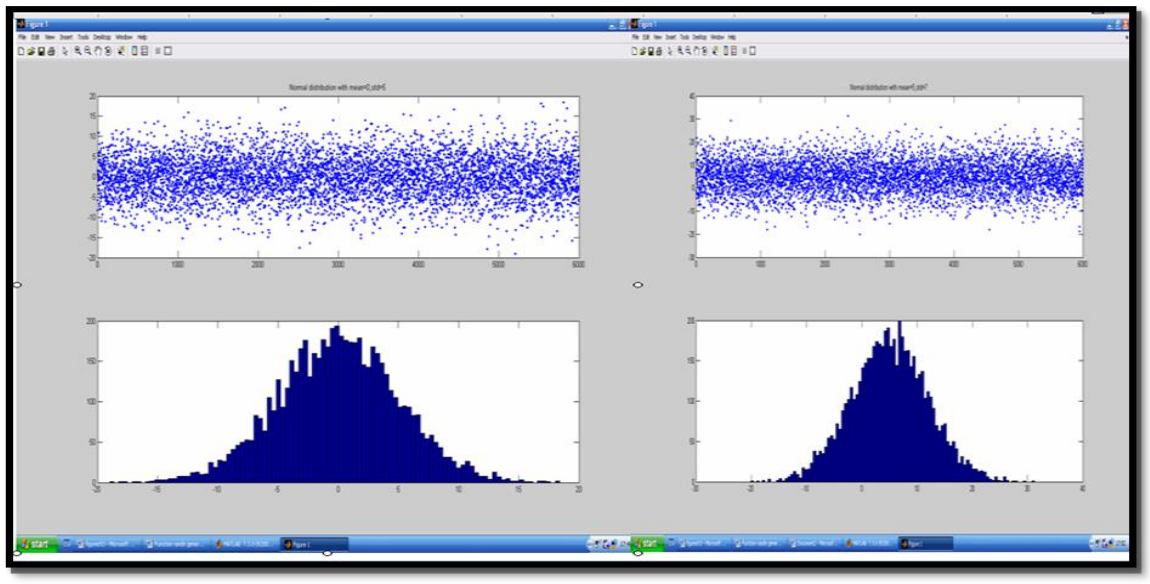


Figure 3.6: Simulated 6000 values from each of N (0, 5) and N (5, 7) and their histogram.

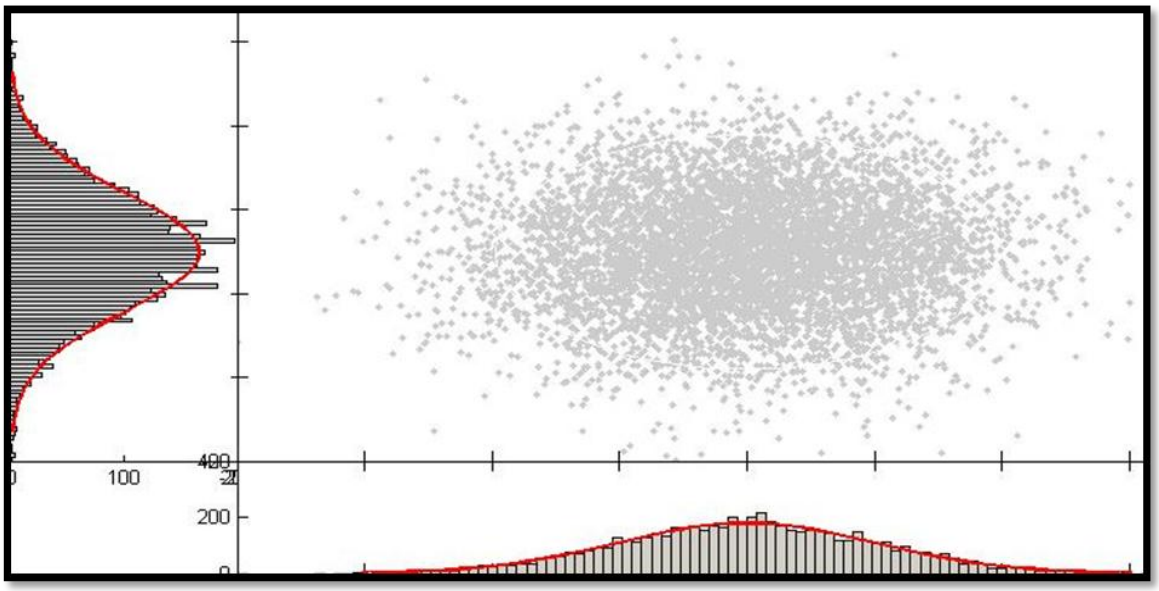


Figure 3.7: Illustration of step 1 of the algorithm; plotting data points and drawing histogram on parallel axes L_x and L_y.

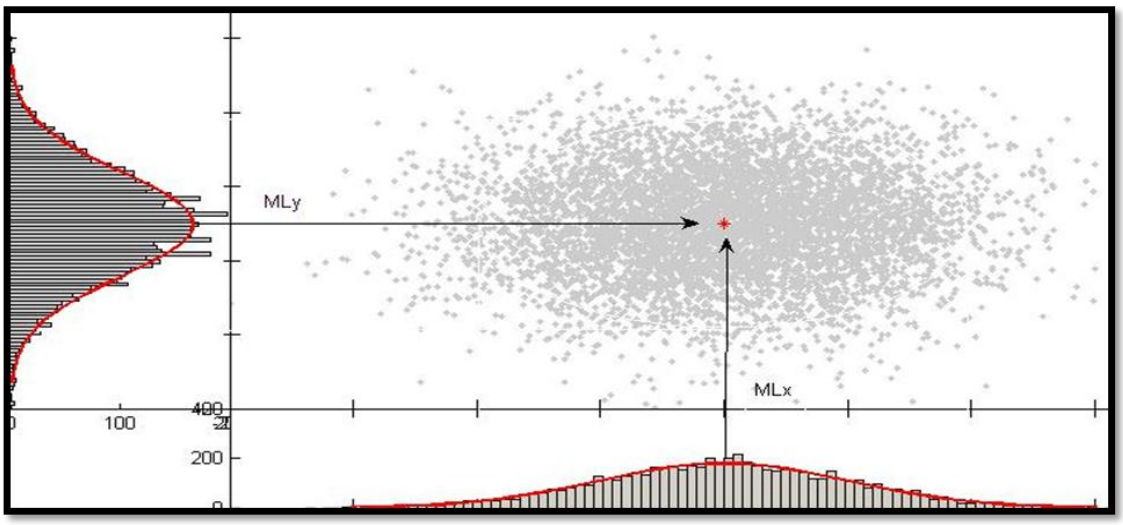


Figure 3.8: Illustration of step 2 of the algorithm; predicted cluster center (Maximum likelihood for the distribution on the L_x and the L_y-axis).

Cluster sizes are determined according to confidence interval requirements; confidence intervals are ranges of numbers that have a high probability of the inclusion of the unknown parameter as an interior point. As illustrated in figure 3.9, the smallest ellipse contains 68% while the largest contains 95% of the observations.

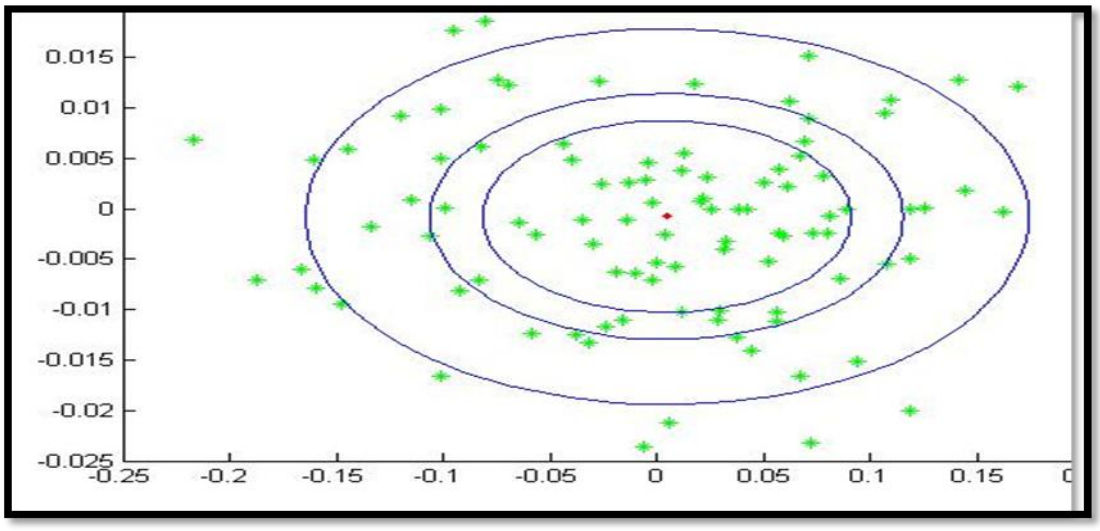


Figure 3.9: Some examples of confidence interval: First cluster with 68%, second cluster with 80% and third with 95%.

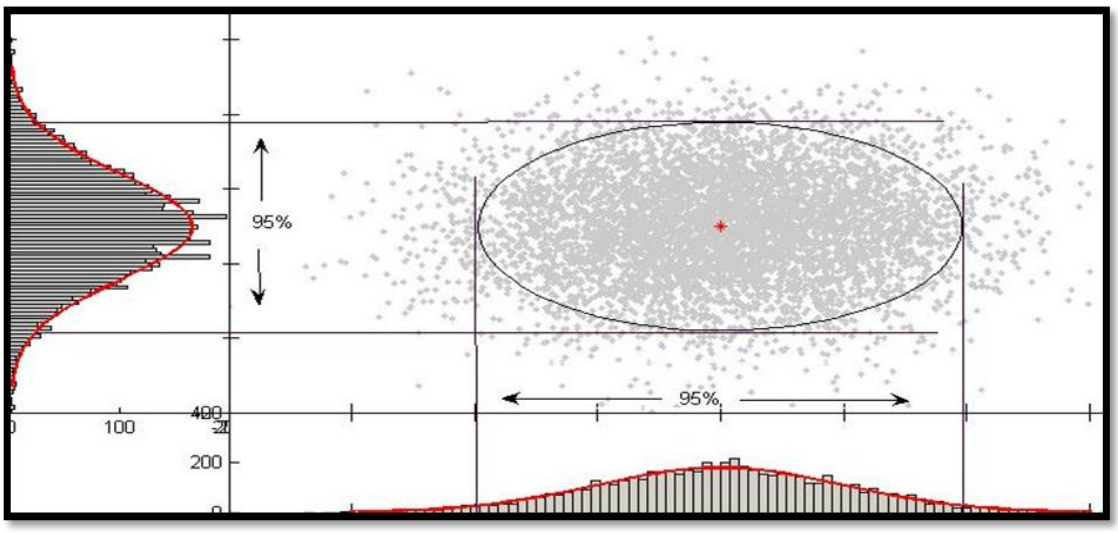


Figure 3.10: Illustration step 3 of the algorithm existence of the cluster.

The model used is that of a mixture of a finite number of probability distributions. The number[1] of mode indicators as the number of distributions in the mixture is illustrated in Figure 3.11.

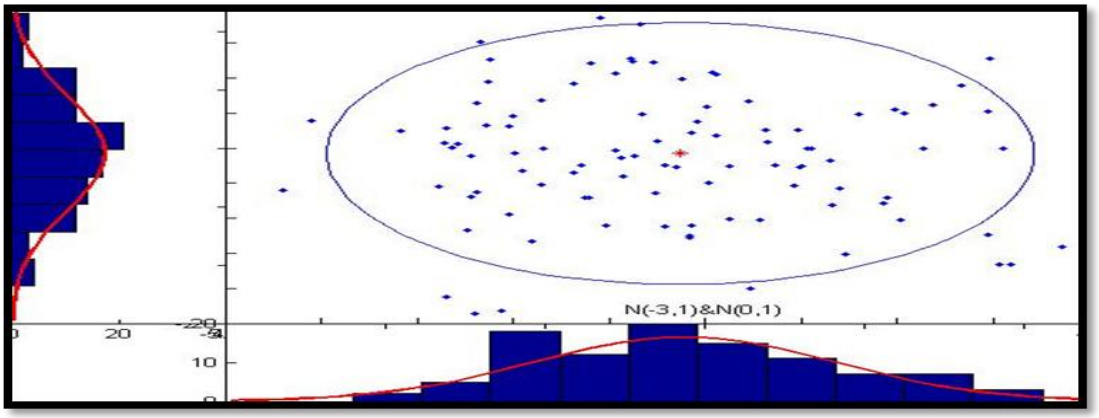


Figure 3.11a: Example, of a number of clusters that indicator as the number of distributions in the mixture. Plotting of 100 values from each of N(0,1) and N(-3,1).

---

[1](Number of probability distribution projected on x-axis) X (Number of probability distribution projected on y-axis).

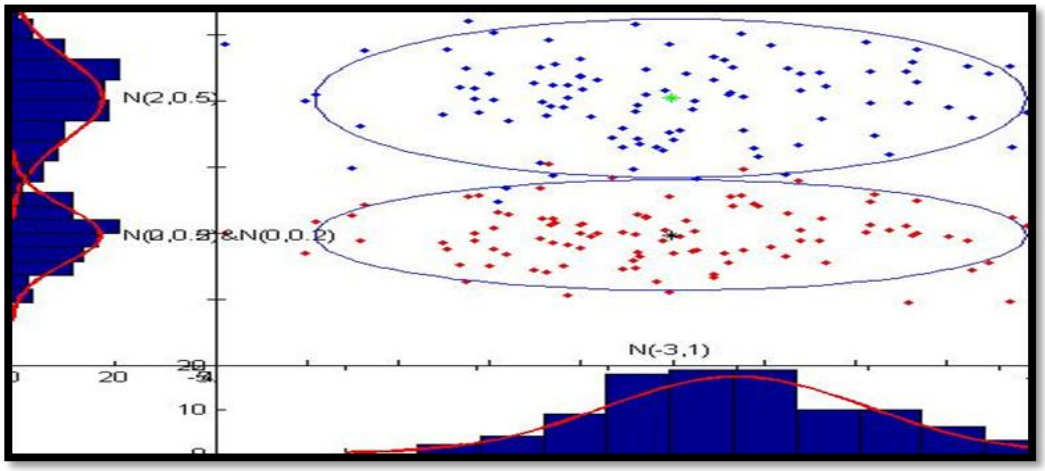


Figure 3.11b: Example, of a number of clusters that indicator as the number of distributions in the mixture. Plotting of 200 values from each of N(-3,1),N(0,0.2) and N(2,0.5).

3.4 Summary and Conclusion

This chapter introduces a new cluster detection methodology, called Salar's Clustering with Significance (SCS). This technique has been developed from knowledge drawn through a hybrid of both statistical and computing techniques. The algorithm for its implementation has employed specific properties of statistical distributions to estimate the position of the centers, and the boundaries of each cluster using subjectively set significance levels.

The algorithm was tested using artificial datasets, with very promising results. Experimental results demonstrate that SCS is especially suitable for large dataset (see Figure 3.10) and even for small sample size (see Figure 3.11). The attribution of the SCS algorithm are: easy to implement; no previous knowledge of the data set requirements; less number of performed steps leads to a reduction in clustering time; and the results provide with the detail information about the distribution of cases within the dataset. It performs reasonably well in terms of: memory

requirements; running time; cluster quality. The algorithm requires that the user first specify the valleys (lower density) in the histograms that are created in the marginal axes for data splitting. This process was used to delimit the concentrated location.

To test the validity of the clustering methodology, chapter 4 includes applying the algorithm to real (crime) datasets.

4 Application of SCS Algorithm on Real Data

This chapter presents the application of SCS algorithm on real (crime) datasets. The promising results obtained from identification of hotspots are compared with other available algorithms such as CLAP, Satscan and GAM. The algorithm is also applied to the rotation of the same real dataset to illustrate its robustness.

4.1 Introduction

The previous chapter the strength of the new clustering algorithm on artificial data sets has been explored. For real data, residual burglary incidence and suggested rotation degrees (30 and 85) of the same real data have been used for testing the effectiveness of the new clustering methodology (see Sections 4.4 and 4.7). One advantage of using rotation in this thesis is to allow the algorithm to be applied to more than one set of real data. That is to demonstrate the robustness of the algorithm. Burglary or 'breaking and entering' is one of the most common crimes worldwide. Burglary is a high volume crime affecting around one in twenty five households annually in UK (Home office 2008).

A new cluster detection methodology was utilized for performing the procedure of identification hotspots (see Section 4.6). For testing the effectiveness of the algorithm the obtained results are compared with results of other available algorithms that are applied to the same real datasets such as CLAP, Satscan and GAM.

4.2 Residential burglary

The target research of choice is residential burglary which takes three points of rationale. First for testing the effectiveness of the new clustering methodology (see Section 4.4). This case is related to clustering and identifying high levels (hotspot) of burglary incidence. The obtained results were utilized in a development of a predictive crime model (see chapter 6). The second point of rationale is that burglary incidents are not randomly distributed; certain types of household characteristics and certain locations suffer from burglary more frequently than others. It aims to assess the risk of households within parcels in study area being subjected to residential burglary. For this analysis a predictive crime model is built. The model employed potential explanatory variables of characteristics of burgled households, such as household composition, socio-economic, and household spaces and accommodation type. Burglary rate constituted as the response variable (see Section 6.4). Burglary is a phenomenon with disparities within a geographical distribution. This case is related to the third point of rationale. GIS was used to construct maps to depict spatial distribution of residential burglary rate in the study region, integration information of different data sources based on common geographic variable (see chapter 5). The obtained results utilized in a predictive crime model are presented in Sections 6.8 and 6.9.

4.3 Data Representation

Cluster detection techniques can be applied to point datasets in order to reveal information relating to the spatial distribution of the cases (for example crime, disease). As was pointed out in section 4.1 the aim of this chapter is to test the validity of the clustering methodology (SCS); i.e. identification of high level of crime (hotspot). Thereby the data utilized in this experiment relies on two main sources of information. These are spatial distribution of burglaries and population data over the same area. These two types of data combine in this case for crime rate measurements within each cluster. Crime rate is a measure of the rate of

occurrence of crimes committed in a given area and time. It is the proportion of crimes committed among a given number of persons and is a useful statistic for many purposes, such as evaluating the effectiveness of crime prevention measures and for comparison purposes between the state of different areas and/or different times within the same area.

Police record crime that has been reported. The Home office issues rules to police covering the recording of data, counting, and classification of crime. The data are organized into a matrix, which is a table organized in rows and columns. Each row contains all the information for one particular record and each column contains information about one particular characteristic describing the data (for example, date, time, location) (Rachel, 2006). Crime data[1] are sensitive in nature, and as such often require security clearances, special permissions in order to acquire the appropriate ethics approval.

In 2003, 10905 cases representing spatial distribution of burglaries were reported to the police during the period of 6 February to 31 October in the study region. The input crime data file, that utilized by MATLAB include the location (x and y co-ordinate) address to one residential burglary incidence. 294310 cases (population size) that cover the same area under study were downloaded from the CASWEB[2] website (CASWEB, 2009) and have been geo coded. The census is a key source for information about the instance, localities, key concepts of population, household and residence. The contribution of population data in this analysis is for measuring crime rate when obtaining clusters.

---

[1] Ethics: No personal data were held and all diagrams illustrating the distribution of crime are distorted to obfuscate the region covered. In addition, all regional labels associated with causal relationships have been removed to ensure anonymity.

[2] CASWEB: Is a web interface to an aggregate statistics and related information from the UK census of population. The latest census was held in 2001 (CASWEB 2009).

4.4 Experimental results on real datasets (crime)

This section presents the application of the SCS algorithm on the real datasets (crime) that was identified in Section 4.3. MATLAB is a software package that is highly effective for the production of graphical displays utilized to implement the SCS algorithm. The following steps are applied to read the input data file from excel within MATLAB: save the input crime data file that identify above into Excel; open the file from the command window menus; then select the desired file let say Sheet1 from import wizand; finally use this command, x=Sheet1(:,:) in command window. The file is now ready for use in MATLAB. Figure 4.1 shows distribution of 10905 data points of burglary incidence in the study area. Projecting the data as shown in Figure 4.2 into their marginal dimensions separately and formulating histograms. This obtains the following general characteristics of the data: Shape, spread or location of the distributions. This process is performed in step 1 of the algorithm. A frequency distribution and property of Normal distribution are used to delimit the concentrated location. It is requisite in this stage to delimit the interval of the produced marginal distributions with respect to marginal dimensions *x* and *y*. Maximum likelihood of the produced marginal distributions is used to estimate the position of the center of the clusters which were generated in step 2 (see Figure 4.3). Figure 4.3 illustrated that the mean of each distribution in marginal dimensions is the center of its graph (property of Normal distribution). For the optimal size of any clusters, 95% is suggested to be a confidence interval for each distribution in marginal dimensions. Figure 4.4 presents the existence of the clusters (step 3). The formula 3.2 in Section 3.2.3 was employed to draw the boundary of each cluster.

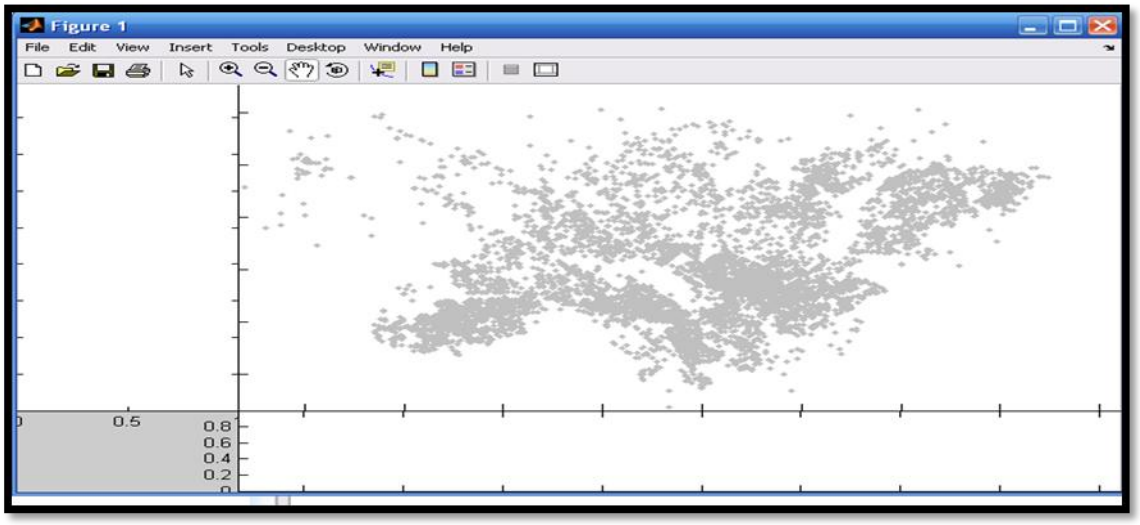


Figure 4.1: Plots showing the distribution of 10905 data points of burglary incidence in the study area.

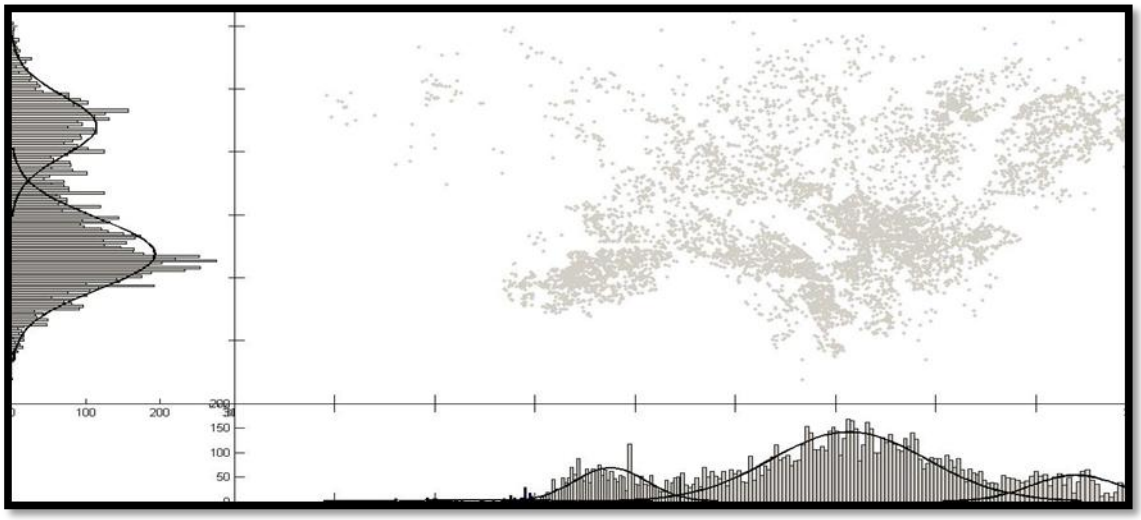


Figure 4.2: Illustration of step 1 of the algorithm; plotting data points and drawing histogram on parallel axes L_x and L_y.

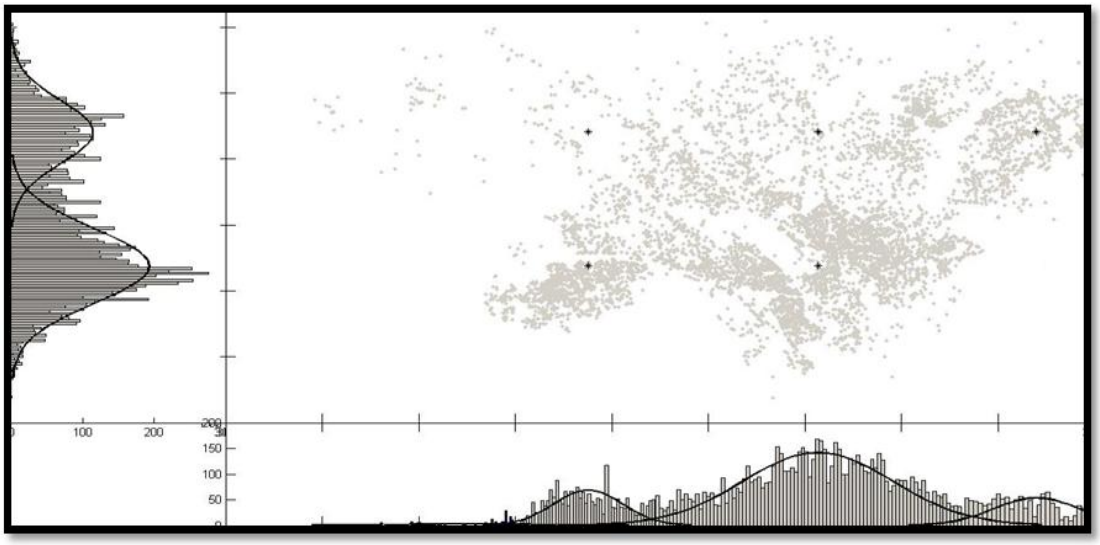


Figure 4.3: Illustration of step 2 of the algorithm determining the center of clusters.

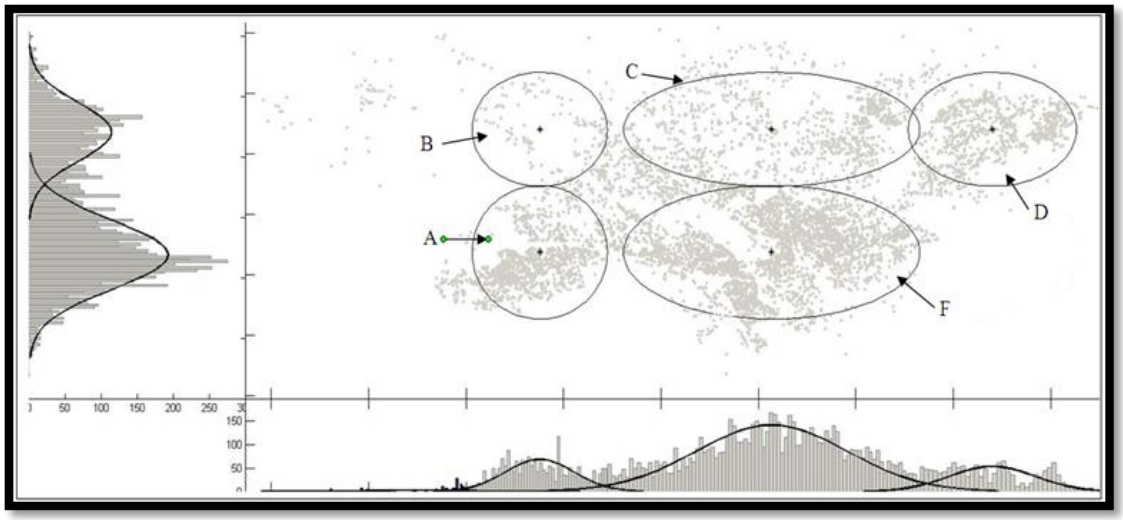


Figure 4.4: Illustration of step 3 of the algorithm existence of the clusters.

4.5 Summary of SCS results

The following output results of clustering (Figure 4.4) were obtained utilizing the SCS algorithm on real (crime) datasets. Determining of the position of the center of the clusters; the boundaries of each cluster (radius); the distances between the outputs clustering, and the number of cases (crime) within each cluster (size). Kullback-Leibler divergence (KL) formula 3.7 was employed to compute the distances between the outputs clustering in Figure 4.4. MATLAB was used to implement the formula 3.7 described in details in Section 3.2.6.

Table 4.1 Summary of the output result of clustering that shown in figure 4.4

Cluster	***A***	***B***	***C***	***D***	***F***
Center (x,y)	(3.135e+00, 1.768e+005)	(3.13e5+005, 1.808e+005)	(3.183e+005, 1.808e+005)	(3.2228e+005, 1.808e+005)	(3.183e+005,1.768e+005)
Radius with respect x & y meters	$x_r = 1400$ $y_r = 2100$	$x_r = 1400$ $y_r = 1900$	$x_r = 3100$ $y_r = 1900$	$x_r = 1700$ $y_r = 1900$	$x_r = 3100$ $y_r = 2100$
KL	KL(A,B) **24.0898** KL(A,C) **41.8950** KL(A,D) **159.4593** KL(A,F) **42.1217**	KL(B,C) **23.7669** KL(B,D) **126.5060** KL(B,F) **29.7583**	KL(C,D) **17.2823** KL(C,F) **14.1894**	KL(D,F) **31.4622**	
Number of crime	1751	290	1646	1246	5336

4.6 Crime hotspots analysis

Hot spots of crime are ''small areas that have statistically significant high levels of crime relative to surrounding areas'' (Chetwin et al., 2005: 26). Crime "hot spot" analysis is one of the important topics in crime spatial pattern analysis. In investigating the spatial autocorrelation of crime incidents, hot spot analysis has been an important approach for the explanation and prediction of crime spatial patterns. Hotspots allowed law enforcement to examine criminal phenomena within the area of concern. This is for more policing concern in a certain location. It allows crime analysts to identify concern in certain location that helps to identify the problem associated with the characteristic of the people within the location. For analysis a crime analyst can use related methodologies, such as statistics or neural networks. Hotspots can be analysed spatially as well as temporally. This leads to significant contribution to crime preventions strategies. The many studies researching "hotspot", for example, Corcoran and Wilson (2003) used hotspots (spatial clusters of crime) for temporal analysis. Spatial and temporal analysis of burglary incidence by Ratcliffe (2001) showed that the housing characteristics of the residential burglary hotspots vary considerably across the city. The highest probability for residential burglaries was the period that most people were at work.

There are several techniques for the determination of hotspots. Each has their advantages and disadvantages. These mainly related to their case of use, visual results and interpretation. For example, spatial and temporal analysis of crime (STAC), generate a set of ellipses that represent the highest concentrations of points. This technique has the advantage of showing conclusive hotspot regions but show a gradual change from a hotspot area to a less dense crime area with no indication of cut-off points. GIS is one of the simplistic methods, requiring minimal GIS skills to create graduated circles. The disadvantage of this method is its draw backs in that these circles can often overlap. This overlaps making it difficult to visually discern patterns of concentrations. From this many research studies have gone into detection of spatial cluster of crime (Ratcliffe 2004).

Section 4.6.1 illustrates the process of identification of hotspots in the study area. The SCS algorithm is used to perform this identification. The main purpose of this process is to identify those locations that suffer from crime more than others. Then, the factors that associated to the characteristic of the people within these identification locations are examined. For this analysis, regression analysis and neural networks are utilized. Building a predictive crime model is discussed in more detail in chapter 6.

4.6.1 Identification of hotspots: SCS algorithm

The SCS algorithm was utilized for performing the procedure of identification of high levels of crime (hotspots) in the area under study. The procedure begins with detection of spatial cluster of distribution of the real (crime) data (Figure 4.5). Then, the crime rates within the clusters that are obtained from the first step are measured. Crime rate in this analysis is expressed as the number of crimes observed in that cluster per the combined population. This necessitates counting the number of crimes and population in a given cluster. The obtained results are illustrated in Table 4.2b. Table 4.2a present the output results of counting number of crime and its combined population size. The results of crime rate are presented as a matrix, let's say C with 10 rows and 18 columns. Each element C (I, J) in the matrix C, act out the information about crime rate within the clusters that display in Figure 4.5. For instance, C (1, 1) =0.011442 was obtaining as: 5 which is the number of crimes (numerator) observed per 437 the combined population (denominator). It is observed from the results presented in Table 4.2b that some values of rates are an outlier. That is, the lack of a residential population in the central business location leads to outlier rates in these locations. For example, the number of crime and population for $C_{8,10}$ are 250 and 10 respectively and then the value of crime rate is 25. 128 clusters have been found among combined crime incident and with population data. The output results are then examined to determine the hotspots. The clusters with high levels of crime (hotspot) are those with a crime rate greater than or equal to crime rate of the study region. In order to clarify the results of the

specific locations of high-rate clusters these clusters were shaded in red (see Figure 4.5). The details of the program can be found in Appendix B.

The objective of figure 4.6 is to represent the location of the distribution of hotspots in the area under study. MATLAB capability was used to colour the case of the high levels (hotspot) of occurrence which are within a red boundary of clusters, figure 4.5, with red and light grey for low levels (cold spot). Finally the boundary of the clusters is removed.

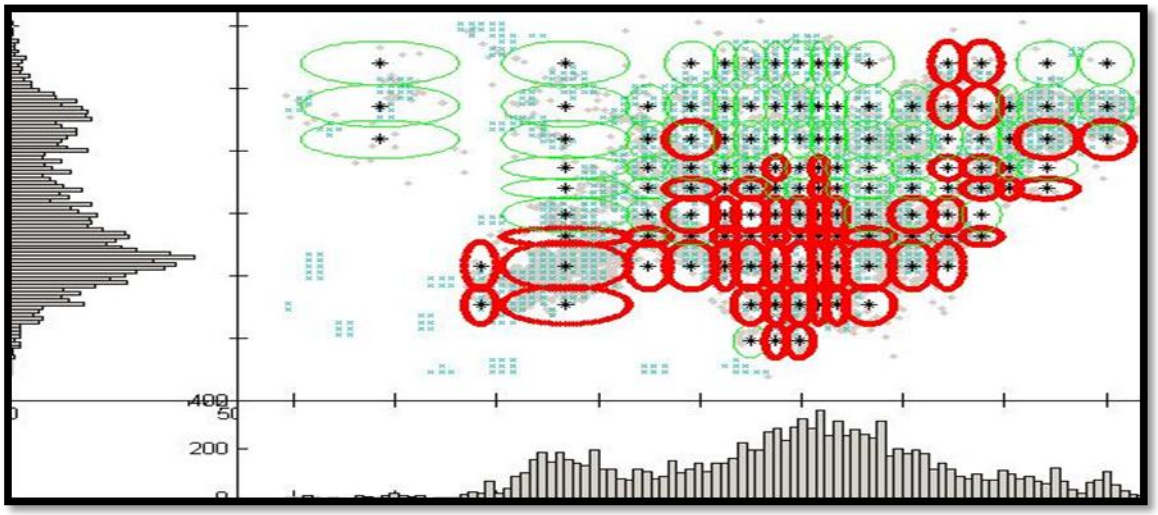


Figure 4.5: Delimit the concentrated location; red shading shows the specific locations of high-rate clusters.

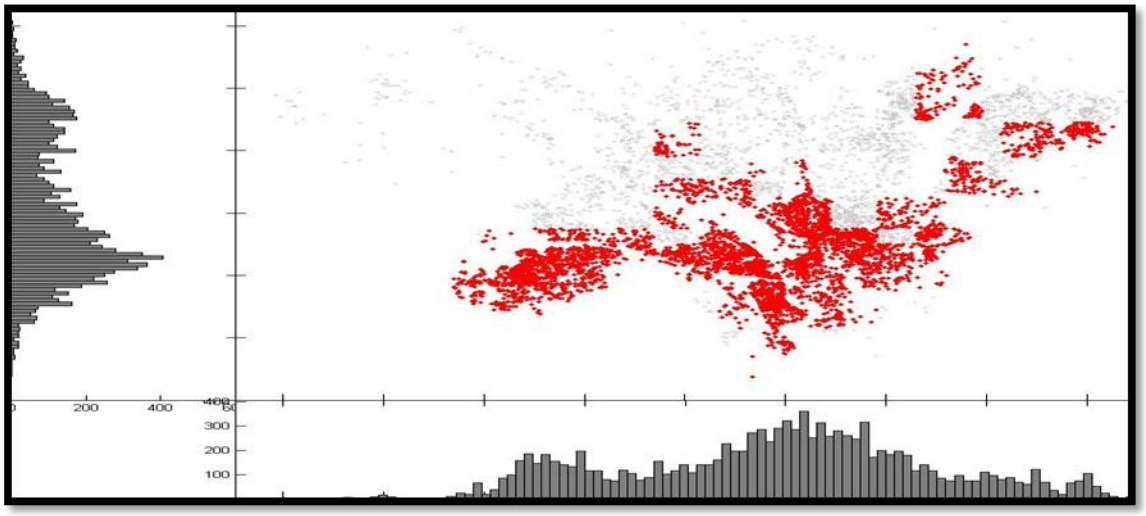


Figure 4.6: Shows the distribution of burglary incidence concentration (hotspot). Hotspots red area showing and low levels show in light gray area.

Table 4.2a: Details of the obtained results of number of crime and combined population for 10x18 clusters; (a)number of crime and (b) number of population within obtained clusters.

(a)

5	4	25	3	2	18	25	26	10	14	7	11	4	11	26	7	5	2
45	0	74	22	57	26	53	42	66	4	8	27	201	122	59	88	221	210
8	3	49	113	120	29	53	14	27	23	17	100	130	5	69	66	204	134
7	1	36	64	73	18	39	52	20	10	9	25	42	2	76	39	31	2
1	0	52	51	79	32	63	64	42	33	22	47	18	12	49	21	29	0
0	0	167	88	85	13	50	34	142	101	173	323	124	70	5	4	4	0
0	1	103	50	41	47	54	8	33	85	119	279	81	86	36	0	0	0
0	57	1226	58	174	160	286	280	219	250	179	438	228	79	9	0	0	0
0	21	165	0	7	11	57	120	226	50	107	109	9	8	0	0	0	0
0	0	0	0	0	0	21	12	48	0	4	13	5	0	0	0	0	0

(b)

437	0	1923	0	729	707	2658	2403	2606	982	1304	386	0	278	78	0	567	238
4176	0	4512	2394	4431	2366	1523	1350	3370	1126	674	2951	6218	2120	785	2674	6527	6466
459	0	2521	4350	2493	2143	3144	1593	1278	688	1355	5021	4035	838	2043	2207	5822	2019
0	0	2152	3724	2468	1681	2241	554	927	282	606	1527	3434	56	2037	1687	1060	0
0	0	3976	1518	2130	598	1681	2272	1243	9	634	2727	1141	542	1218	587	12	0
0	44	8333	2870	1515	301	346	379	3925	2370	4351	9635	2991	637	323	0	0	0
0	0	2555	1217	2161	817	1134	1	59	175	877	3106	3168	1747	56	0	0	0
1100	1073	20537	1042	4776	4124	5542	5240	2976	10	698	4499	6140	1082	0	0	0	0
454	393	4110	0	0	0	342	1456	5733	887	1146	1421	0	0	0	0	0	0
188	0	814	306	0	114	949	56	372	0	0	0	0	0	0	0	0	0

0.0114	0	0.013	0	0.0027	0.0255	0.0094	0.0108	0.0038	0.0143	0.0054	0.0285	0	0.0396	0.3333	0	0.0088	0.0084
0.0108	0	0.0164	0.0092	0.0129	0.011	0.0348	0.0311	0.0196	0.0036	0.0119	0.0092	0.0323	0.0576	0.0752	0.0329	0.0339	0.0325
0.0174	0	0.0194	0.026	0.0481	0.0135	0.0169	0.0088	0.0211	0.0334	0.0126	0.0199	0.0322	0.006	0.0338	0.0299	0.035	0.0664
0	0	0.0167	0.0172	0.0296	0.0107	0.0174	0.0939	0.0216	0.0355	0.0149	0.0164	0.0122	0.0357	0.0373	0.0231	0.0292	0
0	0	0.0131	0.0336	0.0371	0.0535	0.0375	0.0282	0.0338	3.6667	0.0347	0.0172	0.0158	0.0221	0.0402	0.0358	2.4167	0
0	0	0.02	0.0307	0.0561	0.0432	0.1445	0.0897	0.0362	0.0426	0.0398	0.0335	0.0415	0.1099	0.0155	0	0	0
0	0	0.0403	0.0411	0.019	0.0575	0.0476	8	0.5593	0.4857	0.1357	0.0898	0.0256	0.0492	0.6429	0	0	0
0	0.0531	0.0597	0.0557	0.0364	0.0388	0.0516	0.0534	0.0736	25	0.2565	0.0974	0.0371	0.073	0	0	0	0
0	0.0534	0.0401	0	0	0	0.1667	0.0824	0.0394	0.0564	0.0934	0.0767	0	0	0	0	0	0
0	0	0	0	0	0	0.022	0.214	0.129	0	0	0	0	0	0	0	0	0

Table 4.2b Details of the obtained results of crime rate (expressed as the number of crime observed in that cluster per the combined population) for 10X18 clusters. 128 cases have been found among combined crime incident with population size.

4.7 Experimental results on rotation datasets

This section introduces the concept of rotation of data. The results that were derived from the application of rotation datasets are used for testing the effectiveness of the SCS algorithm.

4.7.1 Rotation

The goal of rotation is to simplify and clarify the data structure and this process often allows the analyst to observe clusters more clearly especially when close correlations are present between neighbouring clusters. Criteria for determining better rotations are not always clear (Pedhazur & Schmelkin, 1991: 611). Rotations can improve the interpretability of the results. One advantage of using rotation in this thesis is to allow the algorithm to be applied to more than one set of real data. That is to demonstrate the robustness of the algorithm.

There are two types of rotation that can be done; orthogonal when the new axes are also orthogonal to each other and oblique when the new axes are not required to be orthogonal. In this case study orthogonal rotation was used. The values in the factor transformation matrix consist of sines and cosines of the angle of axis rotation θ. This matrix is multiplied by the matrix of a un-rotated factor, to obtain a matrix of a rotated factor. For the case of two factors the factor transformation matrix would be:

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

X (rotation) = $X' * \cos\theta - Y' * \sin\theta$

(4.1)

Y (rotation) = $X' * \sin\theta + Y' * \cos\theta$

Where X' and Y' are un- rotated datasets.

The angle of rotation is found in an iterative way (Field 2005: 635).

In this study, MATLAB was used to implement formula 4.1. The program can be found in Appendix B.

4.7.2 Implemented SCS algorithm

In this section results derived from the application of rotation the same real (crime) datasets are presented. For instance, rotation of the data points with 30 and 85 degrees are obtained by utilizing formula 4.1. The clustering results (Figure 4.8) are obtained utilizing the SCS algorithm on rotation of real crime data with 85 degree. The results are promising and are compared with the un-rotated output. The clusters named in Figure 4.8 indicate that these clusters are the same as in Figure 4.4.

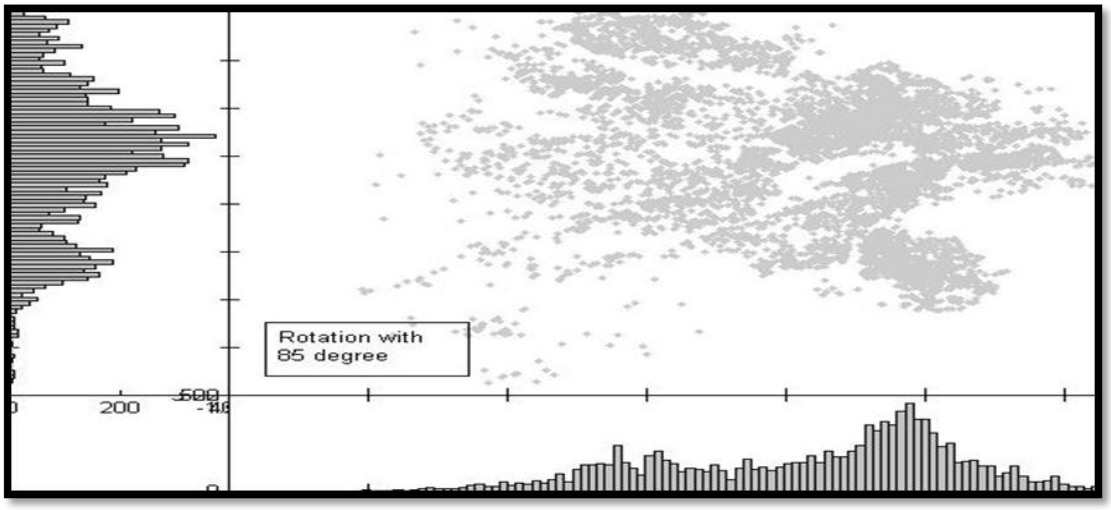


Figure 4.7: Plots showing the distribution of crime datasets. Rotation data with 85 degree utilized.

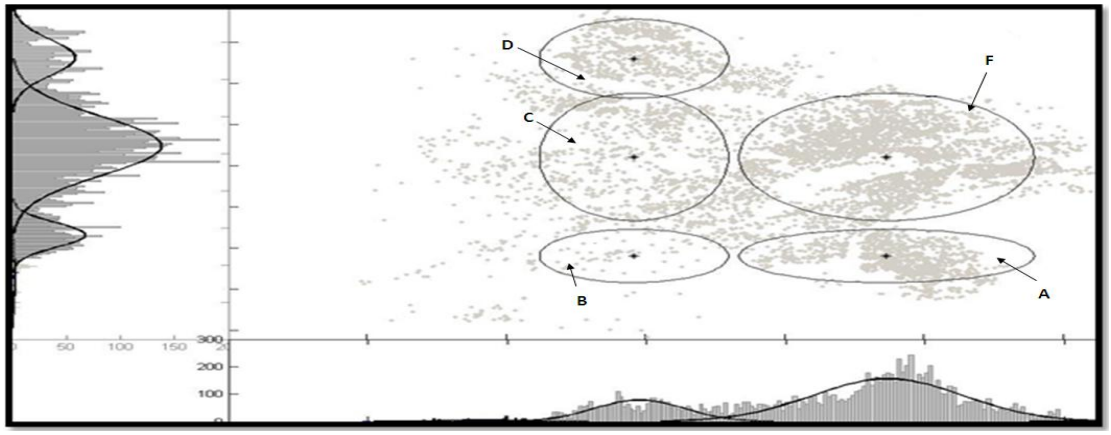


Figure 4.8: Illustrated clustering of the rotation of data points of 85 degrees using the clustering algorithm SCS.

The identification of hotspots results (Figures 4.11 and 4.12) are obtained utilizing the SCS algorithm on rotation of real crime data with 30 and 85 degrees. The promising results obtained for identified hotspots are compared with the un-rotated output Figure 4.6. The outputs are superficially very similar. The Arc view ESRI GIS package was used in chapter 5 for visualization of these results.

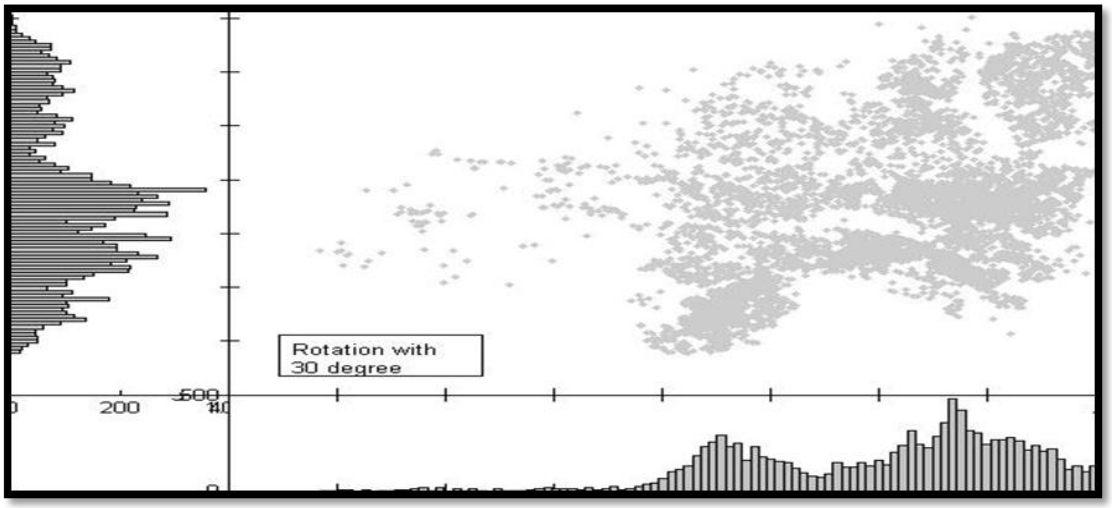


Figure 4.9: Plots showing the distribution of crime datasets. Rotation data with 30 degree utilized.

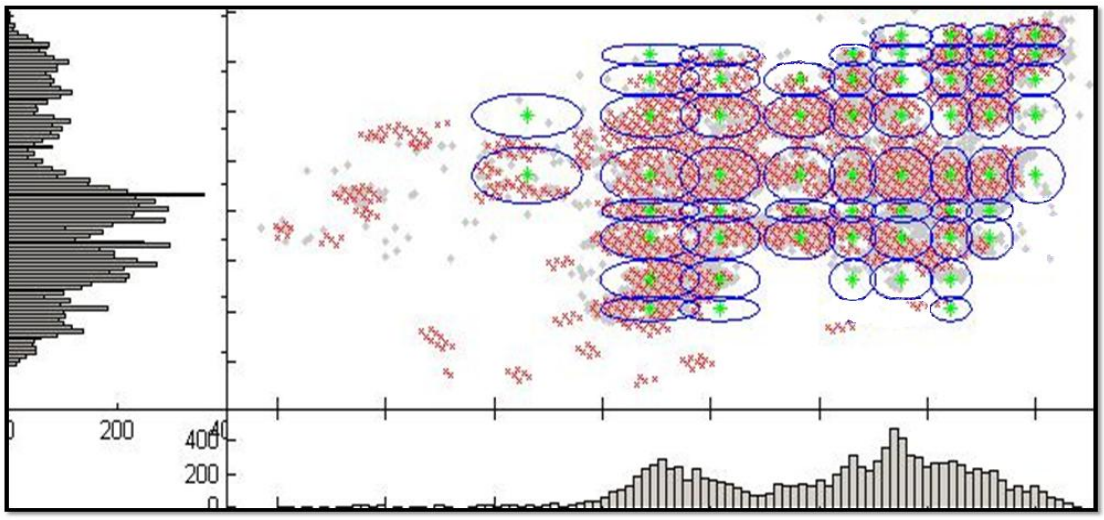


Figure 4.10: Delimit the concentrated location.

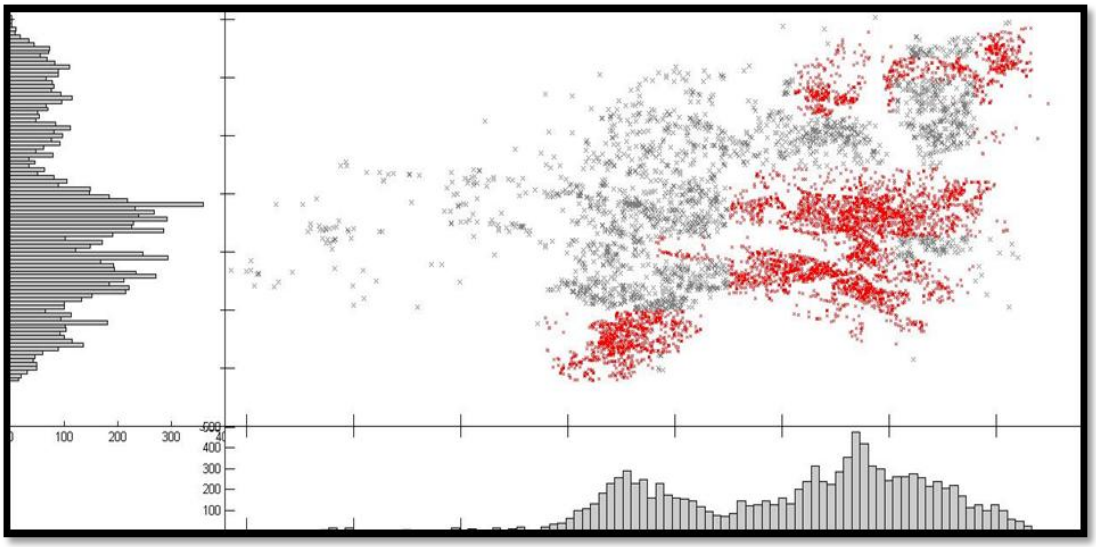


Figure 4.11: Hotspot identification: Rotation of data points of 30 degrees utilized. Hotspots red area showing and low levels show in light gray area.

Table 4.3: Details of the obtained results of crime rate (expressed as the number of crime observed in that cluster per the combined population). Rotation data with 30 degree utilized.

(a) Number of crime

0	0	0	1	3	31	44	45	183
0	7	16	8	33	56	98	140	156
0	50	39	30	189	168	88	150	64
2	98	178	91	152	130	71	133	44
95	387	317	356	389	959	463	327	94
0	123	106	101	86	426	116	122	12
0	292	305	424	360	610	217	98	18
0	811	278	0	29	59	55	0	0
0	262	6	0	0	0	1	0	0

(b) Number of population

0	0	0	0	0	39	1202	1857	4432
0	30	1154	0	667	983	3179	3090	2845
0	3021	4992	1934	4957	3816	4506	5017	1597
161	10209	6817	6500	6514	6778	2623	4069	12
6039	15872	15514	7664	10339	19469	9893	8711	247
33	4599	1752	2594	727	88	430	0	0
127	12501	10105	11153	6478	10858	3935	700	0
171	11185	5151	0	30	422	578	0	0
287	5481	1428	0	23	925	144	0	0

(C) Crime rate

0	0	0	0	0	0.7949	0.0366	0.0242	0.0413
0	0.2333	0.0139	0	0.0495	0.057	0.0308	0.0453	0.0548
0	0.0166	0.0078	0.0155	0.0381	0.044	0.0195	0.0299	0.0401
0.0124	0.0096	0.0261	0.014	0.0233	0.0192	0.0271	0.0327	3.6667
0.0157	0.0244	0.0204	0.0465	0.0376	0.0493	0.0468	0.0375	0.3806
0	0.0267	0.0605	0.0389	0.1183	4.8409	0.2698	0	0
0	0.0234	0.0302	0.038	0.0556	0.0562	0.0551	0.14	0
0	0.0725	0.054	0	0.9667	0.1398	0.0952	0	0
0	0.0478	0.0042	0	0	0	0.0069	0	0

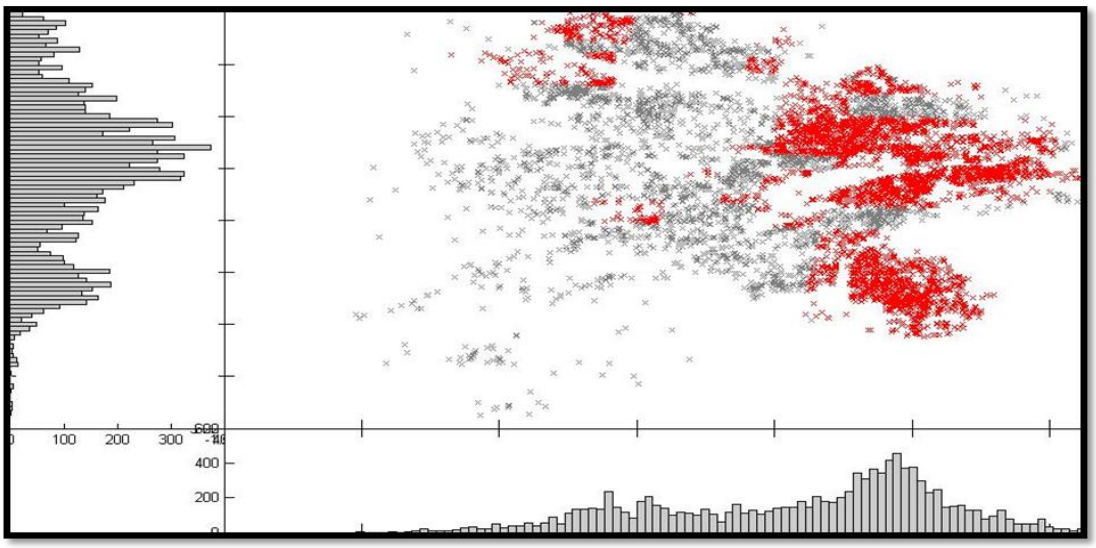


Figure 4.12: Hotspot identification: Rotation of data points of 85 degrees utilized. Hotspots red area showing and low levels show in light gray area.

4.8 Validation of SCS algorithm

Cluster validation refers to ‘‘procedures that evaluate the results of cluster analysis in a quantitative and objective fashion’’ (Jain, 1988: 143). For testing the effectiveness of the algorithm, first the same algorithm uses several datasets. This test demonstrates whether the algorithm can be expected to perform well for all types of data. The second test includes performance comparison in relation to other validated algorithms. If the results of clustering are similar, then this indicates that the algorithms perform well.

To demonstrate that the algorithm performs well, other available algorithms such as CLAP Satscan and GAM were applied to the same real datasets (crime). The four techniques identify the distribution of burglary incidence concentration (hotspot). However each of these techniques follow different principles(outlined in Section 2.1.1) but the final obtained results are very similar, specially SCS with Satscan result. This indicated that the SCS algorithm performs well. This is the main purpose of this comparison. Figure 4.14 shows a representation of final results obtained, while Figure 4.13 shows the location of clusters.

For testing the effectiveness of the SCS algorithm, the methodology was applied to suggested rotation degrees (30 and 85 degrees) of the same real data (crime). The results obtained from identification of hotspots were presented in Figures 4.11 and 4.12, emphasising validity of the algorithms introduced in addition to indications of generalisability to higher dimensions in clustering special data. Experimental results demonstrate that the new methodology perform reasonably well for several real datasets.

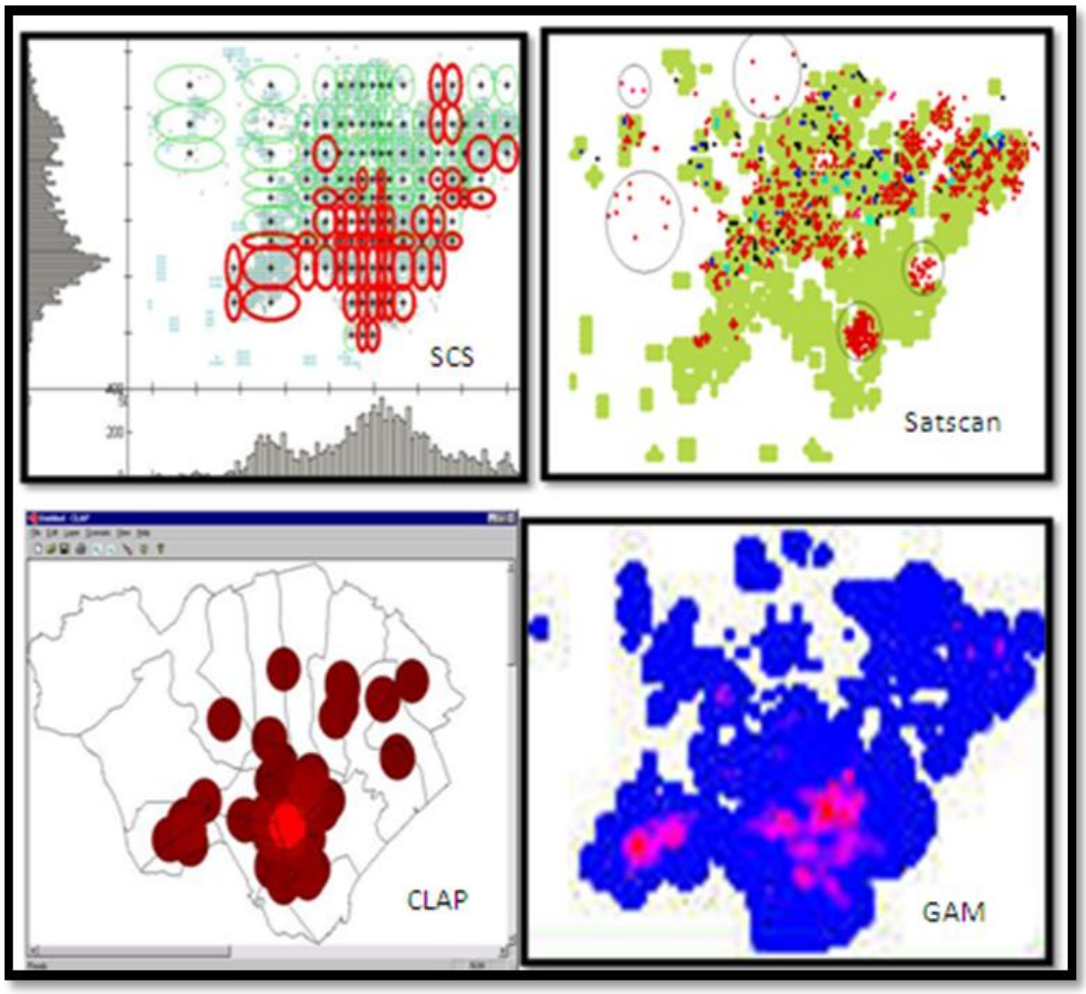


Figure 4.13 Shows the location of clusters. Using the clustering algorithms SCS, satscan, CLAP and GAM.

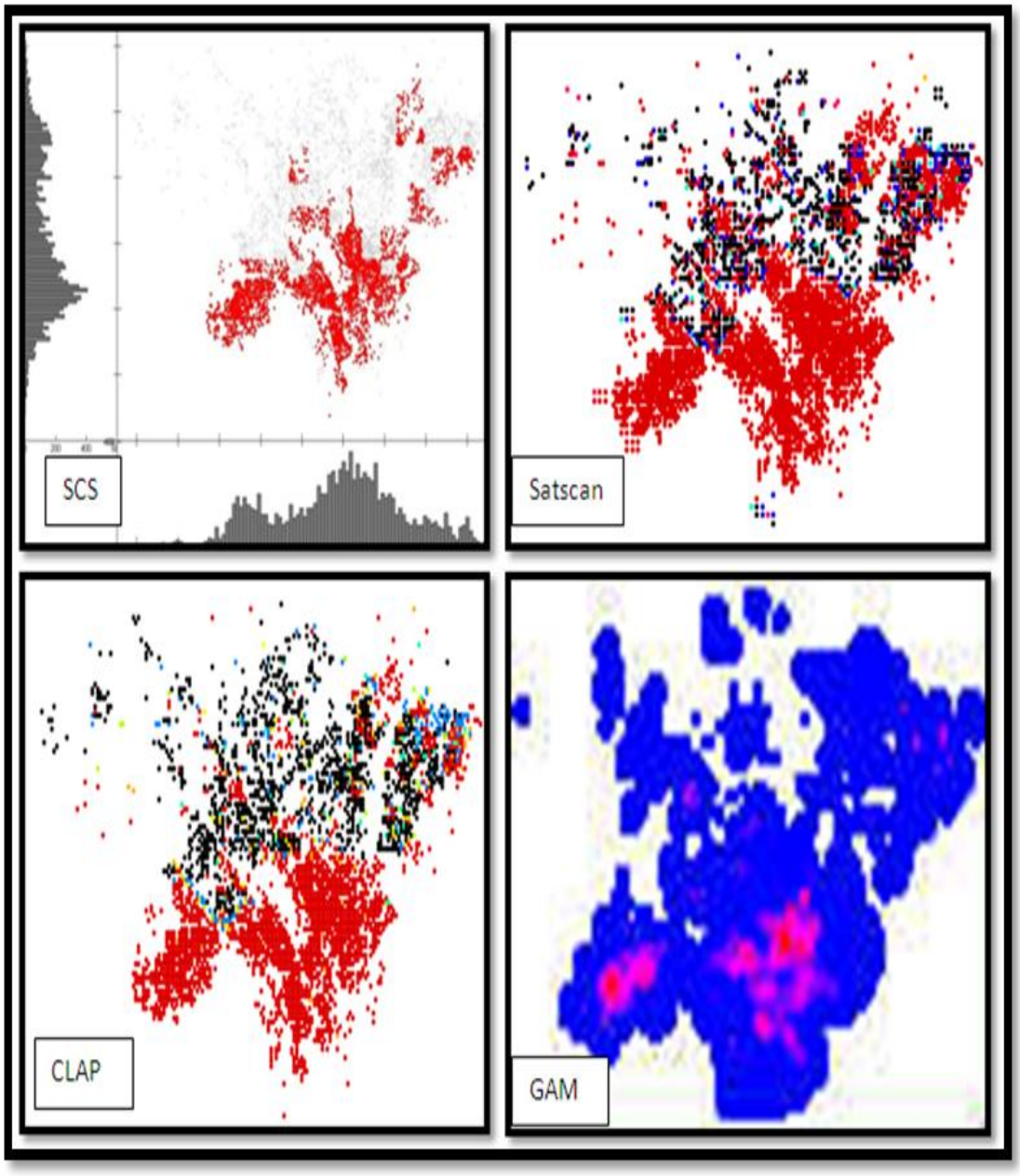


Figure 4.14: Shows the distribution of burglary incidence concentration (hotspot). Using the clustering algorithms SCS, Satscan, CLAP and GAM. Hotspots red area showing.

4.9 Summary and Conclusion

This chapter presents the application of clustering methodology (SCS) on a real (crime) dataset. In addition, results are presented for rotated data (30 and 85 degrees) which was used for testing the effectiveness of the SCS algorithm. The experimental results demonstrated that the SCS algorithm performs well. The promising results obtained for identified hotspots were compared with other available algorithms such as CLAP, Satscan and GAM. The outputs are very similar. The SCS algorithm performed reasonably well in terms of: memory requirements; running time; clustering quality. Therefore, this approach has an improved time complexity, performance and the best quality clustering with the available memory.

Experimental results demonstrate the effect of the new algorithm for large datasets. The work to date has demonstrated a significant theoretical contribution to knowledge in the field of statistical analysis of spatial datasets. It is necessary in several geographic applications to account for a spatially variable background population. It is an efficient and objective algorithm, leading to improvements in statistical data analysis.

The next chapter described the combines of SCS algorithm with Geographical Information Systems (GIS).

5 Utilization of GIS: Crime analysis

This chapter identifies various methods in which GIS have been utilized to interpret burglary distribution within the study area. The new cluster detection technique (SCS) accommodates Geographical Information Systems (GIS) in terms of generating predictive crime models, mapping, displaying distribution of crime in the study region, identifying the total number of crimes within a polygon and integrating information from a variety of sources associated within the observed data.

5.1 Introduction

Since 1990, the extensive usage of GIS has enabled police forces to map and analyse crime data efficiently. Crime mapping is well established in England and Wales in many police forces. Hirschfield, 1999 reported that around two-thirds of police forces and one-third of local authorities in England and Wales had access to a GIS (Home office, 2007). Currently GIS is a standard tool for crime analysis. Several police departments are in the process of implementing GIS. GIS is often credited for providing a valuable analytical tool for the identification and analysis of crime problems as well as the development and assessment of crime. GIS functionality has become widely used in many areas within crime data analysis, such as crime hotspot mapping and cluster detection, repeat victimization and temporal pattern analysis of crime incidents (Wise 2007). GIS has enabled integration of data (both crime and crime related) from a variety of sources, used in terms of crime prediction. Craglia et al. (2001) reported the strengths of GIS-based spatial analysis with census data for modelling high-intensity urban crime area. Hirschfield and Bowers (2001) summarized extensive research contributions of GIS and their practical potential in crime data mapping and analysis. GIS can be used to measure the extend and type of problem within a certain distance a round a particular location. Community characteristics (for example, markets, colleges, parks, alcohol permit locations) can

be routinely displayed on maps while analysis crime patterns to interpret relationship between these characteristics and the crime. For example, crime clusters round locations where alcohol is sold (Johnson, 2000).

5.2 GIS and Crime mapping

Crime mapping is the ''process of using a geographic information system, to conduct spatial analysis of crime problems and other police-related issues'' (Rachel 2005: 37). Crime analysts used maps to communicate analysis results; visualize; analyse the relationship between criminal activity and indicators of disorder; examine patterns of crime at and around specific locations, such as schools and bars. Crime mapping was established before computers were invented. The availability of GIS enables numbers of police departments to experiment with crime mapping in their work. It allows crime analysts to link various types of data source together based on common area, for example, linking census information and crime data for a common area. Presenting this information as layers, gives the analysts the ability to analyse multiple layers of information. Geographic data is available in electronic format such as street and census information; the links between GIS and databases have enabled analysts to create visual images of various types of data in map format. Presenting data in the form of a map is helps visualise the significance of the where, when, and by whom.

Capabilities of GIS for, storage, management, integration, and manipulation of various layers of data helped to advance the field of crime mapping. This has improved the efficiency of police activity.

5.3 Data Characteristics

The detailed information of the spatial dataset utilized by GIS in this chapter is presented in sections 4.3, 4.6.1 and 4.8.1. This included the distribution of 10905 residential burglary incidents reported to police; its rotation within 30 and 85 degrees; 294310 population data was downloaded from CASWEB (CASWEB, 2009);

and 7377 cases out of 10905 incidents are related to hotspots obtained from SCS algorithm. Nine hundred fifty four polygons (census wards) in the area under study are covered within population data, residential burglary and its rotation and 330 polygon within 7377 cases of high level burglary rate(hotspot). The data are represented as points on a map that fall within the boundaries of the polygons. Since this study is based on crime, these data provide for building a predictive crime model, which are presented in chapter 6. GIS has the ability to associate xy coordinates in the map with the address of an incident. This concept is known as geo-coding. Table 5.1 presents typical input information relating to these geo-coded data, crime data and population size that has been used in this analysis. Geographic regions within the study region have different characteristics and therefore lead to different population densities and crime levels (see Figures 5.2 and 5.7). The population size -based approach in this analysis, contributes to measuring crime rate for each polygon. The results were used throughout the process of building crime model (chapter 6). The second use of population size in this analysis is to examine crime rates within variant population sizes. It is essential to know what percentage of the population had suffered from crime in that location.

Table 5.1: A typical input file. Columns refer to: x co-ordinate; y co-ordinate; number of crime; population size.

Crime data			**Population data**		
x	Y	Number of crime	x	y	population
322814	181972	1	320100	178100	76
323090	180985	1	311700	177900	22
315359	182932	1	311900	177900	105
322676	178611	1	312100	177900	22
317947	175632	1	312900	177900	61
319981	177755	1	323300	181100	77
319891	176965	1	323500	181100	63
323938	180850	1	323700	181100	225

5.4 Utilization of GIS in this study

GIS is a computer assisted system for the storage, integration, analysis and display of geographic data. GIS has several advantages in crime analysis. Law enforcement agencies used GIS technology to prevent crime and to co-ordinate policing activities. This includes facilitating plotting the details of an incident, the time, date of the crime and the statistic against the map. Crime analysts use GIS for analysis of large quantities of data to interpret quickly, understand and easily share and visualize data in many ways that reveal relationships. GIS technology has the ability to separate information in layers, and then combine it with other layers of information.

GIS is used as a tool in this thesis to accommodate the new cluster detection technique (SCS) in terms of a predictive crime model. Mapping, displaying distribution of crime in a study region, counting number of points within polygons and integrating information from a variety of sources associated with the observed data (crime data, population data, and census data) are specific applications of GIS described in detail in the following sections.

5.4.1 Mapping

GIS is most often associated with a map. Maps are visual, stimulate the imagination and present the world as simpler and more orderly. Maps have played an essential role within the field of criminology, by spatial representations of crime data. This includes plotting the details of an incident, the time, date of the crime, statistical analyses of the spatial nature of crime and other crime reports. GIS is the best tool for understanding spatial patterns of criminal activity.

Once 10905 cases (burglary incidents) in the area under study have been geo-coded with points, three types of maps were created. A map which shows the distribution of 10905 residential burglary incidents as points on a map reveal the location of burglary incidence that cover 29 parcels (954 polygons) of the study area (see

Figure 5.1). A map in Figure 5.2 shows the distribution of the corresponding population of the area under study. The population of this analysis serves as a background of crime data in the same area that contributes to measuring crime rate per each polygon. Polygons are used since crime rate use an underlying population with which to calculate a rate and this necessitates the use of administrative boundary units, such as polygons. The results were used throughout the building crime model process (chapter 6). The map in Figure 5.3 was obtained by utilizing the SCS result as shown in Figure 4.6. The map shows the distribution of burglary incidence concentration (hotspot). Hotspots are areas of more likelihood of burglary than others. The hotspots in Figure 5.3 were shaded with different colours. Highlighted areas indicate that their crime rates are in high level. Displaying concentrations of criminal activity has allowed law enforcement to examine criminal phenomena within areas of concern and arrange for control and prevention accordingly. This is one of the most common applications of a GIS for crime analysis.

Figures 5.4 and 5.5 are show the maps which are obtained by utilizing rotation of the data points. As mentioned in section 4.8.2, the same spatial dataset (crime) has been rotated for instance, within 30 and 85 degree and SCS algorithm was utilized in order to determine the clusters. The results in Figures 5.4 and 5.5 show that the distribution of high crime rate using rotation data point with 30 and 85 degree respectively in the area under study. The results show that these areas of concern are the same as in Figure 5.3. This indicates that the algorithm performs well.

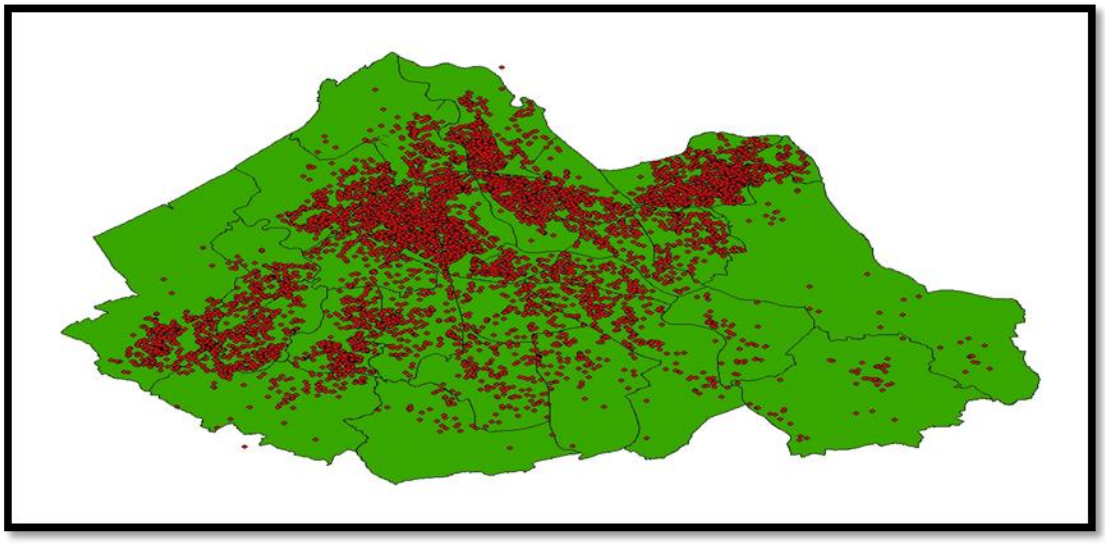


Figure 5.1: Description of distribution of burglary: 10905 cases, over the period 6 February to 31 October 2003 in the area under study, used to identify concentration of crime.

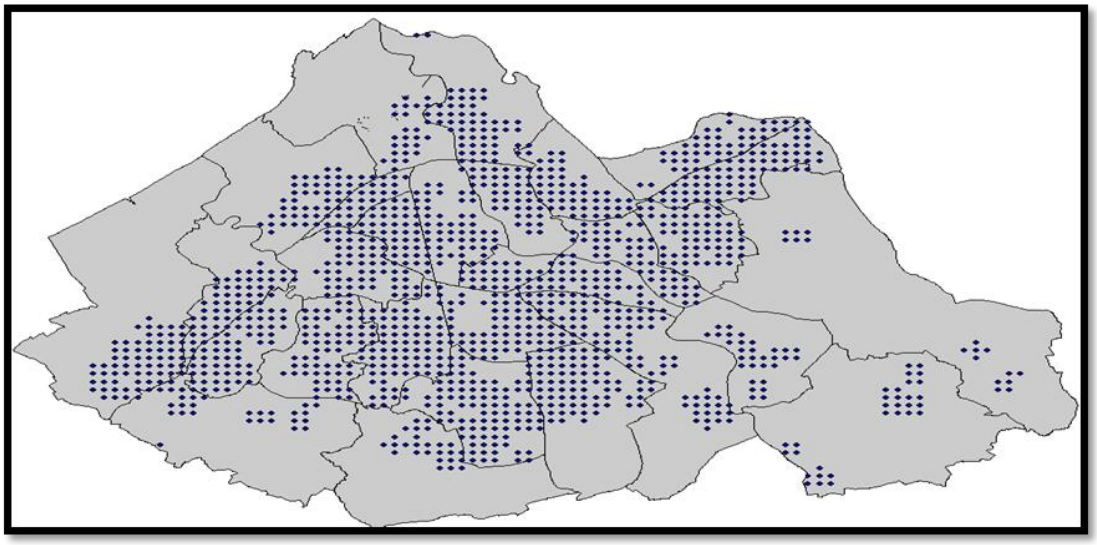


Figure 5.2: shows the corresponding population of the area under study. Its serves as a background of crime data for measuring crime rate within each polygon.

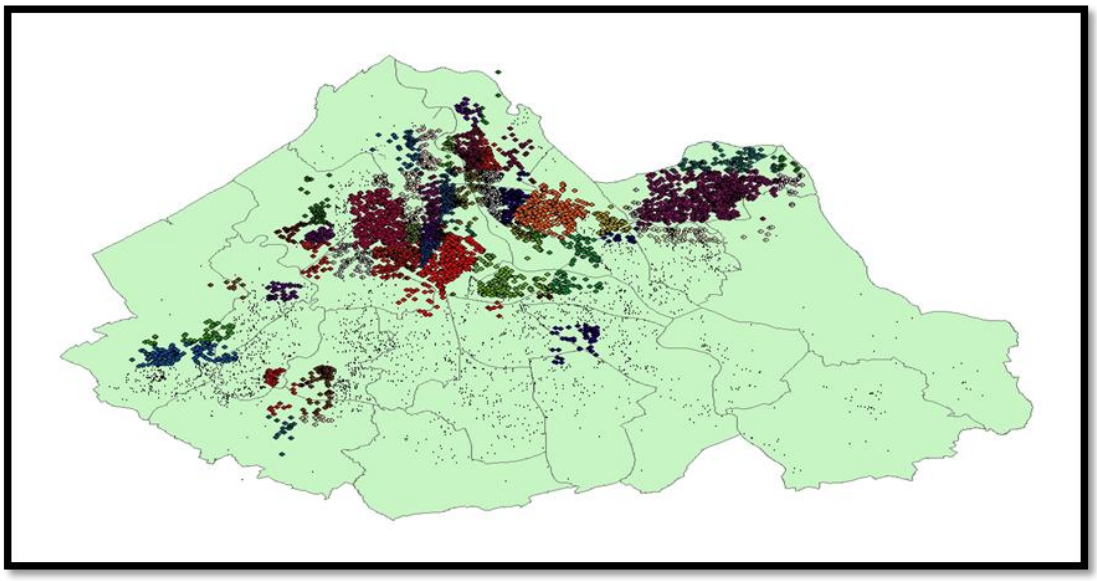


Figure 5.3: Description of distribution of high crime rate clusters in the area under study utilizing the SCS algorithm. The hotspot locations are shown coloured.

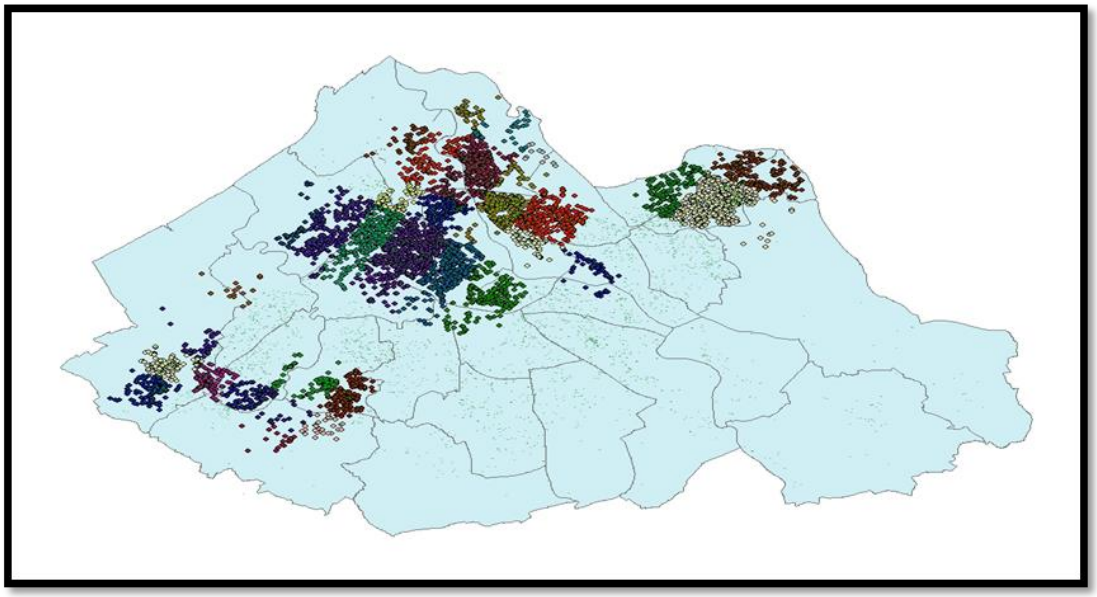


Figure 5.4: Description of distribution of high crime rate (hotspot). The hotspots exhibited are coloured. The map obtained by utilizing rotation data point with 30 degree. It is clear that these concern areas are the same as in figure 5.3.

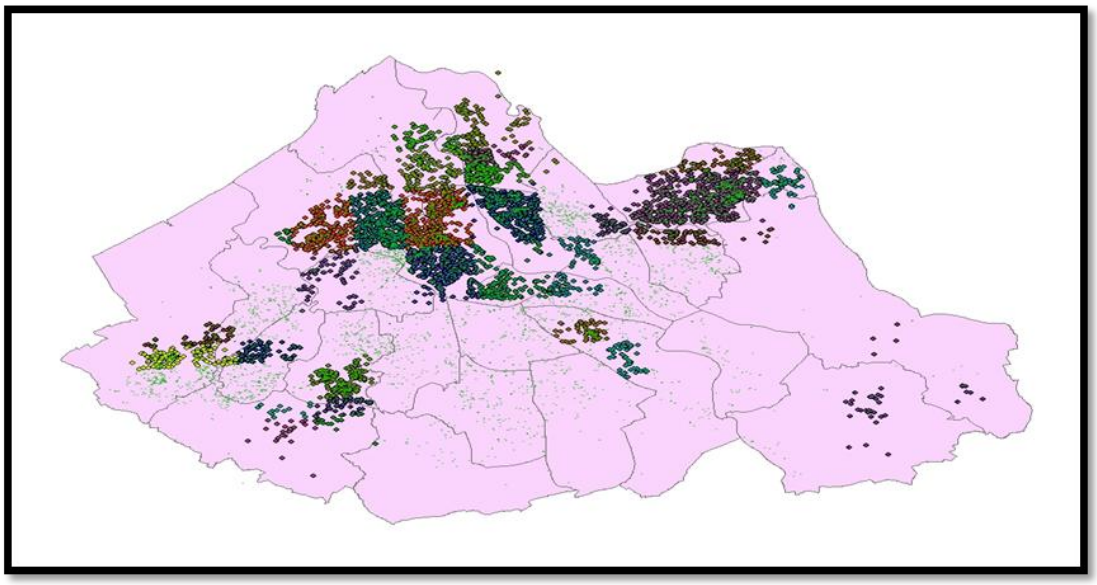


Figure 5.5: Description of distribution of high- crime rate. The hotspots exhibited are coloured. The map obtained by utilizing rotation data point with 85 degree. It is clear that these concern areas are the same as in figure 5.3.

5.4.2 Counting number of crimes within the polygons

The burglary rates in this analysis are expressed as the burglary incidence per number of households in each polygon (census wards). Figure 5.6 shows the distribution of 10905 burglaries in 954 polygons in the study area represented as points on a map that fall within the boundaries of the polygons. Spatial analysis was then conducted using a GIS to identify the total number of crimes within each polygon. The procedure used for conducting this analysis was found in the support section of the ESRI Website (ESRI, 2009). When the total number of crimes within each polygon had been computed, the results were combined with data showing the total number of households in each polygon. The household data was downloaded from the CASWEB Website (CASWEB, 2009). The results of this analysis were then used to obtain the information about the distribution of residential burglary rate within each parcel in study region. Numbers of crime that was obtained within each

polygon transformed into crime rate by division with the number of households within combined polygons. From the census information (CASWEB, 2009) each parcel include numbers of well defined of polygons. Consequently, crime rates within each parcel in the study region were determined. Histograms are used to display the distribution of burglary incidents cross 29 parcels in the study region (see figure 5.7). Figure 5.7 shows the variation in the distribution of actual burglary rate. Classifications of crime levels for middle and low levels are that suggested in this analysis which fall in the polygon are: middle level: [0.0174, 0.0374); low level: [0.0025, 0.0174) but for high level: [0.0374, 0.4] are obtained from identification of hotspot Section 4.6.1. Figure 5.7c shows a typical classification against the polygons in the study region. Classification level of all parcels in the study area can be found in Appendix C. The analysis of level of crime rate will be explained later in section 6.8.3. The results of this analysis were used throughout the model building process in order to predict the spatial distribution of residential burglary in study region (see section 6.8.3).

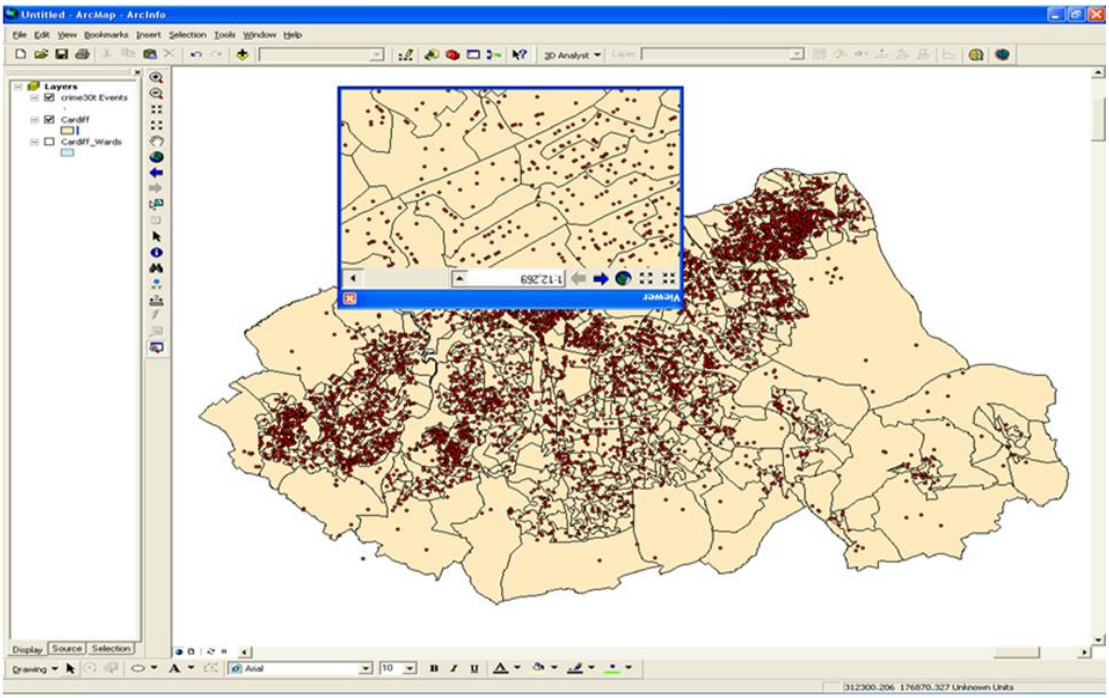


Figure 5.6 shows the distribution of 10905 burglaries in 954 polygons in the study area represented as points on a map that fall within the boundaries of the polygons.

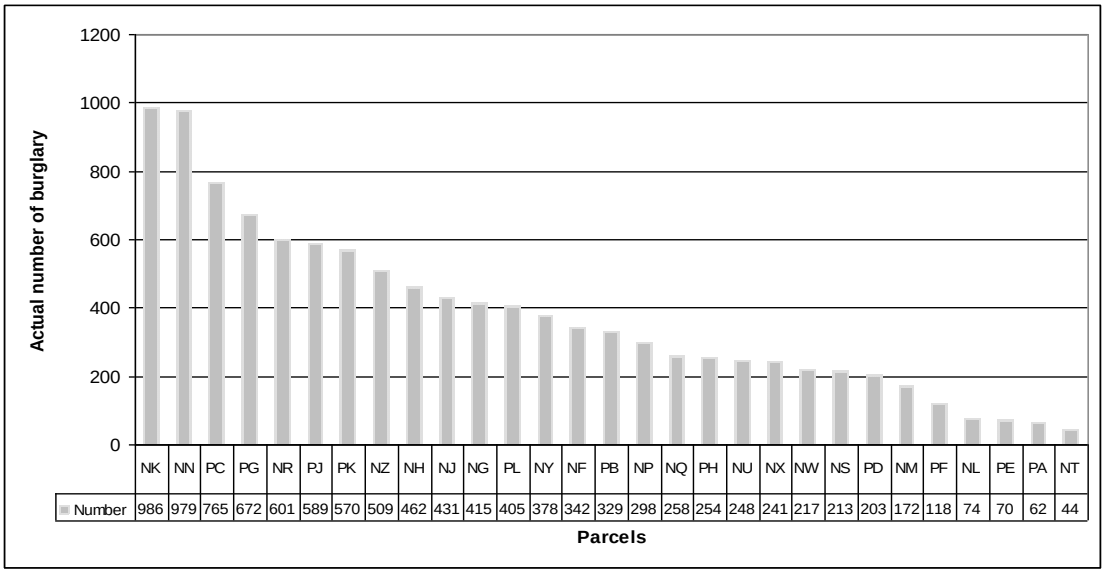


Figure 5.7a: A summary of the number of actual burglary incidents distributed within 29 parcels in study region.

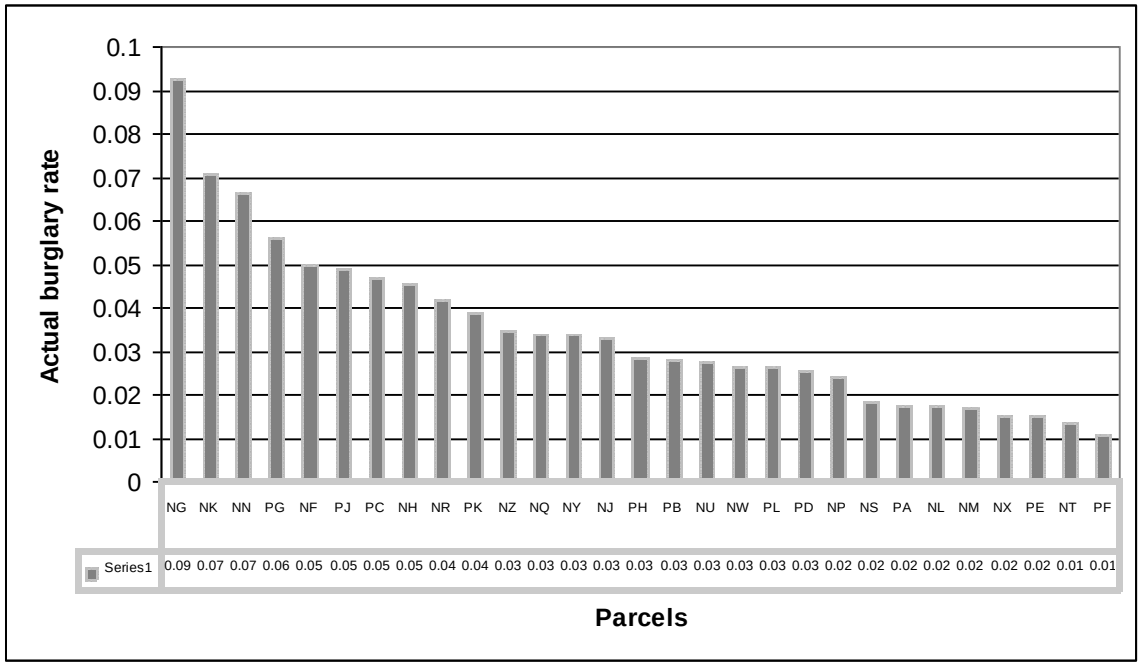
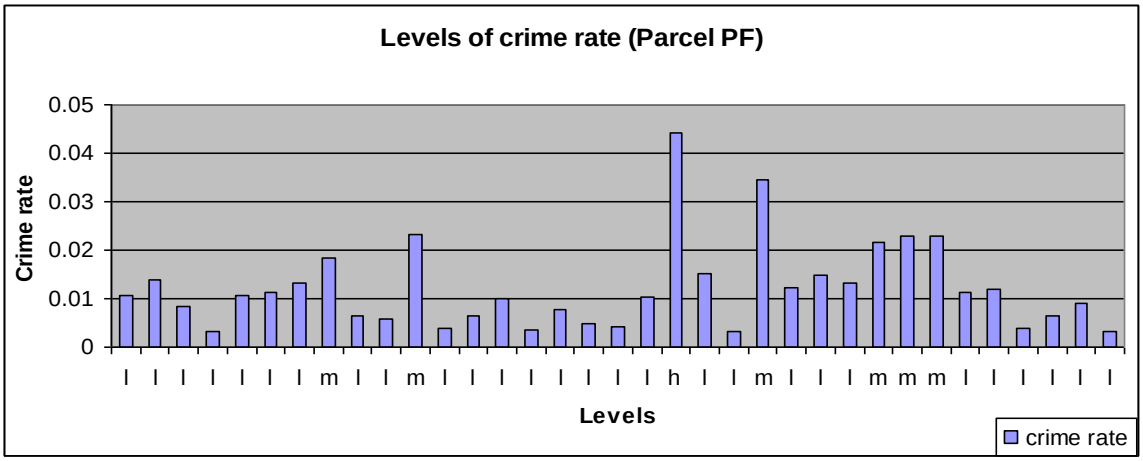
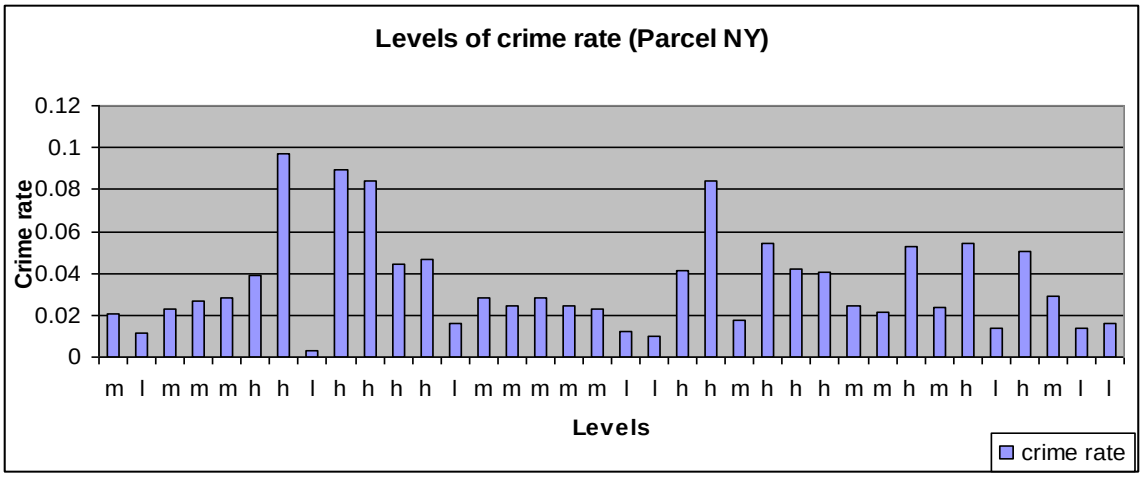


Figure 5.7b: A summary of the distribution of actual burglary rates within 29 parcels in study region. Burglar rates expressed as the burglary incidence per number of households in each parcel.





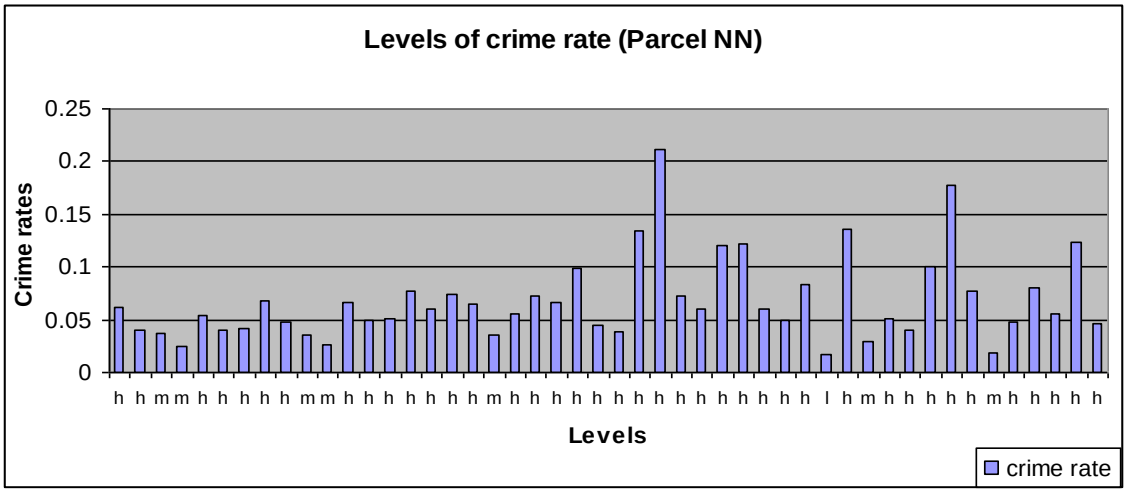


Figure 5.7C: Illustrate the levels of crime rate for some selected parcels in the study region. The levels are high(h), middle(m) and low(l).

5.4.3 Integration of different data layers

One of the fundamental aspects of GIS is the integration of different data layers. The information about a given region can be arranged as a set of maps. Each map displays the information of one characteristic of the region. Thematic map separation is referred to as layers. Each layer overlaps on the others, location matched to its corresponding locations on all the other layers. In the case study presented in this thesis this utility was used to integrate information from a variety source of data. For example, population data and census data that is associated with crime data. The three separated layers shown in the Figures 5.1, 5.2 and 5.3 displayed different information. In the case of a predictive crime model, the analysis requires combination of these layers; layer (population data) with layer (crime data), so every location in layer 1 matched to its corresponding location in layer 2. These two types of data combine in this case for crime rate measurement within each polygon. The location of the bottom layer in Figure 5.8 represented a hotspot. The crime rate in this layer is greater than crime rate in the study region. This layer will integrate with census data for building a predictive crime model. The census data provided a useful source of geo-demographic information. The census is a key source for information about economic, social and demographic composition of small areas. It provides the linkage between population, household, and dwellings at small areas. Thus helping to understanding local economic factors and communities. Linking the available data source was one of the important problems in the spatial analysis. The increasing power of GIS a researcher can easily combine census attribute data with available information for analytical and model purpose for such geographic levels. Awareness of using GIS in the production of census data is growing in all statistical agencies, and simplifies the process of combining census geographic data with attribute data.

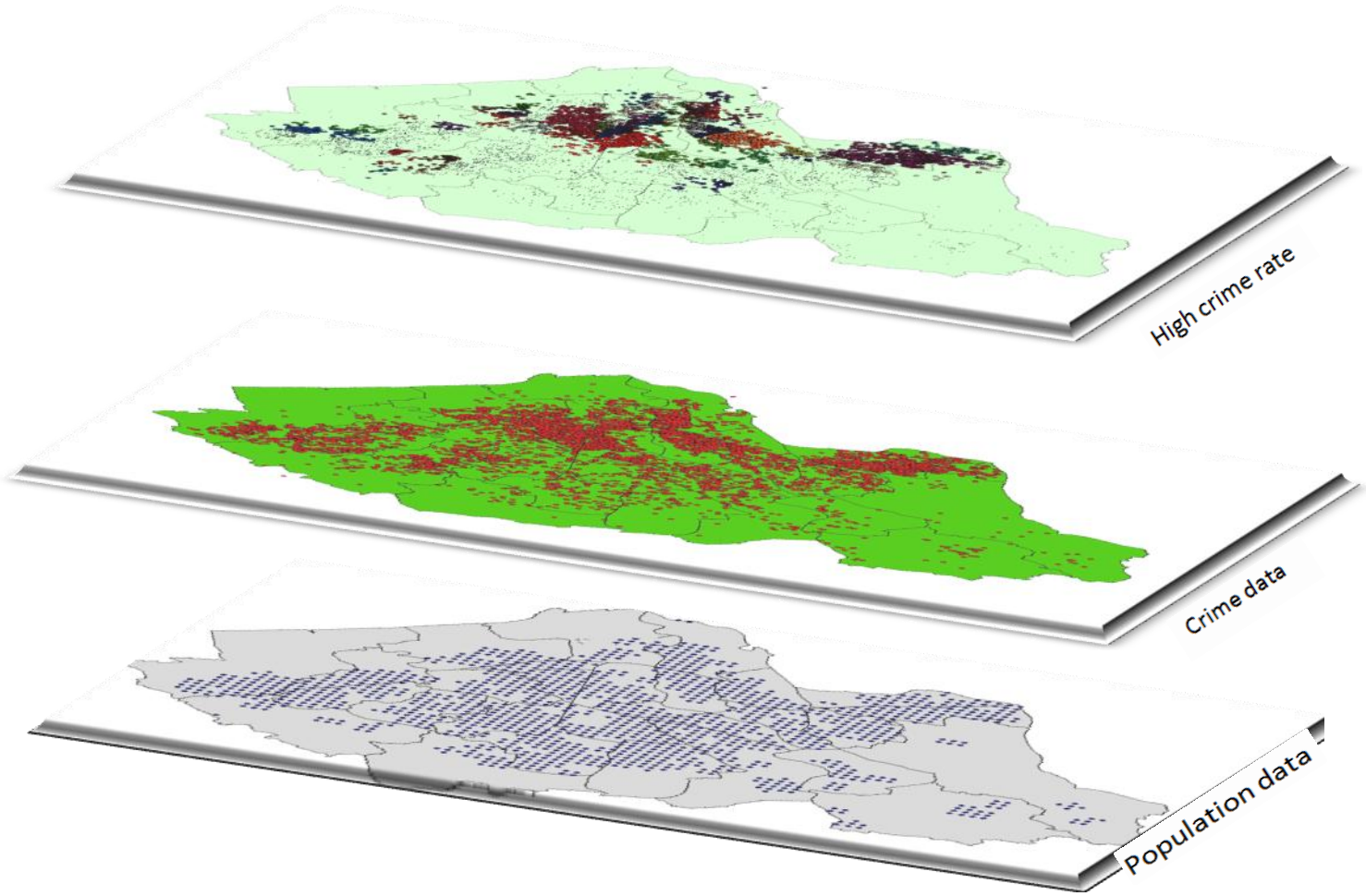


Figure 5.8: The concept of adding layers of geographic information

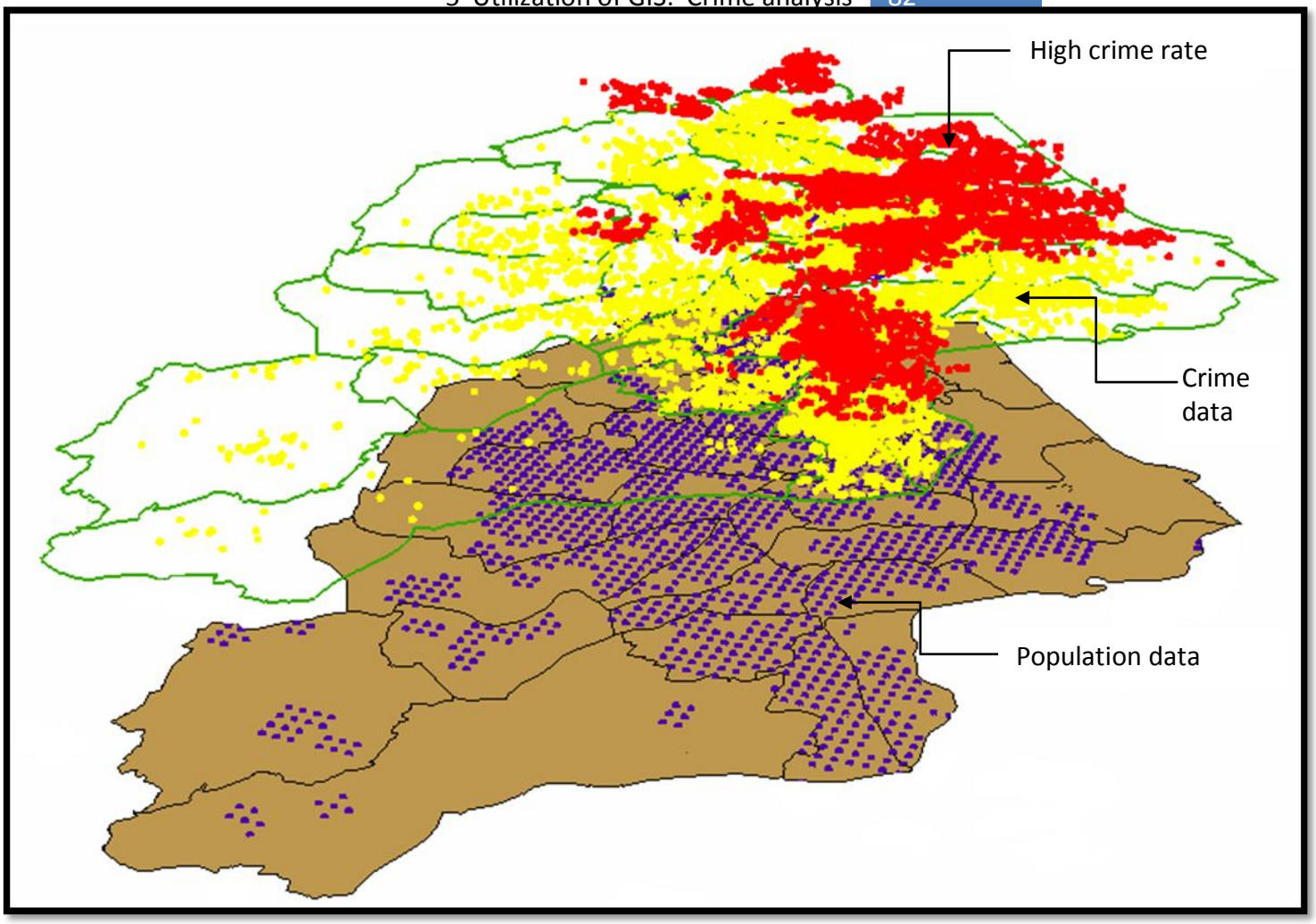


Figure 5.9: Illustration of the utility of GIS to integrate information from a variety of sources such as population, crime data, and high-crime rate in the area under study. 'Integration of figures 5.1, 5.2 and 5.3 '.

5.5 Geographical location of clusters

The predictive crime modelling across the clusters in this thesis is employed 28 potential explanatory variables of characteristics of burgled households. Cluster boundaries (ellipse) were identified using the SCS algorithm (section 4.4). That followed the specification of x and y coordinates of each center location and the limiting points of the major and minor axes of the obtained ellipse from which the geographical location of the obtained clusters were identified on the map of the study region (Figure 5.10).

Polygon (ZIP code) is the smallest spatial unit for which the entire UK Census is publicly available. For example, the parcel (Census tract) NN included 48 polygons. Therefore polygons were selected for spatial aggregation of crime data with the potential characteristics of household data. Thus each cluster includes a set of polygons to be examined. The characteristics of households within each polygon were downloaded from the CASWED. Residential burglary data were then used to calculate rate of burglaries per household within each polygon.

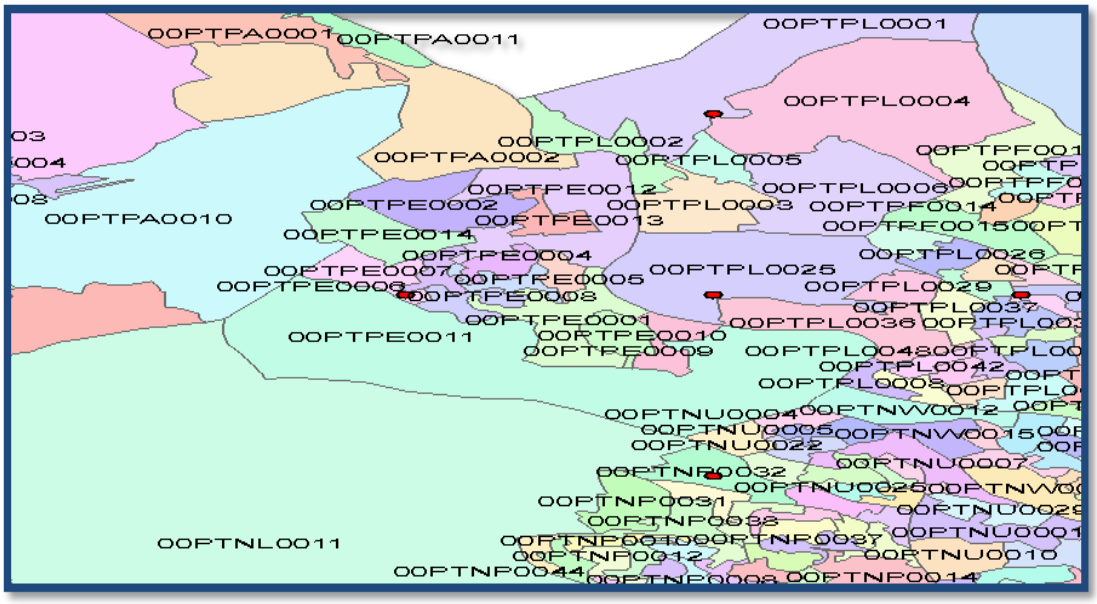


Figure 5.10: Shows the Location of x and y coordinates of center location and the limiting points of the major and minor axes of the obtained ellipse of the cluster in the study region (red point showing).

5.6 Summary and Conclusion

GIS has become one of the most important developments in crime analysis. By combining geographic principles and geo coded location data with crime data and criminological theories, GIS allow the analysis of crime incidents across time and space.

This chapter introduces construction and interpretation of maps using 10905 cases (burglary) that were collected by a police authority in the area under study. Maps are obtained by using the capability of GIS for mapping and organizing data into specific layers, integration of different data layers and determining how many crimes were reported within boundaries of the polygon (census tracts). The resulting rate is combined with multiple data sources which identify in chapter six in order to develop a predictive crime model.

6 PREDICTIVE MODELS

This chapter presents *the building of hybrid predictive models for crimes based on real data. The spatial distributions of residential burglaries are chosen as the foci of this analysis. The model required aggregation of multiple data sources, specifically: crime; population; and census (multivariate) datasets that have been collected. Burglary rate data constituted the response variable in the model. The SCS result algorithm was utilized to identify the spatial level of concentration of burglary incidents in the study region. Both the regression methodology and neural networks have been used for predictive crime modelling. A new methodology, hierarchical neural network for building a predictive crime model is proposed. In this case, statistical methodology combined with ANN to generate a more accurate prediction. Analysis and the results are presented.*

6.1 Introduction

A predictive model is "a simplified representation of reality, comprising a set of relationships, historical information on these relationships, and procedures to project these relationships into the future" (Douglas, 2001: 21).

The purpose of police crime prediction is to support crime prevention and law enforcement. It is an exciting new area, which brings together the disciplines of statistics, machine-learning, criminology and psychology, and database technology. Prediction of crime has recently seen an increase of research and is widely practiced by police. This is due to the establishment of the ecology of crime and using the capabilities of GIS to produce maps depicting crime "hotspots". It enables integration of information from a variety of sources associated with crime data.

In this thesis, spatial distributions of residential burglaries are chosen as the foci of analysis. Burglary, or 'breaking and entering', is one of the most common crimes worldwide. Burglary is a high-volume crime affecting around one in twenty- five households annually in the UK (Home Office 2008).

Several studies have explored the relationship between burglary and possible contributing factors, for instance, poverty (Olligsclaeger 1993); unemployment (Edmark 2005) and house type (Bowers 2004). The predictive crime modelling in this thesis employed 28 potential explanatory variables of characteristics of burgled households, burglary rate constituted the response (dependent) variable.

6.2 Data Representation

Crime is a complex phenomenon. Thus the predicted model required aggregation of multiple data sources, specifically: crime, population, and census (multivariate) datasets that have been collected. The detailed information of these datasets is presented in sections 4.3 and 6.2.2. Census data are a potentially valuable resource for crime analysis. Crime is a social phenomenon and census is the principal national social data resource. Census data describe the characteristics of the population of the UK, and is available via CASWEB (UK Census Website). The recorded burglary incident datasets are derived from a local police authority records.

6.2.1 Data on Burglary

10905 spatial data of actual burglary incidents (Section 4.3) within 7373 hotspot incidents identified by the new cluster detection technique (SCS) in the study region (Figure 4.6) were used in building a predictive crime model. The total numbers of burglaries within each polygon had been computed previously (Section 5.4.2). The results combine the data on census and are then used in a predictive model. Histograms are used to display the distribution of burglary incidents across 29 parcels in the study region (see Figure 5.7).

6.2.2 Data on Census

The National Statistics Office in England and Wales planned one day every ten years for the census, a count and information about all people and households. The latest census was held in 2001. The census outputs are available in a number of forms, especially those covering: Aggregate Statistics, National Statistics Postcode

Directory, Micro data, Interaction data and Digitised boundary datasets. The census data used in this study are aggregate statistics which provide the most complete source of information that is available about the demographic and socio-economic characteristics of the UK population. The aggregate statistical outputs from the 2001 census are available in three main datasets: key statistics (used in this study), standard tables, and census area statistics (UK Census Website).

The characteristics of household data within each polygon were downloaded from the CASWEB Website[1] (CASWEB, 2009). The detail information about these potential characteristics of burgled household can be found in Appendix D. Table 6.1 identifies the potential factors chosen to build a burglary predictive model.

6.3 Neural Networks and statistics

While there are many differences between Artificial Neural Networks (ANNs) and statistics, there are also many similarities. Many ANN models are similar or identical to popular statistical techniques such as generalized linear models, polynomial regression, discriminate analysis[2], and cluster analysis. There are also a few ANN models, such as learning vector quantization, and self-organizing maps, that have no precise statistical equivalent. Neural networks and statistics are not competing methodologies for data analysis; although there is considerable overlap between the two fields.

Statistical methodology is directly applicable to neural networks in a variety of ways, including estimation criteria, optimization algorithms, confidence intervals, diagnostics, and graphical methods. Better communication between the fields of statistical techniques depends on the problem to be solved (Basheer & Hajmeer 2000).

---

[1]Select the 'CAS ward' from the selection 'output geography' then select 'Ethnic Group' from the list of tables that are contained within the key statistic.

[2]Discriminate analysis is a technique use to build a predictive model of group membership based on observed characteristics of each case.

Table 6.1: description of characteristics of households' potential explanatory variables

Age structure	All people		People aged						
			Less than 16	between 16 and 59	greater than 60				
Occupation groups	People aged 16-74 in employment working as								
	Managers and senior officials	Professional occupations	Associate professional and technical occupations	Skilled trades occupations	Personal service occupations	Sales and customer service occupations	Process plant and machine operatives	Elementary occupations	Administrative and secretarial occupations
Qualifications	People aged 16-74								
	No qualifications	Highest qualification							
	People aged 16-74								
Socio Economic Classification	Large employers and higher managerial occupations	Lower managerial and professional occupations	Intermediate Occupations	Unemployment					
Household spaces and accommodatio--n type	All households spaces which are of accommodation type								
	Whole house or bungalow			Flat, maisonette or apartment					
	Detached	Semidetached	Terraced (including end terrace)	Purpose built block of flats or tenement	Part of a converted or shared house (including bed - sits).				
House hold composition	Household comprising								
	One person	All pensioners	Married couple households	Cohabiting couple households	Lone parent households				

6.4 Artificial Neural Network and Multiple Linear Regression Models for Prediction

Building the predictive crime models which are presented in this thesis adopted both ANNs and statistical techniques. Both of these techniques have advantages and disadvantages. The objective of this study is to take advantage of both techniques for best performance of a predictive crime model. Multiple linear regression (MLR) analysis is the statistical technique used in this study to identify potentially significant predictive variables and the level of their contribution in the performance of the model and predicting of future crime. Every statistics computer software package contains a regression component. However, Neural Network analysis generates weights, which are difficult to interpret as they are affected by the program used to generate them. This drawback is one of the most criticized features in neural network models. The main advantage that neural networks offer is they do not require many assumptions before the model can be constructed, for example, nonlinearity. When applying regression, the user must have detailed knowledge about the appropriate non-linear relationship between the input and output variables. However, when applying neural networks, these relationships are determined implicitly by the model, since ANNs are non-linear systems. This property includes robust performance in dealing with noisy data. Another advantage of the ANN approach is that the model allows the inclusion of several input and output variables at the same time. This requires more care with a regression technique.

In this study statistical techniques start their construction model with 28 potential explanatory variables identified as input chosen among characteristics of burgled households. These include Resident Population, Occupation, Qualifications, Socio Economic, Household composition and Household spaces. The identification was shown in Table 6.1. The spatial distribution of actual residential burglary incident rate constituted the response (dependent) variable. The data standardized into rate per household within each polygon (census wards). A hierarchical neural network (HNN) starts their construction model with a number of explanatory variables with

statistical significance which were obtained and identified from regression model results.

6.4.1 Multiple Linear Regression Model

Statisticians and criminologists have been applying their skills and knowledge for a long time to predict when and where the next set of crimes will occur, with varying degrees of success. Multiple linear regression (outlined in section 1.1.4) was chosen in this thesis for a predictive crime model. The regression model is one of the popular methods of modelling and prediction. Regression analysis has major purposes: description, estimation and prediction. This methodology has wide applicability in prediction in a variety of areas (Bolzan 2008; Margaret 2002; carcoran 2003).

6.4.1.1 Regression Modelling Steps

This strategy involves:

- Data collection and preparation;

Any regression analysis to perform the analysis requires data. The characteristics of the data vary with the nature of the study. In the procedures of data collection and preparation it is important that the user is conversant about the theory associated with the subject being analysed. Section 6.2 shows details of the data characteristics which are used in the regression model presented in this thesis.

- Selection of Explanatory variables;

Selection of explanatory variables is an important aspect of regression analysis. The process of model building is to identify those variables which are significant partial contributors to the prediction of the response variable. Backward elimination and forward selection are one of the popular techniques for variable selection. Backwards elimination procedure begins with a regression on all potential explanatory variables. After the regression is run the explanatory variables are examined to determine which variable is non-significant and to be deleted. The statistical tests, t-test, and p-value are used for this purpose. However, forward

selection procedure starts with no explanatory variables. The first variable included in the equation is the one which has the highest simple correlation with a response variable. The significance of the regression coefficient of the first variable is then tested. The equation includes variables which are statistically significant. A search for a second variable is made in the same way. Variables are considered one by one until there is no significant improvement in the model brought about by adding another variable (Kutner 2004).

The regression analysis presented in this thesis used backward elimination. The procedure starts with 28 potential explanatory variables. The p-value used to identify significance of explanatory variable at chosen significance level was 0.05.

- Estimate Unknown Model Parameters and Interpretation;

In multiple linear regression analysis the least-squares criterion is used to estimate the regression coefficients. The regression coefficients B_i measure the partial contribution of each predictor variable to the prediction of the response. If a predictor x_i is changed by one unit, while all the other predictor variables are kept fixed. Then the response variable y will change by B_i. Sings (plus or minus) of regression coefficients refer to the direction of the relationship between the predictor variable and the response. If the coefficient B_i is positive, then the relationship of the predictor x_i to the response is positive, and if the coefficient B_i is negative then the relationship is negative.

- Test for Multicollinearity.

In multiple linear regression analysis, estimation of regression parameters are unstable and have high standard errors, when a predictor variable is a linear combination of other predictors in the model (Feranadez 2003). Variance inflation factors (VIF) “measure how highly correlated each independent variable is with the other predictors in the model” (Kutner 2004:408). O’ Brein (2007) suggested that VIF ≥5 indicates a multicollinearity problem. VIF can be obtained on MINITAB by selecting the options in the regression dialog box.

6.5 Inference about MINITAB Multiple Regression output

Regression analysis developed rapidly with the increasing power computers. Every statistics computer software package contains a regression component. The quantities SSR, SSE and SST are found in the SS column of MINITAB multiple regression output defined as:

SST (total sum of squares) = SSR (sum of squares due to regression) + SSE (sum of squares due to error),

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{6.1}$$

represents the variation in y "explained" by the regression.

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6.2}$$

represents the variation in y left "unexplained" by the regression. The values of R-sq and R-sq(adj) are indictors of how well the regression model agreement. They are mathematically defined as follows:

Coefficient of multiple determinations(R-sq) is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6.3}$$

Where y_i is the actual value, $\hat{y}_i$ the predictive value and $\bar{y}_i$ the mean value. R^2 has a value bounded between 0 and 1. It represents the percentage of variation in the dependent variable y explained by the regression model. The value of R-sq increases as more relevant terms are added to the model even if the new term does not contribute significantly to the model.

R-sq(adj) is defined as follows:

$$R - sq(adj) = 1 - \frac{n-1}{n-(k-1)}(1 - R^2) \tag{6.4}$$

Where n is sample size and k is the number of parameters estimated.

The R-sq(adj) only increases when significant terms are added to the model.

6.6 Measures of Model Adequacy

Any model designed to predict human behaviour, suffers from predicting error as it cannot include all the variables of influence in addition to the difficulty of measuring some phenomena accurately. Once a model is developed, then checking the model adequacy is an important step for identifying models that best fit the data followed by determining which models are the best for prediction purpose. A number of methods are available to estimate prediction errors. The following three traditional techniques are often used: (1) mean absolute deviation (MAD), (2) mean absolute percent error (MAPE) and (3) mean root squared error (MRSE). They are mathematically defined as:

$$MAD = \frac{1}{n}\sum_{i=1}^{n} abs(y_i - \hat{y}_i) \tag{6.5}$$

$$MAPE = \left[\frac{1}{n}\sum_{i=1}^{n} abs(\frac{y_i - \hat{y}_i}{y_i})\right] * 100 \tag{6.6}$$

$$MRSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{6.7}$$

Where y_i actual value, $\hat{y}_i$ predictive value and n is sample size.

Accuracy=100-MAPE (6.8)

That is how closely the estimates provided by the model conform to the actual events being predicted. The model is highly accurate for predicting when its MAPE value is less than 10. Good predicting occurs when MAPE value is between 10 and 20. Predicting is inaccurate when MAPE value greater than 50 and reasonable when the MAPE value is between 20 and 50 (Lewis 1982:40).

6.7 Validation of Regression Models

The final step in the model-building process is the validation process. Model validity refers to the stability and reasonableness of the regression coefficients, and usability of the regression function. Validation of a model is done by testing the model on another independent data set. There are a variety of methods to examine the validity of the regression model, for instance the collection of new data and data splitting. The purpose of collecting new data is to be able to examine whether the regression model developed from the given data is applicable for the new data. The new data are needed to have similar statistical properties as the model data. The corresponding values of the predictor variables need to be similar to those on which the regression model was fitted. One validation method is to re-estimate the regression coefficients using the new data. Then the regression coefficient of the new data is compared to those of the regression model based on the given data. The chosen regression model is applicable if the results are consistent. A second validation method is designed to examine the predictive ability of a model. The predictive ability of a model measure by using the mean of the squared prediction error (MSPR), which defined as:

$$\text{MSPR} = \frac{1}{n^*}\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2 \qquad (6.9)$$

Where:

Y_i is the value of the response variable in the *i* th validation case

$\hat{Y}_i$ is the predicted value for the *i* th validation case based on the model-building data set. n^* is the number of cases in the validation data set.

The mean of the squared prediction error (MSPR) is needed to be closer to MSE. This implies that the error mean square MSE based on the training dataset is a reasonably valid indicator of the predictive ability of the fitted regression model.

The data splitting method for validation of a model can be used when the dataset is large enough to split into two sets. The first set of data with apportion of say, 80%

of the given data called the estimation (training) data is used to estimate the model coefficients. The second dataset with apportion of say, 20% of the data called the validation or prediction set (test set), is used in the same way as the new data to evaluate the reasonableness and predictive ability of the selected model. This validation procedure is called cross-validation (Kutner 2004: 370).

The predictive burglary models in this study were tested with a new data set. The new data are derived from the parcel with a similar level of burglary rate. During the modelling test, the accuracy percentage was achieved for the model and for each polygon (ZIP code) within the parcels test.

6.8 A predictive Crime Model

Multiple linear regressions were chosen for a predictive crime model. A predictive model used data which were introduce in section 6.2 and the potential predictor which was identified in the Table 6.1. Table 6.2 presents typical input information relating to these data that has been used in this analysis. Backward elimination was used for selection of explanatory variables. The p-value is used to identify the significance of the explanatory variable at the chosen significance level 0.05. The least-squares method was used to estimate the regression coefficients. Minitab was used to implement the multiple linear regression analysis. Statistical methods were carried out with this regression model to identify which of these explanatory variables are significantly correlated with the response variable. Analyse the relationships between all the explanatory variables, taking into account the various correlations that may exist between the explanatory variables. The advantage of MLR for identification of significant predictors within the model, allows several social sciences such as sociology and criminology to address questions such as the contribution of the significance explanatory variables to perform the model; the relationship between the significance explanatory within the model and the response. These results are useful for the development of crime prevention strategies.

The sections below illustrate examination of the statistical relationships between the potential of explanatory variables which was chosen among characteristics of burgled households and burglary rate. The analysis focuses on examining the risk of burglary within the specific clusters, applying the model across a number of geographical space parcels in the study region and then selecting the models within three specific levels of the spatial distribution of burglary rate.

Table 6.2: A typical input file. Columns refer to potential predictor variables and rows represent polygon (ZIP code).

Zone Code	*one person*	*allpensioner*	*cohabiting*	*Loneparent*	*Crime rate*
00PTNN0001	0.122	0.0152	0.0183	0.1159	0.061
00PTNN0002	0.1318	0.0405	0.0135	0.1216	0.0405
00PTNN0003	0.0627	0.0495	0.0297	0.0396	0.0363
00PTNN0004	0.1046	0.0268	0.0295	0.0214	0.0241
00PTNN0005	0.0494	0.0148	0.0296	0.0938	0.0543
00PTNN0006	0.0807	0.0202	0.0202	0.1037	0.0403
00PTNN0007	0.0929	0.0353	0.0417	0.0705	0.0417

6.8.1 Example of Regression Modelling Steps

This section outlines the procedures for building a predictive model of the geographical space, Parcel NN in the study region which was chosen as an example to illustrate the regression modelling steps. The procedure starts with 28 potential of characteristics of burgled households as explanatory variables with an actual burglary rate as the response variable. Table 6.3a shows the MINITAB backward elimination, the first stage of the fitting process. The regression result in this stage includes all of the potential explanatory variables. According to VIF, the explanatory variable, people aged less than 16 have the highest VIF =538.247>5, so it dropped from the model. It was considered a problem to the model. This procedure was repeated at each stage according to highest VIF ≥5. Then complete performance the analysis of the model are with remaining explanatory variables with VIF values <5 (see Table 6.3a). Table 6.3b shows the refitting of the model after excluding the explanatory variables from the model according to VIF>= 5. Backward elimination proceeds by sequentially excluding the explanatory variables that adds the least to the model. The variables were with the largest p-value according to decision rule for

p test. The p-value of the explanatory variables: No Qualifications (NQ); semi-detached (h2) and one person are greater than the chosen significance level 0.05. Thus, according to the decision rule for p-test, these non-significant explanatory variables are excluded from the model which is presented in Table 6.3b. The final results for the regression model of burglary rate on significant explanatory variables are shown in Table 6.4.

Table 6.3a: NN model fit at initial stage of backward elimination for prediction burglary rate. The explanatory variable with the highest VIF displayed using dark colour.

Predictor	Coef	SE Coef	T	P	VIF
Constant	-1.128	1.643	-0.69	0.501	
<16	1.299	1.682	0.77	0.45	538.247
16-59	0.222	1.54	0.14	0.887	245.229
>60	0.718	1.558	0.46	0.65	513.66
ManagersOcc	-0.968	1.012	-0.96	0.351	19.379
professionalOcc	-0.468	1.176	-0.4	0.695	25.933
TechnicalOcc	-1.254	1.068	-1.17	0.255	23.729
AdministrativeOcc	-1.0954	0.95	-1.15	0.263	18.498
SkilledtradesOcc	-0.0033	0.7567	0	0.997	4.305
personalserviceOcc	-0.2066	0.7537	-0.27	0.787	2.284
sales&customerservice	-1.0622	0.6757	-1.57	0.132	2.728
machineoperatives	0.7538	0.9337	0.81	0.429	3.039
ElementaryOcc	0.2496	0.5042	0.5	0.626	3.66
NQ	0.7161	0.4226	1.69	0.106	27.943
HQ	0.8111	0.4598	1.76	0.094	80.121
Hocc	0.776	1.003	0.77	0.449	18.878
Loocc	0.7744	0.948	0.82	0.424	77.267
Meocc	1.1749	0.8601	1.37	0.188	13.116
Unemploy	-0.5949	0.6314	-0.94	0.358	9.414
h1	0.0428	0.735	0.06	0.954	44.532
h2	0.1229	0.7099	0.17	0.864	50.096
h3	0.1919	0.6896	0.28	0.784	72.953
f1	0.5107	0.8696	0.59	0.564	148.264
f2	0.075	1.552	0.05	0.962	5.052
one person	-0.1806	0.7455	-0.24	0.811	92.848
allpensioner	-1.594	1.058	-1.51	0.148	6.876
Married	0.6427	0.7414	0.87	0.397	31.313
cohabiting	0.3771	0.8226	0.46	0.652	4.631
Loneparent	0.5847	0.7352	0.8	0.436	29.232

S = 0.0368744 R-Sq = 65.0% R-Sq(adj) = 13.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	28	0.048046	0.001716	1.26	0.303
Residual Error	19	0.025835	0.001360		
Total	47	0.073881			

Table 6.3b: Shows the MINITAB printout of the NN model. 15 explanatory variables was excluded from the model according to VIF>=5. The model includes significance and non-significance explanatory variables. The non-significance explanatory variables NQ; h2; one person displayed using dark colour.

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.060112	0.005172	11.62	0	
SkilledtradesOcc	-0.25619	0.04383	-5.85	0	3.316
personalserviceOcc	0.2393	0.04474	5.35	0	1.548
sales&customerservice	-1.00508	0.04033	-24.92	0	2.056
machineoperatives	0.45451	0.05174	8.78	0	1.535
ElementaryOcc	0.20353	0.03089	6.59	0	3.39
NQ	0.00957	0.01021	0.94	0.367	3.932
Meocc	0.34263	0.02762	12.41	0	3.496
h2	-0.00809	0.01213	-0.67	0.517	2.387
h3	0.03243	0.01024	3.17	0.008	2.495
f2	0.98616	0.06989	14.11	0	2.195
one person	0.010041	0.009024	1.11	0.288	2.965
allpensioner	-0.30304	0.04543	-6.67	0	2.331
cohabiting	-0.41315	0.03221	-12.83	0	1.601

S = 0.00190156 R-Sq = 99.4% R-Sq(adj) = 98.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	13	0.00671587	0.00051661	142.87	0.000
Residual Error	12	0.00004339	0.00000362		
Total	25	0.00675926			

Table 6.4: Shows the finial MINITAB printouts for the NN model. The model presented the significance explanatory variables for the prediction burglary rates.

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.061541	0.002652	23.2	0	
SkilledtradesOcc	-0.29775	0.03809	-7.82	0	2.104
personalserviceOcc	0.24753	0.04801	5.16	0	1.497
sales&customerservice	-1.00206	0.04199	-23.86	0	1.873
machineoperatives	0.46134	0.05125	9	0	1.266
ElementaryOcc	0.20331	0.02536	8.02	0	1.92
Meocc	0.34052	0.0259	13.15	0	2.583
h3	0.040786	0.009523	4.28	0.001	1.813
f2	1.0738	0.06158	17.44	0	1.431
allpensioner	-0.25255	0.03899	-6.48	0	1.443
cohabiting	-0.4317	0.03315	-13.02	0	1.425

S = 0.00207451 R-Sq = 99.0% R-Sq(adj) = 98.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	10	0.00669471	0.00066947	155.56	0.000
Residual Error	15	0.00006455	0.00000430		
Total	25	0.00675926			

The regression equation: Burglary rate = 0.0615 - 0.298 SkilledtradesOcc + 0.248 personalserviceOcc - 1.00 sales&customerservice + 0.461 machineoperatives + 0.203 ElementaryOcc + 0.341 Meocc + 0.0408 h3 + 1.07 f2 - 0.253 allpensioner - 0.432 cohabiting, can be interpreted as providing an estimate of regression coefficients for a given characteristic of burgled households. This equation, for instance, shows that the burglary rate tends to increase by 0.34 for each unit increase in the predictor intermediate occupation (Meocc) while all the other predictor variables are kept fixed. On the other hand if the predictor all pensioners is increased by one unit while all the other predictor variables are kept fixed, then the burglary rate will decrease by 0.25. The analysis reveals a number of predictors that increase the risk of burglary. Living in a terraced house (h3) and shared house (f2) increase risk. In households where the occupations are personal service or machine operative the

risk is high. In addition those in elementary or intermediate occupations are also at risk of burglary. However these predictors decrease the risk of burglary: living in a household in which there are all pensioners; people cohabiting; people with skilled trades' occupation; these with sales and customer service occupations.

6.8.2 Model applied across the clusters

This study shows the benefit of examining the risk of burglary within the cluster. Cluster analysis is an important approach for the explanation and prediction of crime spatial patterns. Within the cluster criminal phenomena are examined within the area of concern. This is because concern in certain location helps to identify the problem associated to the characteristic of the people within their location, and this leads to increase the predictive accuracy of the model. Clustering was widely used in the field of criminology research for instance; Corcoran and Wilson (2003) used spatial clusters for forecasting crime.

For building a predictive regression model for this analysis, modelling utilized data associated with the characteristics of burgled households and actual burglary rate (outline in Section 6.2) within the clusters. Clusters which were identified by SCS (see Section 4.3) and their geographical locations are shown in Figure 6.1(outline in Section 5.5). The final MINITAB printout of regression models across the specified clusters are shown in Figure 6.2. The results presented the significant explanatory variables for prediction of the burglary rate. The predictive burglary models which were obtained from this analysis are tested with new data set (validation data) (see Figure 6.3). The new data are derived from a similar level of burglary rate, for instance, cluster A tested within data that was derived from cluster F. Clusters A and F were previously identified as ''hotspot'' (section 4.6.1). For testing cluster B data was derived from parcel within cluster C which was identified as a ''cold spot'' (section 5.4.2). The accuracy percentage is one of the most important components of the performance properties of a predictive model. During the modelling test an accuracy percentage was achieved for the model and the polygon (zone) of the

parcel's test (see Figure 6.3). The accuracy percentage results which were identified as 68%,,71%,70%,80%,72% for the clusters A,B,C,D,F respectively indicated that the models were reasonable. This is according to the criteria of accuracy test, formula 6.8 (Section 6.6). The model is reasonable when its mean absolute percentage error value is between 20 and 50 (Lewis 1982: 40).

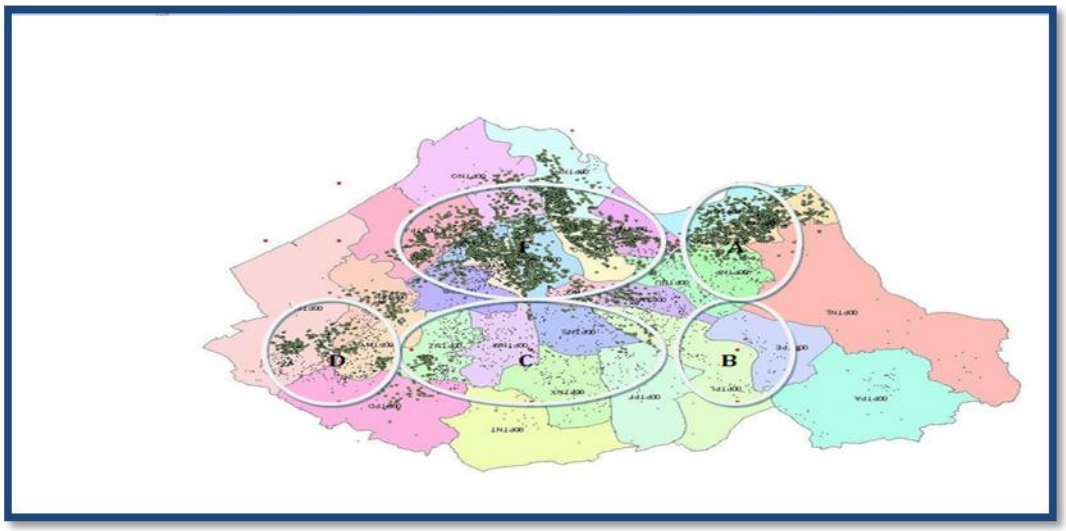


Figure 6.1: Shows the Location of the clusters in the study region using crime dataset. The clustering algorithm SCS was implemented for the specified the clusters.

Cluster A

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.06886	0.01726	3.99	0	
professionalOcc	-0.4281	0.1183	-3.62	0.001	3.806
TechnicalOcc	0.2587	0.1016	2.55	0.013	3.872
machineoperatives	-0.4369	0.133	-3.28	0.002	1.408
h3	0.07801	0.02448	-3.19	0.002	1.678
f2	1.1416	0.2211	5.16	0	1.34
allpensioner	-0.3872	0.1034	-3.74	0	2.197

S = 0.0133760 R-Sq = 66.0% R-Sq(adj) = 59.1%

Cluster B

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.022098	0.008368	2.64	0.018	
professionalOcc	-0.15524	0.03979	-3.9	0.001	2.014
h3	0.05227	0.02284	2.29	0.036	1.694
f2	0.5102	0.2075	2.46	0.026	2.558
one person	0.07572	0.01637	4.63	0	2.1

S = 0.00592016 R-Sq = 89.9% R-Sq(adj) = 83.6%

Cluster C

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.024746	0.007092	3.49	0.001	
TechnicalOcc	-0.10582	0.03126	-3.39	0.001	1.875
sales&customerservice	0.11362	0.04922	2.31	0.022	1.801
Unemploy	0.11583	0.05294	2.19	0.03	1.87
f2	0.24798	0.09986	2.48	0.014	1.238
allpensioner	-0.08793	0.02711	-3.24	0.001	2.853

S = 0.00698176 R-Sq = 44.9% R-Sq(adj) = 38.3%

Figure 6.2: Shows the final Minitab printouts of significance predictors for a prediction regression models across the specified clusters of geographical space in the study region.

Figure 6.2: Continuation**.** The final Minitab printouts of significance predictors for a prediction regression models across the specified clusters of geographical space in the study region.

Cluster D

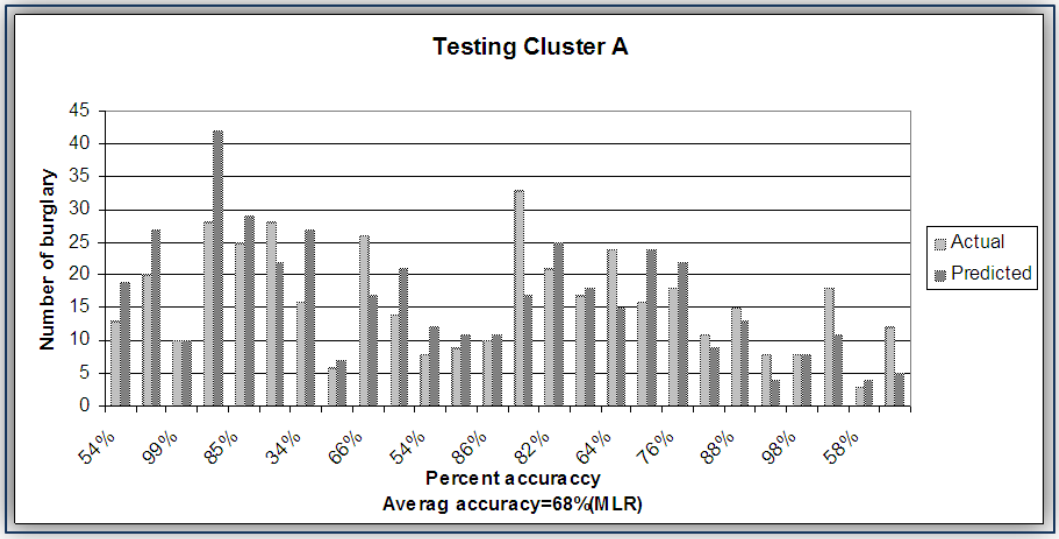
Predictor	Coef	SE Coef	T	P	VIF
Constant	0.008924	0.005441	1.64	0.106	
SkilledtradesOcc	0.2194	0.04867	4.51	0	2.026
sales&customerservice	0.12697	0.04415	2.88	0.005	1.315
Meocc	-0.08266	0.033	-2.5	0.015	2.496
f1	0.07403	0.01582	4.68	0	3.768
Loneparent	0.14946	0.02696	5.54	0	3.63

S = 0.00541410 R-Sq = 77.8% R-Sq(adj) = 73.4%

Cluster F

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.023433	0.005053	4.64	0	
personalserviceOcc	-0.29753	0.06998	-4.25	0	1.281
machineoperatives	0.12636	0.06104	2.07	0.04	1.65
Unemploy	0.0814	0.03254	2.5	0.013	1.935
h2	0.024118	0.009607	2.51	0.013	1.655
f2	0.09261	0.01216	7.61	0	1.702
one person	0.02668	0.01044	2.55	0.011	1.561
allpensioner	-0.18882	0.05103	-3.7	0	2.325
Loneparent	0.19932	0.03477	5.73	0	1.816

S = 0.0101973 R-Sq = 56.3% R-Sq(adj) = 53.3%



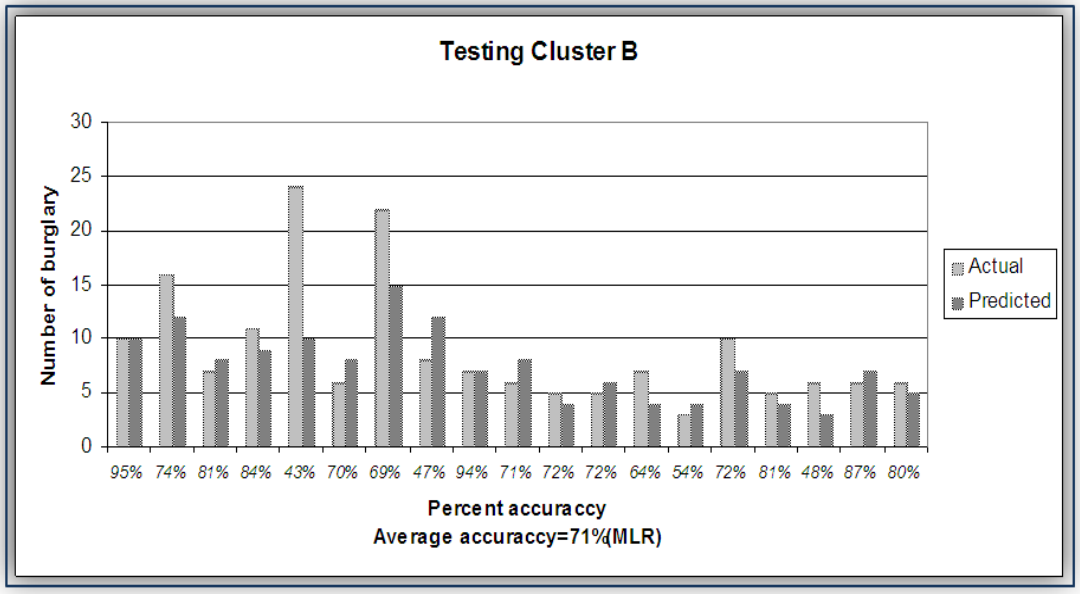
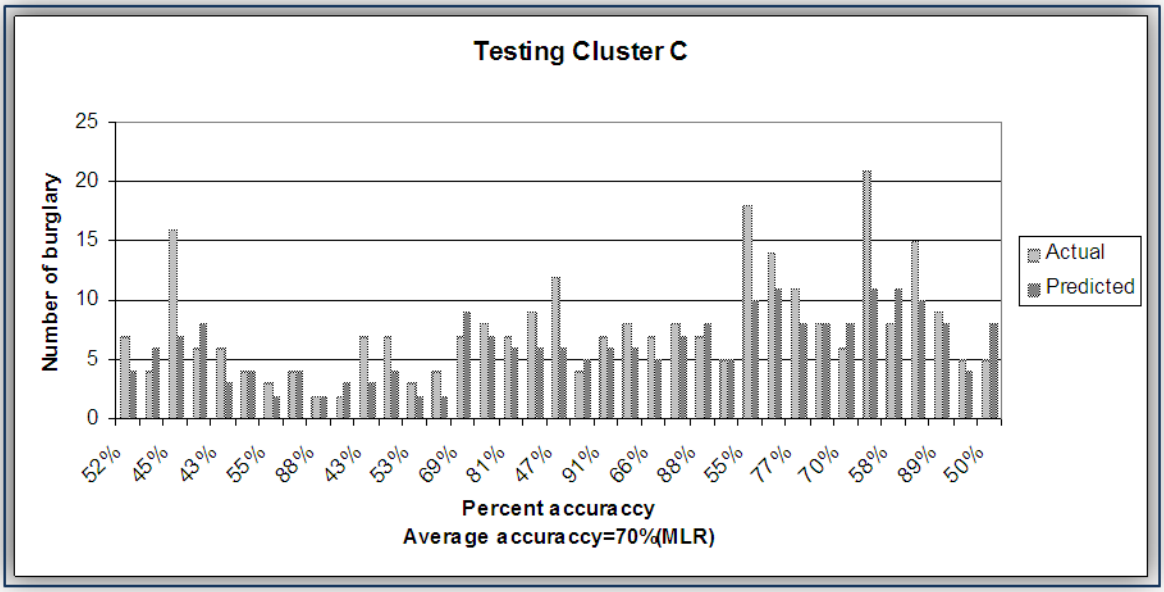
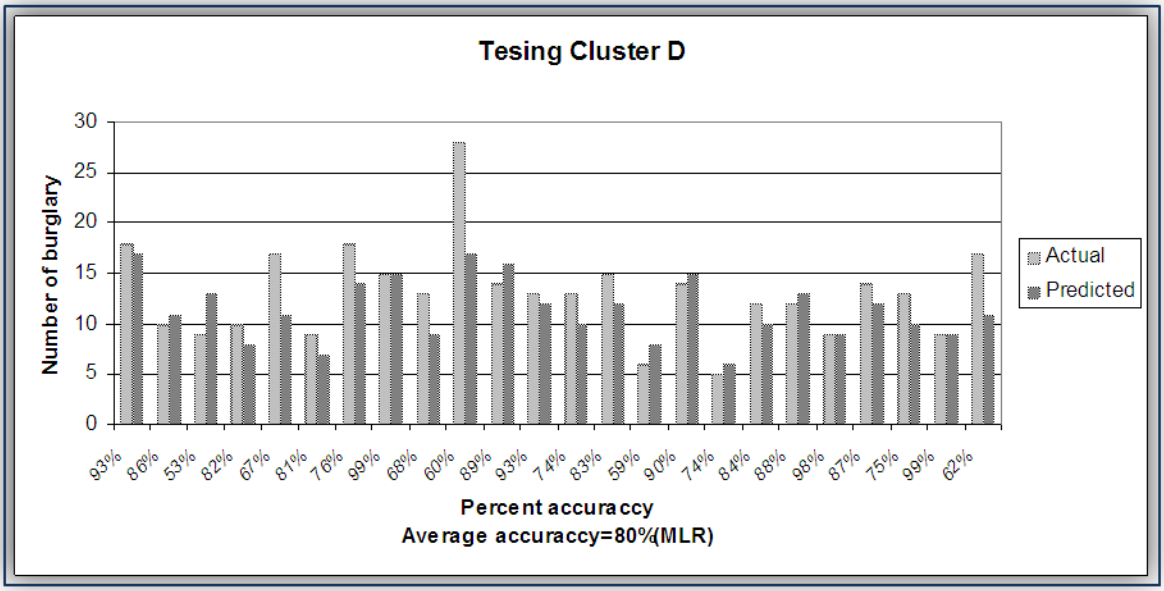


Figure 6.3: Illustration of the validation of the regression models using a new data. The new data have similar statistical properties as the model data. The accuracy percentage for the models and the polygon of the parcel's test are specified. The models across the clusters A and B are resonable. Their MAPE are 32 and 29 respectively.

Figure 6.3: Continuation. Validation of the regression models using a new data. The new data have similar statistical properties as the model data. The models across the clusters C and D are resonable. Their MAPE are 30 and 20 respectively.





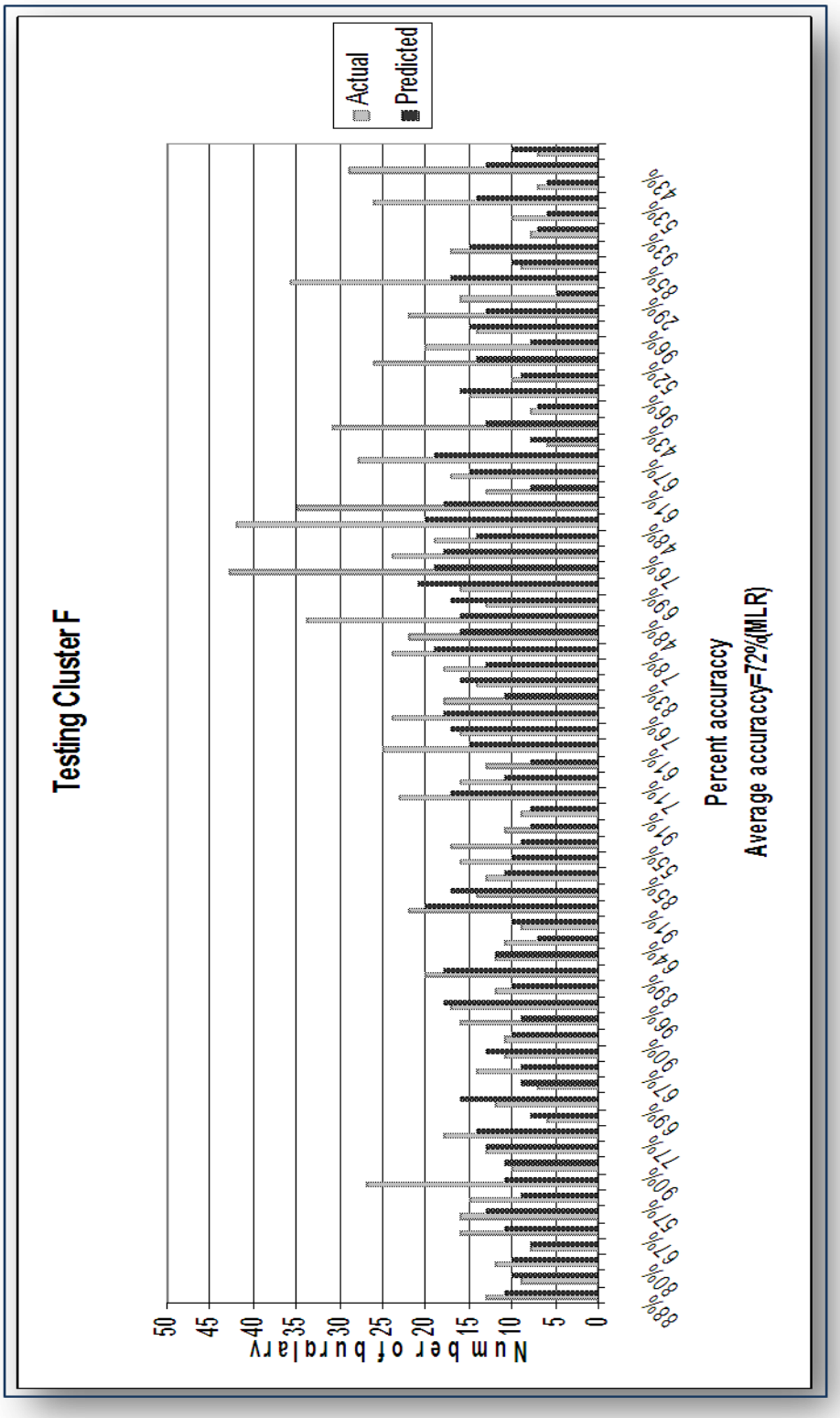


Figure 6.3: Continuation. Validation of the regression models using a new data. The model across a cluster F and it is resonable. Its MAPE is 28.

6.8.2.1 Results and Discussion

Regression models have the ability to identify significant input parameters from a set of potentially relevant parameters. The model results, which were obtained from this analysis and presented in Section 6.8.2, in general indicated that the potential predictors Qualification and Age structure are not significant within the observed datasets (residential burglary) which was recorded for the study region. However, the predictors Household Occupation, Socio-economic, Household space, Household composition are significantly correlated with burglary rate associated with the predictors property. For the influence of Household space, the results indicate that the risk of burglary rate increases within the household living in shared houses (f2). The result reported that the predictor housing types-lats were positively significantly correlated with burglary rate in the study region. In other words the household living in flats suffer from burglary more than in other housing types of living. The influence of household composition were significantly varies according to the composition of the household. For instance, the predictors' 'one person' and 'lone parent' were positively and significantly correlated with burglary rates. Those living in a household in which there was 'one person' and 'lone parent' increase the risk of the burglary rate. On the other hand the predictor 'all pensioners' were negatively significantly correlated with burglary, which decreases the risk of burglary rate. The obtained results are expected since it is logical. According to the theory of the economics of crime, increased unemployment rates lead to higher property crime rates, as is pointed out in Section 2.1.3 (Edmark 2005; Gorr &olligshlagers 2003). This study found that within the 'frame work' of a clusters model, unemployment positively and significantly correlated with burglary rate. Occupancy level significantly correlated with burglary rate positively or negatively related to economic status. Many researchers such as (Bowers 2004; Malczewski 2005; Edmark 2005; Gorr 2003) reported the relationship between economic status and crime. In this analysis the influence of the predictor occupation yield mixed results. This is because the economic levels within the occupation are not specified within the observed data. According to the information about the occupancy level for

some occupation are between high and medium or between medium and low level. This means that the economic levels within the observed data are not well defined. To clarify this point, the technical occupations were negatively significantly correlated with burglary rate within cluster C, while positively significantly within cluster A. But it is evident that the levels of professional occupations are in high occupancy level. The obtained results indicate that the predictor 'professional occupation' was negatively and significantly correlated with burglary rate and decreasing the risk of burglary. The influence of the household occupations related to blending of two factors: occupancy level of the occupation and the time that people are at work. The household occupation with a high level of occupancy decreases the risk of burglary. This is because the household in high level of occupancy can use security, which is effective in reducing the risk of burglary rate. On the other hand the low level of occupancy leads to lack of security. This leads to increase the risk of burglary rate.

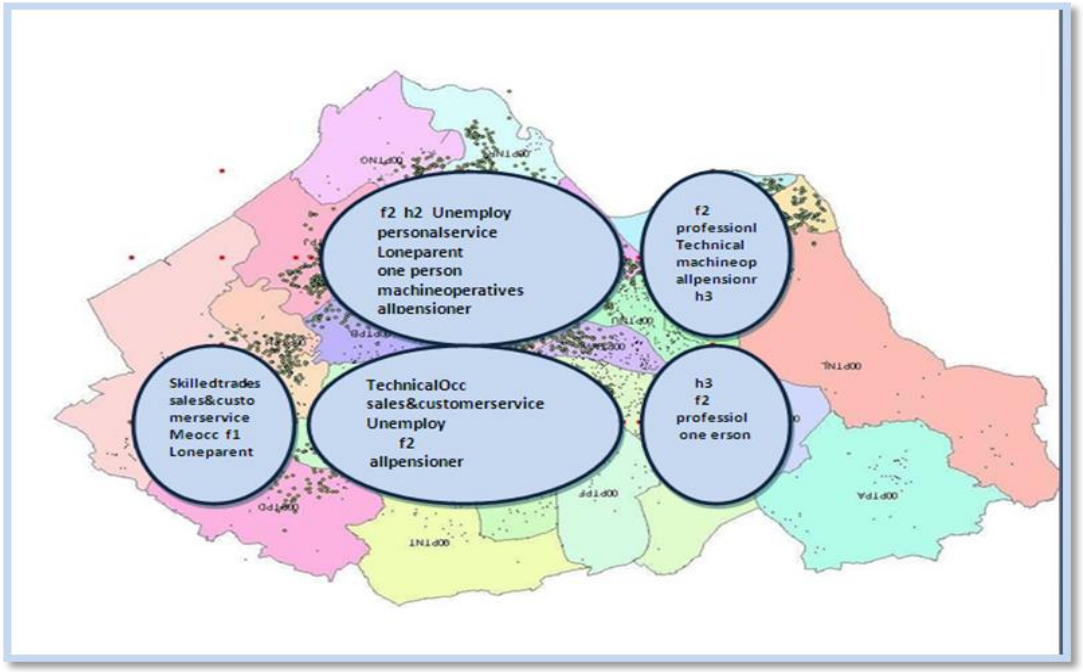


Figure 6.4 Characteristics of burgled household areas. The models of the clusters neighbourhod have approximately similar contributions to significant predictors.

6.8.3 Selection of Multilevel models

The model was first applied across a number of geographical space parcels in the study region. Each parcel's data is modelled separately in order to identify significant predictors of burglaries, through the potential of the examination of different characteristic of burgled households. The final MINITAB output for a number of prediction models of geographical space in the study region are presented in Figure 6.5. The detail of the models can be found in Appendix E. The predictive burglary models which were obtained from this analysis were tested with a new data set and the results are reasonable according to the formula 6.6. Then the models within the specific three levels of the spatial distribution of burglary rate were selected. In the cases of the identification of levels of burglary rate in the study region previously specified in Section 5.4.2**.** The levels are high, middle and low. The models were selected according to the criteria (outline in Sections 6.6). Ten parcels were selected from the high level region and selected four parcels from low region level and four parcels from the middle level region. Table 6.5 provides a summary of the process of model selection. The results indicate that model NN, NY and PF were producing the smallest prediction error according to the criteria MAD, MAPE and MRSE and it has highest value within R-sq and R-sq (adj). Consequently, model NN was selected for predicting future crime in the region of high level of burglary rate. Models NY and PF were selected as representing the predicting of future crime in middle and low level of burglary rate regions respectively. The obtained models were tested with new data. This was derived from a parcel with a similar level of burglary level (see Figure 6.6). The accuracy percentage was achieved for the models and the polygons of the parcel's test.

The models which were obtained from this analysis and which are presented in Figure 6.5 indicate that household occupations are strongly related to burglary rate. The households' occupations with certain occupations are more at risk than others. For instance in the multilevel models the skilled trades occupation are negatively and significantly correlated with burglary rate. That decreases the risk of burglary

rate. Households who are working as Elementary and Personal service occupations are positively significantly correlated with burglary rate. In other words, those households whose work is Elementary and Personal service occupations are at increased risk of burglary rate.

Parcel PF

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.00402	0.005731	-0.7	0.498	
TechnicalOcc	0.08231	0.03004	2.74	0.019	1.573
SkilledtradesOcc	-0.16384	0.05292	-3.1	0.01	2.87
personalserviceOcc	0.22708	0.05618	4.04	0.002	1.864
Machineoperatives	-0.25408	0.08481	-3	0.012	1.792
ElementaryOcc	0.32697	0.05887	5.55	0	1.654
Cohabiting	0.18563	0.06888	2.7	0.021	3.862

S = 0.00161830 R-Sq = 94.5% R-Sq(adj) = 92.2%

Parcel NY

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.14523	0.03309	4.39	0.001	
SkilledtradesOcc	-0.499	0.1728	-2.89	0.012	1.554
sales&customerservice	-0.6365	0.245	-2.6	0.021	2.172
ElementaryOcc	-0.8677	0.1964	-4.42	0.001	2.566

S = 0.0126028 R-Sq = 84.1% R-Sq(adj) = 68.1%

Parcel NJ

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.0454	0.01674	-2.71	0.017	
SkilledtradesOcc	0.585	0.1143	5.12	0	2.775
personalserviceOcc	-0.4492	0.1327	-3.38	0.004	2.806
sales&customerservice	0.3012	0.115	2.62	0.02	2.894
machineoperatives	0.7618	0.1844	4.13	0.001	2.767
Unemploy	0.40883	0.08852	4.62	0	1.954
flat1	0.19175	0.04591	4.18	0.001	2.383
cohabiting	0.28177	0.09469	2.98	0.01	3.32

S = 0.00492369 R-Sq = 91.6% R-Sq(adj) = 82.5%

Figure 6.5: Presents an example of Minitab printouts of regression models of geographical space in the study region.

Table 6.5: This summarises the process for the selection multilevel models.

Selection models are shown dark colour.

High level	MAD	MAPE%	MRSE	Average Accuracy%	R-sq%	R-sq(adj)%
NF	0.005158	17%	4.56E-05	83%	93.90%	82.70%
NH	0.00471	14%	3.03E-05	86%	93.80%	85.10%
NK	0.005707	27%	0.001231	73%	80.70%	61.30%
NR	0.030865	18%	4.79E-05	82%	81.6%	67.3%
PG	0.009769	20%	0.000134	80%	93.7%	87.8%
NG	0.010392	23%	0.000167	77%	72.6%	58.9%
NJ	0.002749	8%	1.13E-05	92%	91.60%	82.50%
NN	0.001022	3%	1.67E-06	97%	99.0%	98.40%
PC	0.001799	22%	0.002284	78%	98.20%	97.40%
PJ	0.00846	39%	0.000111	61%	87.3%	68.8%
Mid level	**MAD**	**MAPE%**	**MRSE**	**Average Accuracy%**	**R-sq%**	**R-sq(adj)%**
PK	0.011245	48%	0.013703	52%	80.7%	61.3%
NZ	0.005851	18%	0.007397	82%	77.20%	61.60%
NY	0.004838	17%	0.005768	83%	84.10%	68.10%
PH	0.011445	50%	0.013733	50%	25.90%	0.00%
Low level	**MAD**	**MAPE%**	**MRSE**	**Average Accuracy%**	**R-sq%**	**R-sq(adj)%**
NX	0.001865	24%	0.002215	76%	85.00%	68.80%
PF	0.001046	11%	0.001237	89%	94.50%	92.20%
NS	0.004927	22%	0.005906	78%	78.90%	52.90%
NM	0.001421	15%	0.001778	85%	95.40%	89.6%

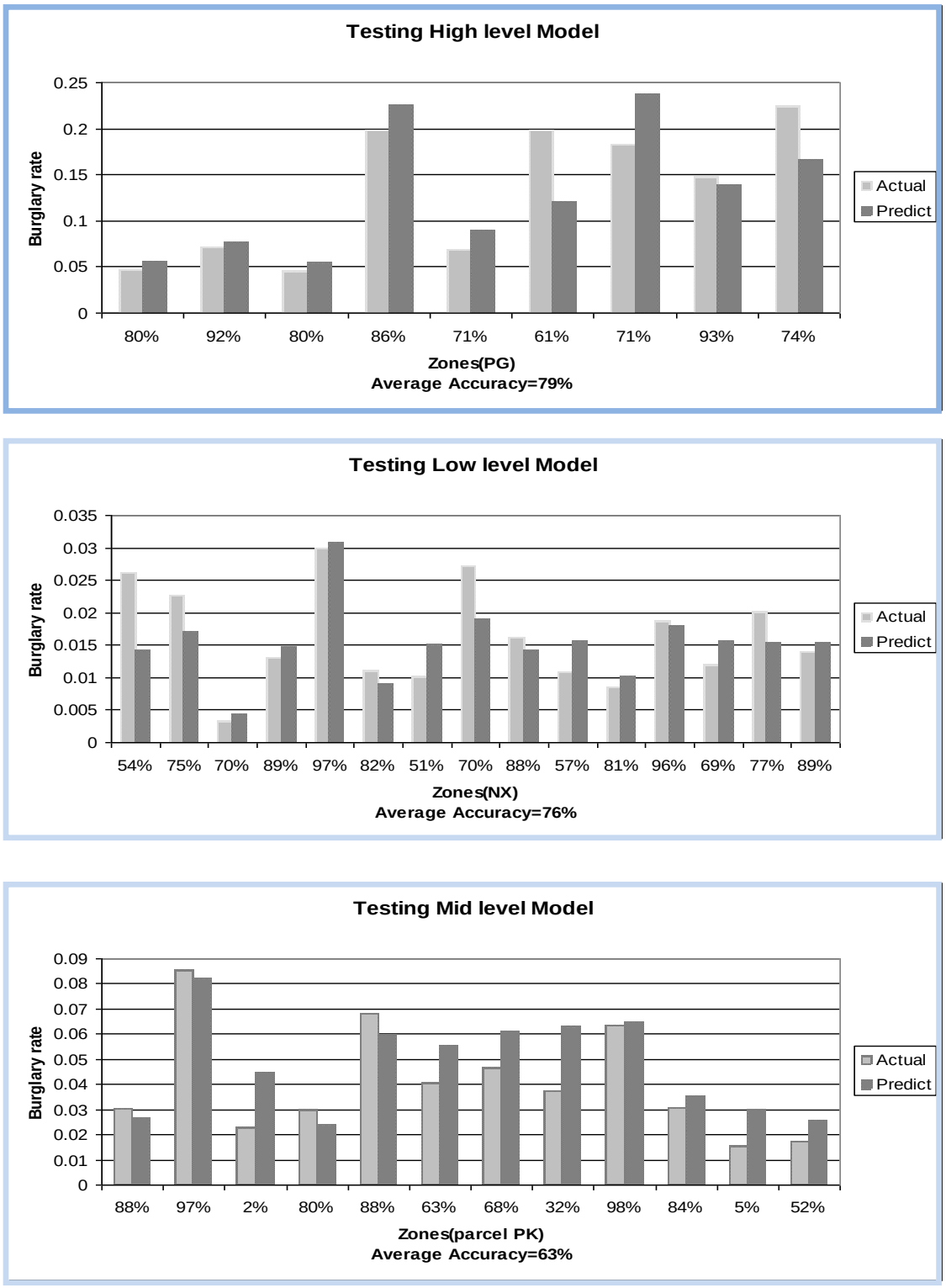


Figure 6.6: Illustration of the validation of the multilevel models using a new dataset. The new data have similar statistical properties as the model data. The accuracy percentages for the models and for each polygon within the parcel are specified. The models are reasonable.

Table 6.6 : Shows the number of significance explanatory variables of characteristics of household within the location of burglary rate levels.

Significance explanatory variables Characteristics of household	high	middle	low
Professional Occupations	5195	6381	7549
Skilledtrades Occupations	3413	3596	3642
Technical Occupations	6029	6504	7225
Personal service Occupations	2634	2452	2456
sales&customer service ocupations	3961	3707	3642
Machine operatives	2681	2541	2114
Intermediate occupations(Meoc)	5992	6608	7383
Unemploy ment	4504	2911	1877
Semidetached (h2)	11076	11796	15707
Terraced (h3)	16210	15898	10746
Tenement (f1)	7782	5249	3563
Shared house (f2)	3844	1986	691
one person	13966	11849	10286
allpensioner	2260	3021	4067
Loneparent	6002	4427	3760

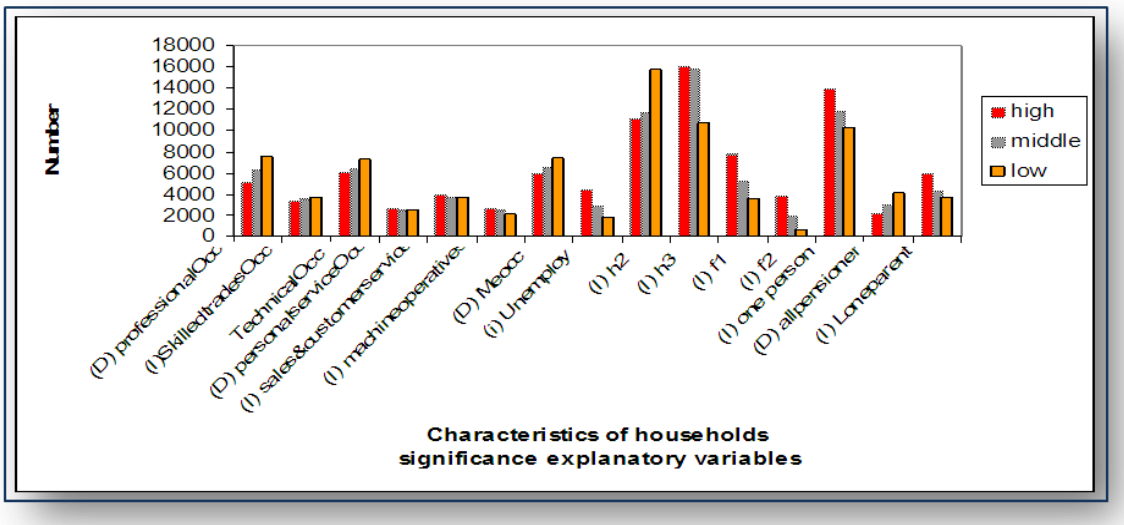


Figure 6.7: Shows how the significant explanatory variables, namely the characteristics of household affect the rate of burglary rate. *D* indicates that the variable decreases the risk of burglary whereas *I* indicates that the variable increases the risk of burglary.

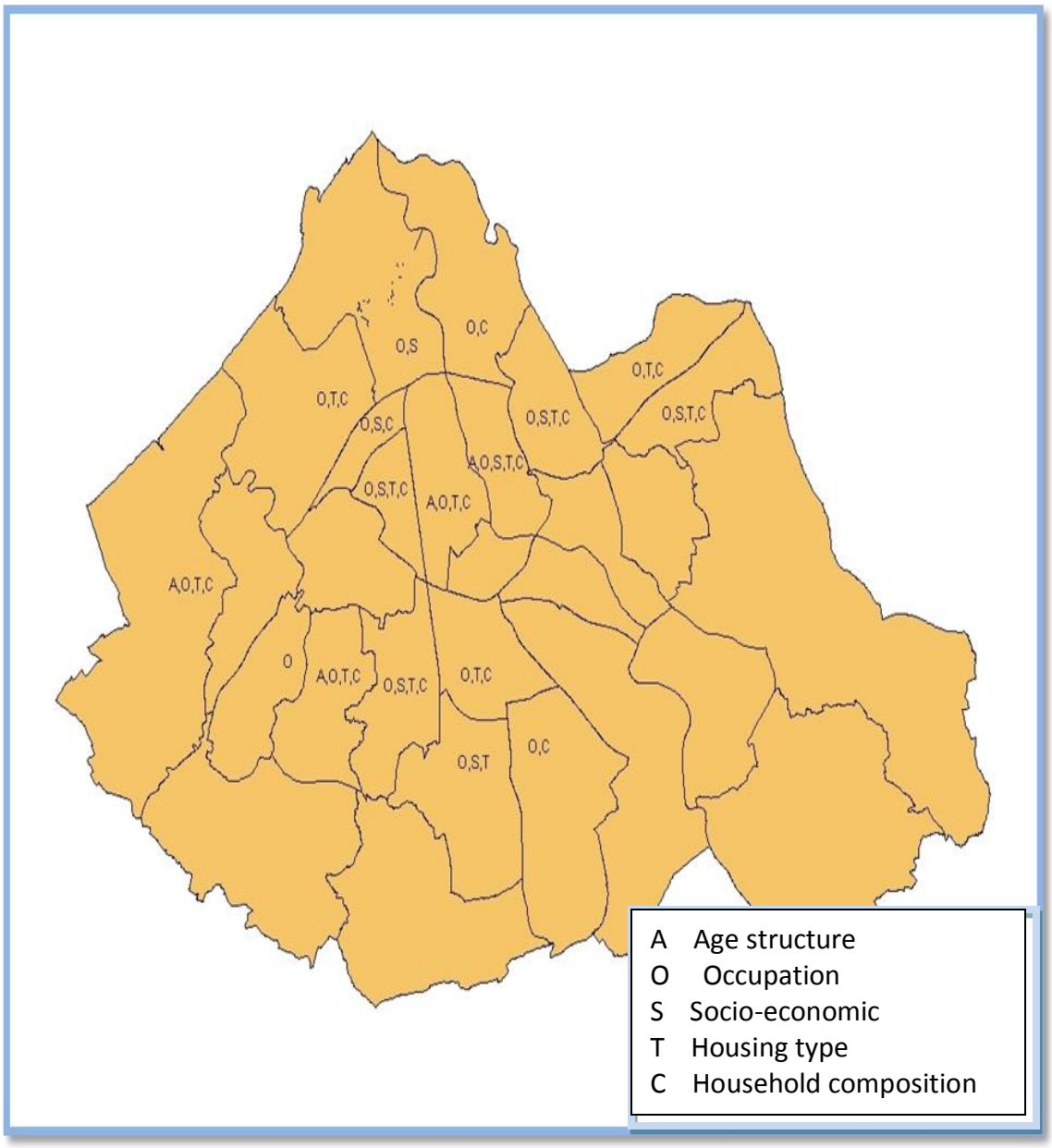


Figure 6.8: Characteristics of burgled household areas. Significance explanatory variables of characteristics of household within parcels in the study region.

6.9 Artificial Neural Models for Prediction

This section outlines Artificial Neural Networks (ANNs) procedures for building a predictive crime model. ANNs are computational modelling tools that attempt to establish a mathematical relationship between input and output (Basheer, 2000). ANNs are accepted in many disciplines for modelling complex- real world problems. ANNs are relevant in crime prediction and have been established in a number of research projects such as Olligschlaeger 1997; Olligschlaeger and Gorr 2001; corcoran and Wilson 2003; Oatley and Ewart 2003.

A hierarchical neural network (HNN) is a neural network architecture in which the problem is divided and solved in more than one step (Mehratra, 1997: 21). HNN has been introduced for improving the performance of the model in terms of time and accuracy. HNNs have found their use in various applications such as medicine, ecology and sociology. Mat (2002) used a hierarchical Radial Basis Function (HRBF) in diagnosing cervical cancer. The hierarchical network was divided into two Radial Basis Functions. The first network performs as a filtering process to the second network. That the second network is fed with certain data. The study reveal that HRBF was increased the performance of single RBF. Corcoran (2003) applied the KSOM network to crime data with clustering of topological ordering then the clusters data formed by KSOM used to train a separate MLP network. Thus, the basics of hierarchical neural network are not new. A HNN for building a predictive crime model which is presented in this thesis is somewhat different. The new methodology is a hierarchical neural network (HNNs) approach with training data for the first network being prepared by a regression methodology. In this case the statistical methodology is combine with a powerful feature of two proposition ANN learning algorithms; unsupervised (Self-Organizing Map SOM) and supervised (back-propagation BP) to generate a more a accurate prediction. The concept of a combination of several methodologies for problem solving has the advantage of improvement the performance of model prediction. Practical works which are described in section 6.9.2 illustrated this point.

An unsupervised SOM network consists of two-layers; input layer which accept multiple inputs, and output layer. The data from the input layer is passed directly to the output layer with no hidden layer. The output layer is formed as a regular grid of cells, which transforms n-dimensional input patterns into one or two dimensions. It has been applied successfully in clustering and visualization of high dimensional data (Kohonan, 1989). The task of a SOM in this study is hybrid networks for, reduces the dimensionality of the input data set into 2 D and is then used as front- end (training set) to BP network in the specified HNN.

A supervised Back Propagation is a multi-layer, feed-forward neural network consisting of three layers: an input layer, a layer with nodes representing the potential influencing factors of a specific problem, a hidden layer and an output layer with nodes that represent the solution of the problem. This supervised learning algorithm is designed to minimize the mean square error between computed output of the network and the desired output. BP network can be used for classification and prediction. The task of BP network in this study is used for prediction burglary rates and to examine the statistical identification of burglary rate levels in the study region.

6.9.1 Methodology

The development of neural network models requires a consideration of a number of issues such as data representation and structuring. This is important in the assessment of a robust model.

6.9.1.1 Data

The dataset was normalized in a 0-1 range, which is needed to construct a neural network model. It is divided in to a training set which is used to train the model and a test set which is used to test the performance of the trained model. The portion of training and test dataset is suggested to be 80% of the dataset for training and 20%

of the dataset for testing the model. A regression methodology (outline in Section 1.1.4) was used in this study for preparing a hierarchical neural network training data. This methodology has the ability to identify explanatory significant variables out of a number of potential explanatory variables which was used to construct the regression model. Training data of HNN in this study starts with a number of explanatory variables with statistical significance which were obtained and identified from regression models results (Section 6.8.2). Non-significant explanatory variables were removed from the data set. This helps to reduce time required to complete the task. The value of an actual output values for prediction are represented by the burglary rate (Section 5.3.2) and for classification the components of each actual binary output vector is summed to 1 (one-of n code).

6.9.1.2 A hierarchical neural network structure

The structure of the HNN which is presented in this thesis consists of five layers: input layer, SOM input layer with *k* specified input neurons; Kohonen layer with neurons which arranged as n rows and m columns; Coordinate layer which become BP input layer in the HNN. The BP network within the HNN consists of one hidden layer and one neuron output layer (see Figure 6.9). There are several parameters which influence the performance of an ANN such as the number of hidden layer and its nodes; transfer function[1]; learning rule[2]; learning runs. There is no formal theory for determining optimal of these parameters. Therefore optimization are made during practical work based on the minimize network error value (form 6.5). It was found from practical work that the choice of the above parameters is important to reduce model's error.

---

[1]Transfer function: is a function that maps a neuron's net output to its actual output.

[2] The Learning rule is the mathematical equation that determines the increment or decrement by which weights of a processing element are changed during the learning phase.

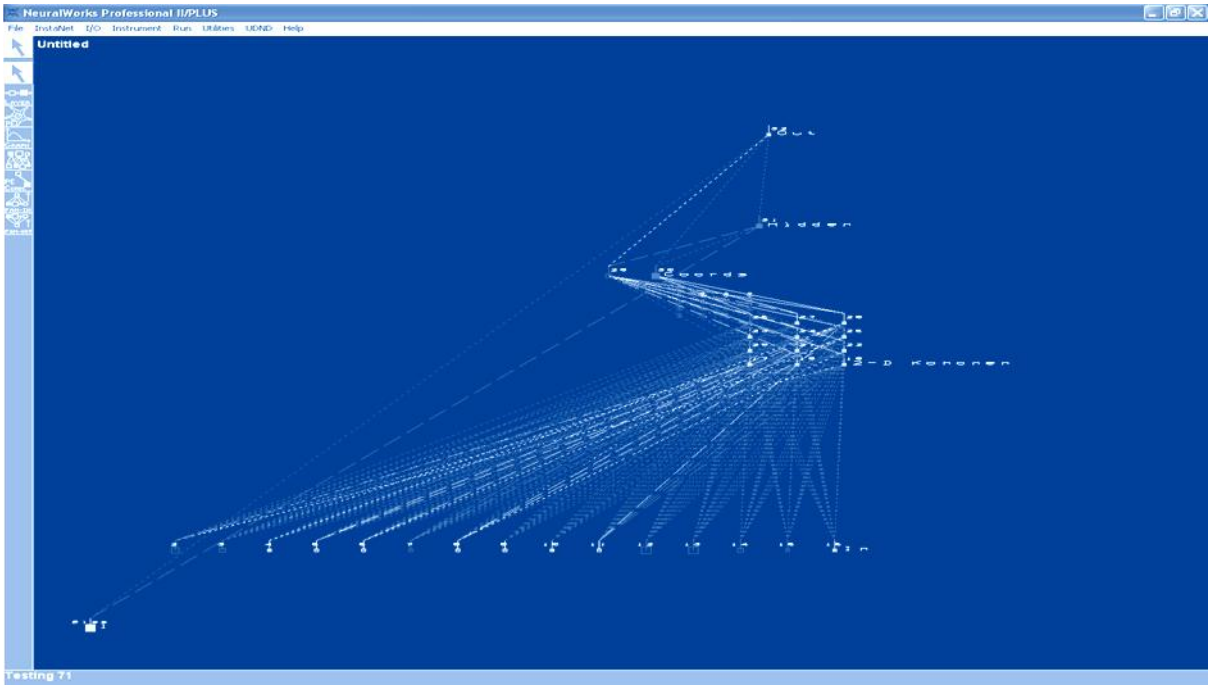


Figure 6.9: Hierarchical neural network architecture of the predictive crime model. Display neurons within the layers. (NeuralWorks Professional II/PLUS software).

6.9.2 Experimental work

A comprehensive software package NeuralWorks Professional II/PLUS was used to develop neural network predictive models. This software was selected because it allows the user to design, test and then implement many different neural networks and easily alter learning parameters during training. In the experiment below, the training and testing process of HNN and BP network are described. Furthermore, detailed comparisons of average percentage accuracy achieved at testing stage by the obtained models are discussed.

Experimental work 1

A supervised BP learning algorithm was selected in this study for classification and predication. A three layered BP network was created first, to examine the levels of burglary rate in the study region. An input layer with 15 neurons represents

explanatory variables with statistical significance (Section 6.8.2). There is no rule for determining the number of hidden layers and neurons. The processes start with Lawreuce and Fredrickson's (1998) suggestion (outlined in section 1.1.3). A best estimation for the number of neurons is a half of the sum of inputs and outputs. Thus the best obtained results for the current models are one hidden layer with 8 neurons. The output layer is with three neurons are represent binary actual output. For example, 1 0 0 represents a low burglary rate level. The selected parameters out of a several options from the dialogue box for back- propagation are shown Figure 6.10. This dialogue box contains a lot of information about the layer, including the learning rule used. In Practice it was found that the network had the smallest AME error with the following selected parameters: transfer function Logistic (sigmoid) and learning rule Extended Delta- Bar- Delta (Ext- DBD) and learning run between 50000 and 90000. A back Propagation neural network architecture associated with the selected parameters is shown in Figure 6.11. The obtained network results are summarized in Table 6.7. The percentage accuracy obtained for a high level of burglary rate is 85% which means that the model is a good prediction. The percentage accuracy obtained for middle and low levels of burglary rate are 56% and 78% respectively. This means that the models in this case are reasonable accurate predictors (Lewis, 1982:40).

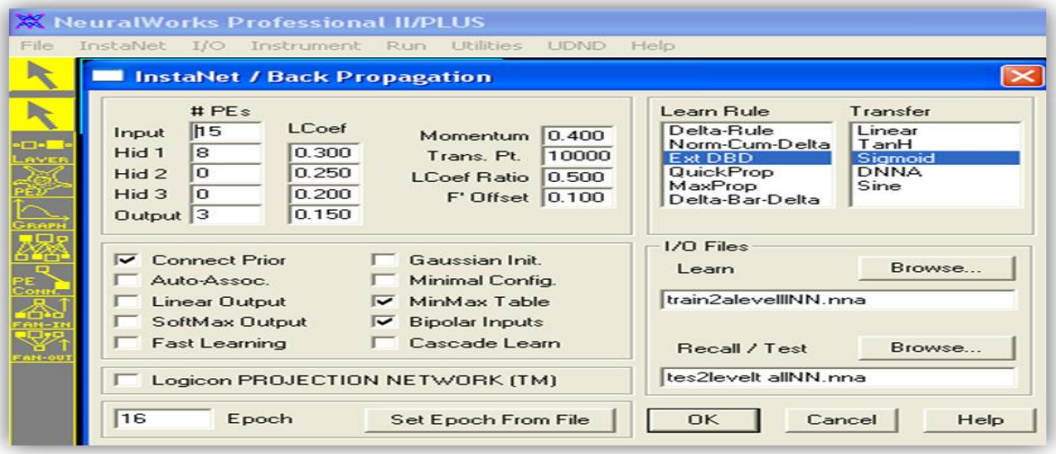


Figure 6.10: Back-propagation dialogue box. Selection parameters used for building classification model,experimental work 1.(software package NeuralWorks Professional II/PLUS).

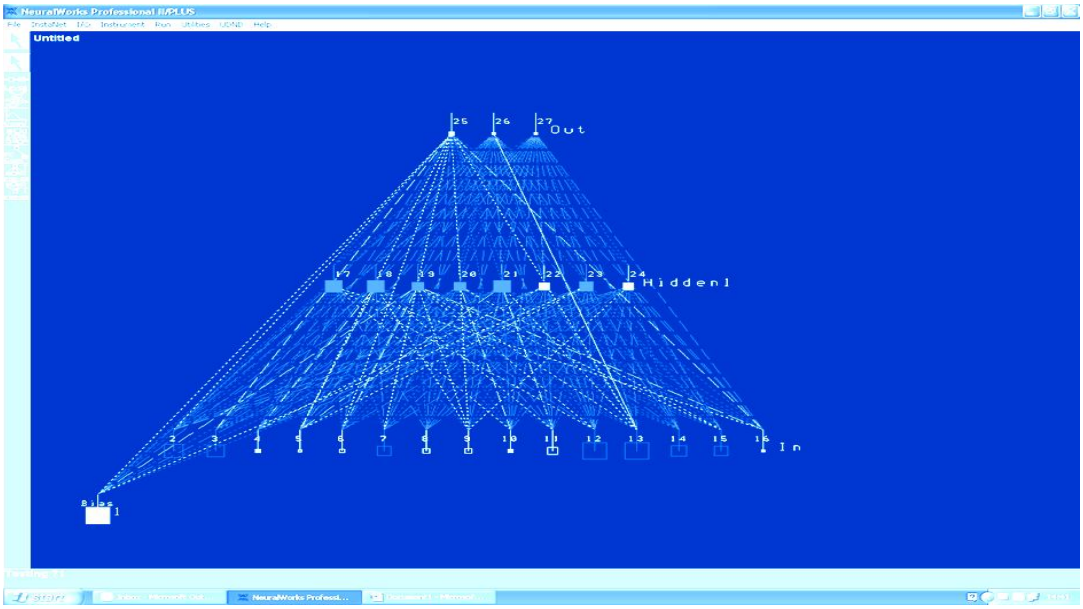


Figure 6.11: Back-propagation network for classification, experimental 1. Display neurons within the layers. (Software package NeuralWorks Professional II/PLUS)

Table 6.7: Results achieved by BP network for examine the levels of burglary rate.

Levels	Average percentage accuracy
Low	78%
Middle	56%
High	85%

Experimental work 2

The non hierarchical neural network used BP for prediction burglary rate in the study region. A BP network is created with three layers. 28 input vectors of potential explanatory variables which were used to construct the regression model are fed into the BP input layer. The model was applied across a number of selection parcels in the study region. In this case the best selected results from BP dialogue

box are one hidden layer with 13 neuron and output layer within one output neuron which represent burglary rate prediction (see Figure 6.12). The obtained network is illustrated in Figure 6.13. Their obtained results for average percentage accuracy are illustrated in Figure 6.14 and summarized in Table 6.8. The results shown that the average percentage accuracy an achieved by BP models at testing stage are 44-69%. Then these obtained results can later be comparing with the results of a HNN new methodology (experimental 3). This is to demonstrate the performance of the new methodology.

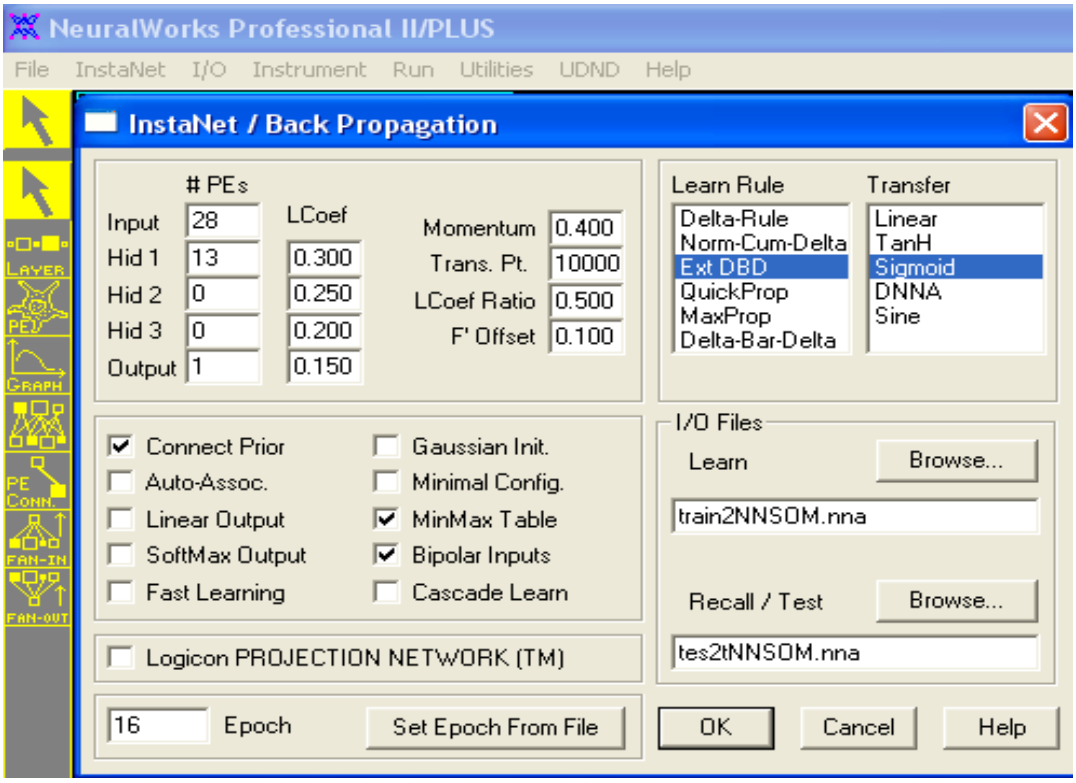


Figure 6.12: Back-propagation dialogue box. Selection parameters used for building predictive model, experimental 2. (Software package NeuralWorks Professional II/PLUS).

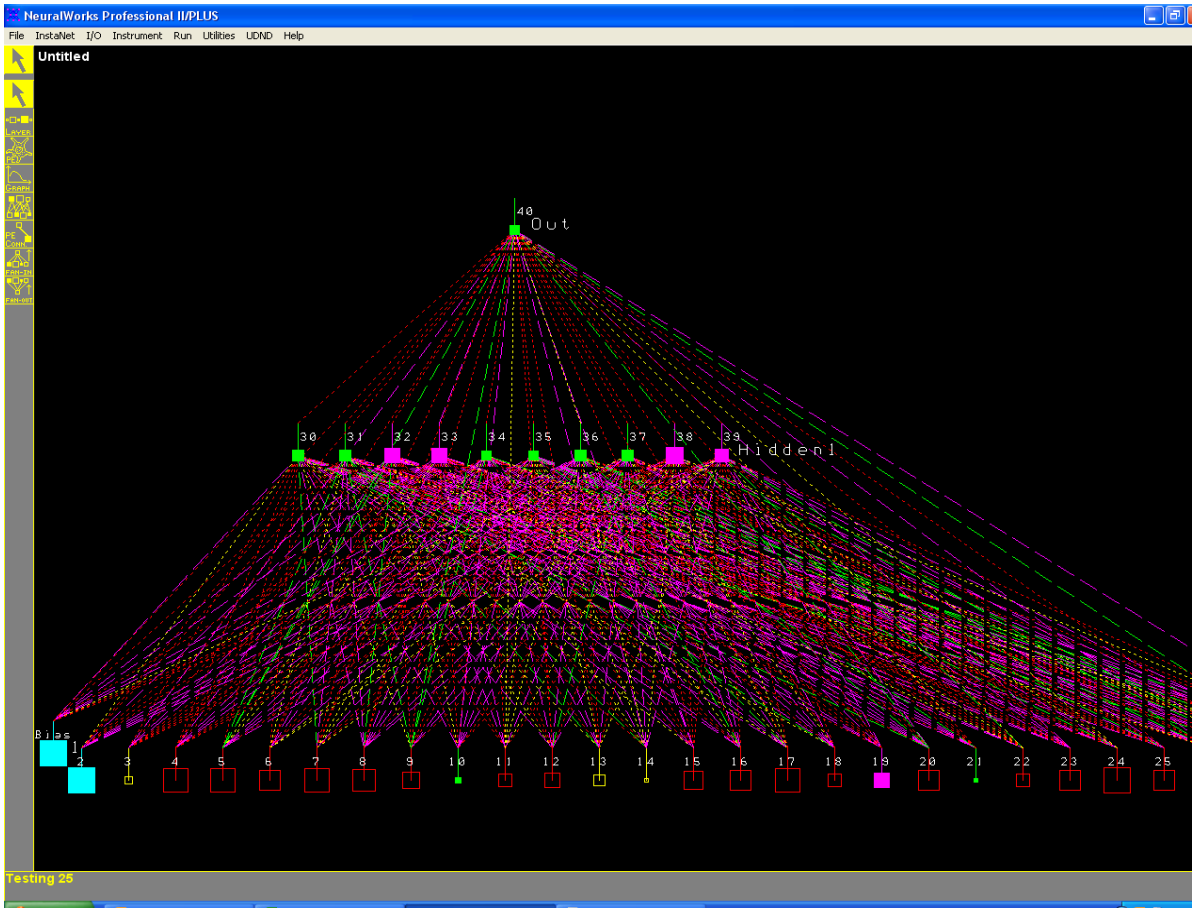
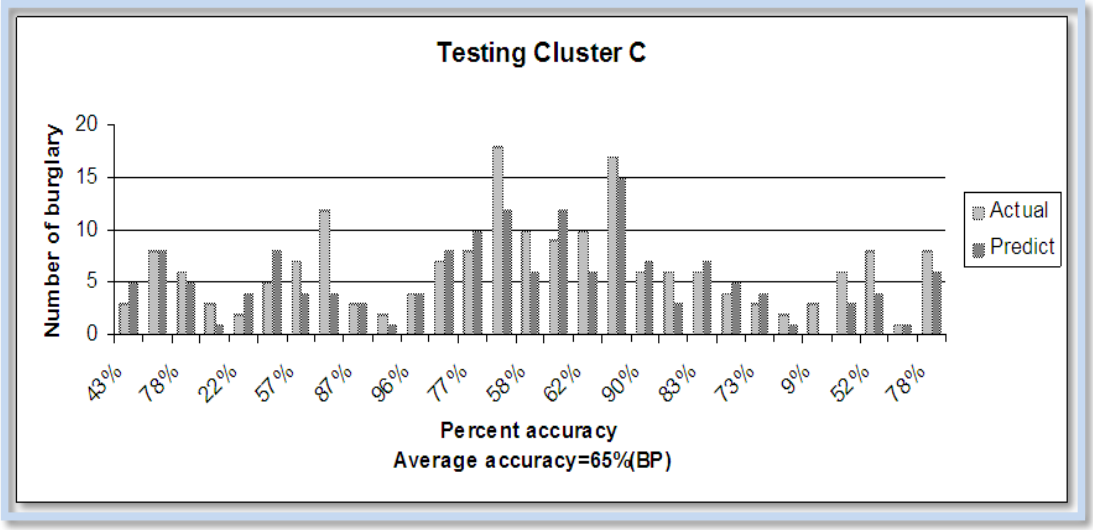


Figure 6.13: Back-propagation network, experimental 2. Display neurons within the layers. (Software package NeuralWorks Professional II/PLUS).

Table 6.8: The average percentage accuracy and mean absolute deviation for selected models which were achieved by BP network.

Models	Average percentage Accuracy	MAD
Cluster A	65%	0.0172
Cluster B	44%	0.01168
Cluster C	65%	0.0067
Cluster D	64%	0.0214
Cluster F	69%	0.0197
Parcel NN	51%	0.02666
Parcel PF	47%	0.0204



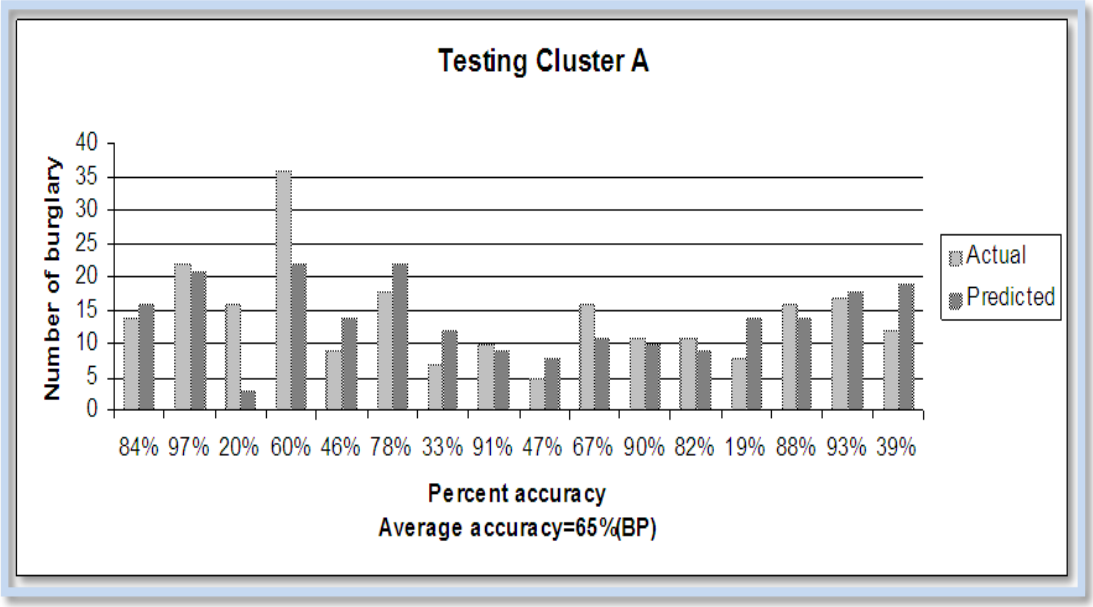
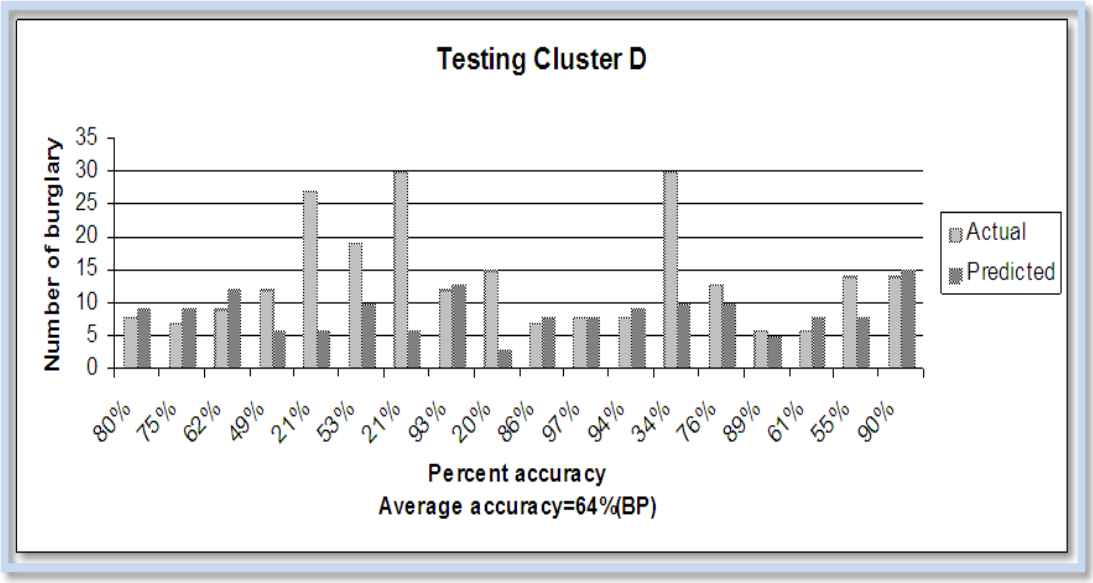


Figure 6.14: Presents an example of BP model results. The accuracy percentages for the models and for each polygon within the parcel are specified.

Figure 6.14: Continuation. BP model results. The accuracy percentages for the models and for each polygon within the parcel are specified.



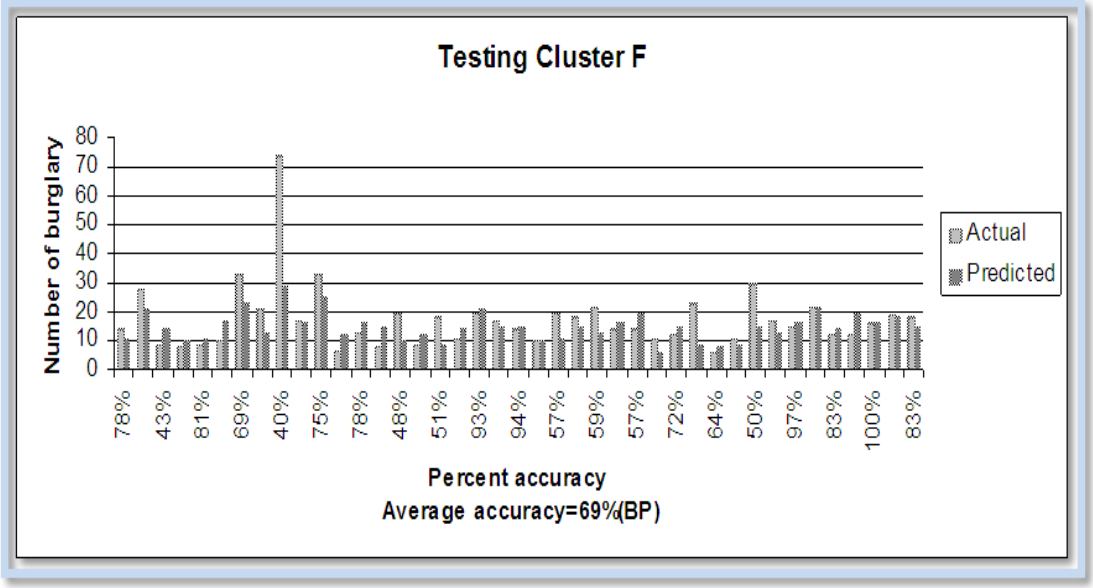
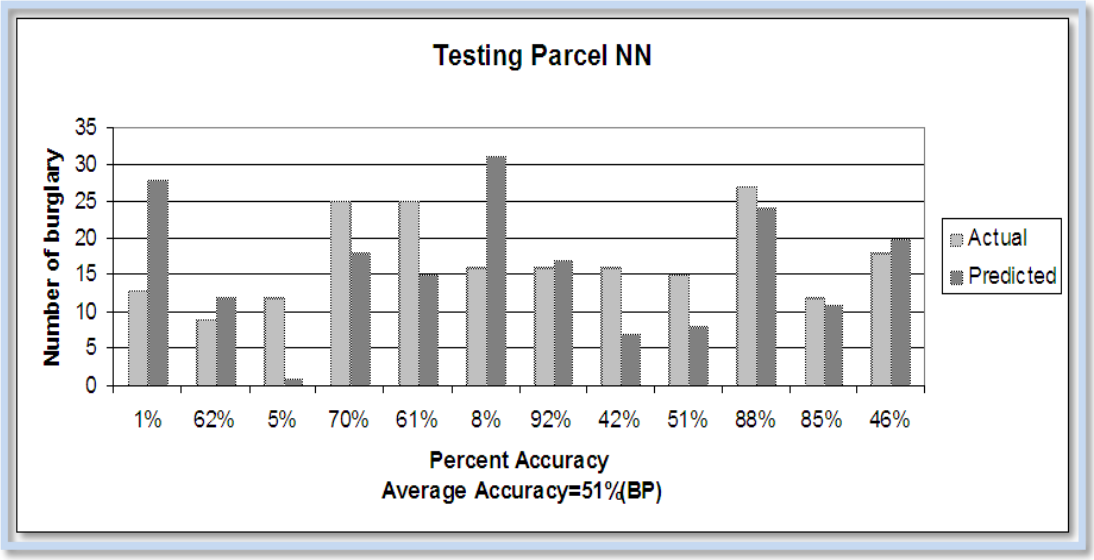


Figure 6.14: Continuation. BP model results. The accuracy percentages for the models and for each polygon within the parcel are specified.



Experimental work 3

In order to illustrate the performance of a new methodology for building a predictive model, the models are applied on the same selection parcel as in experiment 2. A hierarchical neural network (HNN) starts their construction model with a number of explanatory variables with statistical significance. These were obtained and identified from regression model results. The HNN begins training in an unsupervised (SOM) with no specified actual output. This then requires supervised training for the predictive values. The input vectors specified in Section 6.9.1.1 feed into the SOM input layer. The SOM network is trained with different map sizes. During the practical work found that 4x3 is the optimum map size in Kohonen layer is identified. This was based on the minimum values of AME (formula 6.5). The Kohonen layer output feeds into a coordinate layer which gives an (x,y) representation of the winning neurons in the SOM network layer. Figure 6.9 shows how a SOM network can be used to assist in the task of dimension reduction. The dimension reduced in this case from 15 to 2 D. BP network input layers' make use of a coordinate layer with two neurons. Therefore a coordinate layer-become a BP input. Both systems, hierarchical and non-hierarchical, used the back- propagation

algorithm for the prediction of the burglary rate. A back-propagation network in this case consists of one hidden layer and one output neuron. Experiments found that the combination of one hidden layer, a hyperbolic tangent (TanH) (transfer function); Extended Delta-Bar-Delta (Ext. DBD) for learning; and learning run between 3000 and 10000 had the smallest AME error. The summary of the practical selection of these parameters from Self- Organizing Map dialogue box is shown in Figure 6.15. The obtained network results are summarized in Table 6.9.

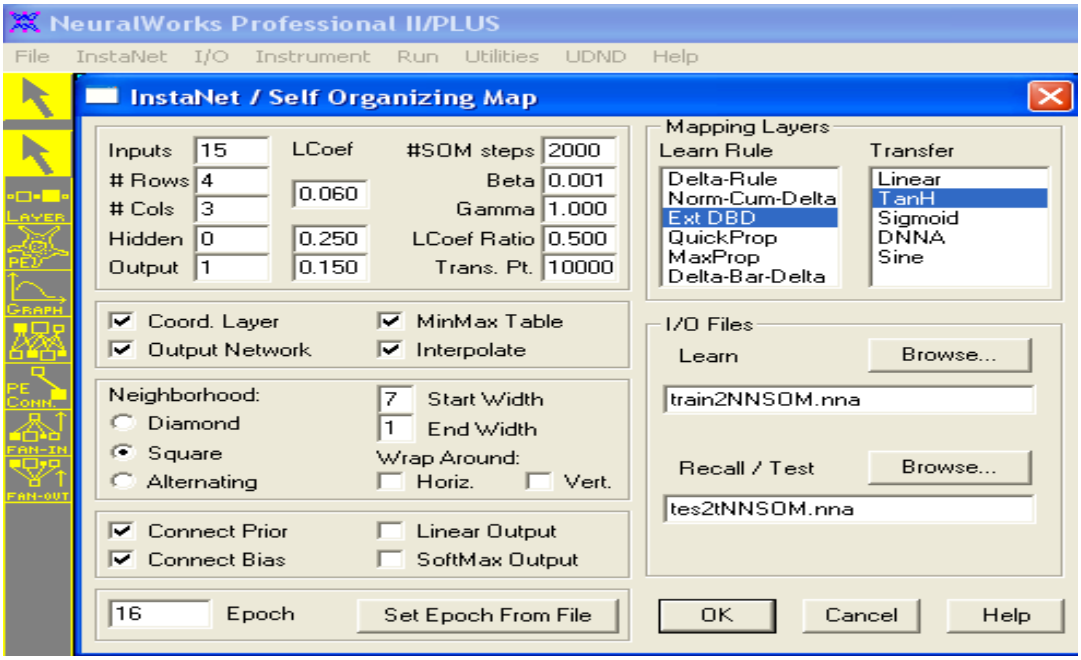


Figure 6.15 Self Organizing Map dialogue box. Selection parameters used for building predictive model, experimental work 3. (Software package NeuralWorks Professional II/PLUS).

Table 6.9: The average percentage accuracy and mean absolute deviation for selected models which were achieved by a new HNN network.

Models	Average percentage accuracy	MAD
Cluster A	75%	0.0150
Cluster B	71%	0.0075
Cluster C	76%	0.00612
Cluster D	74%	0.01666
Cluster F	73%	0.0152
Parcel NN	70%	0.019015
Parcel PF	55%	0.013503

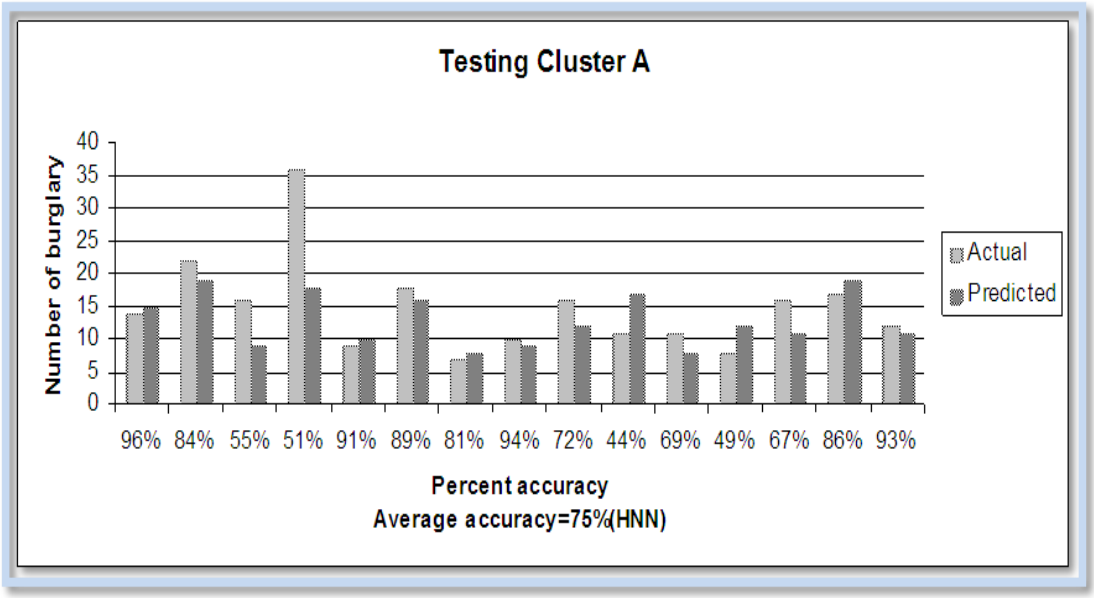
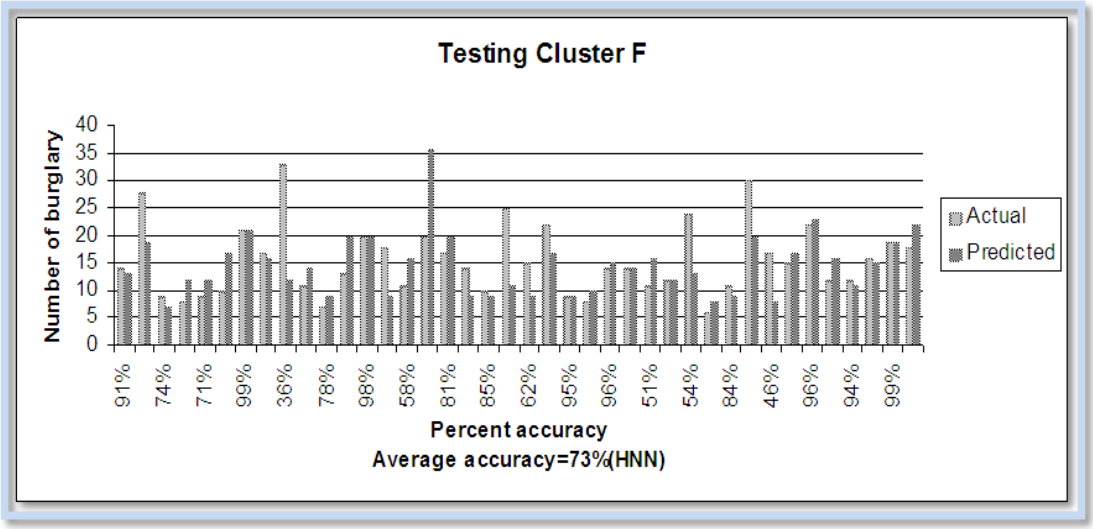


Figure 6.16: Presents an example of a new HNN model results. The accuracy percentages for the models and for each polygon within the parcel are specified.

Figure 6.16: Continuation. HNN models results. The accuracy percentages for the models and for each polygon within the parcel are specified



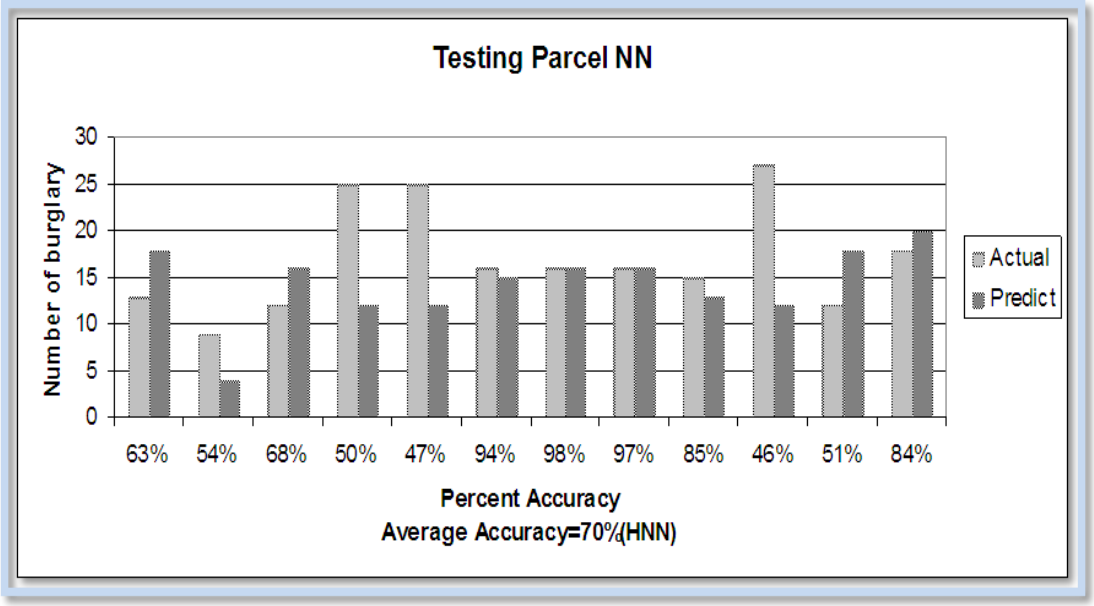


Figure 6.16: Continuation. HNN models results. The accuracy percentages for the models and for each polygon within the parcel are specified.

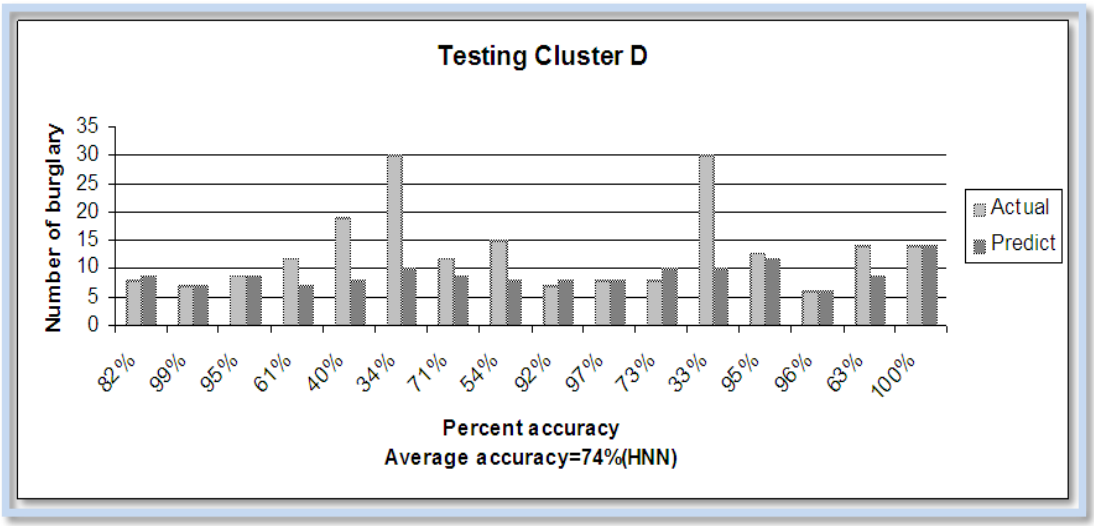


Table 6.10: Summarize the performance of each of the techniques investigated in this analysis: MLR, new HNN and BP.

Models	MLR	New HNN	BP
Cluster A	68%	75%	65%
Cluster B	71%	71%	44%
Cluster C	70%	76%	65%
Cluster D	80%	74%	64%
Cluster F	72%	73%	69%
Parcel NN	79%	70%	51%
Parcel PF	76%	55%	47%

Experimental work 4

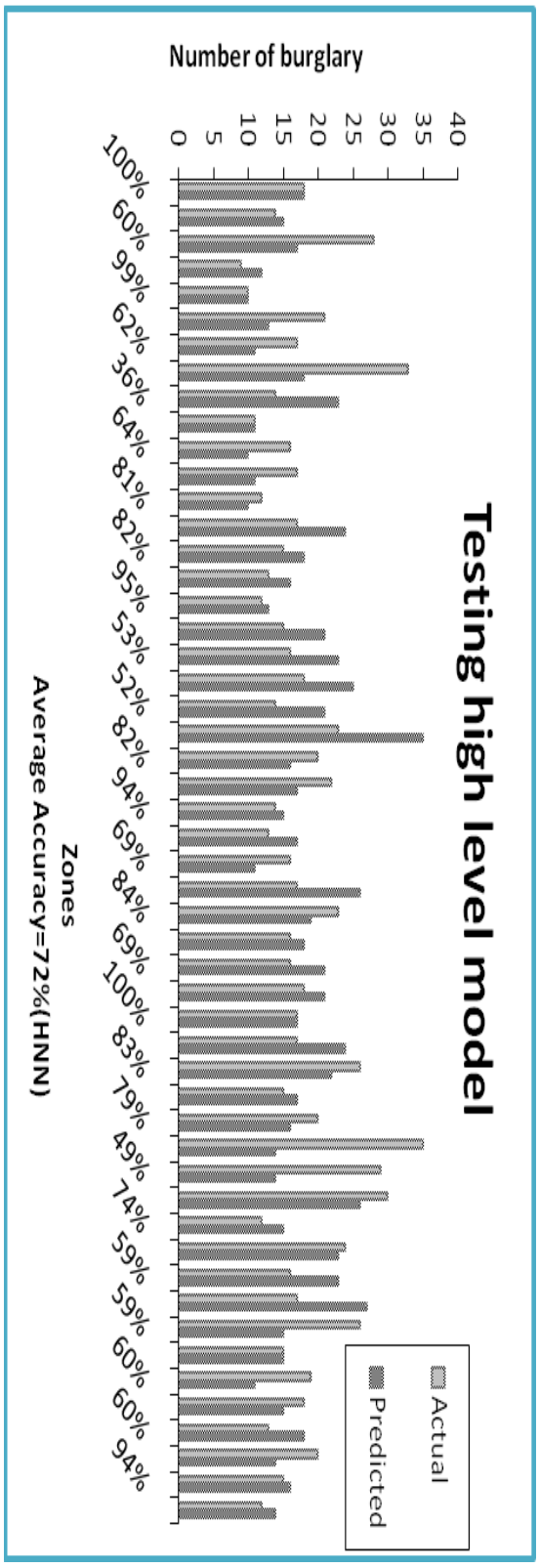
Hotspots analysis has been an important approach for the explanation and prediction of crime spatial patterns. The purpose of this experiment is the prediction of burglary rate and to identify the characteristics of burgled households of the high level 'hotspots' in the study region. In other words, that leads the households within a certain characteristics that have a higher than average risk of victimization. Previously the 'hotspots' within the historical dataset in the study region were identified by the SCS algorithm (discussed in Section 4.6.1). In order to identify and contribution the variables among the characteristics of burgled household, multiple linear regression analysis proposed for this purpose (mentioned in Section 6.8). The obtained results which are shown in table 6.11 indicate that those living in a housed in which there was 'one person' and 'Lone parent' increase the risk of the burglary rate. The risk of burglary rate increases within the household when dealing with detached houses. According to economic levels, personal service occupations positively and significantly affected burglary rate. However, occupations Sales and Customer service negatively and significantly affected burglar rate. The MLR results are used as sufficient data for training a new HNN that is clarified in the experimental work 3. The average percentage accuracy achieved by the new methodology at testing stage is 72% (Figure 6.17). This means that its MAPE is 28%, so the model is reasonable according to Lewis criteria (Section 6.6).

Table 6.11: The final Minitab printouts of significance predictors for a prediction regression models across the 'hotspots' in the study region.

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.03046	0.01183	2.57	0.019	
personalserviceOcc	0.5616	0.1464	3.84	0.001	1.446
sales&customerservice	-0.8512	0.1149	-7.41	0	1.248
h_1	0.5278	0.1016	5.2	0	1.292
one person	0.05892	0.02066	2.85	0.011	1.471
Loneparent	0.21508	0.0618	3.48	0.003	1.922

S = 0.00645414 R-Sq = 81.9% R-Sq(adj) = 74.8%

Figure 6.17: Illustration of the high level 'hotspots' model using a new HNN. The accuracy percentages for the model and for each polygon within the parcel are specified.



6.9.3 Results and discussion

The average prediction performance of a new methodology HNN was compared with non-HNN BP and MLR analysis. The obtained results which are identified in Table 6.10 reveal that there is significant difference in the prediction accuracy between HNN and non-HNN but approximately close to MLR analysis. The average percentage accuracy achieved by the new methodology at testing stage increase 13% compared with the non-hierarchical BP performance. This lead to the acceptance of the hypothesis, that a new methodology has the advantage of improving the performance of a model prediction. The performance is compared in terms of model accuracy and the time which is needed to train the network.

6.10 Summary and Conclusion

This chapter has presented methodologies, used to assist in building predictive crime models, i.e regression methodology, neural network, SCS algorithm and GIS. The methodology was applied to real data on burglary incidence distribution in the study region. Creating these models allows for developing links between social factors and criminals, and the utilization of police resources for crime prevention.

GIS is used in this analysis to integration relevant datasets within census wards. Census datasets that provided a useful source of geo-demographic information were combined in this analysis of population and burglary incidence. Then the obtained results were used in the construction of a predictive model.

The statistical regression model was applied across number of geographical space (parcels) and clusters which were specified by SCS algorithm in the study region (section 4.4). The regression methodology was also used for developing multilevel models of burglary rate. The obtained results from this analysis are based on observed data in the study region. In general the results indicated that the following statistically significance predictors increase the risk of burglary rate: living in the

shared house (f2), living in a household in which there is one person and lone parent, low level of occupancy and unemployed. However these significance predictors decrease the risk of burglary: high level occupancy for instance, the households whose work is professional occupation and living in a household in which all were pensioner. These results are identified and analysed in Section 6.8. Figure 6.7 summarized these significant predictors that decreasing or increasing the risk of burglary. There are correlation between the influence of the number of the above significance predictors and the level of burglary rate. For instance as shown in Figure 6.7, the number of households whose work is professional occupation and living in a household in which all were pensioner were more in low level region of burglary rate than other levels. However the number of households in which were one person and lone parent were more in high level region of burglary rate than other levels. Numbers of shared houses (f2) are more in high-level than other levels. On the other hand as shown in Figure 6.7 the number of a households whose work were skilled trades occupations are approximately the same with respect to the region of burglary levels, which was negatively and significantly correlated within the identified levels (low, middle and high) (Table 6.4 and Figure 6.5).

The statistical methodology in this thesis was combined with an ANN with the aim of developing a new hierarchical neural network (HNNs) approach for generating a more reliable prediction. The training data for the new methodology are prepared by regression methodology. For the demonstration of the performance of the new methodology, both hierarchical and non-hierarchical neural networks were applied on the same selection parcels with MLR. The results reveal that there is significant difference in the prediction accuracy between the new methodology and non-HNN BP-network but it is approximately close to MLR analysis. The results which were achieved by a new HNN methodology achieved without assuming any functional relationship between expletory variables before construct the models. The quality of its result is subjected to the accuracy of the input data which was prepared by regression methodology. However, when applying regression detailed knowledge of

statistical criteria was required before construct the model, for example, linearity and selecting a best subset of predictors. The average percentage accuracy achieved by the new methodology at testing stage increased 13% in average when compared with non hierarchical BP performance. This lead to the acceptance of the hypothesis, that a new methodology has the advantage of improving the performance of a model prediction.

Ameen stated that (cited Frank, 2001: 6) a good model is 'satisfactory in performance relative to the stated objective; logically sound; representative; able to convey information'. Lewis (1982) stated the criteria for judging accuracy (Section 6.6). Thus, the specified obtained models based on the observed data in the study region are reasonable with respect to these criteria.

7 Conclusions and Future Work

This Chapter presents a summary of the research work. The key contributions to knowledge are identified. Suggestions for future research are provided.

7.1 Introduction

At the start of the thesis, the objectives of the research were detailed, with the aim of developing a hybrid modelling approach utilizing relevant principles of Statistics, Geographical Information Systems (GIS), Neural Networks and in general Information Technology for the analysis of observed data. The methodology is applied to real data on crime. The objectives of this research were to:

- Develop a new algorithm based on statistical theory for identifying clusters within spatial data;
- Generate artificial datasets, based on established practice, for use as a proof of concept for a general purpose algorithm for detecting clusters within spatial datasets, which have been used to evaluate the effectiveness of the developed cluster determination algorithm;
- Acquire real world spatial datasets for:
    - testing the developed algorithm and identification of hotspots in the study region;
    - generating predictive models.

- Utilize GIS to accommodate the new cluster detection technique (SCS) in terms of a predictive crime model for:

- mapping;
- display distribution of crime and the corresponding population in a study region;
- visualize the location of obtained clusters which was specified by clustering algorithm SCS;
- display distribution of burglary incidence concentration 'hotspot' which was identified by clustering algorithm SCS;
- identify the total number of cases within polygons(census wards);
- integrating information from a variety of sources such as crime data, population data and census data associated with the observed data.

- Building hybrid predictive models for crimes based on real data (crime).
  - Develop a predictive crime model based on statistical methodology;
  - Develop a new a hierarchical neural network methodology based on statistical methodology and two proposed ANN learning algorithms, unsupervised Self-Organizing Map and supervised back-propagation.

7.1.1 Objective 1

Develop a new algorithm based on statistical theory for identifying clusters within spatial data.

A new cluster detection methodology, called Salar's clustering with significance (SCS) has been developed. This methodology has been developed from knowledge drawn through a hybrid of both statistical and computing techniques. The algorithm for its implementation has employed specific properties of statistical distributions to estimate the position of the centers, and the boundaries of each cluster using subjectively set significance levels. Kullback-Leibler divergence (KL) was employed to compute the distances between the clusters, rather than Euclidean distance. The attribution of the SCS algorithm are: it is easy to implement; no previous knowledge of the data set is required; fewer performed steps lead to a reduction in clustering

time; and the results provide the detail information about the distribution of cases within the dataset. It performs resonably well in terms of memory requirements, running time and cluster quality. It is efficient and objective, leading to improvements in statistical data analysis. In other words, the work to date has demonstrated a significant theoretical contribution to knowledge in the field of statistical analysis of spatial datasets. The algorithm requires that the user first specify the valleys (lower density) in the histograms that are created in the marginal axes for data splitting. This process was used to delimit the concentrated location. The development of the new algorithm is described in Section 3.3.

7.1.2 Objective 2

Generating artificial datasets

This was based on established practice, for use as a proof of concept for a general purpose algorithm for detecting clusters within spatial datasets, which have been used to evaluate the effectiveness of the developed cluster determination algorithm.

The algorithm was tested using artificial datasets, with very promising results. Experimental results demonstrate that SCS is especially suitable for large dataset and even for small sample size. Validation of SCS algorithm is described in Section 3.3.2.

7.1.3 Objective 3

Testing the developed algorithm SCS and identification of hotspots by real world spatial datasets (burglary incidence).

The utility of the clustering methodology (SCS) is demonstrated in experiments applied to several well-known datasets, namely real (crime) datasets; rotation of the

same datasets for instance 30 and 85 degree. The experimental results of validation of the algorithm are described in Section 4.4. The results obtained for identified 'hotspot' were compared with other available algorithms such as CLAP, Satscan and GAM. The outputs are superficially very similar.

7.1.4 Objective 4

Utilization of GIS technology for developing a predictive crime model.

GIS presented in this thesis encompasses the SCS algorithm, statistics and neural networks for developing a hybrid predictive crime model, mapping, visualizing crime data and the corresponding population in the study region, visualizing also the location of obtained clusters and burglary incidence concentration 'hotspots' which was specified by clustering algorithm SCS (object 3). The total number of cases within a census wards are identified and integrated with relevant datasets for this analysis. The census data used throughout this construction provided a useful source of geo-demographic information. The census data in this study were combined crime data and population within census wards. This process is described in Section 5.4.

7.1.5 Objective 5

Building hybrid predictive models for crimes based on real data (crime).

This thesis has presented more than one methodology with the aim of developing a hybrid modelling approach. The methodologies are applied to real data on burglary incidence distribution in the study region. Relevant principles of statistics, GIS, SCS algorithm and ANN are utilized for the analysis of observed data.

The statistical technique, multiple linear regression (MLR) analysis described in this thesis was used to identify potentially significant predictive variables among

characteristics of burgled households and the level of their contribution in the performance of the model and predicting of future crime in the study region. The MLR model involves historical data (crime), population and census data such as Resident Population, Occupation, Qualifications, Socio-economic status, Household composition and Household spaces. The datasets were prepared using statistical and GIS technologies (see objective 4). Regression models were applied across a number of geographical space parcels in the study region and across the identified obtained clusters (object 3). This development of the MLR models is described in Section 6.8.

The statistical methodology in this thesis combined a powerful feature of two proposed ANN learning algorithms. These are the unsupervised Self-Organizing Map and supervised Back-Propagation with the aim of developing a new hierarchical neural network methodology. The methodology involves a training set that was prepared and identified by MLR analysis. So each vector in the network training set represent significant predictive variables of characteristics of burgled households that identified by MLR. The reduction in input vector length produced by the SOM network facilitated the production of a training set for Back-Propagation network. The prediction performance of a new methodology HNN was compared with non-HNN BP and MLR analysis. The obtained results reveal that there is significant difference in the prediction accuracy between HNN and non-HNN, but this is approximately close to MLR analysis. No detailed knowledge of statistical criteria was required before construct a HNN model. The average percentage accuracy achieved by the new methodology at testing stage increase 13% in average when compared with the non hierarchical BP performance. This has lead to the acceptance of the hypothesis that a new methodology has the advantage of improving the performance of a model prediction. The performance is in term of model accuracy and the time which is needed to train the network. This development of the new HNN models is described in Section 6.9.

7.2 Contribution to knowledge

The research presented in this thesis encompasses the disciplines of Cluster detection, Geographical Information System, Crime prediction and Neural Networks. The following contribution to knowledge has made to each of these fields:

7.2.1 Cluster detection

The attribution of a new cluster detection methodology, SCS algorithm are: it is easy to implement; no previous knowledge of the data set is required; the number of clusters is not predetermined; fewer performed steps lead to a reduction in clustering time; and the results provide the detail information, about the distribution of cases within the dataset. It performs resonably well in terms of memory requirements, running time and cluster quality. It is efficient and objective, leading to improvements in statistical data analysis. In other words, the work to date has demonstrated a significant theoretical contribution to knowledge in the field of statistical analysis of spatial datasets.

7.2.2 Geographical Information System

Within the field of GIS, the research has led to development of methodology for modelling data containing multiple functions. The total number of cases within a census wards are identified and integrated with relevant datasets. The obtained datasets were used for predictive crime modelling. SCS also contributes to GIS, visualizing the location of obtained clusters and burglary incidence concentration 'hotspots' which was specified by this clustering algorithm.

7.2.3 Crime prediction

The methodologies, regression analysis, GIS, neural networks and SCS algorithm are combined in this thesis to assist in building predictive crime models. The SCS algorithm was utilized for detecting clusters and to identify the spatial level of concentration of burglary incidence in the study region. GIS was utilized to integrate

relevant information from a variety of sources such as crime data, population data and census data associated with the observed data. Then the obtained results were used in the construction of a predictive model. Multiple linear regression analysis was used to identify potentially significant predictive variables, level of their contribution in the performance of the model and predicting of future crime. The statistical regression model was applied across a number of geographical space (parcels) and clusters which were specified by SCS algorithm in the study region (section 4.4). The regression methodology was also used for developing multilevel models of burglary rate. The obtained results from this analysis are based on observed data in the study region. In general the analysis reveals a number of predictors that increase the risk of burglary. Specifically, living in a household in which there is 'one person' and 'lone parent'. The study aims to provide suggestions in this case, to reduce susceptibility to victimization and guide policy. Their safety may be improved if they are renting or sharing with another instate to living a lone. Secure the home with sturdy windows and doors. A security guard is effective in reducing the risk of burglary rate. Whereas which living in households in which all were pensioner the risk is low. This is because a household more time stay at home therefore decreasing criminal opportunity. Ratcliffe (2001) showed that the highest probability for residential burglaries was the period that most people were at work. In addition the analysis reveals the correlation between the influences of these predictors (one person, lone parent and all pensioners) and the risk of burglary (see Figure 6.7). The number of households in which were one person and lone parent were more in high level region of burglary rate than other levels. The number of households where all were pensioner were more in low level region of burglary rate than other levels. The influence of the household occupations in this study is related to blending of two factors: Occupancy level of the occupation and the time that people are at work. For household where occupations are with a high level of occupancy, such as professional occupation, the risk of burglary is low. Whereas those in elementary or intermediate occupations are at risk of burglary. This may be because the household in high level of occupancy can use security, which is effective

in reducing the risk of burglary rate. However those in low level of occupancy have lack of using security. In the case of unemployed, this study found that within the frame-work of clusters model, unemployment increases risk. Figure 6.7 shows that the numbers of households whose unemployment are more in high level region of burglary rate than other levels. Increased unemployment rates lead to higher property crime rates (Edmark 2005; Gorr &olligshlagers 2003). In addition the analysis reveals the models of the clusters neighborhood have approximately similar contributions to significant predictors.

7.2.4 Neural Network

Within the field of neural network, this research has led to development of a new hierarchical neural networks methodology to improve the performance of the prediction models. The average percentage accuracy achieved by the new methodology at testing stage increase 13% compared with the non hierarchical BP performance.

7.3 Limitation

Naturally the quality of results is subject to the accuracy of the used data. The potential variables that are missing in developing a predictive model thought to be useful and they would have to be collected or measured are:

- Car ownership, migration of people for 1 year before the census, number of students, information on demographic movements, repeat crimes and whether the resident owns their property;
- Unemployment could be measured against the total number of people of employment age within polygon rather than the total number of people within the polygon;

- In the case of policing data. Not all incidents that occur are either reported by the public or recorded by the police;

- One limitation of the process to associate polygon with clusters. That is in some cases overlaying complete polygon in the boundary of the clusters become somewhat problematic.

7.4 Suggestions for Future Work

This section describes potential further areas research that have arisen during this work:

- Clustering

Chapter three presented a new methodology for cluster detection (SCS). Three suggestions for future research are to replicate the procedures of the SCS algorithm within each of the obtained clusters, and to apply rotation to the main diagonal of the ellipse of the obtained clusters. This would simplify and clarify the data structure. This process helps the analyst to observe clusters more clearly. Re-suggest width of confidence intervals according to the clustring result. Applying the algorithm on data which include elements related to temporal elements.

- Data collection and measuring

The development of any model requires consideration of a number of issues such as data representation and structuring. This is important in the assessment of a robust model. It could be important if the collected data include information on demographic movements (for how long the resident has been in that place), Care ownership, migration of people for 1 year before the census, number of students, repeat crimes and whether the resident owns their property. Crime rate could be measured against the house type rather than the number of households. Unemployment could be measured against the total number of people of employment age within polygon. Examining the relationship between crime rate and the level of neighborhoods' economic status may also help

improve models. As a future extension of prediction crime model, create a model using Geographically Weighted Regression (GWR) is a method of analysing spatially varying relationships.

- Neural Networks

Chapter six presented the utility of ANN for building a predictive model. Suggestions for future research in this case are to utilize General Regression Neural Network (GRNN) and Radial basic function (RBF) for building predictive models. Then compare the results obtained using these techniques to assess the best approach.

7.5 Conclusions

This research has developed a hybrid modelling approach. A new cluster detection methodology called Salar's Clustering with Significance (SCS) has been developed, based on knowledge drawn from both statistical and computing techniques. It is easy to implement; no previous knowledge of the data set is required; the number of clusters is not predetermined; fewer performed steps lead to a reduction in clustering time; and the results provide the detail information, about the distribution of cases within the dataset. It performs reasonably well in terms of memory requirements, running time and cluster quality. It is an efficient and objective, leading to improvements in statistical data analysis. Empirically has demonstrated the suitability of the algorithm for large dataset and even for small sample size. Furthermore, the algorithm has the benefit to identify 'hotspots'. The obtained results were compared with the results of other available algorithm such as CLAP, Satscan and GAM. The outputs are very similar. This indicated that the SCS algorithm performs well.

The methodology was applied to real data on burglary incidence distribution in the study region given background population and census datasets to reveal information relating to spatial distribution of the burglary incidence.

GIS was used to develop a methodology for modelling data containing multiple functions. The methodology was used for measuring crime rate and identified the levels of crime within the census ward polygon (the smallest spatial unit for which the entire UK Census is publicly a vailable) in the study region. The methodology can be used to measure crime rate for future work.

Predictive crime models were created using several existing methodologies, such as regression analysis, Geographical Information System, Neural Networks and SCS algorithm. Creating these models allows for developing links between social factors and criminals, and the utilization of police resources for crime prevention. The models were applied across a number of geographical high burglary incidence concentration 'hotspots', a number of geographical space (parcel) and across the clusters. Which were identified by a new methodology (SCS) and GIS was used to visualize their locations in the study region. Hotspots analysis has been an important approach for the explanation and prediction of crime spatial patterns. The households within a certain characteristics that have a higher than average risk of victimization were identified. The accuracy percentage results of the high level 'hotspots' model was 72%, so the model is reasonable according to Lewis criteria. The model was applied across 19 parcels in the study region. Each parcel's data were modelled separately in order to identify significant predictors of burglaries, through the potential of the examination of different characteristic of burgled households. The detail of the models can be found in Appendix E. The predictive burglary models which were obtained from this analysis were tested with a new data set and the results are reasonable according to Lewis criteria. Cluster analysis is an important approach for the explanation and prediction of crime spatial patterns. Within the cluster criminal phenomena are examined within the area of concern. This is because concern in certain location helps to identify the problem associated to the characteristic of the people within their location, and this leads to increase the predictive accuracy of the model. The accuracy percentage results which were identified as 68%,,71%,70%,80%,72% for the clusters A,B,C,D,F respectively

indicated that the models were reasonable according to Lewis criteria. The obtained results from this analysis are based on observed data in the study region. In general the analysis revealed a number of predictors that increases the risk of burglary. Specifically, living in a household in which there is 'one person', 'lone parent' and household in elementary or intermediate occupations. This study found that within the framework of clusters model, unemployment increases risk. For the influence of Household space, the results indicated that the risk of burglary rate increases within the household living in shared houses (f2).

Ameen stated that (cited Frank, 2001: 6) a good model is 'satisfactory in performance relative to the stated objective; logically sound; representative; able to convey information'. Lewis (1982) stated the criteria for judging accuracy (Section 6.6). Thus, the specified obtained models based on the observed data in the study region are reasonable with respect to these criteria.

Some of the limitations of the predictive model include missing or measureing some variables (outline in Section 7.3) may influence the model accuracy. Addressing these issues, each according to the nature of the problem, can helpful developing models as future research.

Within the field of neural networks, a new hierarchical neural networks methodology was developed to generate a more reliable prediction model. The average percentage accuracy achieved by the new methodology at testing stage increase 13% in average compared with the non hierarchical BP performance.

8 References

Adya, M. and Collopy, F. (1998). How Effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation. Journal of Forecasting, 17, 481-495.

Alex, R.P. and Stephen, G.T. (2002). Rational choice and criminal behavior. USA: Routledge.

Alvin, C.R. (2002). Methods of Multivariate Analysis. 2nd ed, USA: John Wiley.

Ameen, J., Neale, R.H. and Abramson, M. (2003). An application of regression analysis to quantify a claim for increased costs. Journal of construction Management and Economics, 21,159-165.

Anderson, D. (1992). Artificial Neural Networks Technology. New York, Rome Laboratory Report RL/C3C 13441-5700.

Andrew, K. (2000). Basics of MATLAB. USA: CRC Press.

Andrew, F.W. (2004). The semi-automated classification of sedimentary organic matter and dinoflagellate cysts in palynological preparations. Thesis(PhD). Glamorgan University.

Armstrong, J.S. (2006). Findings from evidence- based forecasting: Methods for reducing forecast error. International Journal of forecasting, 22(3), 583- 598.

Basheer, I.A. and Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods, 43, 3-31.

Beale, R. and Jackson, T. (1997). Neural computing. 5th ed, London: Institute of Physics publishing.

Begg, R., Kammruzzaman, J. and Sarker, R. (2006). Neural Networks in Healthcare Potential and challenges. UK: IDEA.

Bennett, C.H, Gacs, P. and Zurek, W. (1998). Information Distances. IEEE Transactions on information theory, 44(4), 1407-1423.

Berry, M.J. and Linoff, G.S. (2004). Data Mining Techniques. 2nd ed, Indiana: John Wiley.

Bigi, B. (2003). Using Kulback-Leibler distance for text categorization. 25th European conference on IR Research ECIR, 2633, 305-319.

Bijleveld, C. and Van D.T. (1998). *Longitudinal Data Analysis, Designs, Models and Methods*. London: Sage publications.

Bishop, C.M. (1994). Neural Networks and their application. Review of Scientific Instruments, 65(6), 1803-1832.

Bolzan, A., Machado, R. and Piaia, J. (2008). Egg hatchability prediction by multiple linear regression and artificial neural networks. Revista Brasilira de ciencia Avicola, 10(2).

Bowers, K.B. and Hirschfield, A. (1999). Exploring links between crime and disadvantage in north-west England: an analysis using geographical information systems. INT.J. Geographical information Science, 13(2), 159-184.

Bowers, K.B., Johnson, S., D. and Pease, K. (2004). (Re)Victimis risk, housing type and area: a study of interactions. Journal of Quantitative Criminology, 14307-30.

Bozdogan, H. (2004). Statistical Data Mining and Knowledge Discovery. London: CRC Press Company.

Bradley, P. (1998). Refinig initial points for k-means clustering. Proceedings 15th International conf, on Machine Learning, San Francisco, 91-99.

Carcach, C. and Huntley, C. (2002). Community Participation and Regional Crime. Australian Institute of Criminology, 22.

Carling, A. (1993). Introducing Neural Networks. South Oak Lane: Sigma press.

Chen, C.R. and Ramaswamy, H.S. (2000). Aneuro-Computing approach for modelling of residence time distribution of carrot cubes in a vertical scraped surface heat exchanger. Food Research International, 33,549-556.

Chetwin, A., Johns, K., Barwick, H. and Carswell, S. (2005). Literature Review: Police Practice in Reducing Residential Burglary. Newzealand Ministry of Justice.

Chiehwen, E., Jacobson, H. and Mas, F. (2004). Evaluating the disparity of female breast cancer mortality among racial groups-a spatiotemporal analysis. International Journal of Health Geographics, 4(3).

Cigizoglu, H. and Alp, M. (2006). Generalized regression neural network in modeling river. Advances in engineering software, 37, 63- 68.

Cios, K., Pedrycz, W. and Swiniarski, R. (2000). Data Mining Method for Knowledge Discovery. USA: Kluwer Academic Publishers.

Cohen, J., Gorr, J.W. and Olligschlaeger. (1993). Modelling Street-Level illicit Drug Markets. School of public policy and Management Carriage Mellon University.

Corcoran, J. (2003). Computational techniques for the geo- temporal analysis of crime and disorder data. Thesis (PhD). Glamorgan University.

Corcoran, J., Wilson, I.D., Lewis, O.M. and War, J.A. (2001). Data clustering and Rule Abduction to facilitate crime hotspot prediction. Lecturer Notes in Computer Science, 2206, 807-822.

Corcoran, J.J., Wilson, I.D. and Ware, J.A. (2003). Predicting the geo-temporal variations of crime and disorder. International Journal of Forecasting, 19, 623-634.

Cullen, J.B. and Levitt, S.D. (1999). Crime, Urban Flight, and the Consequences for Cities. The review of economics and statistics, 81(2), 159-169.

Craglia, M., Haining, R. and Signoretta, P. (2001). Modeling high-intensity crime areas in English cities. Urban studies, 38(11), 1921-1941.

Dan, W.P. (1996). Artificial Neural Networks Theory and Applications. London: Prentice Hall.

Daniel, T.L. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. UK: John Wiley.

David, G.K., Lawrence, L.K. and Keith, E.M. (1998). Applied Regression Analysis and other Multivariable Methods. London: Duxbury Press.

Dayhoff, J. (1990). Neural Network Architectures. New work: Van No strand Reinhold.

Deadman, D. (2003). Forecasting residential burglary. International Journal of Forecasting, 19, 567-578.

Deboeck, G. (1998). Financial Applications of Self-Organizing Maps. American Heuristics Electronic Newsletter.

Deelers, S. and Auwatanamonghol, S. (2007). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance. PWASET, 12, 323-328.

Dielman, T.E. (2005). Applied Regression Analysis. USA: Brooks/Cole Thomson Learning.

Douglas, C. F. (2001). Forecasting Tourism Demand: Methods and Strategies. UK: Butterworth-Heinemann.

Edmark, K. (2005). The Effects of Unemployment on Property Crime: Evidence from a period of unusually large swings in the business cycle. Scandinavian Journal of Economics, 107(2), 353-373.

Egeli, B., Ozturan, M. and Badur, B. (2003). Stock Market Prediction Using Artificial Neural Networks. [Online] Available from: http://www.hicbusiness.org/biz2003proceedings/Birgul%20Egeli.pdf.

Ellacott, S. (1996). Neural Networks. UK: ITP.

Everitt, B.S. and Lesse, M. (2001). Cluster Analysis. UK: Arnold.

Fausett, L. (1994). Fundamentals of Neural Networks. USA: Prentice Hall International.

Felson, M. and Poulsen, E. (2003). Simple indicators of crime by time of day. International Journal of forecasting, 19, 598- 601.

Feng, C., Samuel, Y. and Kusiak, A. (2006). Selection and Validation of predictive regression and neural network models based on designed experiments. IIE Transactions, 38, 13-23.

Ferandez, G. (2003). Data mining using SAS application. USA: CRC Press.

Field, A. (2005). Discovering statistics using SPSS. 2nd ed, London: Sag.

Fotheringham, A. and Zhan, F. (1996). A comparison of three exploratory methods for cluster detection in spatial point patterns. Geographical Analysis, 28, 200-218.

Frank, E.H. (2001). Regression Modelling Strategies with Applications to linear Models, logistic Regression, and survival Analysis. USA: Springer.

Fraser, C. (2008). Business statistics for Competitive. USA: Springer.

Gartin, J. and Goor, W. GIS and crime Analysis. [Online] Available from; http://urisa.Org/files/Getis Vol 12No2.pdf.

Geoffrey, M. J. (2004). Current practices in the spatial analysis of cancer. International Journal of Health Geographic, 3, 22-26.

Gorr, W., Anselin, L., and David, C. (2000). Spatial Analyses of crime. Criminal Justice, 4, 213-243.

Gorr, W. and Harries, R. (2003). Introduction to crime forecasting. International Journal of forecasting, 19(4), 551- 555.

Gorr, E.L. and Kurland, K.S. (2007). GIS Tutorial. UK: Transatlantic Publishers Group ltd.

Gorr, W., Olligschlaeger, A. and Thomposon, Y. (2003). Short-term forecasting of crime. International Journal of forecasting, 19, 597- 594.

Grover, V. and Adderley, R. (2006). Review of current crime prediction Techniques. [Online] Available from: http://www.Tech.port.Ac.Uk/staffweb/andersod/Hoc/Reports/Uploads/ViRas9355079.doc

Hartigan, J.A. (1975). Clustering Algorithms. New York: Wiley.

Haykin, S. (1999). Neural Networks. 2nd ed, London: Prentice Hall international.

Hettmansperger, T. and Thomas, H. (2000). Almost nonparametric inference for repeated measures in mixture models. J.R. statist.Soc.B, 62(4). 811-825.

Hirschfield, A. and Bowers, K. (2001). Mapping and Analysing Crime Data. London: Taylor and Francis.

Hollman, J., Tresp, V. and Simula, O. (1999). A Self-Organizing Map for clustering Probabilistic Models. IEEE Transactions on Networks, 2, 946-951.

Hopgood, A. (1993). Knowledge-Based Systems for Engineers and Scientists. London: CRC press.

Jain, A. and Dubes, R. (1988). Algorithms for Clustering Data. London: Prentice-Hall international, 55-133.

Jain, A. K. and Mao, J. (1996). Artificial Neural Networks: A tutorial. IEEE Computer, 29(3), 31-44.

Jain, A., Murty, M. and Flynn, R. (1999). Data Clustering. ACM Computing surveys, 31(3), 264-323.

James, J.N. (2004). Establishing the statistical relationship between population size and UCR crime rate: Its impact and implications. Science Direct, 32(6), 547-555.

Jarvis, P. (2006). Determining Geographical Causal Relationships through the Development of Spatial Cluster Detection and Feature Selection Techniques. Thesis (PhD), Glamorgan University.

Jarvis, P.S., Wilson, I.D. and Kemp, S.E. (2006). The Application of a New Attribute Selection Technique to the Forecasting of Housing Value Using Dependence Modelling. Neural Computing and Application, 15(2), 136-153.

Jimenez, L. (1995). High Dimensional Feature Reduction via Projection Pursuit. Thesis (PhD), Purdue University.

John, E. and David, W. (1995). Crime and place crime prevention studies. UK: Willow Tree Press.

Johson, C.P. (2000). Crime Mapping and Analysis Using GIS. Conference on Geometrics in Electronic Governance, Pune University Campus.

Joseph, T. and David, E. (2001). Artificial Neural Networks in Prostate Cancer. Lab Medical international, 18(3).

Kaufman, L. and Peter, J. (2005). Finding Groups in Data An Introduction to cluster Analysis. USA: John Wiley.

Kenneth, D. (1975). Cluster Analysis. Sociological Methodology, 6(1), 59-128. [Online] Available from: http://link.jstor.org/sici?sici=0081-1750%28 1975%3C59%3ACA%3F2.o-Co%3B2-2.

Kershaw, C., Nicholas, S. and Walker, A. (2008).Crime in England and Wales. Home Office, UK.

Klassen, A., Kulldorff, M. and Curriero, F. (2005). Geographical clustering of prostate cancer grade a diagnosis, before and after adjustment for risk factor. International Journal of health geographic, 4(1).

Kohonen, T. (2001). Self-Organizing Maps. 3rd ed, London: Springer.

Krzanowski, W.J. (1998). An Introduction to statistical Modeling. London: ARNLD.

Kung, S.Y. (1993). Digital Neural Network. USA: PTR Prentice Hall.

Kutner, M.H., Christopher, J.N. and John, N. (2004). Applied linear Regression Models. 4th ed, UK: Mc Graw Hill.

Lawis, C.D. (1982). Industrial and business forecasting methods. London: Butter worth Scientific.

Lawrence, J. and Fredrickson, J. (1998). Brain Marker User's Guid and Reference Manual, 7th ed, California Scientific Nevada City, CA95959.

Lehmann, E.L. and Romano, J.P. (2005). Testing Statistical Hypotheses.3rd ed, USA: Springer.

Likas, A. and Verbeek, J. (2003). The Global k-means clustering algorithm. Pattern Recognition, 36(2), 451-461.

Liu, H. and Brown, D. (2003). Criminal incident prediction using appoint- pattern-based density model. International Journal of Forecasting, 19, 603- 622.

Liu, H. and Yu, X. (2009). Application Research of K-means Clustering Algorithm in Image Retrieval System. Proceedings of Second Symposium International Computer Science and Computational Technology, 274-277.

Lus, L., Fotouhi, F., Deng, Y. and Brown, J. (2004). Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC Bioinformatics, 5, 172-178.

Maheswaran, R. and Craglia, M. (2004). GIS in Public Health Practice. USA: CRC Press LLC.

Malczewski, J. and Poetz, A. (2005). Residential Burglaries and Neighborhood Socioeconomic context in London, Ontario. Professional Geographer, 57(4), 516-529.

Margaret, W.E., David, J.L., Roy, D.A. and Anthony, H. (2002). Data modelling and the application of a neural network approach to the prediction of total construction costs. Construction Management and Economics, 20, 465-472.

Martinez, W.L. (2002). Computational Statistics handbook with MATLAB. London: Chapman and Hall CRC.

MathWorks, Inc. (2007). MATLAB 7 Graphics.

Mat Isa, N.A., Mashor, M.Y. and Othman, N.H. (2002). Diagnosis of Cervical cancer using Hierarchical Radial Basic Function network. Proceedings of the International conference on Artificial Intelligence in Engineering and Technology, Kota, 458-463.

Mawby, R.I. (2001). Burglary. UK: Willan.

Mclachlan, G. and Peel, D. (2000). Finite Mixture Models. USA: John Wiley.

Meredith, W., Introduction GIS. [Online] Available from: http://www.sul.stanford.edu/depts/gis/whatgis.html.

Metrotra, K., Mohan, C.K. and Ranka, S. (1997). Artificial Neural Networks. UK: MIT Press.

Michie, D. and Taylor, C. (1994). *Machine Learning, Neural, and Statistical Classification*. London: Ellis Harwood.

Miller, T.R. and Cohen, M.A. (1996). Victim costs and consequence: A new look, National Institute of Justice Report, NCJ 155282.

Miloslavsky, M. and Mark, J. (2003). Fitting of mixtures with unspecified number of components using cross validation distance estimate. Computational Statistics and Data Analysis, 41, 413-428.

Mirkin, B. (2005). Clustering for Data Mining. USA: Hall/CRC.

Myrvoll, T.A. and Soong, F.K. (2003). On divergence based clustering of Normal distributions and its application to HMM adaptation. Eurospeech-Hongkong, 1517-1520.

Neill, D. (2006). Detection of spatial and spatio-Temporal Clusters. Thesis (PhD), Carnegie Mellon University.

Nick, T. (2002). Analysis for crime prevention. USA: Willan.

Niknam, T., Nayeripour, M. and Bahmani, B. (2007). Application of a New Hybrid optimization Algorithm on Cluster Analysis. European Journal of Operation Research, 3(177), 1400-1408.

Nolan, J. J. (2004). Establishing the statistical relationship between population size and UCR crime rate: Its impact and implications. Journal of Criminal Justice, 32(6), 547-555.

O'Brien, R.M. (2007). A caution Regarding Rules of Thumb for variance Inflation Factors. Quality and Quantity, 41(3), 673-690.

Oatley, G. and Wart, B. (2003). Crimes analysis soft ware: Pins in maps, clustering and Bayes net prediction. Expert systems with Applications, 25, 569- 588.

Oatley, G. , Zelezniknow, J. and Ewart, B. (2005). Matching and predicting Crimes. [Online] Available from: http://popcenter. Org/ Library/ crime prevention / volume % 2004/ 02- Sherman. PDF.

Olligschlaeger, A. and Gorr, W.L. (2001). Crime Hot spot Forecasting: Modeling and Comparative Evaluation. National Institute of Justice.

Olligschlager, A. M. (1997). Artificial Neural Networks and crime Mapping. [Online]Available from: http://www. Popcenter. Org/ Library / crime prevention / volume% 2008? Artificial%20 Neural %20 Networks%20 and %20 crime %20 Mapping. Pdf.

Open University. (2002). T396 Artificial Intelligence for Technology. 6th ed, UK.

Park, Y. and Lek S. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modeling, 160, 265-280.

Pedhazur, E.J. and Schmelkin, L.P. (1991). Measurement, design and analysis: An integrated approach. Lawrence Erlbaum Associates: Hillsdale.

Pelleg, D. (2004). Scalable and Practical Probability Density Estimators For Scientific Anomaly Detection. Thesis (PhD), Carnegie Mellon University.

Rachel, B. (2005). Crime Analysis and Crime Mapping. London: SAGE publications.

Ratcliffe, J.H. (2001). Policing Urban Burglary. Australian Institute of Criminology, 213.

Ratcliffe, J.H. (2004). The hotspot Matrix: A framework for the spatio-Temporal Targeting of crime Reduction. Police Practice and Research, 5(1), 05-23.

Rencher, A. (2002). Methods of Multivariate Analysis. USA: John Wiley.

Rich, W. (1995). Geographic information Science and Crime analysis [online]. Available at http://dusk2.geo.orst.edu/ucgis/web/apps white/crime.html. [Accessed 19 June 2007].

Richards, W., Ameen, J. and Coll, A. (2005). The community General Dental Practioner. British Journal of Health care Management, 11(10), 308-12.

Rigol, J. P., Jarvis, C.H. and Stuart, N. (2001). Artificial neural networks as a tool for spatial interpolation. INT.J. Geographical information Science, 15(4), 323-343.

Romesburg, H. (2004). Cluster Analysis For Researchers. USA: Lulu Press.

Ronald, V.C. and Marcus, F. (1993). Routine activity and Rational choice. UK: Transaction Publishers.

Roncek, D. and Maier, P. (1991). Bars, Blocks, and crimes Revisited: Linking the theory of routine activates to the empiricism of “Hot spots”. Criminology, 29 (4), 725- 753.

Rouse, D.M. (2005). Estimation of Finite Mixture Models. Thesis (MSC), North Carolina state University.

Sandhya, S. (2006). Neural Networks for Applied Sciences and Engineering. USA: Taylor and Francis Group.

Sharma, S. (1996). Applied Multivariate Techniques. New York: John Wiley.

Snee, R.D. (1977). Validation of Regression Models: Methods and Examples. Techno metrics, 19(4), 415-425.

Stanley, L. (2002). Statistical finite mixture Models classification and cluster Analysis [online].Available at http://www.Uic.edu/classes/idsc/ids594/notes.doc. [Accessed 10 February 2007].

Stanley, L. (2001). Notes on Cluster Analysis. [Online] Available from: http:// www.Uic.edu/classes/idsc/ids594/notes.doc.

Stillwell, J. and Clarke, G. (2004). Applied GIS and spatial Analysis. UK: John Wiley.

Swingler, K. (1996). Applying Neural Network. London: Academic press.

Teanby, N.A., Kendall, J.M. and Bann, M.V. (2004). Automation of Shear-Wave splitting measurements using cluster analysis. Bullentin of Seismological Society of America, 94(2), 453-463.

Thomas, D. (2002). Statistical Pattern Recognition for Breast cancer Research: Comparison of Theory Driven General Linear Model Methodologies with Data Driven Artificial Neural Network. Thesis (PhD).

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. statist. Soc. B, 63(2), 411-423.

Tobin, A.D. (2009). Learning MATLAB. USA: SIAM.

Tracey, B. (2001). Development and statistics Directorate. A publication of policing and Reducing crime unit. Home Office Research.

Tseloni, A., Wittebrood, K., Farrell, G. and Pease, K. (2004). Burglary Victimization in England and Wales, the United States and the Netherland. Brit. J. Criminal, 44, 66-91.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-Organizing Map. IEEE Transactions on Neural Networks, 11(3), 586-590.

Von, V. (2004). Statistical Learning with Similarity and Dissimilarity function. Thesis (PhD), Berlin University.

Wang, F. (2005). Geographic information Systems and crime Analysis. London: Idea Group.

Washington, D. (1988). Discriminate Analysis and Clustering. UK: National Academy press.

Wheeler, D. C. (2007). A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003. International Journal of health geographic [online]. Available at http://www.Pubmedcentral. nih.gov/arti clerender.fcgi?artid=1851703. [Accessed 27 March 2007].

Wise, S. and Craglia, M. (2007). GIS and Evidence-Based policy making. USA: CRC Press, 115-116.

Wong, M.A. and Lane, T. (1983). A Kth nearest neighbour clustering procedure. Royal statistical society B, 45, 362-368.

Wortley, R. and Mazerolle, L. (2008). Environmental Criminology and Crime Analysis. UK: Willan.

Yona, G. (1999). Methods for global organization of the protein sequence space. Thesis (PhD), Hebrew University.

Yu, G. (2007). Social Network Analysis Based on BSP Clustering Algorithm. Communications of the IIMA, 7(4), 39-45.

Zhang, G., Patuwo, B. E. and HU, M.Y. (1998). Forecasting with Artificial Neural Networks. International Journal of Forecasting, 14, 35-62.

Zhang, X. and Li, Y. (1993). Self-Organizing Map as a new method for Clustering and Data analysis. International Joint conference on Neural Networks, 3, 2448-24

Appendix A

This Appendix gives MATLAB program on which

- Implement formula 3.7 (KL distance) to measure the distance between two clusters.
- Implement SCS algorithm on an artificial data set.

Program 1:'Kullback-Leibler distance

```
function [p,q,k,d]=KL(x1,y1,x2,y2,meanp,meanq)
p=cov(x1,y1);
q=cov(x2,y2);
a=inv(p)+inv(q);
b=meanp-meanq;
c=b';
k=a*b*c+p*inv(q)+q*inv(p)-2*eye(2,2);
d1=trace(k);
d=d1/2
```

Program 2: Implement SCS algorithm on artificial data set

```
>> %Generating Random variables, using formula 3.1
>> r1=randn(1,6000)*5;
>> r2=randn(1,6000)*7+5;
>> %maximum likelihood for each distribution on the line parallel to x-axis and y-axis
>> [cr1,cm1]=mle(r1);
>> [cr2,cm2]=mle(r2);
>> %standard deviation
>> sx=std(r1);
>> sy=std(r2);
>> figure;
>> %values to create the ellipse
>> t=linspace(0,2*pi);
>> axes('position',[0.2 0.2 0.9 0.9]);
>> %plot the data points into their marginal dimensions
>> plot(r1,r2,'.')
>> hold on;
>> %plot the centers of the clusters
>> plot(cr1,cr2,'*')
>> hold on;
>> %perform step 3
>> plot(cr1(1)+1.96*sx*cos(t),cr2(1)+1.96*sy*sin(t),'color','y');
>> hold on;
>> axes('position',[0 0.2 0.2 0.9]);
>> %draw histogram on ly-axis
```

```
>> histfit(r2,100);
>> hold on;
>> set(gca,'view',[90 270]);
>> axes('position',[0.2 0 0.9 0.2]);
>> %draw histogram on lx-axis
>> histfit(r1,100)
```

Appendix B

This Appendix gives the MATLAB program to:

- Implement SCS algorithm on real datasets for identification of 'hotspots'.
- Creating the rotation of the data points.

1 Implement SCS algorithm on real data sets

```
% crx &cry co-ordinate of crime data.
%  n number of crime datasets.
% px&py co-ordinate of population data.
% pz number of population with respect to co-ordinate px&py.
% np number of population.
% cx 1Xmx vector of maximum likelihood for distributions on the line %parallel to x-axis.
%cy cx 1Xmy vector of maximum likelihood for distributions on the line %parallel to y-
axis.
%Sx vector of standard deviation for distributions on x-axis.
% Sy vector of standard deviation for distributions on y-axis.
% inx 1Xmx vector of delimitation of distributions on x-axis.
% iny 1Xmy vector of delimitation of distributions on y-axis.
% ll crime rate within study region.
%c1 number of crime in each cluster.
% p1 number of population in each cluster.
%active matrix of crime rate.
% indc index of crime rate.


function [c1,p1,active]=plotdata1(crx,cry,n,px,py,pz,np,cx,cy,sx,sy,mx,my,inx,iny,ll)
figure;
t=linspace(0,2*pi);
axes('position',[0.2 0.2 0.9 0.9]);
plot(crx,cry,'.y')
hold on;
plot(px,py,'.b')
hold on;
[c1,p1,active,indc]=crimerate(crx,cry,n,px,py,pz,np,inx,mx,iny,my);
for i=1:mx
        for j=1:my
        if active(i,j)>0
            plot(cx(i),cy(j),'k*')
        hold on;
        else
        end
    end
    hold on;
end
hold on;
for i=1:mx
    for j=1:my
        if active(i,j)>= ll
        plot(cx(i)+1.96*sx(i)*cos(t),cy(j)+1.96*sy(j)*sin(t),'color','r');
```

```
hold on;
        elseif active(i,j)< ll &active(i,j)>0
     plot(cx(i)+1.96*sx(i)*cos(t),cy(j)+1.96*sy(j)*sin(t),'color','g');
hold on;
     end
    hold on;
end
hold on;
end
hold on;
axes('position',[0 0.2 0.2 0.9]);
hist(cry,100);
hold on;
set(gca,'view',[90 270]);
axes('position',[0.2 0 0.9 0.2]);
hist(crx,100);
hold on;
m1=mean(crx)

function[c,p,rate,indc]=crimerate(a,b,n,ap,bp,r,np,inx,n1,iny,n2)
for j=1:n2
    for i=1:n1
        mc=0;
        for k=1:n
            if a(k)>=inx(i) & a(k)<inx(i+1) & b(k)>=iny(j) &b(k)< iny(j+1)
              mc=mc+1;
            else
            end
        end
        c(i,j)=mc;
    end
end
for j=1:n2
    for i=1:n1
        mp=0;
              for k=1:np
            if ap(k)>=inx(i) & ap(k)<inx(i+1) & bp(k)>=iny(j) &bp(k)< iny(j+1)
              mp=mp+r(k);
                        else
            end
        end
        p(i,j)=mp;

    end
end
q=0;
for j=1:n2
    for i=1:n1
    if p(i,j)==0 | c(i,j)==0
        rate(i,j)=0;
    else
        rate(i,j)=c(i,j)/p(i,j);
        q=q+1;
        indc(q,1)=i;
        indc(q,2)=j;
    end
    end
    end
```

```
function[qq,jj]=group(x,y,n,indc,n1,n2,indx,indy)
m=0;
for i=1:n1
   for j=1:n2
      if i=indc(:,1) &j=indc(:,2)
         [a,t]=groupsub(x,y,n,indx,indy,i,j);
         kk=1;
         for i1=m+1:m+t
            qq(i1,1)=a(kk,1);
            qq(i1,2)=a(kk,2);
            kk=kk+1;
         end
         m=i1;
         jj(:,1)=m;
      else
      end
   end
end

function[a,t]=groupsub(x.y,n,indx,indy,i,j)
t=0;
for k=1:n
   if indx(i)<= x(k) & indx(i+1)>=x(k)& indy(j)<=y(k)& y(k)<=indy(j+1)
      t=t+1;
      a(t,1)=x(k);
      a(t,2)=y(k);
   else
   end
end

function [c1,c2,c3,c4,c5,c6,c7,c8,c9,c10]=datasub(data,n,t,g)
l=0;
for j=2:g+1
   k=0;
   for i=1:n
      if t(j-1)<=data(i) &data(i)<t(j)
         k=k+1;
         d(k)=data(i);
      else
      end
   end
   for m=k+1:n
      d(m)=NaN;
   end
   l=l+1;
   a(:,l)=d;
   d=[];
end
c1=a(:,1);
c2=a(:,2);
c3=a(:,3);
c4=a(:,4);
c5=a(:,5);
c6=a(:,6);
c7=a(:,7);
c8=a(:,8);
c9=a(:,9);
c10=a(:,10);
c1(isnan(c1))=[];
c2(isnan(c2))=[];
```

```
c3(isnan(c3))=[];
c4(isnan(c4))=[];
c5(isnan(c5))=[];
c6(isnan(c6))=[];
c7(isnan(c7))=[];
c8(isnan(c8))=[];
c9(isnan(c9))=[];
c10(isnan(c10))=[];
```

2- Function for creating rotation datasets

```
% Given an angle b, creates the corresponding.  Implement formula 4.1
function[r]=rotation2(x,y,b)
t=b*pi/180;
xt=cos(t)*x-sin(t)*y;
yt=sin(t)*x+cos(t)*y;
r(:,1)=xt;
r(:,2)=yt;
```

C1: Details of the obtained results of crime rate (expressed as the number of crime observed in that cluster per the combined population). The rotation of data points of 85 degree utilized and using the clustering algorithm SCS.

Number of crime

14	412	218	77	20	0	0	0	0
17	177	113	128	146	23	0	0	0
58	260	71	46	115	313	44	5	5
4	49	94	31	329	600	111	2	20
43	66	33	42	542	946	163	119	133
40	93	68	118	269	690	204	220	149
26	48	138	75	174	297	11	6	1
3	87	26	74	120	66	0	0	0
25	50	29	43	282	1259	128	2	0
0	4	1	0	3	103	19	0	0

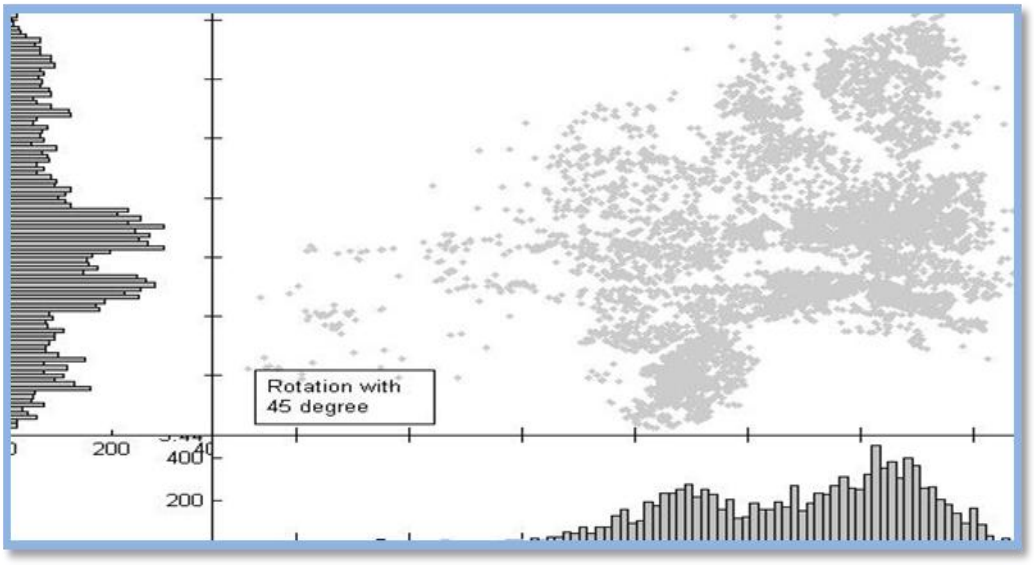
Population

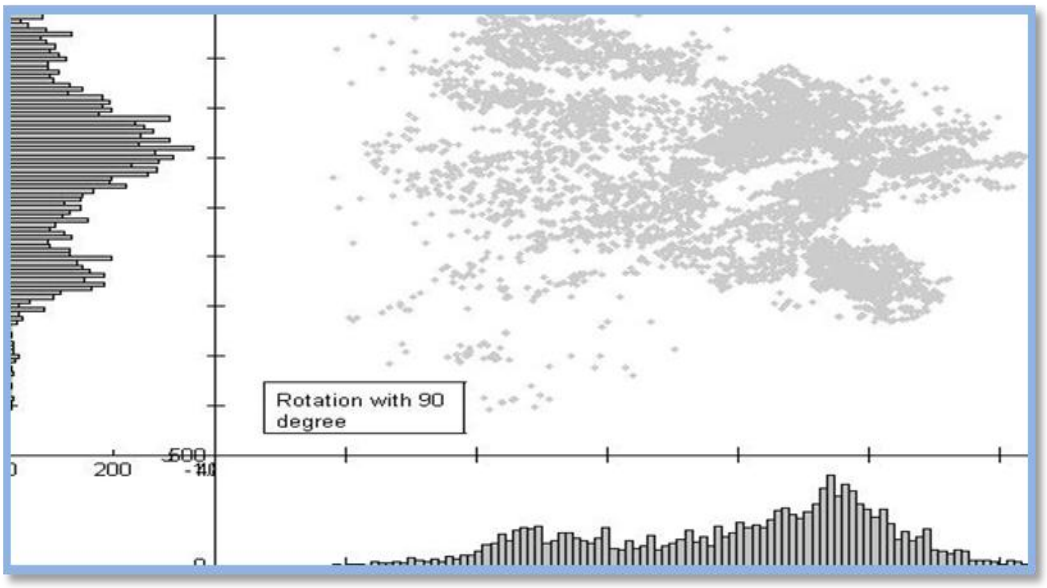
702	10032	5666	1141	0	0	0	0	0
33	5457	2965	5476	4422	56	0	0	0
1641	6478	3406	2591	1746	8531	0	0	0
1061	2895	3427	1944	10238	12003	11	0	0
1773	4510	2379	2044	14366	4773	1089	1903	2100
2273	4506	4699	3778	6174	12210	4775	3026	3598
1387	4550	1849	3705	4243	10027	78	0	773
1089	1997	1928	2226	3761	1505	0	0	46
935	3850	850	2489	13431	20223	3375	255	260
0	0	0	0	254	2213	852	408	814

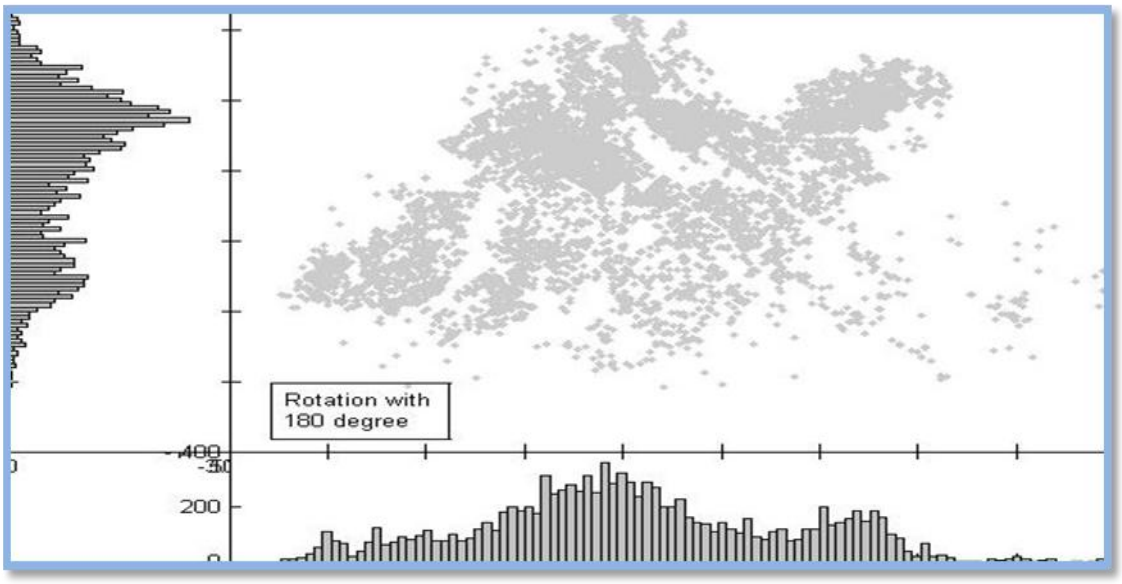
Crime rate

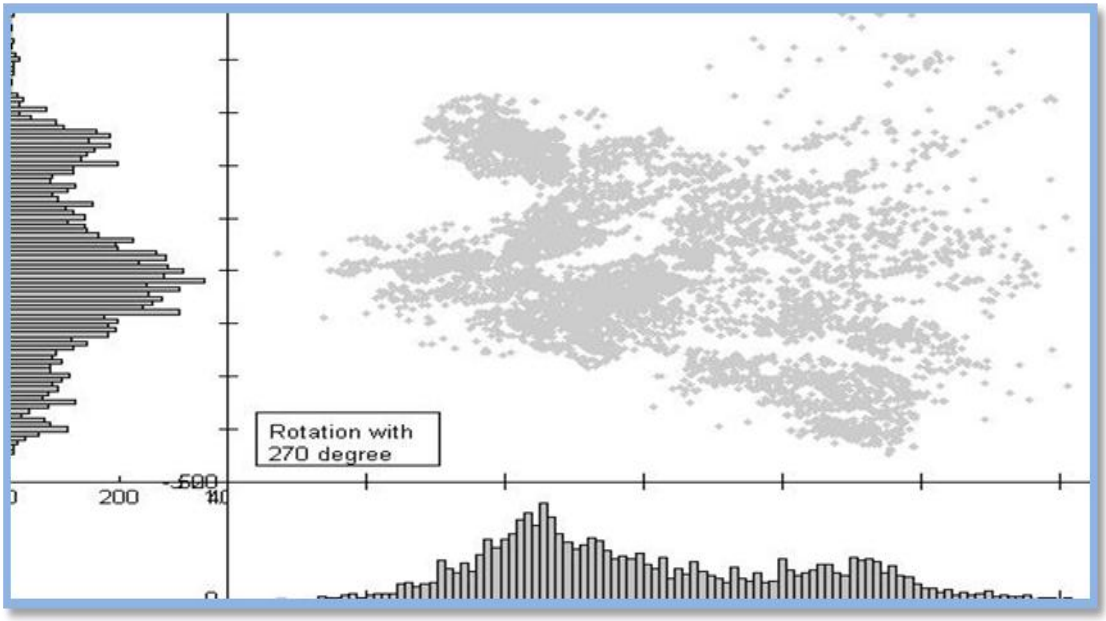
0.0199	0.0411	0.0385	0.0675	0	0	0	0	0
0.5152	0.0324	0.0381	0.0234	0.033	0.4107	0	0	0
0.0353	0.0401	0.0208	0.0178	0.0659	0.0367	0	0	0
0.0038	0.0169	0.0274	0.0159	0.0321	0.05	10.0909	0	0
0.0243	0.0146	0.0139	0.0205	0.0377	0.1982	0.1497	0.0625	0.0633
0.0176	0.0206	0.0145	0.0312	0.0436	0.0565	0.0427	0.0727	0.0414
0.0187	0.0105	0.0746	0.0202	0.041	0.0296	0.141	0	0.0013
0.0028	0.0436	0.0135	0.0332	0.0319	0.0439	0	0	0
0.0267	0.013	0.0341	0.0173	0.021	0.0623	0.0379	0.0078	0
0	0	0	0	0.0118	0.0465	0.0223	0	0

C2: Plots showing the distribution of crime datasets. Rotation data with(45, 90,180 and 270) degree utilized.

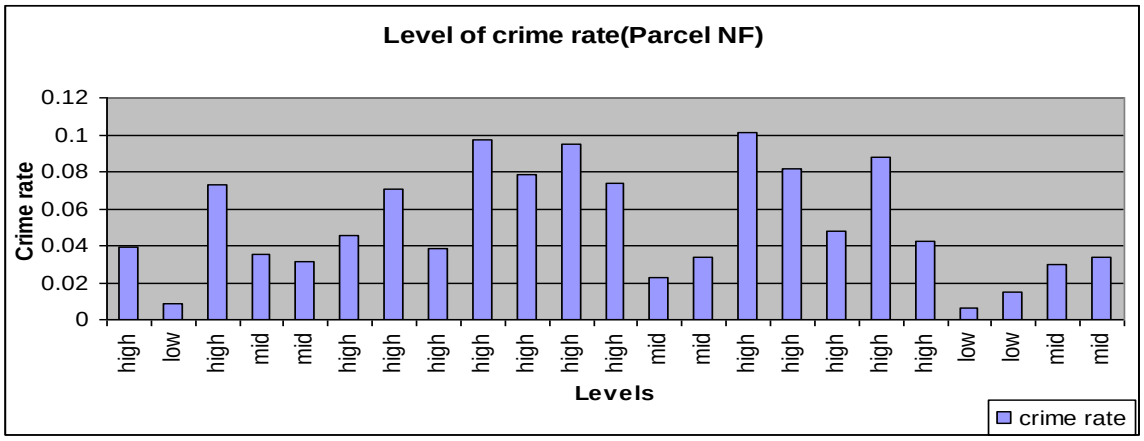


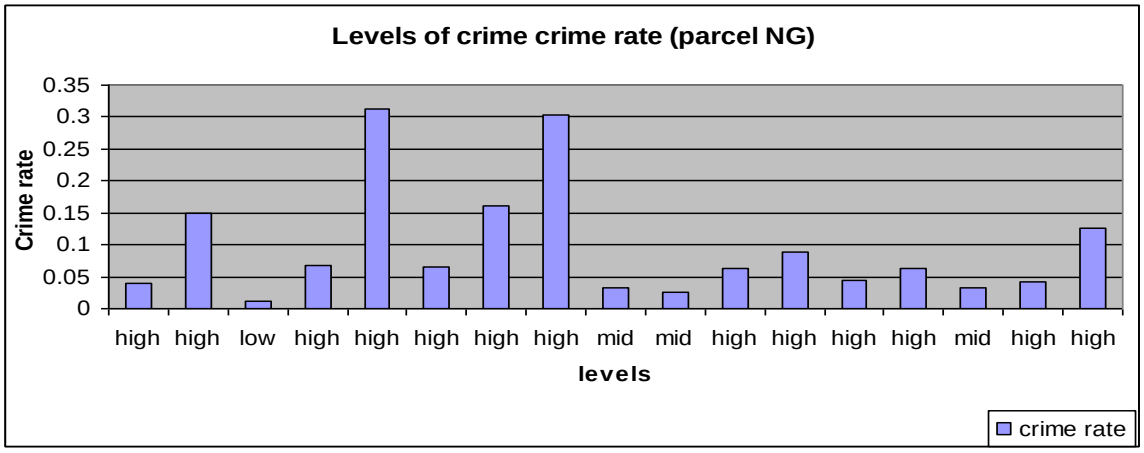


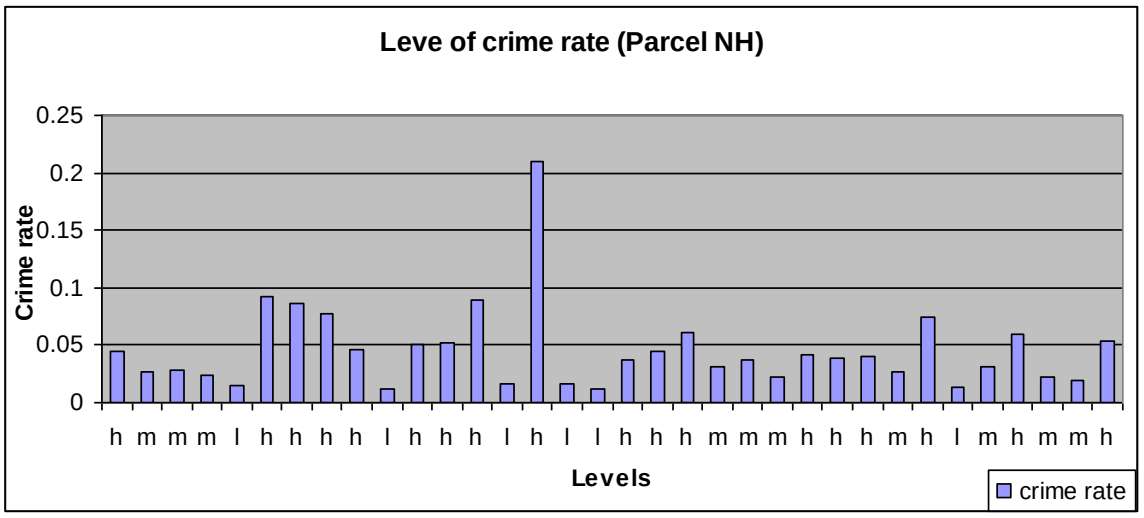

Rotation with
180 degree
200

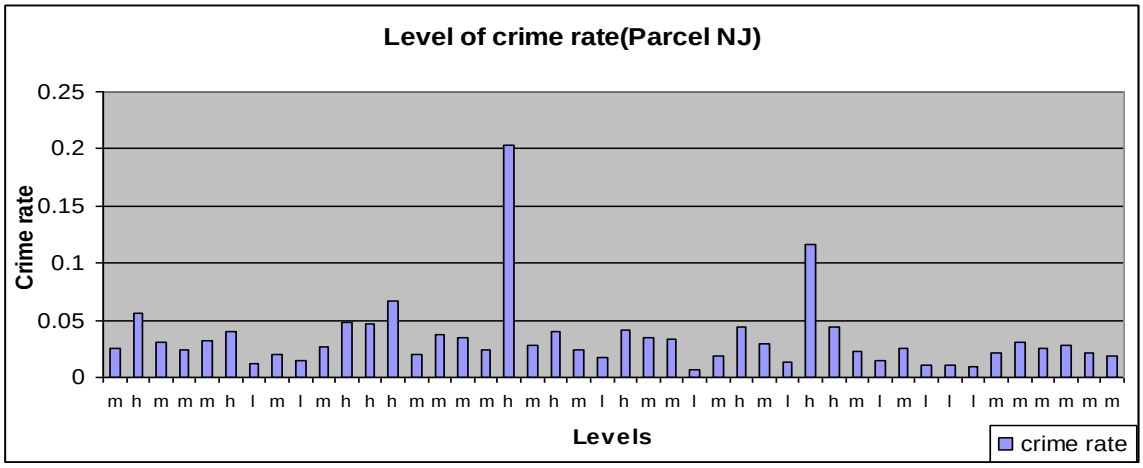

Rotation with
270 degree
200

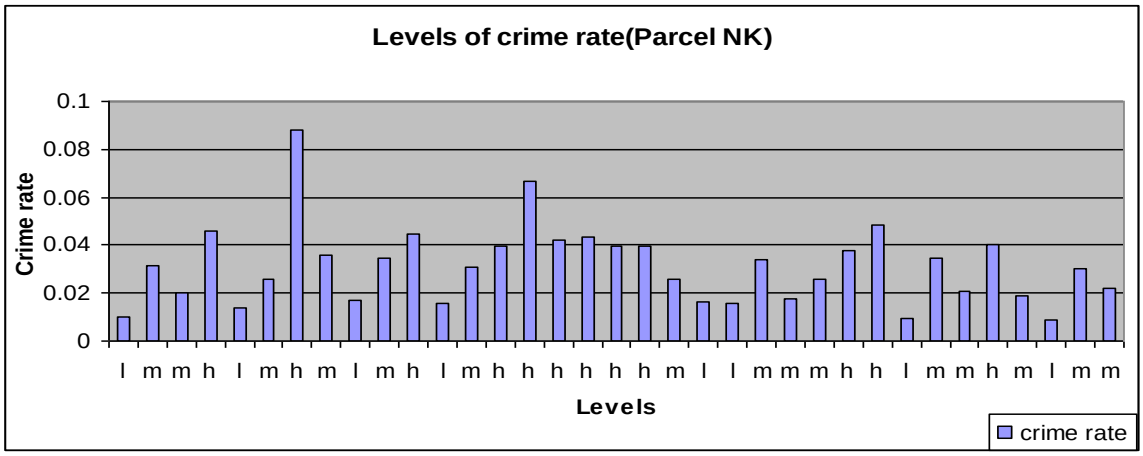
C3: Illustrate the levels of crime rate for selected parcels in the study region. The levels are high(h),middle(m) and low(l)

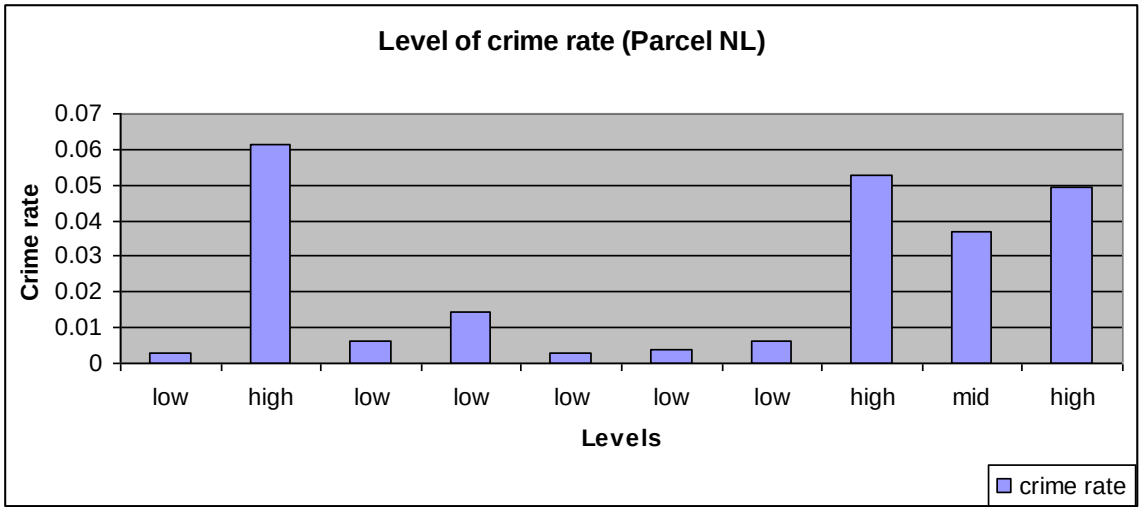


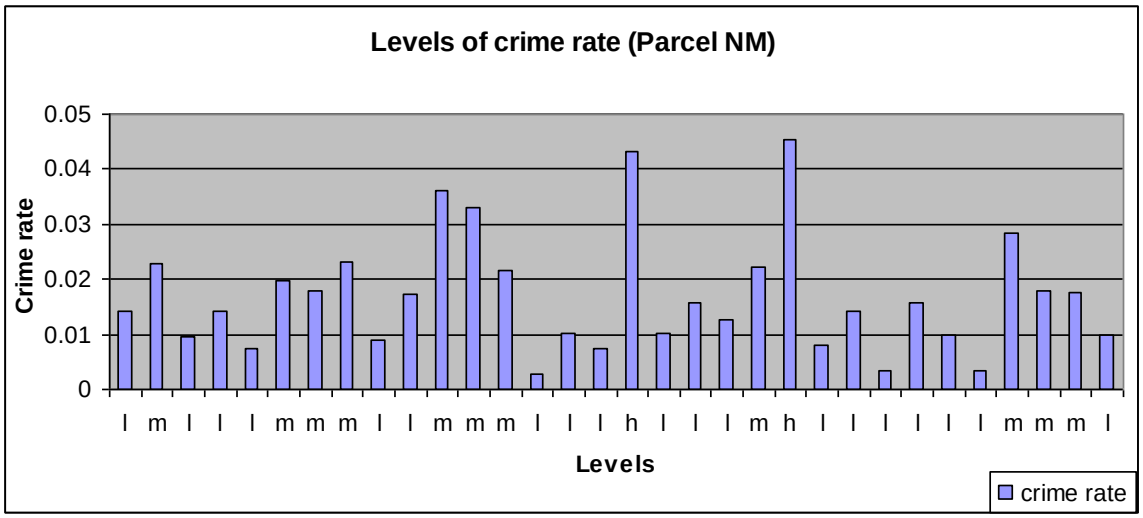


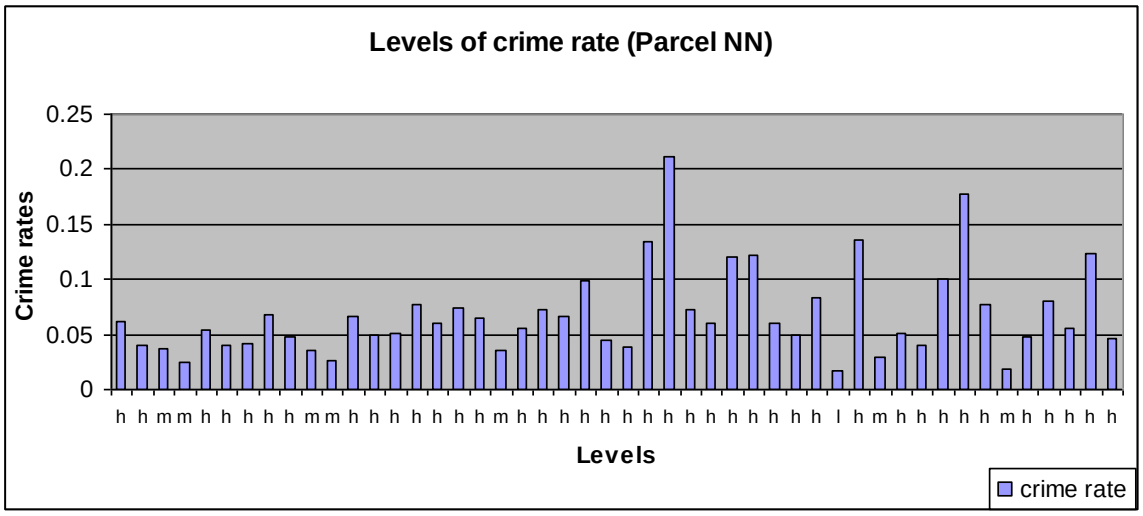


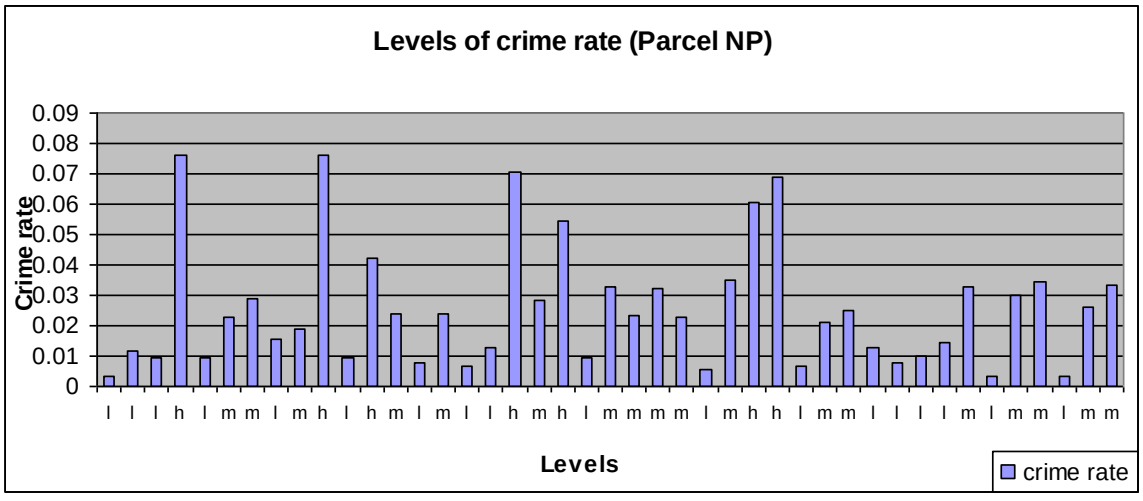

Level of crime rate(Parcel NJ)
Crime rate
0.25
0.2
0.15
0.1
0.05
0
m h m m m h l m l m h h h m m m m h m h m l h m m l m h m l h h m l m l l l m m m m m m
Levels
crime rate

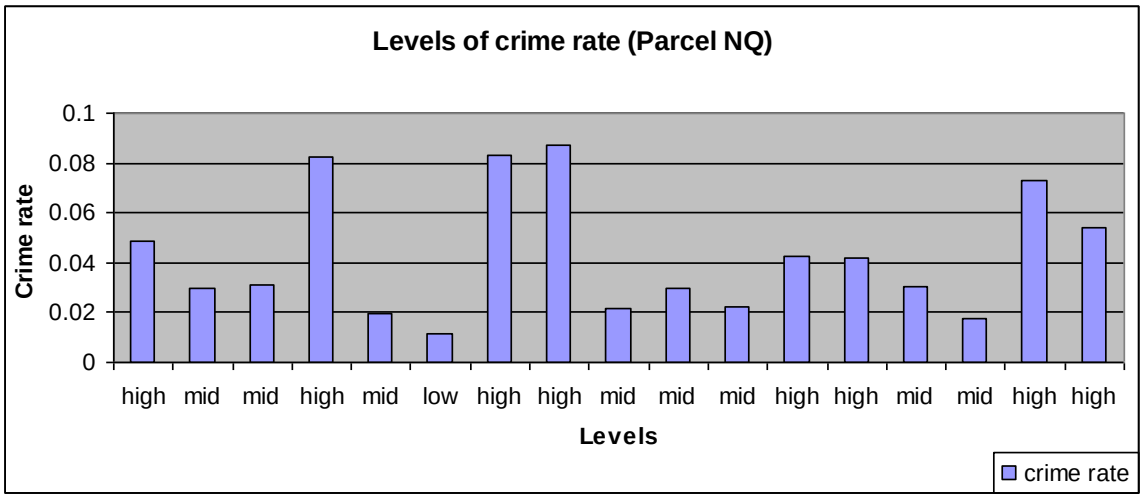

Levels of crime rate(Parcel NK)
Crime rate
0.1
0.08
0.06
0.04
0.02
0
l m m h l m h m l m h l m h h h h h h m l l m m m h h l m m h m l m m
Levels
crime rate

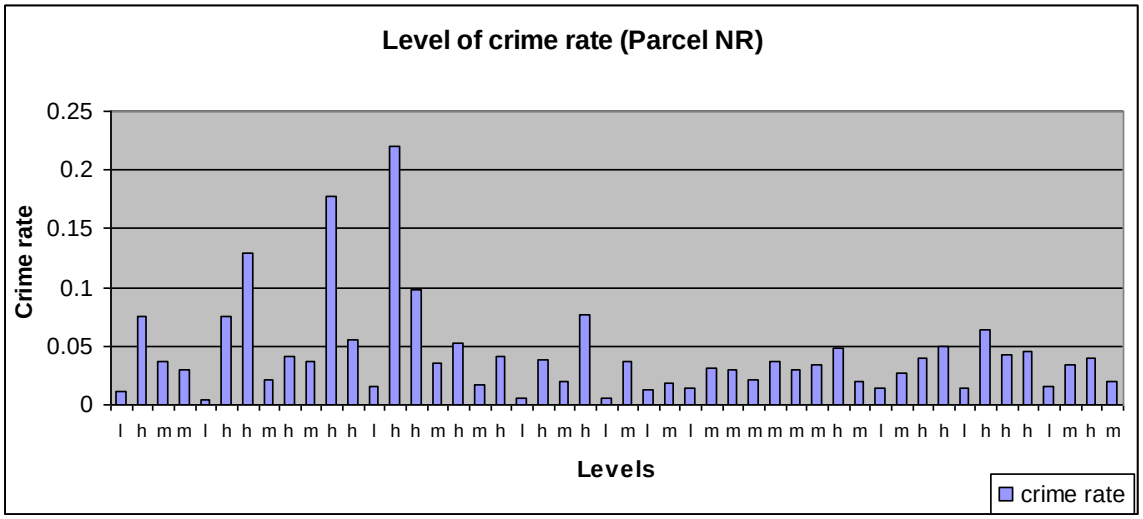

Level of crime rate (Parcel NL)
Crime rate
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
low high low low low low low high mid high
Levels
crime rate

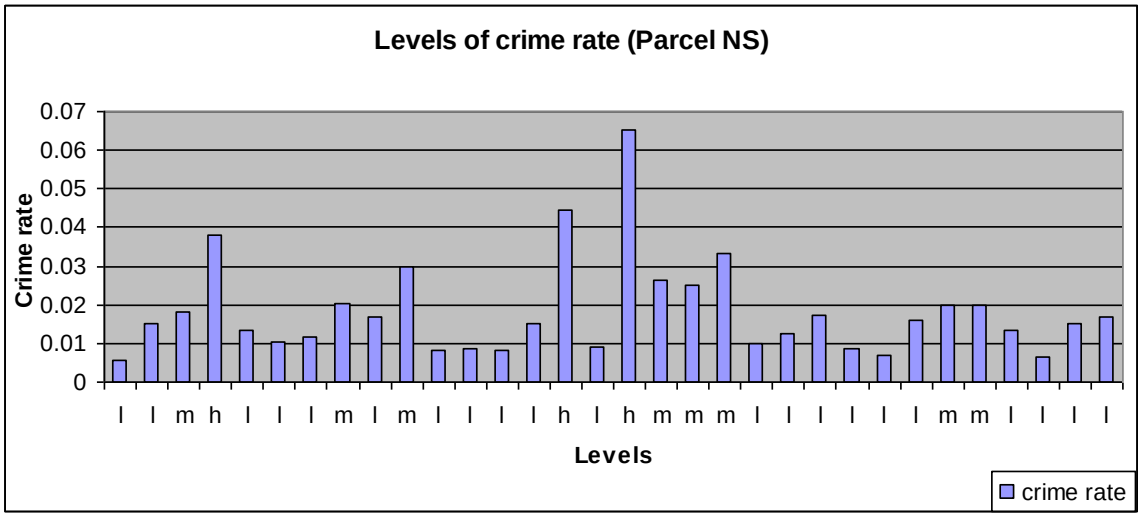

Levels of crime rate (Parcel NM)
Crime rate
0.05
0.04
0.03
0.02
0.01
0
l m l l l m m m l l m m m l l l h l l l m h l l l l l l m m m l
Levels
crime rate

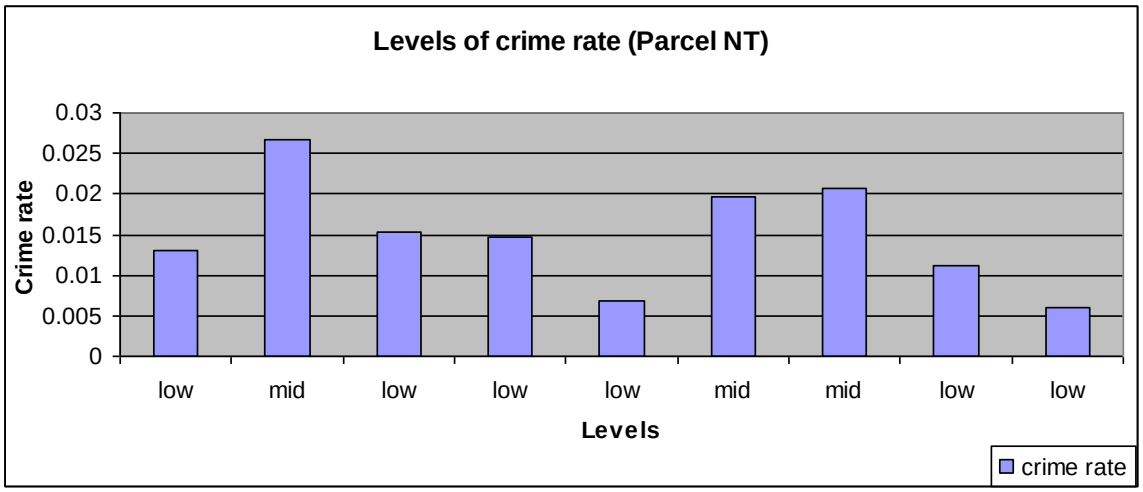

Levels of crime rate (Parcel NN)
Crime rates
0.25
0.2
0.15
0.1
0.05
0
h h m m h h h h h m m h h h h h h h m h h h h h h h h h h h h h h h h h h h h h l h m h h h h h m h h h h h
Levels
crime rate

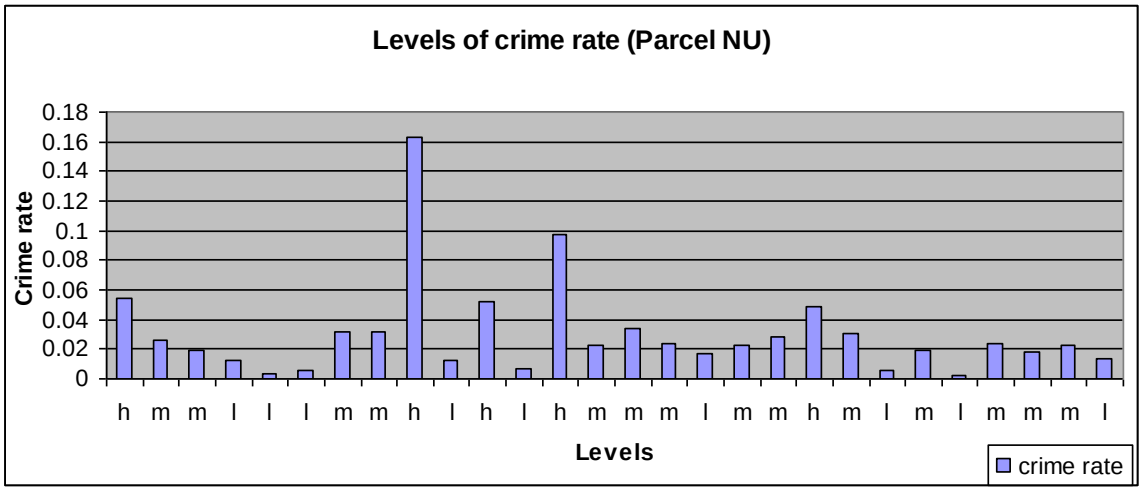

Levels of crime rate (Parcel NP)
Crime rate
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
l l l h l m m l m h l h m l m l l h m h l m m m m l m h h l m m l l l l m l m m l m m
Levels
crime rate

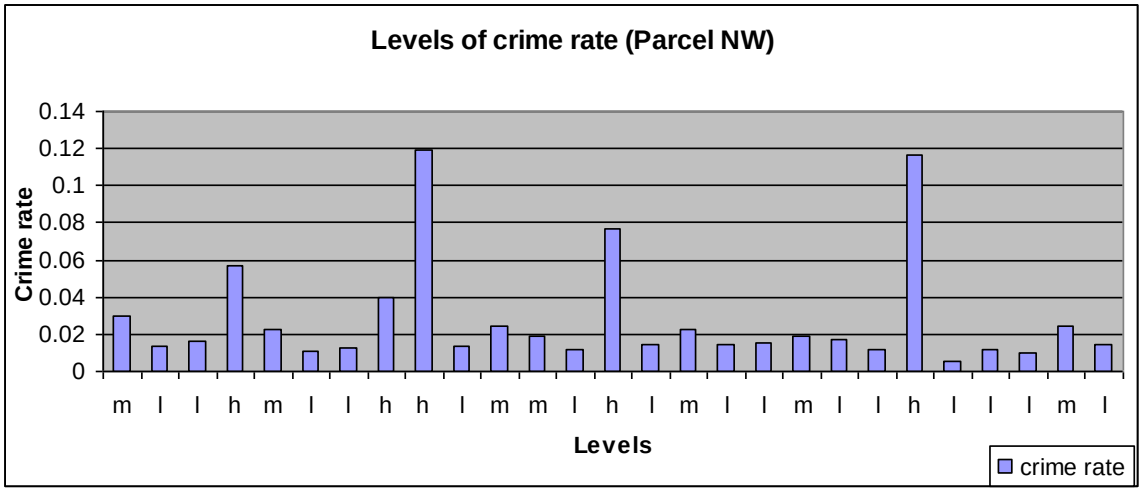

Levels of crime rate (Parcel NQ)
Crime rate
0.1
0.08
0.06
0.04
0.02
0
high mid mid high mid low high high mid mid mid high high mid mid high high
Levels
crime rate

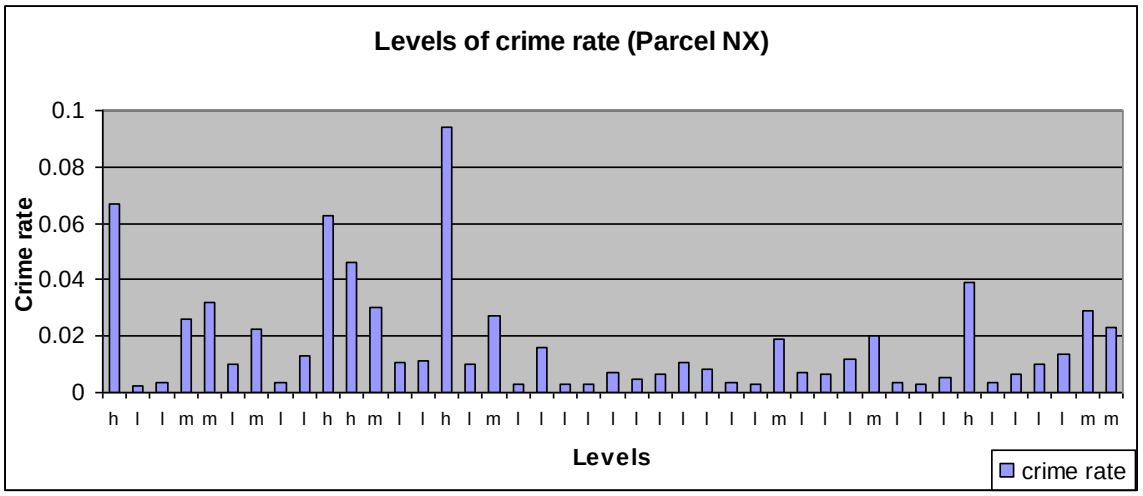

Level of crime rate (Parcel NR)
Crime rate
0.25
0.2
0.15
0.1
0.05
0
Levels
crime rate

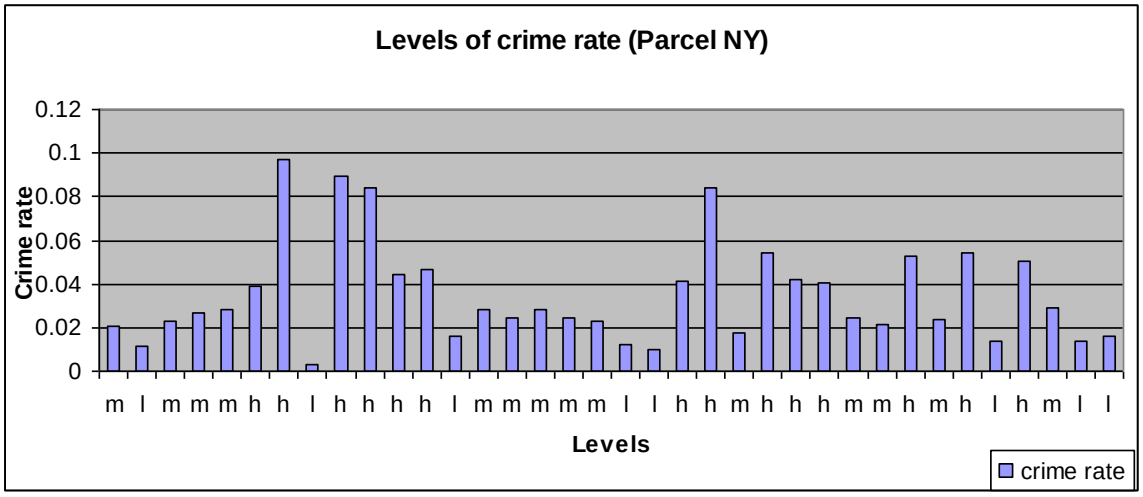

Levels of crime rate (Parcel NS)
Crime rate
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Levels
crime rate

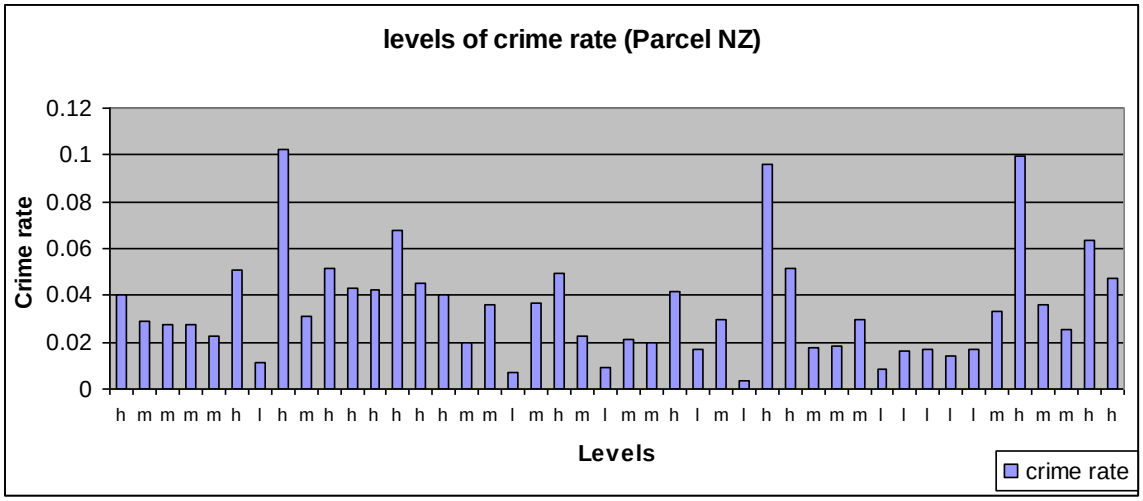

Levels of crime rate (Parcel NT)
Crime rate
0.03
0.025
0.02
0.015
0.01
0.005
0
low mid low low low mid mid low low
Levels
crime rate

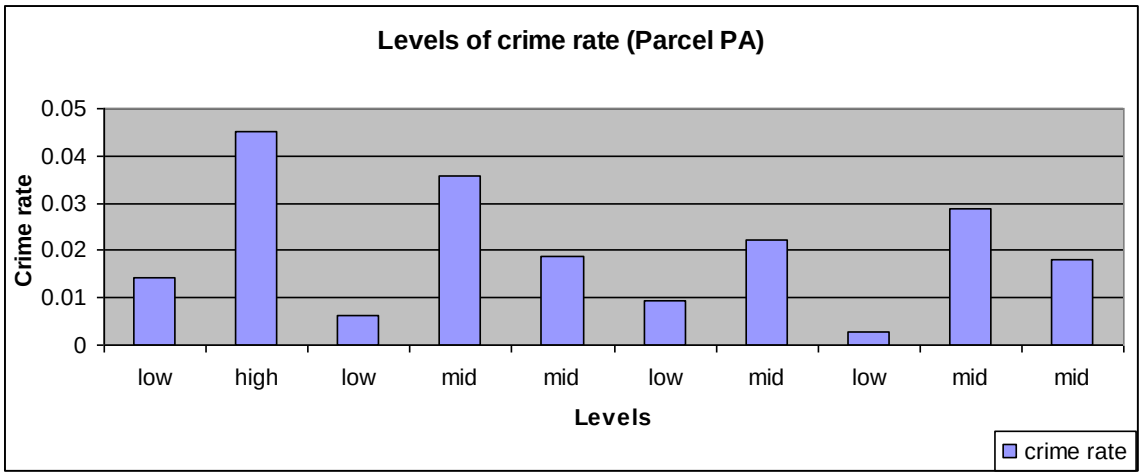

Levels of crime rate (Parcel NU)
Crime rate
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
h m m l l l m m h l h l h m m m l m m h m l m l m m m l
Levels
crime rate

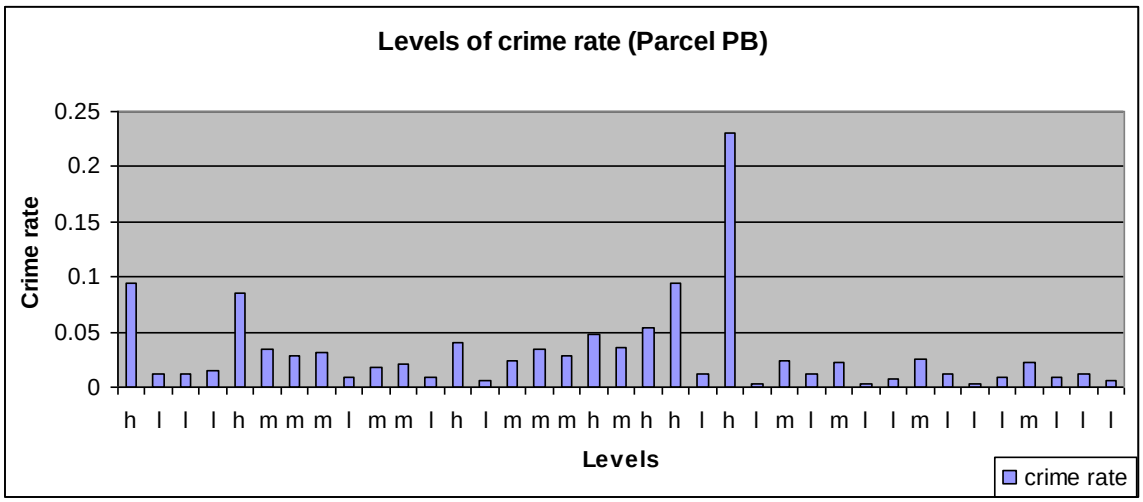

Levels of crime rate (Parcel NW)
Crime rate
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
m l l h m l l h h l m m l h l m l l m l l h l l l m l
Levels
crime rate

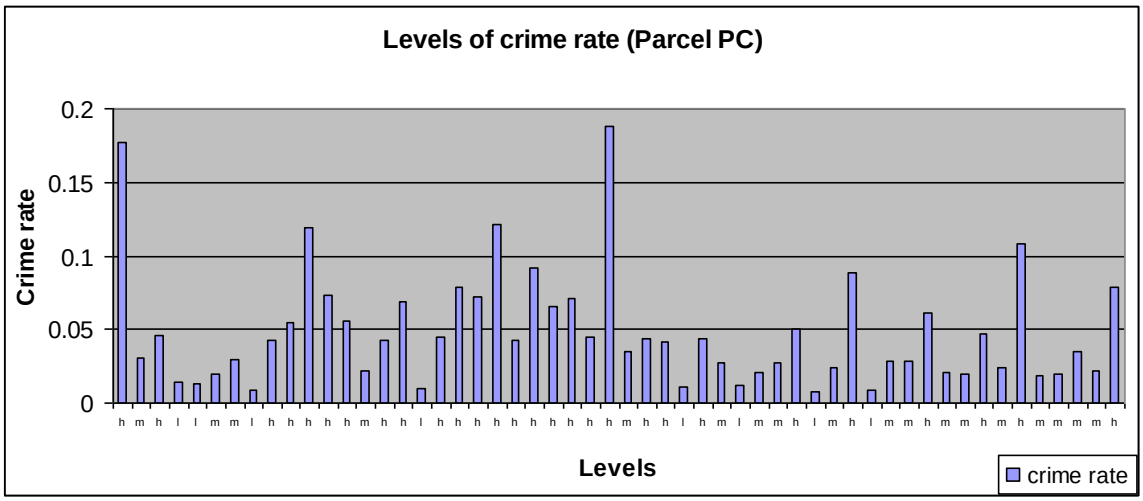

Levels of crime rate (Parcel NX)
Crime rate
0.1
0.08
0.06
0.04
0.02
0
h l l m m l m l l h h m l l h l m l l l l l l l l l l m l l l m l l l h l l l l m m
Levels
crime rate

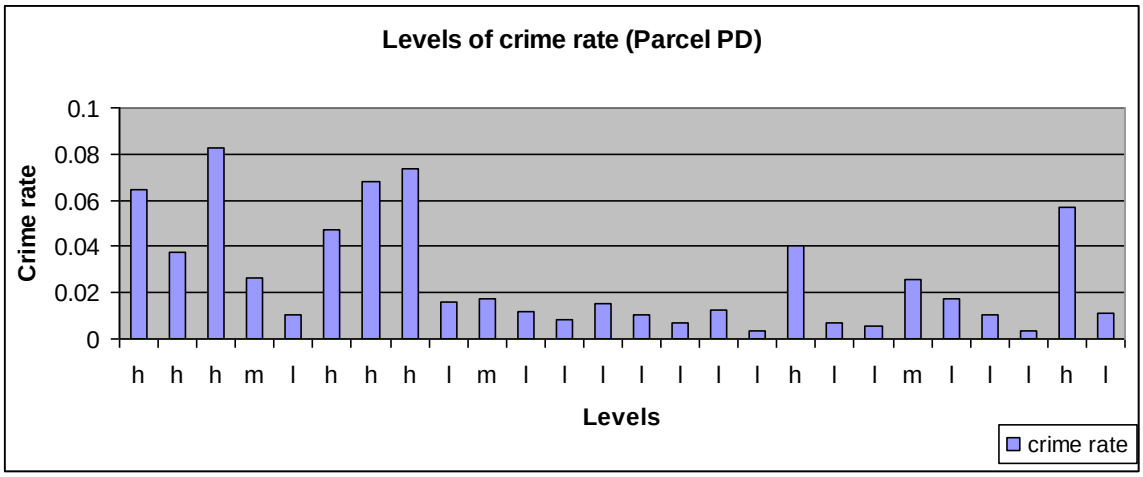

Levels of crime rate (Parcel NY)
Crime rate
0.12
0.1
0.08
0.06
0.04
0.02
0
m l m m m h h l h h h h l m m m m m l l h h m h h h m m h m h l h m l l
Levels
crime rate

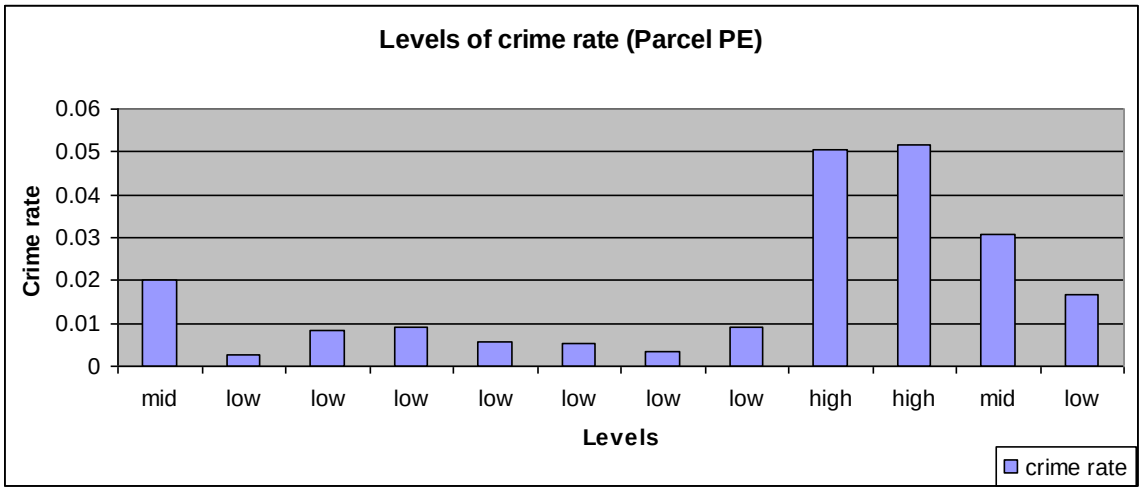

levels of crime rate (Parcel NZ)
Crime rate
0.12
0.1
0.08
0.06
0.04
0.02
0
h m m m m h l h m h h h h h h m m l m h m l m m h l m l h h m m m l l l l l m h m m h h
Levels
crime rate

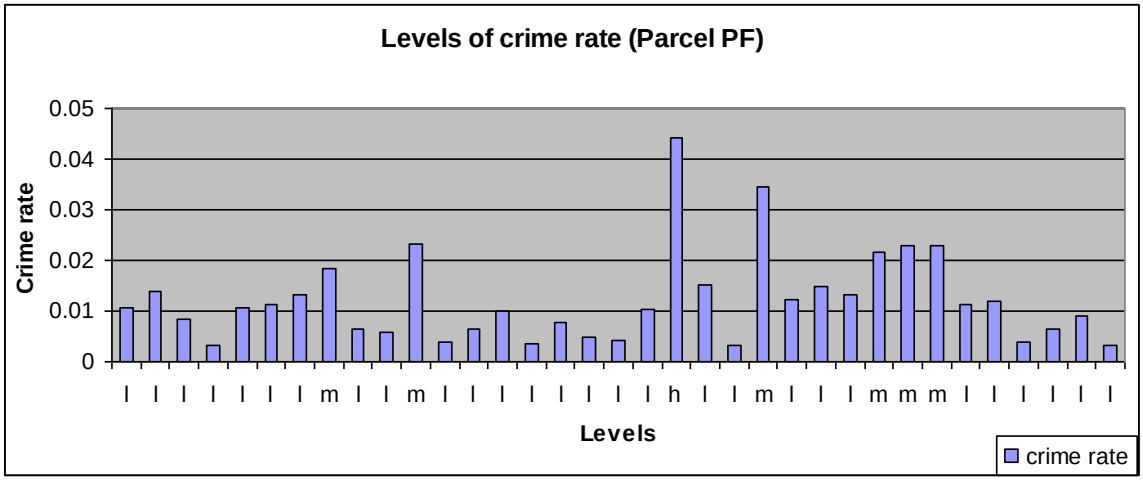

Levels of crime rate (Parcel PA)
Crime rate
0.05
0.04
0.03
0.02
0.01
0
low high low mid mid low mid low mid mid
Levels
crime rate

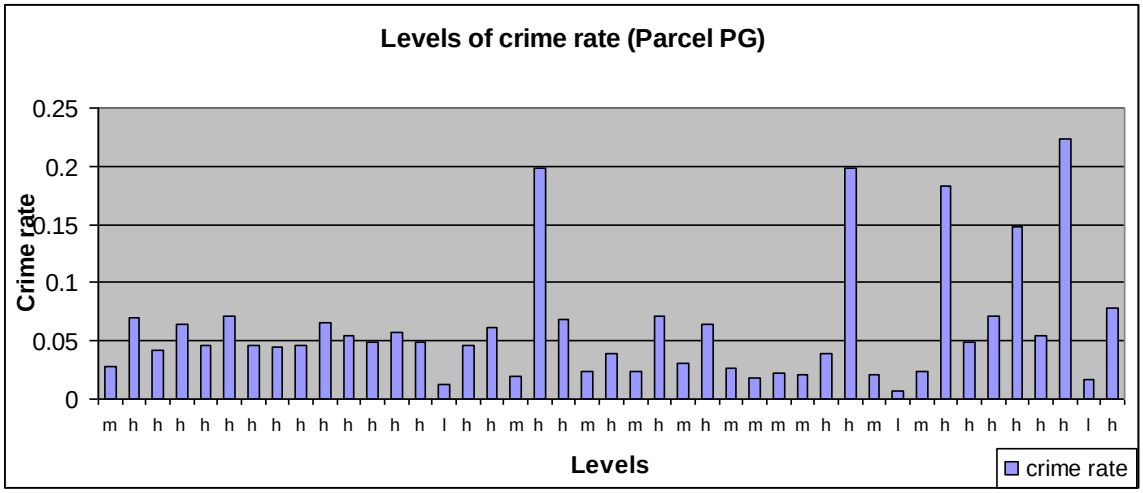

Levels of crime rate (Parcel PB)
Crime rate
0.25
0.2
0.15
0.1
0.05
0
Levels
crime rate

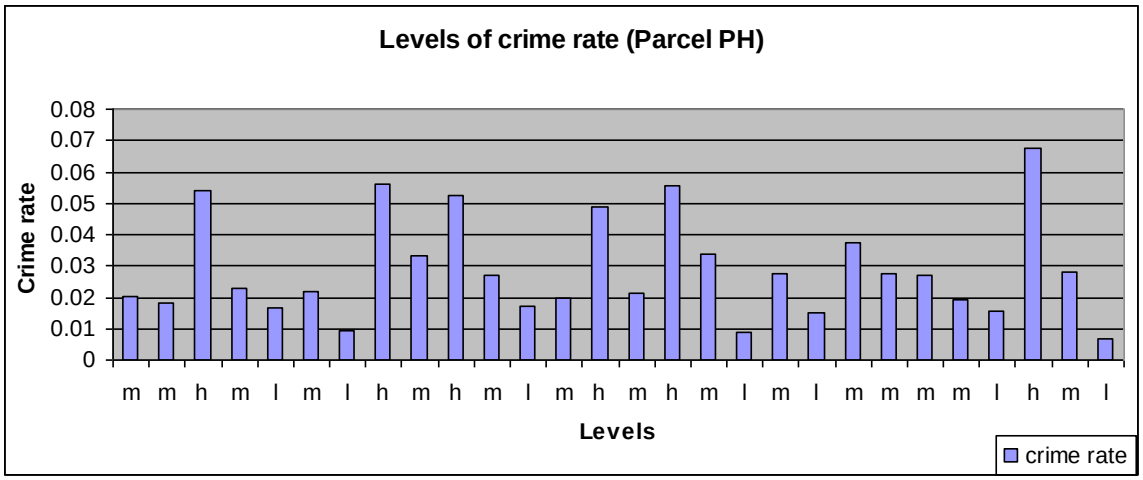

Levels of crime rate (Parcel PC)
Crime rate
0.2
0.15
0.1
0.05
0
Levels
crime rate

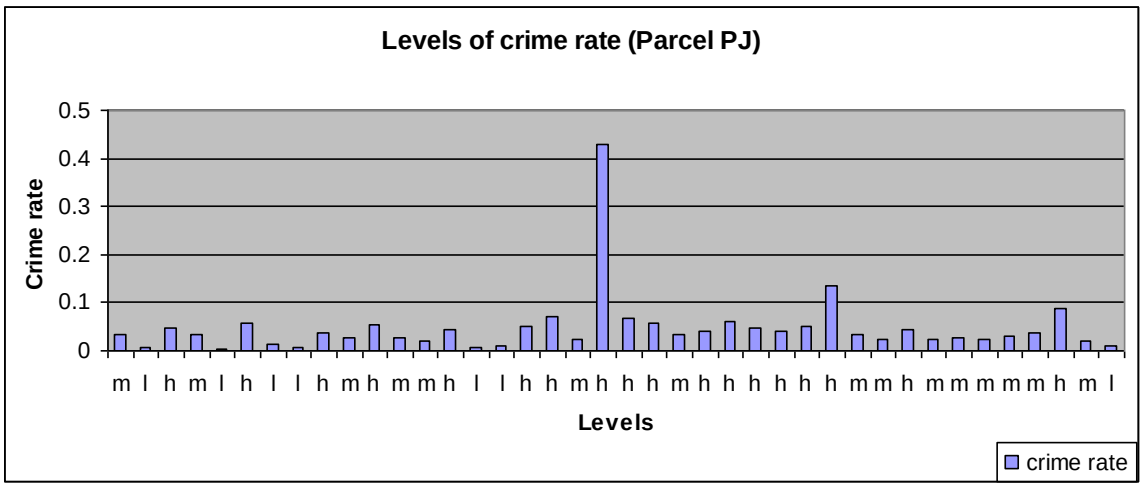

Levels of crime rate (Parcel PD)
Crime rate
0.1
0.08
0.06
0.04
0.02
0
h h h m l h h h l m l l l l l l l h l l m l l l h l
Levels
crime rate

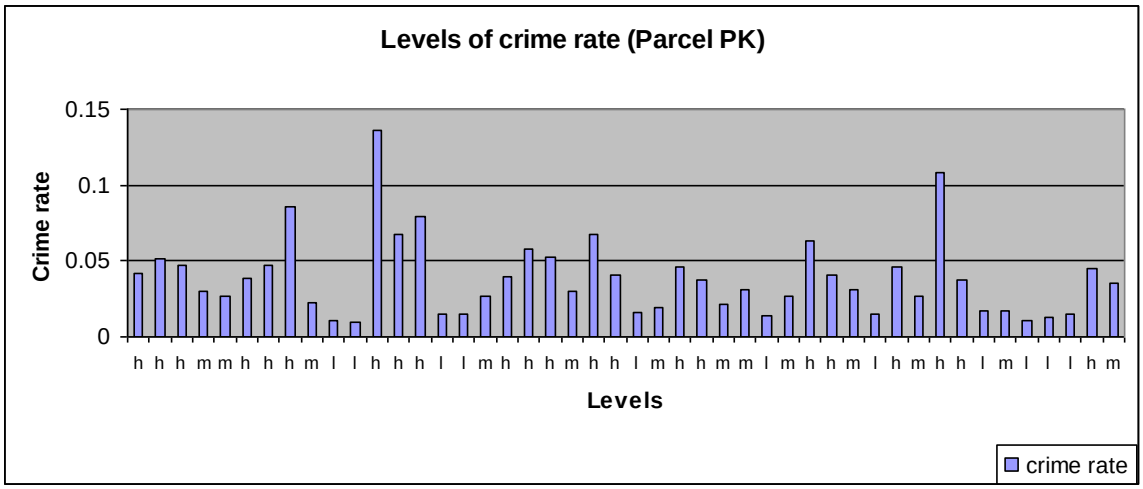

Levels of crime rate (Parcel PE)
Crime rate
0.06
0.05
0.04
0.03
0.02
0.01
0
mid low low low low low low low high high mid low
Levels
crime rate

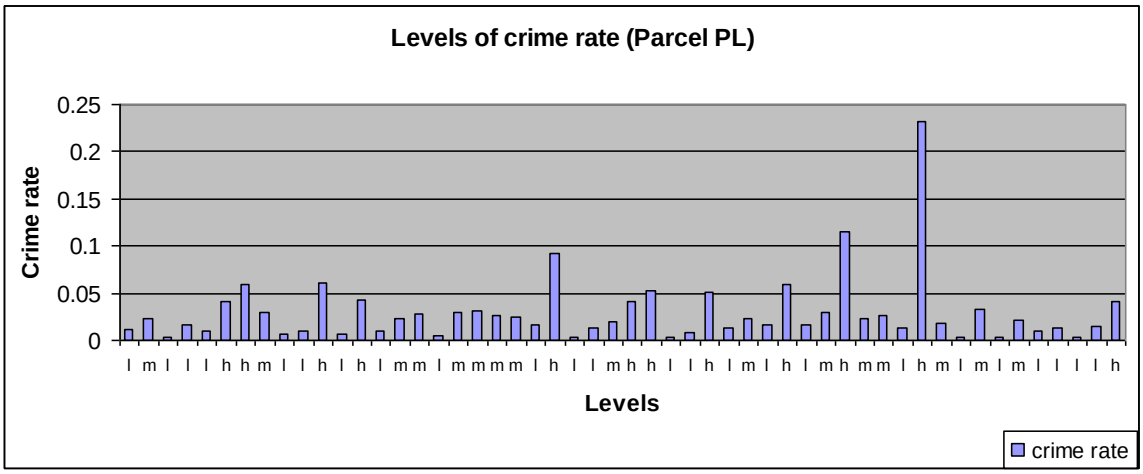

Levels of crime rate (Parcel PF)
Crime rate
0.05
0.04
0.03
0.02
0.01
0
l l l l l l l m l l m l l l l l l l l h l l m l l l m m m l l l l l
Levels
crime rate


Levels of crime rate (Parcel PG)
Crime rate
0.25
0.2
0.15
0.1
0.05
0
m h h h h h h h h h h h h h l h h m h h m h m h m h m h h m m m m h h m h h h h h l h
Levels
crime rate


Levels of crime rate (Parcel PH)
Crime rate
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
m m h m l m l h m h m l m h m h m l m l m m m m l h m l
Levels
crime rate


Levels of crime rate (Parcel PJ)
Crime rate
0.5
0.4
0.3
0.2
0.1
0
m l h m l h l l h m h m m h l l h h m h h h m h h h h h h h m m h m m m m m h m l
Levels
crime rate


Levels of crime rate (Parcel PK)
Crime rate
0.15
0.1
0.05
0
h h h m m h h h m l l h h h l l m h h h m h h l m h h m m l m h h m h h l m l l l h m
Levels
crime rate


Levels of crime rate (Parcel PL)
Crime rate
0.25
0.2
0.15
0.1
0.05
0
l m l l l h h m l l h l h l m m l m m m m l h l l m h h l l h l m l h l m h m m l h m l m l m l l l l h
Levels
crime rate

C4: HowTo Count the number of Point features within a Polygon

Article ID: 30779

Software: **ArcGIS - ArcEditor** 8.3, 9.0, 9.1, 9.2, 9.3, 9.3.1 **ArcGIS - ArcInfo** 8.3, 9.0, 9.1, 9.2, 9.3, 9.3.1 **ArcGIS - ArcView** 8.3, 9.0, 9.1, 9.2, 9.3, 9.3.1

Platforms: **Windows** 2000, XP, 2003Server

Summary
Instructions provided describe how to count how many point features fall within each polygon feature.

Procedure
Create a count field and a spatial join between the point shapefile and the polygon shapefile.

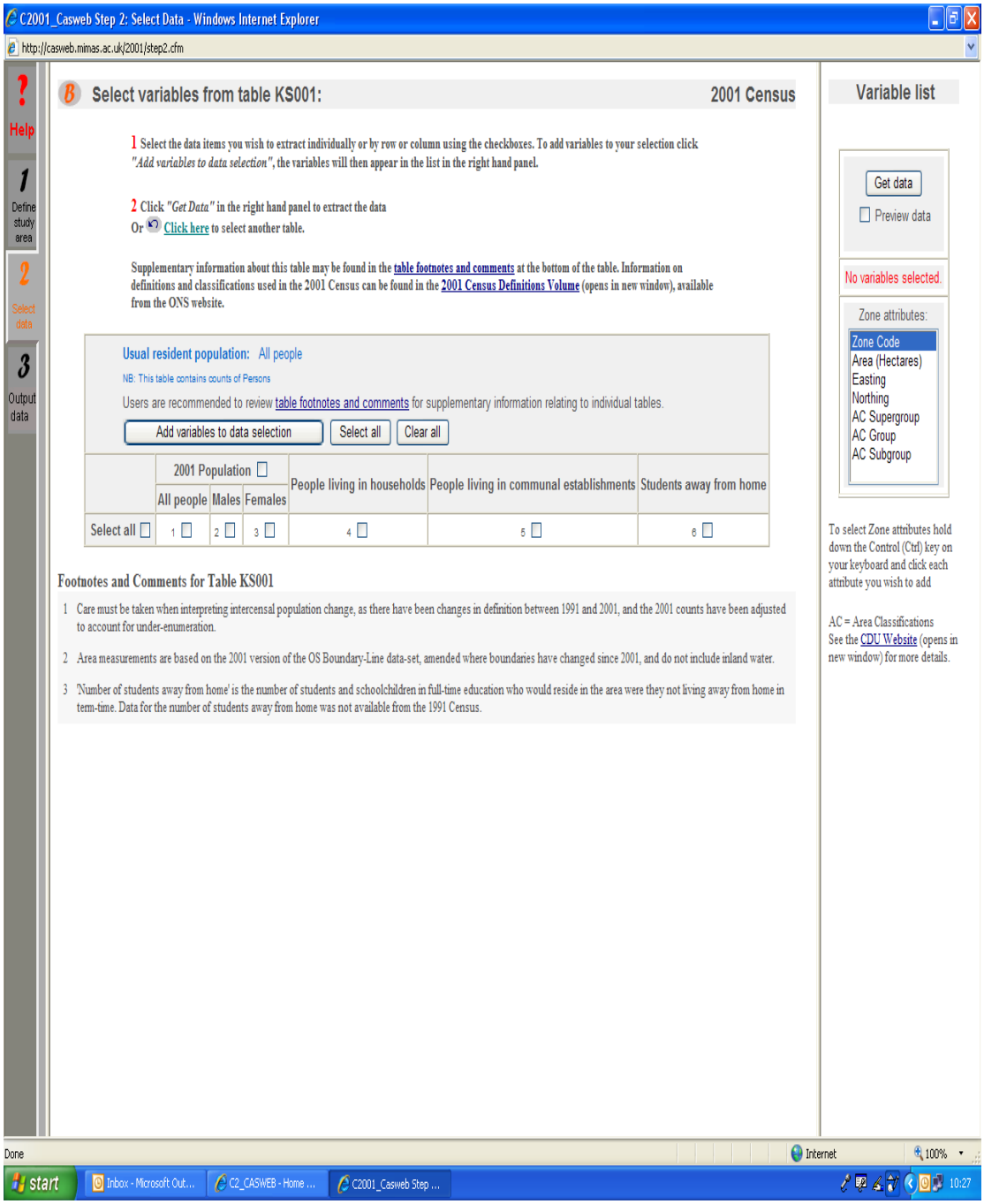
1. Create a field called 'Count' in the attribute table of the point shapefile of type 'Short Integer'.
2. Calculate the Count field equal to 1 by right-clicking on the field name > calculate values. Enter a 1 in the white dialog area and then click OK.
3. Right-click on the polygon shapefile > Joins and Relates > Joins. Click the dropdown list and select 'Join data from another layer based on spatial location'.
4. Specify the point shapefile for Step 1.
5. Select the first bullet, each polygon is given a summary of the numeric attributes, for Step 2, and check the 'Sum' box.
6. At Step 3, specify an output location and then click OK.
7. A polygon shapefile with the 'Count' field indicating how many point features lie within each polygon feature is now present.

Created: 5/8/2006
Last Modified: 12/15/2009

http://support.esri.com/index.cfm?fa=knowledgebase.techarticles.articleShow&d=30779

Appendix D

The detail information about the potential characteristic of burglary household was downloaded from the CASWEB Website.



Usual resident population: All people

NB: This table contains counts of Persons

Users are recommended to review table footnotes and comments for supplementary information relating to individual tables.

Add variables to data selection | Clear all

	2001 Population			People living in households	People living in communal establishments	Students away from home
	All people	Males	Females			
Select all	1	2	3	4	5	6

Footnotes and Comments for Table KS001

Age structure: All people

NB: This table contains counts of Persons

Users are recommended to review table footnotes and comments for supplementary information relating to individual tables.

Add variables to data selection | Clear all

	All people	People aged															
		0-4	5-7	8-9	10-14	15	16-17	18-19	20-24	25-29	30-44	45-59	60-64	65-74	75-84	85-89	90 & over
Select all	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Footnotes and Comments for Table KS002

1 'Age' is age last birthday.

Occupation groups - all persons: All people aged 16-74 in employment

NB: This table contains counts of Persons

Add variables to data selection | Clear all

	All people aged 16-74 in employment	people aged 16-74 in employment working as								
		Managers and senior officials	Professional occupations	Associate professional and technical occupations	Administrative and secretarial occupations	Skilled trades occupations	Personal service occupations	Sales and customer service occupations	Process, plant and machine operatives	Elementary occupations
Select all	1	2	3	4	5	6	7	8	9	10

K12

National Statistics - Socio Economic Classification - all persons:

All people aged 16-74

NB: This table contains counts of Persons

Users are recommended to review table footnotes and comments for supplementary information relating to individual tables.

Add variables to data selection | Clear all

	All people aged 16-74	People aged 16-74											
		Large employers and higher managerial occupations	Higher professional occupations	Lower managerial and professional occupations	Intermediate occupations	Small employers and own account workers	Lower supervisory and technical occupations	Semi-routine occupations	Routine occupations	Never worked	Long-term unemployed	Full-time students	Not classifiable for other reasons
Select all	1	2	3	4	5	6	7	8	9	10	11	12	13

Footnotes and Comments for Table KS014a

Household spaces and accommodation type: All household spaces

NB: This table contains counts of Household spaces

Users are recommended to review table footnotes and comments for supplementary information relating to individual tables.

Add variables to data selection | Clear all

	All household spaces			All household spaces which are of accommodation type						
		With no residents		Whole house or bungalow			Flat, maisonette or apartment			
	With residents	Vacant	Second residence/holiday accommodation	Detached	Semi-detached	Terraced (including end-terrace)	Purpose built block of flats or tenement	Part of a converted or shared house (including bed-sits)	In a commercial building	Caravan or other mobile or temporary structure
Select all	1	2	3	4	5	6	7	8	9	10

Footnotes and Comments for Table KS016

Household composition: All households

NB: This table contains counts of Households

Users are recommended to review table footnotes and comments for supplementary information relating to individual tables.

Add variables to data selection | Clear all

	All households	Households comprising														
		One person	One family and no others										Other households			
					Married couple households			Cohabiting couple households			Lone parent households					
		Pensioner	Other	All pensioners	No children	With dependent children	All children non-dependent	No children	With dependent children	All children non-dependent	With dependent children	All children non-dependent	With dependent children	All student	All pensioner	Other
Select all	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Footnotes and Comments for Table KS020

1 A dependent child is a person in a household aged 0 to 15 (whether or not in a family) or a person aged 16 to 18 who is a full-time student in a family with parent(s).

Qualifications and students: All people aged 16-74

NB: This table contains counts of Persons

Users are recommended to review table footnotes and comments for supplementary information relating to individual tables.

Add variables to data selection | Clear all

	All people aged 16 - 74	People aged 16-74 with						Full-time students and schoolchildren		Full-time students aged 18-74		
		No qualifications	Highest qualification attained				Other qualifications/level unknown	aged 16-17	aged 18-74	Economically active		Economically inactive
			level 1	level 2	level 3	level 4/5				In employment	Unemployed	
Select all	1	2	3	4	5	6	7	8	9	10	11	12

Footnotes and Comments for Table KS013

1 1+ 'O' level passes, 1+ CSE/GCSE any grades, NVQ level 1, Foundation GNVQ.

2 5+ 'O' level passes, 5+ CSEs (grade 1), 5+ GCSEs (grades A-C), School Certificate, 1+ 'A' levels/'AS' levels, NVQ level 2, Intermediate GNVQ.

3 2+ 'A' levels, 4+ 'AS' levels, Higher School Certificate, NVQ level 3, Advanced GNVQ.

4 First degree, Higher degree, NVQ levels 4 and 5, HNC, HND, Qualified Teacher Status, Qualified Medical Doctor, Qualified Dentist, Qualified Nurse, Midwife, Heath Visitor.

Appendix E

E1: This appendix presented the obtained results from MLR models for selected parcels in the study region.

NF

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.06019	0.02705	-2.23	0.068	
ManagersOcc	1.3586	0.2402	5.66	0.001	1.662
Hocc	0.6802	0.2097	3.24	0.018	3.473
allpensioner	-3.4721	0.6867	-5.06	0.002	3.647
Loneparent	0.4567	0.1667	2.74	0.034	2.108

S = 0.0130456 R-Sq = 93.9% R-Sq(adj) = 82.7%

NH

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.05231	0.02124	-2.46	0.034	
professionalOcc	2.0868	0.2938	7.10	0.000	3.398
AdministrativeOcc	-0.9057	0.1869	-4.85	0.001	4.146
sales&customerservice	0.4508	0.1964	2.30	0.045	2.043
machineoperatives	0.9385	0.2005	4.68	0.001	3.093
ElementaryOcc	-0.4210	0.1425	-2.95	0.014	1.762
h3	0.14670	0.05839	2.51	0.031	2.312
flat2	-1.3962	0.3539	-3.94	0.003	2.205
one person	0.21606	0.05237	4.13	0.002	2.705

S = 0.00870974 R-Sq = 93.8% R-Sq(adj) = 85.1%

NJ

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.04545	0.01674	-2.71	0.017	
SkilledtradesOcc	0.5850	0.1143	5.12	0.000	2.775
personalserviceOcc	-0.4492	0.1327	-3.38	0.004	2.806
sales&customerservice	0.3012	0.1150	2.62	0.020	2.894
machineoperatives	0.7618	0.1844	4.13	0.001	2.767
Unemploy	0.40883	0.08852	4.62	0.000	1.954
flat1	0.19175	0.04591	4.18	0.001	2.383
cohabiting	0.28177	0.09469	2.98	0.010	3.320

S = 0.00492369 R-Sq = 91.6% R-Sq(adj) = 82.5%

NK

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.03137	0.02055	-1.53	0.146	
<16	-0.3090	0.1004	-3.08	0.007	3.583
sales&customerservice	0.4638	0.1703	2.72	0.015	2.537
h2	0.5641	0.2618	2.15	0.047	2.339
one person	0.23781	0.09132	2.60	0.019	3.928

S = 0.0104467 R-Sq = 80.7% R-Sq(adj) = 61.3%

NR

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.06354	0.02043	3.11	0.006	
ManagersOcc	0.24898	0.08991	2.77	0.013	1.892
SkilledtradesOcc	-0.3929	0.1023	-3.84	0.001	1.763
personalserviceOcc	0.4382	0.1993	2.20	0.041	1.943
machineoperatives	-0.4287	0.1469	-2.92	0.009	1.684
cohabiting	-0.6538	0.1152	-5.68	0.000	3.139

S = 0.00781864 R-Sq = 81.6% R-Sq(adj) = 67.3%

PC

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.08408	0.01948	4.32	0.005	
ManagersOcc	-0.59581	0.08478	-7.03	0.000	3.102
personalserviceOcc	1.6611	0.1891	8.78	0.000	3.403
sales&customerservice	-0.5108	0.1015	-5.03	0.002	3.818
machineoperatives	-0.4944	0.1236	-4.00	0.007	2.727
Unemploy	-0.7168	0.1008	-7.11	0.000	4.100
h1	0.5869	0.2311	2.54	0.044	1.995
h2	0.95748	0.07929	12.08	0.000	4.454
allpensioner	-0.5697	0.1150	-4.95	0.003	2.230

S = 0.00438474 R-Sq = 98.2% R-Sq(adj) = 97.4%

PG

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.02287	0.02351	-0.97	0.344	
>60	-0.17973	0.04674	-3.85	0.001	2.577
AdministrativeOcc	-0.3560	0.1395	-2.55	0.021	1.717
SkilledtradesOcc	0.4669	0.2165	2.16	0.046	2.421
personalserviceOcc	0.9107	0.1930	4.72	0.000	1.595
Meocc	0.3930	0.1401	2.80	0.012	1.994
h1	1.2774	0.3701	3.45	0.003	1.896
one person	0.24932	0.02565	9.72	0.000	2.164

S = 0.0115639 R-Sq = 93.7% R-Sq(adj) = 87.8%

PJ

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.01797	0.01644	1.09	0.298	
TechnicalOcc	-0.3665	0.1068	-3.43	0.006	2.446
h1	0.7381	0.1924	3.84	0.003	3.326
flat2	0.3549	0.1509	2.35	0.038	1.968
cohabiting	-0.8383	0.2084	-4.02	0.002	2.525

S = 0.00997802 R-Sq = 87.3% R-Sq(adj) = 68.8%

NG

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.00848	0.03292	0.26	0.803	
SkilledtradesOcc	-2.9169	0.7903	-3.69	0.006	1.956
machineoperatives	-2.3470	0.9702	-2.42	0.042	1.197
ElementaryOcc	2.2141	0.5281	4.19	0.003	1.813
meocc	1.2804	0.3711	3.45	0.009	1.562

S = 0.0283116 R-Sq = 72.6% R-Sq(adj) = 58.9%

PK

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.03137	0.02055	-1.53	0.146	
<16	-0.3090	0.1004	-3.08	0.007	3.583
sales&customerservice	0.4638	0.1703	2.72	0.015	2.537
h2	0.5641	0.2618	2.15	0.047	2.339
one person	0.23781	0.09132	2.60	0.019	3.928

S = 0.0104467 R-Sq = 80.7% R-Sq(adj) = 61.3%

NZ

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.02641	0.04108	-0.64	0.527	
16-59	0.22154	0.07842	2.83	0.010	3.218
personalserviceOcc	0.5284	0.1664	3.18	0.004	1.984
h2	0.07825	0.03548	2.21	0.038	2.354
allpensioner	-0.4182	0.1989	-2.10	0.047	1.855
Married	-0.3005	0.1010	-2.98	0.007	4.080
cohabiting	-0.7717	0.1609	-4.80	0.000	1.520

S = 0.00972125 R-Sq = 77.2% R-Sq(adj) = 61.6%

NY

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.14523	0.03309	4.39	0.001	
SkilledtradesOcc	-0.4990	0.1728	-2.89	0.012	1.554
sales&customerservice	-0.6365	0.2450	-2.60	0.021	2.172
ElementaryOcc	-0.8677	0.1964	-4.42	0.001	2.566

S = 0.0126028 R-Sq = 84.1% R-Sq(adj) = 68.1%

NX

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.00600	0.01159	0.52	0.613	
personalserviceOcc	-0.35403	0.08068	-4.39	0.001	1.333
machineoperatives	-0.23434	0.08778	-2.67	0.019	3.016
Unemploy	0.3047	0.1100	2.77	0.016	2.276
h2	0.04011	0.01035	3.87	0.002	2.708

S = 0.00325083 R-Sq = 85.0% R-Sq(adj) = 68.8%

NS

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.06178	0.02609	2.37	0.034	
TechnicalOcc	0.3902	0.1245	3.14	0.008	2.833
machineoperatives	0.4683	0.1905	2.46	0.029	1.685
h2	-0.07598	0.02862	-2.65	0.020	2.720
Loneparent	0.5577	0.2053	2.72	0.018	2.730

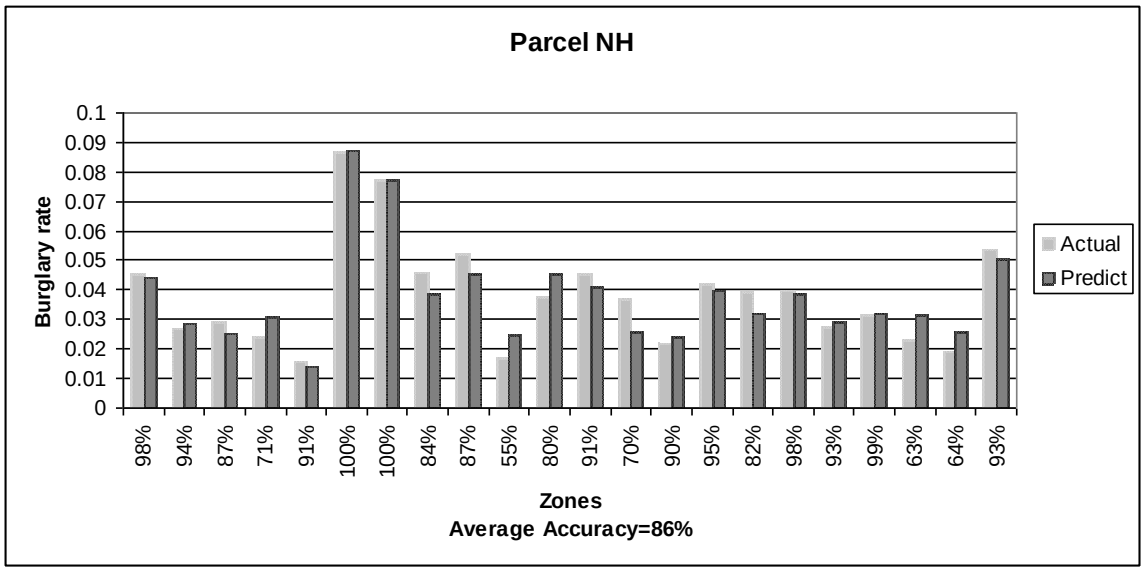
S = 0.00897228 R-Sq = 78.9% R-Sq(adj) = 52.9%

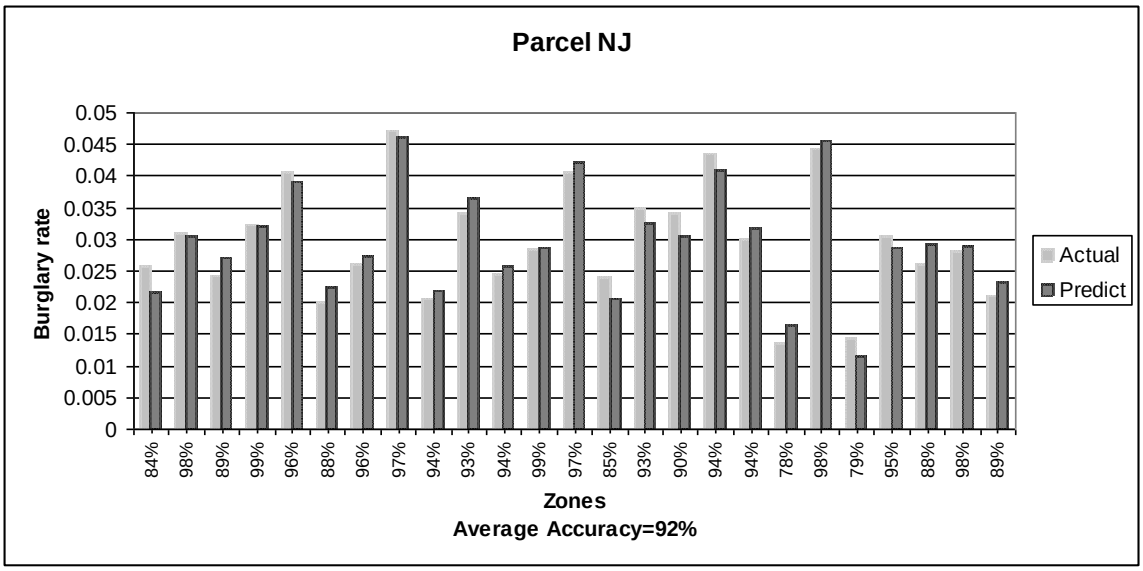
NM

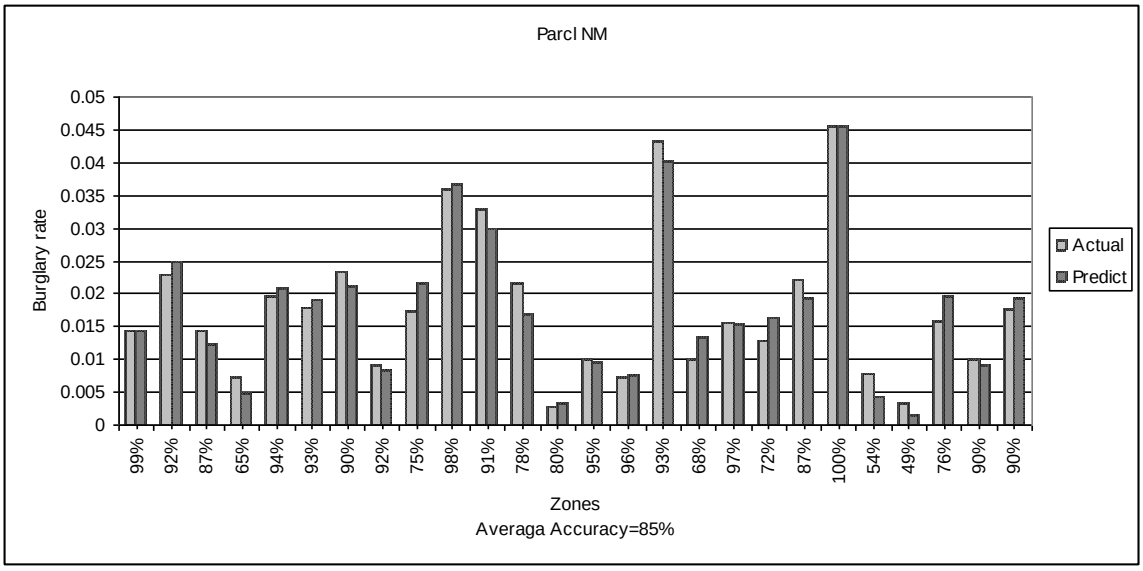
Predictor	Coef	SE Coef	T	P	VIF
Constant	0.13746	0.01233	11.15	0.000	
ManagersOcc	-0.11674	0.04460	-2.62	0.022	1.773
professionalOcc	-0.19364	0.04131	-4.69	0.001	3.514
TechnicalOcc	-0.74099	0.06941	-10.68	0.000	3.097
personalserviceOcc	0.3113	0.1228	2.54	0.026	3.707
ElementaryOcc	-0.38153	0.08750	-4.36	0.001	3.101
Loocc	0.10973	0.04299	2.55	0.025	3.104
Meocc	-0.18735	0.04494	-4.17	0.001	2.173
h2	-0.029005	0.008864	-3.27	0.007	2.249
allpensioner	-0.43796	0.05796	-7.56	0.000	2.447

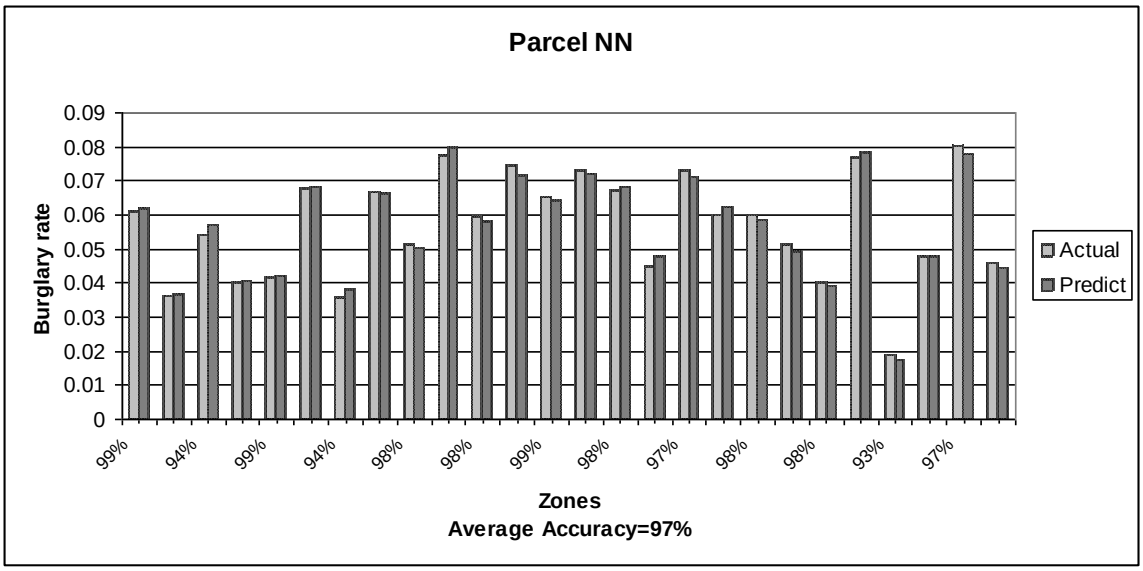
S = 0.00359163 R-Sq = 95.4% R-Sq(adj) = 89.6%

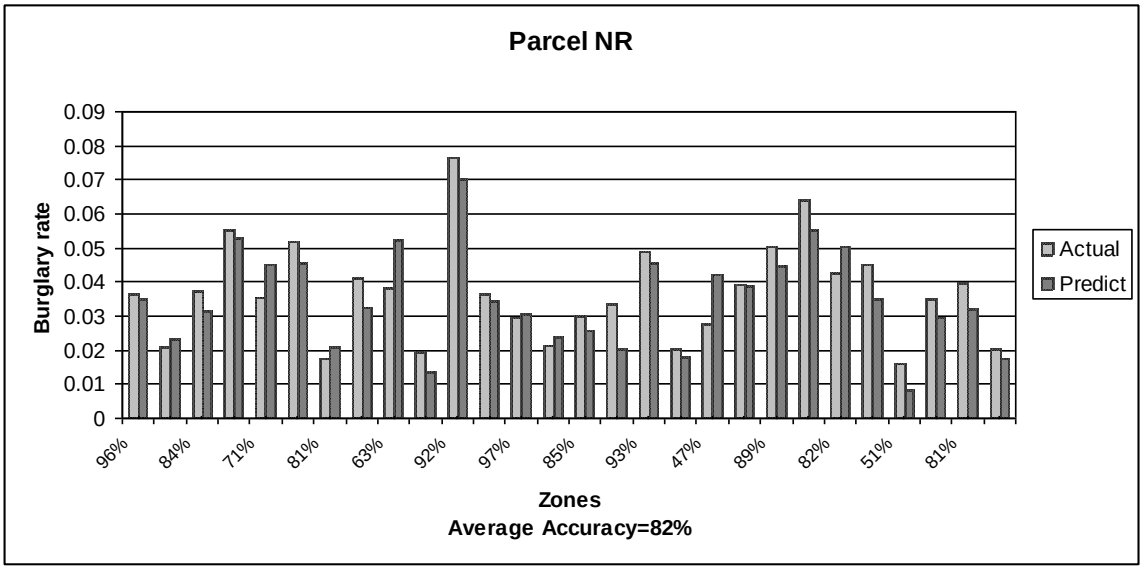
E2: Validation of the regression models within a historical data.

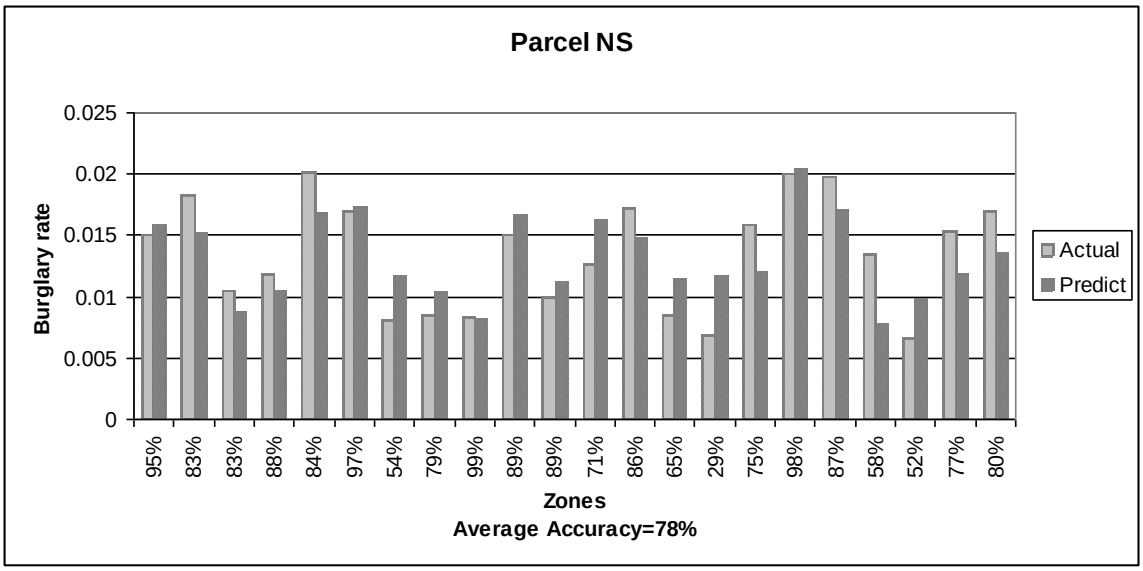


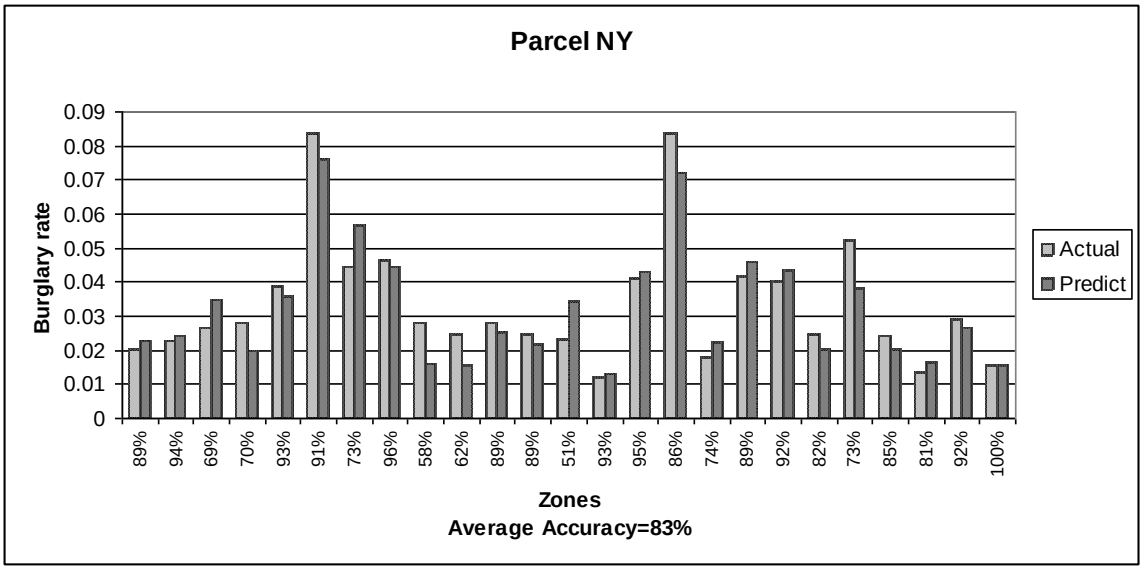

Parcel NJ
Burglary rate
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
84% 98% 89% 99% 96% 88% 96% 97% 94% 93% 94% 99% 97% 85% 93% 90% 94% 94% 78% 98% 79% 95% 88% 98% 89%
Actual
Predict
Zones
Average Accuracy=92%

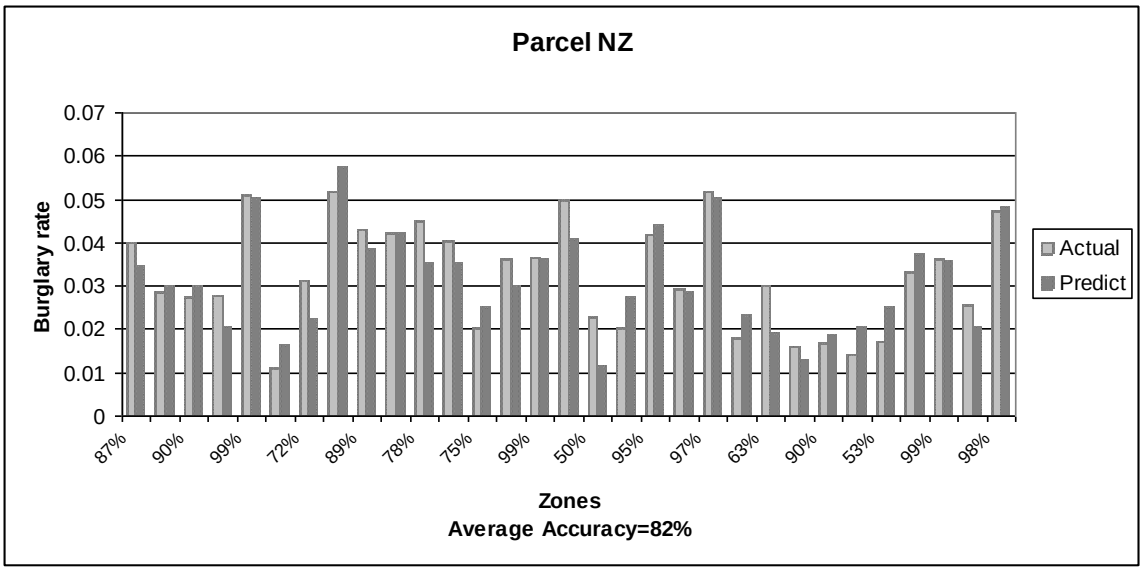

Parcl NM
Burglary rate
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
99% 92% 87% 65% 94% 93% 90% 92% 75% 98% 91% 78% 80% 95% 96% 93% 68% 97% 72% 87% 100% 54% 49% 76% 90% 90%
Actual
Predict
Zones
Averaga Accuracy=85%

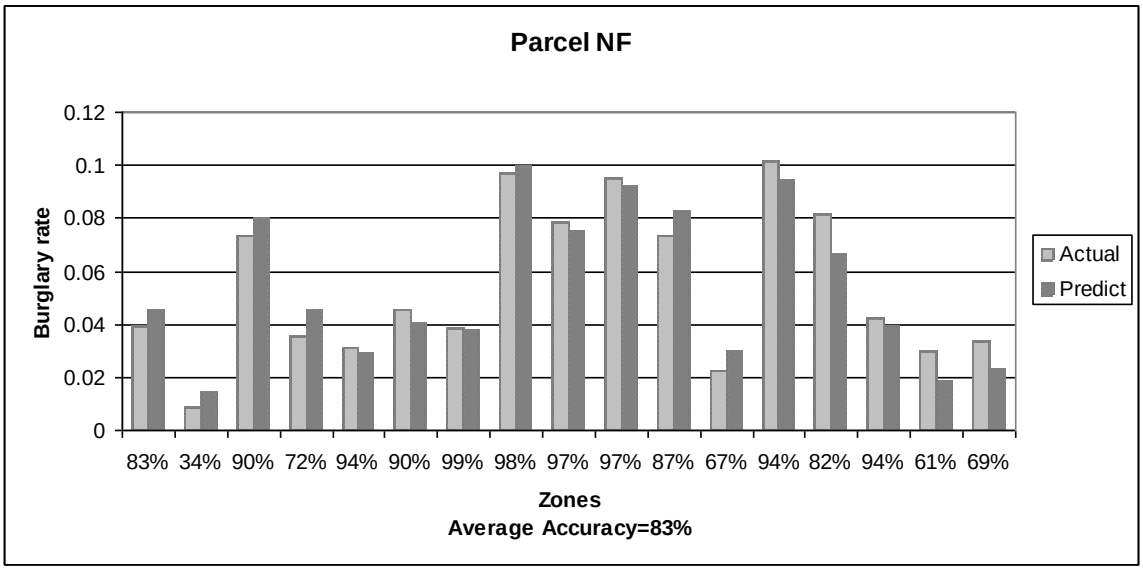

Parcel NN
Burglary rate
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
99%
94%
99%
94%
98%
98%
99%
98%
97%
98%
98%
93%
97%
Actual
Predict
Zones
Average Accuracy=97%

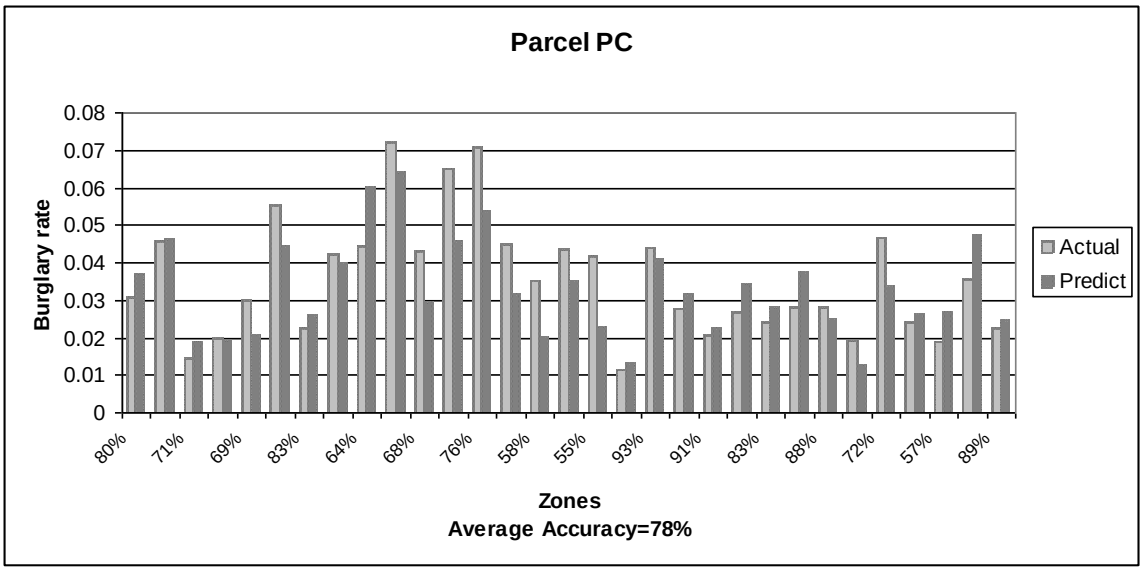

Parcel NR
Burglary rate
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
96%
84%
71%
81%
63%
92%
97%
85%
93%
47%
89%
82%
51%
81%
Actual
Predict
Zones
Average Accuracy=82%

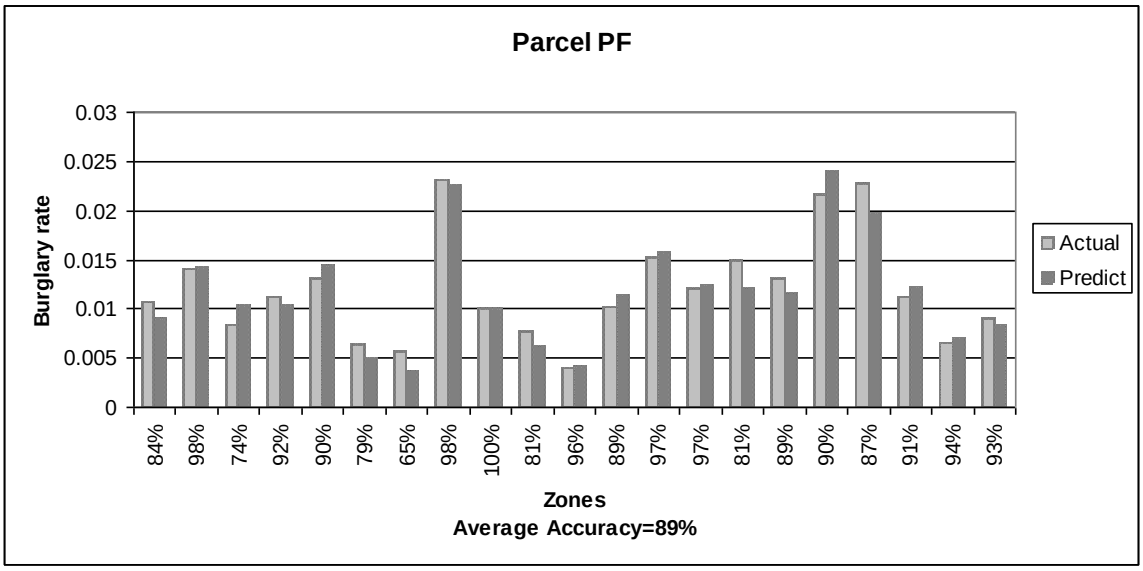

Parcel NS
Burglary rate
0.025
0.02
0.015
0.01
0.005
0
95% 83% 83% 88% 84% 97% 54% 79% 99% 89% 89% 71% 86% 65% 29% 75% 98% 87% 58% 52% 77% 80%
Zones
Average Accuracy=78%
Actual
Predict

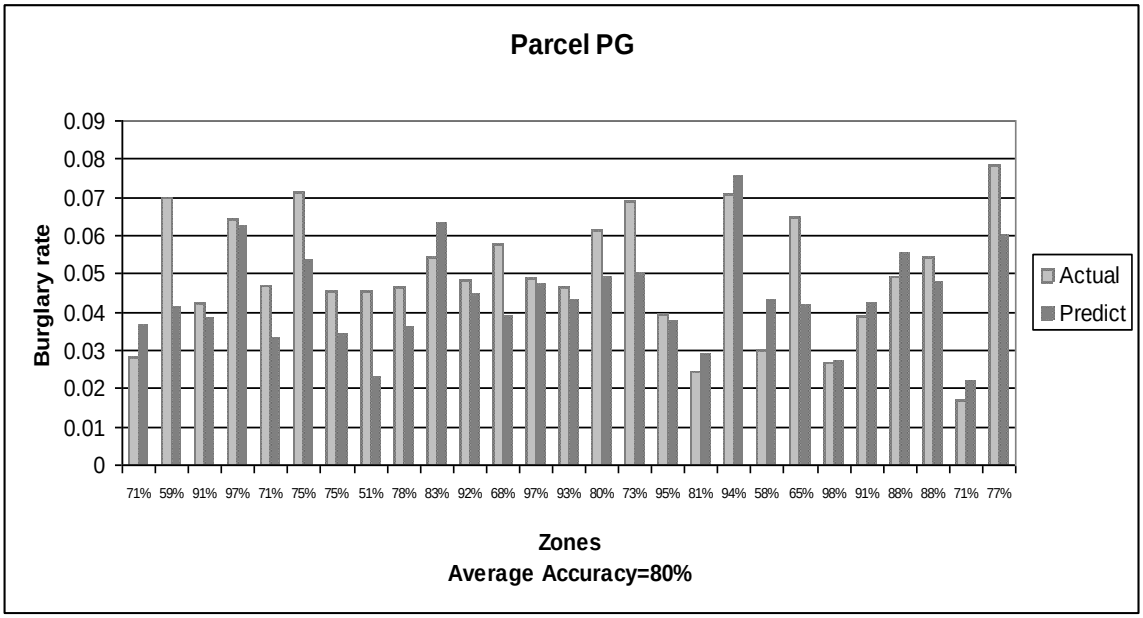

Parcel NY
Burglary rate
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
89% 94% 69% 70% 93% 91% 73% 96% 58% 62% 89% 89% 51% 93% 95% 86% 74% 89% 92% 82% 73% 85% 81% 92% 100%
Zones
Average Accuracy=83%
Actual
Predict

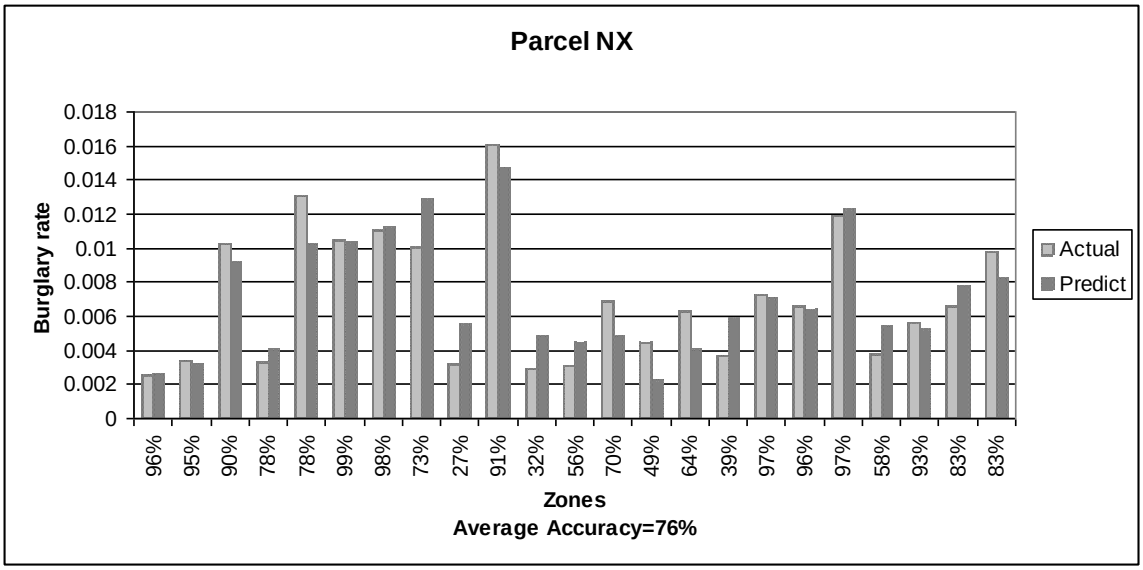

Parcel NZ
Burglary rate
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
87% 90% 99% 72% 89% 78% 75% 99% 50% 95% 97% 63% 90% 53% 99% 98%
Actual
Predict
Zones
Average Accuracy=82%


Parcel NF
Burglary rate
0.12
0.1
0.08
0.06
0.04
0.02
0
83% 34% 90% 72% 94% 90% 99% 98% 97% 97% 87% 67% 94% 82% 94% 61% 69%
Actual
Predict
Zones
Average Accuracy=83%

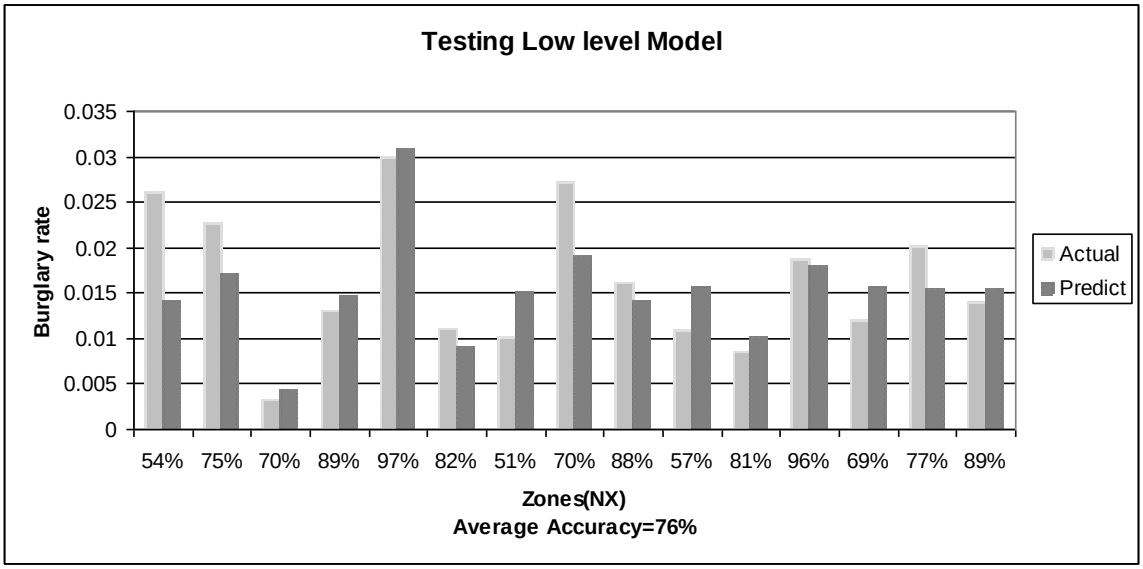

Parcel PC
Burglary rate
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
80%
71%
69%
83%
64%
68%
76%
58%
55%
93%
91%
83%
88%
72%
57%
89%
Actual
Predict
Zones
Average Accuracy=78%

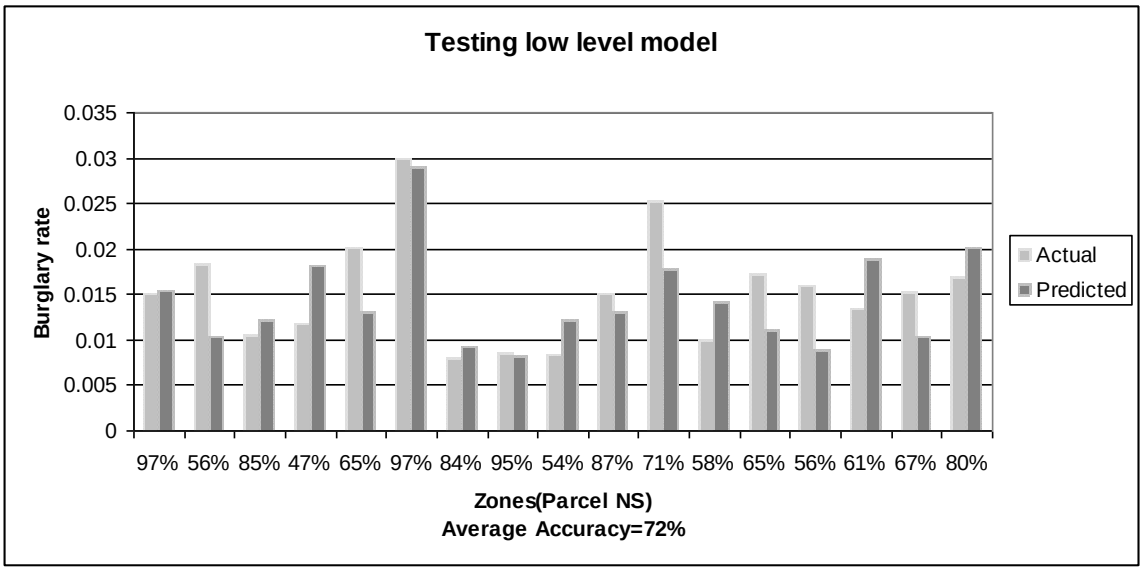

Parcel PF
Burglary rate
0.03
0.025
0.02
0.015
0.01
0.005
0
84%
98%
74%
92%
90%
79%
65%
98%
100%
81%
96%
89%
97%
97%
81%
89%
90%
87%
91%
94%
93%
Actual
Predict
Zones
Average Accuracy=89%

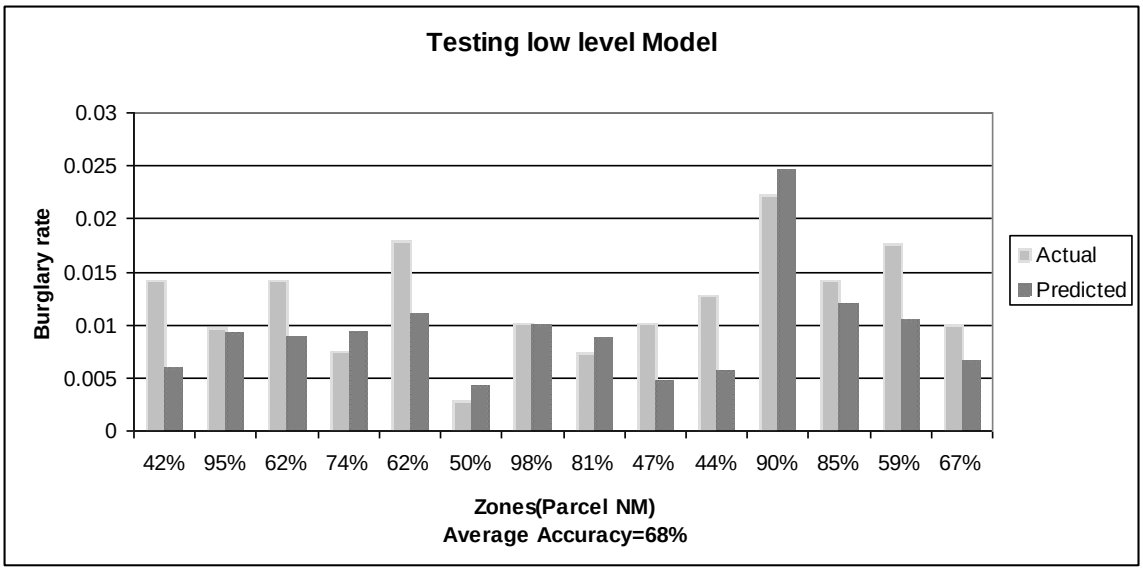

Parcel PG
Burglary rate
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
71% 59% 91% 97% 71% 75% 75% 51% 78% 83% 92% 68% 97% 93% 80% 73% 95% 81% 94% 58% 65% 98% 91% 88% 88% 71% 77%
Actual
Predict
Zones
Average Accuracy=80%

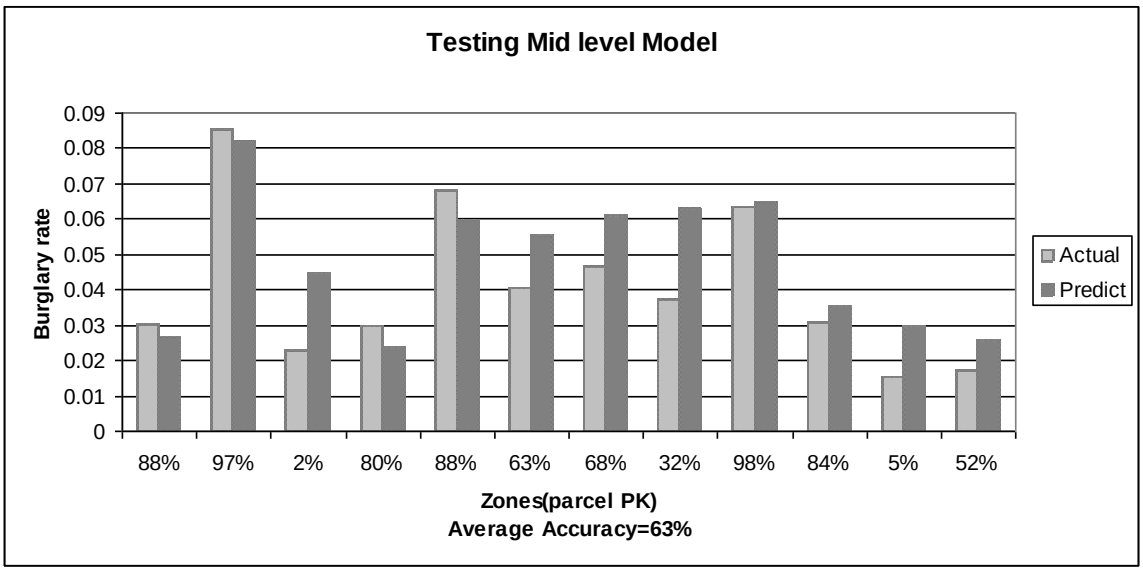

Parcel NX
Burglary rate
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
96% 95% 90% 78% 78% 99% 98% 73% 27% 91% 32% 56% 70% 49% 64% 39% 97% 96% 97% 58% 93% 83% 83%
Actual
Predict
Zones
Average Accuracy=76%

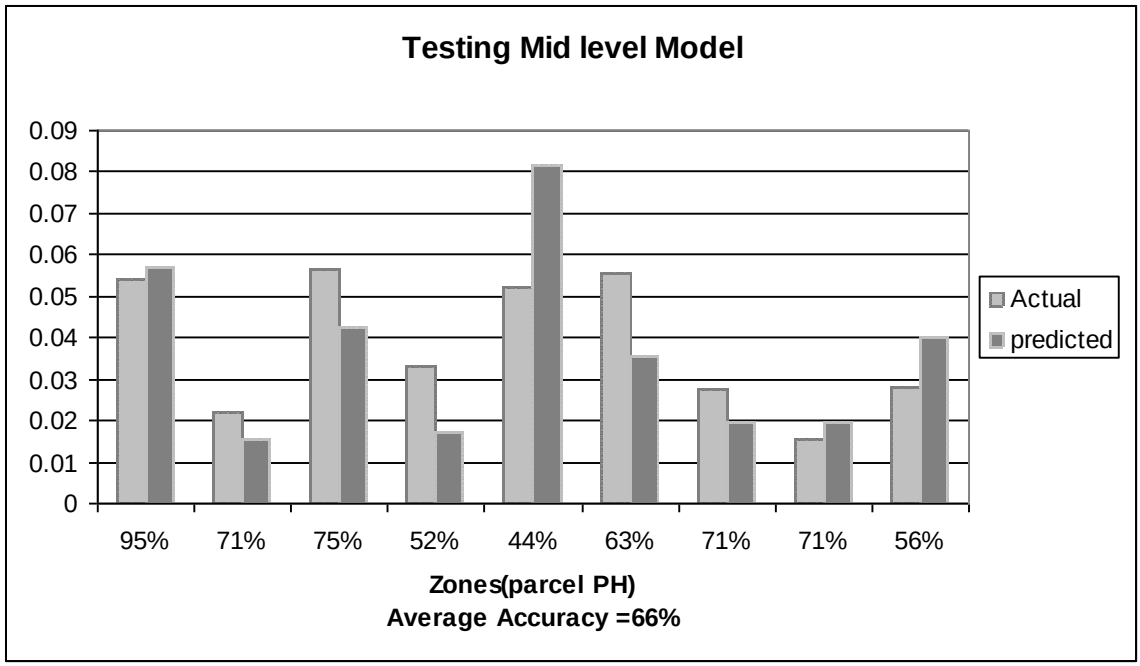
E3: Validation of the regression models within a new data.

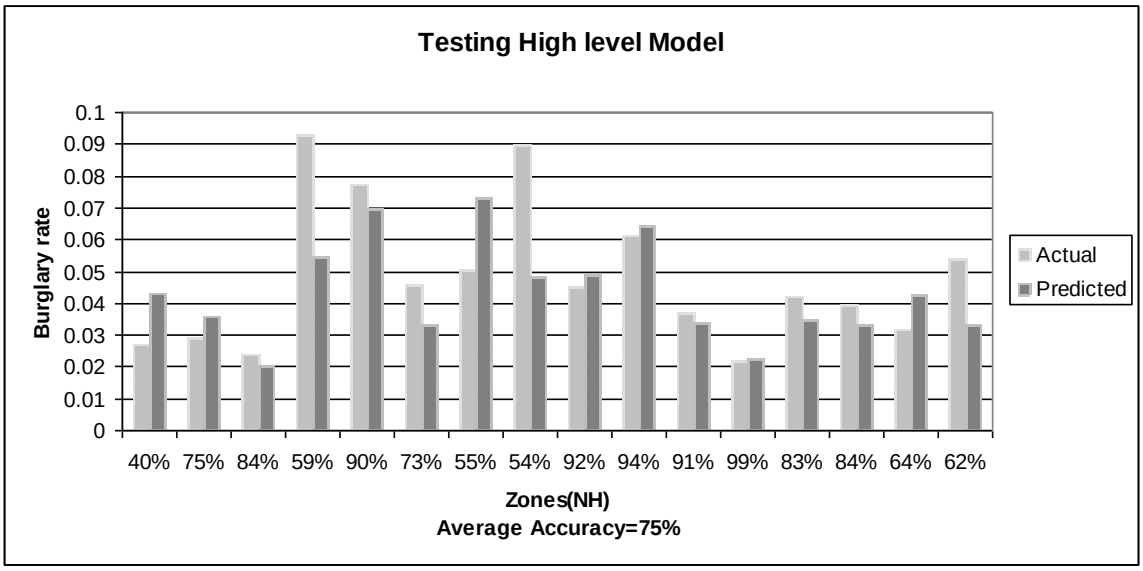


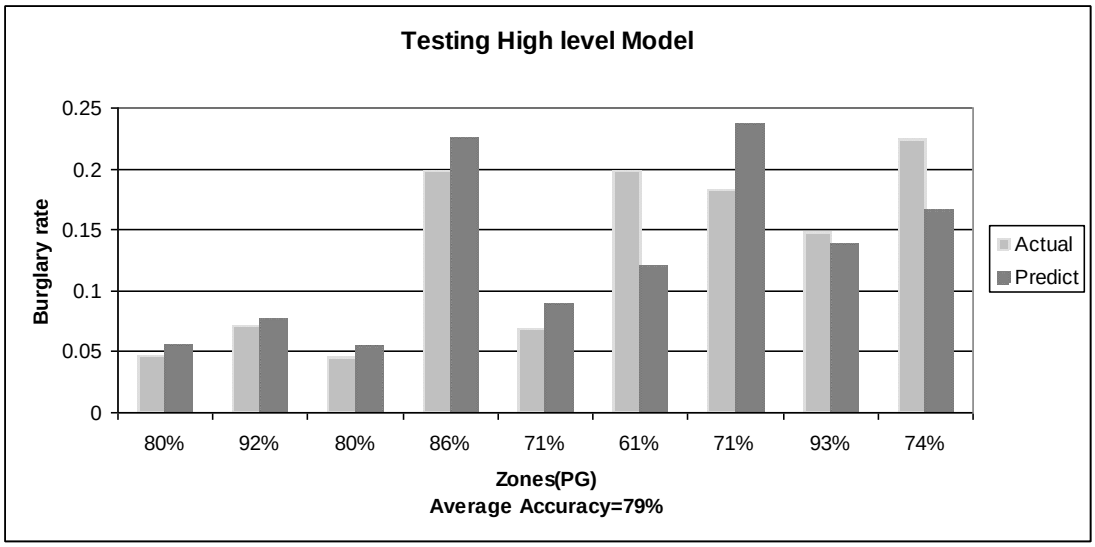


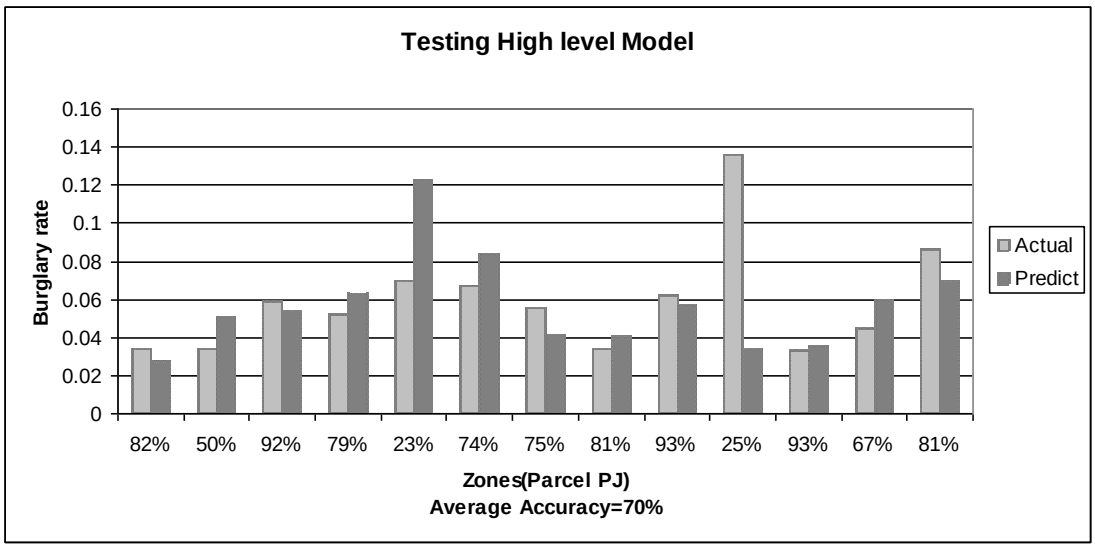

Testing low level Model
Burglary rate
0.03
0.025
0.02
0.015
0.01
0.005
0
42% 95% 62% 74% 62% 50% 98% 81% 47% 44% 90% 85% 59% 67%
Zones(Parcel NM)
Average Accuracy=68%
Actual
Predicted

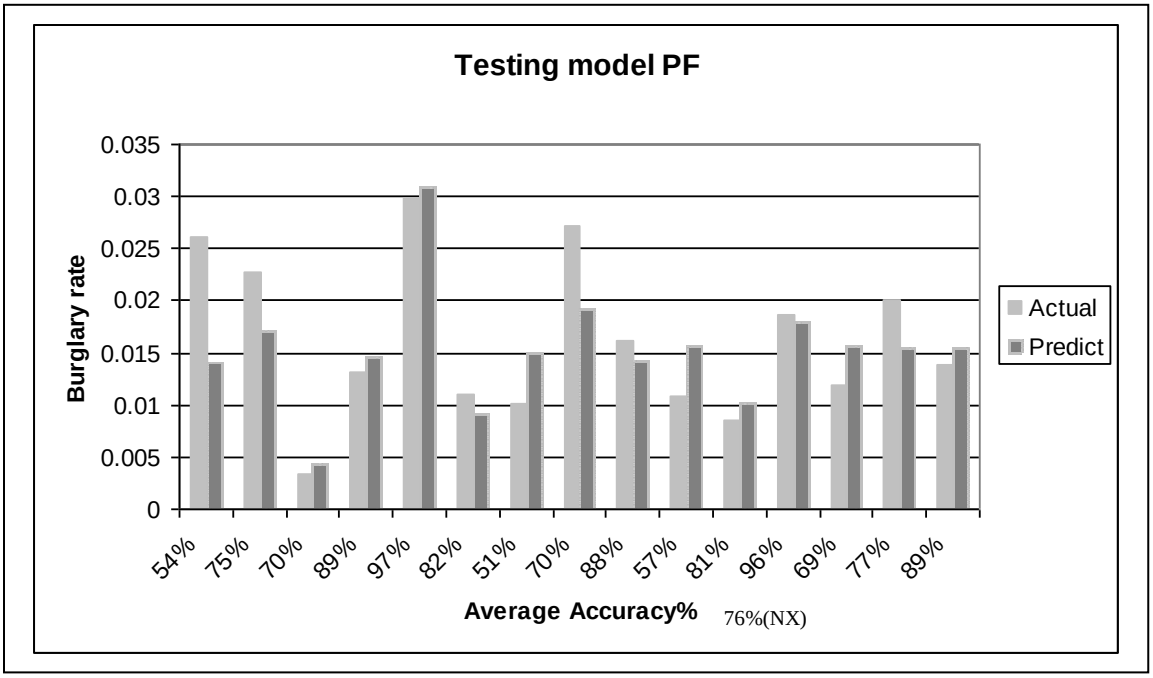

Testing Mid level Model
Burglary rate
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
88% 97% 2% 80% 88% 63% 68% 32% 98% 84% 5% 52%
Zones(parcel PK)
Average Accuracy=63%
Actual
Predict


Testing Mid level Model
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
95% 71% 75% 52% 44% 63% 71% 71% 56%
Actual
predicted
Zones(parcel PH)
Average Accuracy =66%

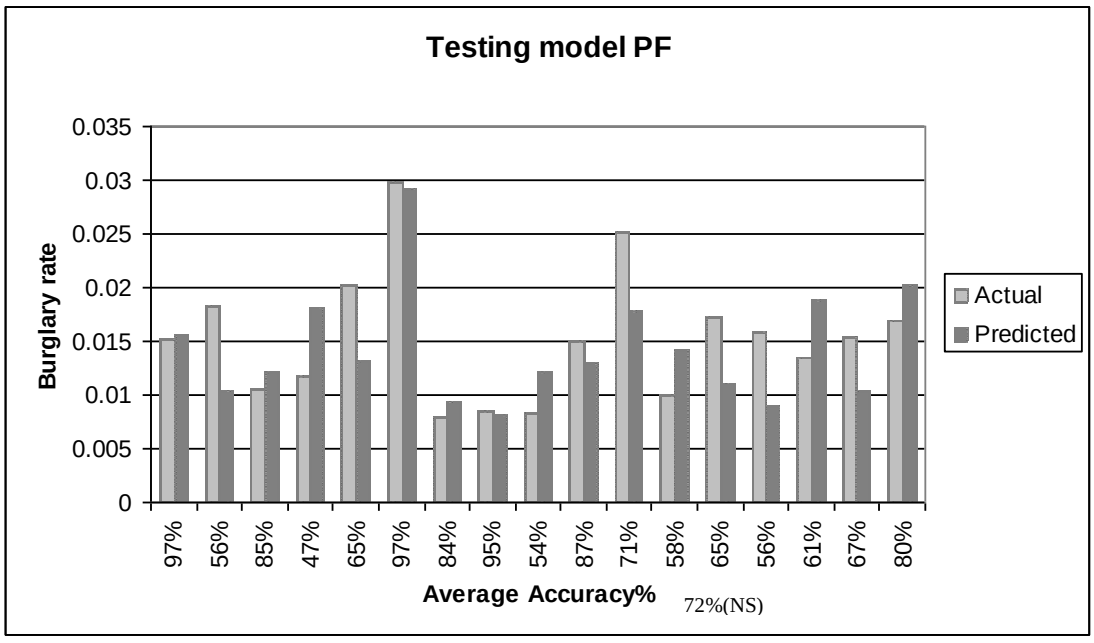

Testing High level Model
Burglary rate
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
40% 75% 84% 59% 90% 73% 55% 54% 92% 94% 91% 99% 83% 84% 64% 62%
Actual
Predicted
Zones(NH)
Average Accuracy=75%


Testing High level Model
Burglary rate
0.25
0.2
0.15
0.1
0.05
0
80%
92%
80%
86%
71%
61%
71%
93%
74%
Actual
Predict
Zones(PG)
Average Accuracy=79%

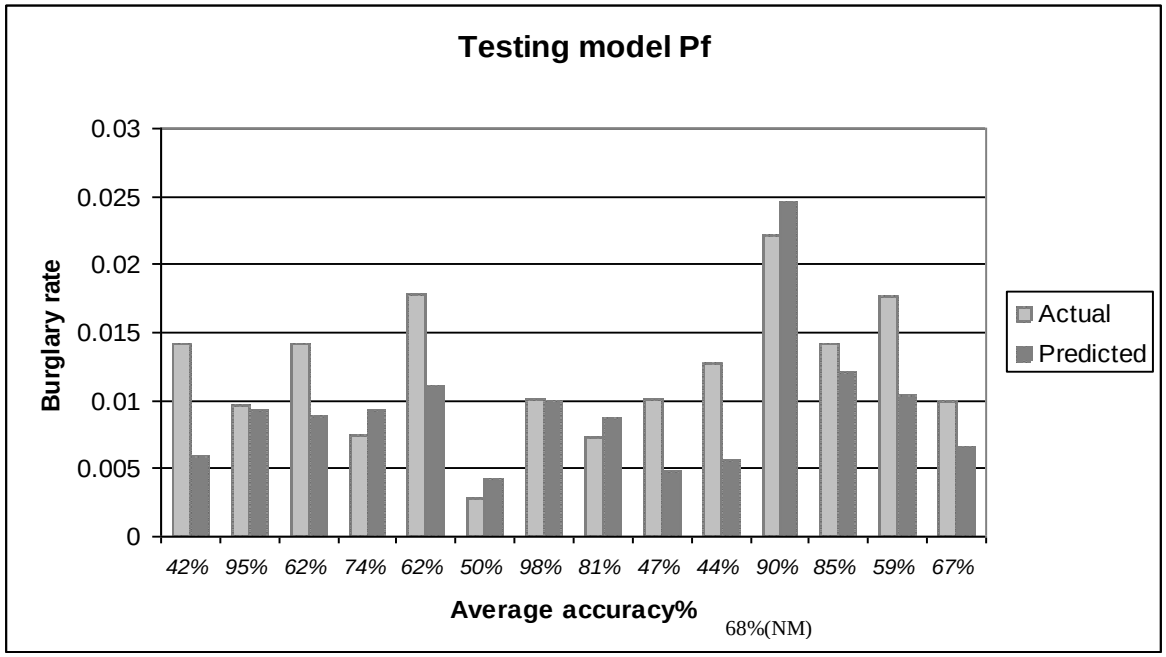

Testing High level Model
Burglary rate
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
82%
50%
92%
79%
23%
74%
75%
81%
93%
25%
93%
67%
81%
Actual
Predict
Zones(Parcel PJ)
Average Accuracy=70%

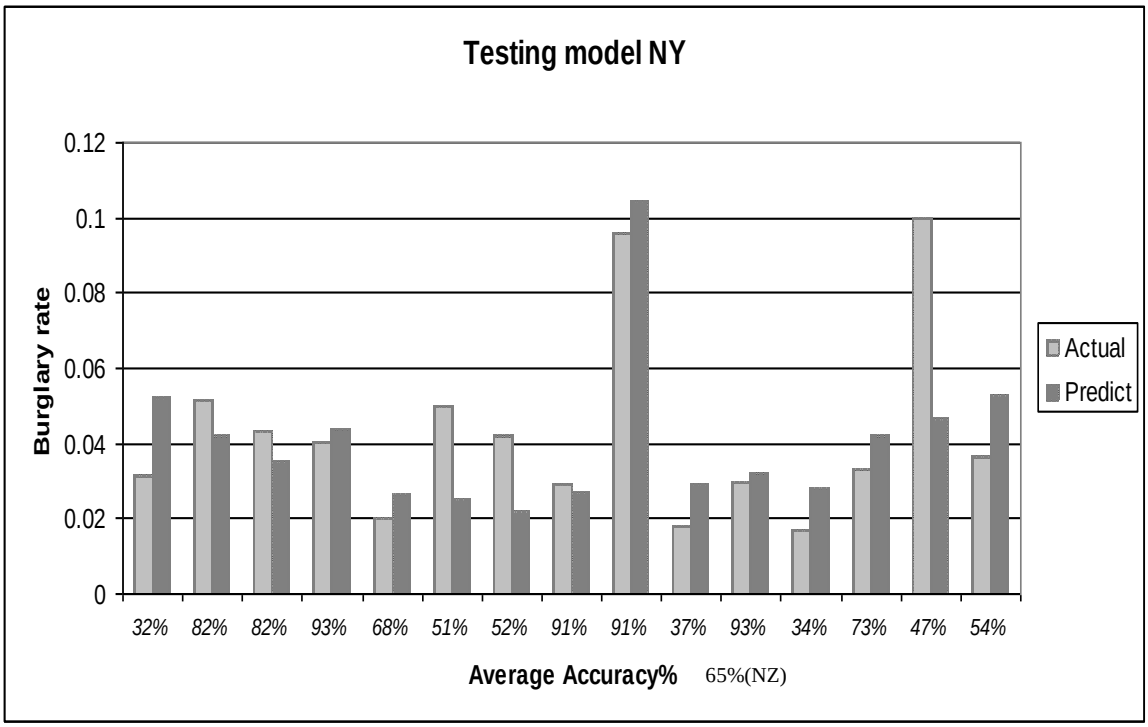

Testing model PF
Burglary rate
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
54% 75% 70% 89% 97% 82% 51% 70% 88% 57% 81% 96% 69% 77% 89%
Average Accuracy%
Actual
Predict

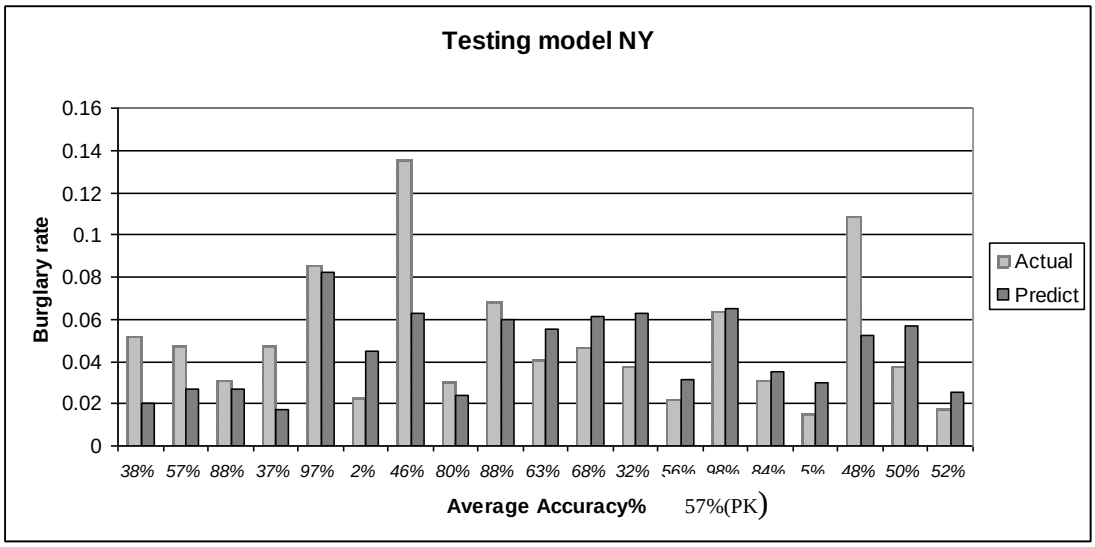
76%(NX)


Testing model PF
Burglary rate
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
97% 56% 85% 47% 65% 97% 84% 95% 54% 87% 71% 58% 65% 56% 61% 67% 80%
Average Accuracy%
Actual
Predicted

72%(NS)


Testing model Pf
Burglary rate
0.03
0.025
0.02
0.015
0.01
0.005
0
42% 95% 62% 74% 62% 50% 98% 81% 47% 44% 90% 85% 59% 67%
Average accuracy%
68%(NM)
Actual
Predicted


Testing model NY
Burglary rate
0.12
0.1
0.08
0.06
0.04
0.02
0
32% 82% 82% 93% 68% 51% 52% 91% 91% 37% 93% 34% 73% 47% 54%
Average Accuracy%
65%(NZ)
Actual
Predict


Testing model NY
Burglary rate
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
38% 57% 88% 37% 97% 2% 46% 80% 88% 63% 68% 32% 56% 98% 84% 5% 48% 50% 52%
Actual
Predict
Average Accuracy% 57%(PK)