ANALYSIS OF SPATIAL DISTRIBUTIONS OF ROAD ACCIDENTS

A thesis
submitted in partial fulfilment of the requirements for the
degree of Doctor of Philosophy in
Transportation Engineering

By

Haran Arampamoorthy

Department of Civil Engineering
University of Canterbury
New Zealand

January 2005

ABSTRACT

Traffic accidents result in life and financial loss to the society. In developing countries traffic fatalities are comparable to other leading causes of death. The need for the analysis of the spatial distribution of traffic accidents, as an aid to select the most appropriate type of accident reduction programme (e.g. site, route and area plans) and assessing the effectiveness of such plans after implementation, is very important. The current practice (e.g. visual examination) for assessing the spatial distribution of accidents is reviewed. In this thesis, various methods for the statistical analysis of spatial distributions of accidents (including quadrat and nearest – neighbour methods) are reviewed and further improvements are described.

Accidents are random events subject to both temporal and spatial variation. The basic variables for accident analysis are; distance and direction of accident locations in terms of North and East co-ordinates, azimuth, and the year of the accident. A new method for analysing the spatial pattern is proposed, whereby detection of a particular pattern will indicate which type of accident reduction programme is most appropriate. The method distinguishes the spatial distribution (point cluster, line cluster, area cluster or a completely spatially random distribution) of accidents in different types of road networks (regular or irregular and dense or sparse). The method can also help assessment of the changes in spatial distributions of accidents.

ACKNOWLEDGMENTS

I thank my supervisor, Associate Professor Alan J. Nicholson, Department of Civil Engineering, for providing valuable guidance on methods for identifying spatial distribution of accidents and the development of a computer program to analyse the accident data and provide helpful information to identify the most appropriate accident reduction plans. His help and advice at all stages of my endeavour is gratefully acknowledged.

I am grateful to my wife for her unsparing support, my two sons, my father and friends, for their encouragement and help at all stages of my post-graduate studies.

I thank the Staff of the Department of Civil Engineering for employing me as Teaching Assistant at various stages of my tenure as a post-graduate. Apart from the financial gain, I acquired experience in other branches of engineering.

CONTENTS PAGE No.

LIST OF FIGURES

LIST OF TABLES

SYMBOLS AND TECHNICAL TERMS

A_c:	The area under the cumulative distribution of events profile for a cluster distribution
A_{csr}:	The area under the cumulative distribution of events profile for a CSR distribution
Accident-centred quadrat:	Quadrat centred on an accident location
A_{line} :	Area under the dissimilarity coefficient profile for a line cluster distribution
Annual accident rate:	Average number of accidents per year
A_{point}:	Area under the dissimilarity coefficient profile for a point cluster distribution
A_r :	The area under the cumulative distribution of events profile for a regular distribution
Area action plan:	An accident reduction plan for a black area
Black area:	An area for which the number of accidents is substantially more than in other similar areas
Black route:	A route or section of route for which the number of accidents is more than for other similar routes
Black spot:	A site, for which the number of accidents is greater than for other similar sites
BSATUQM:	Computer program using quadrat method to identify black spots in spatial distributions of accidents
BRATUQM:	Computer program using quadrat method to identify black routes in spatial distributions of accidents
CBD:	Central Business District
cdf:	Cumulative distribution function
Characteristic length of a cluster:	The maximum length between any two accidents in a cluster
Cluster:	A location in which accidents are concentrated
Cluster locations	Locations where accidents appear to be clustered

Completely Spatially Random:	An accident distribution, which satisfies stationarity and isotropy conditions
Continuum:	An area in which accidents can occur anywhere
Cost density:	Social cost of accidents per unit length of road
cov:	Co-variance
CSR:	Completely Spatially Random
CV:	Coefficient of Variation
Dissimilarity:	A measure used in cluster analysis to show how far apart accidents or groups of accidents are
$d_{(i)}$:	The distance from a test-location to the i^{th} neighbor event
d_{ij}:	The distance between the i^{th} and j^{th} accidents
d^k_i :	The distance to the k^{th} nearest neighbour from the i^{th} accident
$d_{(N)}$:	The distance from a test-location to the N^{th} nearest accident
E[c]:	Expected quadrat count
E(i):	Expected number of events within a distance $d_{(i)}$
E(N):	Expected number of events within a distance $d_{(N)}$
$F_1(X)$:	The cdf for the observed distribution
$F_2(X)$:	The cdf for a CSR distribution
G:	Coefficient of concentration
I:	Moran's index
ICF:	Index of cluster frequency in quadrat analysis
ICR:	Index of mean crowding in quadrat analysis
ICS:	Index of clumping in quadrat analysis
IP:	Index of patchiness in quadrat analysis

Isotropy:	Where the relative locations of any two events depend only on the distance between the two accidents but not in the direction between the two accidents
K:	The Kuiper-Stephens test statistic
$K_{(h)}$:	Expected number of extra events within a circle (radius h) divided by the intensity of the spatial process
K*	A modified K
Kurtosis (KUR):	A measure of the peakiness or flatness of an accident count frequency polygon
Lattice:	Area in which accident can occur only on defined lines (eg. roads)
Line cluster:	A route on which accidents are concentrated
L_{max} :	Maximum distance between two accidents or groups of accidents
L_{min}:	Minimum distance between two accidents or groups of accidents
Location:	A site, route or area
m*:	Mean
m	Number of directions of event concentrations (or modes)
MaxCou:	Maximum quadrat count
MI:	Morisita's Index in quadrat analysis
m_i :	The expected number of annual accidents at the i^{th} location
M_h:	The calculated mean quadrat count for quadrats with radius h
N:	Number of events within a sample

$N_{(h)}$:	The number of distinct pairs of points separated by distance h
NNAT:	A computer program using Nearest-Neighbour Analysis Techniques to analyse spatial distributions of accident
P(i/N):	Expected proportion of events within $d_{(i)}$
Point cluster:	A site at which accidents are concentrated
Proportion of mean count (PMC):	100 times the ratio of the mean count for a particular quadrat radius and the mean count for the maximum quadrat radius
$Q_{(S)}$	The set of observed accident counts for locations (S)
Random/CSR locations	Locations where the nearest neighbours appeared to be randomly distributed
Regular locations	Locations where the nearest neighbours appeared to be regularly spaced
Risk density:	Social cost per vehicle kilometre of travel
Route action plan:	An accident reduction plan appropriate to a black route
R	Measure of concentration (magnitude of the resultant vector)
$\bar{\mathbf{R}}$	The normalized measure of R for unimodal data
$\bar{\mathbf{R}}_{\mathbf{m}}$	The normalised measure of concentration for multimodal data
$\mathbf{s}_i$:	The vector from the origin to the i^{th} position in two-dimensional space
S :	The set of spatial locations $\{s_1, s_2, \ldots, s_i, \ldots, s_n\}$ where accident can occur
Site:	A small area where accidents can occur
Site action plan:	An accident reduction plan appropriate to a black spot
Single accident quadrats:	Quadrats containing single accidents

Skewness (sk):	A measure which indicates more accident counts lie on one side of the mean than on the other side
Stationarity:	Where the probability of an accidents occurring is the same for all possible locations
%SAQ:	The number of single accident a quadrats (per 100 quadrats)
%QAM:	The number of quadrats having a quadrat count greater than the mean quadrat count (per 100 quadrats)
U^2 :	The Watson test statistic
UTAR:	Underlying True Accident Rate
VKT:	Vehicle kilometres of travel
X_{ij}:	The observed accident counts at the i^{th} location during the j^{th} year
$2\gamma(\mathbf{s_i} - \mathbf{s_j})$:	The variogram
λ:	The intensity of a spatial process
μ:	Expected accident count
σ^2:	Variance (var) of accident counts
θ_{ij} :	The direction from the i^{th} accident to the j^{th} accident
θ^k_i :	The direction to the k^{th} nearest-neighbour from the i^{th} accident

Chapter 1

INTRODUCTION

1.1 General

Most traffic accidents may be considered to be random events which depend on time and location. Thus the annual accident count at a particular location will vary from year to year, and for a particular year, the annual accident count will vary from location to location. That means that accident counts are subject to both temporal and spatial variations. Some of the accidents may not be completely random, in that the temporal and spatial variations in their occurrence can be explained in part by variations in the factors involved in accident occurrence.

Accidents are rare events and generally not uniformly or equally distributed over the road system; they are often clustered at sites, along routes or within areas. The basis for a strategic approach to accident reduction by specific engineering measures is to develop a framework within which priorities may be set for implementing measures identified through accident reduction analysis techniques.

There are several types of accident reduction programs [IHT, 1990] including:

1. Single site plans for "the treatment of a specific type / types of accidents clustered at a single location or over a short length of road" (black spots);
2. Route action plans for a "road having above average accidents for that type or class of road";
3. Area action plans for "accidents which are scattered too sparsely";
4. Mass action plans for "locations having a particular type of accidents".

Because limited funds are available, the accident reduction programme are generally designed to give the maximum benefit/cost ratio. Based upon the UK experience, the appropriate objectives for accident reduction plans suggested in [IHT 1990] are:

- 33% average accident reduction at treated sites and first year rate of return of more than 50% for a single site treatment;

- 15% average accident reduction on treated route section and first year rate of return of more than 40% for a route action plan;
- 10% average accident reduction in a treated area and 10%-20% first year rate of return for an area action plan;
- 15% average accident reduction at treated sites and first year rate of return of more than 40% for a mass action plan.

Past experience in New Zealand [LTSA, January 1998] indicates an overall reduction of 28% in crashes between 1985 and 1998 at treated locations. This figure is consistent with the UK-based average accident reduction estimates. The emphasis in New Zealand has been on site plans as well as some route plans. Hence we would expect the accident reduction to be between 33% and 15%, but closer to the higher value.

The guideline indicates that the expected percentage of accident reduction and economic return are higher for site plans than route plans, and higher for route plans than area plans. However, this does not mean that the site plan will be always the best option for an accident reduction plan.

The selection of an accident reduction plan depends on the pattern of accidents; for example, the site plan would be appropriate for accidents clustered at single sites or short lengths of roads (black spots). If the site plan is appropriate it should reduce the level of accident clustering at the site. When the level of accident clustering at sites is reduced then there may be routes that become route clusters. Individual sites along a route cluster may not be identified as black spots for a site plan but when all the sites along the route are aggregated, it may become a black route. An appropriate route plan will further reduce the level of accident clustering. A further reduction of accidents is possible by implementing area plans for areas with relatively high numbers of accidents but which may not contain any black spots or black routes. As suggested by Nicholson (1989 and 1990) "natural progression" may take place from site plans to route plans to area plans, as the nature and extent of accident clustering changes. Hence the best plan depends upon the spatial distribution of accidents.

The cost of an accident reduction treatment will depend on the type of treatment and the type of treatment depends on the type of accidents. It is likely that simple and cheap

treatments will be readily identified for solving the problems at hazardous locations in the early stages of accident reduction work. For example, low cost treatments (e.g. erecting new traffic signs and installing pavement markings) may be appropriate during the initial stages of an accident reduction programme. According to the "law of diminishing returns", each increment of progress generally requires greater effort as one makes progress. Hence in the later stages of an accident reduction programme, it will probably become more difficult to diagnose safety problems and identify cost-effective treatments, and the cost of treatments will generally be greater. In 1988 around 800 fatal accidents occurred. Applications of several accident reduction programmes helped to reduce the number to around 450 in 2003. LTSA [October 2004] proposed to bring down the rate of fatal accidents to not greater than 300 per annum by 2010. It should be noted that the rate of fatal accidents was reduced by only 350 in the 15 year period up to 2003 (i.e. about 23.3 per year). According to the LTSA proposals the fatal accident rate needs to be brought down by 150 in seven years (i.e. 21.4 per year). These rates of accident reduction are very similar. However, because of the law of diminishing returns, it may be very difficult to bring down the road toll as proposed with single site and route plans (the current approach).

Shaikh [1990] noted that between 1983 and 1988, only 3% - 6% of accidents in New Zealand occurred at sites with five or more accidents. In this study, a site was defined as a 70 m square quadrat centred on an accident. He found that accidents are less clustered and more dispersed in New Zealand compared to the UK, as shown below in Table 1.01. So it might be appropriate to focus attention on route and area plans in New Zealand rather than site plans.

Table 1.01: Percentage of accidents in various group sizes (UK and NZ).

Accident group size	Percentage of accidents	
	United Kingdom	New Zealand
1	35-51	50-57
2	16-25	25-29
3	4-17	10-11
4	5-10	4-6
5	15-29	3-6

When considering how to reduce accidents within a large network (e.g. the Christchurch network) the issues to be addressed are:

- Which plan type should be selected initially?
- When should one progress to a different plan type?

To answer these questions the spatial distribution of accidents needs to be assessed and, if an accident reduction plan is implemented, the spatial distribution of accidents should be monitored regularly. This will be useful in selecting the most appropriate accident reduction plan and subsequently assessing whether it has been effective or not.

1.2 Research method

Various exploratory analytical techniques are needed for accident investigation. At the preliminary stage, accident locations need to be analysed using statistical analysis techniques, to identify any strong spatial pattern, such as accidents being clustered at sites, or along roads, or within parts of the study area. If accidents are so clustered then site, routes or area plans should be developed, respectively.

The current practice for the analysis of spatial distributions of accidents depends upon visually examining a map showing the location of accidents in the road network. Such a map is shown in Figure 1.01. The assessment process is very subjective, and relies heavily on exercising judgement, in order to decide whether there is an observable dominant pattern and what it is (that is, whether the accidents are clustered at sites, or along roads, or within particular parts of the study area).

Visually examining a Completely Spatially Random (CSR) distribution of accidents may well lead to a spurious pattern [Cressie 1993], and different observers may not agree on the nature and strength of any spatial pattern. Therefore a technique, which can quantify the nature and strength of any pattern (that is, whether clustered at sites, along roads or within particular parts of the study area), would assist identification of the most appropriate type of accident reduction plan. Quantitative techniques would also help to assess the effectiveness

of the plan after it has been implemented, by determining whether it has made a statistically significant change in the spatial distribution of accidents.

There are several well-known statistical analysis techniques to analyse spatial distributions. These are quadrat analysis [Ripley 1981, Cressie 1993], the nearest-neighbour method [Cressie 1993], and cluster analysis [Michael 1973]. These techniques can be used to identify whether there is a dominant pattern and the nature of the pattern. Once it has been established that accidents are clustered at sites, along roads or within particular areas, it is a straightforward task to identify the most effective accident reduction plan.

Considerable research has been done [Nicholson 1995, 1998 & 1999, Anujah 1997] to evaluate the three statistical techniques (i.e. cluster analysis, nearest-neighbour methods and quadrat analysis) using hypothetical distributions. The problems associated with these analysis methods are discussed in Chapters 4, 5 and 6, and attempts are made to address the problems and to improve the methods of analysis. The methods of analysis are tested using hypothetical distributions. Then the techniques are applied to actual accident distributions.

Three basic types of pattern (accidents clustered at sites, along roads and within an area) are generated via a two-stage "parent and daughter" procedure [Ripley, 1981]. These three basic patterns are mixed with a CSR (Completely Spatially Random) distribution in varying proportions, to produce three test pattern sequences. These test patterns sequences involve gradually increasing the strength (or proportions) of the basic patterns and gradually reducing the spatially random component of the distribution.

The statistical analysis methods can be applied to selected types of accidents. For instance, the spatial distribution of accidents involving vehicles and roadside poles could be analysed, to ascertain the extent to which they are clustered (along particular routes or within a particular area, say), or random. This would assist practising engineers to design a programme to reduce this specific type of accident and to assess the feasibility of such a programme.

1.3 Basic theory

Cressie [1993] suggests that there are three traditional spatial distributions (regular, CSR, and clustered) as shown in Figures 1.02, 1.03 and 1.04. The traditional regular / random / cluster classification allows clustering at points only, but does not allow for the accidents to be concentrated along a road (line cluster) or for accidents to be clustered in areas (area cluster). An alternative classification for accident distributions was proposed by Nicholson [1998]. This classification system is based on the three different types of clusters and accident reduction plans proposed by the IHT [1990]. These three cluster patterns are:

- accidents clustered at sites (point clusters), that is black spots (see Figure 1.05);
- accidents clustered along roads (line clusters), that is black routes (see Figure 1.06);
- accidents clustered in an area (area cluster), that is a black area (see Figure 1.07).

With this classification system, it is a straight-forward task to identify the best accident reduction plans;

- for point clusters, a site plan is best;
- for a line clusters, a route action plan is best;
- for area clusters, an area action plan is best.

The distinction between different types of clusters is based on the relative size of the characteristic dimensions (length and width) of the clusters. The length and width are small for point clusters, the width is small and the length is large for line clusters, and the length and width are large for area clusters, where "small" typically means 70m and "large" typically means 1000m, as discussed in Chapter 2. An area cluster may involve a majority of the events being densely located in a sub-area (see Figure 1.07). Area clusters can comprise events which are regularly spaced or CSR distribution throughout a sub-area.

It is important to understand the production mechanism of accident clusters. The term clustering could be explained with respect to internal cohesion and / or external isolation of events [Ripley, 1981]. Internal cohesion occurs when the events are attracted to each other and external isolation occurs when the events are repulsed. Point clusters occur when the events seem to be attracted to certain sites, line clusters occur when events seem to be attracted to certain paths, and area clusters occur when events seem to be equally attracted to all the locations or randomly distributed within the area, but are not attracted by certain

sites or certain paths. Here the word "attracts" means the probability of accident occurrence is slightly higher in a certain place than in other places because of the road environment conditions, as explained in the Chapter 2. Certain sub-areas seem to 'attract' events (accidents) but within the sub-area the accident locations are not clustered. The objective is to identify whether the occurrences of events are dependent on a certain special characteristic of a site or route or area, that will reveal the type of spatial pattern in the given data and hence the best type of accident reduction plan.

1.3.1 *Classical statistical view of clustering*

Events or objects which have common characteristics are generally considered to form a group. The term cluster has a similar meaning to group. Events, which are concentrated near a specific location (site) are called a point cluster. Events concentrated in a sub-area are called an area-cluster.

Defining the term cluster is a difficult task. Everitt (1974) gives several definitions of a cluster (cited in Jain and Dubes (1988), page1). They are:

1. "A cluster is a set of entities which are alike, and entities from different clusters are not alike".
2. " A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it."
3. " Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points."

The "points" mentioned in the second and the third definitions above refer to the locations of events. The word "events" was used instead of the word "points" because there are an infinite number of points in a space but only a finite number of accident events.

1.3.2 *Main features of spatial analysis techniques*

Three spatial analysis techniques (quadrat analysis, cluster analysis and nearest-neighbour analysis) are used in this research to detect and identify the three basic spatial distributions of accidents (i.e. point cluster, line cluster and area cluster distributions).

The main features of quadrat analysis are:

1. The study area is divided regularly or randomly into sub-areas, which are consistent in size and shape.
2. The events within each of the sub-areas are counted.
3. The distribution of counts are analysed. The counts are Poisson distributed if accidents are completely spatially random.

The main features of cluster analysis are:

1. The events within the given area are arranged together in groups, so that the events within a group are relatively similar.
2. The similarity of events is based on a function of the distances between them.
3. There are various distance functions, which lead to different groupings of events.
4. The groups of events can be considered event clusters.

The main features of nearest-neighbour analysis are:

1. An event is selected within the given area.
2. The distances between the selected event and its neighbouring events are determined.
3. The directions from the selected event to the neighbouring events are determined.
4. The distribution of distances and directions to nearest neighbours are analysed.

1.3.3 *Application of the techniques to accident data analysis*

There are limitations on the spatial statistical analysis of accident data, for example the number of accident data available from some areas may not be sufficient for statistical reliability and the selected time period of the data may not be sufficient. For example, if the selected time period is very short then there is a danger of random fluctuations (such as a high number of accidents at certain sites for a short period followed by a low number of accidents in the next short period) affecting the results. If analysing specific accident types

(e.g. pedestrian accidents), the numbers of accidents might be too small to allow meaningful analysis, unless a very long time period is used. In other cases, traffic management treatment may influence the traffic for a short period and accidents might increase for a short period. The effects of temporal variations need to be considered when analysing spatial distributions. This is discussed in detail in Chapter 2.

The spatial accident analysis involves the application of spatial statistical techniques to identify any spatial pattern in the locations of the accidents. It does not involve going into the details of the factors that lead to the accidents at sites or along routes or particular areas. Once the accident patterns are identified from the spatial accident analysis, the next step would be to consider the factors that lead to the occurrence of those accident clusters, and then formulate and implement accident reduction initiatives.

Identifying the spatial pattern could be helpful in identifying the common factor or factors related to the road geometry or the traffic management issues. For example, we can select the accidents involved in the route cluster and identify the common factors involved with those accidents. Identifying common factors is helpful in selecting the accident reduction plan as briefly discussed above.

Several factors are generally involved in an accident (e.g. factors related to the driver, the environment, the vehicle). If the factor or factors relating to an accident are not strongly dependent on the characteristic of that site then it is likely to be a random accident (e.g. an animal crossing a high speed road and colliding with a vehicle is unlikely to occur again at the same place). On the other hand, if a factor or factors related to some accidents are strongly dependent on the characteristic of the site or the environment of the site (e.g. animals frequently crossing over a short length of a road) then there may well be a number of accidents in the accident record and that site may be considered a point cluster. Similarly, if the factors are strongly dependent on the characteristic of a road (e.g. animals frequently crossing over a long length of a high speed road) then there may well be a high number of accidents spread along the road, which may be considered a line cluster. If the accidents are spread throughout the road network within an area, and the factors are strongly dependent on the characteristics of that network but not on individual sites or routes, then that tends to be an area cluster.

The object of this research is to develop a method for analysing accident distributions, to identify whether there is a pattern and what sort of pattern it is. If there is no pattern in the distribution then the analysis needs to indicate that. Suppose the null hypothesis H_0 is that there is no pattern in the distribution and the distribution is CSR. There is a possibility the result of the statistical analyses will lead to one of two types of error:

- Type I; reject H_0 when H_0 is true;
- Type II; do not reject H_0 when H_0 is not true.

If the method used for identifying the spatial distribution is a powerful test then the result has low probability of type II error. Therefore the statistical analysis method must be carefully designed to work as a powerful test.

The purpose of using spatial statistical analysis techniques is to identify any spatial pattern (i.e. whether there are site, route or area clusters) and to identify the nature of the pattern. In Chapters 4, 5 and 6, three techniques used to identify the hazardous locations are described. The application of the spatial analysis techniques for identifying point, line and area clusters is briefly described.

1.3.4 *Point clusters*

Accidents are often clustered at intersections because of the conflicting traffic movements within a limited road space, but there are accident clusters on links between intersections too. One of the reasons for point clusters is that some sites become dangerous when the environmental conditions (such as traffic, weather and road conditions) change. The number of accidents recorded at a site during a period may indicate how frequently the site becomes dangerous.

If the number of accidents at a site is higher than some threshold value then the site is commonly identified as a hazardous site. This occurs because the factor or factors related to the accidents occurring at the site are strongly related to the characteristics of the site.

The likelihood of accidents occurring at any site may be related to the accident counts at that site (i.e. the calculated probability at that site). The variation in the accident counts at

could indicate the variation in the probability of occurrence of another accident at the sites. The accident counts can be analysed using quadrat analysis, as explained in Chapter 6. The distance from a selected accident position to the nearest neighbour accident positions will also be used, to test whether the selected accident is within a cluster of accidents. If a high proportion of accident positions appear to be clustered at sites, then the dominant pattern is a point cluster. This idea was first applied to accident analysis by Nicholson [1998]. The merit of this method is in testing the immediate neighbourhood to see whether there are other accidents clustered in the vicinity. Cluster analysis can also be used for analysing the distances between accident positions and could be used to investigate the level of point clustering, as described in Chapter 4.

1.3.5 *Line clusters*

A line cluster means the accidents are spread along a long section of road. The reason might be recurring factors that greatly reduce the level of safety for the traffic using the section of road. For example, an accident can occur if the view of passing vehicles is blocked to pedestrians by parked vehicles, or if a parked vehicle exits from the parking space and possibly collides with a passing vehicle. Here, the common factor is the involvement of parked vehicles. If any accident factors recur throughout the route then a combination of single site and mass action plans may be applied. The identification of recurring factors for accidents is an important part of this combination approach. Identifying the recurring factors are somewhat difficult without identifying which road or section of road is a line cluster. Therefore, if we identify a line cluster then it may be easier to identify the recurring factors related to the accidents, which are aggregated to form a line cluster.

The number of accidents, which are higher than normal on that type or class of road or section of road, are a general indication of the line cluster. These roads are commonly called hazardous routes. The accidents are spread throughout such roads or road sections and it is necessary to identify them as a line cluster. The route might have an equal number of accidents at sites or randomly distributed. The randomness of accidents along the route could be tested by quadrat analysis, nearest neighbour or cluster analysis.

The level of line clustering can be estimated by the number of accidents per unit length. The unit length of road could be used as quadrats and the quadrat counts could be analysed using quadrat analysis. If the maximum count for a road section is above the average then that is generally considered a hazardous road section. The alternative method is to test the directions to the nearest neighbour events. If the directions to nearest neighbour events are distributed in a non-uniform manner, then the accidents could be along routes. Nicholson [1998] tested the nearest neighbour events to see whether the directions from randomly selected events positions are distributed non-uniformly. If a high proportion of accidents have a non-uniform distribution of directions to their nearest neighbours, then the dominant pattern could be a line cluster. Cluster analysis can be used for analysing the distance between accident locations and for investigating line clusters, as described in Chapter 4.

1.3.6 *Area clusters*

Traffic volumes in networks will tend to increase in the longer term and traffic safety may be threatened, if there is no upgrading works or other measures (e.g. education or enforcement campaign). Therefore urban and suburban road networks should be assessed, to establish whether the road network is safe with increasing travel demand. This could be a part of a monitoring process to maintain safety objectives and the need to encourage the traffic to use each part of the road network safely.

In some areas, traffic accidents occur all over the road network but are not concentrated at particular sites or along certain roads. The accidents are scattered sparsely throughout the road network or are concentrated fairly evenly at sites. This type of accident occurrence can be identified by monitoring the spatial distribution of accidents.

In general, establishment of a road hierarchy, on the basis of the movement and access functions, leads to an increased clustering of accidents (or reduction in the scatter). This is becauseroads high in the hierarchy (e.g. arterials) have agreater facility for movement and reduced access facility compared to the other roads. This leads to an increase in the clustering of traffic on arterial roads and the associated reduction of traffic on other roads. The increased traffic flows are likely to lead to an increase in the clustering of accidents (which are related to flows) on the arterial roads..

The identification of increasing accident dispersion at an early stage will assist the initiation of appropriate accident reduction measures when maintenance or minor improvement work are being carried out. IHT (1990) noted "considerable benefits can be obtained from a slight, inexpensive change or addition to a maintenance scheme to incorporate accident reduction measures".

One can assess the level of clustering by analysing the proportions of locations with accident counts of 0, 1, 2, etc. If the accidents are randomly distributed between the locations, then the distribution of accident counts will be approximately Poisson (Cressie, 1993). Hence, as the overall intensity of accident occurrence increases (i.e. the mean accident count increases), the variance will also increase, so that the mean and variance are approximately equal. This approach does not, however, involve analysing the spatial relationships between accidents (i.e., one does not know whether the high count locations are close to other high count locations or not).

The spatial relationship between accidents can be investigated using quadrat analysis, the nearest-neighbour method or cluster analysis. The theory and application of these three methods are described in Chapter 4, 5 and 6.

1.3.7 *Types of accident patterns*

The basic types of clusters are point cluster, line cluster and area cluster. Point clusters (see Figure 1.08) have small characteristic lengths compared to line clusters (see Figure 1.09) or area clusters (see Figure 1.10). A line cluster has a long characteristic length and accidents can be arranged in various ways within that length. The various ways are:

1. accidents are random along a line;
2. accidents are regular along a line;
3. accidents are clustered at points along a line.

If the accidents are strongly clustered at points along a line then they might appear to be point clusters.

An area clusters has two long characteristic lengths and accidents in an area cluster can also be arranged in various ways;

1. accidents are random within the cluster-area;
2. accidents are regular within the cluster-area;
3. accidents are clustered at points within the cluster-area;
4. accidents are clustered along lines within the cluster-area;
5. accidents are clustered in sub-areas within the cluster-area.

If accidents are clustered in sub-areas within the cluster-area, each sub-area can be investigated to determine whether the accidents are random, regular or clustered at points or along lines or within parts of the sub-area.

If the accidents are strongly clustered at points then they might appear to be point clusters. If the accidents are strongly clustered along lines then they might be appear to be line clusters. If the accidents are strongly clustered within sub-areas then they might be appear to be area clusters.

As shown in Figure 1.11,

1. the appearance of a pattern depends upon the scale at which it is examined,
2. a cluster which appears to be a line cluster at one scale might look like a number of point clusters, if examined at a more detailed scale and
3. a cluster which appears to be an area cluster at one scale might comprise a number of point clusters or line cluster or area clusters, if examined at a more detailed scale.

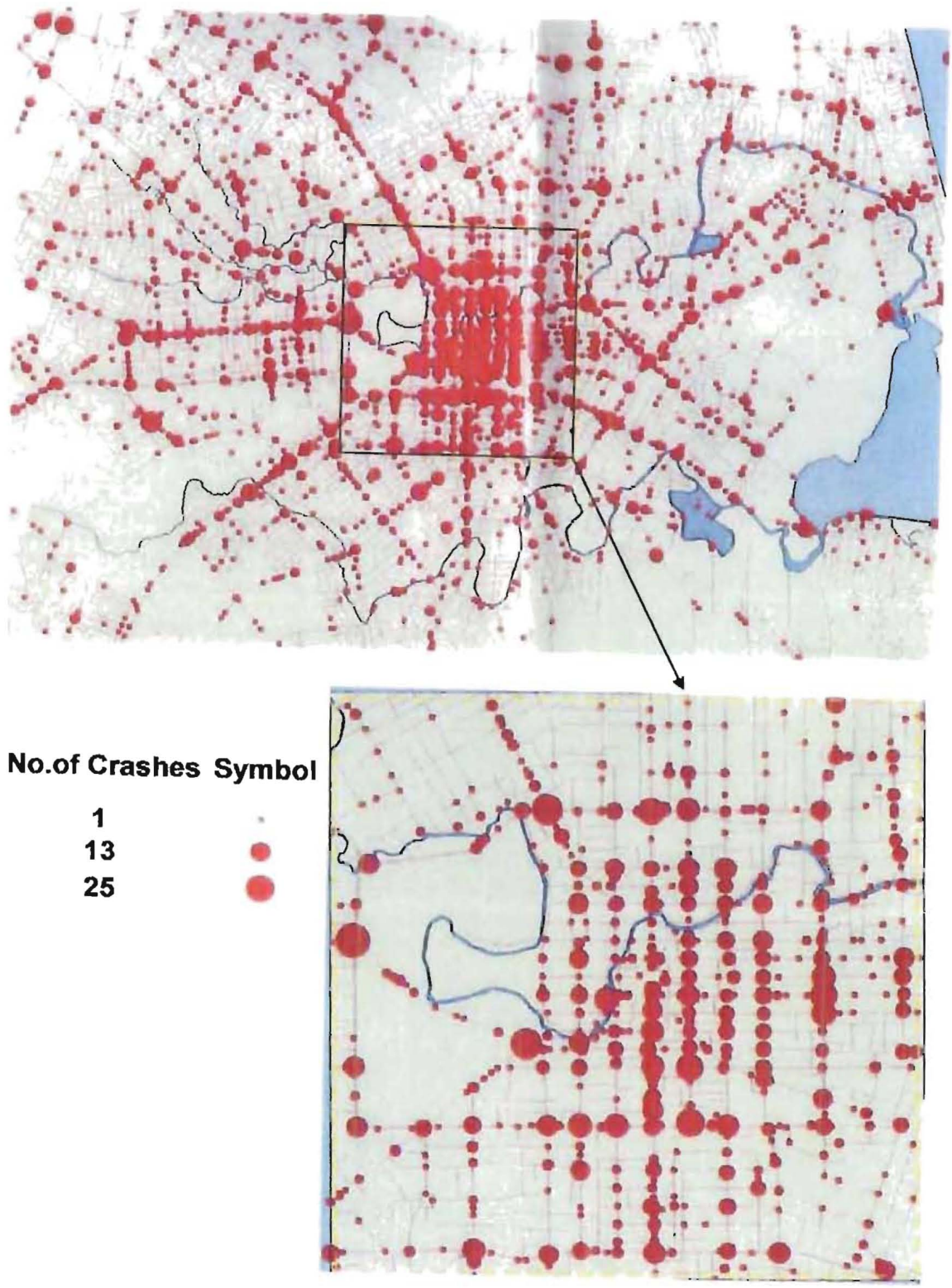


Figure 1.01: Injury accidents in Christchurch (1995-1999)

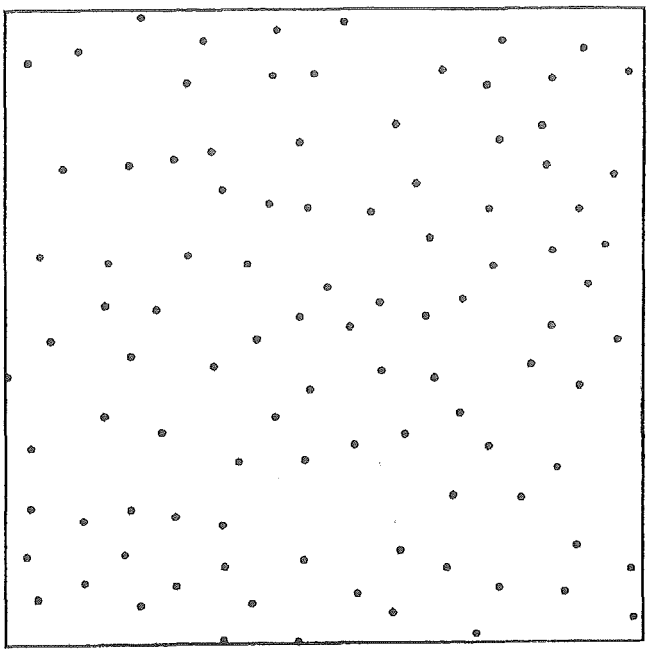


Figure 1.02: A regular distribution

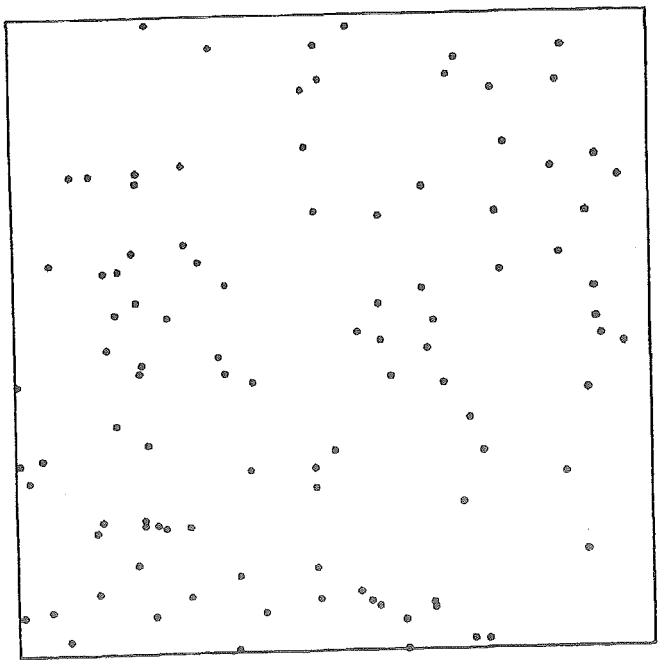


Figure 1.03: Complete spatial randomness.

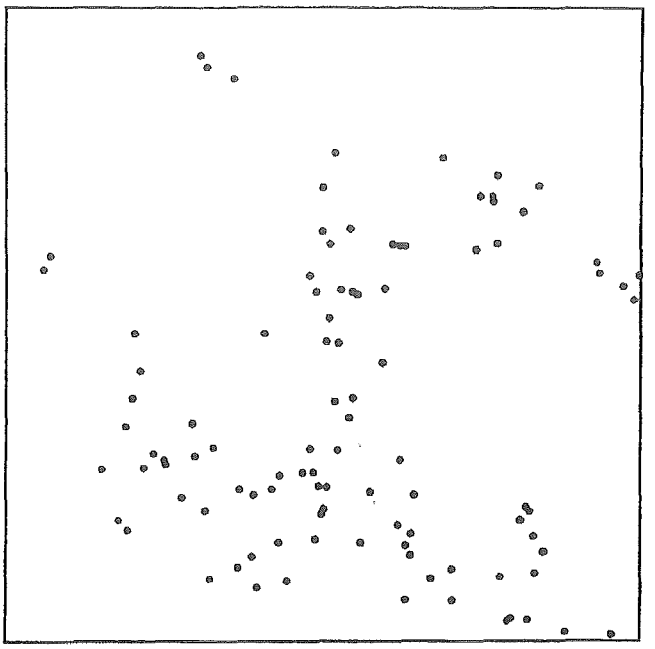


Figure 1.04: A cluster distribution

[Figures 1.02, 1.03 and 1.04 extracted from Cressie 1993]

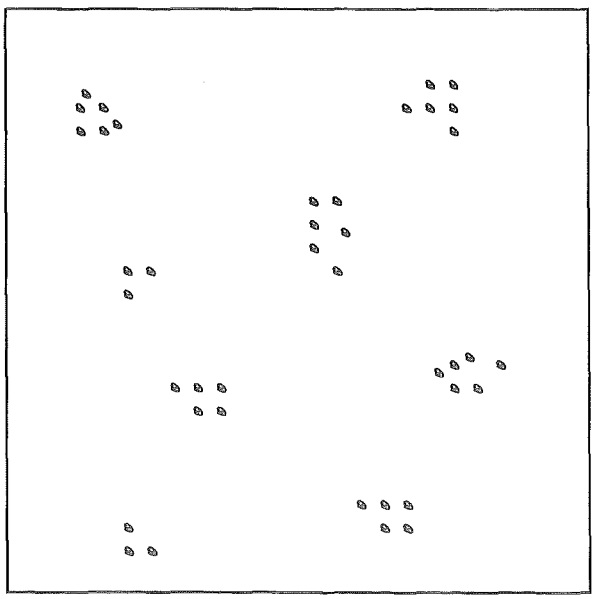


Figure 1.05: Black spots (point cluster)

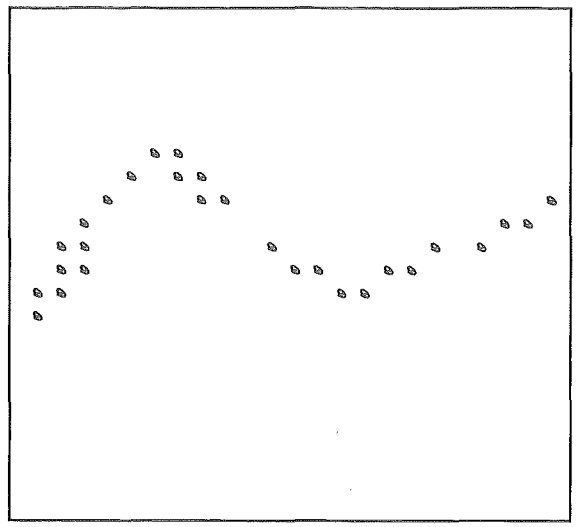


Figure 1.06: Black route (line cluster)

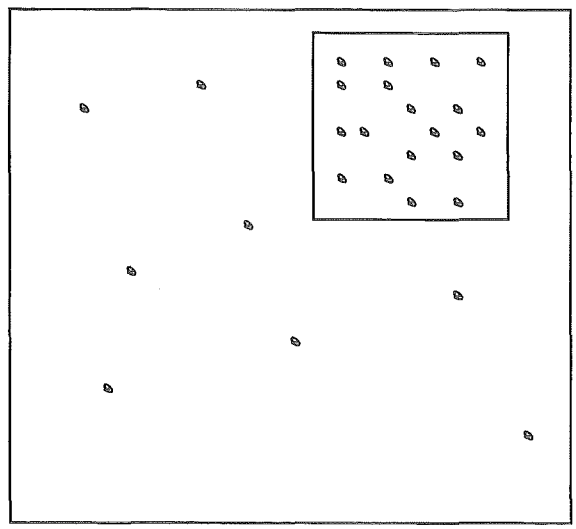
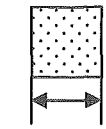


Figure 1.07: Black area (area cluster)



Short length

Figure 1.08: A point cluster

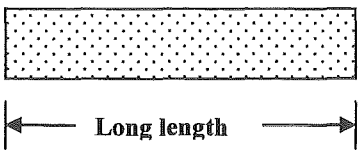


Figure 1.09: A line cluster

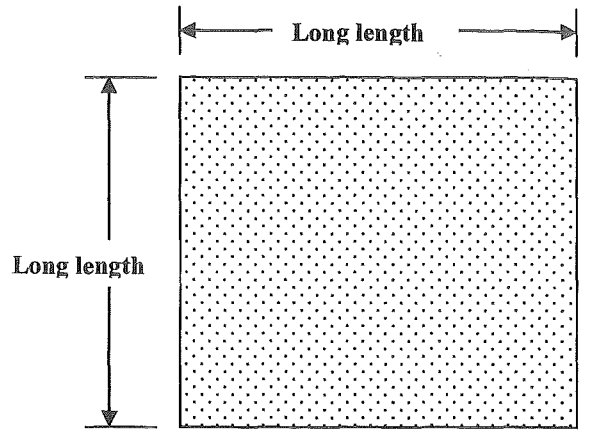


Figure 1.10: An area cluster

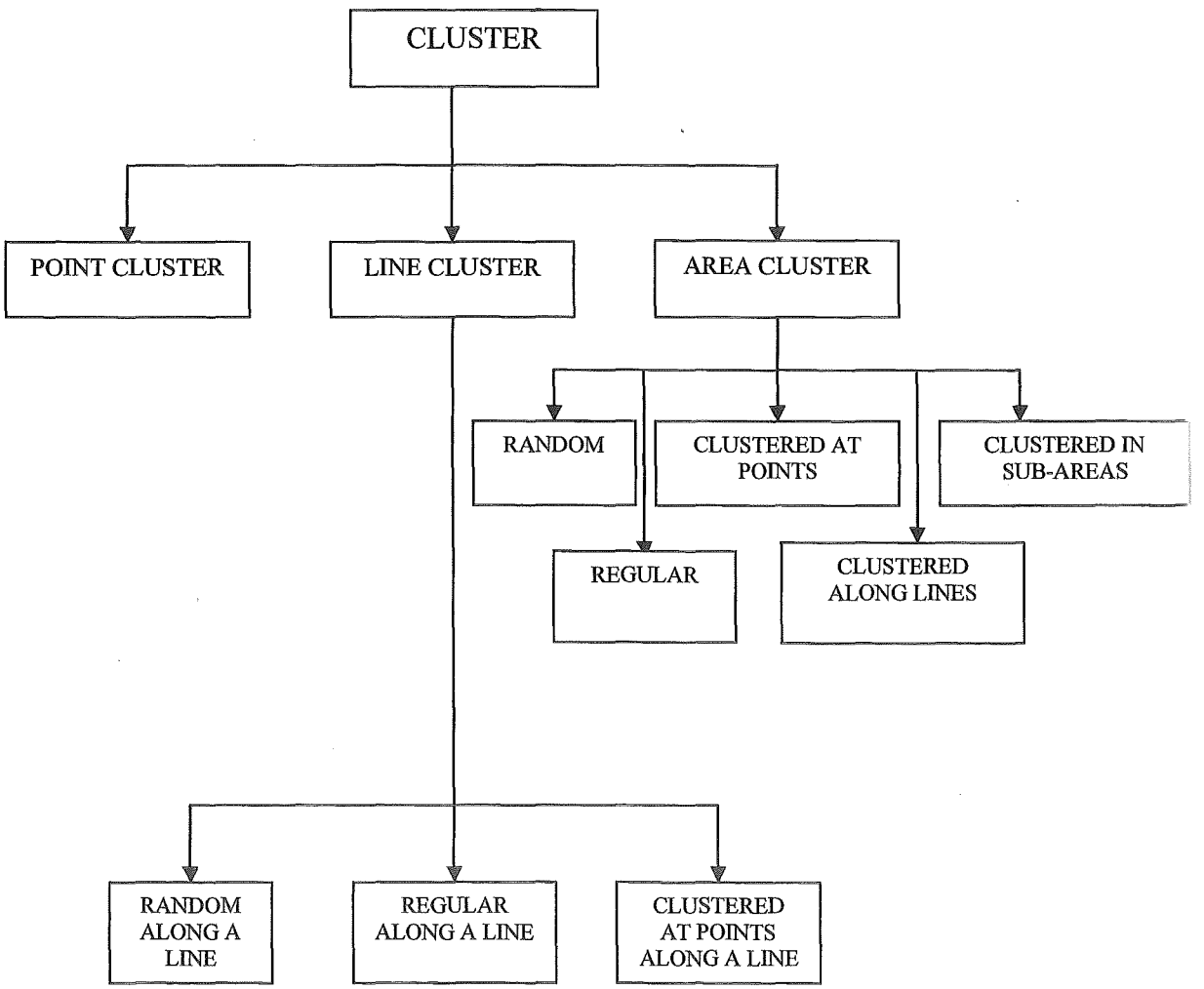


Figure 1.11: Cluster patterns

Chapter 2

CLUSTERING FACTORS AND ANALYSIS ISSUES

2.1 Factors in accident clustering

Accident clusters indicate that some factors related to accidents are dependent on the characteristic of the sites or route or area. If a site appears to be an accident cluster, then there could be single or multiple factors common to all the accidents which occurred at that site. To understand why the accidents might be clustered at certain locations (sites, routes and areas), it is necessary to understand the factors contributing to the accidents, and how these factors are dependent on the characteristics of the locations. This information is useful to understand why the accidents occur in a random, regular or cluster pattern; why the clusters occur as either point, line or area clusters; and why there is variation in the level of clustering over time. A good understanding of the causes of the accident patterns is critical to ensuring that any analysis of the spatial distribution of the accidents is complete and enables proper interpretation of the results.

The first question that comes to mind would be why do accidents occur. Drivers are often blamed for not being alert, not making a proper judgement or not responding to complex situations arising from traffic or road conditions. Generally, accidents are not intentional. If the driver becomes confused, or the demand upon the driver becomes overwhelming and the driver cannot cope with the situation, then an accident may result. There are several factors that lead to accidents and knowledge of these can be useful in understanding the spatial patterns of accidents. These factors can be divided into three groups; vehicle, road environment and driver. The demand upon drivers is determined by the vehicle and environmental factors.

Accidents can occur from the interaction of combinations of vehicle, environment and driver factors. This is shown in Figure 2.01. Driver behaviour is derived from the interaction of human factors (i.e. physiological and psychological factors), the environment and the vehicle. Drivers adjust their behaviour according to the characteristics of their vehicle and the environment.

The focus of this research is on the spatial distribution of accidents, and the road environment (including the road network characteristics) can strongly influence the spatial distributions of accidents. Therefore this chapter focuses on the road environment, as well as its interactions with the vehicle and driver factors.

2.1.1 *Vehicle factors*

There are various types of vehicles on the road. Different vehicles impose different levels of demand upon drivers (i.e. the demand depends on vehicle factors). The important vehicle factors are:

- warning and instrument systems layout (e.g. distraction of driver through monitoring vehicle instruments);
- brakes (e.g. braking system limits the deceleration capability and non-symmetric wheel locking causes loss of control) and tyre characteristics;
- stability (e.g. vehicle height and wheel track);
- size, weight and power (e.g. restriction on acceleration or deceleration).
- visibility limitation of the vehicle design (e.g. restriction of the driver's field of view);
- vehicle lighting (e.g. restriction on illuminated field of view);

These factors can cause difficulties when driving and may lead to accidents. For instance, vehicle pillars may restrict the driver's view of pedestrians or cyclists and may increase the level of accident clustering in those places with frequent pedestrian or cyclist activity.

2.1.2 *Driver factors*

The demand upon drivers depends on flow variables and non-flow variables. Flow related demand upon drivers can vary with time at a place. Some non-flow related demands (such as road geometry) do not change with time but others may vary with time, due to changes in weather conditions (e.g. ice or rain) or visibility conditions (e.g. overcast and dark). The overall demand on drivers varies at different times and at different places, because of changes in the road environment.

The driver performance and the environmental demand are not constant throughout a journey, as illustrated in Figure 2.02, in which one notes a substantial reduction in environmental demand at "B" followed by a sudden increase but the response of the driver was not sufficiently adequate and/or quick. Accidents might occur at "C" where the driver performance curve meets the environmental demand curve. Good drivers will meet the demand by timely changes in the level of alertness with good strength of response, but sometimes driver tiredness or a sudden increase in demand may prevent this.

Environmental demands are not spatially constant over a road network. Generally, the place where the demand is high is a potential hazardous location (i.e. black spot or route or area). Although accidents do not always occur in places where the demand is high, because the driver (and vehicle) can sometimes meet the demand, the frequency of accidents could be high compared to places where the demand is low. The high-demand locations might be the location of accident clusters.

Environmental demands also vary over time. For example, in the case of an intersection which is only busy during peak periods, the demands upon drivers is only high for these short periods. In this case the opportunity for accidents could be very low throughout the day compared to an intersection that is busy most of the day. In such cases the accident frequencies are low, even though the demand upon drivers is high for a short time during the day, and these places are not identified as hazardous locations.

The demand upon drivers cannot be easily quantified but the number of accidents can. Chapman [1973] noted that the exposure is the "number of opportunities for accidents of a certain type in a given time in a given area (i.e. it is the possible number of accidents of that type which could occur in that time in that area)". Chapman also suggested that $A = N \times p$, where A is the expected number of accidents during a time period, N is the number of accident opportunities during that time and p is the conditional probability of an accident occurring (given an opportunity occurs). This relationship indicates that A will be high if N (i.e. number of accident opportunities) is high and/or p (i.e. probability of an accident is high if an exposure occurs) is high. Exposure can be related to various explanatory variables (e.g. traffic flow, travel time, travel distance). The form of the relationships between exposure and flow differs for links and intersections.

2.1.3 *Road environment factors*

The four major components of the road environment related to accidents are:

1) the traffic stream;
2) the road design;
3) the land use adjacent to the road (see Section 2.1.5) and
4) the climatic conditions (see Section 2.2.5).

These environmental factors place a demand on drivers and the drivers place a demand on their vehicles to avoid an accident. Accidents occur when the driver or the vehicle are unable to meet these demands. Each of the components must be looked at in more detail to understand the relationship between accidents and environmental demands.

Traffic stream

The traffic stream involves three main components:

a) flow rate;
b) flow composition;
c) traffic speed.

These components impose demands on drivers (e.g. the probability of an accident might increase as the traffic speed increases).

Flow rate

The relationship between accidents and traffic flow is discussed in detail in Sections 2.1.4 and 2.1.5. In general the frequency of accidents is related to the number of vehicles on the road.

Flow composition

Having different types of vehicles on roads can cause several difficulties for drivers, for example:

- trucks can block the view of another vehicle;
- trucks have lower acceleration and deceleration rates;
- heavy vehicles often travel at lower speeds;
- trucks with trailers can be difficult to overtake.

That is, the vehicle composition on the roads causes a variety of driving difficulties and places demands on drivers. Heavy vehicles sometimes have a different speed limit, which may

increase the occurrence of accidents [Ogden, 1996]. In some places where frequent overtaking occurs, there may be sight distance restrictions which may cause accidents to cluster.

Traffic Speed

The relationship between traffic speed and the occurrence and severity of accidents is clearly established and widely recognized. Higher traffic speeds can lead to an increased frequency and severity of accidents due to loss of control, head-on collisions and accidents involving road side hazards such as trees.

Evans [1985] suggested that speed affects drivers and hence the occurrence of accidents in several ways, including:

- "we have to focus attention much further away";
- "information density is much higher";
- "the variation of speeds between road users is much larger";
- "it is very hard to predict potential danger points with several high speed vehicles";
- "the necessary manoeuvre-time and distances are proportional to the square of the speed";
- "a mistake is more difficult to correct and will have more serious consequences".

Taylor [2000] noted the "reduction in deaths and injuries that is achievable from reductions in speed varies according to the type of road and the average traffic speed" (see Figure 2.03). Taylor, in an empirical study, related the accident frequency to just the mean speed and found that the accident frequency depended upon the mean speed raised to the power of 1.536.

The level and nature of clustering in a network will be somewhat related to the variations in mean speeds within the network. Garber and Gadirau [1988] found that there is a relationship between accidents and speed variance, and accidents may be clustered at places where the speed-variance or mean-speed is high.

Road Design

The American Association of State Highway Officials [1954] mentioned "...any highway feature which happens to be substantially below the standard prevalent on a given highway, introduces a surprise element with resultant higher accident occurrence". Jorgensen and Associates [1978] mentioned about 50 design features which are related to accidents. These

include the characteristics of the carriageway, vertical alignment, horizontal alignment, auxiliary lanes, median, roadside, traffic control and streetlights. These characteristics directly or indirectly influence accident occurrence over different parts of the road network, which may lead to accident clustering at various locations.

Carriageway characteristics

Some design features (e.g. inconsistency of geometric standards) may lead to point, line or area clusters. Road design features directly affect the safe operating speed. If there is a sudden change in safe operating speed for a short distance, then that may lead to frequent accident occurrences in that short length. In this situation driver behaviour is affected by driver expectancy. Krammes [1995] noted that drivers tend to " react to what they expect rather than to the roadway or traffic situation as it actually exists".

Alignment

Charlesworth and Coburn [1957] and the Road Research Laboratory [1965] noted that there was a distinct tendency for accidents to cluster on very sharp curves and accident rates decreased as the average curvature (degrees per unit distance) increased. Mullins [1961] found accidents were concentrated at vertical curves. These accidents could be due to loss of control or sight distance restrictions (i.e. reduction in the length of carriageway visible to a driver). Hedman [1990] noted that the accident rate increases with decreasing sight distance, especially for a single vehicle at night, so low sight distance at a location may lead to a high accident frequency at that location.

Glennon [1987] concluded that grade sections have higher accident rates than level sections and steeper grades have a higher accident rate than milder grades. Paniati and Council [1991] found that most of the accidents are concentrated within about 30m from the summit, especially when the change in grade at the vertical curve exceeds 6%. These types of summits are point clusters.

The effect on safety of combinations of geometric features can be greater than the sum of the effects of the individual features [Kihlberg and Tharp 1968]. Wright and Robertson [1976] and McBean [1982] noted that sites with a combination of downhill gradient and curvature have more accidents than sites with the individual features. That is, the level of clustering may be high where a site has a combination of design features.

Auxiliary lanes

Where there is a long up-hill road section, heavy vehicles will slow down while climbing. In the absence of a climbing lane, there are likely to be risky overtaking manoeuvres leading to accidents. The section of road may therefore turn out to be a line cluster of accidents. Hoban [1982] found that there was a 25% reduction in accidents on rural roads where passing lanes were installed. It appears that if a rural road section is provided with passing lanes, then the number of accidents on that road and the level of line clustering may be reduced. However, if the passing lane is not terminated properly, then it may result in accidents where the passing lane merges with the main lanes, leading to a high level of point clustering.

Median

Median barriers may reduce head-on collisions, but may increase same-direction accidents. Instead of crossing centreline and colliding with the vehicles travelling in the opposite direction, the same vehicle may hit the median barrier and collide with following vehicles. If we analyse some specific accident types (e.g. head-on collisions or same direction collisions) then median barriers may change the spatial pattern of that types of accidents.

Roadside

The nature of the roadside is important when a vehicle leaves the roadway and needs to be brought back under control. This is made worse by roadside hazards (e.g. poles, trees, drainage inlets/outlets or deep ditches). Any of these types of roadside hazard can contribute to line clusters or point clusters. Guardrails may be used to deflect vehicles away from the roadside hazard, and the LTSA [2002] reported a 45% reduction in fatal and serious injury crashes where flexible guardrails were installed. McLean [1996] noted that guardrails are sometimes installed in situations where the accident rate with the guardrail is higher than the accident rate with an unprotected roadside. That is, a location with a guardrail may sometimes be a point or line cluster.

Traffic control

Some pedestrian facilities (e.g. zebra crossing, central refuges, pedestrian crossing) contribute to accident clustering. The presence of pedestrian facilities is associated with the land use adjacent to the road.

Summersgill and Layfield [1996] found that nose–to-tail and lane-changing vehicle accidents are more frequent on urban road links with zebra crossings and single-vehicle accidents involving vehicles hitting refuges increase when the frequency of central refuges increases. That is, the level of clustering may increase when the frequency of central refuges increases. Summersgill et al. [1996] noted that there were more accidents at three-arm priority junctions on urban single-carriageway roads if there was a pedestrian crossing. It appears that accidents tend to be clustered at the locations of such facilities.

Traffic calming activities (e.g. street closures and speed reducing devices) at a site or route or an area may affect the level of point, line and area clustering. Traffic management schemes within a specific area may affect the level of area clustering. Traffic management methods typically involve establishing a hierarchical road network, by removing through traffic (e.g. by street closures and speed reducing devices) from minor roads (e.g. residential streets or roads having an access function only), and improving the main road to accommodate the diverted traffic without additional delays.

Streetlights

Street lighting may also affect accident clustering. Street lighting is useful to identify the objects or pedestrians on roads but some lighting arrangements can cause drivers to misjudge and run off roads. This may lead to accidents occurring in clusters at those locations where the lighting arrangement is deficient.

Land use adjacent to the road

Brindle [1993] noted that problems arise from the conflict between the movement and access functions. The movement function involves long distance travel normally at a higher speed, while the access function involves people leaving or entering properties and parking alongside the road. The access function is dependent on the land use adjacent to the road. If both movement and access functions happen along the same stretch of road, then the frequency of accidents could be high and the road may be a line cluster. If there are only one or two points of access to the road (e.g. access to off-street parking areas) then there could be point clusters. If there are frequent access points to the network in an area (e.g. access to premises in the central business districts) then there could be an area cluster. If there is a car park then there will be traffic moving in and out of the car park. There might be kerb-side parking along the road.

The land use adjacent to the road will also influence the pedestrian activity. For instance, high pedestrian activity can be expected if there is a supermarket or shops on both sides of a road.

2.1.4 *Accident clustering at intersections*

Common types of accidents at intersections include crossing, turning, merging and rear-end accidents. Over 25 % of the accidents in New Zealand occur at intersections [LTSA, March 2000], and 39% of urban crashes were at intersections. The conflicting flows at intersections make demands upon drivers. Different layouts mean different exposure because the number of conflict points can differ between layouts.

High exposure plays an important role in accident frequency at intersections, and Chapman [1973] showed that the exposure depends upon the product of the intersecting flows. However, Turner and Nicholson [1996], Maycock and Hall [1984], Hakkert and Mahalel [1978] and Tanner [1953] found that the number of accidents is very closely related to the square root of the product of the conflicting flows. This means that the accident rate (per exposure) decreases as the flow increases. Tanner suggested that driver attention increases as the demand (i.e. level of conflict) increases.

Navin et al. [2000] noted that rear-end accidents represent approximately 21% of all accidents in British Columbia and 36% of all accidents in Vancouver. They noted that rear-end accidents are most common at urban signalised intersections. For example, if a vehicle in a traffic stream suddenly stops because of a signal change, the vehicle following may collide with it. That is, the presence of signals may affect the level of clustering at intersections.

2.1.5 *Accident clustering along links*

Common link accident types include single-vehicle accidents, rear-end accidents and head-on accidents. Single vehicle accidents depend on the number of vehicles travelling within the unit length of road, and a commonly used exposure measure is vehicle-kilometres of travel (VKT). The number of vehicles involved in an accident depends on the number of vehicles present in

the vicinity. However, if the number of vehicles in the vicinity increases then the driver's attention may increase and the accident rate (per vehicle) may reduce. Hauer [1995] noted that the relationship between the number of accidents and the traffic flow is non-linear, and Mountain et al. [1996] studied six carriageway types and found that accidents on highway links are not proportional to link exposure (VKT) as commonly assumed.

Chapman [1973] discussed how accident exposure is related to link accidents. Exposure can be related to various explanatory variables (eg. flow rate, travel time, travel distance), and Transfund New Zealand (2000) noted that accident occurrence is related to the flow rate. If traffic flows are clustered on particular roads then accidents will tend to be clustered on those roads.

Silcock and Worsey [1982] studied the relationship between accidents and flow rate, and vehicle-kilometres of travel (i.e. a measure of exposure). In this study the roads were categorized according to the adjoining land use and the carriageway type. By doing this, stronger relationships between accidents and flow rates were achieved. This suggests that the adjacent land use may influence the spatial distribution of accidents.

In general, accident frequency is related to the geometric standard of roads. Ogden [1996] noted that a higher geometric standard is designed to facilitate "high design speed, full control of access from abutting property, forgiving roadsides, entry and exit at grade-separated interchanges, and opposing directions of traffic separated by a median". Motorways or freeways are examples of the highest geometric standard and are much safer per vehicle-kilometre of travel than other roads.

Olmstead [2001] found that a freeway management system reduces property damage, minor injury accidents, rear end accidents and sideswipe accidents but it does not reduce major injury, fatal accidents and single vehicle accidents. If we analyse the major injury and fatal accident data only for a road network, which include freeways, arterials and distributor roads, then the results may therefore indicate the freeways as line clusters.

In general vehicle speeds are related to the horizontal and vertical alignments of the road. Advisory speed signs are frequently located along rural roads or main highways on the approaches to selected curves, and drivers need to vary speed to negotiate the curves safely.

Accidents caused by the failure to vary speed are more likely to occur at highway curves than straight road sections. The accident rate is related to the sight distance on two-lane rural roads, as shown in Figure 2.04, and poor sight distance is associated with vertical or horizontal curves, which may be point clusters.

2.2 Issues in the analysis of clustering

In this thesis we consider the spatial distribution of traffic accidents only on the road network and not within private property or places where there are no roads. If we consider accident locations throughout an area, then the result may be biased towards a cluster pattern because a large part of the area may not have vehicular traffic and may hence have no accidents. Areas where there is no vehicular traffic must be omitted in spatial data analysis and a method for doing this is explained in Chapter 6. Other important issues related to the analysis are discussed in the remainder of this chapter.

2.2.1 *Identification of accident clusters*

This thesis analyses spatial distributions, to identify spatial clustering patterns that may exist, by investigating the specific locations of accidents. The identification of clusters need not entail considering the detail of each and every accident. Searching accident patterns and identifying clusters will help identify the underlying problem. If we consider all accidents rather than clusters, then identifying any underlying common problem is more difficult, because there are several general factors involved in each accident.

There are different criteria for identifying each of the accident cluster types (point, line and area clusters). The criteria for identifying clusters, cost effective treatment, the characteristic length of each cluster, and the period for which data are analysed, are important issues. They need to be understood before identifying clusters, and are useful when analysing the spatial distribution of accidents, in order to select a cost effective accident reduction plan.

2.2.2 *Criteria for identifying clusters types*

Point clusters

The " potential accident reduction" method [McGuigan 1981 and 1982] used for identifying black spots, is based on the difference between the expected number of accidents and the actual number of accidents at locations. For example, if the expected number of accidents is higher than the observed number, then priority is not given to treating that location. The potential accident reduction method can be used to rank the locations. The expected number of accidents per year for each location can be estimated using accident-flow relationships such as those outlined in Sections 2.1.4 and 2.1.5. The accident-flow relationship has some uncertainty. Therefore, the ranking of locations might not be accurate because the expected number of accidents may not be estimated accurately for individual locations. Maher and Mountain [1988] concluded that the potential accident reduction method is not as good as simply using the number of accidents, so the potential accident reduction method was not used in this thesis.

The two basic measures for identifying black spots are the number of accidents, and the number of accidents per exposure. If one or both of these measures at a site exceed selected values or thresholds, then that site may be identified as a black spot. It was argued [IHT 1987, DTp 1986] that the first measure is biased towards large flow rate accident locations (for example, locations with a high traffic flow rate might have a high number of accidents) and the second measure is biased towards small flow rate locations (for example, accident locations with a low traffic flow rate might have a high accident rate). Therefore the combined measure (i.e. the number and rate of accidents) has been recommended for identifying not only black spots but also black routes and black areas.

The observed number and rate of accidents before the treatment and the estimated values after the treatment are useful in evaluating the remedial treatment (i.e. deciding whether the remedial treatment reduced the accidents at a treated hazardous location and whether it produced an acceptable economic rate of return). The flow rate could change after the remedial treatment and this might cause an accident reduction. Therefore, considering the number of accidents only may not correctly reflect the effect of the remedial treatment. So, the combined method is better than using only one of the two single measures for hazardous location identification and the evaluation of a remedial treatment.

It is easier to use the number of accidents than the rate of accidents to identify hazardous locations, since the exposure details required for the latter are often not readily available because they are not regularly recorded or updated. Exposure details and accident rate are calculated using traffic flow. Traffic volumes are not readily available for the majority of low volume local roads and rural roads (excluding state highways). Data that is available (in RAMM) may only be an estimate, and not from an actual count. Use of such data may lead to erroneous conclusions. For this reason the number of accidents was used instead of accident rate for point cluster identification.

Route clusters

Suppose there are many random accidents throughout a road or a long section of that road; this could be considered a black route (line cluster). It is sensible to identify black routes using the following two criteria:

- number of accidents per unit length (generally in kilometres);
- number of accidents per vehicle kilometre of travel (VKT).

Ideally both criteria should be used, because each of the criteria is biased towards either low or high traffic flow, as discussed above. However, since accurate traffic volume data are not always available, the number of accidents per unit length of road was used for route cluster identification.

Urban or rural roads can be separately categorised into groups. For example, a simple two-category system is:

1. main roads (volume and speed are generally high);
2. minor roads (volume and/or speed are generally low).

Each of these groups can be analysed separately using the number of accidents. The emphasis here is the indirect use of traffic volume and speed. While this may appear to be a simplistic approach, it might be useful in the absence of traffic volume data.

Area clusters

IHT [1999] identifies black areas by analysing the accident distribution throughout an area, which "may be determined within routes forming the main road network, by administrative boundaries or, for example by selecting 1km squares". Lynam et al. [1988] suggested an area of 7 km^2. If a part of the selected area is a Central Business District (CBD) and the other part

is a suburb, then the selected area for analysis may not be appropriate because the CBD area might well appear to be an area cluster. Therefore the area selected for analysing accident distribution should have similar traffic and road environment characteristics (e.g. a CBD or a suburb or a rural area or industrial area).

The number and/or rate of accidents are generally used to identify black areas. The accident rate could be:

- the number of accidents per unit area;
- the number of accidents per person living and/or working in the area;
- the number of accidents per unit length of road in the area;
- the number of accidents per vehicle owned by persons living or working in the area;
- the number of accidents per Vehicle Kilometre of Travel (VKT) in the area.

The first measure does not consider the length of road or traffic volume but the second measure considers the population. The population is not directly proportional to the traffic volume or length of road, which may differ between similar size areas. Therefore, the number of accidents per length of road within an area could be used to identify black areas, but it will not take into account the traffic flow. Since vehicles owned by different persons may travel different distances in different parts of the area, this measure is not directly proportional to the traffic volume or travel distance. The last measure, the number of accidents per VKT, considers traffic volume and length of road and is the most appropriate method, if traffic volume data are available.

Clusters are generally identified using accident numbers or rates. Two other measures that can be used identify clusters are cost density and risk. These two measures were introduced in the NZ Road Safety Strategy [October 2000]. The cost density is the accident cost (or social cost) per km of road and the risk is the social cost per VKT (vehicle–km travelled). Accident cost is a function of accident severity.

Considering all the points raised in this section, especially the fact that traffic volume data are not readily available, the number of accidents is used in this thesis for the identification of all three types of clusters. However, traffic volume can be indirectly taken into account, by analyzing main and minor roads separately.

2.2.3 *Identifying clusters with cost-effective treatments*

The cost-effectiveness of treatment is a function of the social cost of accidents and the cost of treatment. Major injury accidents and fatality accidents have a higher financial cost than non-injury or minor injury accidents. Therefore, selecting hazardous location because of a high number and/or rate of serious accidents could be justified from the economic viewpoint.

If we need to improve the benefit-cost ratio of accident treatment, then the cost density must be lowered as much as possible. The cost of making a road safe is roughly proportional to the length of road, and the benefit is roughly proportional to the social cost. A high cost density indicates that the road is economical to treat but high-risk locations may not be. The high-risk locations need to be treated to ensure fairness or equity for the frequent users of those risk locations. Therefore both cost density and risk are needed to identify the locations which should be treated, and the traffic volume will be needed when using the risk measure for analysis.

In some cases sites with a high number and/or rate of a particular type of accident, with no injury or fatalities, could be selected for treatment. For example, if a site has a high number of right-of-way accidents then a high accident reduction could be achieved by the erection of a give-way sign, and this is a single low cost treatment for that site. If a location has a high number and/or rate of several types of accidents then remedial treatment might not be cost effective, because several types of remedial treatments may be needed for a high accident reduction. For example, if an intersection has several types of accidents (e.g. right-turn, pedestrian and rear-end accidents), then the remedial treatments are: introduction of traffic signals, improved skid resistance, and channelization. The accident reduction even if high may not be sufficient for the treatments to be cost effective. Therefore, a decision on whether to undertake remedial treatment depends on the number and/or rate of accidents, the accident severity, the number of factors causing the accidents, and the cost and the effectiveness of treatment.

The identification of hazardous locations for treatment should include both the number and type of accidents (e.g. alcohol-related accidents) occurring and the cost-effectiveness of the possible treatment.

2.2.4 *Characteristic length of a cluster*

The level and nature of accident clustering revealed by an analysis of spatial distributions will depend upon the scale and shape of clusters. If we view a plot of accidents in a road network then we may see that accidents are clustered at particular sites (e.g. road curves or intersections) but within an accident-clustered road curve there might be random accidents. It is important to define the scale or size before identifying the cluster. Thomas [1995] mentioned that " both problems (size/scale, shape) have been acknowledged for a long time, but researchers have mainly concentrated on the assessment of their effects on statistical measures such as variance or correlation coefficients: when decreasing the scale of the analysis, the correlation usually increases monotonically towards unity, and the variance decreases". Therefore, we need to define the scale or size with more care (e.g. the characteristic lengths of clusters) before undertaking analysis of spatial distributions.

The characteristic lengths of clusters are quite different for black spots, black routes and black areas. The characteristic length of a black spot is relatively small, such as the width of an intersection including approaches or the length of a dangerous curve. The characteristic length of a black route must be a relatively long section of a road, which may contain intersections and curves but no black spots. That is, a black spot could be a short length of road (l_p), shown in Figure 2.05, or an intersection (Figure 2.06), while a black route could be a long or substantial length ($l_l >> l_p$) of road (Figure 2.07), and a black area could be ($l_a \times l_p$) sq.km, as shown in Figure 2.08. Considering Figure 2.05, if we count the number of accidents (n = 3 say) within a small length of road (l_c), which is smaller than the actual characteristic length (l_p) of a cluster, then the count is smaller than the actual count (n = 5 say) of that cluster. In this case we count only a portion of the cluster. The extent of clustering is related to characteristic length. If we use smaller lengths than the length (l_c) then we may count single accidents and conclude that there is no clustering. Therefore, a correct size of characteristic length is important to distinguish between clusters and non-clusters distributions.

Previous studies may provide some guidance on the possible range of characteristic lengths for each type of cluster (point, line and area clusters). The section lengths discussed in IHT [1990] are:

- black spot: "The location may be a single junction, a small area 200-400m in diameter, or a short length of road 300-400m",

- black route: "Usually the search process uses a highway unit of 0.5 - 1.5km in length",
- black area: " The boundaries of areas to be treated may be determined by routes forming the main road network, by administrative boundaries or for example by selecting 1km squares."

Zegeer [1982] noted that in the USA the characteristic length for black spots is 30-500m and for black routes is 500-2500m lengths. Silcock and Smyth's [1984] survey of UK practice, indicated that a characteristic length of 30m was commonly used. Dalby [1987] studied the spatial distribution of urban road accidents, with accidents occurring within 20m being grouped. In NZ [Nicholson 1995], it is standard practice to group together those accidents occurring within 35m of the centre of an intersection, giving a characteristic length of 70m for black spots.

The factors to be considered when deciding the desirable characteristic length are:

- roadway and traffic characteristics need to be fairly uniform within the characteristic length;
- the level of precision and degree of error in the measurement of an accident location;
- effective range of a hazard; an accident could appear to occur far from where the accident was triggered, especially if the speeds of vehicles are high;
- statistical reliability; if the selected characteristic length of a cluster is smaller than the actual characteristic length then the chance of accident within that length will be one or nil and this will suggest no clustering. If the selected characteristic length is larger than the actual characteristic length of any clusters, then separate clusters might appear as a single cluster.

Closely located clusters, which cannot be clearly distinguished, could be called joint clusters. Examples of joint clusters are shown in Figures 2.09 a, b and c. In these joint clusters defining the characteristic length is difficult, because it varies between the clusters. It may be better to use a range of characteristic lengths instead of relying on a particular characteristic length for the analysis. Identifying the range of characteristic length is useful for analysis where the data contains point clusters with different characteristic lengths.

A cluster of accidents at a particular site may be a single type. For example vehicles may collide with a pole, which is near the edge of the road, as shown in Figure 2.10, or an accident

cluster appears at a site where the sight distance is reduced by an obstacle, as shown in Figure 2.11. Some types of accidents are more likely to occur in compact clusters, and the characteristic cluster lengths may depend upon the types of accidents in the clusters. For example in Figure 2.10 the characteristic length is small while in Figure 2.11 the characteristic length is large.

Classifying and investigating the common types of accidents will be useful in deciding on a suitable range of characteristic lengths, which could be used for analysis. In this regard it is better to find out the percentage contribution of the various accident types to the total number of accidents. This will be useful in deciding whether clusters which have a small characteristic length could be neglected. It is reasonable to neglect a particular accident type if the percentage is very small.

Key information obtained from the accident record could be considered for identifying the characteristic length of each cluster type. OECD [1999] reported that rural road accidents in 1980 and 1996 contributed 55% and 60% of total accidents. These figures are derived from accident data for the OECD member countries, including NZ. The OECD study also found that 80 % or more of fatal accidents on rural roads consisted of:

- single vehicle accidents, especially running off the road (35% or more);
- head-on collisions (25% or more);
- intersection accidents (around 20%).

More than 25% of motorist deaths in Sweden are due to collisions with stationary objects on the roadside. OECD [1999] referred to SETRA/CETUR [1992] which reported that the highest mortality rate was for collision with trees (25.1 fatalities per 100 accidents) followed by collision with utility poles (17.1 fatalities per 100 accidents). OECD [1999] found that about 40% of fatalities in rural roads occur when vehicles collide with stationary obstacles such as poles, trees, culverts or bridges, road signs and advertising posts. Since the percentage of vehicles colliding with such objects is reasonably high, the characteristic length of this type of accident may be considered for analysis to clearly identify compact clusters (as in Figure 2.10).

The factors affecting roadside hazard accidents include: pole or tree density; the distance of the pole or tree from the edge of the road; whether poles or trees are located on the outside or

inside of horizontal curves; low road friction coefficient. Loss of vehicle control is common for all these accidents. This could be due to the driver falling asleep, or inappropriate (or excessive) speed together with poor road surface conditions (e.g. potholes, water or ice). The accident clusters caused by on-road hazards depend on the number of roadside hazards near each other in specific locations. Roadside features may cause clusters, random or regular accidents distributions. Regular accidents might be from lampposts, cluster of accidents from a lamppost close to the road edge on the outside of a circular curve, and random accidents from a feature like a long concrete median barrier.

A hypothetical example of roadside accident cluster is when drivers lose control and collide with a single roadside hazard (e.g. a tree or pole) because of a badly designed or constructed road curve. The recorded accidents must occur within a small area and hence the characteristic length of the cluster is small. If along a badly designed or constructed road curve there are several roadside hazards, collisions may occur over a long segment of that curve. In this case the characteristic length might be set to cover all the recorded accidents.

In some rural areas, roadside pole density is high. A driver may fall asleep near one pole resulting in an accident. Such accidents are likely to be random because falling asleep is generally random. Random accidents can also occur, for example, anywhere on a section of road which has a long concrete median barrier, due to vehicles initially hitting the median barrier and colliding with vehicles following, or alongside. These accidents will tend to be random within the section, unless the loss of vehicle control was because of reasons specific to a particular position or positions.

The characteristic length of accident sites should be small. If a large characteristic length is used, isolated accident sites may join together and the chances of point clusters or isolated features being masked are high. This will mean a loss in confidence in the statistical reliability of the analysis.

Generally, accidents up to 20-30 m along each approach road are considered as part of an intersection site, because accidents within these approaches are often associated with the intersection (e.g. rear end accidents). Therefore, to identify intersection accident sites, the characteristic length must be at least 70m (say 25 m up each approach road and 20m on average for the intersection). Some roadside hazard accidents are clustered in less than a

characteristic length of 30m. To allow for the variation in characteristic length, a range of values, from 5 to 70m, should be used.

IHT [1990] suggested a characteristic length between 0.5 and 1.5 km as reasonable for identifying black routes. Defining the characteristic length for detecting area clusters is far more complicated and depends on the distance between roads, traffic volume, and total road lengths. One may also consider the location of administrative boundaries.

2.2.5 *Analysis period*

When analysing accident data to identify the spatial variation in accidents, the data must be selected for a specific analysis period. The number of accidents and the percentage of each accident type within an area are subject to temporal variations, which need to be considered during data analysis. The proportions of accident types and the proportions of accidents during peak /off-peak periods, day/night, and winter/summer, are often different in different years.

A classic example of seasonal variation in accident occurrence is the variation in the proportion of skidding accidents in Britain during a year. Skid resistance is subject to seasonal variation, as shown in Figures 2.12 and 2.13. The percentage of the accidents on dry roads does not vary much with the season, but on wet roads there are big variations (Figure 2.12). How the percentage of accidents involving skidding on wet roads is correlated with skid resistance can be noted from Figure 2.13. The data needs to cover the full range of weather.

The cyclical variation in accidents in NZ from year to year is shown in Figure 2.14. A period less than one year is not a suitable duration for analysis, as it does not take account of the seasonal variation. When considering the annual variation shown in Figure 2.15, a period such as one year is not sufficient for a full cycle. Sometimes there is a large variation in the weather conditions from year-to-year. Annual variation may be allowed for if we consider more than one year of accident data. Zegeer [1982] noted that in the USA, one and three years of accident data are used to identify hazardous locations.

The spatial pattern of accidents depends on the length of time of the data. For example, in a one-year period, different sites may appear to be black spots in different periods because of temporal variation in accident occurrence at different sites. For example, the spatial pattern of accidents may change if the length of time is increased as shown in Figure 2.16. If the expected number of accidents per year is quite small then the statistical reliability is small. When the expected number of accidents per year increases, the number of years required for statistical reliability decreases.

Silcock and Smyth [1984] mentioned that in the UK the length of time selected for analysis ranges from one month up to five years, but three years is the most frequently used period. Zegeer [1982] also noted that the use of the three-year period is frequent, but he recommended one or three years time period. It is common practice to assume that the occurrence of traffic accidents at a site is governed by a Poisson distribution. Nicholson [1986b, 1987] mentioned that from a statistical viewpoint, the analysis of actual accident count data shows that a five year period of accident data is most suitable.

The greater the period, the greater the expected accident count for a single site, route or area. The number of accidents at a site may be few and vary from year to year. For statistical reliability, the number of accidents must be sufficient and this requires a longer period. However the disadvantages of a longer period are the inclusion of changes in road layout, traffic flow or road geometry. Therefore for analysis, one and three year periods are better than a shorter period, if the data is sufficient.

The statistical reliability for identifying black routes is not so critical compared to black spot identification, because the accident counts of sites on the road are added together to calculate the accident count for the route. The total accident counts for a route are more than that of individual sites. Therefore, shorter periods (one year, say) may be sufficient for black route analysis. Similarly, for the identification of black areas, a one-year analysis period may be sufficient, if accidents in a large region are being analysed.

A five year analysis period has been chosen for the case study described in Chapter 9.

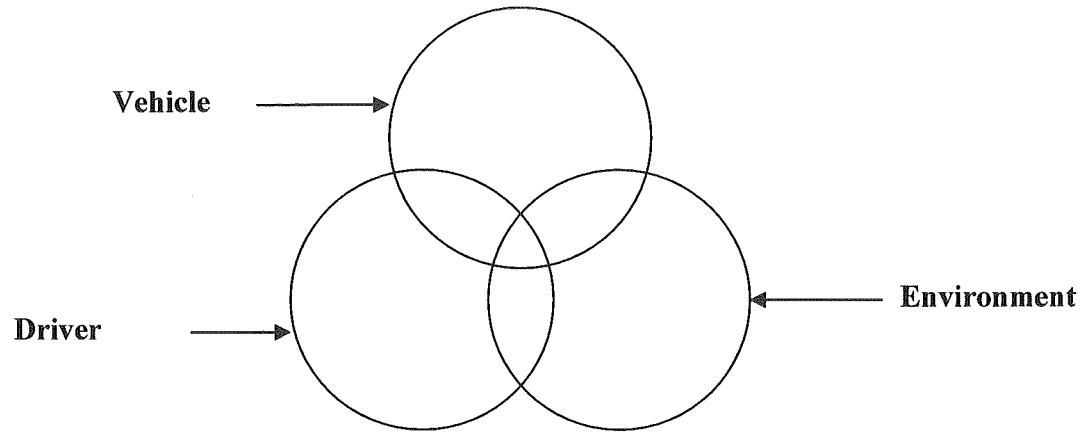


Figure 2.01: Factors that influence the accident clusters

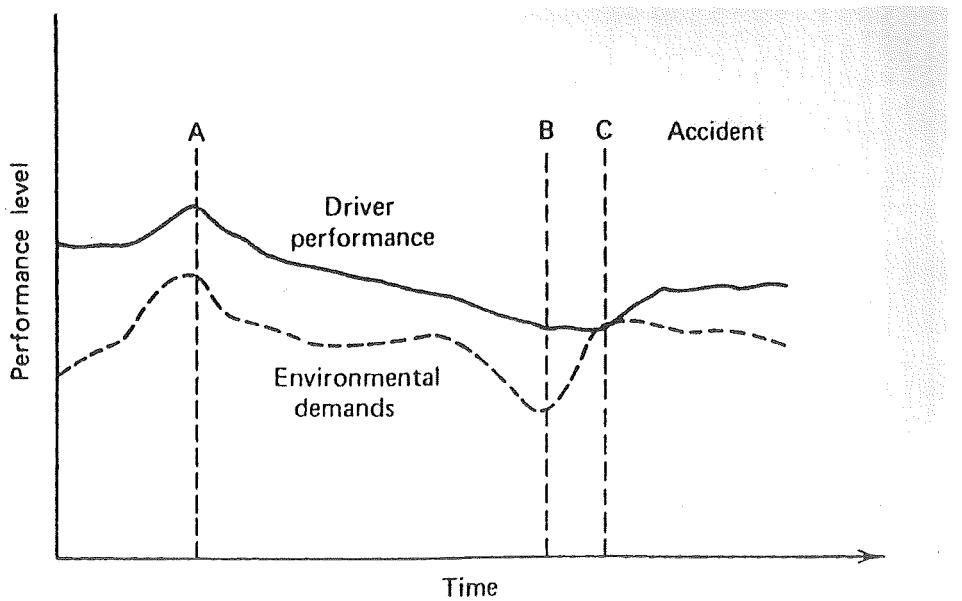


Figure 2.02: Interaction between environmental demand and driver performance
(Figure extracted from David Shinar [1968])

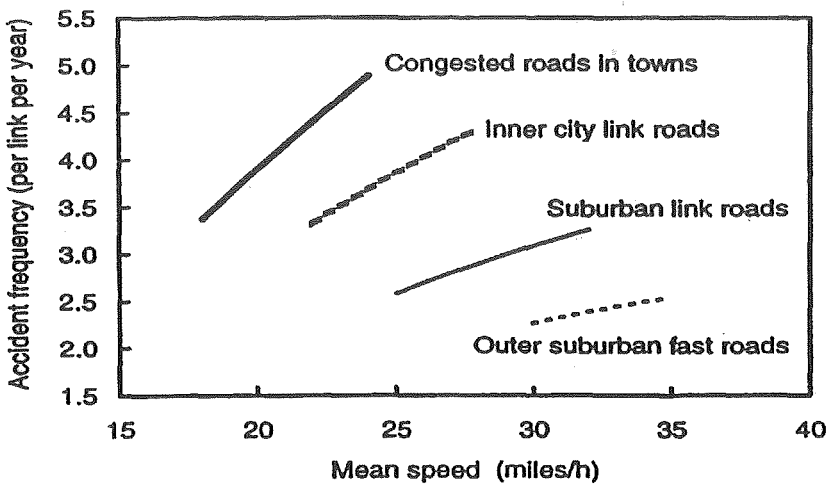


Figure 2.03: Mean speed and accident frequency profile
(Figure extracted from Taylor et. al. [2000])

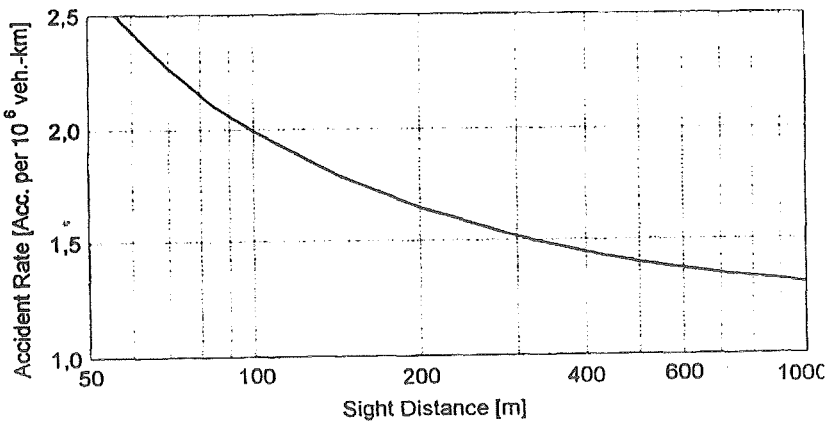


Figure 2.04: Accident rate as a function of sight distance on two-lane rural roads (Figure extracted from Ruediger [1999])

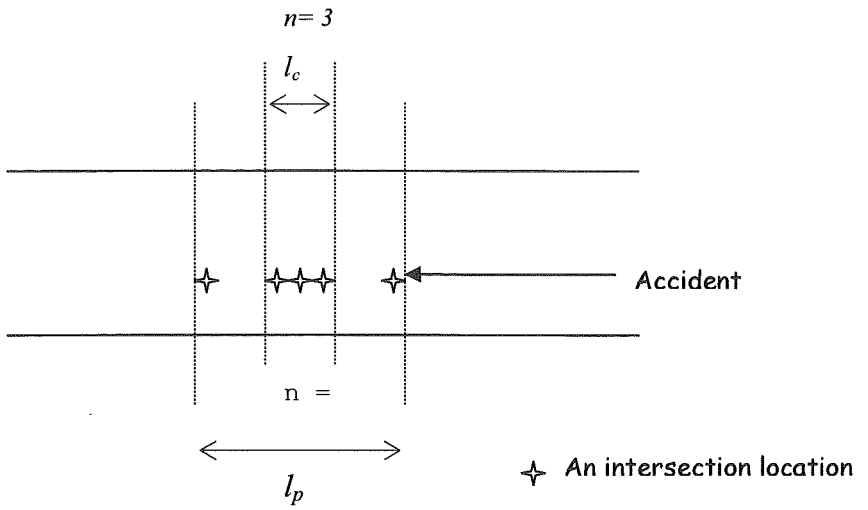


Fig 2.05: A section of a road as a black spot

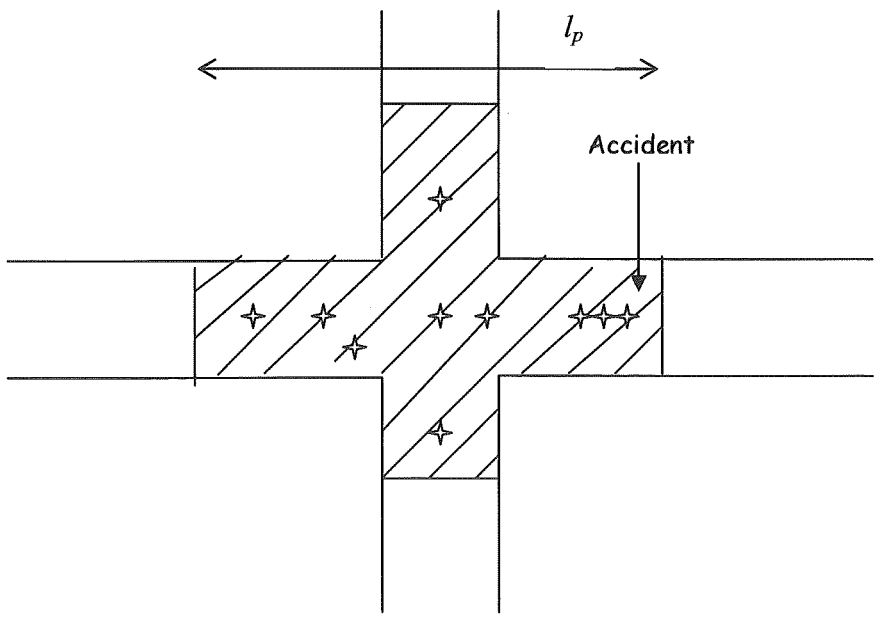


Fig 2.06: An intersection as a black spot

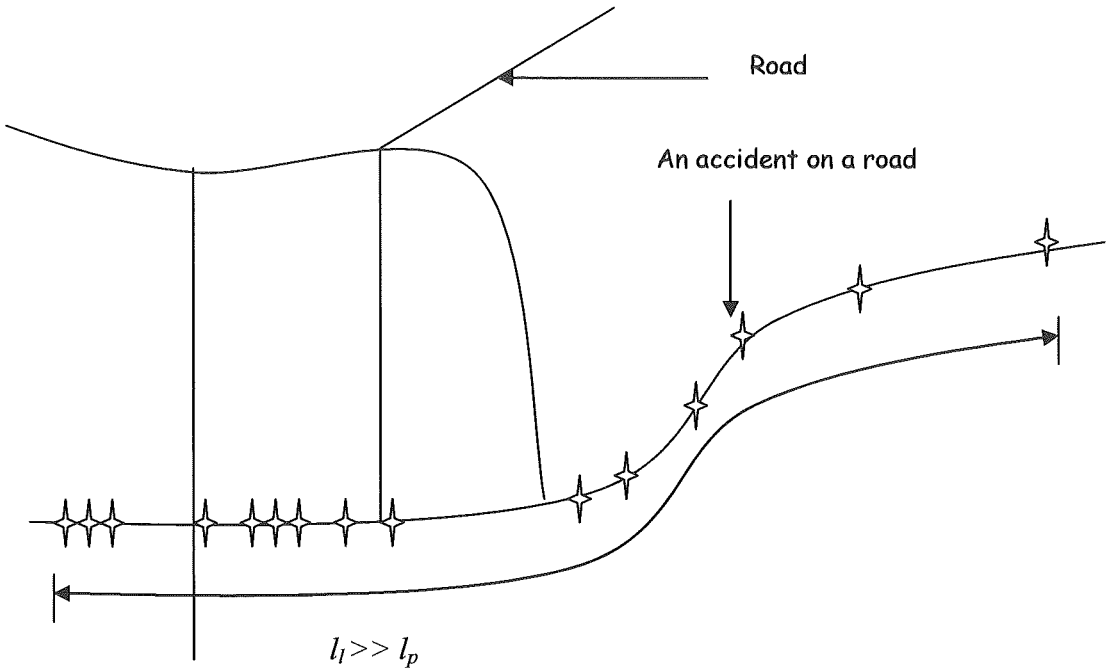


Figure 2.07: A route cluster

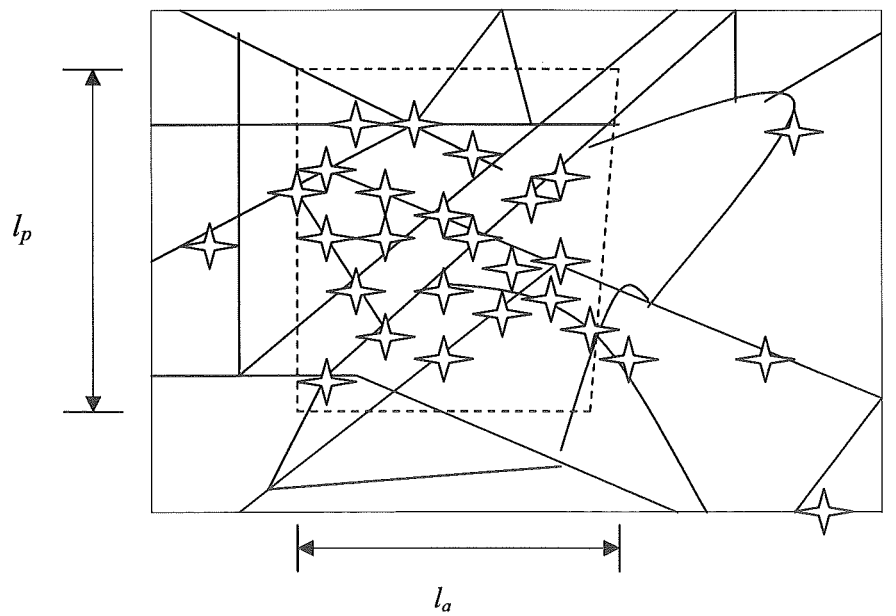


Fig 2.08: A black area with several roads

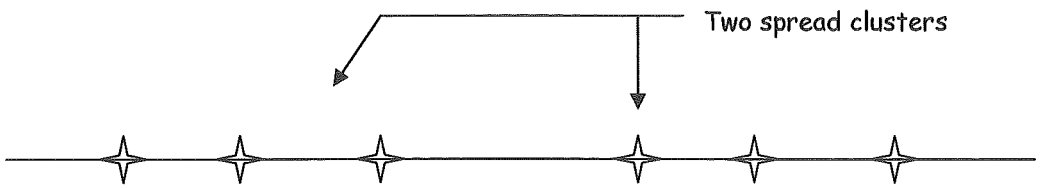


Fig 2.09 (a): Joint clusters

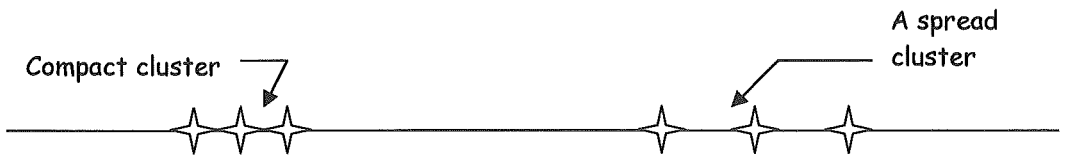


Fig 2.09 (b): Joint cluster (a compact cluster and a spread cluster)

Fig 2.09 (c): Joint cluster (a spread cluster and an isolated accident from a random process)

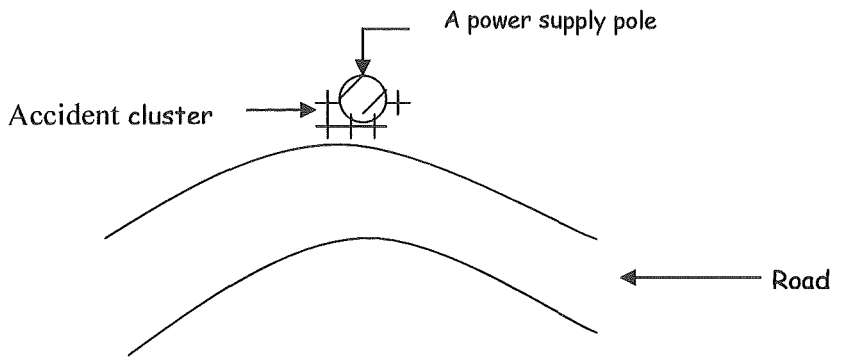


Fig 2.10: Accident site (vehicles collide with a pole)

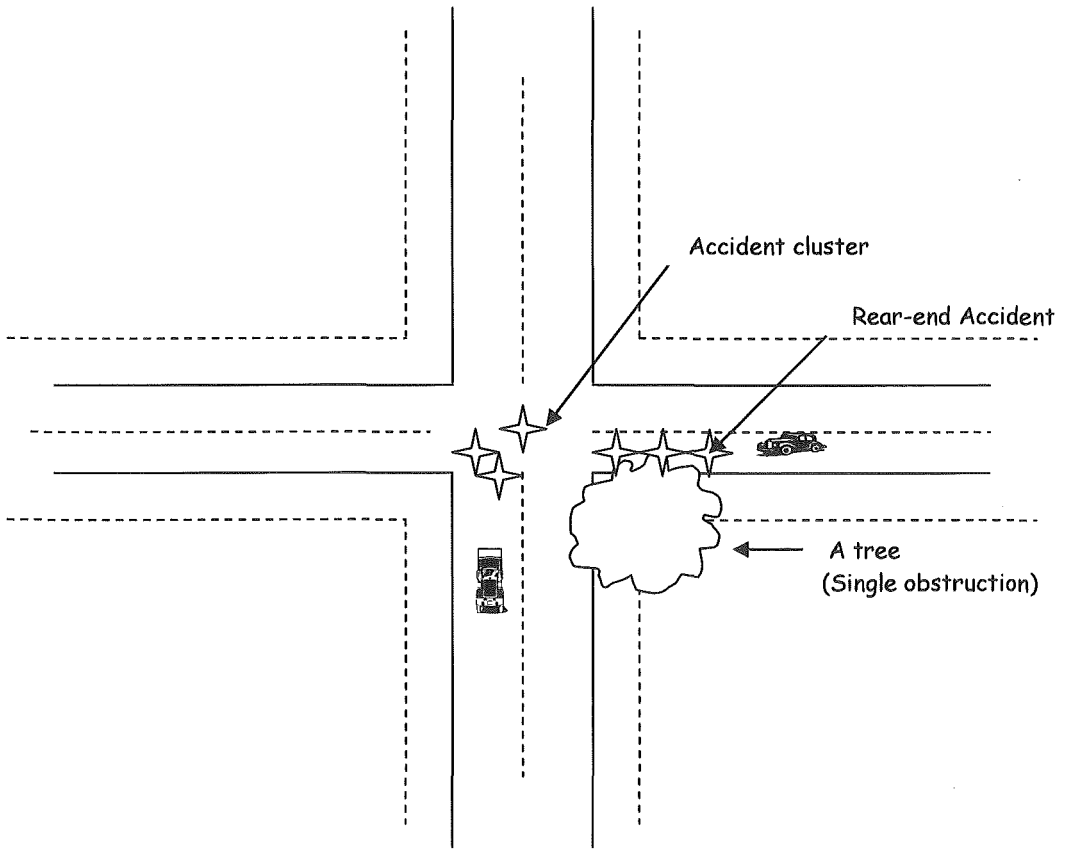


Fig 2.11: Accident site (sight obstructed by a tree)

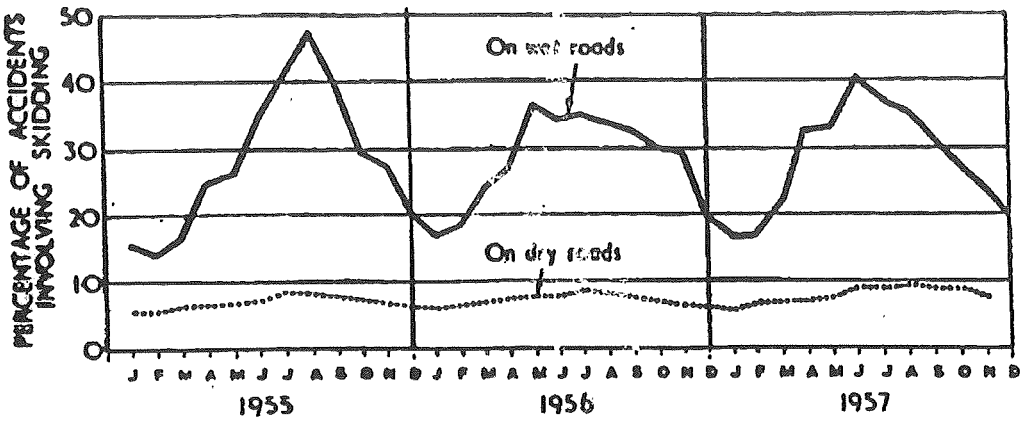


Figure 2.12: Seasonal variation in skidding accidents (data for injury accidents in Great Britain 1955-57)

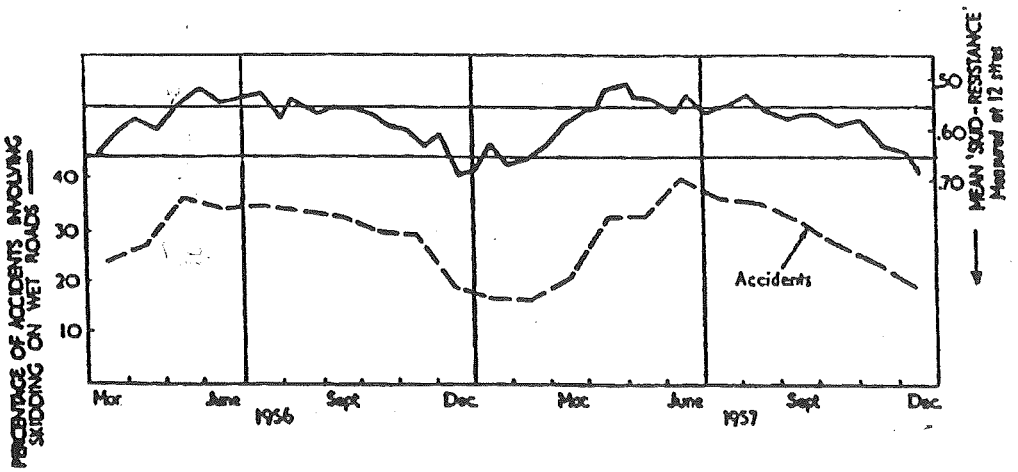


Figure 2.13: Seasonal variation in skidding accidents and skidding resistance

(Figures 2.12 and 2.13 extracted from Road Research Laboratory [1963])

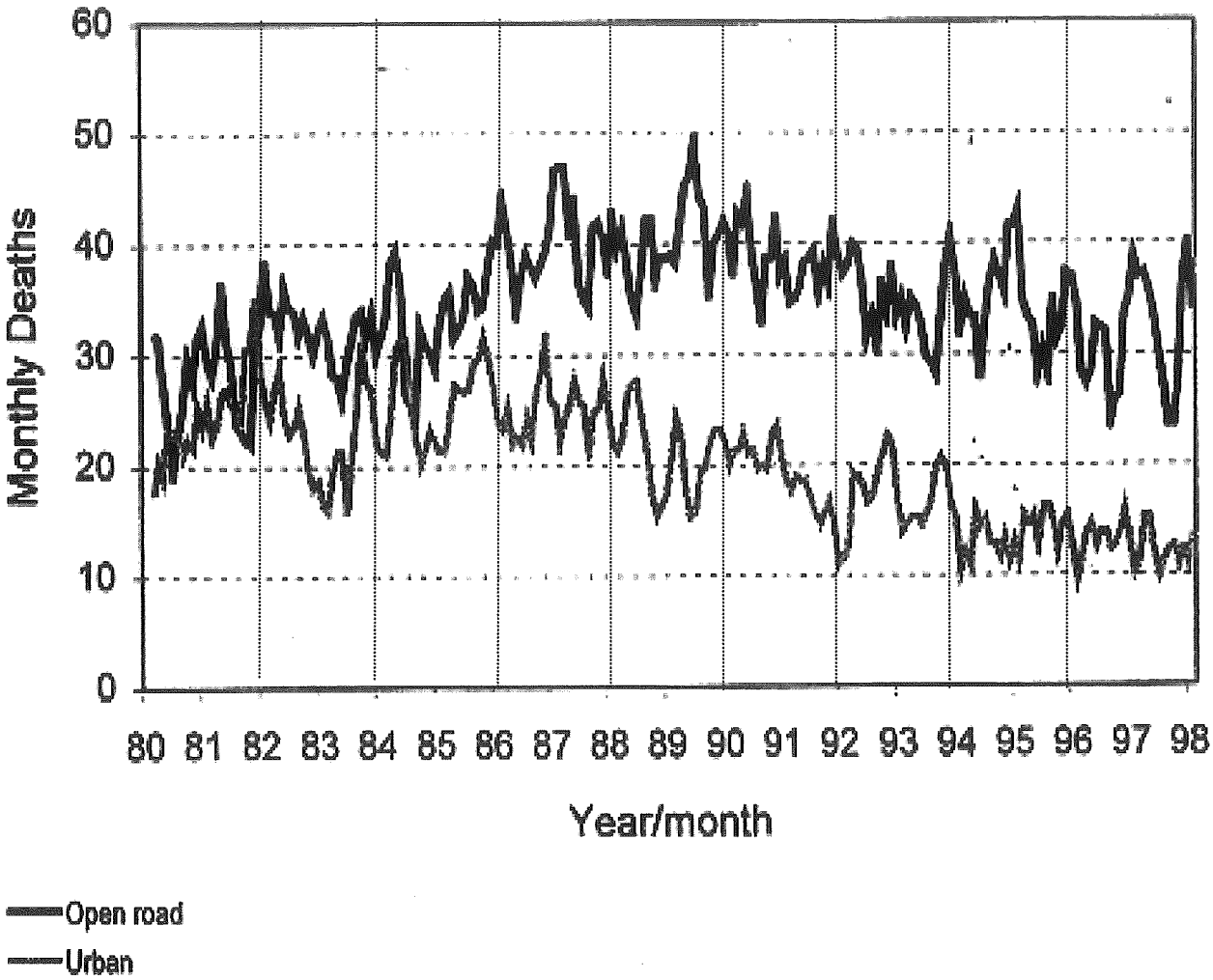


Figure 2.14: Graph of number of deaths in each month in open road and urban road from 1980 to 1998 (Extracted from LTSA [1998])

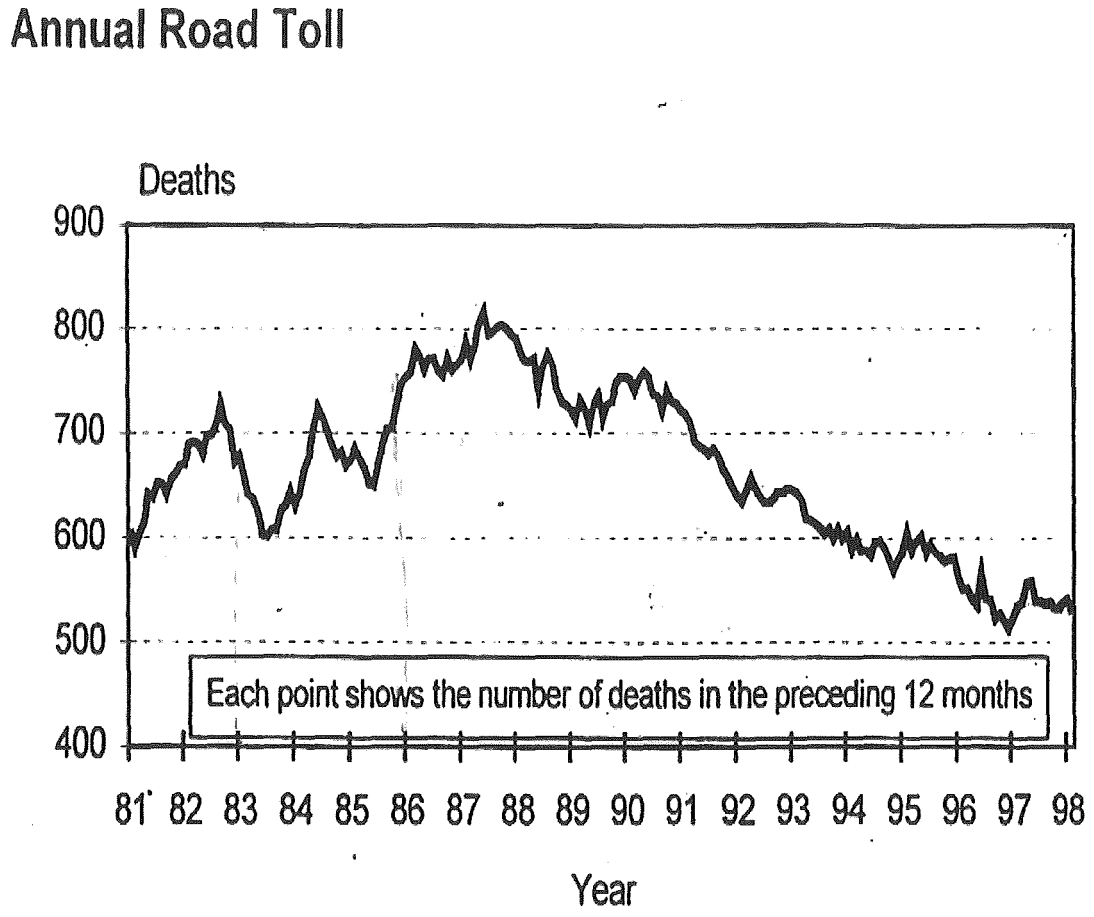


Figure 2.15: Annual road toll (Extracted from LTSA [1998])

Year 1 Apparent point cluster

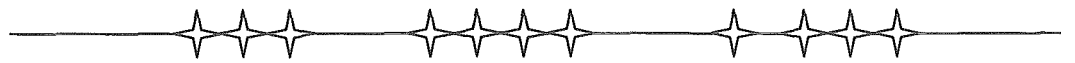
(a)

Year 2 Apparent random accidents

(b)

Year 3 Apparent point clusters

(c)



Year 1, 2 and 3 Apparent strong point clusters

(d)

Figure 2.16: Accident distribution for a section of road for different time periods

Chapter 3

LITERATURE SURVEY

3.1 General

As discussed in Chapter 2, accidents tend to occur in clusters. Hoque and Andreassen [1986] investigated accidents which occur in clusters using "...x% of accidents occur at sites with z or more accidents and x% of accidents occur at y% of the sites". Nicholson [1989] noted that the level of clustering increases as x increases and/or y decreases, or x and/or z increases. Comparisons of the levels of clustering for different areas are difficult because there are no standard values for these numbers (x, y, and z). Nicholson noted the absence of ".... quantitative definitions of cluster and a defined breakpoint between cluster and non-cluster distributions of accidents".

Nicholson [1989] assessed the level of accident clustering in a network using the accident count profile, count frequency distribution, count cumulative frequency distribution and count concentration curve shown in Figures 3.01, 3.02, 3.03 and 3.04. Figure 3.04 is analogous to the Lorenz curve (used by economists). These profiles indicate the level to clustering, which can be identified from the deviation between the "actual" profile and the "perfect equality" profile (i.e. the accident counts for all locations are equal), but they are less concise than numerical indices. Assessing the level of accident clustering could be used to help identify the most appropriate accident reduction programme and assess the effect of an accident reduction programme.

Nicholson also discussed various numerical indices. A simple index of inequality is the "range", which is the difference between the maximum and minimum accident counts. This index is not very useful because of the absence of detailed information (i.e. the mean counts and the frequency of counts). Another index is the "relative mean deviation", which is the ratio of the shaded area in Figure 3.01 to the area under the perfect equality profile. This index is also not very useful because the index value may be the same for some quite different distributions of accident counts. Another index of clustering is the "variance of counts", which is also not very useful because there may be changes in the distribution of

accident counts without any changes in the variance. Another problem with the last two indices is that the same proportional change in the accident count at every site would give a change in the index value. Because of these reasons, Nicholson [1989] suggested two indices, the coefficient of variation (CV) and the coefficient of concentration (G).

Nicholson noted that CV and G are "accident-total independent" (i.e. if the accident count for each site is increased by the same proportion and the total number of accidents is increased, there is no change in the index value) and "population-total independent" (i.e. if there are two sets of sites with the same distributions and the same values of the clustering index, the sum of the two sets also has the same value of the clustering index). Accident-total independence is important, because increasing the period of accident data will affect the accident count but will not affect the index. Nicholson found that the CV value is more sensitive than G when measuring different levels of clustering, but G is the better indicator in terms of indicating accident clustering. Therefore, Nicholson suggested that it is sensible to use both CV and G as indicators of changes in accident clustering.

Nicholson [1990, 1995] described the use of four indices (the coefficients of variation and of concentration, and two indices based on information theory) for analysing the spatial distribution of accidents. Interpretation of these indices is difficult when allowing for random variations in the accident counts from year to year. Nicholson [1995] noted that the four indices can indicate a reduction in accident clustering when the number of multiple-accident sites increases.

Shaikh [1990] studied accident clustering in five urban areas of New Zealand. The number of accidents within 70m squares, centred on intersections or the mid-points of any group of non-intersection accidents, was used to calculate these four indices. The analysis showed substantial temporal and spatial variations in the indices of clustering. The interpretation of those variations is difficult, because the extent to which they are due to randomness in accident occurrence is not clear.

Nicholson [1987] showed how to calculate the "Underlying True Accident Rate" (UTAR) at a site from the annual accident counts for a stationary accident occurrence process. He also noted that if the accident count data covers five or more years then the UTAR is generally very similar to the "observed accident rate". The UTAR calculated from a five-year period

or more could be used to calculate the indices of clustering, taking account of the randomness of the accident counts. If all the sites have approximately the same UTAR, then the indices will indicate less clustering.

Nicholson [1990, 1995] noted that the problem with using the four indices is that while they may indicate considerable non-uniformity in the distribution of accidents between sites, they do not provide information about the spatial relationships between clusters. Accident clusters may be spatially close or concentrated along a route; such information cannot be readily detected using the four indices. Hence the usefulness of the indices is limited.

Table 3.01 is an example of an accident count matrix. The matrix contains the accident counts of I locations over a period of J years. Let X_{ij} be the observed accident count at the i^{th} location during the j^{th} year, where i = 1,2,3 ...,I and j = 1,2,3,...,J. Let m_i be the expected number of annual accidents at the i^{th} location. In each year if the expected number of accidents at a location i is the same as in other years, and equals m_i, then the count process is stationary. But the observed annual accident counts need not to be equal. The level of clustering of accident counts can differ from the level of clustering of UTARs.

Table 3.01 : Accident count matrix

Location	Years								Underlying True Accident Rate
	1	2	3	4	...	j	...	J	
1	X_{11}	X_{12}	X_{13}	X_{14}	...	X_{1j}	...	X_{1J}	m_1
2	X_{21}	X_{22}	X_{23}	X_{24}	...	X_{2j}	...	X_{2J}	m_2
3	X_{31}	X_{32}	X_{33}	X_{34}	...	X_{3j}	...	X_{3J}	m_3
4	X_{41}	X_{42}	X_{43}	X_{44}	...	X_{4j}	...	X_{4J}	m_4
:	:	:	:	:	...	:	...	:	:
i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	...	X_{ij}	...	X_{iJ}	m_i
:	:	:	:	:	...		...	:	:
I	X_{I1}	X_{I2}	X_{I3}	X_{I4}	...	X_{Ij}	...	X_{IJ}	m_I

If the accident counts at sites vary randomly then that location may appear to be hazardous in one year but not in the following year. It is important to analyse UTARs rather than the annual accident count. Analysing the accident counts for n years is a good approximation to analysing the UTARs if n is large.

Nicholson [1986a] found evidence that the accident counts at some locations were not random. Nicholson noted that the "existence of regular fluctuation (....above, below, above, below,....) about the mean annual accident count has been observed". In this case also analysing the accident counts for n years is a good approximation to analysing the UTARs if n is large.

Maher [1987] suggested that the UTARs at neighbouring sites will be similar and may not be independent. This means there will be positive correlation of the UTARs at neighbouring sites because of common factors (e.g. traffic flows, road layouts, forms of control). Positive correlation of UTARs at neighbouring sites would mean less variation in the underlying true accident rates (and the accident counts) between neighbouring sites than between randomly selected sites.

Loveday [1991] mentioned the note in Boyle and Wright [1984] " high-risk sites tend to cluster together" (that is, black spots tend to cluster together). Loveday [1989 and 1991] investigated accident data from 15 London boroughs to identify the evidence of spatial correlation between the accident counts from an "abstract network" rather than a "real network". In this study of an "abstract network" (see Figure 3.06) both the nodes and links in the "real network" (see Figure 3.05) were considered as nodes and the correlation between the observed accident counts was calculated using Moran's index (I) defined as follows;

$$I = \frac{\left(n \sum_i \sum_j (m_i - \bar{m})(m_j - \bar{m}) \right)}{\left(\left(\sum_i \sum_j \delta_{ij} \right) \left(\sum_i (m_i - \bar{m}) \right) \right)}$$

where $m_1, \ldots, m_n$ are accident counts at the n nodes, $\bar{m}$ is the mean accident count for all the n nodes and $\delta_{ij} = 1$ if node i and j are connected by a link in the abstract network, otherwise $\delta_{ij} = 0$. This study showed that the index values for all boroughs were significantly

greater than zero (that is, the spatial correlation was statistically significant). In this method, the precise location of each individual accident was not considered. The method involves pair-wise comparisons of the accident counts at accident sites. The relative positions of sites or the distance between each accident position were not considered. Fischer et al. [1996] mentioned a note in O'Loughlin et al. [1994] ".....in an analysis of the Weimar elections in 1930 in Germany, a highly significant Moran's index at the level of 921 electoral districts in effect hides several distinct local patterns of spatial clustering and complete spatial randomness". This approach seems inappropriate for a detailed analysis of the spatial distributions of accidents to detect local patterns.

3.2 Spatial processes

Let $\mathbf{s_i}$ be the vector from the origin to the i^{th} position in two-dimensional space. Let $\mathbf{S} = \{\mathbf{s_1}, \mathbf{s_2}, \ldots, \mathbf{s_i}, \ldots, \mathbf{s_n}\}$ be the set of spatial locations where accidents can occur, and $\mathbf{Q} = \{ Q_{(S1)}, Q_{(S2)}, \ldots, Q_{(Si)}, \ldots, Q_{(Sn)}) \}$ be the set of observed accident counts for those locations. According to Cressie [1993],

$$E(Q_{(s_i)} - Q_{(s_j)})^2 = (\mu_{(s_i)} - \mu_{(s_j)})^2 + \text{cov}[Q_{(s_i)}, Q_{(s_j)})] \qquad (3.1)$$

If the set of locations (S) and the set of counts (Q) are random then the spatial process is random. Cressie noted that, if a process is completely spatially random then the spatial process must satisfy both the stationarity and isotropy conditions, as follows:

1. Stationarity condition: the expected value and variance of $\mathbf{Q_{(S)}}$ are both constant, i.e.,

   $$E[Q_{(S)}] = \mu$$

   $$\text{var}[Q_{(S)}] = \sigma^2$$

   and , it follows from equation 3.1 that $E(Q_{(s_i)} - Q_{(s_j)})^2 = \text{cov}[Q_{(s_i)}, Q_{(s_j)})]$ because $\mu_{(s_i)} = \mu_{(s_j)}$

2. Isotropy condition: the covariance ($\text{cov}[Q_{(s_i)}, Q_{(s_j)})]$) depends only upon the magnitude of the distance between the points $\mathbf{s_i}$ and $\mathbf{s_j}$ and not on the direction between the points.

   Cressie [1993] noted that for all $\mathbf{s_i}$ and $\mathbf{s_j}$,

   $$\text{var}[Q_{(s_i)} - Q_{(s_j)}] = 2\gamma(s_i - s_j) \qquad (3.2)$$

   where γ is a function of $(s_i - s_j)$. The function $2\gamma(s_i - s_j)$ is called the variogram.

Cressie also noted that the classical estimator of the variogram proposed by Matheron [1962] is

$$2\gamma(\mathbf{h}) = \frac{1}{N_{(\mathbf{h})}} \left(\sum_{i=1}^{N_{(\mathbf{h})}} \left(Q_{(s_i)} - Q_{(s_j)} \right)^2 \right) \qquad (3.3)$$

where $N_{(\mathbf{h})}$ is the number of distinct pairs of points separated by **h**, where $\mathbf{h} = (s_i - s_j)$. The magnitude of **h** is called the lag distance.

Cressie suggested that there are three distinct types of spatial model (namely geostatistical, lattice and point processes) depending upon whether the data are:

- continuous or discrete; i.e. related to points in space or spatial aggregations (e.g. areas);
- related to locations that are regular or irregular;
- related to locations from a spatial continuum or a discrete set.

Because accidents occur on road networks, the lattice model would appear to be most appropriate for the analysis of spatial distributions of accidents. For a road network, if sufficiently dense, then this model may be approximated with the continuum model and a point distribution model.

A point distribution could be a combination of CSR (Figure 1.03), regular (Figure 1.02) and/or clustered (Figure 1.04). Nicholson [1995, 1998] classified point distributions in a different way. For a CSR distribution (Figure 1.02), the spatial process must exhibit both stationarity and isotropy. The other processes (which are non-random) are:

1. non-stationary and isotropic (Figure 3.07);
2. stationary and anisotropic (Figure 3.08);
3. non-stationary and anisotropic (Figure 3.09).

This type of classification of accident distribution is better than the traditional classification (CSR, regular and cluster), because the Nicholson classification system helps to identify the most appropriate accident reduction plan. Nicholson also noted the three basic types of non-random accident distributions and the appropriate accident reduction plans are:

1. for non-stationary and isotropic (Figure 3.07) the appropriate type of plan would be a site plan;
2. for stationary and anisotropic (Figure 3.08) or stationary and isotropic (Figure 1.03) the appropriate type of plan would be an area plan;

3. for non-stationary and anisotropic (Figure 3.09); the appropriate type of plan would be a route plan.

Testing the spatial dependence is an important part of the spatial analysis of accidents. If the spatial variable (i.e. the accident count) depends on the characteristic of the accident site (local location) but not the area (global location) then the spatial pattern of accidents is point cluster, and if the variable does not depend on the characteristic of local location (i.e. not depend on sites or route) but is spatially dependent, then the distribution is an area cluster (i.e. random or regular counts depend on global location which is an area having several sites). If the accidents counts depend on the characteristic of route then the distribution is a line cluster.

3.2.1 *Spatial correlation and covariance*

These techniques are concerned with exploring spatial covariance, which is helpful in identifying whether neighbouring values (e.g. the accident counts at neighbouring sites) are correlated. There are two standard tests for correlation: one is based on statistics such as Moran's index, and the other (known as the variogram method) is based on plotting the mean square difference between counts at different lag-distance along particular directions against lag distances. Moran's index has already been discussed, and the variogram method is discussed in the next section.

3.2.2 *Variogram Method*

A process is spatially stationary and isotropic if the expected accident count and the variance of accident counts at sites are constant, and if the covariance depends upon the relative location, but not the absolute location. If the variogram is direction-invariant, that is it depends only upon the Euclidean distance between the points, but not the direction between the points, then the spatial process is isotropic. Nicholson [1999] discussed the three classical accident distributions (point cluster, line cluster and area cluster distributions) in terms of combinations of stationary or non-stationary and isotropic or anisotropic process. These processes can be identified using the variogram.

The classical estimator of the variogram is given in Equation 3.3. A plot of $\gamma(\mathbf{h})$ against $\mathbf{h}$ is known as the experimental variogram. Alley [1993] used a variogram to investigate spatial variation of regional ground-water quality. An example of variogram estimation is shown in Figure. 3.10. The property of soil or chemical property of ground-water can be measured from sample locations (see Figure 3.10). The sample locations have a regular lag distance but accident sites generally do not occur at a regular lag distance in one direction.

Problem arises when alignment of a road does not match alignment of the quadrats. The variogram method may be used for analysing accidents in a regular grid road network but the intention in this thesis is to develop a method to analyse accidents in any type of road network.

3.2.3 *Lattice and continuum models*

Generally road accident data can be considered to be lattice data, and the lattice may be regular or irregular. Measuring distance on a lattice is more difficult than measuring a distance in a continuum. Lattice distance should be measured along the roads. For example, if we consider a sample event on the middle of a long link (Figure 3.11a) then neighbouring events within the two arms (i.e. both sides of the sample event) should be considered. If we select a sample event near an intersection (Figure 3.11b), to analyse the distances to neighbouring events then events within the four arms of the intersection should be considered. In these two cases deciding on a consistent length is difficult. In order to analyse spatial distributions, details of the network configuration will be required for a lattice model. Nicholson [1999] discussed all these practical problems and suggested that lattice techniques are more complex and more difficult to implement. However, it may be reasonable to approximate the lattice with the continuum, where the road network is relatively dense, and Nicholson [1995] carried out preliminary investigations on the errors associated with using such an approximation. He concluded that the approximation is good if the network is a regular grid with block sizes not more than about 250 metres.

3.3 Quadrat method

The quadrat method involves counting the number of events (accidents) within subsets of the selected area for investigation. Traditionally these subsets are rectangular (hence the name quadrats), although other shapes (including circles) are possible.

With the systematic sampling method, the selected area is divided into a regular grid pattern of quadrats and then the number of accidents is counted in each quadrat. If the spatial process is completely spatial random then the distribution of quadrat counts is expected to follow a Poisson distribution [Ripley 1981]. Let m* be the mean of the counts (m_1, m_2, ... m_n) from a set of n quadrat samples and let the variance be s^2. The distribution of quadrat counts may be tested for randomness by using either the χ^2 or Kolmogorov-Smirnov tests. Ripley noted that the statistic (s^2 (n-1) / m*) is approximately $\chi^2_{(n-1)}$ distributed, and if the statistic is sufficiently large then the null hypothesis (i.e. Poisson distribution) can be rejected.

For a Poisson distribution, the variance-to-mean ratio is equal to one, but the converse is not true. There are many frequency distributions that have the variance equal to the mean, and that is the reason why Ripley [1981] suggested the above statistical test, rather than testing the value of the variance-to-mean ratio.

Dale [1999] noted that with the quadrat method, the quadrat count distribution may follow a Poisson distribution, but the distribution of events may not be CSR. That is the frequency of quadrat counts may follow a Poisson distribution but the events may be distributed with a certain pattern. Consider two following cases of spatial distributions.

- Case I, random quadrats in which quadrat counts increase diagonally from the origin (Figure 3.12a).

- Case II, regular quadrats in which quadrat counts degrease from the centroid (Figure 3.12b).

The frequencies of the quadrat counts for the two cases both follow a Poisson distribution, but the events are not CSR. This is not necessarily true for all quadrat sizes. Further investigations into quadrat sizes are included in Chapter 6 and 7.

Ripley [1981] discussed the following six indices to investigate spatial distribution.

(1) Coefficient of variation (CV) $= \sigma / m^*$

(2) Index of clumping ICS $= (\sigma^2 / m^*) - 1$

(3) Index of cluster frequency ICF $= \sigma^2 / ICS$

(4) Index of mean crowding ICR $= \sigma^2 + ICS$

(5) Index of patchiness IP $= 1/ICF + 1$

(6) Morisita's index MI $= n \,.\, m^* \, IP / (n \,.\, m^* - 1)$

These indices may be used to measure the departure from CSR. The interpretation of some of the above indices requires knowledge of the size of the area of clusters when examining the exploratory data analysis [Cressie 1993]. This is not possible without good prior information about the characteristic length of cluster (i.e., information related to the size of the accident clusters). This was discussed in Chapter 2 and is discussed further in Chapter 6.

Thomas [1995] studied accident distributions in motorway segments in Belgium and identified the accident count distributions for different segment lengths. The method was very similar to the quadrat method, with the quadrats being the motorway segments and the quadrat size being the segment length. Thomas concluded that there are three distinct types of segments:

a. 100m road segments, with the accident counts being approximately Poisson distributed,
b. 300 to 2000m road segments, with the accident counts being intermediate empirical distributed (i.e. not a Poisson but may be a negative binomial distribution), and
c. More than 2000m segments, with accidents being normally distributed.

In interpreting the spatial analysis results the segment length should be considered. According to type a, accidents are randomly distributed for short segments and there are no point clusters. According to type c accidents are clustered along lines for long segments, which may be black route segments.

Thomas stated that, "there is no reason for the process to be different for other accident data sets". It should be noted, however, that the selection of motorway segments between entries and exits do not consider all type of accidents like accidents at intersection, pedestrian or

cycle accidents. The removal of a large element of variation in the driving environment may lead to conclusions that are not valid for analysis of all type of accidents.

Some specific types of accidents (e.g. pedestrian or turning vehicle) that occur frequently in point clusters do not occur in motorways. A high standard of geometric design eliminates several types of accidents in motorways (discussed in Chapter 2). The accidents which contribute to point clusters (e.g. pedestrian or intersection accidents) are largely eliminated and hence the usual pattern of accidents on motorway segments may be line clusters rather than point clusters. This is consistent with the results of the Thomas study.

3.4 Paired-quadrat method

The paired-quadrat method indicates the effect of spacing between quadrat centres on the paired-quadrat-variance. This method is similar to the estimated variogram method (Figure 3.10). Cressie [1993] mentioned that the " paired-quadrat-variance method considers only the effect of spacing between quadrat centres". In this method, random quadrat pairs separated by a lag distance h (the distance between the centres of quadrat pairs) are selected and the number of events within the quadrat are used to estimate the paired-quadrat-variance. The paired-quadrat variance for quadrats separated by h is calculated using Equation 3.3.

Cressie noted, "...under a random arrangement of quadrat counts, the variance estimates are expected to be approximately constant". The paired-quadrat variance for the accident distribution will not give any indication of the effect of the spacing between the quadrat centres. Therefore, this method will not be useful for identifying the spatial dependence (i.e. clusters) in spatial accident data.

3.5 Nearest-neighbour method

This method uses the distances and the directions of nearest neighbours to investigate spatial distribution. Upton and Fingleton [1989] described various techniques, including the methods of Kuiper, Watson and Rayleigh, for analysing nearest neighbour direction. These three tests

and the Kolmogorov-Smirnov test for the nearest-neighbour distance distribution were discussed by Nicholson [1998, 1999].

Nicholson [1999] considered the four spatial distributions shown in Figures 3.13 a, b and Figures 3.14 a, b. In Figure 3.13 a and b, the distribution of the distances from the sample event to neighbouring events are similar but the distribution of directions are different. In Figure 3.14 a and b, the direction distributions are similar but the distance distributions are different. Nicholson further noted, "....the analysis of distances and directions would highlight non-stationarity and anisotropy respectively." Therefore, it is worth analysing both distance and direction of nearest neighbouring events.

Figure 3.15 shows the sample point, sample events, neighbouring events, and the following distances:

1. the distance between sample event and 1st neighbouring event (W_1);
2. the distance between sample event and 2nd neighbouring event (W_2);
3. the distance between sample point and 1st neighbouring event (X_1);
4. the distance between sample point and 2nd neighbouring event (X_2);
5. the distance between event and 1st neighbouring events (Y_1);
6. the distance between event and 2nd neighbouring events (Y_2);
7. the distance between event and 1st neighbouring events (Z_1) in the "half-plane not containing the sample point" and
8. the distance between event and 2nd neighbouring events (Z_2) in the "half-plane not containing the sample point".

Cressie [1993] used some combinations of the distances shown in Figure 3.15 and identified 17 indices of nearest neighbour distance statistics from various sources and mentioned that these indices consider only small-scale neighbours (i.e. first and second nearest neighbour distances). He further stated that to identify larger scale patterns it is necessary to consider higher order neighbours and that the 17 indices of nearest neighbour distance statistics "...cannot be generally recommended for mapped data". Since the number of events in accident clusters are often more than three, these indices are not helpful for accident analysis.

The "K" function described in Ripley [1981], is $K_{(h)}$ equal to the expected number of events within a distance h of a randomly selected event divided by λ, where the λ is the intensity of

the spatial process. If the process is Poisson then $K_{(h)} = \pi h^2$. Nicholson [1995] noted that for "...complete spatial randomness in a continuum the value of $K_{(h)} = \pi h^2$, and if events are spaced evenly then $K_{(h)} < \pi h^2$, and for a cluster distribution $K_{(h)} > \pi h^2$ ".

Nicholson discussed the application of this method for accident analysis. After observing the sample distribution of distances from a sample event to the first N (say) nearest neighbours, it is necessary to test the distribution to see whether it is different from the expected CSR distribution. If the actual distribution is not different from the expected distribution then the spatial process is stationary. The two commonly used tests for comparing distributions are the Kolmogorov-Smirnov test and chi-square test. For small N, the Kolmogorov-Smirnov test is more powerful than the chi-square test. Mood et al. [1974] and Press et al. [1992] showed that the Kolmogorov-Smirnov test can be used for $N \geq 4$. When analysing accident distributions, if N is relatively small, then the Kolmogorov-Smirnov test is suitable.

Nicholson [1995, 1999] mentioned two basic methods for obtaining distributions of distances and directions. In the first method it is necessary to specify the radius of the circle centred on a randomly selected accident and the distributions of distances and directions to neighbouring accident locations within the circle need to be tested. The second method involves specifying the number of neighbours to be considered for each randomly selected accident and distributions of directions and distances to those neighbours need to be tested. Nicholson noted that the advantage of the second method is that it affords better control over the number of observations upon which the distributions of distances and directions are based, and more efficient statistical testing. The disadvantage of this approach is that the distance to the n^{th} nearest neighbour can vary, according to the density of accidents within the region for the period. Each selected sample event with a specified number of nearest neighbours does not have a constant area.

Levine et al. [1995] studied the spatial patterns of motor vehicle accidents in Honolulu in 1990. They used a nearest-neighbour index, which measures the average distance from an event to the nearest event. This index was used to identify clustering or dispersion. This method considers only a small scale pattern (i.e. first nearest-neighbour distances from each event) and neglects any larger scale pattern (i.e. up to n^{th} nearest-neighbour distances). The number of events in the accident clusters is usually more than three. Cressie [1993] did not

recommend this method because to identify larger scale patterns it is necessary to consider higher order neighbours. This method does not consider the possibility of the presence of line cluster. Three other methods used by Levine et al. to identify the spatial concentration of events are: mean centre (i.e. mean latitude and mean longitude), standard distance deviation and "standard deviational ellipse" (i.e. calculate "two standard deviations-one along a transformed axis of maximum concentration and one along an axis which is orthogonal to this"). These methods used by Levine et al. are not useful since the results depend on the space between the roads and are not sufficiently accurate.

3.6 Cluster analysis method

The classical cluster analysis method [Everitt 1974] and [Jain and Dubes 1988] was used to identify clusters by analysing the similarity of events, with the similarity decreasing as the distance between them increases and the dissimilarity of events increasing as the distances between the events increase. Various criteria (e.g. single-linkage, complete-linkage and average-linkages) can be used for identifying clusters. These were discussed by Anujah [1997] and Nicholson [1998], and are explained in Chapter 4.

Nicholson [1998] discussed two plots of dissimilarity coefficient versus the number of clusters, for the single linkage and complete linkage methods (Figures 3.16 and 3.17) where the dissimilarity coefficient was defined as "...the sum (over the clusters) of the distances between accidents within each cluster". The notation used in these figures are nstiso (non-stationary and isotropic distribution, e.g. Figure 3.07), staiso (stationary and isotropic distribution, e.g. Figure 1.03), nstani (non-stationary and anisotropic distribution, e.g. Figure 3.09) and staani (stationary and anisotropic distribution, e.g. Figure 3.08). The figures clearly show different profiles for the regular (staani), CSR (staiso), line cluster (nstani) and point cluster (nstiso) distributions. Nicholson concluded that the profiles may be useful for identifying cluster patterns and the cluster analysis method is investigated further in this thesis.

3.7 Edge effects and corrections

Edge effects will be a problem for both the quadrat method and nearest neighbour method. The selected quadrat must lie within the study area. If an event (accident) is close to the boundary of the selected region, then the nearest neighbours are chosen only within the selected region, although there may be some closer events outside the boundary. In this case, if the nearest neighbour is taken to be the closest event within the region, the nearest neighbour distance will be greater for sample events near the boundary of the region than for events near the centre of the region. Cressie [1993] suggested three general approaches to correct to this:

(1) having a buffer zone along the border and within the study area, with no sample event being selected inside the buffer zone (Figure. 3.18);

(2) assuming that the region is surrounded by eight identical regions, as shown in Figure 3.19 or

(3) calculating a correction factors for statistics or indices.

Nicholson [1999] reviewed the advantages and disadvantages of the three approaches. The first approach is applicable when the study area is not rectangular but only part of the events within the study area (inside the perimeter only, Figure 3.18) are analysed. The second approach is applicable when the study area is rectangular but there is "some reduction in the strength of the linear clustering, due to discontinuity at the boundary" if a line cluster, as shown in Figure 3.20, is not parallel to the boundary. The third approach is not useful because the corrections are related to specific situations and do not apply generally. Hence, one of the first two of the above methods should be selected depending on the circumstances.

3.8 Area identification

Different parts of the study area may require different types of accident reduction plan. If we have the spatial distribution of accidents for a large area, then we might identify sub-areas with spatial distributions which are distinctly different from those in other sub-areas (see Figure 3.20).

Nicholson [1999] suggested that the information (e.g. random distribution of neighbouring event's distances) about the vicinity of each sample event needs to be stored separately. The location of the sample accidents where there is evidence of line clustering could be analysed, to see whether they are close together and continuous. This will indicate one or more sub-areas where route plans would be most appropriate (see Figure 3.20). The vicinities of sample events which are close together and are point clusters are suitable for site plans. The vicinities of sample events which are close together and are regularly or randomly distributed are suitable for area plans.

3.9 Analysis of three statistical techniques

Cressie [1993] and Nicholson [1995] mentioned that the traditional quadrat analysis results can vary with different choices of quadrat size and shape. The result from an arbitrary choice of size and shape of quadrat, which does not consider the relative position of events within the quadrats, may not be reliable. Thus it appears that the quadrat method might be useful only for identifying the level of clustering in an area, but the nearest neighbour method, which can distinguish cluster patterns from CSR, is more powerful for detecting accident patterns. Further analysis of the nearest-neighbour method is given in Chapters 5 and 7.

Nicholson [1995] tested results from a point cluster distribution (with an equal number of events in each clusters) and concluded that:

- the nearest-neighbour distance method indicates strong evidence of clustering when the number of nearest neighbours considered in the analysis just exceeds the cluster size and
- the nearest-neighbour direction method indicates strong evidence of non-uniformity in direction distribution (e.g. line cluster), and the method is also sensitive to the number of nearest neighbours selected and the relative size of the clusters.

The above conclusions indicate that the results from both methods of analysis (i.e. distance and direction) are sensitive to the number of nearest neighbour selections. Nicholson [1995] also noted that the results from different types of spatial distribution suggest that special consideration is necessary when selecting the number of nearest neighbours for analysis.

Anujah [1997] analysed spatial distributions comprising mixtures of point and line clusters with some spatially random accidents, using cluster analysis, quadrat analysis and the nearest-neighbour method. Three basic hypothetical distributions similar to Figure 3.07, 3.09 and 1.03 were generated for the purpose of assessment and the mixing proportions shown in Table 3.02 were used.

Table 3.02: Mixture formed by typical distributions

Mixture	Total number of accidents	Number of accidents from		
		CSR distributions	Point cluster distributions	Line cluster distributions
1	156	100 (64%)	56 (36%)	0
2	156	100 (64%)	0	56 (36%)
3	106	50 (47%)	56 (53%)	0
4	106	50 (47%)	0	56 (53%)

Anujah arrived at the following conclusions.

(1) The single linkage method is not sensitive to patterns at the initial stage of cluster forming but it becomes sensitive afterwards, so the process should be continued to the end (i.e. until all events become a single cluster) to determine whether line, point cluster or CSR distribution exist.

(2) The result from the single linkage method shows that a mixture of a line cluster distribution and a CSR distribution indicate the existence of line cluster irrespective of a higher (53%) or a lower (36%) percentage of line cluster in the mixture.

(3) The nearest-neighbour method indicates clear evidence of a line cluster when the proportion of line cluster is higher (53%) than the proportion of CSR distribution in the mixture.

(4) The single linkage method does not indicate the existence of line cluster from the mixture of CSR distribution with a higher percentage (64%) of point cluster, but the nearest neighbour method indicates the existence of line cluster.

(5) Considerable judgment is needed and the analysis is time consuming for the nearest neighbour and cluster analysis methods.

(6) Although the quadrat method is easier, it gives satisfactory results only for the basic distributions. The results from the quadrat method indicate the existence of point clusters when the mixture is 64% CSR and 36% point cluster.

Nicholson's and Anujah's preliminary conclusions indicates that the three statistical techniques are helpful for analysing spatial distributions of accidents. Further research is necessary to distinguish between the different types of spatial distribution, like mixtures of the basic distributions (e.g. point clusters, line clusters and CSR distributions) and accident distribution in a sparse road network.

3.10 Geographic Information Systems

It is recognised that the analysis of spatial distributions of accidents is a task that can be performed using a Geographical Information System (GIS). Fischer [1996] noted that a GIS incorporates many features, such as relational data base management, graphical algorithms, interpolation, zoning and simplified network analysis (termed spatial analysis and modelling). Fischer noted the lack of spatial analytical techniques in GIS, which need to be strengthened in exploratory spatial data analysis, including the "search for data characteristics such as trend, spatial patterns and associations".

In recent years, many studies using GIS in accident analysis have been reported (eg. LaScala [2000] , Austin [1995], Goh [1993] and Peled and Hakkert [1993]). These studies used a GIS to map locations for data validation or data corrections. If a simple numerical index indicating the level of clustering (i.e. departure from CSR distribution) is calculated with assumptions (i.e. accident data in a continuum, homogeneous data sampled from the complex road network) then verification of numerical indices may be needed for using the plot of accident locations.

Austin [1995] used GIS with the aim of removing the errors that exist in accident reports. The GIS database contains location features such as road class, road number, district, speed

limit, pedestrian crossing facilities, junction control, junction details and carriageway type and markings. The accident record were checked for error in the locational variables and the data were corrected. Analysis with incorrectly entered data will lead to misinterpretation of spatial pattern. The data should be checked for validity before any type of statistical analysis.

In some situations we need to retrieve detail of the various attributes of the area (e.g. percentage of area where there are no roads, like sea or lake area) for interpreting accident analysis results. To summarise the results some simple statistics, graphs, histograms, scatter plots and box plots are linked with GIS technology. Fotheringham [1994] noted the slow progress in some simple statistical graphics linked with GIS. Fotheringham further noted that some of the most commonly used commercial GIS systems such as ARC/INFO, SPANS, or GENAMAP, offer little support for basic statistical summarization.

To view simple statistical summaries and plots of data at the same time the data plotted graphically in the form of map is of great value. GIS technology is also useful for checking the validity of data on the location of accidents. Therefore, for accident analysis, GIS with improved spatial statistical analysis functions would be useful.

Spatial data analysis is a relatively new and rapidly growing area of research. The growth is largely prompted by increasing use of GIS. To date, very few road safety studies using spatial data analysis methods have been done. In my research, different forms of accident plots were plotted on road maps and investigated and the crash analysis system [LTSA, 2000] was used to retrieve and plot the accident data, and this was analysed.

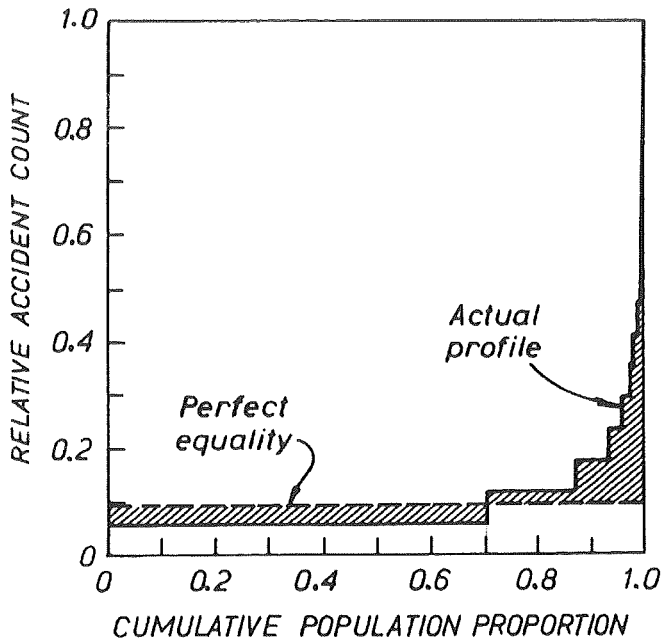


Figure 3.01: Accident count profile

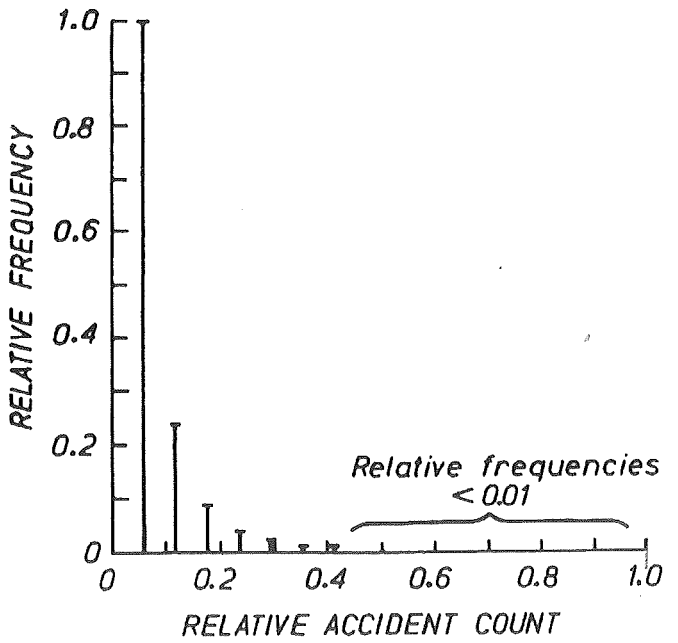


Figure 3.02: Accident count frequency distribution

[Figures 3.01 and 3.02 extracted from Nicholson [1989]]

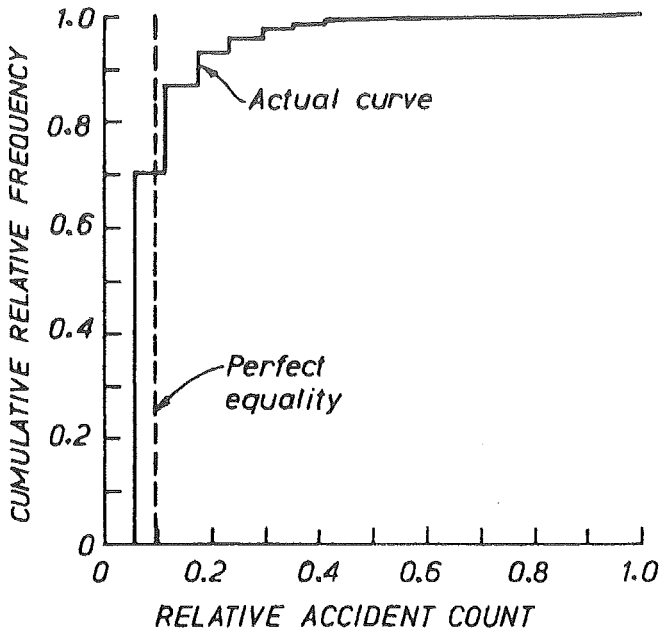


Figure 3.03: Accident count cumulative frequency distribution

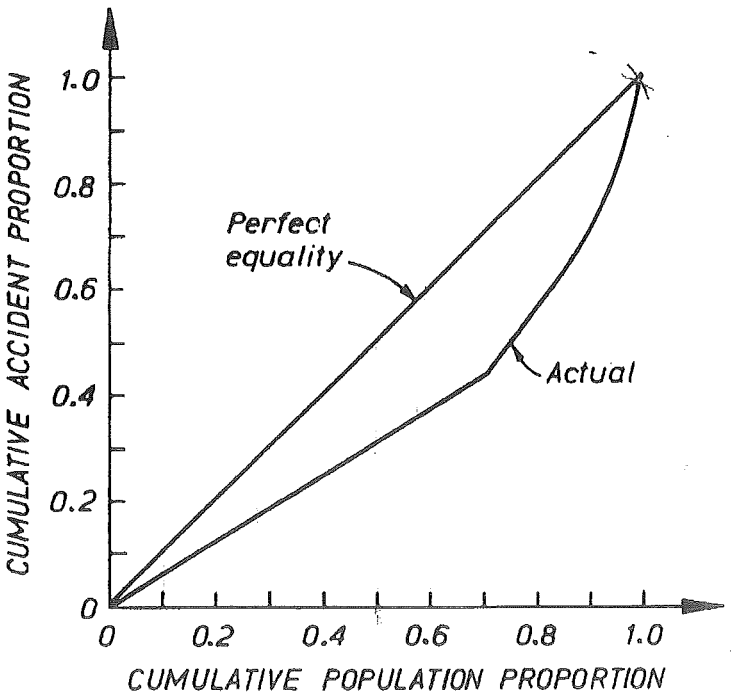


Figure 3.04: Accident count concentration curve

[Figures 3.03 and 3.04 extracted from Nicholson [1989]]

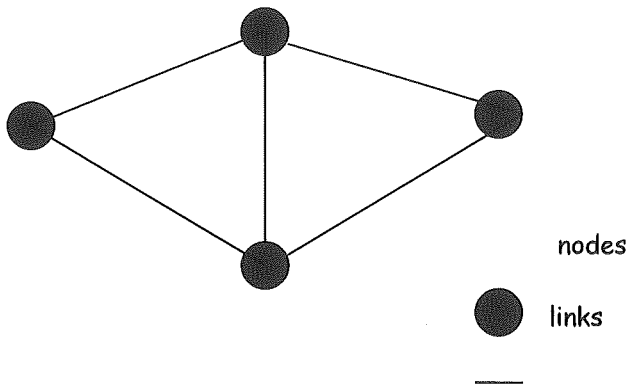


Fig 3.05: A simple road network

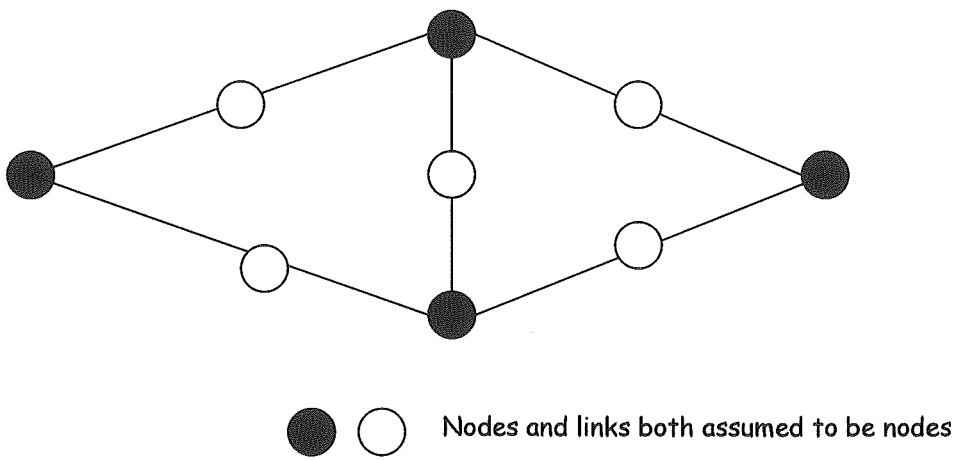


Fig 3.06: Abstract road network

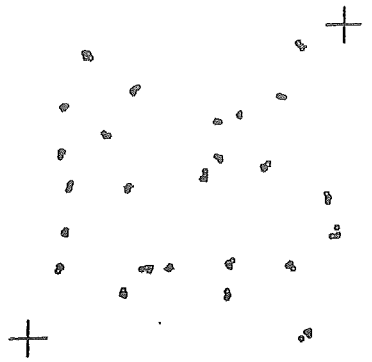


Figure 3.07: A non-stationary and isotropic distribution

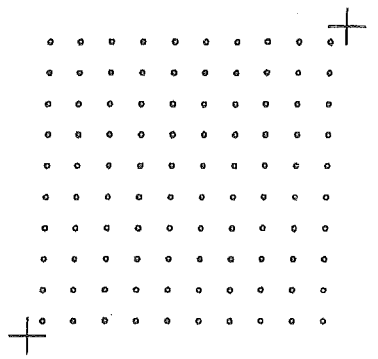


Figure 3.08: A stationary and anisotropic distribution

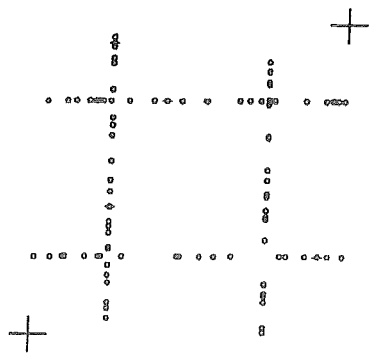
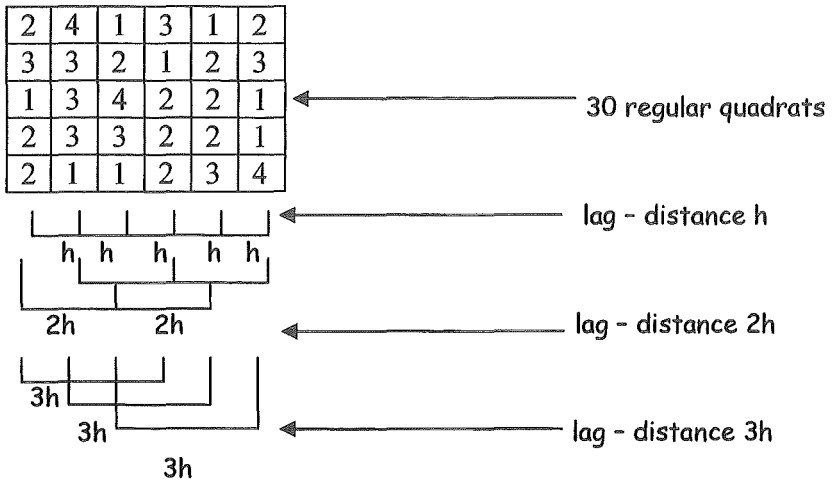


Figure 3.09: A non-stationary and anisotropic distribution

[Figures and 3.07, 308 and 3.09 extracted from Nicholson [1999]]



$N_{(h)} = 25$
$N_{(2h)} = 20$
$N_{(3h)} = 15$

$$2\gamma(h) = 1/25\{(2-1)^2+(1-1)^2+(1-2)^2+(2-3)^2 +(3-4)^2+(2-3)^2+(3-3)^2+(3-2)^2 +(2-2)^2+(2-1)^2+(1-3)^2+(3-4)^2 +(4-2)^2+(2-2)^2+(2-1)^2+(3-3)^2 +(3-2)^2+(2-1)^2+(1-2)^2+(2-3)^2 +(2-4)^2+(4-1)^2+(1-3)^2+(3-1)^2+(1-2)^2\} = 1.72$$

$$2\gamma(2h) = 1/20\{(2-1)^2+(1-3)^2+(1-2)^2+(2-4)^2 +(2-3)^2+(3-2)^2+(3-2)^2+(2-1)^2 +(1-4)^2+(4-2)^2+(3-2)^2+(2-1)^2 +(3-2)^2+(2-2)^2+(3-1)^2+(1-3)^2 +(2-1)^2+(1-1)^2+(4-3)^2+(3-2)1^2\} = 2.05$$

$$2\gamma(3h) = 1/15\{(2-2)^2+(1-3)^2+(1-4)^2 +(2-2)^2+(3-2)^2+(3-1)^2 +(1-2)^2+(3-2)^2+(4-1)^2 +(3-1)^2+(3-2)^2+(2-3)^2 +(2-3)^2+(4-1)^2+(1-2)^2\} = 3.1$$

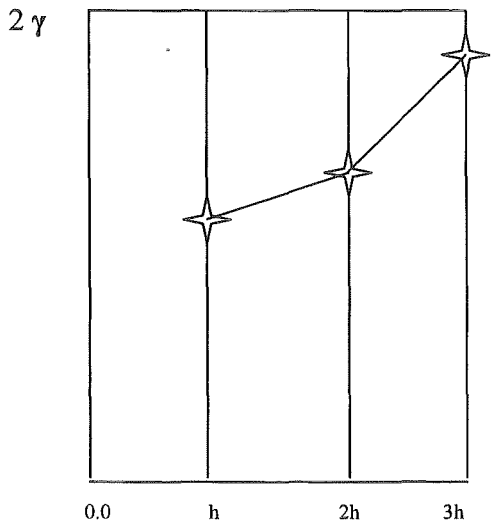
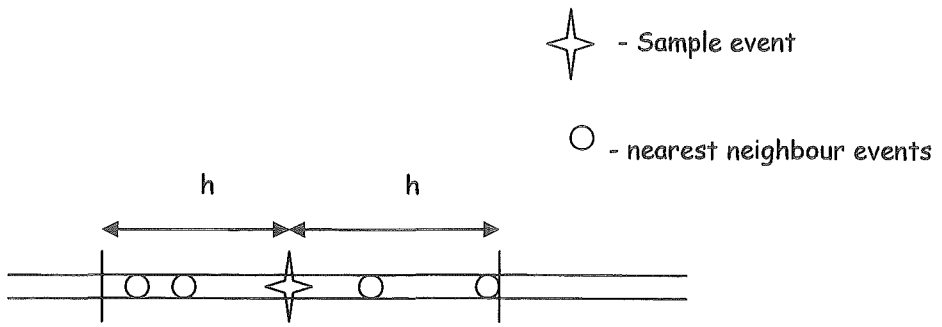
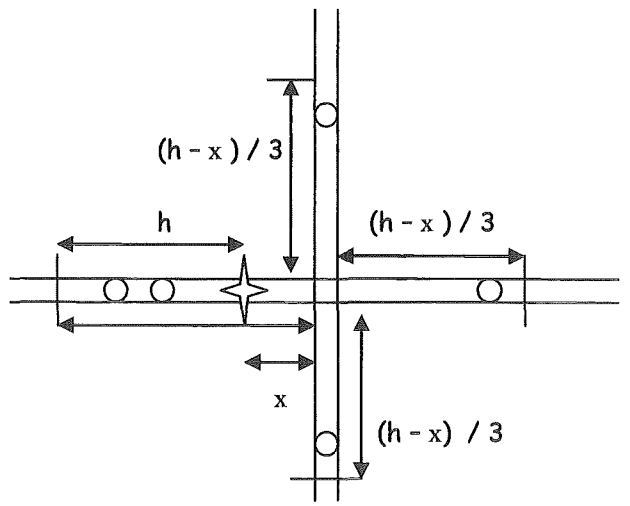


Fig 3.10: Example of variogram estimation

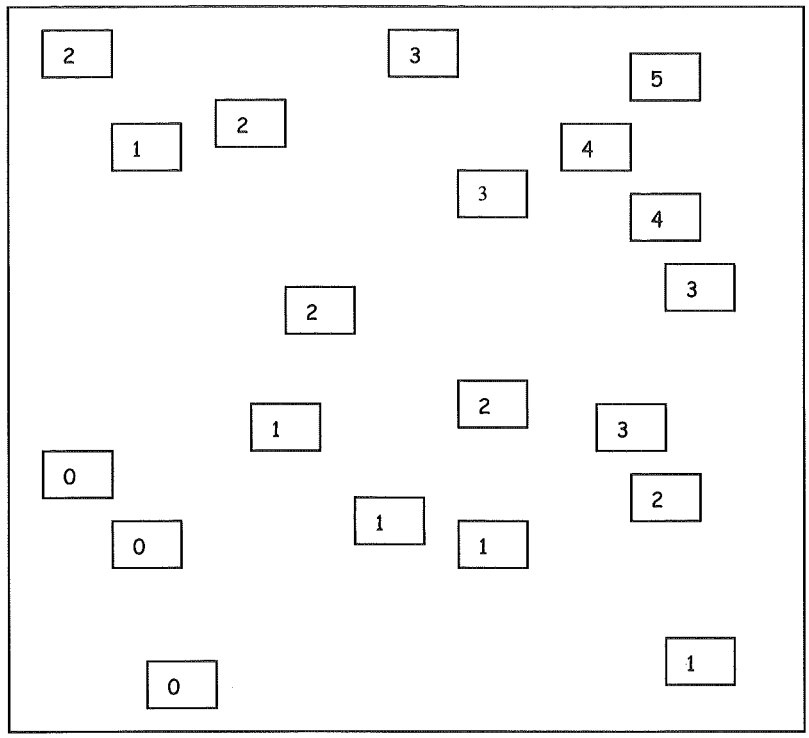


(a) Sample event far from intersection



(b) Sample event near an intersection

Fig 3.11: Nearest-neighbour events within distance h from a sample event.



Number	Frequency
0	3
1	5
2	5
3	4
4	2
5	1

Figure 3.12a: Random quadrats

[Extracted from Dale [1999]]

0	0	0	1	1	2	2	1	0	0
0	1	1	2	2	3	2	2	1	0
1	1	2	3	3	4	3	3	2	1
1	1	2	3	4	5	4	3	3	2
1	2	3	4	5	6	5	4	3	2
1	2	3	4	4	5	4	3	3	2
1	1	2	3	3	4	3	3	2	2
1	1	2	2	2	3	2	2	2	1
0	1	1	2	2	2	2	1	1	0
0	0	1	1	1	1	1	0	0	0

Number	Frequency
0	14
1	27
2	27
3	18
4	9
5	4
6	1

Figure 3.12b: Regular quadrats

[Extracted from Dale [1999]]

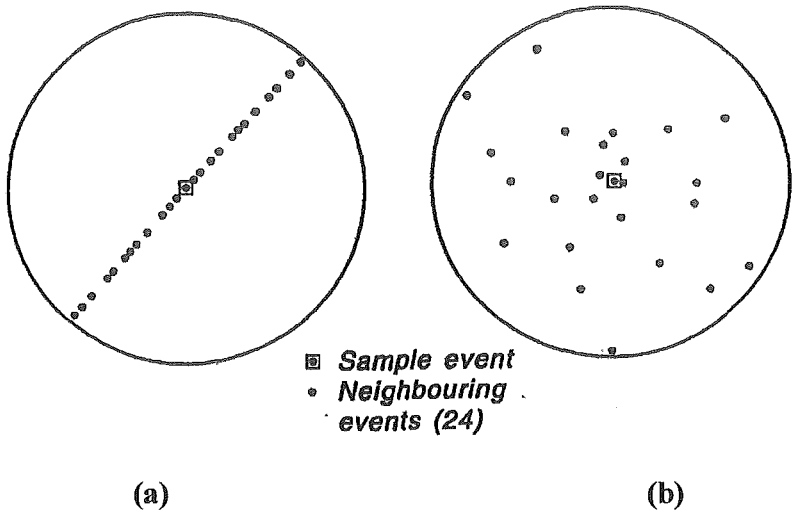


Figure 3.13: Similar distance distribution

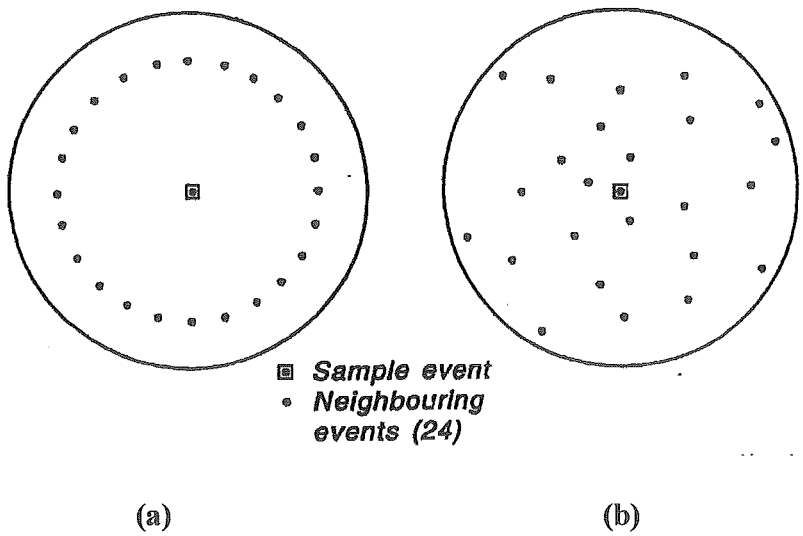


Figure 3.14: Similar direction distribution

[Figures 3.13 and 3.14 copied from Nicholson [1999]]

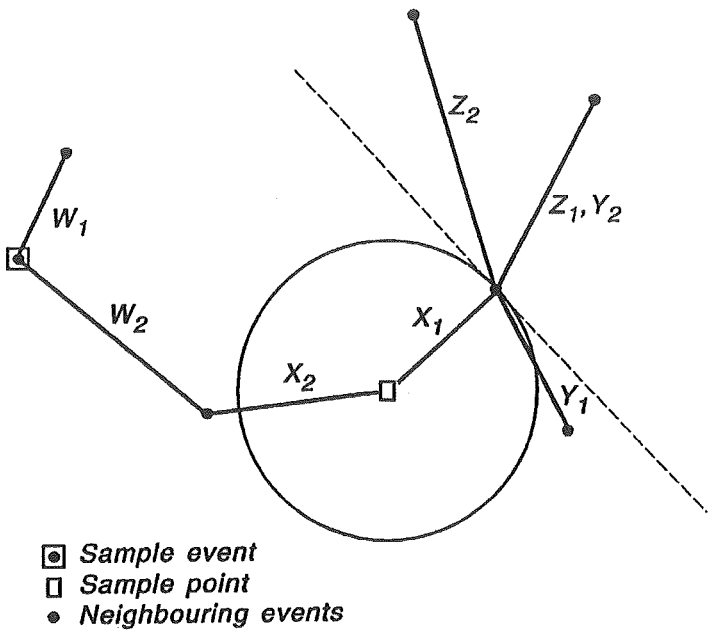


Fig 3.15: Types of nearest-neighbour distances
[Extracted from Nicholson [1999]]

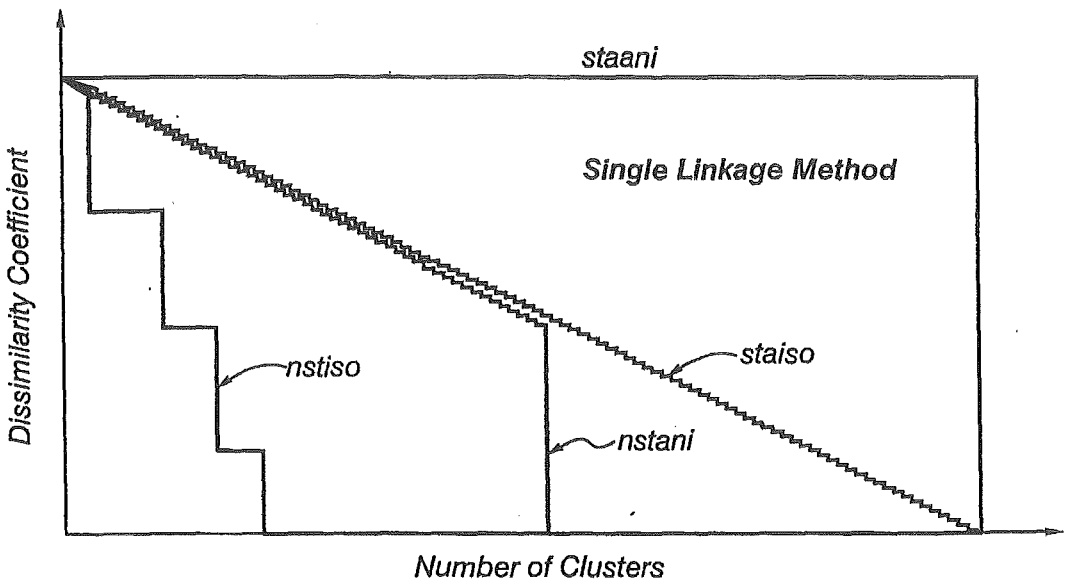


Figure 3.16: Dissimilarity coefficient profile (single-linkage method)

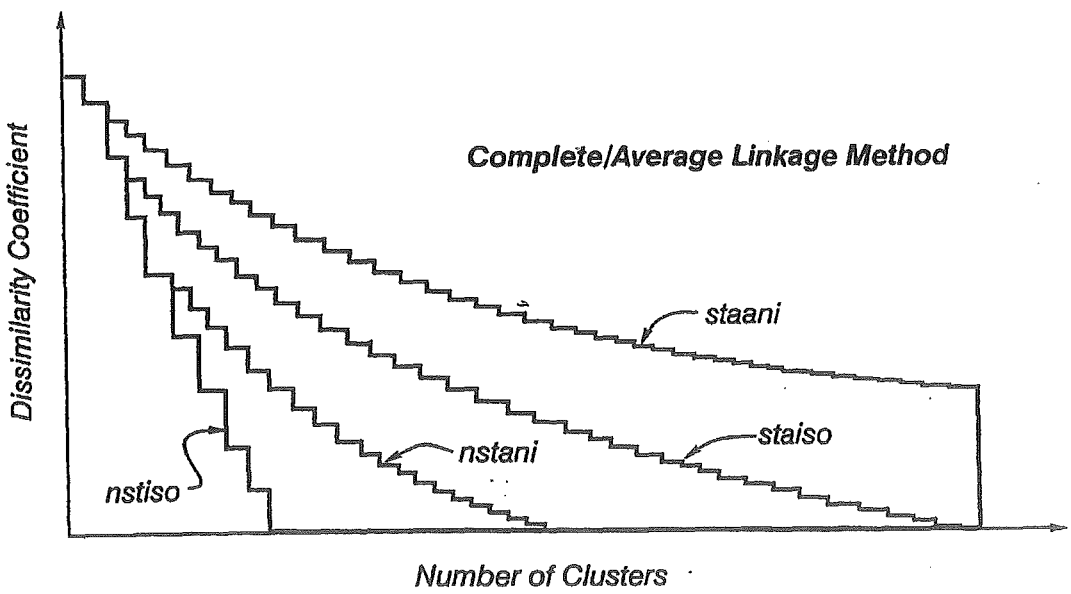


Figure 3.17: Dissimilarity coefficient profile (complete and average linkage methods)

[Figures extracted from Nicholson [1998]]

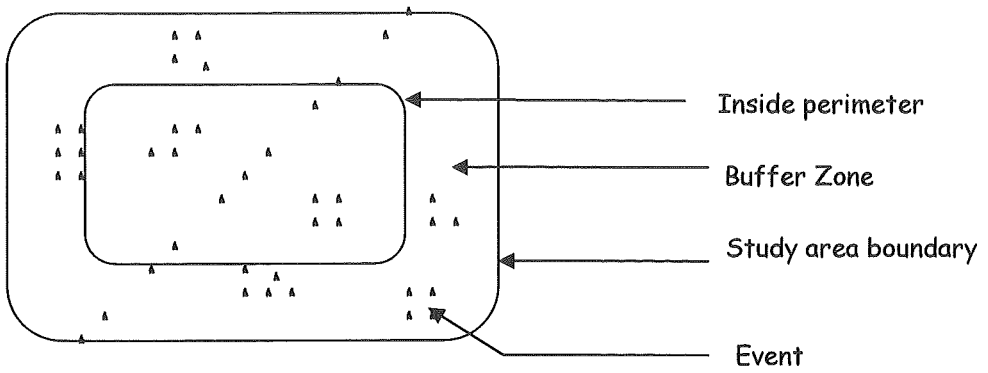


Fig 3.18: Buffer zone shown in the study area

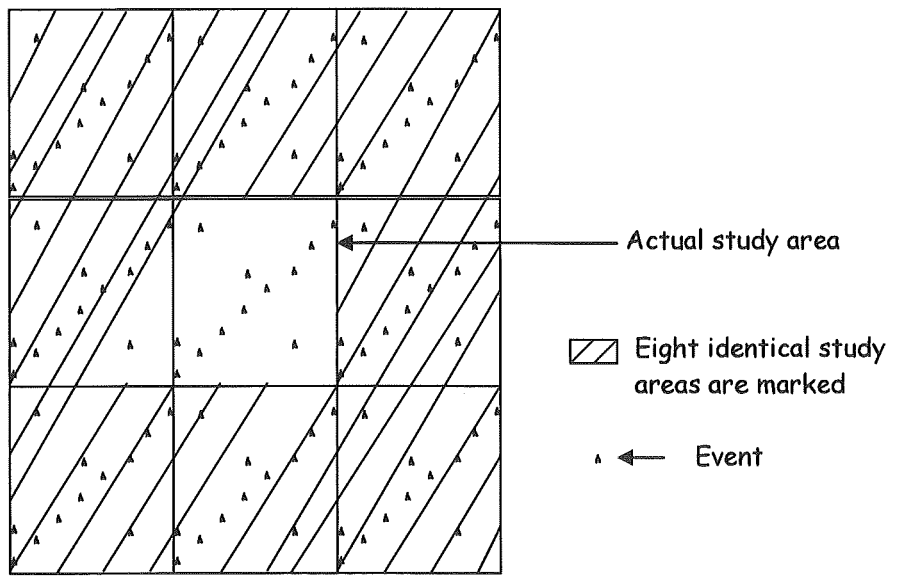
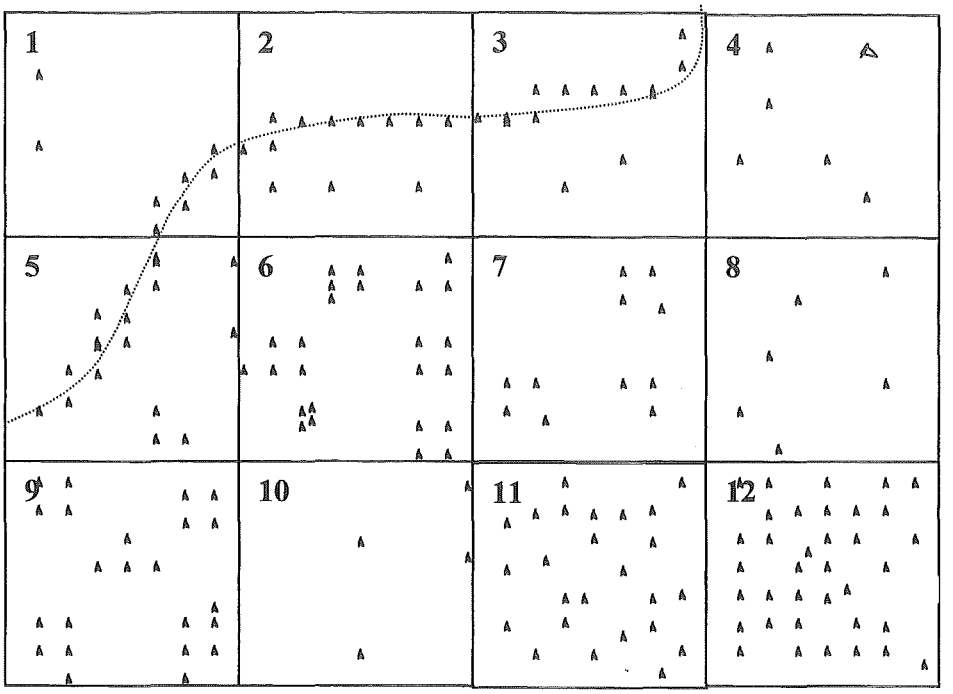


Fig 3.19: Study area surrounded by eight identical study areas
[Figure extracted from Nicholson [1999]]



Accident location

Area clusters: sub areas 11 and 12
Line clusters: sub areas 5, 2, 3 and 4
Point clusters: sub areas 6,7 and 9

Fig 3.20 Accident patterns in sub-areas

Chapter 4

CLUSTER ANALYSIS

4.1 Introduction

Cluster analysis can be used for studying and understanding the structure in data on the characteristics of objects, by grouping the characteristics in a sensible way according to their level of similarity or dissimilarity. The selection of the characteristics of objects should be based on the aim of the analysis. Various techniques for grouping the characteristics of data have been employed in different disciplines such as sociology, biology, medical sciences, market research, archaeology and psychology. For analysing accident clustering, an important attribute is the distance between the positions of events.

Visual examination may be used in cluster analysis. Two aspects to be addressed are the time involved in processing and the consistency of results obtained by different individuals. Cluster analysis techniques have several advantages over visual examination. A program based on such techniques can focus consistently on the purpose to form groups. Humans may identify clusters which have very distinct differences in two-dimensional space, but if the differences are not very distinct, the results obtained may differ from person to person. Results may be influenced by previous cluster images in the human memory, especially if the group of objects are not well separated. The total processing time for grouping may depend on the number of elements in the data. The same person may form different groups within different time frames to process the same data, when the number of data items is very large (more than about one thousand) and the groups are not well separated. One of the governing factors is the skill in visual examination. Clusters from the data identified by different individuals may not be the same, because judgement of the similarity among the objects varies among different individuals, especially when the groups are not well separated. However a computer program based on cluster analysis techniques can form consistent groups in a fraction of the time required by a visual examination process. Therefore in general, the reliability, speed, and consistency of a cluster analysis algorithm is far better than for a visual examination process, and does not need as much judgmental skill.

To identify the measure of similarity, it is necessary to decide which characters or features are the important attributes of the objects. There may be several attributes that could be considered, with a numerical score for each attribute. The similarity measures could be evaluated by averaging the scores, or from a single important attribute.

For the objective of this study (to find a spatial pattern in accident data), only the locations of accidents are considered. Therefore the distance between accidents is considered as the single important attribute, which is used for grouping the accidents. This chapter is focused on the methods used in cluster analysis to identify the spatial patterns of accidents.

4.2 Hierarchical clustering techniques

Objects can be classified into two types of clusters, namely non-exclusive (over-lapping clusters) and exclusive (distinct clusters). The exclusive cluster method can be further categorised as extrinsic when the objects have category labels, and as intrinsic when we need to judge how the object can be assigned during the clustering process and the objects are not group-labelled. Accident data are assumed to be exclusive and intrinsic, and accident data analysis is best done using an exclusive and intrinsic cluster method.

Hierarchical and partitional methods are able to be used to analyse intrinsically classified objects. Hierarchical clustering techniques involve rearranging the data into a nested pattern of clusters, whereas a partitional clustering method involves rearranging the data in a single partition step. This thesis is focused on hierarchical cluster techniques because there is provision for identifying line and point clusters.

The hierarchical classification techniques may be sub-divided into two types; divisive and agglomerative methods. In the divisive method, the procedure begins with a single cluster, which contains all items C ≡ {C1, C2, C3, ----- Cn} and divides them progressively until the clusters cannot be divided further. Finally, we would have a cluster set of items such as {C1}, {C2}, {C3}, -----{Cn} and the original set. The agglomerative method is the opposite of the divisive method, and combines n items such as {C1}, {C2}, {C3}, -----{Cn} into a single cluster C ≡ {C1, C2, C3, ----- Cn} in a series of steps. The agglomerative and divisive

methods are illustrated in Figure 4.01. This diagram illustrates the joined or disjoined clusters at each successive stage. This diagram is known as a dendrogram and is a graphical way of representing the structure in the data. Both the hierarchical classification methods are aimed at finding the structure in the data.

4.3 Agglomerative cluster analysis

The agglomerative method is better for accident analysis because accident data relate to individual accident positions. The four main agglomerative cluster analysis methods are the single-linkage, complete-linkage, group average and Ward's methods. The theories underlying the single-linkage and the complete-linkage methods are explained in Section 4.3.1 and 4.3.2 respectively. The applications of these two methods are explained with an example in Section 4.3.3. A special property of the single linkage method (i.e. the chaining process) is explained in Section 4.3.4. The sensitivity to accident location variation is discussed in Section 4.3.5. The theories underlying the group average and Ward's methods are explained in Section 4.3.6.

4.3.1 *Single-linkage*

In this method the clusters are joined and merged to form a new cluster by considering the single nearest-neighbour, as shown in Figure 4.02. For the above reason the single linkage method has a tendency to form long thin clusters, called 'chain clusters'. The clusters A and B will be combined first because L_{min} (AB) < L_{min}(BC) < L_{min} (AC).

In some publications this method is referred to as the nearest- neighbour method. In this thesis another method is referred to as the nearest- neighbour method, so to avoid confusion this technique is referred to as the single-linkage method.

The four important items of information for accident analysis that can be obtained or identified from the chained clusters are:

1. the number of chain clusters;
2. the length of each chain cluster;

3. the intensity (number of accidents / length) of each chain cluster;
4. the part with greatest intensity for each chain cluster.

The first item is useful for finding the number of roads or sections of roads which have a reasonably high intensity of accident occurrence. The second and third items are useful for identifying the fourth information item. The fourth item enables identification of the higher intensity parts of the high-intensity sections (chain clusters). The investigators can monitor how the accident intensity in a particular road is growing and to see whether there are any changes in the chain clusters over time. Further investigation of these properties is not necessary unless it is useful for identifying any spatial pattern.

4.3.2 *Complete-linkage*

While the single-linkage method merges the clusters with the smallest minimum distance between their members, the complete-linkage method merges the clusters with the smallest maximum distance between their members. An example of cluster formation using the complete-linkage method is shown in Figure 4.03. The clusters A and B will be combined first because L_{max} (AB) < L_{max}(BC) < L_{max} (AC).

When considering the neighbours for the clustering process, in this method the completeness of the two clusters is examined, but in the single linkage method only the nearest two members from each cluster are considered. In practice, the distance between the far ends of two small chain clusters will not be the closest distance and for this reason the complete-linkage does not have a tendency to create a long thin cluster.

4.3.3 *Application of single-linkage and complete-linkage method*

The three basic steps are explained for the spatial distribution which shows the locations of five accidents, as shown in Figure 4.04.

STEP 1: Obtain data matrix

The locations of those accidents are shown in the following matrix, referred to as the data matrix. The numbers shown in bold font are the elements of the data matrix.

Accident sequence number	1	2	3	4	5
X co-ordinates	4	5	7	2	3
Y co-ordinates	3	2	3	5	8

Data matrix for Example I

STEP 2: Compute the proximity matrix

The city-block distance or Euclidean distance can be used to compute the proximity matrix. Anujah [1997] and Nicholson [1999] have shown that the cluster analysis results using the city-block distance or Euclidean distance are the same. In this thesis Euclidean distance is used. The distance between two accident locations is known as the distance coefficient. For example the distance between the accidents 1 and 2 is denoted by d_{12}, and is

$$d_{12} = \sqrt{((4-5)^2 + (3-2)^2)} = 1.4$$

The proximity matrix (D), containing the distance coefficients (distances shown below in bold font) is as follows:

Accident sequence number	1	2	3	4	5
1	**0**	**1.4**	**3.0**	**2.8**	**5.1**
2	-	**0**	**2.2**	**4.2**	**6.3**
3	-	-	**0**	**5.4**	**6.4**
4	-	-	-	**0**	**3.2**
5	-	-	-	-	**0**

Proximity matrix D = [d (i,j)] for Example I.

In this example, proximity matrix and distance are equal.

STEP 3: Execute the clustering method

In this example the agglomerative hierarchical method is used. The clustering method starts with each object regarded as a single separate cluster and in the example we begin with five clusters. There are a series of steps in the agglomerative hierarchical method. In each clustering step, the most similar objects (in our case the closest accident locations)

will be merged to form a single cluster and hence the existing number of clusters is reduced by one. This procedure is repeated until one has a single cluster, which contains all five accident locations from the five single accident clusters.

The following notations are used in the proximity matrix updating algorithm:

D = [d(i,j)] is the n x n proximity matrix.

$L_{(k)}$ is the level of the k^{th} clustering, where the clustering is assigned in a sequential order 0,1,2, 3,….,k,….,(n-1).

d [(r),(s)] is the proximity between cluster (r) and cluster (s).

The five-step algorithm given by Jain & Dubes [1988:73] is as follows:

(1) begin with the disjoint clustering having level L(0) and sequence number m = 0;

(2) find the least dissimilar pair of clusters in the current clustering, say pair {(r), (s)}, according to

$$d[(r),(s)] = \min\{d[(i),(j)]\}$$

where the minimum is over all pairs of clusters in the current clustering;

(3) increase the sequence number: m to m+1 and merge the clusters (r) and (s) into a single cluster to form the next clustering m+1 set the level of this clustering to

$$L(m+1) = d[(r),(s)];$$

(4) update the proximity matrix, D, by deleting the row and columns corresponding to cluster (r) and (s) and adding a row and column corresponding to the newly formed cluster; the proximity between the new cluster, denoted (r , s) and the old cluster (k) is defined as:

for the single link method,

$$d[(k),(r,s)] = \min\{d[(k),(r)], d[(k),(s)]\}$$

for the complete-link method,

$$d[(k),(r,s)] = \max\{d[(k),(r)], d[(k),(s)]\}$$

(5) if all objects are in one cluster, stop or else go to the 2nd step of the algorithm.

The above algorithm for clustering objects is applied to the example. The proximity matrix is rearranged step-by-step. The single-linkage and complete-linkage methods are shown one by one and at the end of each method the dendrogram is shown.

Single-linkage method applied to Example I

The proximity matrix is:

	1	2	3	4	5
1	**0**	**1.4**	**3.0**	**2.8**	**5.1**
2	-	**0**	**2.2**	**4.2**	**6.3**
3	-	-	**0**	**5.4**	**6.4**
4	-	-	-	**0**	**3.2**
5	-	-	-	-	**0**

L(1) =minimum distance in the whole set (underlined above) =

d(1,2) = 1.4

Hence, L(1) = 1.4 and {1,2} becomes a cluster

d[(3),(1,2)] = min(d(1,3), d(2,3)) = min(3, 2.2) = d(2,3) = 2.2

d[(4),(1,2)] = min(d(1,4), d(2,4)) = min(2.8, 4.2) = d(1,4) = 2.8

d[(5),(1,2)] = min(d(1,5), d(2,5)) = min(5.1, 6.3) = d(1,5) = 5.1

The 1st row of the proximity matrix is now changed because of the new cluster (labelled as 12), and results in the updated proximity matrix

	12	3	4	5
12	**0**	**2.2**	**2.8**	**5.1**
3		**0**	**5.4**	**6.4**
4			**0**	**3.2**
5				**0**

The minimum distance in the above matrix is 2.2 (the distance between cluster 12 and cluster 3). Hence, L(2) = 2.2 and {1,2,3} becomes a cluster, and the updated proximity matrix is

	123	4	5
123	**0**	**2.8**	**5.1**
4		**0**	**3.2**
5			**0**

The minimum distance in the above matrix is 2.8 (the distance between cluster 123 and cluster 4). Hence, L(3) = 2.8 and {1,2,3,4}becomes a cluster, and the updated proximity matrix is

	1234	5
1234	**0**	**3.2**
5		**0**

Hence L(4) = 3.2 and {1,2,3,4,5} becomes a cluster. The dendrogram showing the agglomeration process is shown in Figure 4.05.

Complete-linkage method applied to Example I

The proximity matrix for example I (see page 87) is used.

L(1) =min distance in the whole set = d(1,2) =1.4

Hence, L(1) = 1.4 and {1,2}becomes a cluster

d[(3),(1,2)] = max(d(1,3), d(2,3)) = max(3.0, 2.2) = d(1,3) = 3.0

d[(4),(1,2)] = max(d(1,4), d(2,4)) = max(2.8, 4.2) = d(2,4) = 4.2

d[(5),(1,2)] = max(d(1,5), d(2,5)) = max(5.1, 6.3) = d(2,5) = 6.3

The 1st row of the proximity matrix is changed because of the new cluster (labelled as 12), which results in the updated proximity matrix

	12	3	4	5
12	**0**	**3.0**	**4.2**	**6.3**
3		**0**	**5.4**	**6.4**
4			**0**	**3.2**
5				**0**

The minimum distance in the above matrix is 3 (the distance between cluster 12 and cluster 3). Hence, L(2) =3 and {1,2,3} becomes a cluster, and the updated proximity matrix is

	123	4	5
123	**0**	**5.4**	**6.4**
4		**0**	**3.2**
5			**0**

The minimum distance in the above matrix is 3.2 (the distance between cluster 4 and cluster 5). Hence, L(3) =3.2 and {4,5} becomes a cluster, and the updated proximity matrix is

	123	45
123	**0**	**6.4**
45		**0**

Hence L(4)=6.4 and {(1,2,3),(4,5)} becomes a cluster. The dendogram showing the agglomeration process in shown in Figure 4.06.

Figures 4.05 and 4.06 indicate the difference in the agglomeration order. With the single-linkage method, the first two members (1,2) are joined, the result (12) joined with the third (3) and so on. With the complete-linkage method two separate clusters (123 and 45) are formed and then the two clusters are merged together.

4.3.4 *Single-linkage method and chaining*

Understanding the chaining cluster process in the single-linkage method is useful for identifying the spatial pattern of accidents. In the first step of the clustering process, closest locations are joined to form a cluster. Then the next nearest location is joined with the previously formed cluster and this process is repeated till a single cluster is formed. So, the members of the first cluster increase progressively in each step and finally form a long thin cluster. Figure 4.05 illustrates a simple chaining process.

A partial chaining process is also possible. This occurs if two or more chains are formed separately and finally joined together to become a long chain. In this process, two or more chains can be formed simultaneously at different locations, and may finally become a long chain with some branches. The simultaneous formation of thin clusters at different locations is called 'partial chaining', which is described in Section 4.3.5. The advantages and disadvantages of the single linkage method compared with other methods are explained in Section 4.3.6.

4.3.5 *Sensitivity to accident location variation*

Consider another example, shown in Figure 4.07. Here the location of one accident (accident 1) is slightly different from that shown in Figure 4.04.

The data matrix is:

Accident sequence number	1	2	3	4	5
X co-ordinates	**5**	**5**	**7**	**2**	**3**
Y co-ordinates	**3**	**2**	**3**	**5**	**8**

The proximity matrix is:

Accident sequence number	1	2	3	4	5
1	**0**	**1.0**	**2.0**	**3.6**	**5.4**
2	-	**0**	**2.2**	**4.2**	**6.3**
3	-	-	**0**	**5.4**	**6.4**
4	-	-	-	**0**	**3.2**
5	-	-	-	-	**0**

This proximity matrix will be analysed by using single-linkage and complete-linkage methods.

Single-linkage method applied to Example II

In the above matrix,

L(1) =min distance in the whole set = d(1,2) =1.0

Hence, L(1) = 1 and {1,2} becomes a cluster

d[(3),(1,2)] = min(d(1,3), d(2,3)) = min(2.0, 2.2) = d(1,3) = 2.0

d[(4),(1,2)] = min(d(1,4), d(2,4)) = min(3.6, 4.2) = d(1,4) = 3.6

d[(5),(1,2)] = min(d(1,5), d(2,5)) = min(5.4, 6.3) = d(1,5) = 5.4

The new proximity matrix is:

	12	3	4	5
12	**0**	**2.0**	**3.6**	**5.4**
3		**0**	**5.4**	**6.4**
4			**0**	**3.2**
5				**0**

Change to the proximity matrix occurs only in the1st row. The minimum distance is 2.0 (the distance between cluster 12 and cluster 3). Hence, L(2) =2 and {1,2,3} becomes a cluster, and the updated proximity matrix is

	123	4	5
123	**0**	**3.6**	**5.4**
4		**0**	**3.2**
5			**0**

The minimum distance is 3.2 (the distance between cluster 4 and cluster 5). Hence, L(3) =3.2 and {4,5} becomes a cluster, and the updated proximity matrix is

	123	45
123	**0**	**3.6**
45		**0**

The minimum distance is 3.6 (the distance between cluster 123 and cluster 45). Hence, L(4) = 3.6 and {1,2,3,4,5} becomes a cluster. The dendrogram showing the agglomeration process is shown in Figure 4.08. The agglomeration order shown in Figure 4.08 is different from the agglomeration order shown in Figure 4.05.

The dendrogram for partial chaining (Figure 4.08) is similar to the dendrogram for point clusters. The dendrogram for simple chaining (Figure 4.05) is different from the dendrogram shown in Figure 4.08. Partial chaining, which involves separate branches of the tree of the dendrogram, occurs for point clusters. That is, partial chaining can occur for both point and line clusters, and causes a practical difficulty in differentiating between point clusters and line clusters.

Complete-linkage method applied to Example II

The proximity matrix for Example II (shown in Figure 4.07) is used.

L(1) = min distance in the whole set = d(1,2) = 1.0

Hence, L(1) = 1 and {1,2} becomes a cluster

d[(3),(1,2)] = max(d(1,3), d(2,3)) = max(2.0, 2.2) = d(2,3) = 2.2

d[(4),(1,2)] = max(d(1,4), d(2,4)) = max(3.6, 4.2) = d(2,4) = 4.2

d[(5),(1,2)] = max(d(1,5), d(2,5)) = max(5.4, 6.3) = d(2,5) = 6.3

Change occurs to the proximity matrix only in the 1st row. The new proximity matrix is

	12	3	4	5
12	**0**	**2.2**	**4.2**	**6.3**
3		**0**	**5.4**	**6.4**
4			**0**	**3.2**
5				**0**

The minimum distance is 2.2 (the distance between cluster 12 and cluster 3). Hence, L(2) =2.2 and {1,2,3} becomes a cluster, and the updated proximity matrix is

	123	4	5
123	**0**	**5.4**	**6.4**
4		**0**	**3.2**
5			**0**

The minimum distance is 3.2 (the distance between cluster 4 and cluster 5). Hence, L(3) =3.2 and {4,5} becomes a cluster, and the updated proximity matrix is

	123	45
123	**0**	**6.4**
45		**0**

L(3) = 3.6 and {1,2,3,4,5} becomes a cluster. The dendrogram showing the agglomeration process is shown in Figure 4.09.

Two clusters are clearly identified using the complete-linkage and single-linkage dendrograms at the level of 3.6. In Example II both methods yield very similar results. If there are two distinct clusters, then both the single-linkage and complete-linkage methods distinguish the clusters. In the single-linkage method each cluster is identified from the dendrogram when the shortest distance between the events in the two clusters is sufficiently large. In the complete-linkage method to join two clusters the longest distance between events in the two clusters is considered. The largest distance is not changed in the two examples, and therefore the branches of the tree (i.e. dendrogram) do not change.

The different dendrogram structures in Figure 4.05 and 4.06 show that the single-linkage method can be sensitive to small changes in the spatial distribution and raises the issue of sensitivity of the methods when the data contains "noise". "Noise" means the uncertainty or error in fixing the location of events (it is well known that the Police sometimes get the location wrong or estimate the location approximately). The only difference in the two examples is the change in location of accident number 1 and this is the reason for the results from the single-linkage method being completely different (i.e. members are joined at a different "level", and the structure of the tree is different). There is no difference in the complete-linkage results (i.e. L(3) = 6.4, and the dendrogram structure is the same for each example). The dendrogram for the single-linkage method (applied to Example I) indicates one long cluster, and the dendrogram for the complete-linkage method (applied to Example I) indicates two separate clusters. Therefore, there is a possibility that the hierarchical cluster structure can change dramatically with a small change in the data like an error when entering the accident locations. Everitt [1993] noted that "Baker (1974) and Hubert (1974) both produce evidence that complete linkage clustering is less sensitive to particular types of observational errors than single linkage".

4.3.6 *Group-average and Ward's methods*

Two methods which are less likely to be influenced by the above mentioned error or noise are the group-average and Ward's methods, because of the matrix updating formulae for the two methods, which are as follows [Everitt, 1993].

For the group-average method:

$$d[(k),(r,s)] = \left(\frac{n_r}{n_r + n_s}\right) d(k, r) + \left(\frac{n_s}{n_r + n_s}\right) d(k, s). \quad \text{-------} \quad (4.3.6.1)$$

where n_k , n_r and n_s denote the number of objects in clusters k , r and s respectively and d(k, r) and d(k, s) denote the distance between the clusters k and r, and k and s, respectively.

For Ward's method:

$$d[(k),(r,s)] = \left(\frac{n_r + n_k}{n_r + n_s + n_k}\right) d(k, r) + \left(\frac{n_s + n_k}{n_r + n_s + n_k}\right) d(k, s) - \left(\frac{n_k}{n_r + n_s + n_k}\right) d(r, s). \quad \text{----------} \quad (4.3.6.2)$$

where d(k, r), d(k, s) and d(r, s) are the square of the distances.

The five-step algorithms for the single-linkage and the complete-linkage methods are illustrated in Example I. To use the group-average or Ward's methods, the equation given in Step 4 (page 88) needs to be replaced by Equation 4.3.6.1 or 4.3.6.2, respectively. This is the only difference in the proximity matrix-updating algorithm. In these formulae, the distances between all the group members is considered, rather than the distances between single members from each group. Therefore, they are less sensitive to small changes in the location of events, and do not have a tendency to form chain clusters, because of the 'group neighbour property'.

In the single-linkage method each cluster is identified by the longest distance needed to connect any member of the cluster to any other member of the cluster, and in the complete-linkage method each cluster is identified by the longest distance needed to connect every member of a cluster to every other member. These two methods rely on the extreme values for identifying the clusters, but in the group-average method each cluster is identified by the average of all distances between members within the clusters.

In the group average method the distance between two clusters is taken to be the average of the distance between all pairs of individuals from the separate groups as shown in Figure 4.10.

Anderberg [1973] noted that " the Ward objective is to find at each stage those two clusters whose merger gives the minimum increase in the total within-group error sum of squares E" where E is the total within-group error sum of squares for the collection of clusters. Ward's method is good for clusters of approximately equal size but performs poorly when the clusters are of different sizes. The group-average method is good for unequal size clusters [Anderberg 1973, Everitt 1993]. They also noted that Ward's method and the group-average method are affected to a lesser extent by "noise" in the data.

The four main cluster analysis methods are investigated further in Chapter 7.

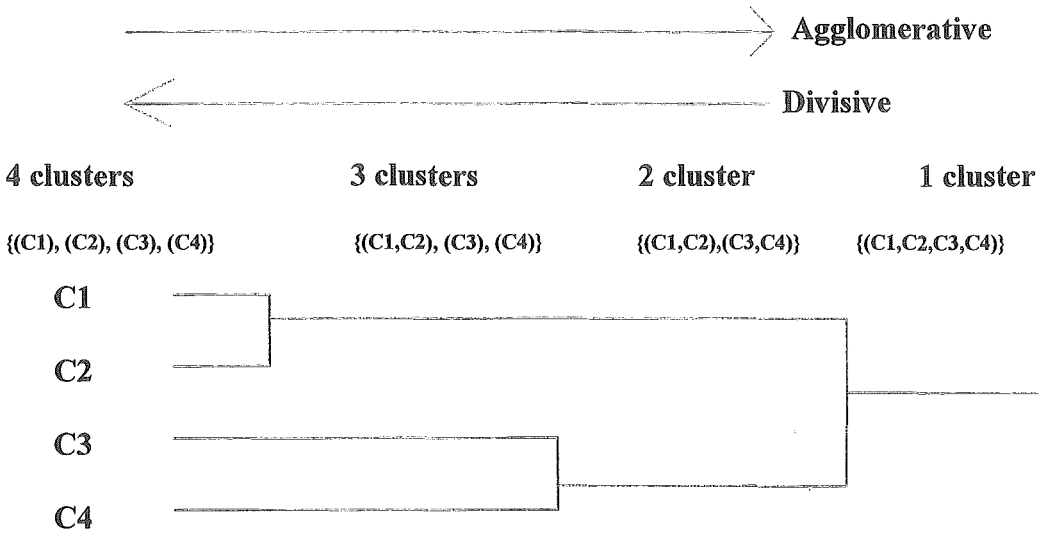


Figure 4.01: Dendrogram

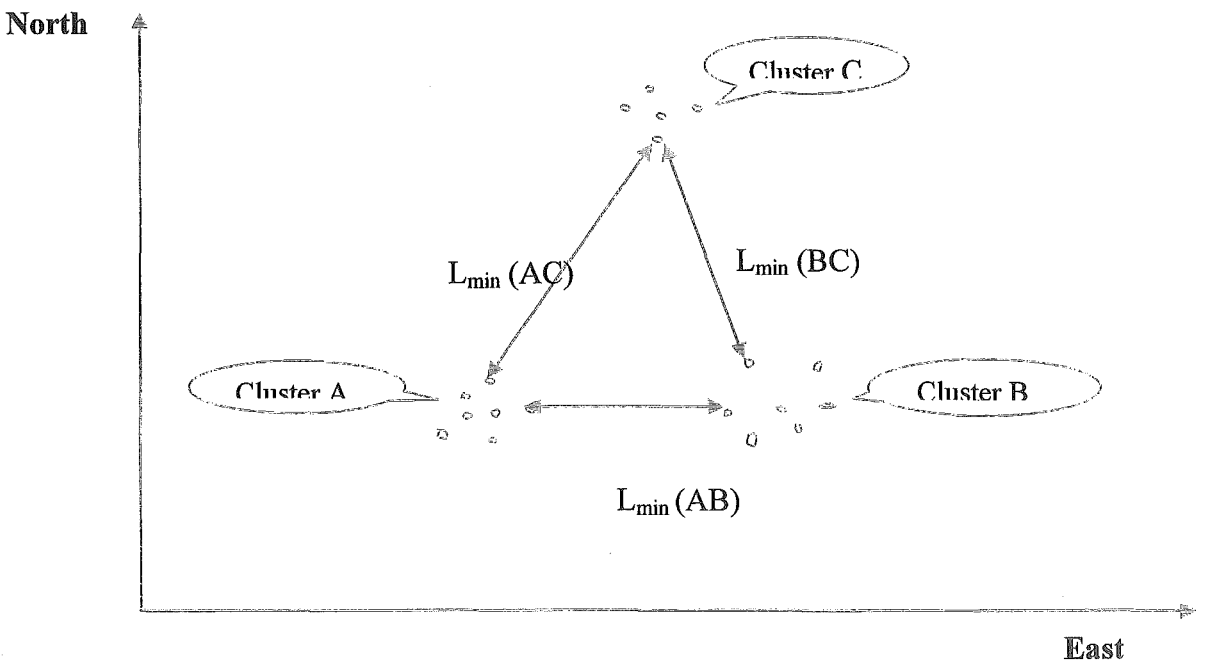


Figure 4.02: Three clusters A, B and C, with the closest members for each pair of clusters marked

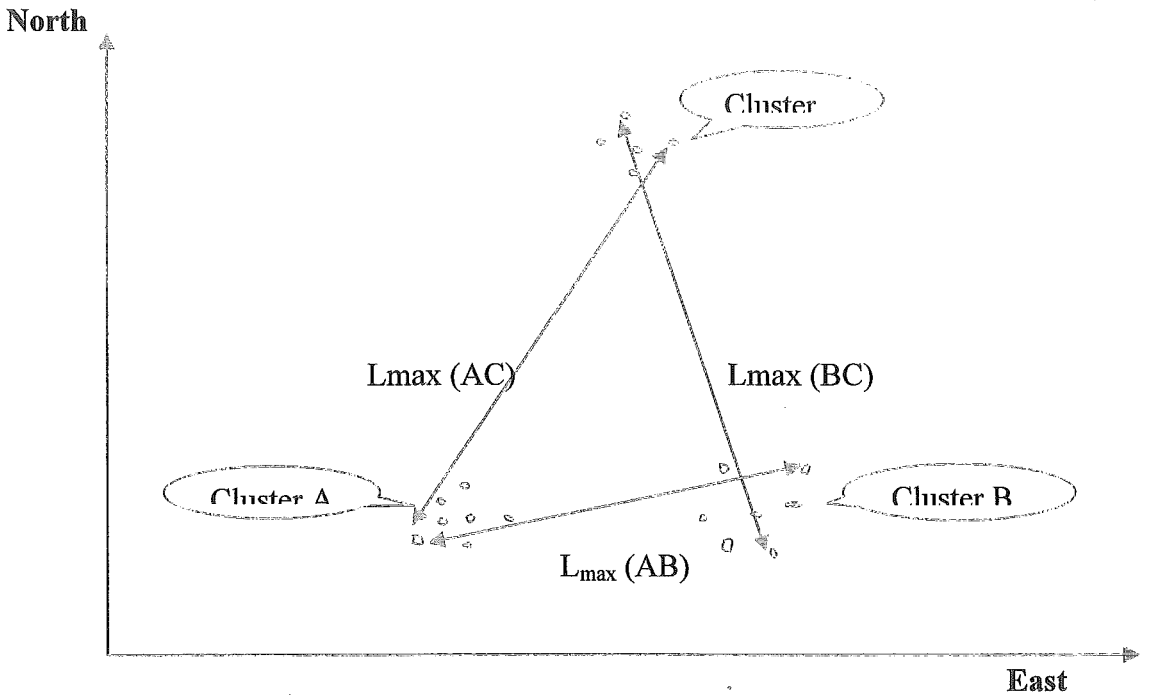


Figure 4.03: Three clusters A, B and C with the distant members for each pair of clusters marked

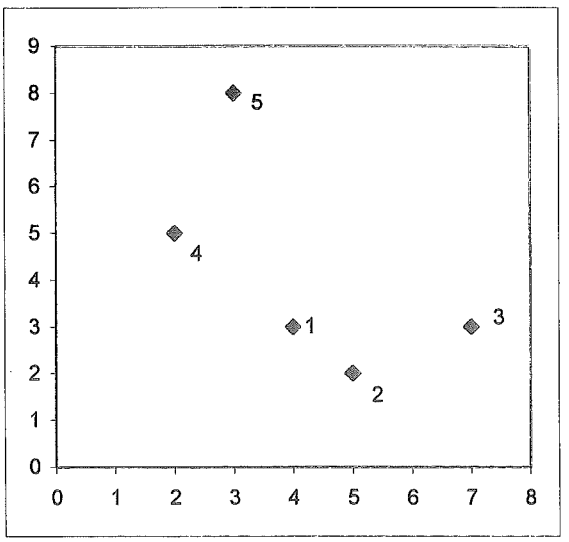


Figure 4.04: Location Plot for Example I

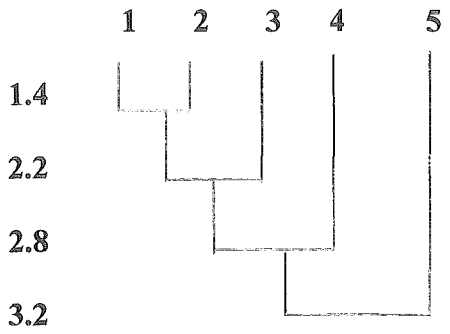


Figure 4.05: Dendrogram (Single-linkage method applied to Example I)

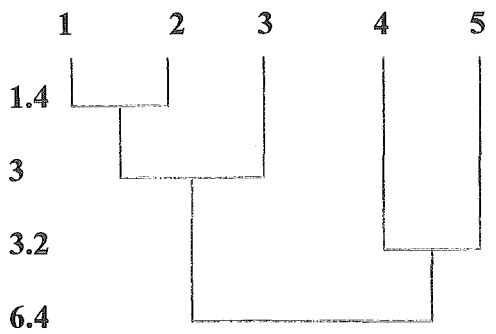


Figure 4.06: Dendrogram (Complete-linkage method applied to Example I)

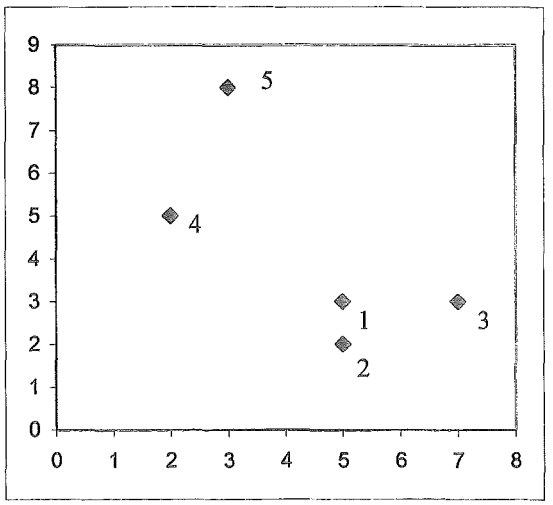


Figure 4.07: Location Plot for Example II

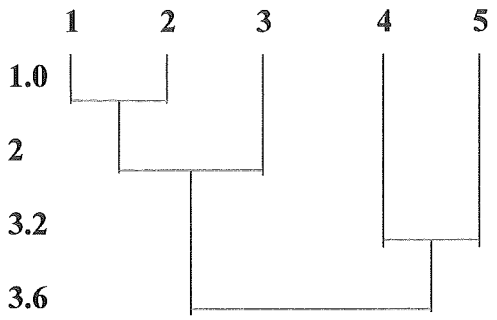


Figure 4.08: Dendrogram (Single-linkage method applied to Example II)

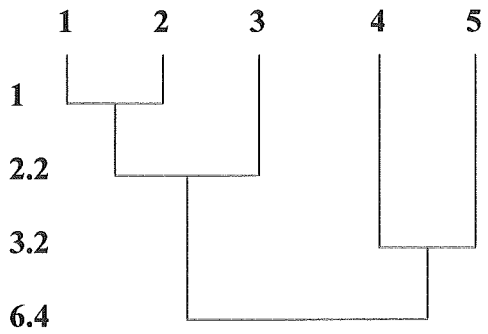
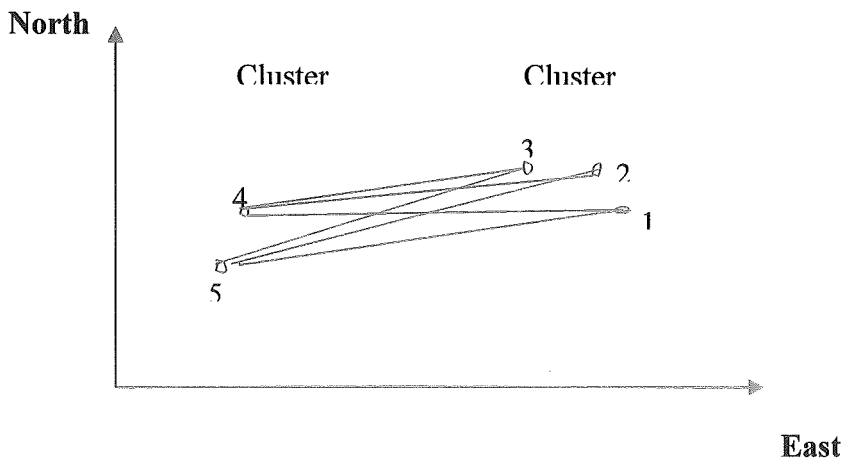


Figure 4.09: Dendrogram (Complete-linkage method applied to Example II)



$$d_{AB} = (d_{14} + d_{24} + d_{34} + d_{15} + d_{25} + d_{35})/6$$

Figure 4.10: Group average distance

Chapter 5

NEAREST-NEIGHBOUR ANALYSIS

5.1 Introduction

In cluster analysis, the objects are grouped according to the distance between events or groups of events, but in the nearest-neighbour analysis the distances and/or directions from an event (or sample point) to the neighbouring events are considered. The difference between the quadrat analysis and nearest-neighbour analysis is that in the quadrat analysis counts of events are grouped according to the quadrat in which they are located, but in the nearest-neighbour analysis event-to-event (or point-to-event) distances and/or directions (i.e. the distances and/or directions from accident locations or sample points to nearby accident locations) are considered. Selecting a sample point has a practical difficulty because the sample point should sensibly be within roads. In addition, the analysis results depend on the selection of points. For these reasons, sample events will be used instead of sample points. The nearest-neighbour analysis analyses the distances and/or directions to neighbouring events, and the distance and direction distributions are discussed separately. The nearest-neighbour distance analysis discussed in this chapter is based on Cressie [1993] and Ripley [1981] and the nearest-neighbour direction analysis is based on Upton and Fingleton [1989].

In the present study, the nearest-neighbour analysis involves a sampling process in which the number of events in each sample is constant. The events within the sample are investigated to search for evidence of the three basic spatial patterns by analysing the nearest-neighbour distance and direction distributions. Each sample is based on a selected event, which is called the test-location, and the nearest neighbours are the rest of the events within the selected sample as shown in Figure 5.01. The distance and direction distributions of the nearest neighbours within the sample are used to test for randomness. Hence the test result depends only on the distribution of events within the sample. The smaller the spatial range of the sample, the more localised is the testing.

The test method identifies whether there is a pattern of events within the sample (i.e. whether the test-location is within a random or clustered space of nearest neighbour events and whether the neighbouring events are located in a particular direction within the sample space).

For example, six nearest-neighbour events and the test location are shown in Figure 5.01. If we consider the six nearest-neighbour events then it might appear to clearly show that the test event is located in a cluster location, but if we consider five nearest neighbour events, then it might appear that the events around the test-location are located randomly. If we selected the 6^{th} event as the test-location then that location might appear as a location around which accidents are randomly distributed. Therefore, the crucial part in this analysis is to decide the number of events in each sample and the selection of test-locations from the given data. Details are discussed in Section 5.2.2.

The test-locations must be within the roads, because the position of the test-location influences the test result. If we randomly select events as test-locations and do not select all the events, then the result of the analysis may vary according to which events are selected as test-locations. To avoid inconsistency in the result and the complexity of randomly selecting test-locations, each event may be selected as a test-location.

Different statistical tests are used for analysing distance and direction distributions and the accuracy of the test result depends on the number of nearest-neighbour events. The role of the number of events within the sample is discussed in Sections 5.2 and 5.3 in greater detail.

The nearest-neighbour analysis involves the use of precise information about the locations of accidents, and the co-ordinates of each location are used to calculate the distances between events. The method traditionally involved the analysis of distances (but not directions) to nearest-neighbours. Nicholson [1999] constructed two sets of distributions (Figures 3.13 & 3.14) and showed that it is worthwhile analysing both the distances and directions to nearest-neighbour events.

5.2 Analysis of nearest-neighbour distances

The two traditional distance analysis methods, cluster and nearest-neighbour analysis, make use of precise information about the locations of events for analysis. Cluster analysis does not take full account of the number of nearest-neighbours but the nearest-neighbour distance analysis does that and hence is better in that respect. The nearest-neighbour distance analysis can be used to analyse any number (e.g. up to five) of nearest-neighbour distances. In general, the nearest-neighbour analysis is a powerful tool for analysing the spatial patterns, but judgement is important when selecting the number of nearest neighbours and interpreting the results. The nearest-neighbour distance distribution can be compared with the benchmark distance distribution for a CSR spatial distribution.

In the nearest-neighbour distance analysis, distances are computed from an event (accident) to other events (accidents) and summarized. All the events are numbered from 1 to n, where n is the total number of events. The distance matrix is recorded in Table 5.01. The lower diagonal half of the matrix is used for direction, as explained in Section 5.3.

Table 5.01: Distance and direction matrix

Events	1	2	3	4	...	i	...	n
1	0	d_{12}	d_{13}	d_{14}	...	d_{1j}	...	d_{1n}
2	θ_{21}	0	d_{23}	d_{24}	...	d_{2j}	...	d_{2n}
3	θ_{31}	θ_{32}	0	d_{34}	...	d_{3j}	...	d_{3n}
4	θ_{41}	θ_{42}	θ_{43}	0	...	d_{4j}	...	d_{4n}
:	:	:	:	:	0	:	...	:
i	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	...	0	...	d_{in}
:	:	:	:	:	...		0	:
n	θ_{n1}	θ_{n2}	θ_{n3}	θ_{n4}	...	θ_{nj}	...	0

Let d_{ij} be the distance between the i^{th} event and the j^{th} event, where i and j indicates a row or a column of the matrix. Suppose, the distance between the 3^{rd} event to other events is noted as a single row matrix (e.g. $[\ d_{3,j}] \equiv [\ d_{3,1}\ ,\ d_{3,2},\ldots,\ d_{3,j},\ldots,\ d_{3,n}]$). Note that the first nearest neighbour from the 3^{rd} event may not be $d_{3,1}$. Let d^k_i be the distance to the k^{th} nearest neighbours from the i^{th} event (the test-location). The first k

nearest neighbour from the test-location (i.e. the i^{th} event) is denoted as $[D^k_i] = [d^1_i , d^2_i, d^3_i, , , , , , d^k_i]$, where $d^1_i < d^2_i < d^3_i < d^k_i$.

5.2.1 *Testing distance distributions*

The aim is to test whether accident (events) in the vicinity of a test location are clustered, regular or random. Nicholson [1999] introduced a method to identify clustered, regular and random distributions based on the K function (discussed in Chapter 3). Nicholson considered two different types of distance distributions, one a regular pattern and the other a clustered pattern. In each sample, N nearest neighbours were considered. The two distributions were compared with the expected distance distribution for a CSR distribution, which has equal density (i.e. the density is equal to N/A, where A is the sample area and N is the number of events within the sample). The cumulative proportion of events is plotted against the proportion of distance for the two distributions (i.e. cluster and regular) and the same density of a CSR distribution, as shown in Figure 5.02.

Suppose N is the number of events within a sample, and $d_{(N)}$ is the distance from a test-location to the N^{th} nearest-neighbour event, and $d_{(i)}$ is the distance from the test-location to the i^{th} neighbour event and λ is the density of the CSR distribution.

The expected number of events within the distance $d_{(N)}$ is

$$E(N) = \lambda \pi (d_{(N)})^2 \qquad \text{---------- 5.2.1.1}$$

and the expected number of events within the distance $d_{(i)}$ is

$$E(i) = \lambda \pi \{d_{(i)}\}^2 \qquad \text{---------- 5.2.1.2}$$

From equation 5.2.1.1 and 5.2.1.2, the expected proportion of events within d(i) is

$$P(i) = E(i) / E(N) = \{d_{(i)} / d_{(N)}\}^2 \qquad \text{---------- 5.2.1.3}$$

P(i) is the expected cumulative proportion of events and $\{d_{(i)} / d_{(N)}\}$ is the proportion of distance (to the Nth nearest-neighbour). The expected curve for CSR is given by equation 5.2.1.3 and is shown in Figure 5.02.

The observed distance distribution is compared with the same-density CSR distribution. The aim is to test how well the observed cumulative distribution function (cdf) fits with the cdf for a CSR distribution. This test is known as the Kolmogorov-Smirnov goodness-of-fit test [Mood et al. 1974, Benjamin and Cornell 1970, Press et al. 1992].

The maximum deviation of the absolute values of the cdf for the observed distribution and the cdf for the CSR distribution must be computed. The computed value (D) is given by the following equation.

$$D = \text{MAX}\ \left[\left|F_1(X) - F_2(X)\right|\right]$$

where $F_1(X)$ is the cdf for the observed distribution, $F_2(X)$ is the cdf for a CSR distribution, X varies between 0 and 1, and n is the number of observations or events within the sample. The value of D can be compared with the critical value for the chosen significance level. If D is less than or equal to the critical value then the observed distribution is similar to CSR, otherwise the observed pattern is not CSR.

Benjamin and Cornell [1970 : pg 467] noted: "It has been found ... that the distribution of this sample statistic is independent of the hypothesized distribution of X". They also mentioned that the parameter (n) depends on the distribution of the sample statistics D. Press et al. [1992] mentioned that the test result is independent of the X-axis scale. The same authors noted that the effective number of data points is $n \left(= \frac{n_1 n_2}{n_1 + n_2} \right)$, where n_1 is the number of data points for $F_1(X)$ and n_2 is the number of points for $F_2(X)$, and the test is asymptotically accurate provided $n \geq 4$. This means that for finite n_1, as $n_2 \rightarrow \infty$ we need $n_1 \geq 4$. The sample size for $F_2(X)$ is thus very large and therefore the number of data points for the observed distribution must be greater or equal to four to get reasonably accurate results. Press et al. [1992] noted that "the Kolmogorov- Smirnov test is good at finding shifts in a probability distribution, especially changes in the median value, but it is not always good at finding spreads, which affect the tails of the probability distribution, and which may leave the median unchanged". That is, the test is unable to detect small variations of CSR very close to the test location and very

close to the n^{th} nearest-neighbour (presumably because the cdf's are constrained at both ends).

Nicholson [1999] investigated the ability of the Kolmogorov- Smirnov test to detect two types of deviation from CSR (i.e. excessive clustering or regularity), where the sample accident is surrounded by 24 other accidents. The cumulative distribution functions of distances for regular, cluster and CSR distributions are shown in Figure 5.02. This shows that the discrepancy for the pattern with excessive clustering is substantially greater than for the pattern with excessive regularity.

The areas under the three curves (cluster, regular and CSR) in Figure 5.02, are A_c, A_r and A_{csr}, and $A_c > A_{csr} > A_r$. From the test result it appears that if the area under the observed curve is greater than the area under the expected CSR distribution curve then the distribution is clustered, but if the area under the observed curve is less than the area under the expected CSR distribution curve then the distribution is regular. This method can be used to identify cluster or regular locations.

The two methods (i.e. comparison of the areas under the expected and observed curves and application of the Kolmogorov- Smirnov test) are both required for detecting the spatial distribution of accidents where there is excessive clustering or regularity. It is advantageous to use the Kolmogorov- Smirnov test to calculate the percentage of accident locations where the neighbouring locations are from a CSR distribution. This method can be used with a small number of events (>4) in each sample and the test method only investigates the neighbouring locations. The advantage of comparing neighbouring accident locations rather than locations which are far apart, is that the neighbouring accidents could be related to the characteristic of the test location. The test results will help to identify whether the neighbouring accidents within the sample depend on the characteristic of the test location, which are related to the site of the test-location. If the test location is within a cluster then the suggestion is that the site needs spot treatment for accident reduction. This method is discussed in detail because of its usefulness in identifying the most appropriate accident reduction plan. In the following example, the influence of the number of nearest neighbours on the results is investigated.

5.2.2 *Sensitivity of the nearest-neighbour test*

The aim of this analysis is to show how the number of events within the sample and the test location will affect the sensitivity of the nearest-neighbour test result. Consider the spatial distribution shown in Figure 5.03 in which two event clusters and an isolated event are plotted.

Different numbers of nearest-neighbours (N= 8, 9 and 14) are considered to study the sensitivity of the nearest neighbour analysis to the variation in the number of nearest-neighbours (N). The cumulative proportion of events versus the proportion of distance calculated from the first event (test-location) up to the 8^{th}, 9^{th} and 14^{th} nearest neighbours are plotted in Figure 5.04.

It can be seen from Figure 5.04 that the areas under the curve and the maximum distance between the CSR curve and the actual distribution curves differ substantially as the number of nearest-neighbours (N) changes. A similar method was used to investigate the case with event 19 as the test location. The result indicated that the distribution is not clustered, and shows that the result is sensitive to the selection of test-location.

Each of the 19 events was selected as the test-location, with the distribution shown in Figure 5.03 embedded in eight identical distributions to take account of edge effects. The results were:

1. events appeared as part of a cluster when $N < 9$;
2. 90 % of events appeared as part of a cluster when $9 < N < 12$;
3. 68% of events appeared as part of a cluster when $N = 13$;
4. less than 50% of events appeared as part of a cluster when $14 < N < 18$.

The analysis indicates that the result is very sensitive to both the number of nearest-neighbours and the position of the test-location. Therefore, considerable care must be exercised when selecting the number of nearest-neighbours and the test locations, and when interpreting the results.

Some conclusions can be drawn by visual examination of Figure 5.03. The visual

examination result depends on the sample window size (i.e. the number of nearest-neighbours). If the sample area is just enough to cover a cluster (i.e. small area), then the locations of events within that sample do not appear to be clustered. If the sample window is larger than the cluster-size, then the locations within the sample appear to be clustered.

Since accidents occur only on vehicle paths, the analysis results could be influenced by the amount of space between roads. Nicholson [1995] investigated the error associated with the approximation of a lattice with a continuum, and concluded that the continuum approximation is appropriate for lattices, where the road network is a regular grid with the block size less than 250m (i.e. where the road network is relatively dense or the space between the roads is small). The reason why the errors become large as the block as the block size increases is discussed below.

The distance between two locations can be defined in two different ways. They are the Euclidean (straight-line) distance and the distance along the road. The Euclidean distance depends on the space between the roads if the two locations are in two different roads (as shown in Figures 5.05a and b). The objective is to find an analysis method, which is sensitive to the accident pattern on roads but not to the space between the roads.

When we have accident locations, there is a practical difficulty in finding the distance between two locations along the road. This difficulty could be minimised by selecting a small number of events in each sample by reducing the area of the sample (as shown in Figures 5.05 (a), 5.05 (b)).

Figure 5.05(a) shows two intersecting roads, d_2 is a reasonable approximation to the distance measured along the roads when the two locations are within the small sample, but d_1 is not a good approximation to the distance measured along the roads when the two locations are outside the small circle but within the large circle. Where both the test location and sample event are close to the junction of intersecting roads, the direct distance (straight line) can be used, but not if the test location and sample event are

not close to the junction. For a given intensity, the greater the number of events, the greater the area, and the less is the accuracy from using the straightline distance.

Figure 5.05(b) shows two non-intersecting roads with five accidents near the test location (location A) and two neighbouring accidents on the near by road. Let location B be the location of one of those two. For the test location at location A, the nearest-neighbour distances for the small sample are measured along the road but for the large sample, the distances from A to B and the distance from A to C are influenced by the distance between the roads. In this case it is better to consider a small number of nearest-neighbours. For the test location at event B, the test-location is isolated and is far from other accidents. In this case the shortest distance is influenced by the space between the roads.

Consider the same road network with two accidents near the test-location (location A) and four neighbouring accidents on the near by road. Let location B be the location of one of those four as shown in Figure 5.06. The test-location is A and six nearest events are analysed.

The cumulative distance distribution function for 6 nearest neighbours is shown in Figure 5.07, the area under the actual curve is smaller than the area under the CSR curve, indicating that the test-location A appears to be regular, when the events around it are actually clustered. In this case the distance between the two roads affects the result and the conclusion. The distance d between A and B (Figure 5.06) is influenced by the distance between the two roads. Whether the test location is clustered or regular or random, the distance between the two roads can influence the result.

The unexpected result in Figure 5.07 is because the number of nearest neighbours analysed (6) is greater than the cluster size. If the number of nearest neighbours selected for analysis is 4 then the expected result (i.e. a cluster distribution) can be obtained from the analysis. Therefore for analysis, the right selection of the number of nearest neighbours will improve the reliability of the results.

The notation "D-" indicates the maximum discrepancy in the cumulative proportion of events between the two curves (profile of cumulative proportion of events when the actual curve is below CSR curve). The notation "D+" indicates the maximum vertical discrepancy in the cumulative proportion of events between the two curves when the actual curve is above CSR curve. The notation "d+" indicates the maximum horizontal discrepancy proportion of distance between the two curves when the actual curve is above CSR curve. The notation "d-" indicates the maximum horizontal discrepancy proportion of distance between the two curves when the actual curve is below CSR curve. The notation "d(ave)" indicates the absolute average horizontal discrepancy between the two curves. These variables are used for further investigation as explained below.

To improve the analysis results two different methods were examined.
In method one:
if D+ > D- then the test location appears to be within non-clustered events;
if D+ < D- then the test location appears to be within clustered events.
In method two:
if d+ > d(ave) then the test location appears to be within non-clustered events;
if d+ < d(ave) then the test location appears to be within clustered events,
where d(ave) = $(d_1+d_2+,.....d_N)/N$, and N is the number of nearest-neighbours. The two methods are more sensitive to clustering than regularity, so are not any better than the previous method. Hence the two methods were not investigated further.

Figure 5.08 is the same as Figure 5.06 but with greater number of accidents, which are distributed differently. As expected the result for the isolated test-location A in Figure 5.08 indicates that this is a non-cluster location when considering six nearest-neighbour locations. The result for six nearest-neighbours from the test-location B indicates that location B is within a cluster, as expected. In both these cases, the space between the two roads does not affect the result, even though the distance between the two locations (A and B, B and C) depends on the distance between the two roads at this location.

For the spatial distribution shown in Figure 5.08, consider the cases where:

(1) AC = 350m , AB = 300m or

(2) AC = 600m, AB = 300m.

Let the test location be location A. The difference between the two cases is the distance of the 7th nearest-neighbour from location A. The cumulative proportion of events against the proportion of distance from the event A to the 7th nearest neighbour is plotted for each case in Figure 5.09.

If the distance from A to C is large compared to the distance from A to B (ie., AC >> AB), then for the test-location A, the results for seven nearest neighbours, do indicate a cluster location. If AC $\approx$ AB, do the results indicate a non-cluster location for 7 nearest neighbours. Although this problem may occur only occasionally, such problems need to be considered when selecting the number of nearest neighbour events (to minimise the likelihood of such errors), and in interpreting the results.

The nearest-neighbour distance analysis will indicate the appropriate pattern for test locations in most cases. Since the analysis will be done for all possible locations (i.e. for all the events), one or two special cases will not affect the overall result.

Further discussion of the nearest-neighbour distance analysis is covered in Chapter 7.

5.3 Analysis of nearest-neighbour directions

The importance of nearest-neighbour direction analysis was described in Chapter 3. Upton and Fingleton [1989] described a range of techniques and these were investigated by Nicholson [1995 & 1999]. The techniques involved analysing the nearest-neighbour directions from the test-location. The direction from event "i" to event "j" is $\theta_{i,j}$ as shown in Figure 5.10. Note that in this figure the direction is measured clockwise from the north direction. The direction from every pair of events is recorded in the lower diagonal part of the matrix shown in Table 5.01. The upper diagonal part of the matrix contains the distances as explained in Section 5.2.

Let θ_{ij} be the direction from the ith event to the jth event, where i and j indicates the row and column of the matrix respectively. Suppose, the direction from the 3rd event to other events is denoted as a single row matrix (e.g. $[\theta_{3,i}] \equiv [\theta_{3,1}, \theta_{3,2}, \ldots, \theta_{3,i}, \ldots, \theta_{3,n}]$). Note that the first nearest-neighbour direction from the 3rd event may or may not be

$\theta_{3,1}$. Let θ^k_i be the direction to the k^{th} nearest-neighbour from the i^{th} event (test-location). The vector of directions to the first k nearest-neighbours from the test-location (i.e. the i^{th} event) is denoted as $[\theta^k_i] = [\theta^1_i , \theta^2_i, \theta^3_i, \ldots, \theta^k_i]$, where the distance $d^1_i < d^2_i < d^3_i < \ldots < d^k_i$

Consider the situation where there are 17 nearest-neighbours with directions as shown in Table 5.02, in which the first column indicates the nearest-neighbour bearing in degrees and the second column indicates the frequency of observation. The directions to nearest-neighbours bearings are defined in the range from 0° to 360°. The number of observations in each bearing can be plotted in a circular histogram, as shown in Figure 5.11, in which the frequency of observation represents the length of the bar. In this figure, the nearest-neighbours are nearly on a straight line in the east-west direction.

Table 5.02: Nearest neighbour bearings and frequency.

Bearing(deg)	Frequency
85	1
88	2
90	4
92	1
265	1
268	2
270	4
275	1

The nearest-neighbour direction techniques differ from the analysis of nearest-neighbour distances. The first approach is to treat each direction observation θ^k_i (k = 1,2, . . ., n) as equivalent to a unit vector (i.e. a vector with a unit length) in the direction θ^k_i (the clockwise angle from north direction). If the displacement in the north direction is X, (i.e., $X= \sum_{k=1}^{n} \cos\theta_i^k$) and displacement in the east direction is Y (i.e. $Y = \sum_{k=1}^{n} \sin\theta_i^k$) then $R = \sqrt{(X^2 + Y^2)}$

The magnitude of the sum of the vectors (i.e. the magnitude of the resultant vector R) is

$$R = \{(\sum_{k=1}^{n} \cos\theta_i^k\ )^2 + (\sum_{k=1}^{n} \sin\theta_i^k\ )^2\}^{0.5}$$

where $0 \leq R \leq n$ and that because of the cyclic nature of the sine and cosine, data can cancel out.

The value of R depends on the number of observations (n). The normalized measure of concentration is $\overline{R} = R/n$, so that $0 \leq \overline{R} \leq 1$. The $\overline{R}$ calculated from Table 5.02 is 0.004, which is close to zero. This is because the observations are in two directions (i.e. east and west). In this calculation the $\overline{R}$-value is small because sum of the data is also small.

Upton and Fingleton [1989] noted that "...in unimodal samples, the size of R gives an indication of the degree of concentration of the individual observations about the preferred direction; the larger the value of R, the greater the concentration". The value of R is a useful measure of concentration for unimodal data but not for multimodal data. If the data come from a multimodal distribution, then the directions need to be multiplied by number of modes (say m). The measure of concentration is then;

$$\overline{R}_m = \{(\sum_{k=1}^{n} \cos(m\theta_i^k))^2 + (\sum_{k=1}^{n} \sin(m\theta_i^k)\ )^2\}^{0.5}\ /n,$$

From the data in Table 5.02, the nearest-neighbours are distributed in two modes (i.e. m = 2). The calculation of $\overline{R}_m$ (m=1,2) is shown in Table 5.03.

When the data are not unimodal then we may get a very low value of $\overline{R}$. If we allow for multi-modality, we may get a large value of $\overline{R}_m$. If the $\overline{R}_m$ value is large then the directions to nearest neighbours are non-uniformly distributed. In other words, if the distribution of directions to the nearest neighbours is sufficiently non-uniform, then it can be concluded that the accidents are not uniformly distributed (i.e. the accident process is not isotropic). Pearson and Hartley [1972], mention that goodness-of-fit test statistics based on cumulative distribution functions (including the Kolomogorov – Smirnov statistics) are not suitable for circular data, as they depend on the origin (i.e. the direction that is assigned the value zero).

Table 5.03: Calculation of $\bar{R}_1$ and $\bar{R}_2$

Range (deg)	Frequency (f)	Cos θ	Sin θ	f Cos 2θ	f Sin 2θ
85	1	0.0872	0.9962	-0.9848	0.1736
88	2	0.0349	0.9994	-1.9951	0.1395
90	4	0.0000	1.0000	-4.0000	0.0000
92	1	-0.0349	0.9994	-0.9976	-0.0698
265	1	-0.0872	-0.9962	-0.9848	0.1736
268	2	-0.0349	-0.9994	-1.9951	0.1395
270	4	0.0000	-1.0000	-4.0000	0.0000
275	1	0.0872	-0.9962	-0.9848	-0.1736
sum	16	0.0523	0.0032	-15.9422	0.3829

$m = 1$ $\quad \bar{R}_1 = \quad 0.0033$

$m = 2$ $\quad \bar{R}_2 = \quad 0.9868$

Upton and Fingleton [1989] and Nicholson [1995, 1999] described three tests for testing the uniformity of circular data. The two modified versions of the Kolmogorov-Smirnov goodness-of-fit test (the Kuiper-Stephen and Watson-Stephens tests) are based on the cumulative distribution function. The third test is the Rayleigh-Wilkie test, which is based on the R value. Upton and Fingleton noted that the accuracy of the tests depend on the circumstances. The three tests are described in the following sections.

5.3.1 *The Rayleigh-Wilkie test*

The Rayleigh test can be used to test whether the value of R is sufficiently large to justify rejecting the hypothesis of uniformity. Upton and Fingleton noted that the expected value of R is $\sqrt{(n\pi)}/2$. If R is sufficiently greater than this value then the null hypothesis of an isotropic process can be excluded. Upton and Fingleton noted that the Rayleigh statistic (i.e. $T = 2R^2/n$), which is approximately χ^2 distributed, relies on the number of observations being greater than or equal to 100. Stephens [1969] has given a table of exact critical values of (R/n) and showed that the Rayleigh statistic provides good accuracy for $n \geq 20$. Upton and Fingleton noted that Wilkie [1983] shows that the

probability of obtaining an R value greater than or equal to the observed R is approximately

$$\exp\{[1 + 4n + 4(n^2 - R^2)]^{0.5} - (1 + 2n)\},$$

which implies "the exceedance probability associated with obtaining a value equal or greater than R from a sample of size n taken from a uniform distribution". The number of observations n must be greater than or equal to five. That is, this test is used for samples as small as five. If R is sufficiently large, the null hypothesis (i.e. an isotropic process) can be rejected.

The Rayleigh test is valid when the alternative to uniformity is a unimodal distribution. However, because the R value can become zero for a multimodal distribution, when the opposite modes balance each other, the data should be scaled appropriately, and the value of R^2 based on the scaled data should be used. Nicholson [1995 & 1999] proposed a method to estimate the value of m. The $\overline{R}_m$ value can be calculated for different values of m and the m value giving the maximum $\overline{R}_m$ value is taken as the number of modes. For the data in Table 5.04, the $\overline{R}_m$ value is a maximum when m is two.

Table 5.04: Calculation of m and $\overline{R}_m$ for data in Table 5.03

m	$\overline{R}_m$
1	0.0033
2	**0.9967**
3	0.0098
4	0.9868
5	0.0163
6	0.9706
7	0.0228
8	0.9485

5.3.2 *The Kuiper-Stephens test*

This test is a modified version of the Kolmogorov-Smirnov test, which was discussed in the previous section in relation to nearest-neighbour distance analysis. The basis of the Kuiper-Stephens test is the cumulative distribution function of the observed nearest neighbour directions (θ_i) compared with the uniform distribution function.

Upton and Fingleton noted that the Kuiper-Stephens test statistic (K) is given by

$K = (D^+ + D^-)\sqrt{(n)}$

where $D^+ = Max_\theta \{0, F(\theta) - F_n(\theta)\}$, $D^- = Max_\theta \{0, F_n(\theta) - F(\theta)\}$, (see Figure 5.07),

$F(\theta)$ = Pr[obtaining value between 0 and θ]

If the distribution is uniform then $F(\theta)$ is equal to $\theta /(2\pi)$.

$F_n(\theta)$ = observed sample proportion of values between zero and θ.

Stephens [1970] suggested a modified K statistic,

$K^* = (D^+ + D^-)\sqrt{(n)}(1+.155/\sqrt{(n)} + 0.24/n)$.

This does not require an extensive table of percentage points and can be used for $n \geq 5$.

Table 5.05: The percentage points of the K* value (extracted from Upton and Fingleton [1989])

	Significance level		
n	**10%**	**5%**	**1%**
5	1.63	1.75	1.97
10	1.62	1.74	1.99
20	1.61	1.74	2
100	1.62	1.75	2
∞	1.62	1.75	2

If the calculated value of K* is sufficiently greater than the critical value then the null hypothesis (an isotropic process) can be rejected. It can be seen that the critical values do not differ very much for varying sample sizes.

5.3.3 *The Watson-Stephens test*

The Watson-Stephens test compares the cumulative distribution function of the nearest neighbour directions (θ_i) with the uniform distribution function. This test is based on the discrepancies between the cumulative distribution functions at each (θ_i) and not just on the maximum discrepancy.

Upton and Fingleton noted that the Watson test statistics is

$U^2 = \sum [F(\theta_i)]^2 - \{\sum_i [(2i-1)F(\theta_i)]\}/n + n[(1/3)(\bar{F} - 0.5)^2]$,

where $\bar{F} = \sum_i F(\theta_i)/n$.

They also noted that the expected value of U^2 is 1/12. If the calculated value of the Watson test statistic is sufficiently greater than 1/12 then the null hypothesis (an isotropic process) can be rejected.

Stephens [1970] suggested a modified statistic

$U^* = \{U^2 - 0.1/n + 0.1/n^2\}\{1.0 + 0.8/n\}$.

This does not require an extensive table of percentage points and the test can be used for $n \geq 5$. The following table indicates the critical values of U* for five different significance levels.

Table 5.06: Critical values of U* for five different significance levels.

Significance level	15%	10%	5%	2.5%	1.0%
Approximate critical value	0.131	0.152	0.187	0.221	0.268

If the calculated value of U* is greater than the critical value then the null hypothesis (i.e. an isotropic process) can be rejected. Stephens [1970] showed that the critical value is extremely stable for varying sample sizes.

Nicholson (1994) assessed the relative abilities of the tests to detect non-uniformity in the direction to nearest neighbours. Three simple test patterns were constructed, with a centrally located accident surrounded by various sized groups of accidents and arranged so that there are two, four and eight evenly spaced modes. Varying group sizes for each mode were used for the analysis. From this test method it was found that the Rayleigh test of R is much more powerful than the Kuiper and Watson tests. However the Rayleigh test is extremely sensitive to the estimate of the number of modes.

These nearest neighbour direction tests are discussed further in Chapter 7.

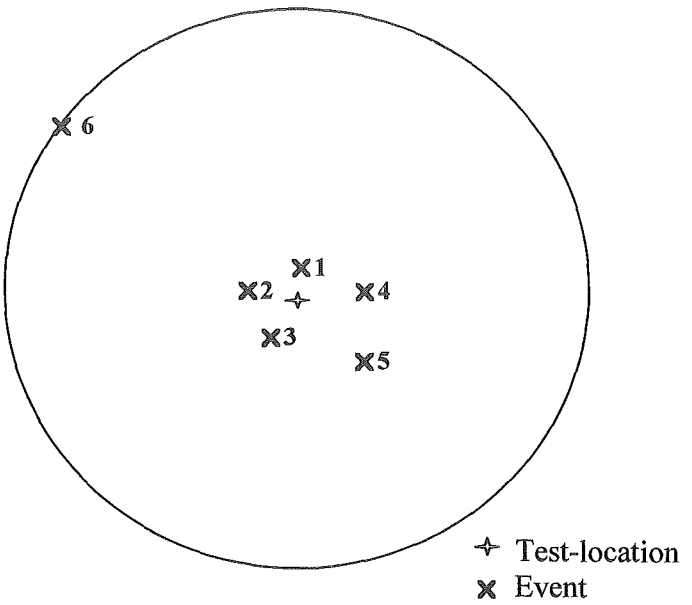


Fig 5.01: Six nearest neighbour events and a test-location (a selected event) shown

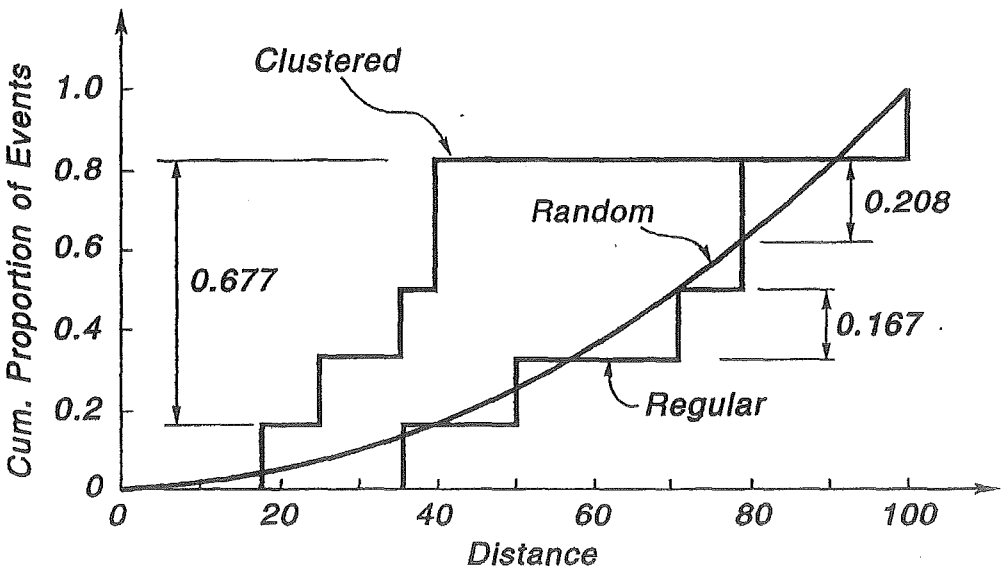


Figure 5.02: Cumulative proportion of events against proportion of distance for cluster, regular and CSR distribution.

[Figure 5.02 extracted from Nicholson [1999]]

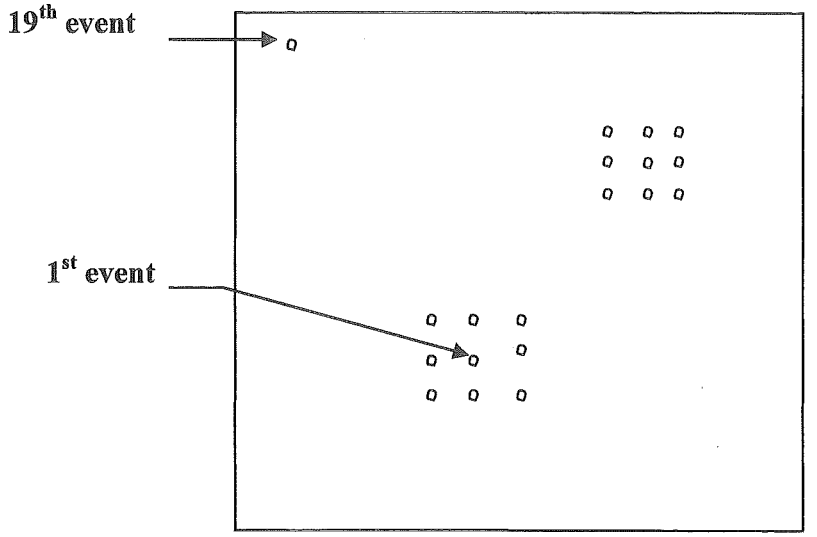


Figure 5.03: Location plot of 19 events

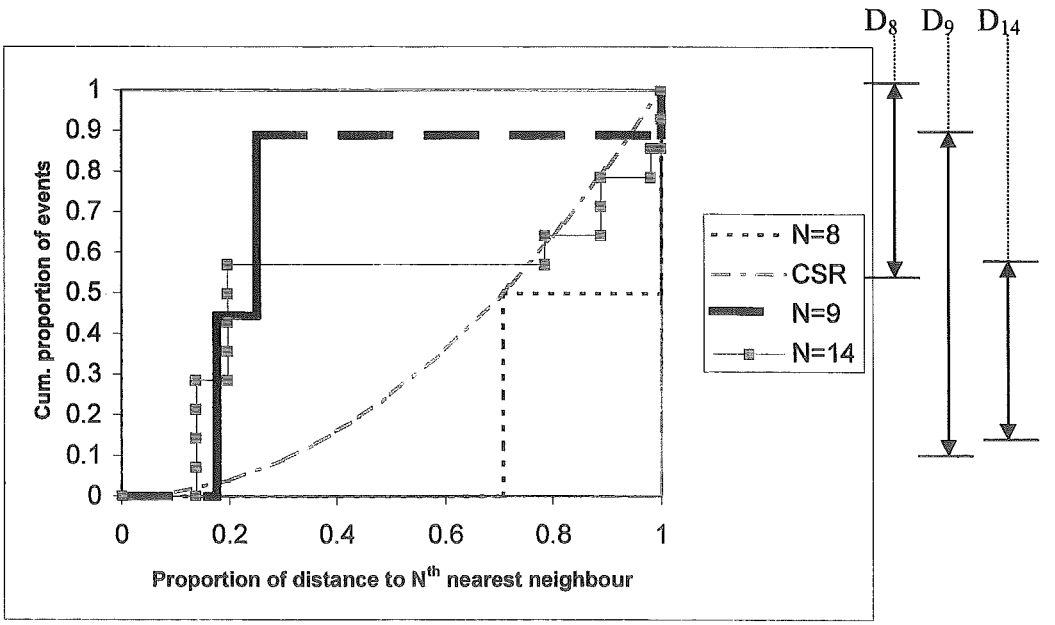


Figure 5.04: Cumulative distribution function

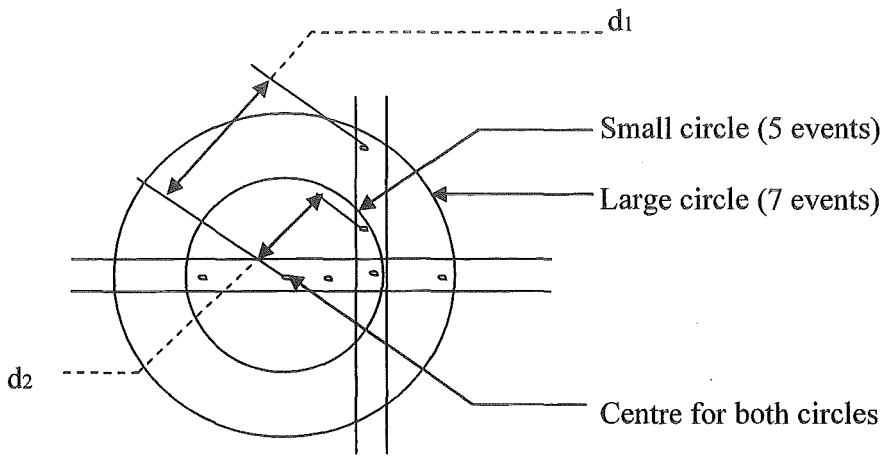


Figure 5.05 (a): Accident locations at an intersection.

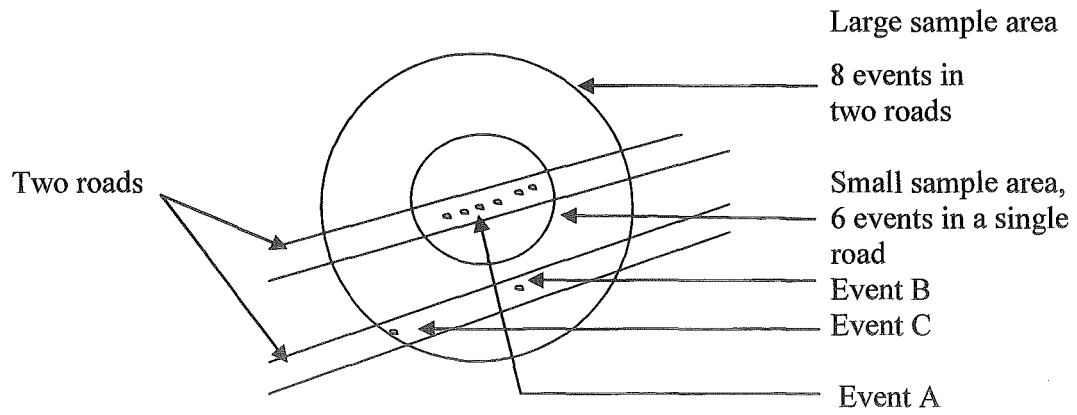


Figure 5.05 (b):Test-location and nearest-neighbour locations at non-intersecting roads.

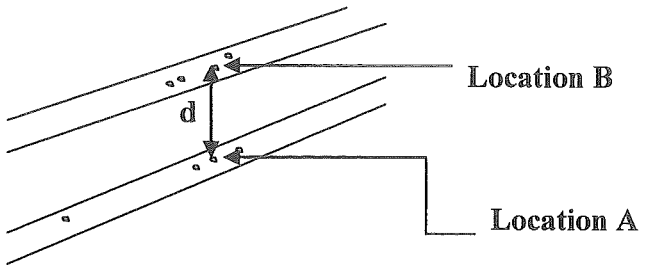


Figure 5.06: Test-location and nearest-neighbour locations of accidents

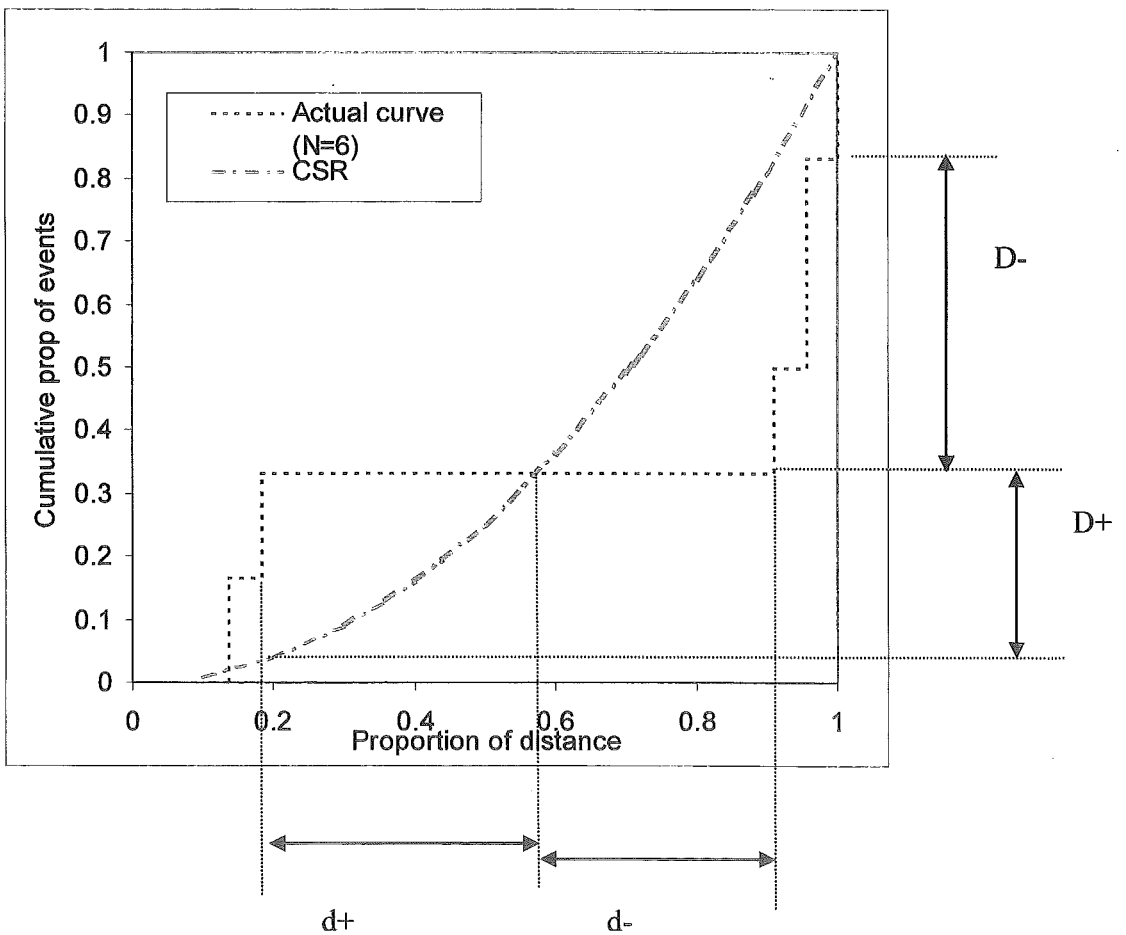


Figure 5.07: Cumulative distribution function for Figure 5.06

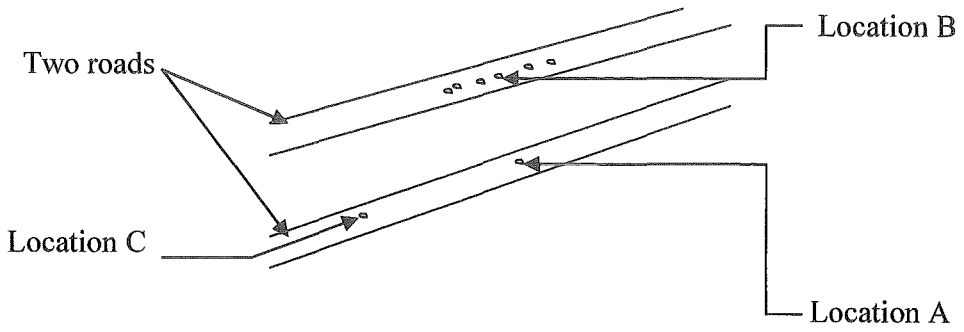


Figure 5.08: Accident locations (events) in a road network

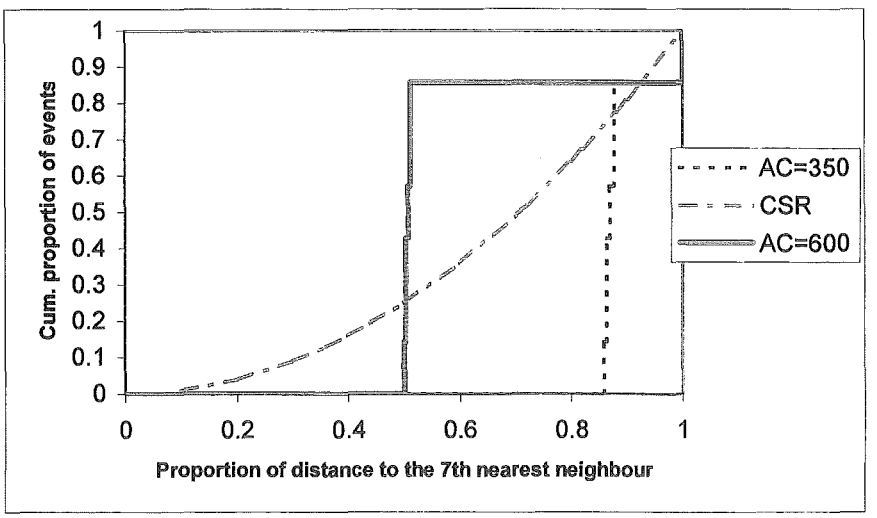


Figure 5.09: Cumulative distribution functions for the test-location A (shown in Figure 5.08)

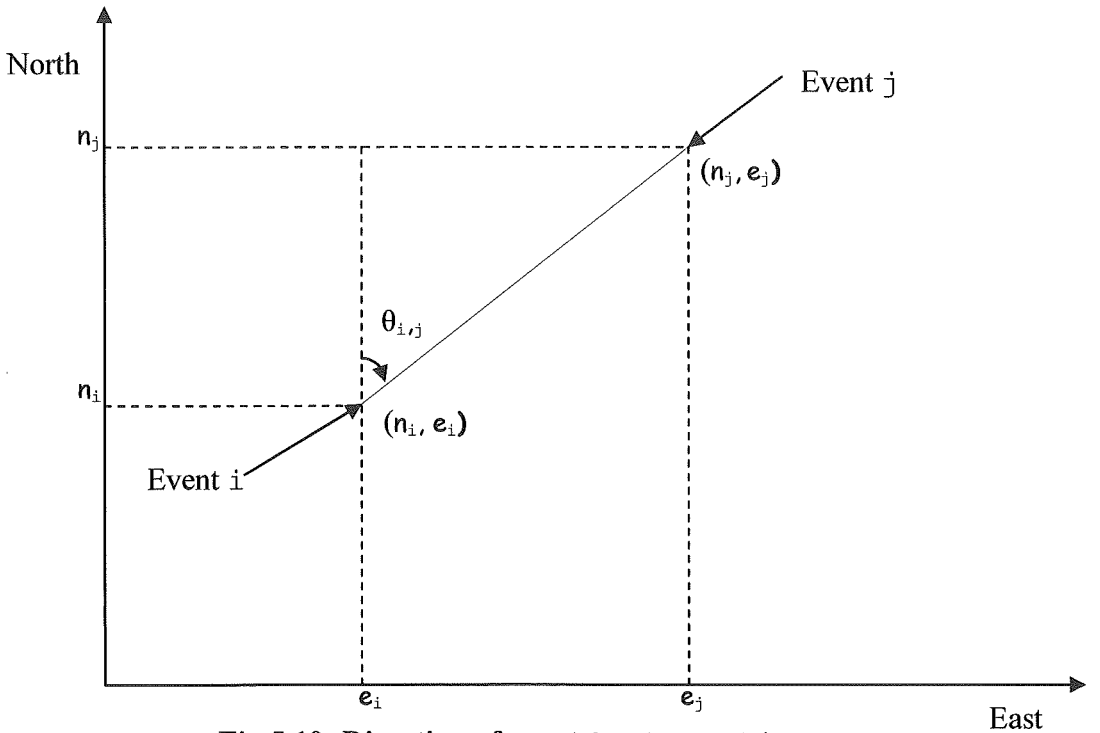


Fig 5.10: Direction of event i to event j

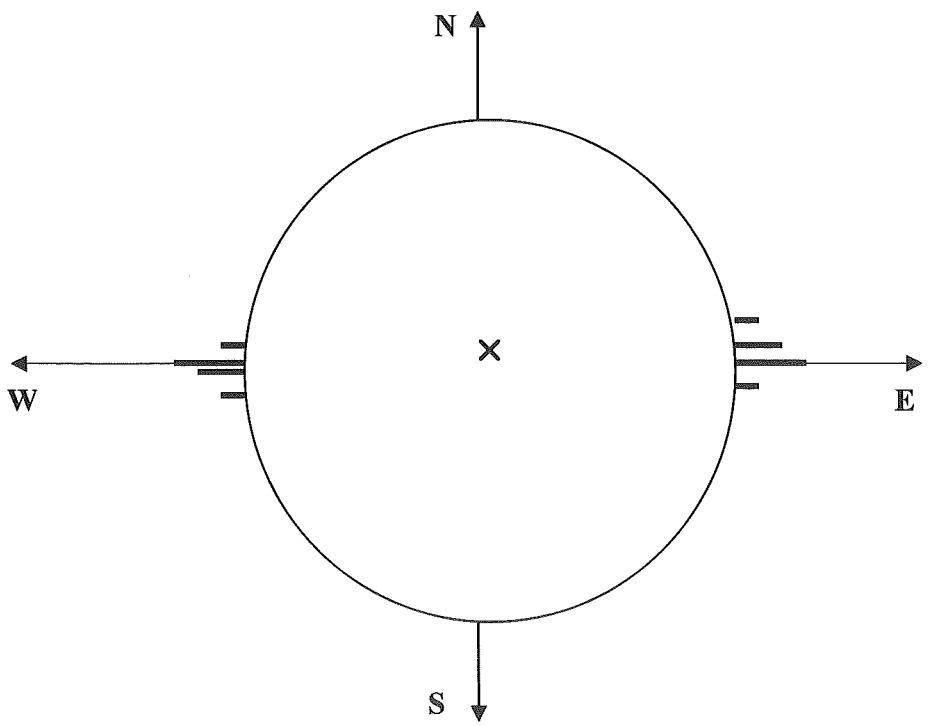


Fig 5.11: Circular histogram for direction distribution

[Figures 5.10 and 5.11 extracted from Upton and Fingleton [1989]]

Chapter 6

QUADRAT ANALYSIS

6.1 Introduction

A spatial distribution can be analysed by sampling the data and analysing the samples. In quadrat analysis the data are divided into small groups that belong to sub-areas. Sampling can be either regular or random. With regular sampling the area is systematically divided into regular quadrats, which are consistent in size and shape. Regular quadrats cover the whole area but random quadrats generally do not. Random quadrats are also consistent in size and shape, but are located randomly within the area. Whichever way the quadrats are chosen, the objects or events are counted within each quadrat, giving the quadrat counts. This process finally converts the spatial information in the data into a two dimensional array, which contains the sequence numbers of quadrats and the respective counts. The array gives the spatial distribution of counts of given areas. Completely spatially random (CSR) is the benchmark for assessing spatial distributions. The array from the given data can be compared with the array for a CSR distribution with the expected intensity equal to that of the given data.

Spatial distributions of accident locations can be identified by the same method. Accident location data are convert to a two dimensional array, which contains counts of accidents within each quadrat and the corresponding sequence number of the quadrats. To avoid shortcomings while applying the quadrat analysis method for accident distribution, and to make use of prior knowledge of the accident data (i.e. the tendency to cluster at points, along routes, or in some sub areas), this method is modified for analysing accident data and some powerful analysis tools are applied. The details are discussed in this chapter.

6.2 Problems and alternative methods

Cressie [1993] and Ripley [1981] pointed out the following problems associated with the quadrat method: the selection of quadrat positions and the size of the quadrats is arbitrary and affects the analysis results;

- the relative positions of events are not considered;
- the loss of information due to the single scale measurement from the pattern (i.e. single quadrat size).

Both of these authors were concerned mostly with field measurements related to forestry, but the occurrence of traffic accidents at sites is completely different from the occurrence of trees in forests, for which the data set contains accurate coordinates of tree locations in a two dimensional continuum. In accident investigation there are two separate issues, namely the location where accidents can occur and the precision of locating an accident. Trees can grow in most of the area in the field but a traffic accident is generally restricted to those places with vehicular traffic (e.g. public car parks or roads but not private property, grass field or buildings). The exact location of a tree is a fixed point, but the location of a traffic accident is not a precise point. It is within an area called the accident site and is often hard to locate precisely. That is, tree distributions in forests and accident distribution on roads are two somewhat different cases.

Ripley [1981: pg130] mentions that "quadrats are reported to be difficult to use in forestry and distance methods have been used". Although the location of a tree is a fixed point, its exact co-ordinate may not be available, but for accident analysis the approximate co-ordinates of the accident locations are often available. Anujah [1997] concluded that the quadrat method gives satisfactory results for the basic distributions (i.e., point clusters, line clusters and CSR) but it could not properly identify combinations of point clusters and CSR. This could be because a single arbitrary quadrat size was chosen in that study. Nicholson [1999] mentioned that the nearest neighbour analysis result is sensitive to the number of nearest-neighbours. Similarly, the quadrat analysis result is also sensitive to the quadrat size (i.e. the number of events within the quadrats is related to quadrat size). These type of problems need to be considered to find an improved method. The problems are:

1. the size of quadrat; the results depend on the quadrat size;

2. the position of quadrats; the results depend on the position of quadrats which are selected on or outside of the roads and the accuracy of quadrat counts; errors in accident co-ordinates affect the accuracy of quadrat counts.

When developing a new method of analysis we need to consider the above points. The problems with quadrat analysis are considered one by one and alternative methods for accident analysis are discussed. The discussion focuses on methods which will be useful for identifying and monitoring accident patterns.

6.2.1 *Quadrat positioning*

Regular or random positioning of quadrats is possible. Using regular quadrats is easy and convenient, but the average count from regular quadrats is influenced by empty (accident-free) quadrats. This will cause problems in the assessment of the spatial pattern of accidents, because of areas that do not have vehicle access (e.g. grassland, buildings or lakes) will not have accidents. If we assume that all an area, which includes a sparse road network and non-vehicle access areas, can have traffic accidents, then the analysis results will be influenced by the ratio of the vehicle-access area to the non-vehicle access area. The analysis is aimed at determining whether the accidents are concentrated or sparse on the road network, but not the concentration of the road network. If the sparseness of the road network influences the results, then any decision based on the analysis results may well be incorrect. The random quadrats method has the same problem. The positioning of quadrats needs to be decided carefully and an alternative method, which can focus on the vehicle access areas, needs to be found.

6.2.2 *Quadrats on roads*

The quadrat centres should be chosen on the places where vehicle access is available. If regular quadrats are selected on roads then empty-quadrats can occur and influence the analysis result. Suppose, a road has very little traffic compared to other roads, then there is much less chance of an accident on that road (or section of the road) within a two or three year period. The regular quadrat method may tend to measure the spatial distribution of the road network, rather than the spatial distribution of accident-locations on the road network.

To investigate the spatial arrangement of accident-locations and to ensure selected quadrat include roads, accident-centred quadrats have been used from now on, rather than random or regular quadrats.

6.3 Accident-centred quadrat method

The character of each accident location (i.e. whether locally-dense or locally-sparse) is analysed using the accident-centred quadrat method, with the quadrat counts indicating whether the location is locally dense or sparse. Consider the example of two clusters with 24 accident locations, as shown in Figure 6.01a. Consider a quadrat centred at a location (A) with radius of 5 units. The radius is increased by nineteen steps (5 units up at each step) to 100 units, with the number of events within the quadrat being noted at each step. The plot of counts versus quadrat radius is shown in Figure 6.01b. A random distribution of accident locations is shown in Figure 6.02a and the corresponding plot of counts versus quadrat radius is shown in Figure 6.02b for the quadrat centre at the location B.

In Figure 6.01b, up to 20 units radius the counts increase, from 20 units to 85 units radius there is little change, but after 85 units radius the counts increase again. This variation of count indicates that the first cluster (i.e. the cluster around point A) ended near 25 units radius, and the next cluster started at 85 units distance from the point A. In this plot the two clusters are clearly indicated, but in Figure 6.02b there is no indication of clusters because the counts continue to increase steadily with the radius. Figure 6.02b indicates locally-sparse counts and Figure 6.01b indicates locally-dense counts. This indicates that the selected event (A) is within a cluster.

The mean and the variance of the quadrat counts were calculated for 5 units radius quadrats, which were centred on each of the 24 accident locations. This was repeated for quadrat radii of 10,15, ... , 100 units, and the mean and the variance of the counts were plotted against the radii. This was done for each of the two spatial distributions shown in Figures 6.01a and 6.02a. The two plots are shown in Figures 6.01c and 6.02c. It may be noted in Figure 6.01c that there is no increment in the mean count from 35 to 80 units radius, but in Figure 6.02c there is a steady increase in the mean count from 20 to 100 units radius. It appears that the

slope of the mean count line is a useful indicator to distinguish point cluster and CSR distribution.

In Figure 6.01c, both the variance and mean increased between 5 and 20 units of radius, which indicates that within 20 units of radius the points are distributed differently. Between 20 units and 40 units of radius the variance in quadrat counts decrease, which means the differences in the quadrat counts are decreasing. The variance between 40 to 75 units of radius is almost zero, which does not mean that the points are regular, but it means the quadrat counts are almost all the same. This also indicates that events in each cluster are within circles of 35 units radius. When the radius increases from 80 units up to 100 units the variance increases rapidly and approaches the mean. This occurs because the quadrats start covering other neighbouring clusters.

In Figure 6.02c, the variance is below the mean up to 30 units radius, is above the mean up to 85 units, and then is below the mean. This implies that the variance in quadrat counts up to 30 units radius is small and the variance in quadrat counts from 30 up to 85 units radius is high. From 85 units radius up to 100 units radius the variance decreases. Figure 6.01c indicates equal count of quadrats at the 5 unit radius and between the radius 40 and 70 units but Figure 6.02c indicates equal quadrat counts between the radius 5 and 10. The mean and variance of quadrat counts indicates that it is not regular and is not clustered. Randomly distributed accidents may indicate little clustering and little regularity for different quadrat sizes, but there is no indication of excessive cluster or regularity. A detailed explanation on the performance of the variance and the mean profile is given in Chapter 7 and 8.

The coefficient of variation (CV) for the distributions shown in Figures 6.01a and 6.02a is between 0 and 0.6 (less than one). In the point cluster distribution the CV is less than one because two similar sizes of clusters were selected and so the variation in quadrat counts is small. If the data contains point clusters and random events, or different size of clusters, then there could be a high variation in the quadrat count and CV. In practical situations, some accident locations in road networks are random, regular and clustered but there may be only clusters, only random locations or only regular locations. Therefore it is necessary to consider analysing when the data contain random, mixture of clustered and regular distributions. The CV might be more sensitive when analysing mixed distributions of accidents. Mixed distributions are investigated in Chapter 7, using the CV.

Figure 6.01c, shows the variances for most of the quadrat radii are nearly zero. If the mean is high and the variance is nearly zero then level of clustering of all the accident locations are similar. The sites of clusters are contributing equally to the accident total as is occurring in Figure 6.01a (two equal site clusters), and Figure 6.01c shows that the variance is nearly zero for the quadrat radii between 30 and 80 units. If accident counts for all the sites of clusters contribute equally to the accident total for a road network, then it is difficult to arrange the sites in the order in which accident reduction plans will be implemented. In this case further investigation (e.g. analysing the type of accidents) may be needed.

Bailey [1995] and Ripley [1981] mentioned that if the index of clumping (ICS), defined in section 3.3, is greater than zero then clustering is indicated but if ICS is less than zero then regularity is indicated. The coefficient of variation (CV) indicates any excessive cluster or regularity. If a spatial pattern is more regular then the quadrat counts will be more uniformity and will therefore have a relatively small variance, when compared to the size of the mean. If there are clusters, then some quadrats will have large counts and some will have small counts, and the variance of the quadrat counts will be relatively large. When considering Figures 6.01c and 6.02c, the CV alone will not give the details such as how the mean and variance varies with quadrat radius. As discussed in Chapter 2, the CV is a useful index when the overall density of accident location changes in the road network, and these three indices (CV, mean, variance) are used for further investigation in Chapter 7.

Two cases need to be considered when CV is greater than one;
Case I: the percentage of the quadrat counts which are above the mean, is greater than the percentage of the quadrat counts, which are well below the mean (Table 6.01).
Case II: the percentage of the quadrat counts, which are above the mean, is lower than the percentage of the quadrat counts, which are well below the mean (Table 6.02).

In our case, we assume single isolated accidents are random accidents, while the clusters give much large quadrat counts. Therefore the computed percentage of single-accident quadrats (i.e. the number of single accident quadrats/ total number of accidents) is approximately equal to the estimated percentage of isolated accidents which could be considered to be random.

The two examples shown in Tables 6.01 and 6.02 are used to explain how to identify the Case I and Case II, using the percentage of single accident quadrats, and two more indices;

1. the maximum count (i.e. maximum count identified from the quadrats having the maximum number of accidents) and
2. the percentage of quadrats having accidents above the mean accident count.

Note that the quadrat count distributions shown in Table 6.01 and 6.02 are not from CSR distributions and are not from the distributions in Figures 6.01a and 6.02a.

Table 6.01:Quadrat count distribution, Case I

Accident count	Frequency of quadrat
6	**6**
5	**5**
3	**3**
1	**3**

Mean = 4.29 Variance = 120.57 CV = 2.56

Variance > mean

Percentage of single accident quadrats = 17

Percentage of quadrats having count above mean = 64

Maximum count = 6

Level of clustering = ICS = (Variance / Mean)-1 = 27.10

64% of quadrats have counts between 4.29 and 6 (above mean)

17% of quadrats have a count of one.

Note that in Table 6.01 the difference between the maximum and the mean is 1.71, which is less than the mean, and the percentage of quadrats having accident counts greater than the mean is higher than the percentage of quadrats having single accidents, and this indicates that point clusters are dominant in this distribution.

Table 6.02: Quadrat count distribution, Case II

Accident counts	Frequency of quadrat
5	5
3	3
1	11

Mean = 2.37 Variance = 56.42 CV = 3.17

Variance > mean

Percentage of single accident quadrats = 58

Percentage of quadrats having a count above mean = 42

Maximum count = 5

It appears random accidents are dominant

Level of clustering = ICS = (Variance / Mean) -1 = 22.82

42% of quadrats have a count between 2.4 and 5 (above average)

58% of accident quadrats have a count of one

Note that in Table 6.02 the difference between the maximum and the mean is 2.63, which is greater than the mean and the percentage of quadrats having single accidents is higher than the percentage of quadrats having accident counts more than the mean. This indicates that random accidents are dominant in this distribution. In both cases the variance is greater than the mean. The mean and the variance for Case I is higher than for Case II but the CV for Case I is lower than for Case II. The two cases can be compared using six indices: mean, variance, CV and the following three indices:

1. percentage of quadrats having single accidents (PSA);
2. the maximum count (MaxCou) and
3. percentage of quadrats having above mean count (QAM -%).

The index ICS shows the cluster size differences between the two cases (i.e. ICS for case I is higher than the case II), but the 3 indices (PSA, MaxCou and QAM - %) are providing more information. The quadrat size is an important factor for deciding the percentage of random accident locations, as noted in the discussion of the characteristic length of clusters in Chapter 2. Therefore the three indices (PSA, MaxCou and QAM -%) are further investigated for varying quadrat radii, using the distributions shown in Figures 6.01a and 6.02a.

The PSA was computed from the distribution shown in Figure 6.01a, for 5, 10, 15,.....,55 units radius quadrats. This computation was also done for the distribution shown in Figure 6.02a, and Table 6.03 shows the results.

Table 6.03: Percentage of quadrats with single accident

Quadrat radius (units)	% of quadrats with single accident	
	Point clusters	Random events
5	100	100
10	76.47	100
15	4.76	95
20	0	93.75
25	0	64.71
30	0	42.11
35	0	30.43
40	0	23.81
45	0	9.09
50	0	4.76
55	0	0

The 2rd column in Table 6.03 indicates that for point clusters, there are no quadrats with a single accident when the quadrat radius exceeds 15 units. The 3nd column indicates that for random events, there is no quadrat with a single accident when the quadrat radius exceeds 50 units. Table 6.03 shows that the PSA for the point cluster distribution is less than for random distribution for most of the quadrat radii, and hence the PSA may help to distinguish between the point cluster and random accident distributions.

For the distribution shown in Figure 6.01a, Figure 6.01d show the variation (with quadrat radius) of the maximum quadrat counts, the percentage of quadrats with single accidents, and the percentage of quadrat having counts greater than the mean. Figure 6.02d shows the results obtained from the distribution in Figure 6.02a. For the cluster distribution in Figure 6.01a, the maximum count and the percentage of quadrats having counts greater than the mean are almost constant for radii between 35 and 70 units. For the random distribution in Figure 6.02d, the quadrat counts is not constant for the quadrat radius in the range 5 to 100 units. The profile of PSA and QAM - % (see Figures 6.01d and 6.02d) show notable difference between the point cluster and random distributions, and hence the PSA and QAM - % are investigated further in Chapter 7. The characteristic length (35 units) and the maximum number of events (11) in a cluster can be noted from Figure 6.01d.

The two distributions shown in Figures 6.01a and 6.02a are not mixed distributions. To identify the dominant distribution in a mixed distribution we may investigate the quadrat radius with 50% of quadrats having single accidents. To identify an unknown accident distribution by investigating the minimum quadrat radius which gives zero PSA, could lead to a wrong conclusion about the distribution, if a point cluster distribution has an isolated event (an error in the data). For any distribution (mixed or un-mixed) investigating the quadrat radius with 50% of quadrats having single accidents from the plotted PSA profile may be helpful. As an example in Figure 6.01d the cluster distribution, the PSA is 50 when the quadrat radius is approximately 12 units. This radius is higher (i.e. approximately 28 units as shown in Figure 6.02d) for the random distribution.

As explained in the above paragraphs, the following five indices are used to analyse spatial distribution.

1. mean count (MEAN)
2. variance of count (VARIANCE)
3. percentage of quadrats with a single accident (PSA)
4. percentage of locations having counts above the mean (QAM-%)
5. maximum count (MaxCou)

The percentage of single accident quadrats is used to decide whether the percentage of locations is random or cluster, but the maximum and mean count may still be needed. Suppose most of the accident locations are random and a substantial proportion of accidents locations are clustered at sites, and one or two locations have a very high maximum count.

In this case, the appropriate accident reduction plan could be the single site plan for those locations having high accident counts, because the single site plan generally gives a better rate of return, as discussed in Chapter 1. This decision would need to be confirmed with further investigation, to obtain details of the cost of treatment and expected reduction of accidents. If the proportion of accident locations having clustered accidents is lower than random accident locations in a mixed distribution (i.e. point cluster and random distribution), then the site action plan may not be appropriate, but the locations having very high maximum count suggests that the first preference is the site action plan.

Imagine another situation in which most of accident locations are clustered and a substantial proportion of accident locations are random, and the maximum and the mean counts are very low, then further detailed investigation would be necessary to identify the appropriate accident reduction plan. The reason might be that the cheapest accident reduction treatments (site treatments) have already been implemented. When considering the cost effective treatment, the route or area action plan could be appropriate. So the indices (i.e. mean and maximum counts) may be useful for selecting the appropriate accident reduction plan.

To identify line clusters, the maximum count for each length of road section needs to be investigated. If the high accident-count sections (length $\geq$500m) are not close to each other and are spread through out the road network, then those sections are indicated as line clusters. If the high accident count sections are close to each other and are spread within a portion of the road network, then that portion of the road network is indicated as an area cluster (see Figure 3.20). The accident-centred quadrat method can be used to identify a sequence of increasing accident count road sections and to find the section which has maximum count for different quadrat radii. In the method introduced here, the accidents are counted on the same road of the quadrat centre, as shown in Figure 6.03. Although the figure shows straight roads the analysis results will not affected if the roads are not straight. The mean quadrat count will give a measure of line clustering. The maximum concentration (i.e. accident count per unit length of road) within the length of road in the road network could be found using the same method. The objective is to identify black routes, that is roads having numbers of accidents above the average for the type or class of road.

6.3.1 *K function*

This is one of the more powerful tools for analysing spatial distributions [Cressie 1993]. The only information used is the distance between accident locations, although distance does not fully represent the spatial information. The K-function was discussed in Chapter 3 and the application of the K-function in nearest neighbour distance method was discussed in Chapter 5. The application of the K-function in the accident-centred quadrat method is discussed here.

The disadvantage of that the K-function approach is that it cannot differentiate line clusters and CSR [Nicholson 1995]. If the accident data are on lines (i.e. roads), then the test method needs to first determine this and then analyse whether the accidents are concentrated at certain locations on lines. The intensity (λ) must represent the total population and depends on the number of accidents and size of the selected road network. Estimating K(h) involves the estimation of the intensity (λ) and the expected number of events.

To analyse accident distributions the accident-centred quadrats can be combined with the K-function method. For a quadrat radius of h, the mean count (M_h) is calculated from the following formula: $M_h = \left(\sum_{i=1}^{n} Q(i)\right)/n$, where Q_i is the quadrat count for the quadrat with radius h and centred on the i^{th} accident, n is the total number of accident locations within the road network being investigated and $E(Q_i)$ is the expected quadrat count. In practice $E(Q_i)$ is difficult to calculate, so the expected count ($E(Q_i)$) is approximated by the mean count (M_h).

For quadrat radius h, $M_h \approx E(Q_i)$ for large n. In addition, $E(Q_i) = \lambda K(h)$, where K(h) is the expected accident count (i.e. $E(Q_i)$) divided by the intensity (λ). That is, $M_h = \lambda K(h)$.

If the process is random then the estimated K(h) equals πh^2, and it follows that $M_h = \lambda \pi h^2$, where $\lambda \pi$ is constant for an intensity of λ.

If we plot M_h against h then the expected profile will be a quadratic form.

The mean count M_h is plotted in Figures 6.01c and 6.02c for a clustered distribution and for a random distribution. The mean profile in Figure 6.02c is similar to Figure 6.04, which was extracted from Jain and Dubes [1988: 212]. Jain and Dubes noted that Diggle [1983] demonstrated how a plot of estimated-K(t) versus radius t can be compared with the theoretical-K(t) function for a Poisson process, to test the random position hypothesis (H_0) using Figure 6.04. This figure illustrates the upper and lower envelopes, which is a confidence band around the theoretical K(t) function for random distributions. If the estimated K(t) lies within this band for a quadrat radius of t then H_0 is not rejected.

Let the maximum quadrat radius used for an analysis be r and the mean count for that radius be M_r. In Figure 6.05 the normalised plot of M_h / M_r against h/r is shown for the cluster distribution and a random distribution shown in Figures 6.01a, 6.02a respectively. In this figure the cluster profile is substantially different from the random profile. This method of analysis is further investigated in Chapter 7.

The number of clusters can be estimated approximately from the number of quadrat counts, which are greater than the mean (M, say). If Q(i) is greater than expected quadrat count ($E(Q_n)$) then the i^{th} location "appears to be a clustered location". The term "appears to be a cluster location" means that, there is a possibility the location might not be a cluster. In CSR distribution the variance of the quadrat counts is equal to the mean but not zero. That is, some counts are more than the mean and some are less than the mean. That is the reason why one of the five indices introduced above is the percentage of quadrats having a count above the mean, rather than using the notation as the percentage of cluster locations.

If we consider line quadrats (i.e. the quadrats being sections of the centre line of roads, discussed in section 3.3.6) the expected quadrat counts is approximately proportional to quadrat length then the distribution is possibly a line cluster. If we plot the mean number of counts against h it is likely to be a straight line going through the origin and the gradient is likely to be λ. If the expected quadrat count is nearly proportional to the square of quadrat length then it may be point cluster or CSR distribution. If it deviates excessively from a straight line then it is a clustered or regular distribution as shown in Figure 6.06. This figure shows that for a quadrat radius between zero and "a", the distribution is regular, while for a

quadrat radius between "a" and "b" the distribution is clustered. This is further analysed in Chapter 7.

6.3.2 *Skewness and kurtosis*

The shape of frequency distributions generally described by skew and kurtosis was mentioned in Frank and Althoen [1994]. The skewness and kurtosis could be used to investigate the frequency of sparse locations and dense locations. If the number of dense locations is less than the number of sparse locations then the frequency distribution is skewed to the right (positive skew), as shown in Figure 6.07a. If the quadrat counts distribution is a Poisson distribution then that quadrat count distribution is positively skewed. This will happen when more observations lie below the mean than above the mean. If the number of sparse locations is less than the number of dense locations then the frequency distribution is skewed to the left (negatively skewed), as shown in Figure 6.07b. This will happen when more observations lie above the mean than below the mean. If the number of sparse locations and the number of dense locations is equal then the frequency distribution is not skewed, as shown in Figure 6.07c. The effect of quadrat radius on the shape of the count distribution is discussed in Chapter 7.

The discussion in Chapter 1 about "natural progression from site plan to route plans to area plans", means that at the start accidents are generally clustered at sites, and after the single site plans are applied the accidents become more dispersed. In this process the level of point clustering reduces and the frequency of sparse locations becomes higher than that of dense locations. Hence, the frequency polygon may be expected to change from skewed to the left to skewed to the right. The skewness could be used to identify the accident distribution.

A general form of skew (SK) is given by

$$SK = \left(\sum_{i-1}^{n}(Q_i - \overline{Q})^3\right) f_i \Big/ N \times s^3$$

where Q_i is the i^{th} quadrat count, $\overline{Q}$ is mean quadrat count, f_i is the frequency of Q_i , N is the number of counts and s is the standard deviation. The Pearson skewness coefficient (sk) is defined,

$$sk = 3(\bar{Q} - \tilde{Q}) / s$$

where $\tilde{Q}$ is the median of the quadrat count.

The SK is a third order property in count. Ripley [1981] noted that " there is some evidence [Julesz, 1975] that the human eye is most aware of the second order properties of planar point patterns". Investigating the third order property by viewing location plots, seems to be difficult. The skewness is used to compare the frequencies of the lower accident counts with the frequencies of higher accident counts, where the lower and higher accident counts are relative to the mean count of the distribution.

The fourth order property of a distribution is given by kurtosis (KUR). It is a term that refers to the sharpness of the peak of a frequency distribution. Frank and Althoen [1994] noted that the kurtosis is given by,

$$\text{KUR} = \left(\sum_{i-1}^{n} (Q_i - \bar{Q})^4 \right) f_i \Big/ N \times s^4$$

For normal distribution KUR is equal to 3. If the frequency distribution is more peaked than the normal distribution (i.e. leptokurtic) then KUR>3 and if the frequency distribution is flatter than the normal distribution (i.e. platykurtic) then KUR<3.

A few examples indicating different values of kurtosis are shown in Figures 6.08a, b, c, d, e and f. Frank and Althoen [1994] stated that the following important points should be noted when investigating the frequency polygon by using the kurtosis and skewness.

1. The values in the shoulders of a distribution exert more influence than does the peak.
2. The values in the shoulders of a distribution have a greater effect on the denominator and the values in the tails have greater effect on the numerator of the equation noted above for KUR.
3. The higher moments involve higher powers of the input data and are less robust than lower moments.

For these reasons, considerable caution is necessary while interpreting the results.

Consider the two frequency distributions shown in Figures 6.09a and 6.09b, which are derived from the two spatial distributions shown in Figures 6.01a and 6.02a. In Figure 6.09a, the frequency distribution is skewed to the left, the skewness is less than zero and kurtosis is

greater than three. This means the number of cluster locations is higher than the number of sparse locations. In Figure 6.09b the frequency distribution is skewed to the right, the skewness is greater than zero and kurtosis is less than three. This means cluster locations are fewer than sparse locations.

The kurtosis may be used to investigate the distribution of quadrat counts. For example, if most of the sites in a road network have very low accident counts which are approximately equal then the frequency distribution tends to peak at that count, as shown in Figure 6.08a. In this case an area action plan will be appropriate. If the frequency distribution shows more flatness than a normal distribution then the appropriate plan could be a single site plane. The reason is some sites having relatively high numbers of accidents and some have relatively low numbers and the distribution may indicate that the accident count distribution is not a CSR. The frequency distribution shows flatness (i.e. platykurtic, rectangular and slightly bimodal) when there are several sizes of clusters in the distributions.

A highly bimodal frequency distribution suggests two distinct populations of quadrats; one with low accident counts and one with high accident counts. The latter are candidates for black spot treatment, if the quadrats are small. The size of the quadrats plays an extremely important role in determining which accident reduction plan is best. If the quadrats are large, then even if the accident counts are large, it is difficult to conclude whether the quadrats include a few 'black spots' or lots of 'grey spots'. Therefore quadrat size needs to be small to arrive at a conclusion.

6.3.3 *Non-overlapping accident-centred quadrats*

Overlapping together with non-overlapping quadrats were analysed in the previous sections but in this and the next sections non-overlapping quadrats are considered as a special case.

The discussion on Sections 6.3, 6.3.1 and 6.3.2 were mainly focused on the spatial characteristic of each accident. In that case, the analysis method needs to accommodate both overlapping and non-overlapping accident-centred quadrats. The quadrats centred on an isolated accident will be a non-overlapping quadrat and the quadrats centred on clustered accident locations will be overlapping quadrats.

Stuart and Keith [1987] mentioned that " when we consider sampling from a finite population of size N, the notion of randomness implies that each individual has the same chance of selection". When investigating the accident occurrence at accident locations, whether they are clustered or random locations, then overlapping quadrats can be accommodated in the samples. When the investigation is focused on sites, whether there are clustered accident sites or random accident sites, then the overlapping quadrats need to be eliminated from the sample. The overlapping quadrats could possibly repeat the information by partly or completely covering the same sites. The method must produce a set of sub-areas (i.e. quadrats) in which each sub-area has approximately equal probability of being selected, and only then will the calculated mean or variance be an unbiased estimate. Therefore the non-overlapping quadrat method is used to analyse the distribution of accident counts at different quadrats locations, where accident occurred.

6.3.4 *Truncated Poisson distribution*

Cressie [1993] and Ripley [1981] noted that for CSR distributions the non-overlapping quadrat counts will be Poisson distributed. In this research accident-centred quadrat are used, hence zero quadrat counts are not possible. Therefore the accident-centred quadrat count frequency distribution cannot be compared with the Poisson distribution but it can be compared with truncated Poisson distribution.

To test the randomness of the quadrat counts, the distribution of counts is usually compared with the Poisson distribution. The accident-centred non-overlapping quadrats do not have zero counts. Therefore the distribution of counts should be compared with the truncated Poisson distribution, which is the Poisson distribution truncated between one and zero.
The probability mass function for the Poisson distribution is,

$f_x(x) = (e^{-m} m^x)/x!$ for x = 0,1,2,3..... (Eq. 6.1)

where m is the expected value of x.

The probability mass function for the truncated Poisson distribution is;

$f_x(x) = (e^{-m} m^x)/[x!(1-e^{-m})]$ for x = 1,2,3,..... (Eq. 6.2)

The non-overlapping accident-centred quadrat count x can take a value greater than zero. Nicholson [1995] noted that for the truncated Poisson distribution

the expected value of x = m / $(1-e^{-m})$,

the variance of x =m $[1-(m+1)\ e^{-m}]/(1-e^{-m})^2$.

Note from the above two expressions that the variance is not equal to the mean for the truncated Poisson distribution. The coefficient of variation is;

CV = $\{m[1-(m+1)\ e^{-m}]/(1-e^{-m})^2\}^{½}$ / [m / $(1-e^{-m})$].

The method used to compare the non-overlapping quadrat counts distribution with the truncated Poisson distribution is as follows. The accidents are ordered randomly and then the accident-centred quadrats are selected without overlapping from that order. Upton and Fingleton (1985) and Griffith (1988) discussed the role of the sample size, and suggested that when the number of quadrats is less than 30, then the quadrat counts may well fail to provide any useful information about the properties of the spatial distribution. That is, the number of quadrats needs to be greater than 30. From the quadrat counts the mean and the variance are calculated. These two indices are used to analyse whether there is any excessive clustering or regularity. The quadrat diameter and the total length of the road network with accidents influence the number of non-overlapping quadrat samples. The observed distribution of counts is tested for randomness, using the Chi-square test with the expected truncated Poisson distribution.

Table 6.04: Comparison of quadrat counts (CSR and truncated Poisson distributions)

Counts	Observed frequency (O)	Expected Probability	Expected Frequency (e)	$\{(o-e)^2\}/e$
1	27	0.26	26	0.04
2	41	0.29	29	4.97
3	19	0.23	23	0.70
4	6	0.13	13	3.77
5,6,7	7	0.09	9	0.44
				9.91

Consider the quadrat count distribution obtained shown (Table 6.04). The quadrat counts are shown in the first column and the frequency of quadrats with the corresponding counts are

shown in the second column. The expected probability is calculated from the truncated Poisson distribution (equation 6.2), with an expected value (m) is equal to 2.31 (the mean of the quadrat counts in Table 6.04). The result indicates that the truncated Poisson distribution is not a valid model for the count distribution at the 10 % significance level with 3 (i.e. the number of classes, minus one, minus the number of model parameters) degrees of freedom.

6.3.5 *Quadrat shape and size*

Quadrats can be any shape (e.g. circular, square or polygon). The road edges are usually two parallel lines, so the rectangular shape is a sensible choice. It is necessary to consider the practical problems in locating a quadrat in which the two edges of each quadrat are parallel to the road edges, without knowing the details of road edges (i.e. coordinates). If the aim is to identify whether an accident location is dense or sparse then it is more practical to have the quadrats boundaries at a constant distance from the centre, so the shape of the quadrats will necessarily be a circle.

There is no major problem when the diameter of the quadrat is only slightly greater than the road width, but the quadrat size is a big concern when the diameter is substantially greater than the road width. A small size (approximately 70m) is useful for identifying the point cluster distribution or CSR distribution. To identify the line cluster distribution, the quadrat sizes need to be comparatively larger (the characteristic length of a line cluster is discussed in Section 2.3.3). If the quadrat size is comparatively larger then the quadrats will extend too far outside the road (more than in Figure 6.10). The shaded areas shown must be neglected because these are outside of road. In that case, it is necessary to use the length of roads as quadrats. For identifying the line clusters from an accident distribution, which deviates from CSR distribution, the quadrats are considered as the road sections rather than a circle (illustrated in Figure 6.03).

The LTSA accident data system locates the accidents approximately on the centreline of roads. There is no real problem to analyse accident data by considering the quadrats as lines (i.e. zero width) with the quadrats being sections of the centreline of roads. To achieve this the road name information in the accident data are used.

To identify spatial distributions, several quadrat sizes are needed. When the area of the quadrat increases then the spacing of the information also increases because a large quadrat covers a large distance. If the analysis needs to identify, whether the accident locations are clustered or non-clustered, then the quadrat area needs to be larger than a cluster area. The practical difficulty is to find cluster size (i.e. area or length) because it is unknown at the stage of explanatory data analysis. In this case, the quadrat analysis may be done for different quadrat sizes. The range of the quadrat size may be based on the speed limits permitted on the roads and the types of accidents (see the discussion on the characteristic length of clusters in Chapter 2).

6.3.6 *Advantages and disadvantages of using accident-centred quadrats*

There are two issues to be considered. In clustered patterns, if the quadrat size selected for analysis is too small and the quadrats are centred on accidents, then the count for each quadrat could be one and the variance of the counts would be zero (unless empty quadrats are also considered).

Section of the roads are not included in the analysis because accidents did not occur on those sections. The size of the quadrat can increase to a reasonable extent and accommodate the road sections which did not have accident. If the quadrats are large then a bigger part of the quadrat area is outside the road. In this case the quadrats will be line quadrats, which is the quadrats being section of the centre line of roads as discussed in Section 3.3.6.

The analysis must always take into consideration the accuracy of the accident co-ordinates. The exact coordinates of an accident cannot be identified with certainty. So it is better to consider each accident as occurring within a small area, with the estimate of the distance between two crash locations being subject to error. In distance analysis, for example, the error in the co-ordinates of accident location may impact on cluster analysis results, but this is not considered further in this thesis.

6.3.7 *Dispersion measures*

The six indices (I, ICS, ICF, ICR, IP, MI) which may used to analyse quadrat count data were briefly discussed in section 3.3. Cressie [1993] mentioned that the results are dependent on the size of the quadrats, and the characteristic length of clusters is an important issue when selecting the quadrat size. The accident-centred quadrat method may be used to investigate these six indices.

Quadrats with zero counts are not possible with this method, and this needs to be considered when interpreting the results. The exclusion of zero count quadrats in the frequency of accident-centred quadrat counts will influence all the indices. Several CSR distributions were used to investigate these indices and the results are discussed in Chapter 7.

For Poisson distributed counts, the relative variance index I, which equals (variance / mean), is one. The second index is called the index of clumping or index of cluster size (ICS) and equals (I-1). If the variance is high compared to the mean, then I is high and so ICS is high, which indicates that the spatial distribution is clustered. If the variance is low compared to the mean then I is low and so ICS is negative, which indicates the spatial distribution is regular. For example, if the events are regularly spaced, the quadrat count does not vary, so the variance is zero, I is zero and so ICS is –1.

The third index is called the index of cluster frequency (ICF) and equals (mean / ICS). Ripley [1981] mentioned that " the ICF should measure the mean number of clusters per quadrat and so is proportional to the area A of the quadrat, whereas ICS is independent of the quadrat size or shape". This will be investigated using accident-centred quadrats in Chapter 7.

The fourth index is called the "index of mean crowding" (ICR) and equals (mean + ICS). Ripley noted that Lloyd [1967] "considered the ICR to represent the number of individuals sharing a quadrat with a typical individual, the two terms representing those in other clusters and those in the same cluster".

The fifth index is called the "index of patchiness" (IP) and equals (1 + 1/ICF) and the sixth index is called "Morisita's index" (MI) and equals [mean × n × IP / (n × mean – 1)]. The IP

and MI, measure the variability in intensity between the patches. Both will be approximately equal in value when n is high.

The purpose of this research is to identify the appropriate accident reduction plan so that the first preference will be given to black spots, when the location has a high degree of clustering. Suppose a location has 9 accidents then the average accident reduction (30%, as discussed in Chapter 1) is around 3, but if the cluster size is 4 the average accident reduction (30%) is one, which is not of much benefit. Therefore the index of cluster size (ICS) and the index of cluster frequency (ICF) may be useful. The index I and ICS are not much different (ICS = I-1), and hence only the five indices ICS, ICF, ICR, MI and IP were tested with the hypothetical distributions, with the results being discussed in Chapter 7. These five indices are investigated with various sizes of quadrats.

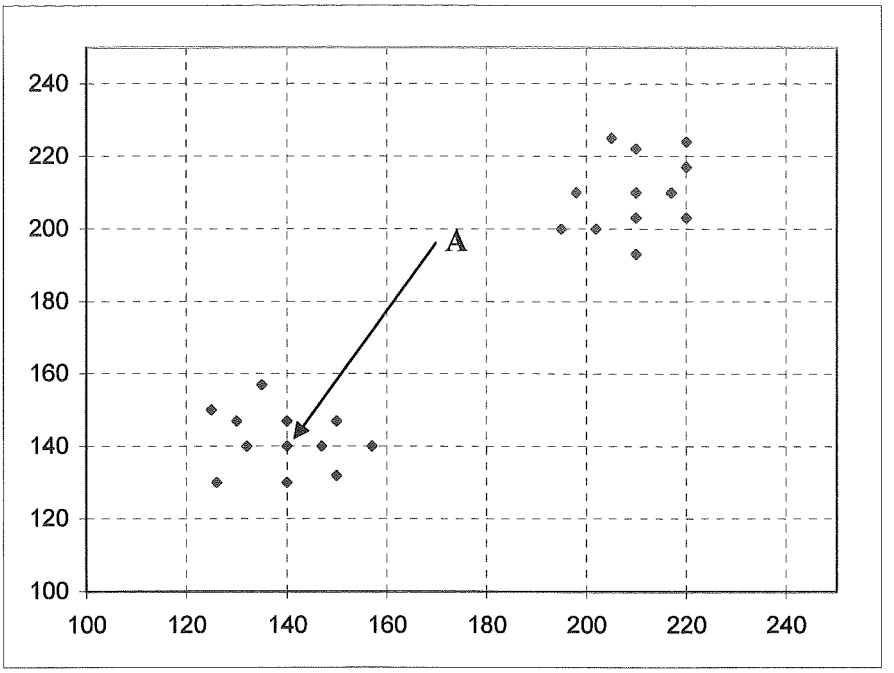


Figure 6.01a: Two point clusters (each 12 accidents)

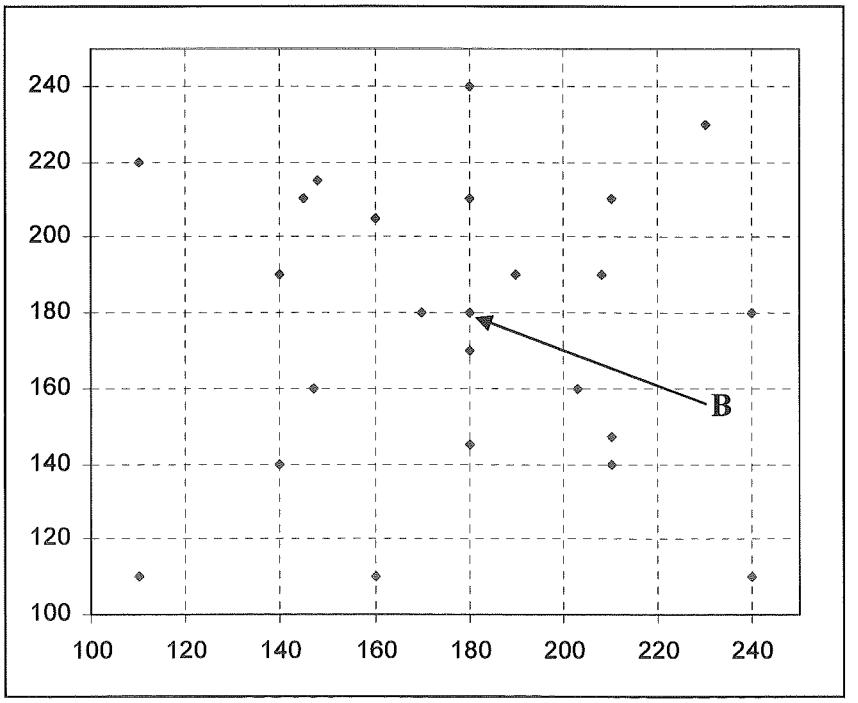


Figure 6.02a: Random locations (24 accidents)

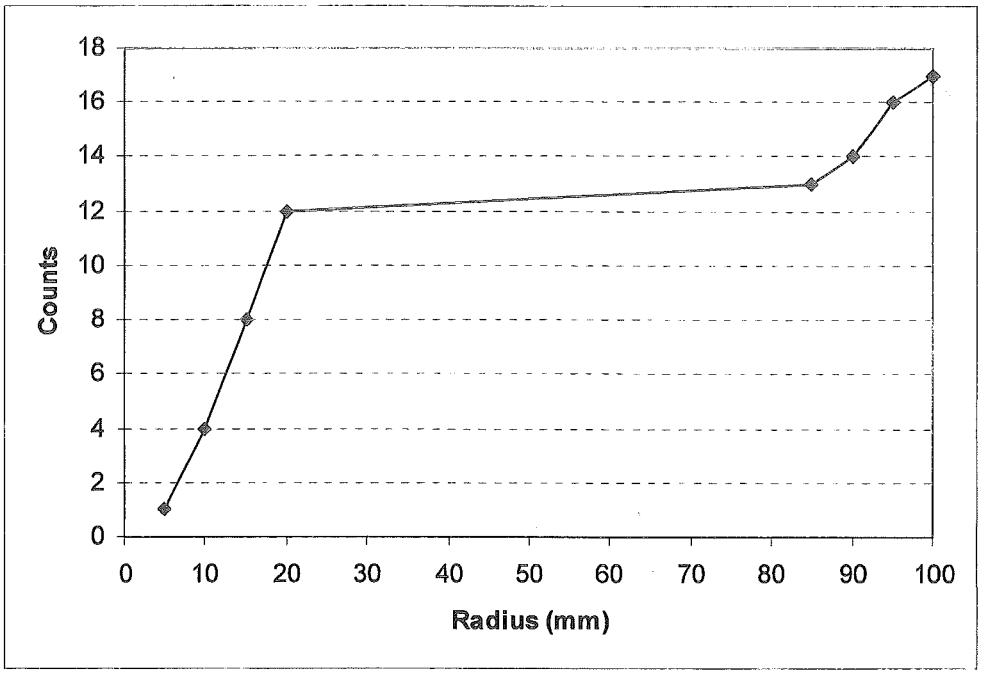


Figure 6.01b: Accident counts against radius of quadrats for point clusters (Figure 6.01a).

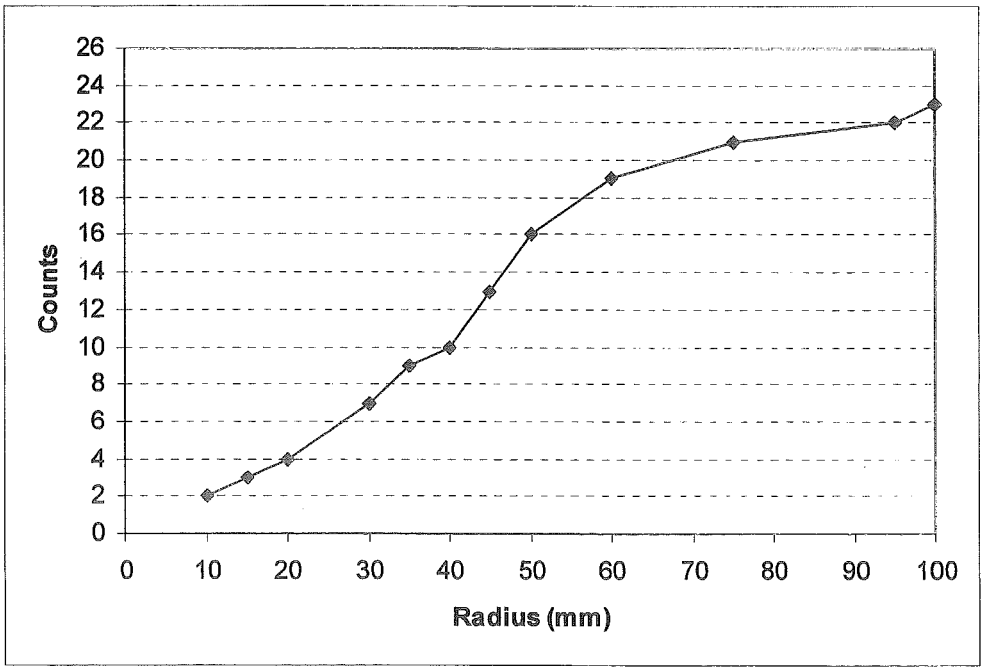


Figure 6.02b: Accident counts against radius of quadrats for random accidents (Figure 6.02a).

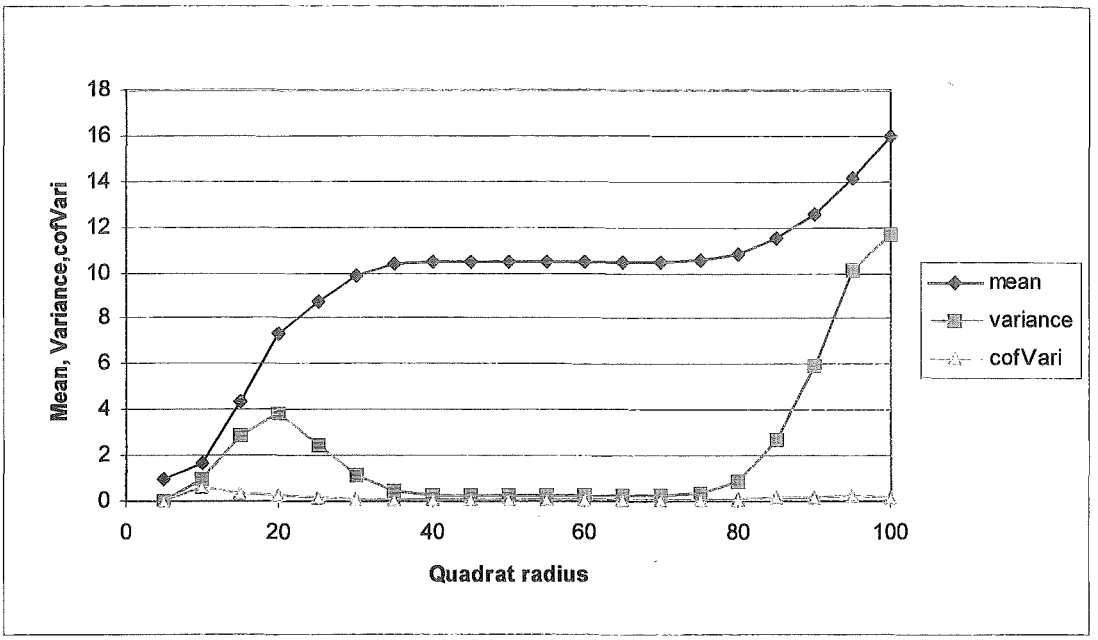


Figure 6.01c Variation of mean, variance and coefficient of variance with quadrat radius for a point cluster distribution (Figure 6.01a)

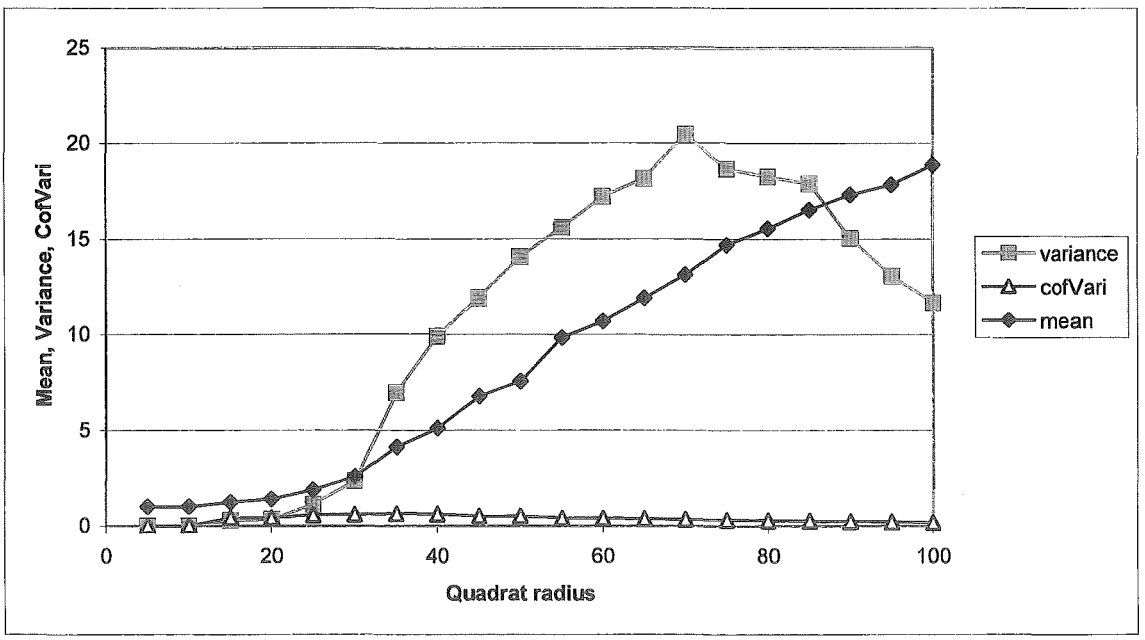


Figure 6.02c Variation of mean, variance and coefficient of variance with quadrat radius for a random distribution (Figure 6.02a)

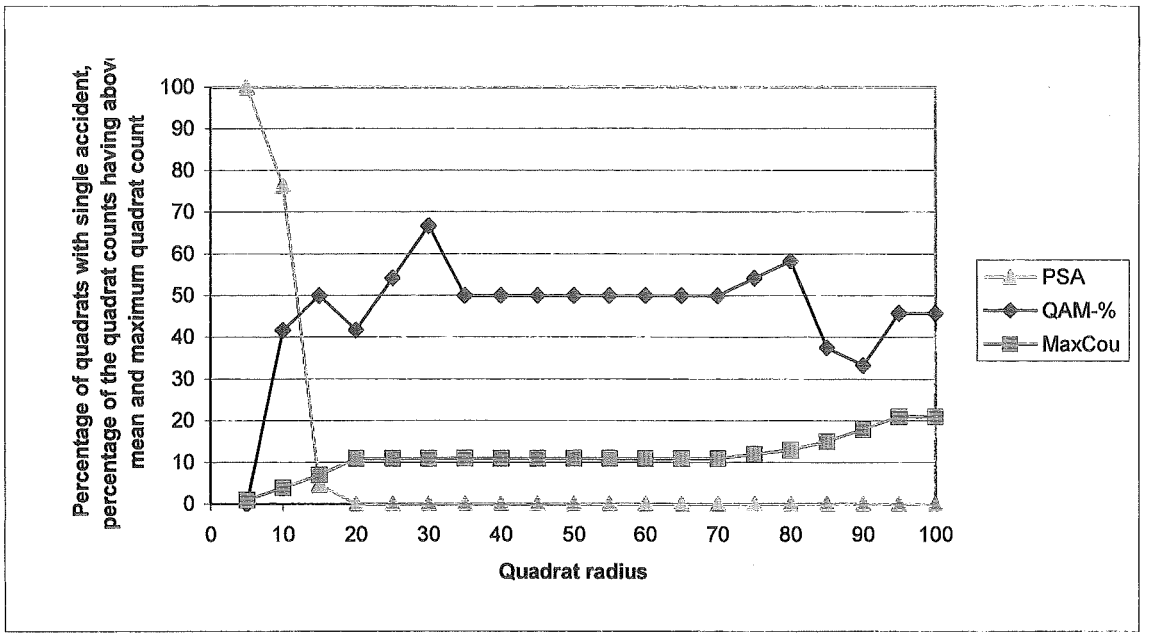


Figure 6.01d Variation of PSA , QAM% and MaxCou with quadrat radius for a point cluster distribution (Figure 6.01a)

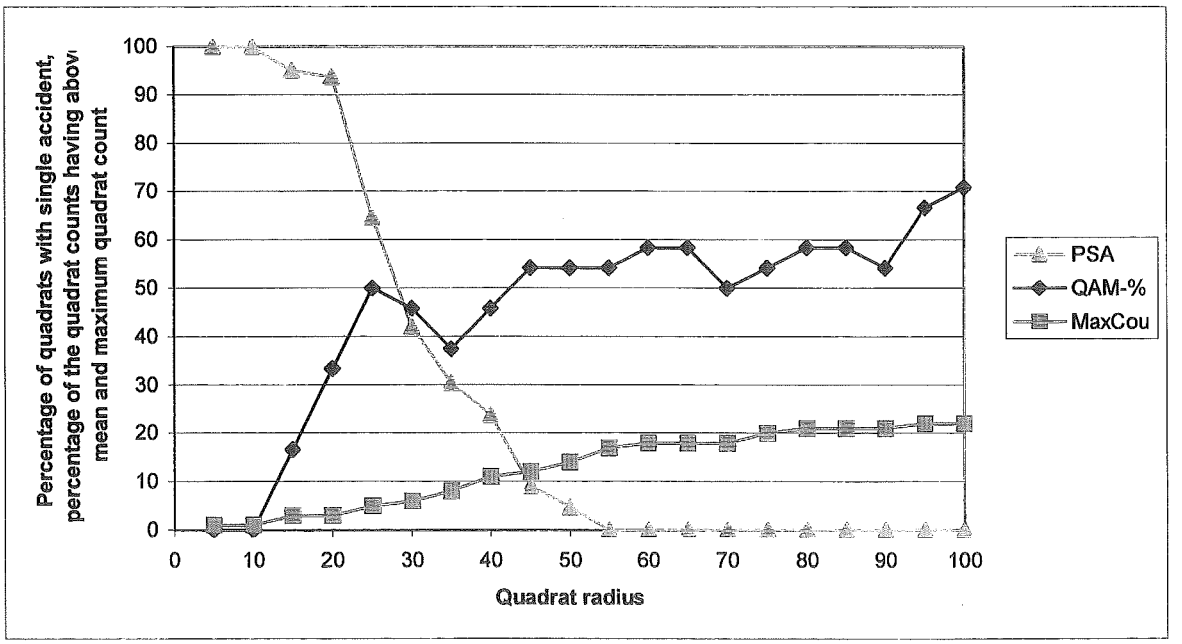
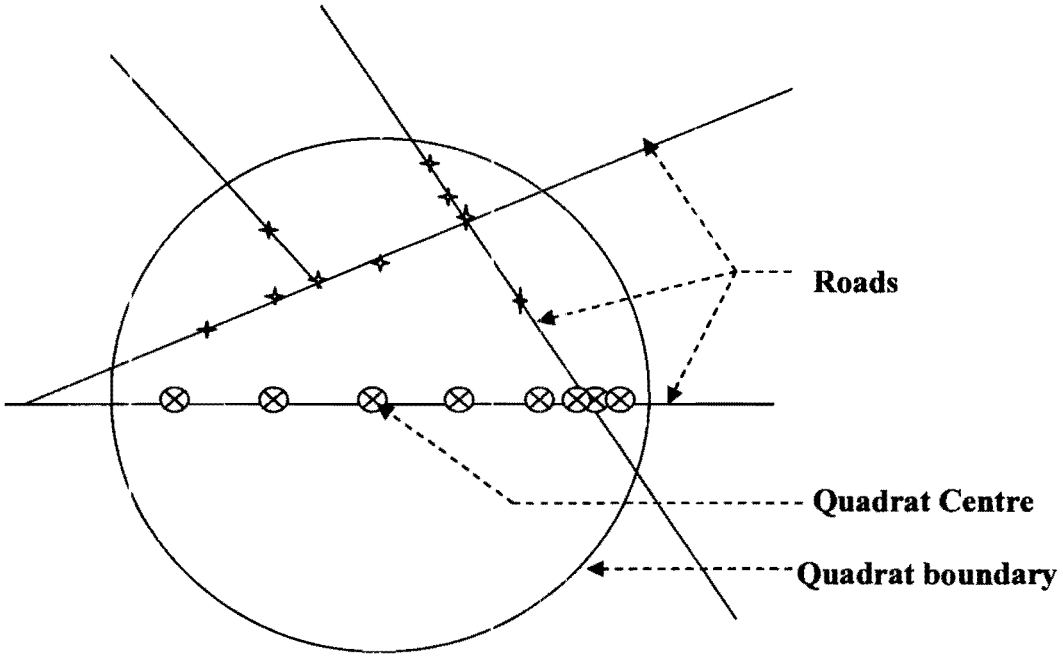


Figure 6.02d Variation of PSA , QAM% and MaxCou with quadrat radius for a random distribution (Figure 6.02a)

PSA-% - percentage of quadrats with single accident %QAM - percentage of the quadrats having above mean MaxCou - maximum quadrat count



⊗-----Counted accidents in the same road where quadrat centre located

+-------Uncounted accidents in other roads

Figure 6.03: Example for finding accident count from a single road

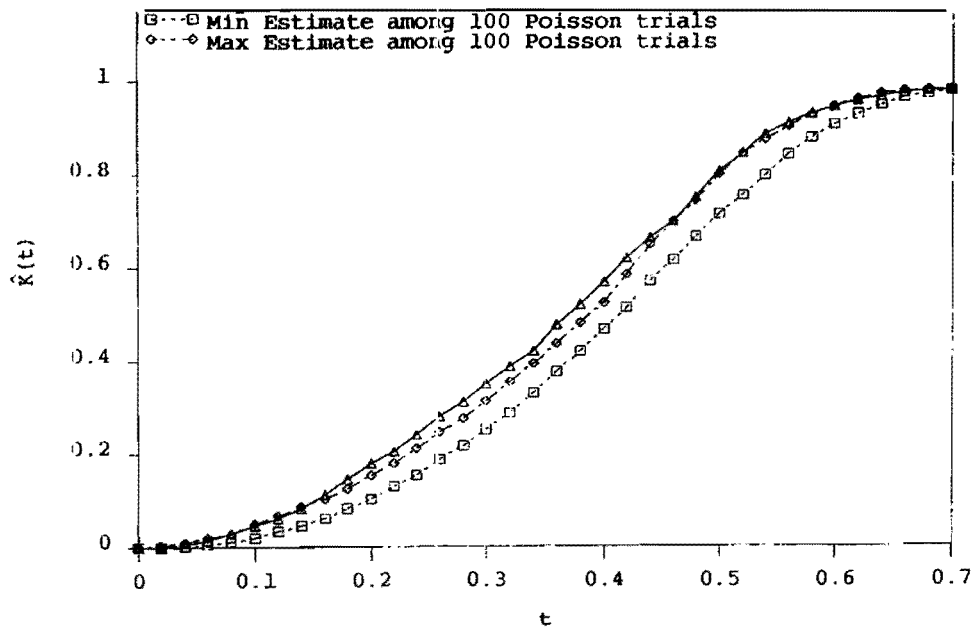


Figure 6.04: Example of K(t) function for Poisson trials.

(Figure 6.04 extracted from Jain and Dubes [1988])

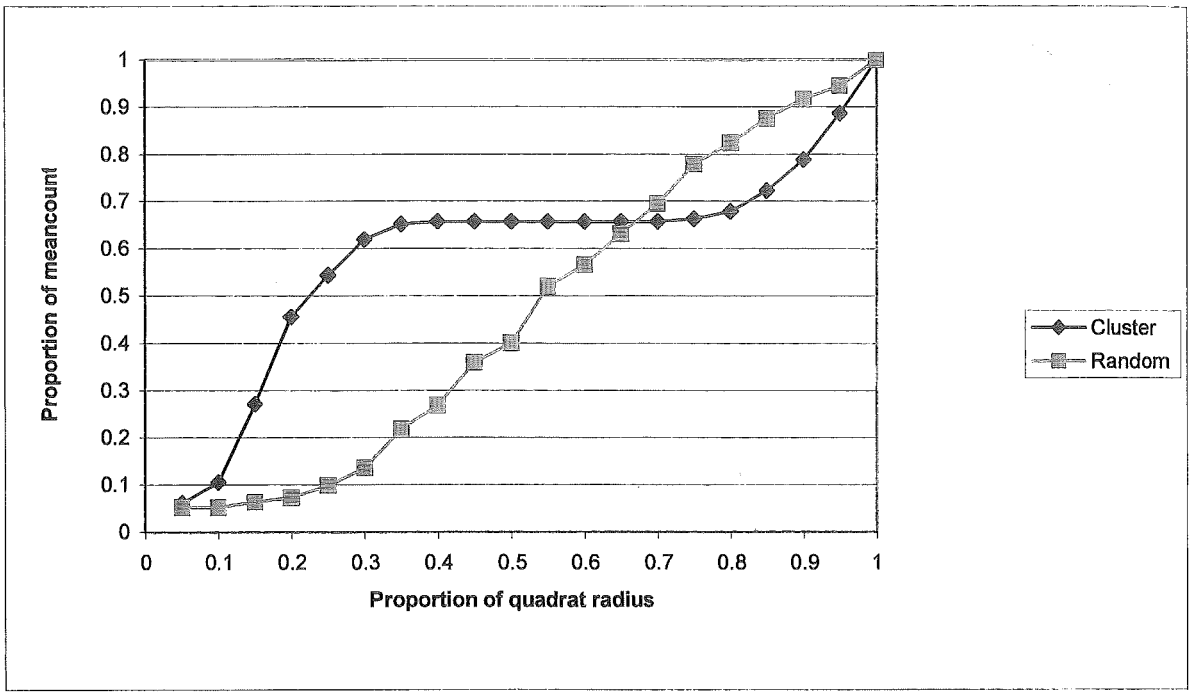


Figure 6.05: Variation of the proportion of mean versus the proportion of quadrat radius for the cluster distribution and random distribution.

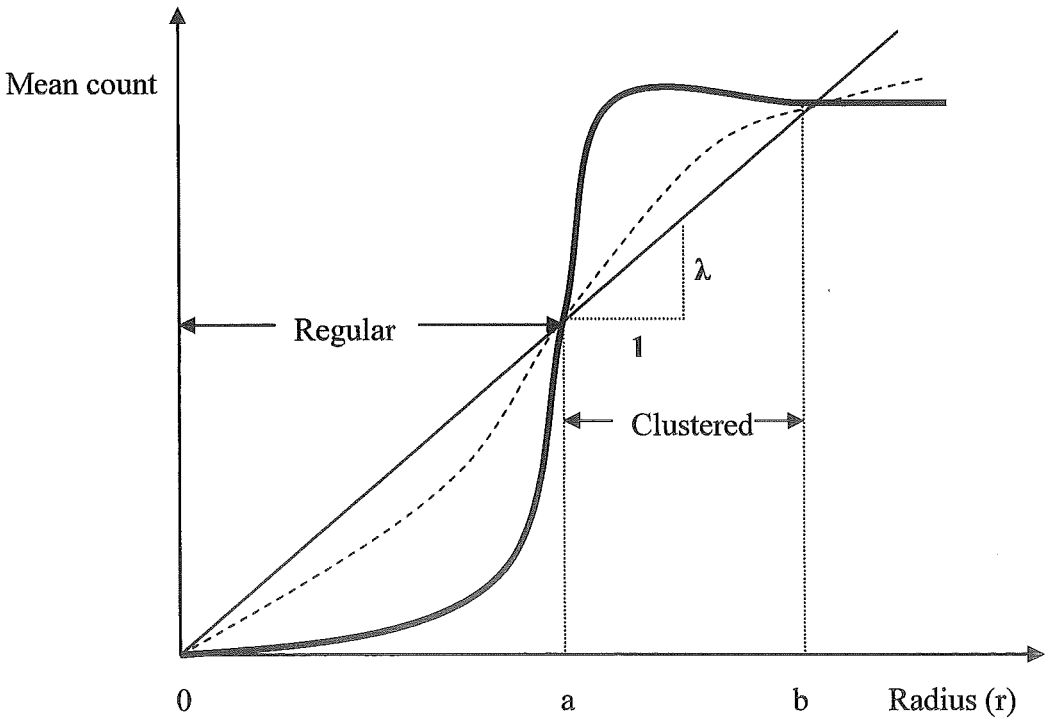


Figure 6.06: Mean quadrat count against quadrat radius

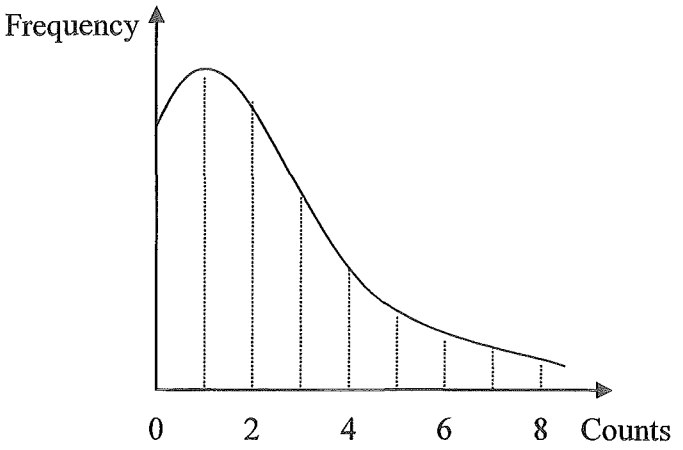


Figure 6.07a: Frequency against quadrat counts (positively skewed)

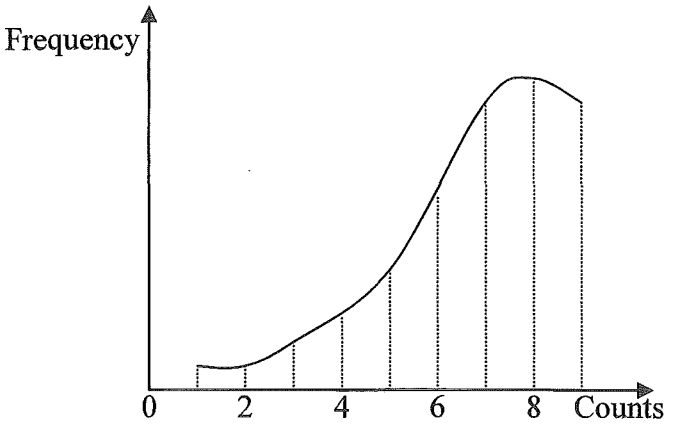


Figure 6.07b: Frequency against quadrat counts (negatively skewed)

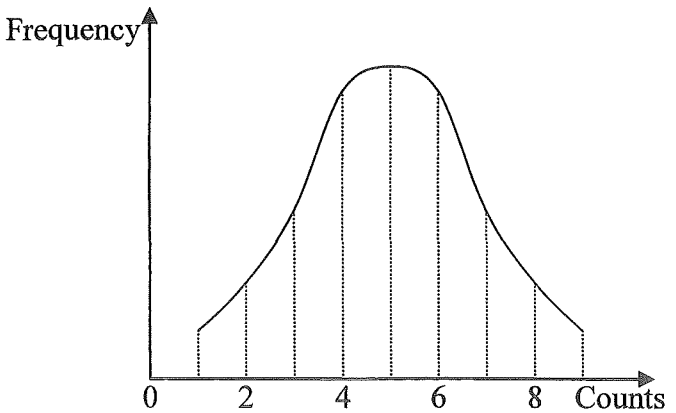
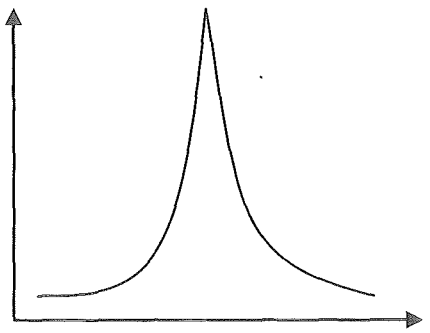
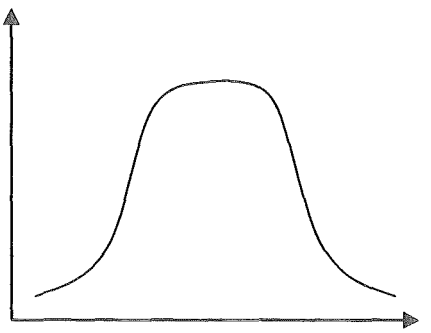


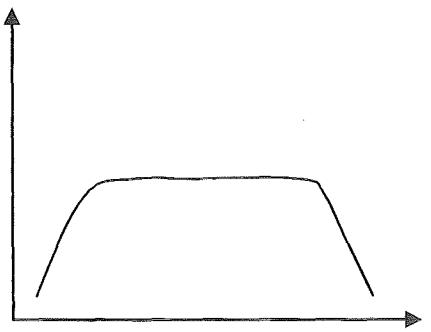
Figure 6.07c: Frequency against quadrat counts (zero skewed)



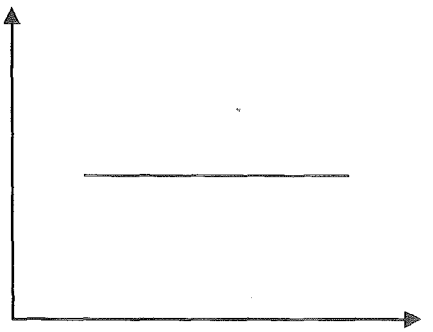
a. Leptokuritic distribution
(KURTOSIS > 3)



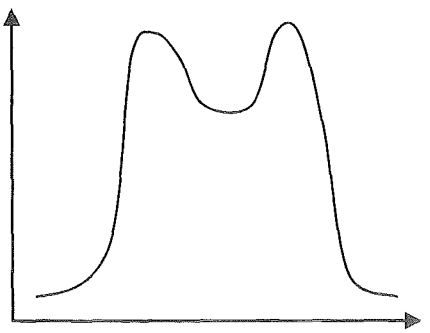
b. Mesokurtic distribution
(KURTOSIS ≈ 3)



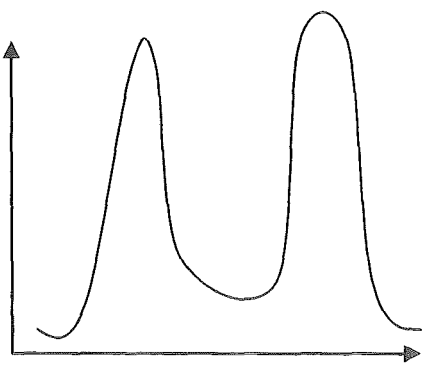
c. Platykurtic distribution
(KURTOSIS < 3)



d. Rectangular distribution
(KURTOSIS < 3)

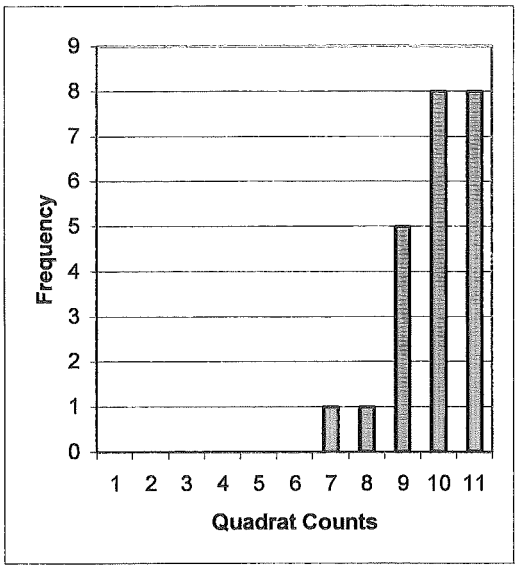


e. Slightly bimodal distribution
(KURTOSIS ≈ 1)



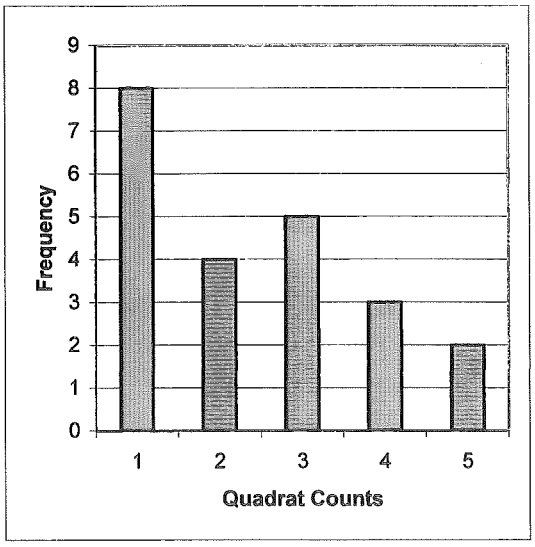
f. Highly bimodal distribution
(KURTOSIS ≈ 1)

Figure 6.08: Distribution exhibiting various values of kurtosis.



mean	9.9
variance	1.1
std.deviation	1.1
skewness	-0.9
kurtosis	3.5

Figure 6.09a: Frequency polygon of quadrat counts for a cluster pattern shown in Figure 6.01a



mean	2.7
variance	2.4
std.deviation	1.5
skewness	0.4
kurtosis	2.1

Figure 6.09b: Frequency polygon of quadrat counts for a random pattern shown in Figure 6.02a

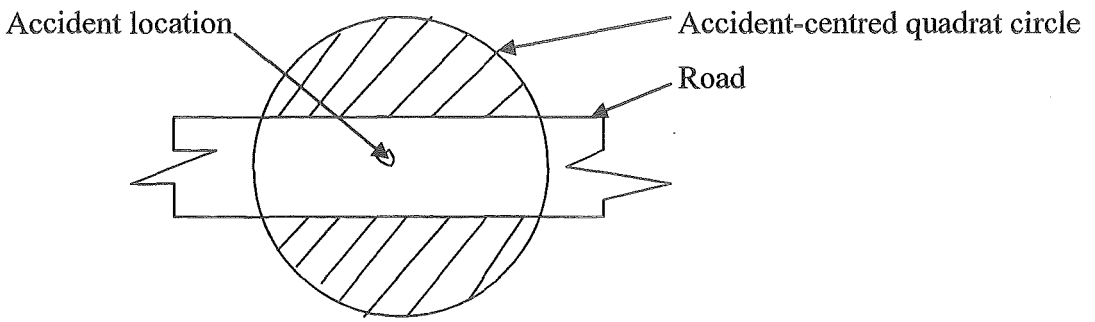


Figure 6.10: A quadrat circle on a road

Chapter 7

ANALYSIS RESULTS FOR HYPOTHETICAL DISRIBUTIONS

7.1 Description of data

The cluster analysis, nearest-neighbour analysis and quadrat analysis techniques were tested using four basic hypothetical distributions, as follows:

1. completely spatially random (CSR);
2. line cluster;
3. point cluster;
4. regular.

The distributions are shown in Figures 7.01, 7.02, 7.03 and 7.04, respectively. Each of these figures shows 100 event locations within an area of 300×300 sq. units. Another four basic hypothetical distributions were generated for testing the three analysis techniques with a higher number of events but lower density (Figures 7.05, 7.06, 7.07 and 7.08). Each of these figures shows 400 events locations within an area of 1000×1000 sq. units. These two sets of distributions were used to assess the usefulness of the three analysis techniques for identifying whether a spatial pattern (if any) exists and the nature of that pattern. For those analysis techniques which were identified as useful, some additional hypothetical distributions and mixtures of the basic distributions were used for further investigations.

In the following sections of this chapter, the cluster analysis results, nearest-neighbour analysis results and quadrat analysis results are discussed respectively.

7.2 Cluster analysis results

Four cluster analysis techniques (i.e. single linkage, complete-linkage, group average and Ward's method) were discussed in Chapter 4. Initially each of the four techniques was tested using the two sets of four basic hypothetical distributions described above. A method was identified for interpreting the analysis results and distinguishing between the distributions.

Then the techniques were tested using the mixed distributions, in which the proportion of events of each basic distribution was known.

7.2.1 *Basic distributions*

The single linkage, complete-linkage, group average and Ward's techniques were used to analyse the four basic distributions shown in Figures 7.01, 7.02, 7.03 and 7.04. In the point cluster distribution, each cluster has four events. For the completely spatially random distribution shown in Figure 7.01, the profiles of dissimilarity coefficient versus the number of clusters, computed using the four techniques, are shown in Figures 7.09a, 7.09b, 7.09c and 7.09d respectively. Figure 7.10 a, b, c and d are the corresponding plots for the line cluster distribution shown in Figure 7.02 (line cluster distribution). Similarly Figures 7.11a, b, c, and d are for Figure 7.03 (point cluster distribution) and Figures 7.12a, b, c, and d are for Figure 7.04 (regular distribution).

The four plotted profiles obtained using the single-linkage method for each of the four basic distributions have different shapes. The average distance between the plotted profile and the x-axis is decreasing from CSR to line cluster and then to point cluster distribution. This is very similar to the results of Anujah [1997] and Nicholson [1998], except for the regular distribution. The plotted profile obtained for the regular distribution using the single linkage technique shown in Anuja [1997] indicates the nearest-neighbour distance from each event is not constant, but the location plot indicates the nearest-neighbour distances from the events are constant. If this distance is constant then the dissimilarity coefficient should not vary and it should be 100. The location plot shown in Figure 7.04 indicates a constant nearest-neighbour distance, consistent with Figure 7.12a, which indicates a constant dissimilarity coefficient.

The profile in Figure 7.09a is similar to an inverted S-curve, which is for CSR distribution but the Figures 7.10a and 7.11a are very similar to a hyperbola (i.e. $xy = a^2/2$ where x and y are greater than one and a is greater than zero) in the first quadrat. The distance a is measured from the origin to the vertex of the curve. The value of *"a"* decreases when the pattern changes from a line cluster to a point cluster. Identifying the vertex is a difficult task because the dissimilarity coefficient profile is not a true hyperbola. Therefore a different

method was used to identify these patterns. The area (A) between the dissimilarity coefficient profile and the x-axis, is used to indicate the difference in the four figures (Figures 7.09a, 7.10a, 7.11a and 7.12a). The areas under the dissimilarity coefficient profiles decrease from CSR to line cluster to point cluster to regular distribution, respectively (i.e. $A_{csr} > A_{line} > A_{point} > A_{reg}$).

Figures 7.09b, 7.10b, 7.11b and 7.12b, computed using the complete-linkage method, also indicate different profiles, but they are not as different as those for the single-linkage method. The plots obtained using the group average and the Ward's methods are quite similar to the plots obtained using the complete linkage method. The areas computed using each method for each of the distributions are noted in Table 7.01.

Table 7.01: Areas below dissimilarity coefficient profiles

Methods	Area (A) for each distribution			
	Regular	CSR	Line clusters	Point clusters
Single-linkage	0	4036.68	2068.57	1120.71
Complete-linkage	1018.59	1290.21	1023.28	1013.23
Group average	1396.09	1911.76	1160.6	891.03
Ward's	1640.78	1242.75	741.6	632.59

The areas computed using the single-linkage method differ substantially for the different distributions but the areas computed using the other three methods do not show such pronounced differences. The trend for A to reduce from CSR to line cluster to point cluster distribution is evident, but the value of A calculated by other methods except single-linkage method, increases from point cluster to regular distribution.

Another example of the four basic distributions (i.e. CSR, line cluster, point cluster and regular distributions) is shown in Figures 7.05, 7.06, 7.07 and 7.08. In the point cluster distribution, each of the clusters again has four events. The number of events per cluster is chosen as a constant because previous studies have mentioned that Ward's method works well with equal sized clusters (as discussed in Chapter 4). In this example the location of 400 events are plotted within 1000×1000 sq. unit. In this example the number of events are increased but the average intensity is reduced.

The four distributions shown in the Figures 7.05, 7.06, 7.07 and 7.08, are analysed using the four cluster techniques and the profiles of dissimilarity coefficient versus the proportion of events not in clusters are shown in Figures 7.13, 7.14, 7.15 and 7.16. The plotted profile computed using the single-linkage method, shown in Figure 7.13a, clearly indicates similarity to an S-curve. The other three profiles shown in Figure 7.13 are very similar to each other. The profiles shown in Figures 7.13a, 7.14a and 7.15a are plotted in Figure 7.17a and these three profiles exhibit fairly different shapes. The profiles shown in Figures 7.13b, 7.14b and 7.15b are plotted in Figure 7.17b and these three profiles exhibit quite similar shapes. The profiles computed using the complete-linkage, the group-average and the Ward's method are very similar to each other, as shown in Figures 7.13, 7.14, 7.15 and 7.16.

The plot shown in Figure 7.16a is not included in Figure 7.17a. The plotted profile using the single-linkage method for regular distribution shown in Figure 7.16a is very similar to the plot shown in Figure 7.12a. These two figures indicate that each of the regular distributions (shown in Figures 7.04 and 7.08) has a constant nearest-neighbour distance. The results shown in Figures 7.12a and Figure 7.16a for the regular distributions are easily predictable because the events are distributed at a constant distance, therefore the focus at the moment is upon the other three distributions (i.e. the CSR, line cluster and point cluster distributions).

Figure 7.17 shows the extent to which the plotted profiles are different in shape for each of the three type distributions (i.e. CSR, line cluster and point cluster distributions). The plots computed using the single-linkage method indicate a different shape, but the shapes of the plots computed using complete-linkage methods are not much different. This figure indicates that the single linkage method is helpful for identifying the type of pattern from the unknown distribution.

The areas (A) under the dissimilarity coefficient profiles computed from Figures 7.13, 7.14 and 7.15 are tabulated in Table 7.02. Although the total number of events has increased in this example, the area computed for each of the four techniques for the CSR distribution is higher than for the line cluster and point cluster distributions. The area computed using Ward's method for the regular distribution is higher than for the other three distributions. The areas calculated using single-linkage techniques show a clear indication of a decreasing trend from CSR to line cluster to point cluster to regular distributions. The area computed

using the complete-linkage or group-average or Ward's methods do not indicate a similar trend.

Table 7.02: Areas below dissimilarity coefficient profiles

Methods	Area (A) for each distribution			
	Regular	CSR	Line cluster	Point cluster
Single-linkage	0	3361.98	1711.34	1263.12
Complete-linkage	519.99	679.86	405.44	432.81
Group average	737.22	953.38	458.57	620.44
Ward's	1640.78	605.8	283.58	387.04

The areas computed using the single-linkage method shown in Tables 7.01 and 7.02 are zero for the regular distribution. This is an indication that the events are distributed at constant distance for regular distributions (Figures 7.04 and 7.08) and can be easily identified. The results plotted in Figure 7.12 and Figure 7.16 for the two regular distributions are easily predictable because in the regular distribution cluster level increases for a constant distance. Therefore for the time being we analyse the other three distributions (i.e. the CSR, point cluster and line cluster distributions).

Tables 7.01 and 7.02 indicate that the single-linkage cluster technique is helpful compared to the other cluster techniques. The group-average and Ward's methods in the first example indicates $A_{line} > A_{point}$, but the second example indicates $A_{line} < A_{point}$. Because of this unexpected result, these methods are not used for further investigations. For complete-linkage the percentage difference in the area (A) computed for the line and point cluster distributions is considerably less than the percentage difference in the area between CSR and other three (i.e. point cluster, line cluster regular distributions). Therefore, for further analysis the single-linkage and complete-linkage techniques are used.

7.2.1.1 *Analysing the confidence band*

The results obtained until now are for particular examples of each type of distribution and do not allow for variations between different examples of each type of distribution. Therefore twenty five examples of the three distribution types (i.e. CSR, line cluster, point cluster distributions) were randomly generated. Twenty-five plotted profiles were obtained for each of the three distribution types, using the single-linkage and complete-linkage methods. The profiles for the single-linkage method for the CSR, line cluster and point cluster distributions are shown in Figures 7.18a, 7.19a and 7.20a. The profiles for the complete-linkage method for the CSR, line cluster and point cluster distributions, are shown in Figures 7.18b, 7.19b and 7.20b. The intention was to see whether the envelope containing the profile for one type of the distribution, was distinctly different from the other two envelopes. If so, then it will be helpful for identifying the type of pattern in a distribution. The dissimilarity coefficient profiles in Figures 7.18a, 7.19a and 7.20a are plotted in Figure 7.21a to identify the extent of overlapping in the envelopes. The dissimilarity coefficient profiles in Figures 7.18b, 7.19b and 7.20b are also plotted in Figure 7.21b. The envelopes do not overlap in Figure 7.21a, to the extent that they do overlap in Figure 7.21b. Therefore the complete-linkage method is not investigated further, because the envelopes obtained from the complete-linkage method for the three distributions do not show substantial differences, but the envelopes obtained from the single-linkage method show substantial differences.

In Figure 7.21a, a large portion of the envelope obtained for CSR distribution using the single-linkage method is quite separate from the envelopes obtained for the point cluster and line cluster distributions but the envelopes obtained for the line cluster and point cluster distributions do overlap each other. It appears that this method is not helpful for identifying line cluster or point cluster distributions. Further analysis was carried out to verify this.

The variance and the mean of the areas under the 25 dissimilarity coefficient profiles for the single-linkage method, for each of the three distributions, are shown in Table 7.03. This table will be used in the next section.

Table 7.03: Mean and variance of the areas under dissimilarity coefficient profiles for single-linkage method

Distribution	Area under the dissimilarity coefficient profiles			
	Mean	Variance	Maximum	Minimum
CSR	M_{csr} = 4064.0	S_{csr}^2 = 377793.3	5314.0	2694.8
Line cluster	M_l = 2254.9	S_l^2 =186623.7	2983.7	1297.2
Point cluster	M_p = 2006.4	S_p^2 =91919.3	2681.4	1382.3

7.2.1.2 *Inference for distributions*

A conclusion can be inferred from the procedure for comparing the variance of two normal populations based on the F-statistics. If the two variances are significantly different then the two means can be tested using the t-test to see whether the means are significantly different. The Table 7.04 shows the F(calculated), F(critical), t(calculated) and t(critical) values. The F(calculated) and t(calculated) values are based on the data in Table 7.03.

Table 7.04: Calculated and critical values for F-test and t-test

Distributions	F(calculated)	F(24,24,0.05)	t(calculated)	t(48,0.05)
CSR/Line	2.02	1.98	12.04	1.68
CSR/Point	4.11	1.98	13.69	1.68
Line/Point	2.03	1.98	2.35	1.68

The calculated F and t statistics (Table 7.04) exceed the critical values, indicating that each of the 25 sets of samples are from three different distributions. Hence, three different confidence limits can be identified for the distributions (CSR, line cluster, point cluster).

7.2.1.3 *Confidence interval estimation*

The confidence interval estimation method for a normally distributed population with an unknown variance was used to calculate the confidence limits for the CSR distribution. The

mean and variance of the area under the computed dissimilarity coefficient profile are given in Table 7.03.

The 95% confidence interval for the area under the dissimilarity coefficient profile computed using the single-linkage method for the CSR distributions is in the range $3810 \leq A_{csr} \leq 4318$. Similarly, for the line cluster distributions the confidence interval is in the range $2077 \leq A_{line} \leq 2433$ and for the point cluster distributions the confidence interval is in the range $1881 \leq A_{point} \leq 2132$. The 95%, 90% and 85% confidence intervals, maximum and minimum and the mean areas are calculated for the 25 examples of the three distributions used in Section 7.2.1. The calculated values are plotted in Figure 7.22, which gives a visual indication of the confidence band for each of the three distributions. The three confidence intervals for 95%, 90% and 85% are shown in Figure 7.22a, b and c respectively.

Figures 7.22a, b and c indicate that the confidence band calculated for the CSR distributions, when compared with the confidence bands calculated for the line cluster and point cluster distributions, is substantially different, but the confidence bands calculated for the point cluster and line cluster distributions are not substantially different. The confidence bands calculated from the area under the dissimilarity coefficient profile are very close to each other for the point and line cluster distributions. This can be verified from Figure 7.21a, in which the envelopes obtained for the line and point cluster distributions overlap.

The 95% and 90% confidence intervals (shown in Figures 7.22a and b) for the point cluster and line cluster distributions overlap ($2132 > 2077$ and $2110 > 2107$), but the 85% confidence intervals (shown in Figure 7.22c) do not overlap ($2097 < 2126$). The lower the confidence level, the more narrow the confidence intervals, and the less likely they are to overlap. For the 85% confidence level, the confidence bands of the three basic distributions do not overlap. At this confidence level, the technique might help to identify the three basic patterns.

7.2.2 *Mixed distributions*

Accident data may be a mixture of CSR and the other distributions (regular, point or line cluster). Therefore it is necessary to test the methods with mixed distributions. As

previously mentioned in Section 7.2, for the time being the regular distribution is not considered. Several combinations of the three distributions (CSR, point cluster and line cluster distribution) were tested, the mixing proportions are tabulated in Table 7.05.

Table 7.05: Proportions of each distribution for cluster analysis

Case	Mixture Name	Proportion of distributions		
		CSR	Point cluster	Line cluster
I	60R-40P	60	40	0
	50R-50P	50	50	0
	40R-60P	40	60	0
II	60R-40L	60	0	40
	50R-50L	50	0	50
	40R-60L	40	0	60
III	40P-60L	0	40	60
	50P-50L	0	50	50
	60P-40L	0	60	40
IV	50R-20P-30L	50	20	30
	20R-30P-50L	20	30	50
	30R-20P-50L	30	20	50

The mixtures noted in the above table were analysed using the single-linkage method and the areas under the dissimilarity profiles were computed. The results (i.e. the maximum, minimum, 85% confidence level and mean area) are plotted in Figure 7.23. In this figure, R, P and L indicate CSR, point cluster and line cluster, respectively.

Case I

In this mixture CSR and point cluster distributions are present. The area computed for 60R-40P mixture is 2416 which is just outside to the 85% confidence limit for the line cluster (i.e. 2383 < 2416 > 2126) although there is no line cluster in that mixture. The result for the mixed distribution (50R-50P) is within the confidence limits of line cluster distributions but the line cluster is not present in the mixed distributions. The area calculated for the mixture 40R-60P is 1960 which is within the 85% confidence limits for point clusters (i.e. 2097 > 1960 > 1916). This indicates that the mixture 40R-60P is a point cluster distribution. These results do not correctly indicate the presence of the CSR distribution in the mixed

distribution. In this case, it is difficult to conclude whether the higher proportion in the mixture is point cluster or line cluster or CSR distribution, but the results indicate that the distribution is not a purely random distribution.

Case II

In this mixture CSR and line cluster distributions are present. The results from the mixture of the line and CSR distributions (60R-40L and 50R-50L) are also within the 85% confidence limit for point or line cluster distributions but the point cluster is not present in the mixtures. If we consider the two mixtures (60R-40L and 50R-50L) the computed areas are within the 85% confidence limits for the line cluster distributions but the results do not correctly indicate the proportion of events of the CSR distribution in the mixture. The area computed for the mixture 40R-60L is 1776 which is below the 85% confidence limit of point cluster, although there is no point cluster in that mixture. In this case, it is difficult to conclude whether the higher proportion in the mixture is point cluster or line cluster or CSR distribution, but the results indicate that the distribution is not a purely random distribution.

Case III

In this mixture point and line cluster distributions are present. The areas calculated for the mixtures (60L-40P, 50L-50P and 40L-60P) are less than the 85% lower limit for line or point cluster distributions. The three different proportions of the mixtures indicate that this method does not help to identify the higher proportion present in the mixed distribution. It can be concluded that the distributions are not random, but one cannot determine whether it is a mixture with point clusters or line clusters or both.

Case IV

In this mixture all three distributions are present in different proportions. The computed areas are higher than the computed area for Cases I, II and III. The area computed for 50R-20P-30L is 2847 which is lower than the area computed for 20R-30P-50L. The proportion of events from the CSR distribution is higher in mixture 50R-20P-30L than 20R-30P-50L but the computed area for 50R-20P-30L less than the area for 20R-30P-50L. The mixture (50R-20P-30L) gives an unexpected result. In Case I and II, the computed area for the mixture with the highest proportion of CSR distribution is higher than the computed area for the other two. In Case IV for 50R-20P-30L the opposite is true. The area computed for the mixture 30R-20P-50L is 3976 which is within the 85% confidence limit (i.e. $3881 < 3978 <$

4247), and fails to indicate the presence of a high proportion of the line cluster distribution or low proportion of the point cluster distribution, and indicates that the events are from a CSR distribution. A pure CSR distribution cannot be identified because the results indicate that the mixture (30R-20P-50L) and a pure CSR distribution can have equal area under the dissimilarity coefficient profile.

These are unexpected results, and do not help to identify the distribution with the higher proportion in the mixture. The analysis result is from one example of each mixture. If we analysed several examples of each mixture then the result may vary, but distinguishing point cluster or line cluster distributions is difficult because the confidence limit for the area under the dissimilarity coefficient profile for the distributions (point cluster or line cluster) are close to each other. The method developed so far does not distinguish between a pure CSR distribution and a distribution with a high proportion of the line cluster in the CSR, point and line cluster mixed distributions. The usefulness of this method to distinguish mixed distributions is not promising.

7.2.3 *Discussion of cluster analysis results*

Tables 7.01 and 7.02, and Figure 7.21a indicate that the area under the dissimilarity coefficient profiles obtained using the single-linkage method for the CSR distribution is distinguishable from the other three (i.e. line cluster, point cluster and regular distributions). The nearest-neighbour distance of each event is not a constant in the CSR distribution and hence events can be grouped according to the distance.

The overall results for this method are not helpful in identifying the basic distributions or mixed distributions. For example in Case IV, the pure CSR distribution cannot be distinguished from a mixed distribution. The 85% confidence limit for a CSR distribution is higher than the line cluster and point cluster distributions, as shown in Figure 7.23.

Unreliable results were obtained for Case I, Case II and Case III. For example, the area under the dissimilarity coefficient profiles obtained for a mixture of the CSR distribution and point cluster distribution (50R-50P) is within the confidence limit for a line cluster distribution, but the line cluster is not present in the mixture. The major component in the

mixture cannot be identified using the single-linkage cluster method. In Section 4.3.5 the sensitivity of single-linkage techniques when "noise" is present was discussed. This is probably a reason why the single linkage technique does not perform well when analysing mixed distributions.

The envelopes plotted in Figure 7.21a indicate differences for each distribution but there is no clear indication of separate envelopes for the different types of distribution. The upper and lower boundaries of the envelopes of point and line cluster overlap each other, as can be noted from Figures 7.21 a and b. A large portion of envelope obtained for CSR distribution is non-overlapping.

If we analyse a mixed distribution then the envelope identified is not helpful for identifying the major component in the mixture. For example the mixed distribution indicates unexpected results for the 30R-20P-50L, 60R-40P and 60R-40Lmixtures. To identify the accident reduction program it is necessary to identify whether the accident locations are CSR or a high proportion of the accidents is from a point cluster or line cluster distribution. Therefore, further investigation of the area under the dissimilarity coefficient profile is not very helpful for accident analysis.

Figure 7.21b indicates that the complete linkage method is even less helpful for distinguishing the three distributions (i.e. CSR, line cluster and point cluster) because these three distributions have very similar profiles. The other two techniques (the group average and Ward's methods) also have very similar dissimilarity coefficient profiles for the CSR, line cluster and point cluster distributions. Therefore all four cluster analysis techniques are not very helpful for analysing the accident data.

7.3 Nearest-neighbour analysis results

A crucial task is to decide the number of nearest-neighbours for analysis, as the results are sensitive to variations in the number of nearest-neighbours (as explained in Chapter 5). Nicholson [1995] noted the direction and distance tests are sensitive to the relative size of clusters and the number of nearest-neighbours. The number of nearest-neighbours must be greater than four for the Rayleigh and K-S tests, and greater than 19 for the Kuiper-Watson

test (as explained in Chapter 5). To test the reliability and sensitivity, the number of nearest-neighbours chosen for this investigation was varied from two to (N-1), where N is the total number of events in the study area.

The confidence level for the statistical tests needs to be decided. For example, for the CSR distance distribution we need to allow for some clustering and regularity. For the line cluster, we need to allow some deviation of nearest-neighbour directions, because the road centerline might have small deviations over a short distance. How far do we need to allow for such matters and what is the significance level? To help answer this question we need to analyse the results for hypothetical distribution for the four selected values of significance levels 0.005, 0.05, 0.1 and 0.15.

7.3.1 *Basic distributions*

The four basic distributions shown in Figures 7.01, 7.02, 7.03 and 7.04 were analysed using the distance and direction methods. In this analysis initially the significance level 0.005 is used, and one by one the number of nearest-neighbours was increased from two to (N-1) for each test location (see Figure 5.01). Each of the events was chosen as the test location. This procedure was repeated for significance levels of 0.05, 0.1 and 0.15. The proportion of events indicating regular, clustered and 'unusual' distance distributions were plotted against the number of nearest-neighbours, for each of the four basic distributions. The proportion of locations indicating unusual distances is the sum of the proportion of locations with regularly distributed accidents and the proportion of locations with clustered accidents. The proportion of events indicating directional uniformity of their nearest-neighbour directions, using the three statistical tests for directions (i.e. the Rayleigh, Kuiper and Watson tests), were plotted against the number of nearest neighbour. The plotted profile for the CSR, point cluster, line cluster and regular distributions are shown in Figures 7.24, 7.25, 7.26 and 7.27. In these figures the plots a, c, e and g show the distance analysis results and the plots b, d, f and h show the direction analysis results, for significance levels 0.005, 0.05, 0.1 or 0.15 respectively.

The following observations are made from Figures 7.24, 7.25, 7.26 and 7.27.

1. The plot 7.24 a, c, e and g indicate that the proportion of cluster locations increases rapidly when the number of nearest-neighbours selected for analysis increases from about 80 to 99. Hence the results are not reliable in this range. For the number of nearest neighbours between 60 and 80, in each of these plots the variation is small. The results for the number of nearest neighbours between 60 and 99 do not give any special information compared with the results for the number of nearest neighbours between 2 and 60. Therefore it is appropriate to analyse from 2 to 60 for the rest of this Section.

When considering the number of nearest neighbours (N) around 40, the nearest-neighbour distance distributions indicating unusual nearest-neighbour distances less than 15 is shown in plot e (i.e. for $\alpha = 0.1$) but the nearest-neighbour distance distribution indicating the unusual nearest-neighbour distances is greater than 20 for the CSR distribution analysed (see plot g for $\alpha = 0.15$). The proportion of events indicating non-uniform nearest-neighbour directions in plot h is mostly in the range 0 and 30 but in plot f varies between 0 and 30. These results indicate that the significance level of 0.1 will be appropriate because the plots are similar for $\alpha = 0.1$ and 0.15.

2. The plots a and c shown in Figure 7.25 are not sufficiently enough to identify point clusters, because the plots indicating the proportions of cluster locations are around 50% or below for a point cluster distribution analysed with the number of nearest-neighbour between two and 60. Plots e and g indicate a maximum of 70% for the number of nearest-neighbours of 4. The significance levels 0.1 and 0.15 indicate appropriate results (i.e. cluster locations > 50%) for distance analysis. There is not much difference between the plot e and g compared to the plots a and c. Most statisticians prefer to use the significance level of 0.05 and 0.1 rather than 0.15. The proportion of events indicating cluster locations increased from plot a, to plot c then plot e but there is no notable difference between plot e and g. Considering all these points, the significance level 0.1 is appropriate.

3. In Figure 7.25e the cluster and regular lines are closer to each other, but this does not mean that equal proportions of events are regularly spaced and clustered. The regularly spaced or clustered events can be noted when N is equal to five. The cluster line indicates 70% and the regular line indicates 0% when N is four. This indicates that a large proportion of locations have four events, and that four events at each locations are not regularly spaced.

4. In plot e of Figure 7.25, when the number of nearest neighbour is four, 70% of events are cluster locations, 0% regular location and 30% CSR locations. The distribution is point cluster, even though the result indicates that around 30% of events are from a CSR distribution. The reason for this unexpected result is overlapping clusters (see the location plot Figure 7.03). There are seven clusters (each comprising four events) near the location (648040, 266610), which might appear to be an overlapping cluster with 28 events, indicating around 28 % of CSR locations. If clusters overlap then the number of events per overlapped cluster will be increased. If the number of nearest-neighbours analysed is less than the number of events per overlapped cluster, then the result could indicate that the test locations are random, even when the test location is clearly within a cluster. This is explained in Section 5.2.2, and is the reason that the test result for point cluster distribution indicates 28% CSR locations.

5. If we analyse say 20 nearest-neighbours without considering a range from 2 to 99 then the result is around 20% cluster location, 20% regular location and 60% CSR location for a significance level of 0.1 (see Figure 7.25e). In this case we miss the correct result for four nearest-neighbours and if the number of nearest-neighbours considered is 20 then we may misinterpret the information that there is a higher proportion (60%) of events from CSR compared to the proportion of cluster or regular events in this distribution. The information noted in point three and four clearly indicates why it is necessary to analyse using a range of values for the number of nearest neighbours.

6. The plotted lines for cluster or regular are very similar in the Figures 7.26 c, e and g, but in plot a, the cluster line is almost entirely below 50%. For consistency with CSR and point cluster distributions, the 0.1 significance level can be used for analysing line cluster distributions too. In Figure 7.26 e the regular line is not near the cluster line, but in Figure 7.25e the cluster and regular lines are close to each other for the number of nearest-neighbours between 10 and 40. This could help to distinguish between point cluster and line cluster distributions, but it needs further analysis because this indication might change when analysing a distribution which has a range of point cluster sizes (i.e. the area or number of events in each cluster).

7. Figures 7.27a, c, e and g indicate either 100% regular and 0% cluster, or 0% regular and 100% cluster for N from 2 to 99. These plots mainly indicate that the analysis results are

highly sensitive to N. Figures 7.27b, d, f and h indicate the proportion of events having uniform nearest-neighbour direction. The plots shown in Figure 7.27 for a regular distribution indicate any of the four significance levels can be used because there is no substantial difference between the plots using different significance levels. For convenience, the significance level 0.1 can be selected for further analysis.

8. In Figure 7.25 or 7.26, the proportion of events indicating non-uniform nearest-neighbour directions increased (big change) from plot b to plot d but the difference between plots d, f and h is not large. A big difference noted between Rayleigh and other test in Figures 7.25b and 7.26b.

9. Selecting a constant significance level will help to identify the distribution from the unknown distribution. When all the above reasons are considered, the significance level 0.1 seems most appropriate. From now on a significance level of 0.1 is used for further analysis and discussion.

10. The test results for a CSR distribution (see Figure 7.24, plot f) indicate a very low proportion of events with non-uniform directions to neighbours (i.e. the Rayleigh test indicates less than 20%, while the Kuiper and Watson tests indicate less than 10% on average). This clearly indicates that the distribution is anisotropic.

11. Figure 7.25 f shows an unexpectedly high proportion of direction test results where the proportion of events indicating non-uniform directions is around 80% for a point cluster distribution. The proportion of events indicating non-uniform directions of about 95% is noted from the plot obtained for line cluster distribution (see Figure 7.26 plots f). These results indicate that the proportion of events indicating non-uniform directions is higher for a line cluster distribution than for a point cluster distribution.

The plotted test results for the point cluster distribution (see Figure 7.25 f) indicate a high proportion of events (about 80%) with neighbours which are distributed in non-uniform directions. This unexpected direction test result for the point cluster distribution is because the direction test cannot distinguish between the two patterns shown in Figures 7.28a and 7.28b. The distance and direction distributions for the point cluster distribution (Figure 7.28a) are shown in Figures 7.28c and d, and for the line cluster distribution (Figure 7.28b)

are shown in Figures 7.28 e and f. The distance distributions for the two distributions shown in Figures 7.28a and 7.28b are different, but the direction distributions are similar (i.e. both direction profiles approaches 100%). Therefore the sensitivity of the direction test is not enough to distinguish between the point and line cluster distributions, but the distance test result seems better able to distinguish the two distributions than the direction test.

An important point should be noted from the test results from the distance analysis result shown in Figures 7.25 and 7.26, and that is that the distance analysis provides the capability to distinguish between point cluster and line cluster distributions. In Figure 7.25e the cluster profile exhibits a sudden increase for four to five number of nearest-neighbours and then a sudden drop, because the distance test is sensitive to the number of nearest-neighbours, which depends on number of events per cluster (as explained in Section 5.2.2), and the significance level. This type of sudden increase and drop is not noted for the distance analysis results for the line cluster distribution (Figure 7.26e), because the number of events per cluster is very large for line clusters compared to point clusters. The point cluster distribution used for this analysis has four events per cluster, and this is the reason for the sudden increase when N is four. In the line cluster distribution the number of events per line cluster is twenty-five and this is the reason the cluster profile (see Figure 7.26e) increased gradually from N two to thirty without any large sudden increase and/or decrease. This suggests that the nearest-neighbour method can distinguish between the point cluster, line cluster and the CSR distributions.

In this example, the events per cluster are four (i.e. a constant) for the point cluster distribution. This might be a reason why the line cluster and point cluster can be easily identified. In practice, the number of events per point cluster can vary in real accident data. Therefore, it is necessary to analyse a hypothetical distribution with various numbers of events per cluster and then compare the results with the results for the line cluster distribution.

Examples of eight hypothetical distributions, in which the events per cluster vary from 2 to 20, were analysed. Four of the examples of location-plots for dense point clusters (i.e. events are close to each other within each cluster) are shown in Figures 7.29 a, b, c and d and the other four examples of location-plots shown in Figures 7.30 a, b, c and d are sparse point clusters (i.e. events are far from each other within each cluster). The difference between the

dense point cluster and sparse point cluster distributions is that the area of a dense point cluster is less than the area of a sparse point cluster, but the number of events per cluster is equal. The four dense point cluster distributions (Figure 7.29) were analysed using the distance test and direction test with a significance level of 0.1. The results are shown in Figure 7.31. The four sparse point cluster distributions (Figure 7.30) were also analysed using distance and direction tests, and the results are plotted in Figure 7.32.

The plots are for the number of nearest-neighbours varying from four to 60. The following observations can be noted from the results shown in Figures 7.31 and 7.32.

1. For the number of nearest neighbours up to 20, whenever the cluster line indicated a peak the regular line indicated a trough. The peaks shown in Figures 7.31a, c, e and g, and Figures 7.32a, c, e and g, are not prominent compared to the peak in Figure 7.25e.
2. The distance test results help to distinguish point cluster and line cluster distribution when considering equal numbers of events in the point clusters, but the test is less helpful when considering non-equal numbers of events in the point clusters.
3. Overall the usefulness of the direction or distance test to distinguish between point cluster and line cluster distributions is minimal compared to cluster analysis.

The plot for a point cluster distribution with constant events per cluster indicates a high proportion of events with non-uniform neighbour directionality (e.g. Figure 7.25f). To verify the conclusions made from the direction test results and distance test results for the distributions shown in Figures 7.01 to 7.03, another set of less dense distributions (shown in Figures 7.05 to 7.07) were investigated. The total number of events shown in each of the Figures 7.05, 7.06 and 7.07 is 400, and the number of nearest neighbours used for analysis ranged from 2 to 240. The analysis result for distance and direction distributions are plotted in Figure 7.33 and the results are compared with previous results.

The proportion of events indicating non-uniform neighbour directions approaches 80% for point cluster distributions (see Figures 7.25f and 7.33f) but for line cluster distribution it approaches 100% (see Figures 7.26f and 7.33d). The proportion of events indicating non-uniform neighbour directions for point cluster distributions approaches around 100 (see Figures 7.31b, f and h) when all these test results (e.g. Figures 7.26f and 7.31f) are

compared, the difference in the direction test results between point cluster distribution and line cluster distribution is not substantial.

For the number of nearest neighbour up to 60 % of the events in the distribution, there is no substantial difference between the CSR distribution test results plotted in Figures 7.24 e and f compared to Figures 7.33 a and b respectively. There is almost no difference in the distance and direction test results when comparing dense and less dense distributions (Figures 7.01 to 7.08). When the number of nearest-neighbours is four, the proportion of events indicating cluster locations was reduced to 40% (see Figure 7.33e) compared with 70% in Figure 7.25e. The 40% shown in Figure 7.33e is because a high proportion of point clusters overlap and the number of events in those overlapping clusters influences the test results. The analysis results shown in Figures 7.25 e and f, 7.32 and 7.33 e and f for point cluster distributions and the analysis result shown in Figures 7.26 e and f and 7.33 c and d for line cluster distributions indicate that the line and point cluster distributions can not be readily distinguished using the nearest-neighbour method. This confirms the comment by Nicholson [1995], that is for point cluster distribution "the results for the direction tests indicate strong evidence of non-uniformity in direction; this is probably due to nearby clusters appearing to be modes...".

7.3.2 *Mixed distributions*

The nearest-neighbour techniques can identify CSR distributions but not the line or point cluster. Hence, mixtures of point and line cluster distribution were not investigated. Firstly, the nearest-neighbour distance and direction test were used to investigate the mixtures of CSR and point cluster distributions, CSR and line cluster distributions, and CSR, point cluster and line cluster distributions, as shown in Table 7.06.

The nearest-neighbour analysis results are plotted in Figure 7.34 for CSR and point cluster mixtures, Figure 7.35 for CSR and line cluster mixtures, and Figure 7.36 for CSR, point cluster and line cluster mixtures.

Table 7.06: Proportions of each distribution for nearest-neighbour analysis

Mixture Name	Proportion of mixer of distributions		
	CSR	Point cluster	Line cluster
60R-40P	60	40	0
50R-50P	50	50	0
40R-60P	40	60	0
60R-40L	60	0	40
50R-50L	50	0	50
40R-60L	40	0	60
50R-20P-30L	50	20	30
20R-30P-50L	20	30	50
30R-20P-50L	30	20	50

A rapid increase in the proportion of events indicating non-random nearest-neighbour distances was noted as the number of nearest-neighbours increased from 3 to 6 (see Figures 7.34 a, c and d). The proportion of events indicating cluster locations is 54 for the 60R-40P mixture (Figure 7.34a), and 37 for the 50R-50P and 40R-60P mixtures (Figures 7.34c and d) when the number of nearest neighbour is six. The direction distributions also indicate the mixture differences. For the direction test results (Figures 7.34 b, d and f); the three lines (the results from the Rayleigh, Kuiper and Watson tests) approach 100 for the first distribution, approach 90 for the second mixture and approach 80 for the third mixture. Considerable judgement is thus necessary to distinguish between these mixtures. The differences are quite small and may not be statistically significant.

In Figure 7.35, the peak of the proportion of events indicating non-uniform directions is around 80% for plot b, but around 60% for plots d and f. The maximum of the proportion of events indicating clustered nearest-neighbour distance distribution reduced from 45% to 35%. These direction and distance test results indicate the reduction in the proportion of CSR in the mixed distributions, but considerable judgment is necessary to distinguish between these mixtures. The differences are quite small and may not be statistically significant. Figure 7.36 does not indicate the proportion of events from each of the distribution in the mixture. Figures 7.34, 7.35 and 7.36 clearly indicate that the events are not 100% CSR. The plots for different mixtures look quite similar, so it is difficult to

identify the components of the mixtures (and the proportions of the components) from the plots.

7.3.3 *Discussion of nearest-neighbour analysis results*

The following points are noted from the nearest-neighbour analysis with hypothetical distributions.

1. The analysis results shown in Figures 7.24 to 7.27 indicate that the 0.1 confidence level is appropriate for distinguishing between the different hypothetical distributions. Figure 5.04, 7.24, 7.25 and 7.26 indicate that the analysis results are sensitive to the number of nearest-neighbours and the relative size of clusters (i.e. number of events per cluster). Generally accidents are clustered with variable size, therefore the number of nearest neighbours for analysis must be varied from four (as discussed in Section 5.2.1) to a number which is at least just more than the maximum number of events per cluster in the data. To decide the number of nearest neighbours, an iterative approach is probably necessary.
2. One of the advantages of this method is that it indicates the number of events per cluster (e.g. Figure 7.25 e indicates the cluster size is four and 70% of events are from this type of cluster).
3. The nearest-neighbour technique clearly identifies the CSR distribution. For example the proportion of events indicating a non-random nearest-neighbour distance distribution is less than 20 and non-uniform nearest-neighbour directions is less than 25 for CSR distributions shown in Figures 7.24e and 7.24f. The cluster analysis result for one of the mixture (30 events from CSR, 20 events from point cluster and 50 events from line cluster distribution) indicates the distribution is CSR (see Figure 7.23) but the nearest-neighbour analysis result for the same mixture does not indicate a CSR distribution (see Figures 7.36 c and d).
4. Reasonable care should be taken while interpreting the analysis result. For example, the proportion of events indicating clusters in Figure 7.25e is 70% and in Figure 7.33e is 40%, when the distributions used for analysis were 100% point clusters. The possible reasons for the discrepancy between the actual proportion (100%) and the results (70% or 40%) may be:
    1. overlapping clusters affect the results,

2. some proportion of events may be from CSR or line cluster distribution,
3. the number of events per cluster may be varied.

In Figures 7.25e and 7.33e, the possible reason for the indication of the proportion of clustered events less than 100% is because of overlapping clusters, which was clearly identified from the location plots. If the situation was the third, then the distance distribution plots (Figures 7.25e and 7.33e) would have indicated several small peaks in the plotted cluster-profile while troughs in the plotted regular-profile. Other ways of testing these indications are covered in the next section.

The nearest-neighbour analysis method is very useful compared to the cluster analysis method. It is important to note that considerable judgement (e.g. see bullet point 1 and 4 above) is needed when analysing the nearest-neighbour results.

7.4 Quadrat analysis results

The accident-centred quadrat method, which was introduced in Chapter 6, is designed for investigating accident clusters. The buffer zone approach was used to ensure the quadrats do not lie outside the area of investigation, as was discussed in Chapter 3. The width of the buffer zone is equal to the radius of the quadrats. In this section, the performance of the accident-centred quadrat method is analysed for several hypothetical distributions and using 14 different indices.

7.4.1 *Basic distributions*

In the initial stage of testing the indices, four basic hypothetical distributions (i.e., CSR, line cluster, point cluster and regular) were used. The four location plots shown in Figures 7.01, 7.02, 7.03 and 7.04 were analysed using the 14 indices noted below;

1. the mean count (MEAN),
2. variance of count (VARIANCE),
3. coefficient of variation (CV),
4. percentage of quadrats with single accident (SAQ-%),

5. percentage of quadrats having counts above the mean (% QAM),
6. the maximum count (MaxCou),
7. skewness,
8. kurtosis,
9. index of clumping (ICS),
10. index of cluster frequency (ICF),
11. index of mean crowding (ICR),
12. index of patchiness (IP),
13. Morisita index (MI) and
14. proportion of mean count (PMC).

The MEAN, VARIANCE, CV, SAQ-%, %QAM, MaxCou were explained in Section 6.3, skewness and kurtosis were explained in Section 6.3.2, and ICS, ICF, ICR, IP, MI and PMC were discussed in Section 6.3.1. The indices were calculated for accident-centred quadrat counts, for quadrats with radii ranging from 3 to 60 units. Overlapping quadrats were also used in this analysis. The results are plotted in Figures 7.37, 7.38, 7.39, 7.40 and 7.41.

The mean and the variance calculated from quadrat counts are plotted against radius in Figures 7.37a, c, e and g for the CSR, line cluster, point cluster and regular distributions respectively. In the case of the CSR distribution, the variance is slightly less than the mean, for the radius between three and thirty-five. When the radius increases above thirty-five, the difference between the mean and variance increases. The possible reasons are:

1. an increase in the proportion of overlapping area in each quadrat;
2. because the quadrats were centred on accidents, the frequency of quadrat counts in the lower range was small (e.g. zero counts cannot occur) compared with randomly centred quadrat count frequency. The smallest quadrat count will depend on the quadrat radius. As the quadrat radius increases, the frequency of single accident quadrats will decrease. If the radius increases further then the two accident quadrats will decrease or may not be present, and so on. That is, when the quadrat radius increases the mean quadrat count increases in the accident-centred quadrat count distribution.

Due to the above two reasons, the mean shown in Figure 7.37a increased when the radius was increased, but the variance in the quadrat counts was decreased. To minimise these two effects on the results, we will consider only small radius quadrats. The point should be noted

that the increments in the mean and variance are equal when the quadrat radius increases to a certain limit. This is further explained in Section 7.4.4.

For the results shown in Figure 7.37 c for the line cluster distribution, the mean and variance plots diverge and then converge, and the variance is higher than the mean for most of the radius range, which indicates that the distribution is clustered and is not CSR or regular. If the distribution was regular then the mean plot would show step increments (see Figure 7.37g).

The results in Figure 7.37e for the point cluster distribution shows that the variance is below the mean only from three to 17 units of quadrat radius, but beyond 17 units the variance is higher than the mean. The variance being less than the mean indicates that the events are regular or random for the quadrat radius less than 17 units but when the quadrat radius is greater than 17 units, the quadrat counts indicates the events are clustered. If the spatial distribution is regular, then there are step increments in the mean, as can be noted in Figure 7.37g. Step increments are not noted for the quadrat radius between 3 to 17 units. Therefore, it appears the events are distributed randomly within 17 units radius of quadrats, but the events are clustered when the quadrat radius lies between 17 units and 60 units. The point cluster distribution appears as a regular distribution when the quadrat radius is from 5 to 12, but when the radius is greater than 17, the variance is greater than the mean. Figure 7.37e indicates that in most cases, the events in each cluster are within 17 units. This is a notable distinction between a CSR and a point cluster distribution.

In Figure 7.37 c, the variance is less than the mean for very small quadrat radii. However, the difference between the mean and variance seems quite substantial, especially in Figure 7.37 e. Therefore point cluster or line cluster distributions may not be distinguished using this approach because the differences in Figures 7.37 c and e are not substantial. The results shown in Figure 7.37 g for the regular distribution indicates the distribution can be identified (i.e. the special profiles; variance is almost zero and step increment in the mean line).

Among Figures 7.37 b, d, f and h, there are notable changes in the ICF line but the changes in the ICS line are not notable. The ICR line for plot h is quite different than for others. There are step changes in ICF and ICR for the regular distribution. The ICF is negative for the CSR distribution but for the line or point cluster distributions, the ICF starts with

negative value and then suddenly changes to positive value. The ICF line crosses the x-axis at the radius of 20 units in Figure 7.37f. This indicates again that the clusters appear to lie approximately within a 20 units radius circle. This type of change does not appear in Figure 7.37 b. The point cluster and CSR distributions are clearly distinguishable in Figures 7.37 b and f. It appears that the point cluster and line cluster may not be easily distinguished using this method.

The coefficient of variation (CV), index of patchiness (IP) and Morisita index (MI) are plotted against the quadrat radius in Figure 7.38, in which plots a, c, e and g are for the CSR, line cluster, point cluster and regular distributions respectively. The plots indicate that the MI and the IP are equal and this is explained in Chapter6. The three indices (i.e., MI, IP and CV) do not help to distinguish the three distributions (CSR, line cluster, point cluster) because the difference in the Figures 7.38 a, c and e are small, but there are distinctive step changes in indices MI and IP in Figure 7.38 g, (i.e. for the regular distribution).

The proportion of single accident quadrats (SAQ), maximum counts (MaxCou) and the proportion of the quadrats with counts greater than the mean (QAM), are plotted against the quadrat radius in Figures 7.38 b, d, f and h for the CSR, line cluster, point cluster and regular distributions respectively. In the Figures 7.38b, d and f there are notable changes in the SAQ line. For small quadrat radius, the SAQ for the CSR distribution is very high (almost 100%), the SAQ for the line cluster distribution is moderately high (almost 70%) and the SAQ for the point cluster distribution is low (about 40%). Step changes can be noted in SAQ and "MaxCou" for the regular distribution (plot h). In Figures 7.38 b, d and f the "MaxCou" line does not display any notable differences. The difference between the QAM line for the CSR distribution and the QAM line for the point cluster distribution is notable, but the difference between the QAM line for the line cluster distribution and the CSR distribution is not notable. There is a regular fluctuation of the QAM line, with reducing amplitude as the radius increases from 20 to 60 units for the CSR distribution, but the fluctuation is irregular for the point cluster distribution.

In Figure 7.38f, the QAM line drops suddenly when the radius changes from 10 to 11units, and this phenomenon is now explored. For the distribution shown in Figure 7.03, the mean count, variance, SAQ, MaxCou and QAM for different quadrat radii are given in Table 7.07.

Table 7.07: Comparison of indices for different quadrat radii

Index	Quadrat radius (units)							
	7	8	9	10	11	12	13	14
Mean count	3.2	3.5	3.7	3.9	4.1	4.2	4.3	4.5
variance	0.9	0.7	0.6	0.7	0.8	0.8	1.2	1.6
% SAQ	6.25	3.19	1.09	1.15	0	0	0	0
MaxCou	5	5	6	7	8	8	9	10
% QAM	41.67	58.51	71.74	78.16	13.79	13.79	16.28	16.67

In Table 7.07 it is noted that for a quadrat radius of 10 units, 78% of quadrat counts are between 7 and 3.9 (i.e. 4 events per quadrat), and only 1.15% of quadrats are single accident quadrats, and the variance of the counts is 0.7. When the quadrat radius becomes 11, the QAM reduces to 13.79%, the maximum count increases by one, the mean count increases by 0.3 and the proportion of single accident quadrats becomes 0%. This indicates that less than 14% of quadrats have more than four events. Figure 7.38f indicates the QAM is around 20 when the radii are between 14 and 27. This is another indication that there is a high proportion of events with four events per cluster and that there are few clusters with more than four events per cluster. Several drops in maximum count can be noted in 7.38f and the reason being the effect of edge correction (i.e. when the radius increases, a few accident centred quadrat and cluster are not considered because they are in the buffer zone).

The maximum events per cluster is about 20 (i.e., the maximum number of events per quadrat is around 20) for radii between 26 to 44. This is another indication that a bigger sized cluster (around 20 events), within a circle with radius of about 30 units, is present in the point cluster distribution shown in Figure 7.03. When the radius is greater than 10, the SAQ is zero. The mean number of events per quadrat for a radius of 11 is 4.1 and there is no single event quadrat for this quadrat radius. Therefore, it appears that the events are not random but are point clusters, that each point cluster is within a radius of approximately 11 units and that the number of events per cluster is four. If this is the case then the variance should be nearly zero for the quadrat radius of 11 units. The variance starts to increase when the quadrat radius is greater than 12 (Figure 7.37e). When the radius is 11m, the variance (0.8) is much lower than the mean (4.1), which justifies the above conclusion (i.e. all the events are from clusters having four events per cluster).

Figure 7.25e indicates 70% of events are random when four nearest-neighbours were analysed for the same point cluster distribution (Figure 7.03). A point noted in Section 7.3.1 is that 28% of events appear to be either from a CSR distribution or from overlapped clusters. From the previous paragraph it appears that 28% of events are from overlapped clusters and not from a CSR distribution. This is investigated further with the help of frequency polygons for the quadrat counts.

Quadrat-count- frequency polygons for quadrat radii of 7, 10, 12, 14, 20 and 27 units, were selected for investigation because at these radii there are some apparent changes in the analysis results shown in Figures 7.37e and 7.38 f, for the point cluster distribution (Figure 7.03). As indicated in Figure 7.37e, the rate of variance changes for a radius of 12, and the mean and the variance are equal when the radius is 20. A trough and a peak in the QAM line for the quadrat radii 7 and 10 (see Figure 7.38f) were chosen for the investigation. Also, a notable increase in the rate of change of QAM is apparent when the radius is 27. The frequency polygons for these radii are shown in Figure 7.39.

Figures 7.39 a – f show that the shape of the frequency polygon changes when the quadrat radius changes. The shape of the polygon is negatively skewed when the quadrat radius is 7 units, fairly symmetrical when the radius is 10 and positively skewed when the radii are 12, 14, 20 and 27. The highest quadrat count frequency (approximately 70) occurs when the quadrat radius is 12 units and the count is four. This coincides with the low QAM value (13.79) for a radius of 12 units (see Table 7.07).

Figure 7.39a is a negatively skewed polygon, which indicates a point cluster distribution as discussed in Section 6.3.2. Figures 7.39c, d, e and f indicate positively skewed polygons but the frequencies of quadrats with 1, 2 or 3 events are very low. Although they are positively skewed polygons it does not necessarily mean that the events are from a CSR distribution.

The frequency polygon shown in Figure 7.39e is positively skewed. The mean and variance of quadrat counts are equal when the quadrat radius is 20 as shown in Figure 7.37e. Although the mean being equal to the variance is expected for a CSR distribution, it does not mean that the events are from a CSR distribution, because the above indications (positively skewed and mean equal to variance) do not simultaneously hold for radii other

than 20, as shown in Figure 7.39a. The above results indicate that several quadrat radii should be used in order to make a sound judgement about a distribution.

The frequency polygons in Figure 7.39 show how the shape changes from negatively skewed to zero skewness and then to positively skewed when the quadrat radius changes. Figure 7.39c indicates that most of the events are from a point cluster distribution with four events per cluster, and not from a CSR distribution. This is useful to answer the question which arose in Section 7.3.1 (i.e. whether the 28% of events are from a CSR or a point cluster distribution). The shape of the frequency polygons is investigated with the help of the skewness and kurtosis. The plots of skewness and kurtosis against quadrat radius are shown in Figures 7.41b, d, f and h for the CSR, line and point cluster and regular distributions respectively. Figure 7.41f indicates that the skewness changes from negative to positive when the range of the radius is 9 to 11 with a zero skewness at the quadrat radius 10. There is a +ve to –ve change in skewness at quadrat radius 5, but this is small compared to the –ve to +ve change at quadrat radius 10. This information indicates that a detailed investigation of the frequency polygon is useful for the quadrat radius around 10.

A distribution more peaked and with more concave shoulders than the normal distribution is known as leptokurtic with the kurtosis being greater than three. The frequency polygons shown in Figure 7.39 b, c, d, e and f appear to be leptokurtic. The kurtosis depends critically on the tails of the frequency distributions [Frank & Althoen 1994], as discussed in Chapter 6. If the quadrat counts are more from the shoulders than from the tails of the distribution, then kurtosis tends to indicate a lower value than expected. This is applicable for skewness as well but the effect is less for kurtosis. Hence there is a need to check the frequency polygon rather than relying on the indices of skewness and kurtosis.

For the frequency polygons for a point cluster distribution shown in Figure 7.39, the hypothesis (i.e. the distribution is Poisson at the significance level of 0.005) is not rejected but they do not look like Poisson. For the point cluster distribution shown in Figure 7.03, the frequency polygon in Figure 7.39a indicates negative skewness (-0.52) and the rest of the frequency polygons (Figures 7.39 b, c, d, e and f) indicate positive skewness. For a CSR distribution, all the frequency polygons shown in Figure 7.40 indicates positive skewness.

Frequency polygons are shown in Figure 7.40 for the CSR distribution (Figure 7.01). The frequency polygon indicates that the frequency of single-event quadrats decrease when the quadrat radius increases and the frequency polygon is still positively skewed. Figure 7.40 illustrates how the shape of the frequency polygon changes with increasing quadrat radius. In all the distributions shown in Figure 7.40, the hypothesis (i.e. the distribution is Poisson at the significance level of 0.005) is not rejected and they look Poisson distribution.

For quadrat radius between 7 and 70 units, Figure 7.41b indicates that the kurtosis is between 0 and 3, but for the line cluster distribution the kurtosis is between 1 and 9 (Figure 7.41d) and for point cluster distribution the kurtosis is between 1 and 11 (Figure 7.41f). The kurtosis or the skewness generally indicates the shape of the frequency polygon but the frequency polygon needs further investigation before a conclusion is reached. Figures 7.39 and 7.40 show that investigation of several frequency polygons with different quadrat radii are necessary, because of the sensitivity of the polygon shape to changes in the quadrat radius.

Graphs of mean count against quadrat radius were normalized in the following manner. The ratio of the mean count (mean count for a quadrat radius divided by the mean count for the maximum quadrat radius selected (say 30 units in this set of hypothetical distribution but in real road accident data the maximum quadrat radius need to be selected so that the neighbouring road accidents is not included in the quadrats) against the ratio of radius (quadrat radius/maximum quadrat radius selected) is plotted in Figures 7.41a, c, e and g for the CSR, line cluster, point cluster and regular distributions respectively. In this plot the radii of quadrats is varied (say 3, 4, 5, 6,..30 units).

From the theory the expected curves:

- quadratic relationship for the CSR or the point cluster distribution and
- linear relation ship for line cluster distribution

are discussed in Section 6.3.1.

The plot for the CSR distribution shown in Figure 7.41a indicates a concave quadratic relationship, which is similar to the profile shown for random distribution in Figure 6.02e (i.e. the theory discussed in Section 6.3.1). The plot for the line cluster distribution shown in

Figure 7.41c indicates a linear relationship, which is similar to the profile expected in the theory discussed in Section 6.3.1. The plot 7.33e indicates a convex quadratic relationship, which is a similar relationship to the profile (Figure 6.02e) for the point cluster. The plotted profile in Figure 7.41g indicates step changes because the distances between events are equally distributed.

It was noted that the best-fit line in Figure 7.41a for the CSR distribution is a quadratic relationship, in Figure 7.41c for the line cluster distribution is a linear relationship, in Figure 7.41e for the point cluster distribution is a quadratic and in Figure 7.41g for the regular distribution is a linear relationship. The above-mentioned relationships for CSR, line cluster, point cluster and regular distributions were obtained from the regression analysis using the statistical software (SPSS) and the R-square value and F-ratio are shown in Table 7.08. The significance levels are less than 0.001 for all F-ratios.

Table 7.08: Comparison of linear and quadratic relationships for profiles (Figures 7.41a, c, e and g)

Distribution	Relationship	R-square	Df	F- ratio
CSR	Linear	0.939	1,26	401
	Quadratic	0.998	1,25	8196
Line cluster	Linear	0.999	1,26	22255.2
	Quadratic	0.999	1,25	12915.4
Point cluster	Linear	0.89	1,26	209.63
	Quadratic	0.967	1,25	361.4
Regular	Linear	0.736	1,26	72
	Quadratic	0.761	1,25	40

The R-square values for linear and quadratic relationships are almost the same. The table indicates that the F- ratios for the linear and quadratic relationships are very high. It appears the convex, linear, concave or uniform (i.e. flatness) shape depends on the distribution.

These relationships are further investigated using the distributions shown in Figures 7.05, 7.06, 7.07 and 7.08 in the later part of this section.

Consider now the four basic distributions shown in Figures 7.05, 7.06, 7.07 and 7.08. The results of accident-centred quadrat analysis on each are plotted in Figures 7.42, 7.43, 7.44 and 7.45. Figure 7.42a again indicates that the variance is below the mean as in Figure 7.37a. The important point which should be noted from Figures 7.37a and 7.42a is that for the CSR distribution the variance is approximately equal to the mean for small quadrat radius and the variance is less than the mean for large quadrat radius, but for the line cluster distribution or point cluster distribution the variance is generally greater than the mean for large quadrat radius (see Figures 7.37c, e and Figures 7.42c, e). Figures 7.37g and 7.42g indicate that the variance is equal to zero and the step increment in the mean for the regular distribution.

The reason why the mean and the variance diverge in Figure 7.37a was discussed earlier in this section. In this Figure the variance profile starts to diverge at the quadrat radius of approximately 34 units but in Figure 7.42a the variance profile start to diverge at the quadrat radius of approximately 134 units. The possible reasons for the increase in the critical quadrat radius from 34 units to 134 units are:

- If we consider the quadrats in which the radius is 34 units, the proportion of overlapping area in each quadrat considered for the results shown in Figure 7.42a was less than the quadrat considered for the results shown in Figure 7.37a.
- the quadrat centres were selected only on accidents, so the frequency of quadrat counts in the lower range were less in the quadrats taken for the results shown in Figure 7.37a compared with the quadrats considered for the results shown in Figure 7.42a.
- The density of events in the distribution shown in Figures 7.01 is higher than the distribution shown in Figure 7.05.

The most possible reason is the density differences; the quadrats were taken from a small area of (300 × 300 units) in Figure 7.01, compared with an area of (1000 ×1000 units) in Figure 7.05.

The ICF is less than or equal to zero for the CSR distribution, as shown in Figure 7.42b, which is similar to Figure 7.37b. The changes in the shape of the frequency polygon (Figure 7.45) with the quadrat radius for the CSR and point cluster distributions are similar to those shown in Figures 7.39 and 7.40. There is no notable difference between the plotted results for the two sets of basic distributions (Figures 7.01, 7.02, 7.03, 7.04 and Figures 7.05, 7.06, 7.07, 7.08). Only selected details of Figures 7.42, 7.43, 7.44 and 7.45 are discussed instead of repeating the same points made earlier. The results indicate that the effect of density (i.e. number of events per sq. m.) does not influence the variation in the indices between each of the four basic distributions in the first and second sets.

Using the accident-centred quadrat method, the reason for an unexpected result obtained from nearest-neighbour analysis in Section 7.3 can be explained. The nearest-neighbour analysis result (Figure 7.33e) indicates that 40% of events are from a point cluster distribution, but all the events are from a point cluster distribution shown in Figure 7.06. The three possible reasons why the indicated result is as low as 40% are:

1. a high proportion of events are from a CSR distribution,
2. a large number of clusters are close to each other (mixed clusters or overlapped clusters explained in Chapter 2) in the distribution,
3. variation in cluster size, i.e. a different number of events in each cluster in the distributions.

The reasons are discussed further, using the plotted results from the accident centred quadrat method.

The variance is close to one for the quadrat radius up to 20 (Figure 7.42e). When the quadrat radius is between 43 and 190 units, the variance is greater than the mean. The SAQ profile (Figure 7.43f) is less than 20 when the quadrat radius is 10 units. As the quadrat radius approaches 33 units the SAQ becomes approximately zero. These indicate that most of the events (about 80%) are very close to a neighbour (within 10 units) and that at least one neighbour is within 33 units for each event. When the quadrat radius is 23, Figure 7.43f indicates that the QAM is 70.35%, the maximum count is 8, the mean count is 3.92 and the SAQ is only 4.58. These results indicate that a negligible proportion of events are from the CSR distribution and the first reason is not applicable.

Figure 7.41d and 7.44d are slightly different because the density difference in both distributions. Figure 7.41f and 7.44f are slightly different because of the overlapping quadrats and the density different in the two distributions.

Consider the frequency polygon shown in Figure 7.45e, where the frequency of four events per quadrats is about 200 when the quadrat radius is 25 units. When the quadrat radius is 50 units (Figure 7.45h), the number of four events per quadrats is 103 and the frequency of eight events per quadrat is 61. When the quadrat radius equals 75 units (Figure 7.45i), there are peaks at the quadrat counts of 4, 8, 12, 16 and 20, which are multiples of four. The peaks at 4, 8 and 12 are more pronounced than the peaks at 16 and 20. There is a location in Figure 7.05 where there are 20 events within a quadrat of radius 75. The peaks shown in Figure 7.45i indicate that a number of clusters are very close to each other. Around five clusters are within 75 units of a selected events. This is an indication that a reasonable number of clusters are very close to each other. The closest clusters have influenced the results (40 % of cluster events which is very low proportion) because the nearest-neighbour result is influenced by the relative number of nearest-neighbours and the number of events per cluster. If two clusters are very close to each other and the number of nearest neighbours analysed is less than the total number of events in the two clusters, then the results will indicate that a reasonable proportion of events are from a CSR distribution.

Table 7.09: Comparison of linear and quadratic relationships for profiles (Figures 7.44a, c, e and g)

Distribution	Relationship	R-square	Df	F- ratio
CSR	Linear	0.947	1,149	2646
	Quadratic	>0.999	1,148	1228278
Line cluster	Linear	0.999	1,149	100239
	Quadratic	0.999	1,148	60462
Point cluster	Linear	0.952	1,149	2981
	Quadratic	0.999	1,148	92904
Regular	Linear	0.906	1,149	1443
	Quadratic	0.96	1,148	1770

Regression analysis results tabulated in the table 7.09 were computed using statistical analysis software (SPSS). This table was used to identify the best-fit line in Figures 7.44a, c, e and g. The high F-ratio values are highlighted in the table. All significance levels are less than 0.001. The profiles shown in Figures 7.44a, c and e again indicate that the profile obtained for the CSR distribution is a quadratic relationship, the profile obtained for the line cluster distribution is a linear relationship and the profile obtained for the point cluster distribution is quadratic. The profile obtained for regular distribution is also quadratic fitted to step increments.

Table 7.08 and Table 7.09 indicate that for all four basic distributions, the F-ratio for quadratic and linear relationship are very high. In Figure 7.44e, a convex curve (i.e. similar to Figure 7.37e) can be noted when the proportion of radius is between 0 and 30. In this range the Figure 7.44g shows a straight line with more flatness compared to the Figure 7.44c. The results indicate that the maximum quadrat radius (150 units) selected for these calculations influences the result. If the maximum quadrat radius is chosen properly then concave relationship for CSR distribution and convex curve for point cluster distribution will be observed. The reason for the choice of a high quadrat radius was to highlight the effect of the quadrat radius on the results. Further research is suggested in Chapter 9 to identify the appropriate radius, which is possibly a little greater than the characteristic length of an average point cluster to identify the accident distributions.

So far in this section, the distributions have equal cluster size (events per cluster is constant). Consider now the eight point cluster distributions with variable cluster size (area of cluster), the location plots of which are shown in Figures 7.29 and 7.30. The analysis results are plotted in Figures 7.46, 7.47, 7.48, 7.49, 7.50 and 7.51 for the distribution plots in 7.29 and 7.30.

In Figure 7.46, the variance profile intersects the mean profile at different quadrat radii in each plot. The ICF profiles show the mean and variance lines intersecting each other, because the mean and variance profiles are close to each other for some quadrat radii and the points of intersection are not very clear. Different quadrat radii are selected from the plot (Figure 7.46a, c, e and g) for further analysis using frequency polygons.

In Figure 7.46a, the quadrat radius of 30 units is selected for further analysis because the variance starts to increase above the mean when the quadrat radius increases from 30 units. Figure 7.46g indicates that the variance starts to diverge when the quadrat radius is 17 units. Therefore, the quadrat radius of 17 units is selected for further analysis. Figures 7.46 c and e do not indicate clear intersections or coincide of the mean and variance profiles.

Figures 7.47 b, d, f and h show irregular fluctuation in the QAM profile. The SAQ approaches 0% at the quadrat radius 32 units in Figure 7.47b, 20 units in Figure 7.47d, 25 units in Figure 7.47f and 18 units in Figure 7.47h. All the profiles shown in Figures 7.48 a, c, e and g are reasonably linear. The reason for different shape in the maximum quadrat size selected for the calculation was explained earlier in this section. It means that all the four distributions are from cluster but not from CSR distributions. The plots shown in Figures 7.47 a, c, e and g, and in Figures 7.48 b, d, f and h do not exhibit any distinctive features which may help to distinguish between the different type of distributions.

Similar conclusions mentioned in the above paragraph are applicable to Figures 7.49, 7.50 and 7.51. It appears that the mean and variance line intersect only twice in Figure 7.49a but in Figures 7.49 c, e and g the variance and the mean are approximately equal for most of the quadrat radii and the intersecting points are not clear. Therefore the distribution plot 7.30a is selected for further analysis using frequency polygons. In Figure 7.50b, the maximum count (MaxCou) is constant for the quadrat radius between 20 and 25 units, and between 40 and 50 units. So, the frequency polygon for the quadrat radii of 20 units and 40 units were selected to investigate the cluster variation.

In Figures 7.38f and 7.43f, the QAM profile is visibly irregular but Figures 7.47 b, d, f, h or Figures 7.50b, d, f and h the irregularity of the QAM profile is less visible. In Figure 7.38 b and Figure 7.43 b generally a regular nature can be noted for CSR distribution. In Figure 7.38f or Figure 7.43f there is a sudden drop in QAM profile at a certain quadrat radius and is then nearly stable but in Figures 7.47 b, d, f, h or Figures 7.50 b, d, f and h the sudden drop and stable nature in QAM profile occur several times for different quadrat radius. The reason is due to the various size of cluster (from 2 to 20 events per cluster and the area of cluster is different) in these point cluster distributions and the analysis results are shown in Figures 7.47 b, d, f, h or Figures 7.50 b, d, f and h but the results shown in Figures 7.38f and 7.43f a is for single sized cluster (4 event per cluster and each cluster is within a constant

area) distribution. In Figures 7.37e and 7.42e, the variance and the mean profiles clearly intersect only twice but in Figures 7.46 a, c, e or Figures 7.49 c, e, g. the variance line is close to the mean line at several quadrat radii. This indicates several sizes of clusters are present in the distributions.

Four frequency polygons are shown in Figure 7.52. Figure 7.52a is for a quadrat radius of 30m for the distribution shown in Figure 7.29a, Figure 7.52b is for a quadrat radius of 17m for the distribution shown in Figure 7.29d and Figures 7.52c and d are for the quadrat radii of 20 and 40m for the distribution shown in Figure 7.30a. Each of the four frequency polygons have more than one peak. These peaks indicate different numbers of events per cluster in each of the distributions shown in Figures 7.29a, 7.29d and 7.30a.

The results indicate that whether the distributions involve dense (events in each cluster is very close to each other) or sparse (events in each cluster are far away from each other) point clusters does not influence the overall plotted results. It should be clear that the variance line should be below the mean line for CSR distributions, but for line or point cluster distributions the variance line will intersect the mean line. The reason that the variance is not equal to the mean for the CSR distribution is explained later in this chapter. The following points are noted from Figures 7.46 and 7.49.

1. The variance and mean lines being very close and coinciding each other (e.g. Figure 7.46c) does not mean that the spatial distribution is CSR. The distribution could be a point cluster distribution with variable cluster size (based on the number of events per cluster).
2. If the variance is close to zero it means the quadrat counts are fairly constant. For example, the variance line does not approach zero when the quadrat radius is greater than 5 units (Figures 7.49a, c, e and g), but in Figure 7.37e the variance is approximately zero for the quadrat radius up to 12 units for constant size point cluster distribution.
3. In Figure 7.46 the mean and variance lines coincid once for a small quadrat radius but for large quadrat radius the variance is distinctly greater than mean. This result is possible when analysing distributions, which have single or several sizes of clusters.

7.4.2 *Mixed distributions*

Four mixtures 70R-30P, 60R-40P, 50R-50P and 40R-60P are shown in Figure 7.53 a, b, c and d respectively. The last three mixtures were analysed in Sections 7.2 and 7.3 using cluster analysis and nearest-neighbour method. The results for the four mixtures (added to highlight the differences in the results) are shown in 7.54, 7.55 and 7.56. These indicate the effect of increasing the proportion of events from the point cluster distribution in the mixed distribution. The following points can be observed from Figures 7.54, 7.55 and 7.56 when the proportion of events from the point cluster distribution is increased in a mixed distribution.

1. On average the difference between the mean and the variance increases in each of Figures 7.54 a to c to e and to g.
2. The mean number of events in the quadrat for a radius of 70 units increases in each of Figures 7.54 a to c to e and to g.
3. The quadrat radius of SAQ-50 (i.e. 50% of events indicating single accident quadrats) increases in each of Figures 7.55 b to d to f and to h.
4. The size of the sudden drop in the QAM increases in each of Figures 7.55 b to d to f and to h.
5. The skewness of the count distribution become slightly more negative, and the kurtosis become slightly more positive for the quadrat radius of 10 units in each of Figures 7.56 b to d to f and to h.

The information noted in 1- 5 can be used to monitor the proportions of events from CSR or point cluster distributions.

7.4.3 *Non-overlapping accident-centred quadrat analysis*

Non-overlapping accident centred quadrat counts can be tested with the truncated Poisson distribution, as discussed in Section 6.3.4, and are computed for the distribution shown in Figure 7.05, for a quadrat radius of 30 units. The process for selecting non-overlapping quadrats is as follows.

- Number each accident randomly
- Select the first numbered accident as the centre of the first accident-centred quadrat.

- Find out whether the second numbered accident-centred quadrat is overlapping with the first selected accident-centred quadrat.
- If it does not overlap with the first selected quadrat then that is the second selected accident-centred quadrat.
- If the second numbered accident-centred quadrat does overlap with the first selected quadrat then discard the second quadrat and find out which is the next numbered accident-centred quadrat which does not overlap the first quadrat (that quadrat becomes the second selected non-overlapping quadrat).
- Next, select the third quadrat which does not overlap with the selected two non-overlapping quadrats.
- This process is continued till the number of non-overlapping quadrat selected is 30 and is the first set of non-overlapping quadrats.

The above process is continued until 30 sets of quadrats were selected. The quadrats within each set are non-overlapping. The accident count frequency was computed separately for each set. The set of 30 frequency distributions were tested using the Chi-square test to find out whether the distribution is well-described by the truncated Poisson distribution. It was found that the hypothesis (that the counts are truncated Poisson distribution) was not rejected for 22 of the 30 sets of quadrat counts. The result indicates that the zero truncated Poisson distribution is not a good model for the count frequency for non-overlapping accident-centred quadrats with radius of 30 units. This is discussed again in Section 7.4.4.

7.4.4 *Discussion of quadrat analysis results*

The overall results indicate the accident centred quadrats provide a useful method for analysing the accident distribution, but to interpret the results, analysis of several graphs are needed. It appears that the six indices ICS, ICR, IP, MI, CV and PMC do not provide sufficient information to help identify the characteristics of the distribution. The other indices (i.e. mean, variance, ICF, SAQ, QAM and MaxCou) are useful indicators and help to identify the characteristics of the distribution.

For a CSR distribution, the mean and variance plots indicate that the mean is very close to the variance up to a certain radius (e.g. Figure 7.42a). If the radius is large however, there will be few (if any) quadrats with only one accident, the frequency of large quadrat counts will increase, and the effect on the variance is unclear.

Figures 7.46 a, c, e, g and Figures 7.49 c, e, g indicate that if the mean and the variance lines are close together and intersect, then the distribution will have variable sized (events per cluster) point clusters. If the variance line is very close to the mean line then that is not sufficient to conclude that the distribution is CSR. If the distribution is CSR then the mean and variance line will be close but will not intersect. If the proportion of events from a CSR distribution decreases in the mixture of point cluster and CSR distributions, then the variance line moves away from the mean line (e.g. Figures 7.54 a, c, e and g).

The identification of line clusters (i.e. black route) was discussed in Chapter 6 and will be discussed in Chapter 8. The quadrat analysis method is not helpful in identifying line clusters without including further details such as the name of the road where the accidents occurred. The results from the accident-centred quadrat method indicate that the method helps to distinguish between the point cluster and CSR distributions, and to monitor the progress of point clustering or CSR in distributions.

7.5 Comparing statistical techniques with visual examination

Cressie [1993] noted that, due to the subjective nature of the visual examination approach, observers may well disagree as to the existence and nature of any pattern in an accident distribution. The CSR distribution can appear to be a highly clustered distribution for less trained and less experienced examiners. That is, the traditional method of assessing spatial pattern by visual examination is not dependable. It is believed that the analytical method developed in this thesis is more dependable and examples are included in the appendices as illustrations. Readers are invited to assess the distributions visually and to compare with the results with the statistical techniques results. A sequence of mixtures of point cluster and CSR distribution is shown in Appendix A, Figures A.01 to A.10 and the corresponding analysis results are in Appendix A, Figures A.11 to A.18. A sequence of mixtures of line

cluster and CSR distributions are shown in Appendix B, Figures B.01 to B.11 and the corresponding analysis results are in Appendix B, Figures B.12 to B.19.

In Appendix A, Figures A.11 to A.14, the area between the mean profile and variance profile increases when the proportion of point clusters increases. The proportion of single accident quadrats (%SAQ) profile and the proportion of multiple accident quadrat (%MAQ) profile also show changes when the proportion of point clusters increases. Arranging the plots shown in Appendix A, Figures A.01 to A.10 in order of increasing point clustering through visual examination (i.e. viewing the location plots and ordering the figures according the proportion of events that are from the point cluster distribution) is difficult compared to using the analysis results shown in Appendix A, Figures A.11 to A.14. The nearest neighbour analysis result shown in Appendix A, Figures A.15 to A.18 also show the difference in the mixtures.

In Appendix B, Figures B.12 to B.15, when the line clusters increases, the area between the mean profile and the variance profile also increases noticeably, but it is difficult to observe substantial differences between the plots. When the line clustering increases, the proportion of single accident quadrats profile and the proportion of multiple accident quadrat profile indicate a small change.

Arranging the plots shown in Appendix B, Figures B.01 to B.10 in order of increasing line clustering through visual examination (i.e. viewing the location plots and ordering the figures according the proportion of events that are from the line cluster distribution) is difficult compared to using the analysis results shown in Appendix B, Figures B.11 to B.14. The nearest neighbour analysis result shown in Appendix B, Figures B.15 to B.18 also show the difference in the mixtures.

7.6 Summary

The overall test results from the three techniques (i.e. cluster analysis, nearest-neighbour analysis and quadrat analysis) show that the single-linkage cluster analysis method helps in identifying a line cluster (i.e. black route) but the nearest-neighbour and quadrat techniques are very helpful in distinguishing between the point cluster and CSR distributions. All the

test results from the hypothetical distributions indicate that identifying the dominant pattern in a mixture of three basic distributions is difficult. Identifying line clusters using additional data (road names) was discussed in Chapter 6, and a method involving using road name data will be discussed further in Chapter 8.

The cluster analysis results (discussed in Section 7.2) indicate that the method is helpful in analysing accident distributions without additional details such as the road names. The cluster analysis results (Figures 7.22a and 7.22c) indicate that CSR distributions can be identified with 95% confidence limits, and line and point cluster distributions can be identified with 85% confidence limits. Figure 7.23 indicates that identifying the major component in the mixed distribution is difficult and further research to improve the sensitivity will be suggested in Chapter 9. Therefore cluster analysis is not used in Chapter 8.

The results from the nearest-neighbour and accident-centred quadrat methods are useful for analysing accident distributions. The results are plotted in graphical form but the interpretation of results requires considerable judgement and careful attention.

To identify a line cluster (black route) using the accident-centred quadrat method, additional data (the road name where each accident occurs) is needed. The characteristic of a black route was discussed in Chapter 1. As mentioned in IHT [1986] a road having more than the average number of accidents is a black route for that class of road. Identification of a road (or road section) having a high intensity of accidents involved analysing accident-centred quadrat counts. The accidents need to be counted from the road where the quadrat centre is located and the accidents which occur on other roads within the quadrats should not be counted. This is illustrated in Figure 6.03. The quadrat radii are from 250m to 750m, depending on the road network. The maximum quadrat count indicates which road or section of the road has a high number of accidents. If a road has a high accident density then high density quadrats (i.e. the quadrats having high number of accidents compared to the rest of the quadrats for each radius) will be indicated for that road. If the high intensity quadrats are spread all over the network then the route plan is not appropriate.

The flowchart shown in Figure 7.58 shows how the analysis should be done using the quadrat and nearest-neighbour techniques. If large clusters (line clusters) are present in the

accident data then the analysis results from the nearest-neighbour or accident-centred quadrat (radius between 5 and 150m depending on whether road network is dense or sparse, could indicate that a very high proportion of accidents are from a CSR distribution. Selecting the number of nearest-neighbours or quadrat radius depends on the relative cluster size (number of events or area per cluster). If we analyse using the number of nearest-neighbours or quadrat radius less than the cluster size, then the result may indicate that the distribution is CSR (discussed in Section 7.3, 7.4). The number of events or area per cluster in the line cluster distribution is greater than the number of events or area per cluster in the point cluster distribution. Therefore we need to identify line clusters first and then check the CSR distribution at the last stage, and the orders are shown in the flowchart (Figure 7.58). If the distribution is CSR then there is no need to analyse accident data further.

Even if a line cluster is found in the accident distribution, we should make sure that the accident distribution is not CSR. To confirm that the distribution is not CSR, nearest-neighbour or accident-centred quadrat method will be used. We still need to bear in mind that if there are line clusters then the analysis method could indicate a very high proportion of events from the CSR distribution.

Nicholson [1995] noted that if the road network is dense (i.e. the space between the roads is relatively small block size $\leq$100m) then a continuum is a good approximation to a lattice, reasonably ok up to 250m and if the block size is greater than 500m then a continuum approximation is not good. This means that if the road network is relatively dense (block size is less than 100m) we are able to use the nearest-neighbour analysis.

The difficulty of using the lattice distance between accidents in a road network was discussed in Chapters 2 and 3. This requires further research. In a continuum, the straight-line distance is measured between two accidents. This distance may be affected by the space between two roads. This is a reason why the space between roads limits the use of nearest-neighbour analysis. This limitation is applicable to cluster analysis also, because straight-line distances are used in cluster analysis. However, the accident-centred quadrat method does not involve quadrats covering the space between roads (and without accidents). It appears that the accident-centred quadrat method can be used for the block size greater than 250m.

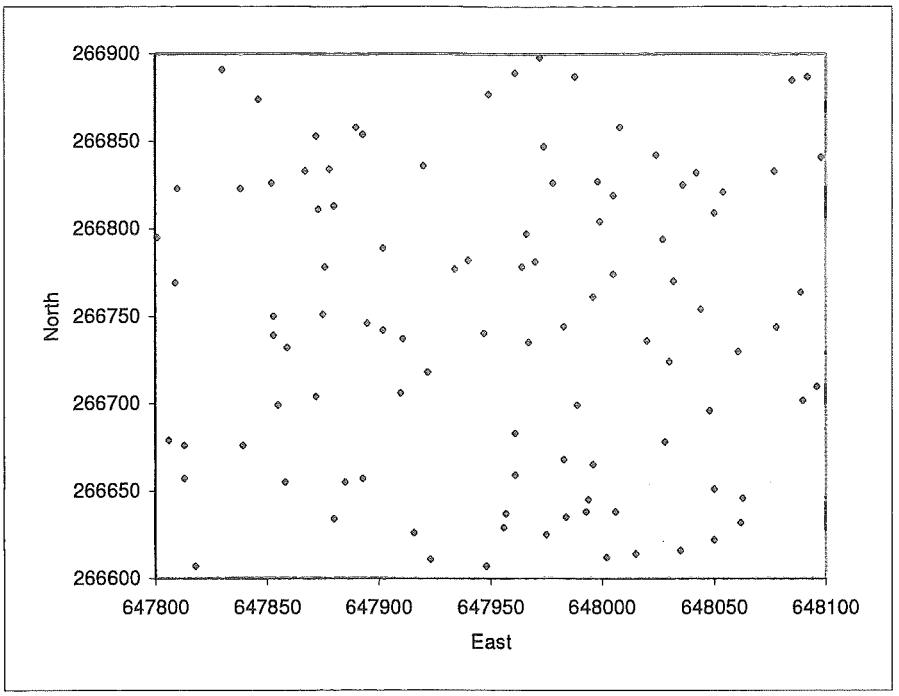


Figure7.01: CSR distribution (100 events in 300^2 sq.units)

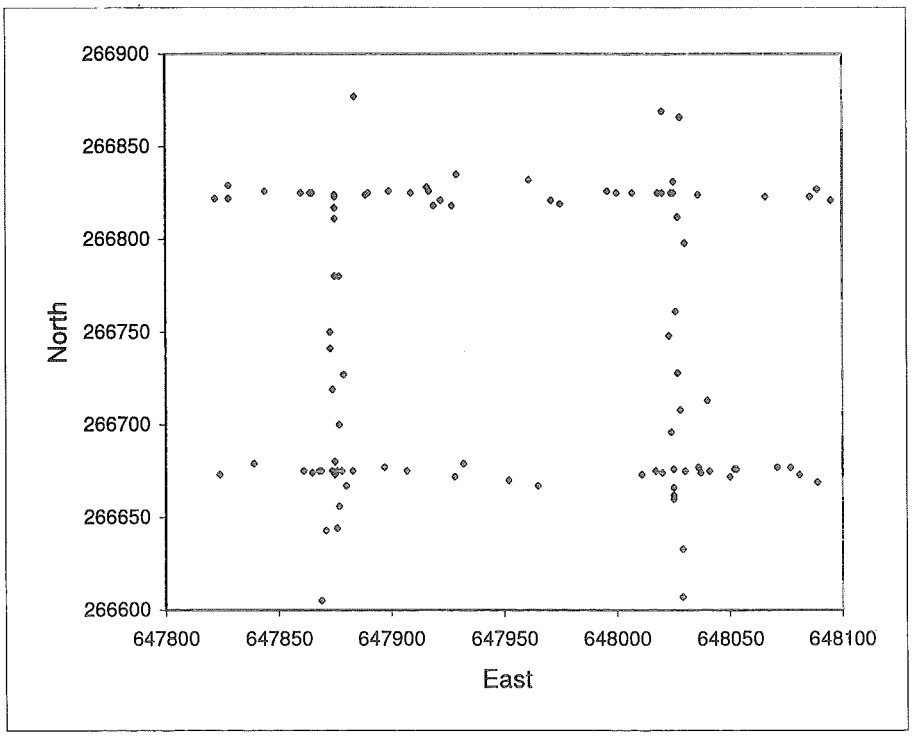


Figure7.02: Line cluster distribution (100 events in 300^2 sq.units)

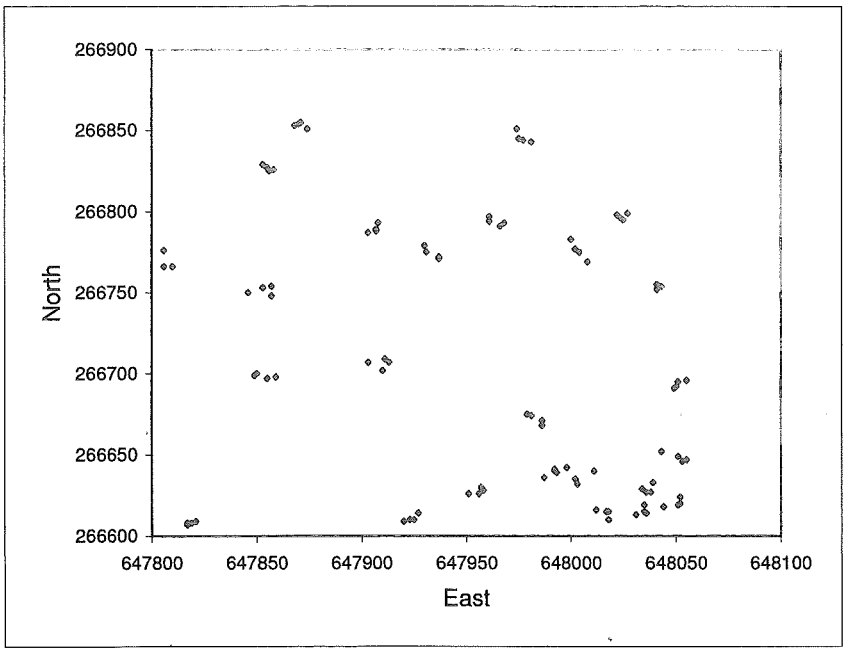


Figure7.03: Point cluster distribution (100 events in 300^2 sq.units)

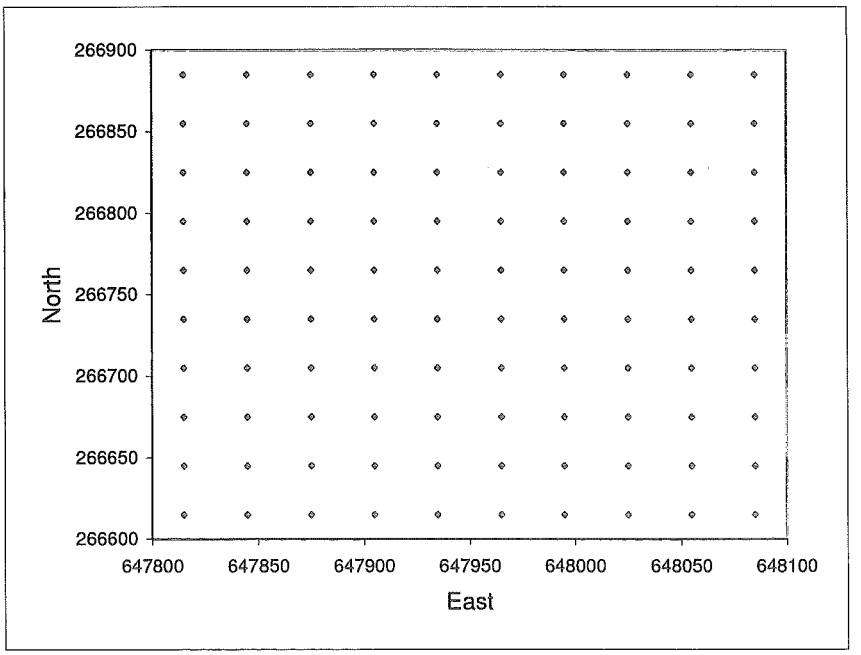


Figure7.04: Regular distribution (100 events in 300^2 sq.units)

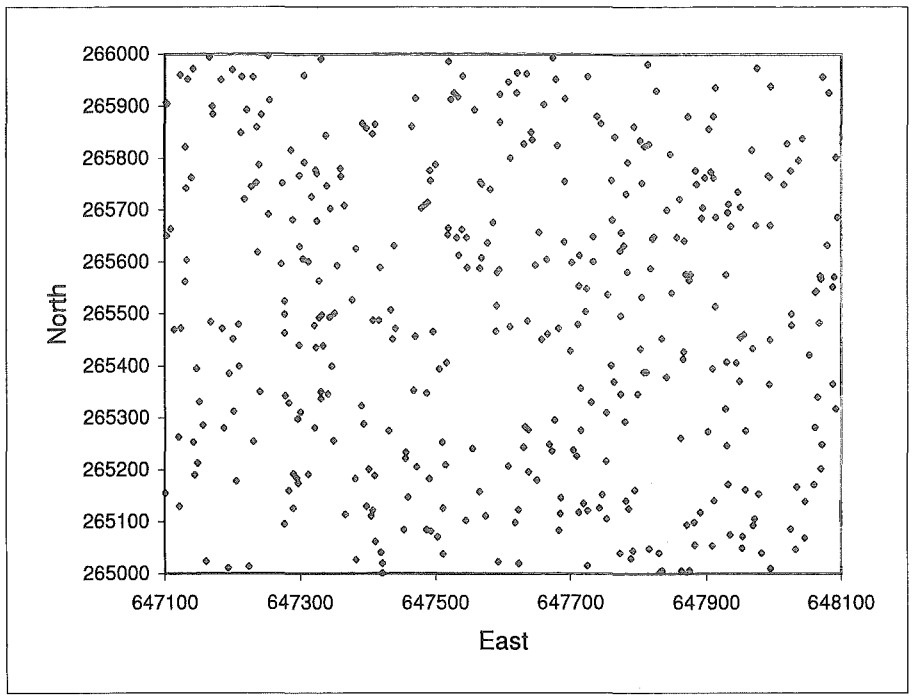


Figure7.05: CSR distribution (400 events in 1000^2 sq.units)

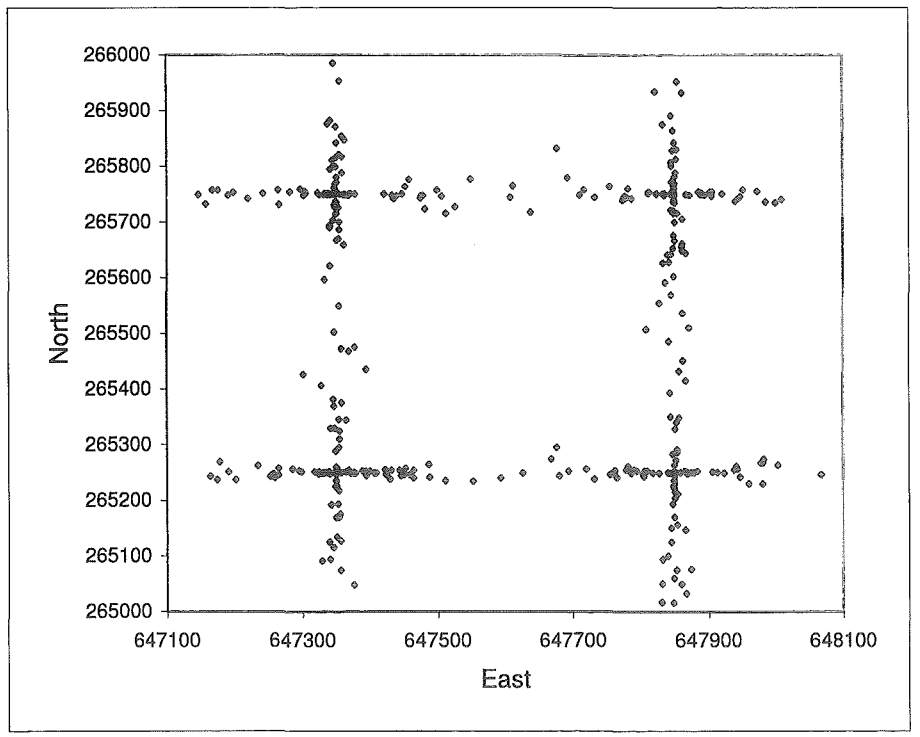


Figure7.06: Line cluster distribution (400 events in 1000^2 sq.units)

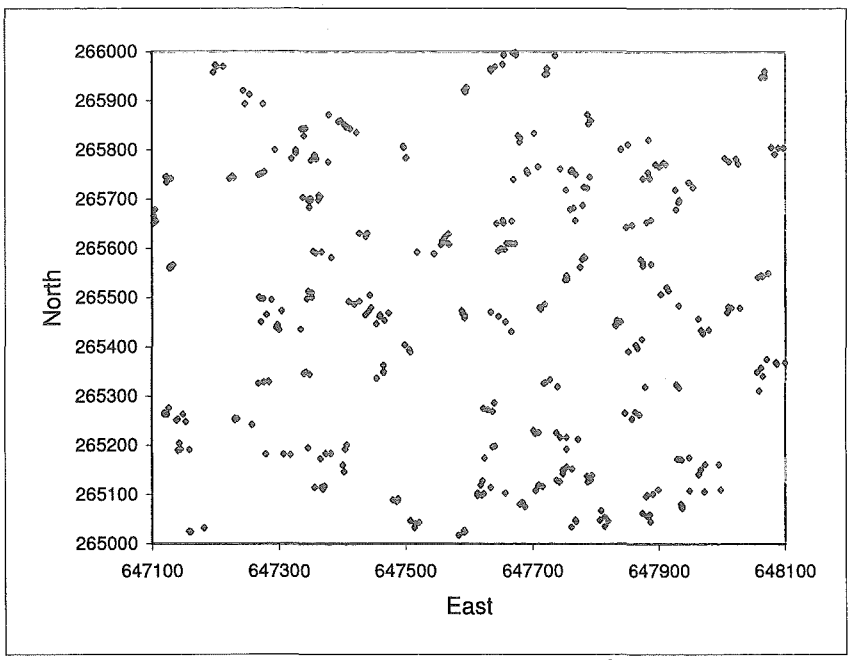


Figure7.07: Point cluster distribution (400 events in 1000^2 sq.units)

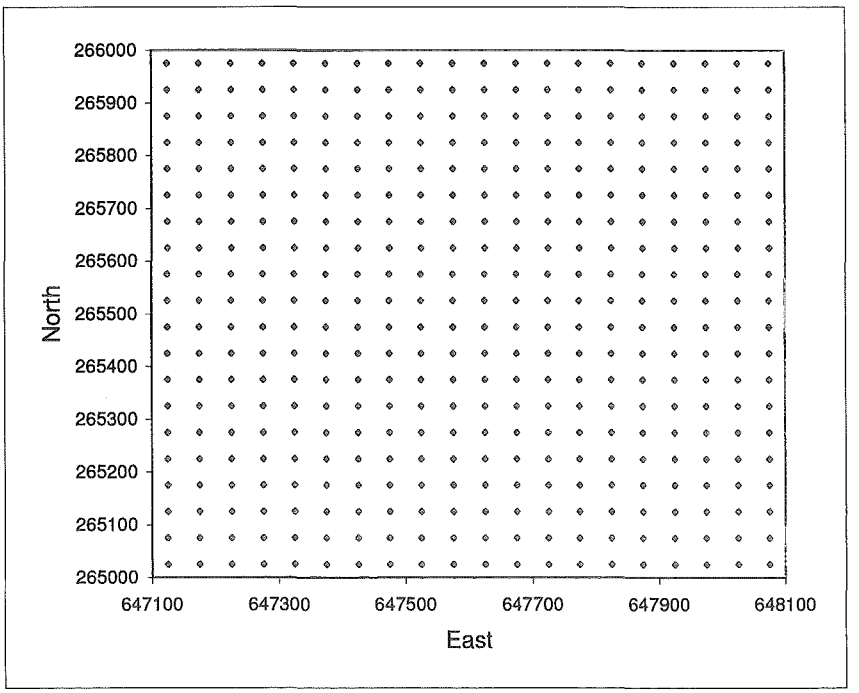
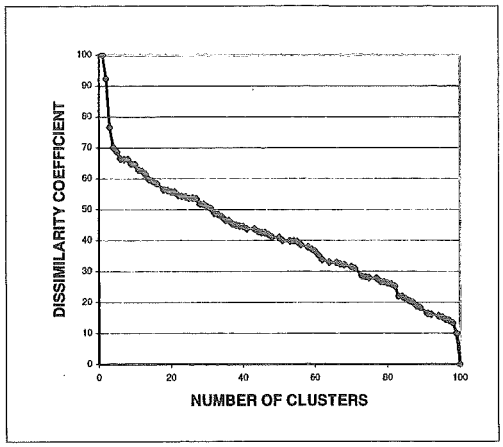
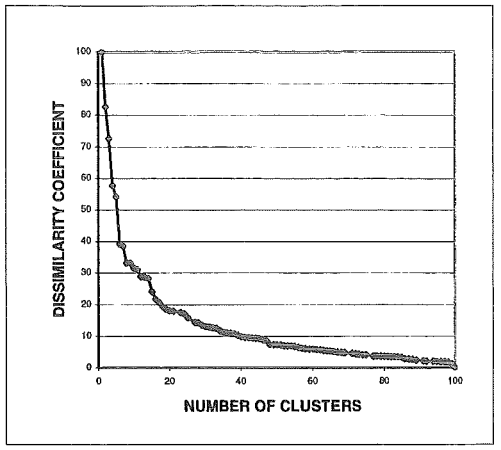


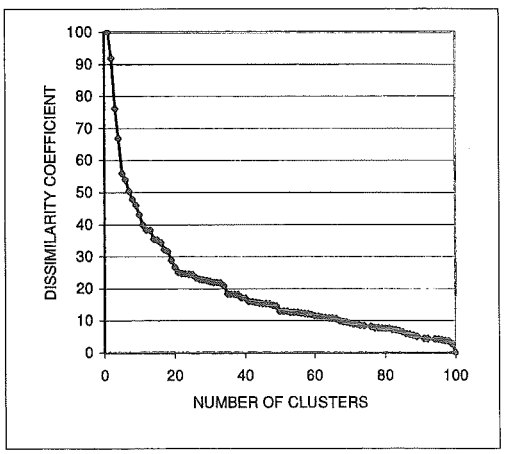
Figure7.08: Regular distribution (400 events in 1000^2 sq.units)



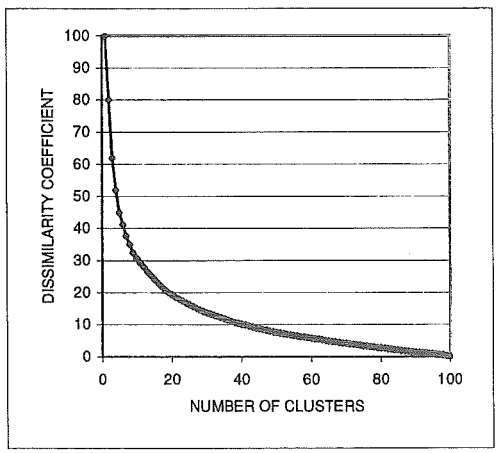
(a) Single-linkage technique



(b) Complete-linkage technique

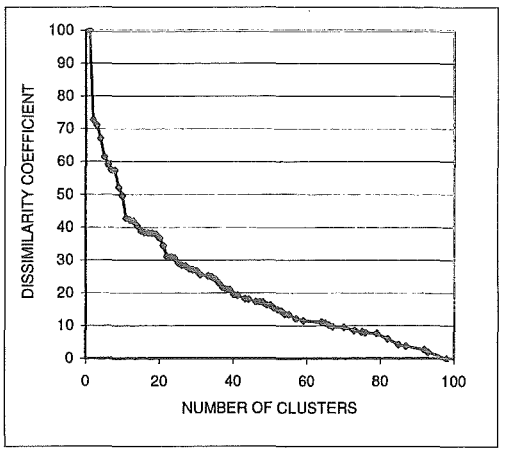


(c) Group average technique

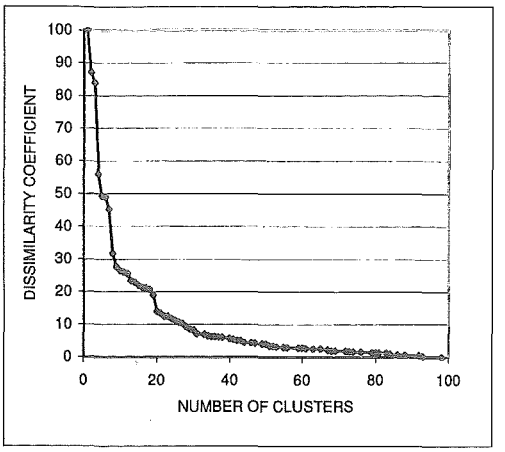


(d) Ward's technique

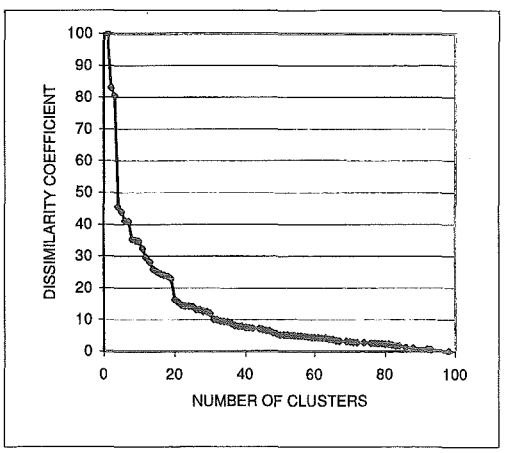
Figure 7.09: Variation of dissimilarity coefficient with number of clusters for completely spatially random distribution (Figure 7.01)



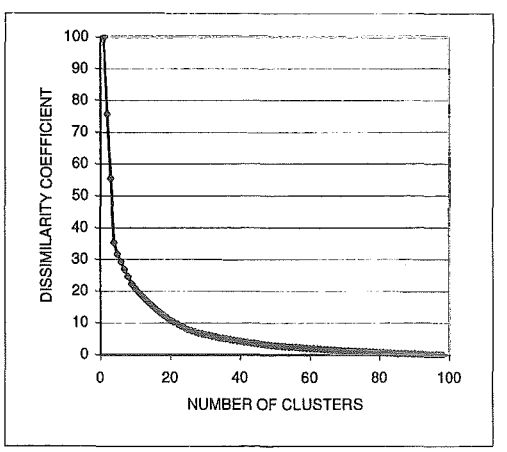
(a) Single-linkage technique



(b) Complete-linkage technique

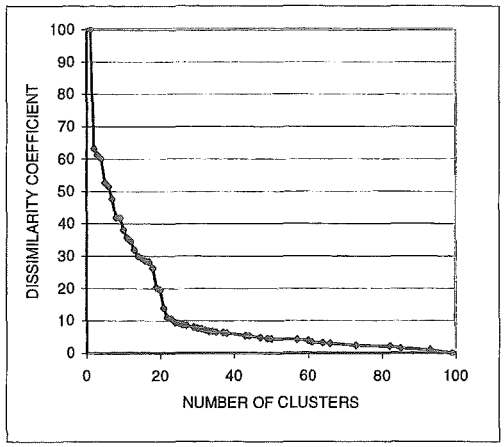


(c) Group average technique

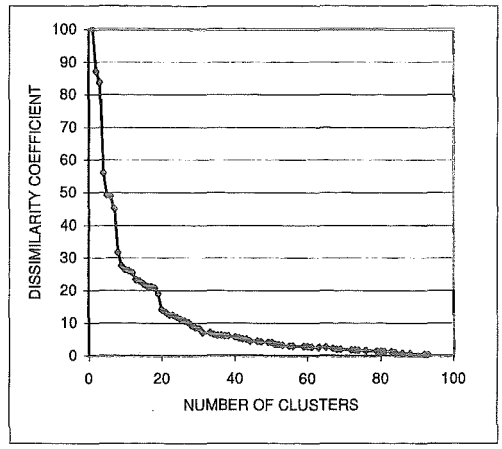


(d) Ward's technique

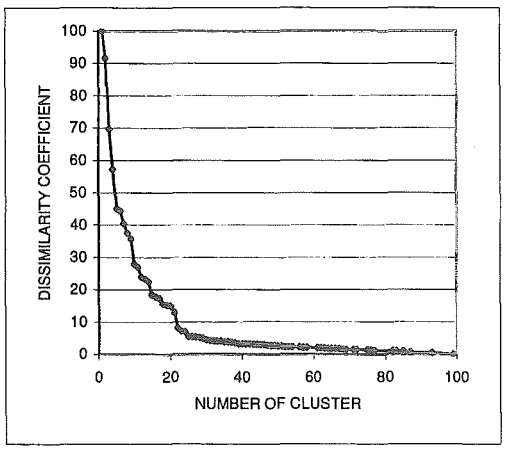
Figure 7.10: Variation of dissimilarity coefficient with number of clusters for line cluster (Figure 7.02)



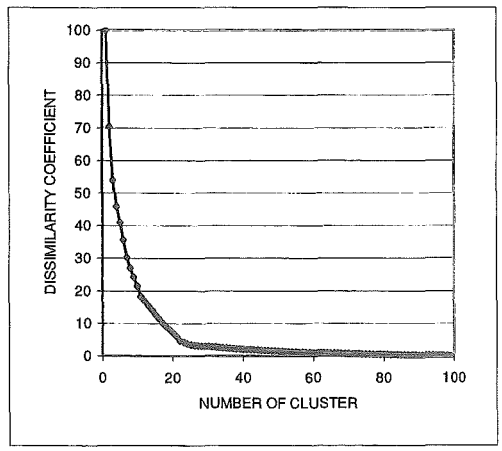
(a) Single-linkage technique



(b) Complete-linkage technique

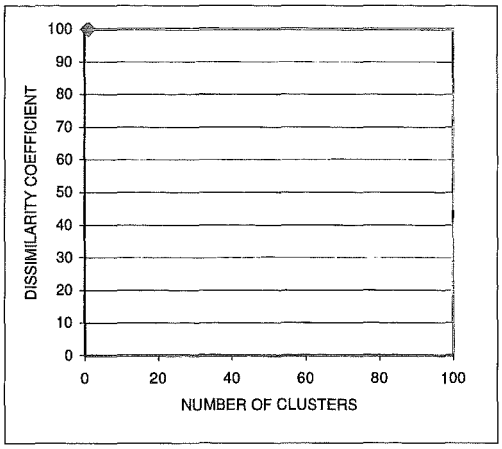


(c) Group average technique

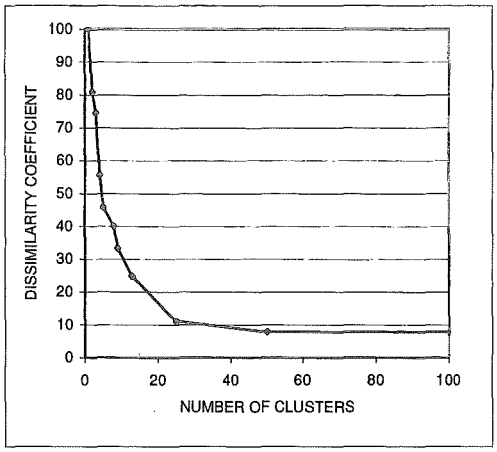


(d) Ward's technique

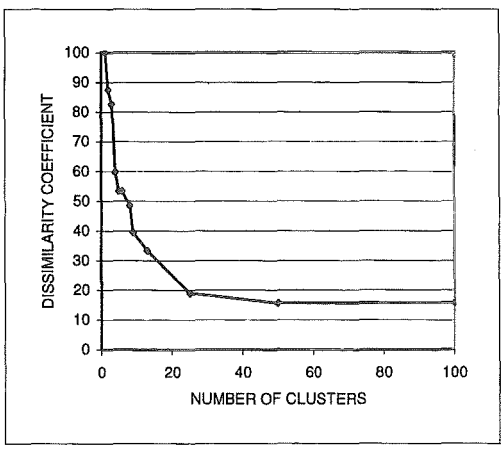
Figure 7.11: Variation of dissimilarity coefficient with number of clusters for point cluster distribution (Figure 7.03)



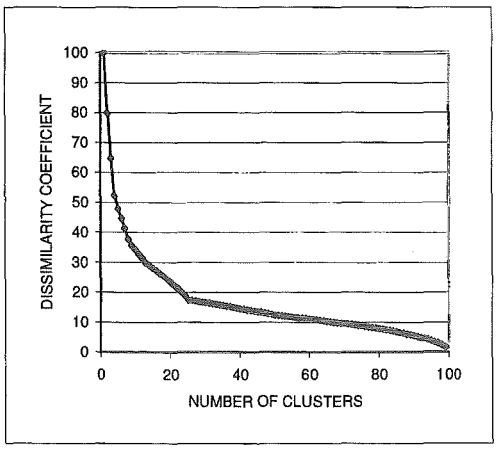
(a) Single-linkage technique



(b) Complete-linkage technique

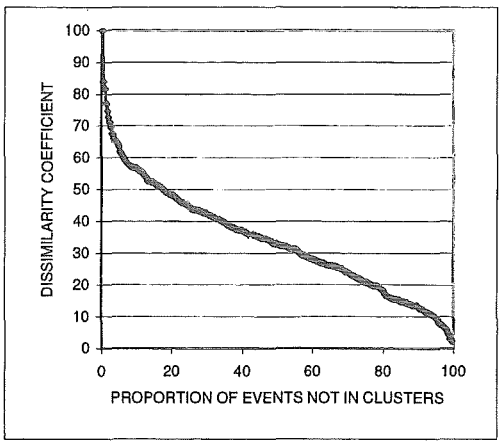


(c) Group average technique

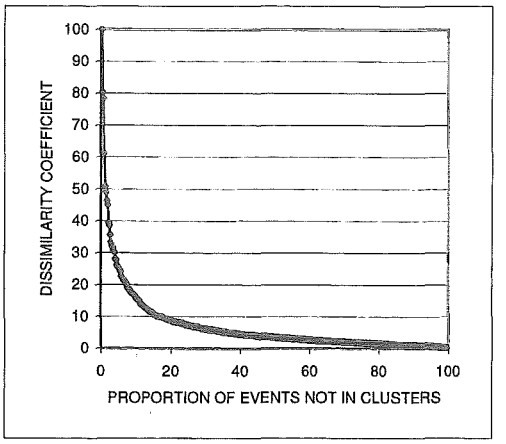


(d) Ward's technique

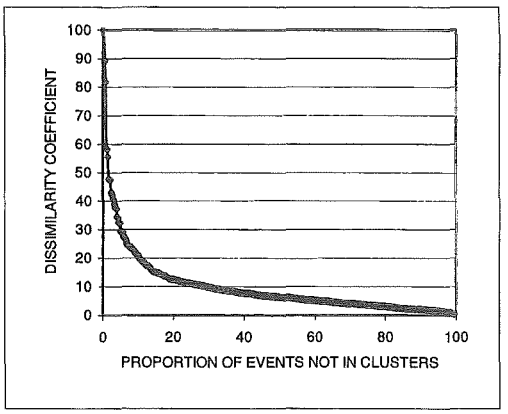
Figure 7.12: Variation of dissimilarity coefficient with number of clusters for regular distribution (Figure 7.04)



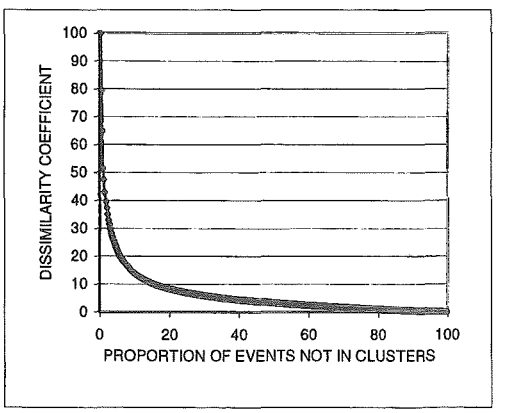
(a) Single-linkage technique



(b) Complete-linkage technique

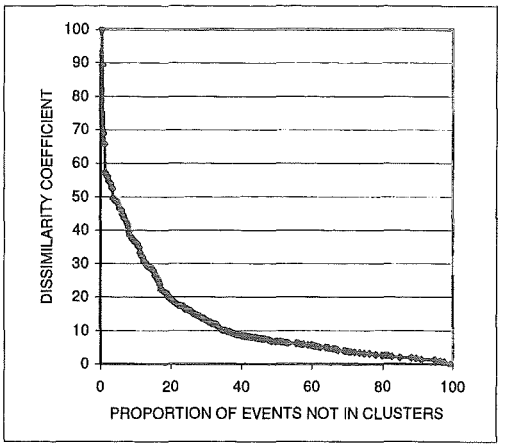


(c) Group average technique

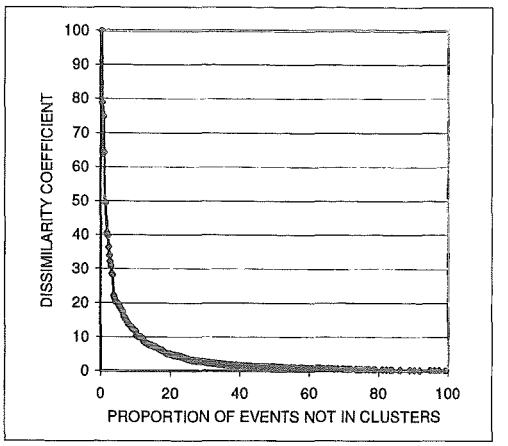


(d) Ward's technique

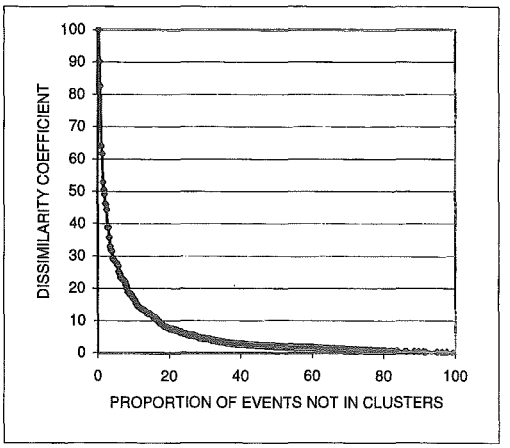
Figure 7.13: Variation of dissimilarity coefficient with number of clusters for completely spatially random distribution (Figure 7.05)



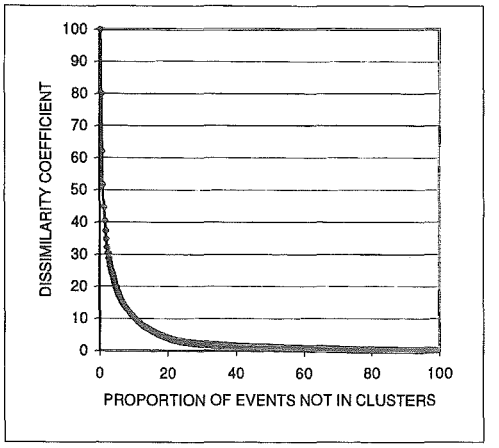
(a) Single-linkage technique



(b) Complete-linkage technique

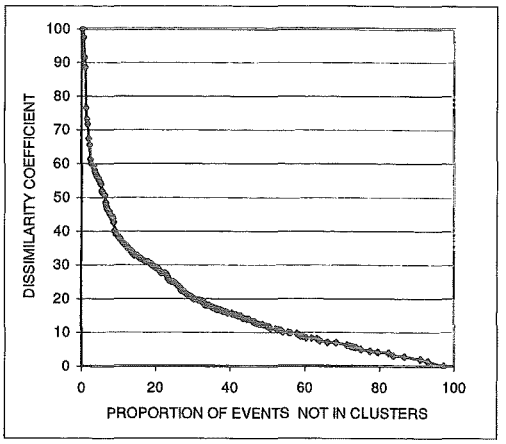


(c) Group average technique

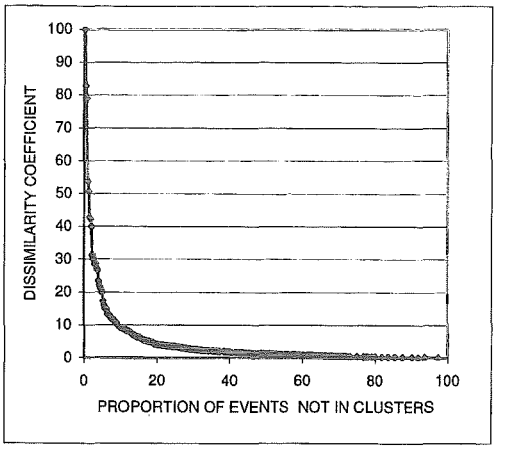


(d) Ward's technique

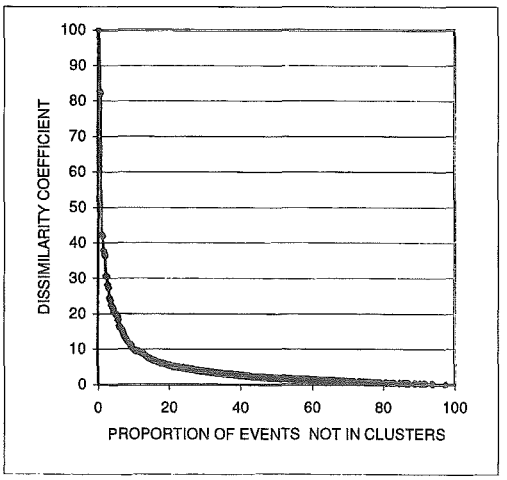
Figure 7.14: Variation of dissimilarity coefficient with number of clusters for point cluster distribution (Figure 7.06)



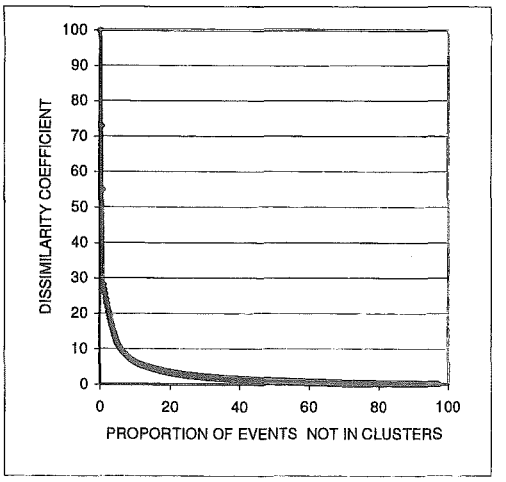
(a) Single-linkage technique



(b) Complete-linkage technique

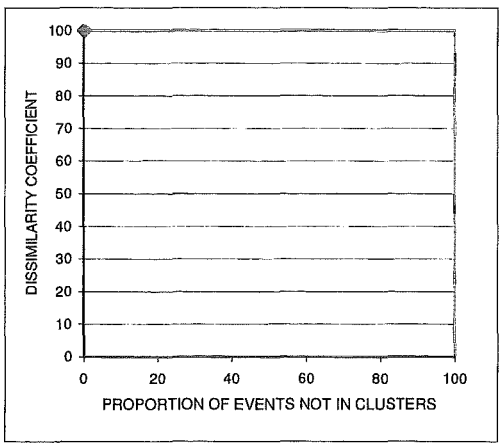


(c) Group average technique

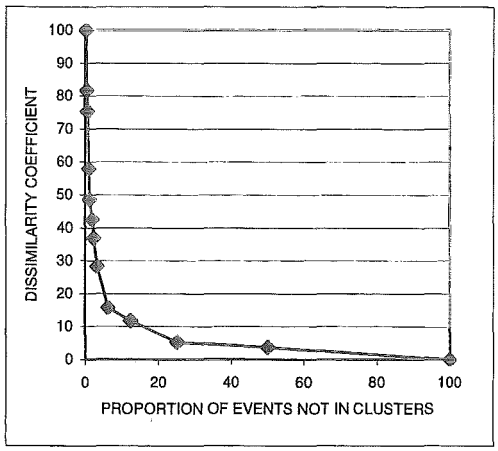


(d) Ward's technique

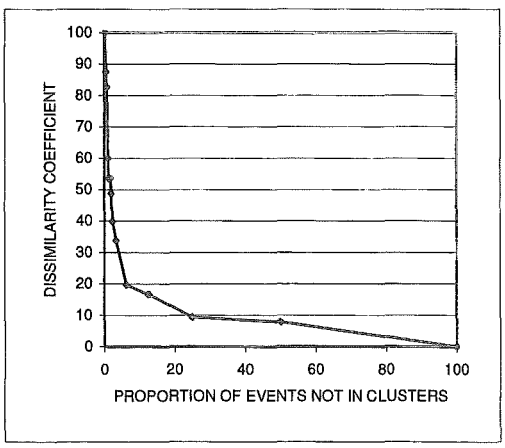
Fifgure 7.15: Variation of dissimilarity coefficient with number of clusters for line cluster distribution (Figure 7.07)



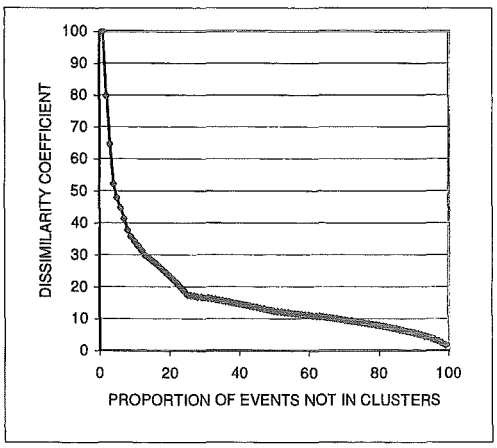
(a) Single-linkage technique



(b) Complete-linkage technique



(c) Group average technique



(d) Ward's technique

Figure 7.16: Variation of dissimilarity coefficient with number of clusters for regular distribution (Figure 7.08)

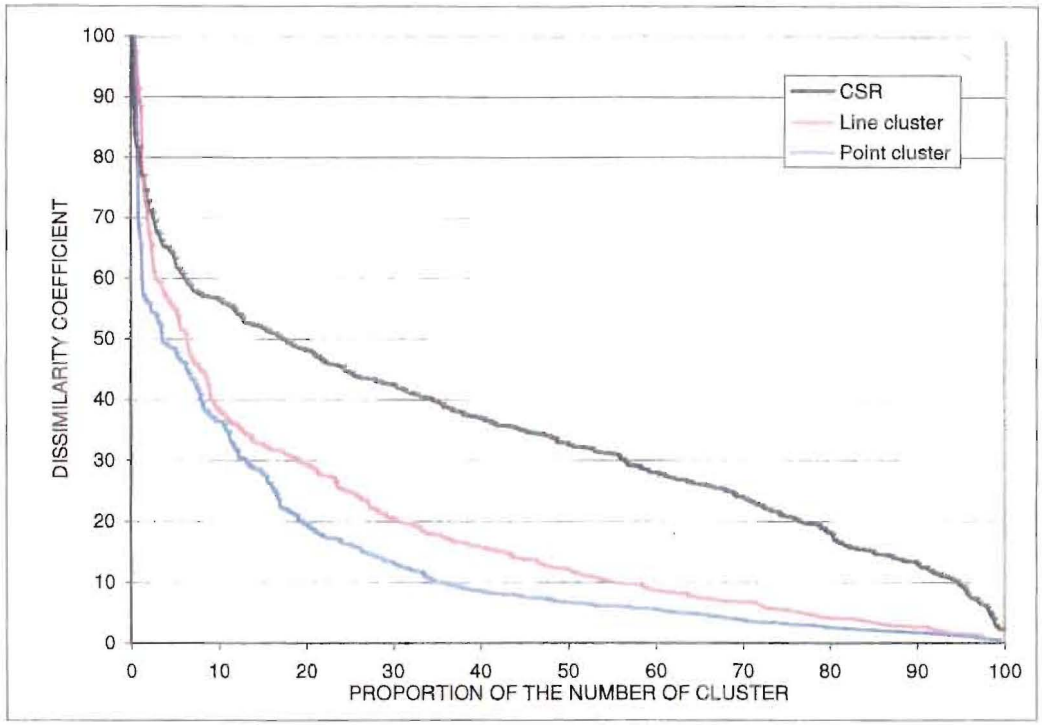


Figure 7.17a: Variation of dissimilarity coefficient variation with the number of clusters using single-linkage technique for three different distributions

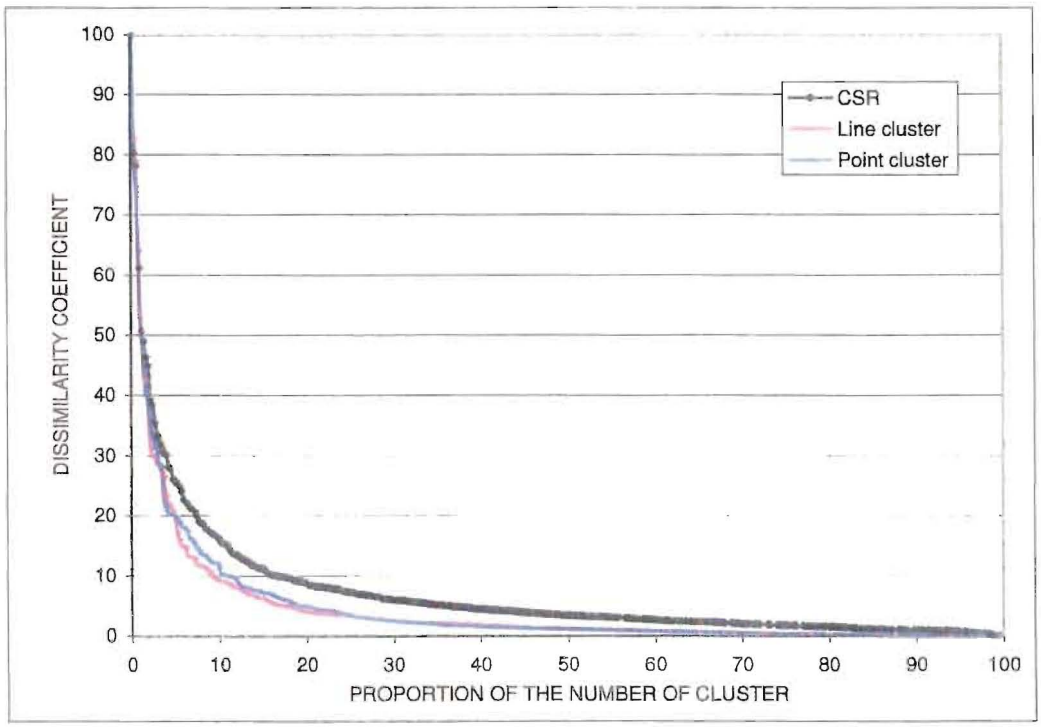


Figure 7.17b: Variation of dissimilarity coefficient variation with the number of clusters using complete-linkage technique for three different distributions

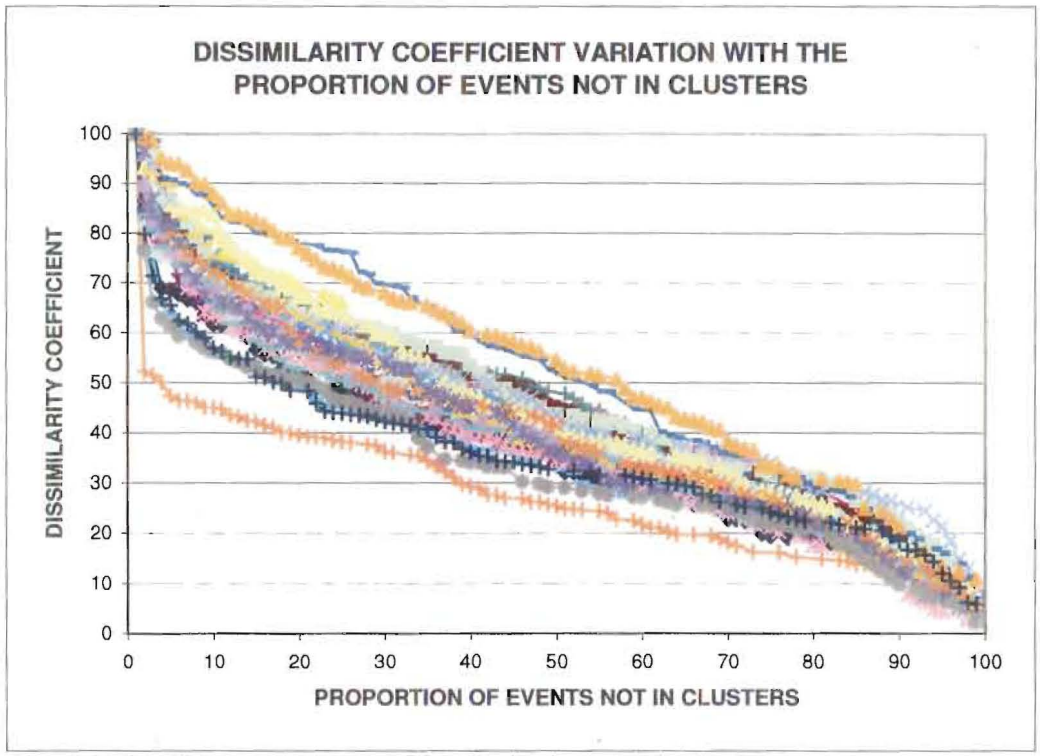


Figure 7.18a: The envelope for 25 CSR distributions obtained using single-linkage technique

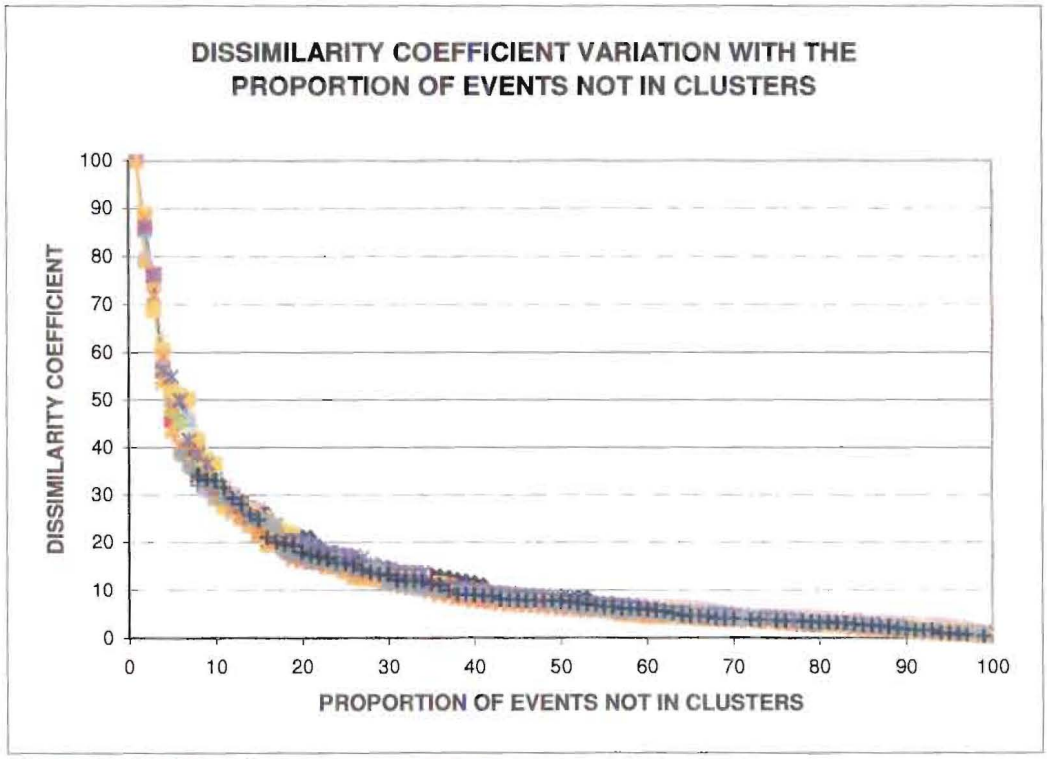


Figure 7.18b: The envelope for 25 CSR distributions obtained using complete-linkage technique

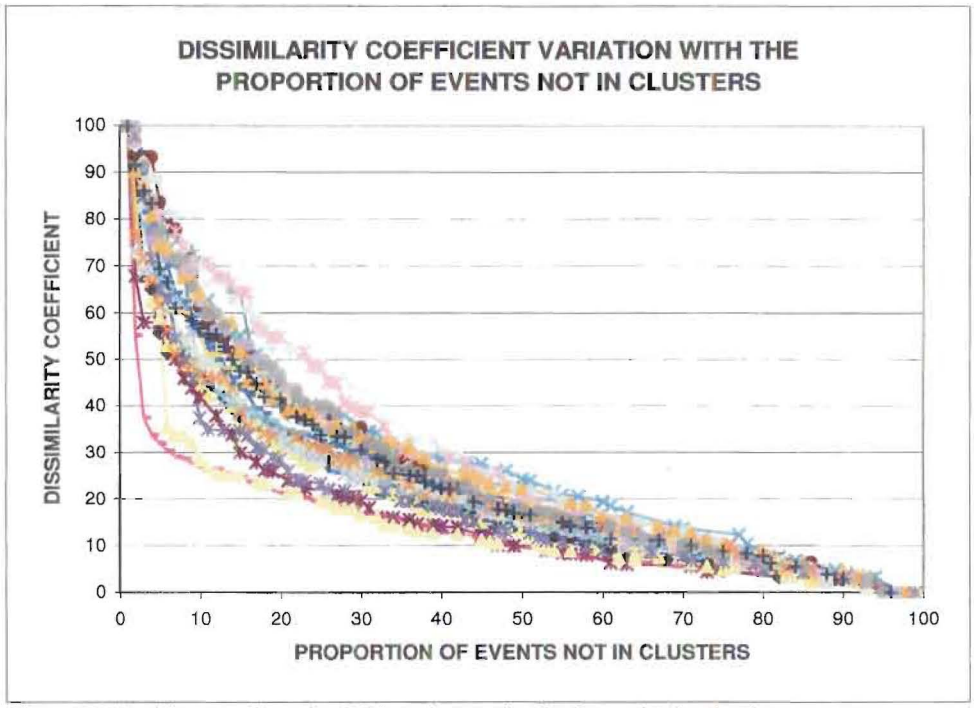


Figure 7.19a: The envelope for 25 line cluster distributions obtained using single-linkage technique

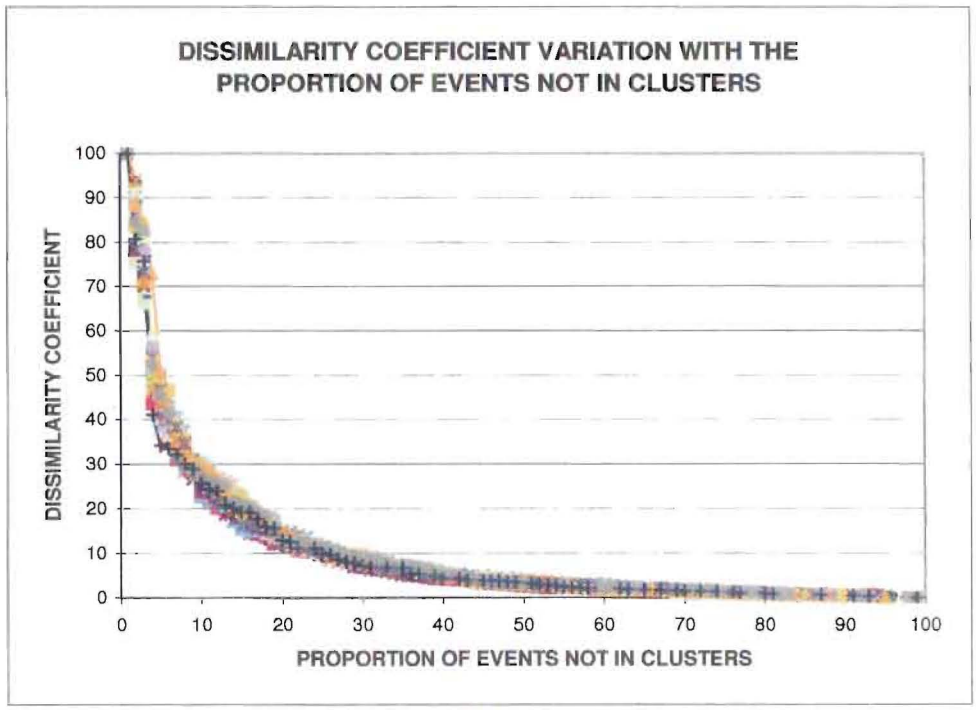


Figure 7.19b: The envelope for 25 line cluster distributions obtained using complete-linkage technique.

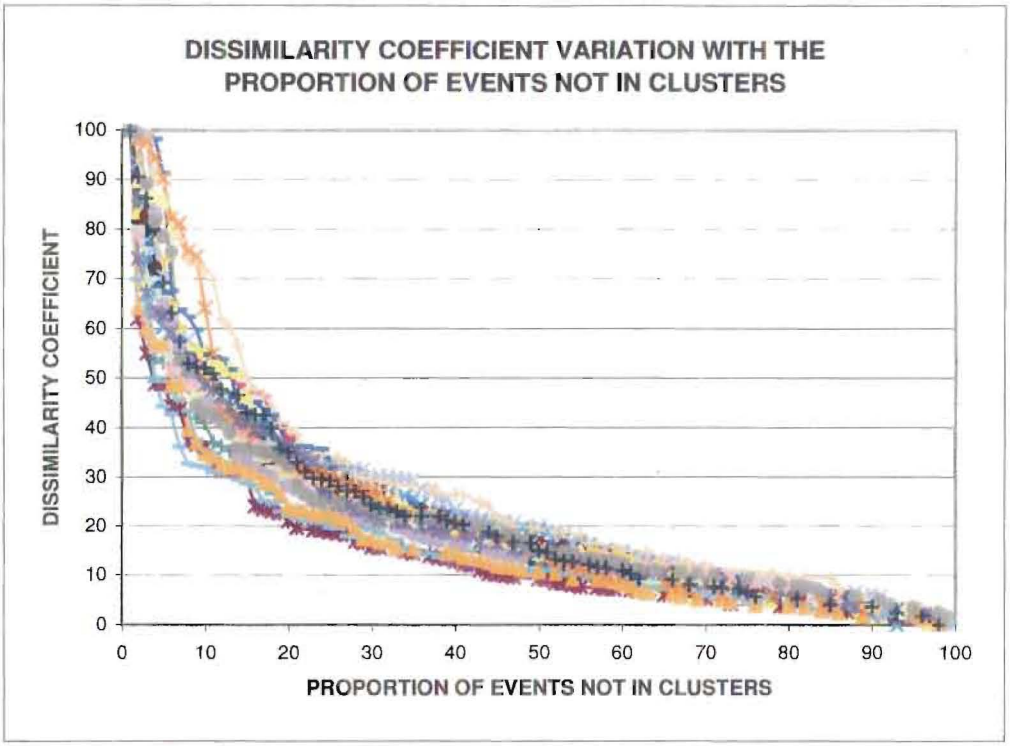


Figure 7.20a: The envelope for 25 point cluster distributions obtained using single-linkage technique.

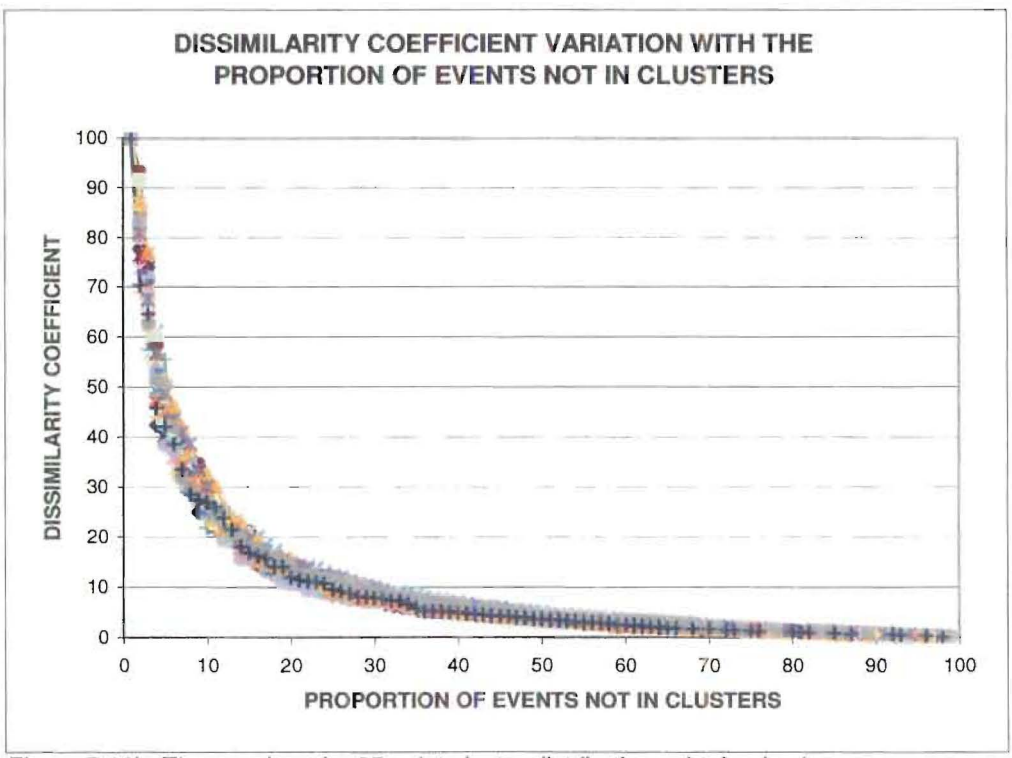


Figure 7.20b: The envelope for 25 point cluster distributions obtained using complete-linkage technique.

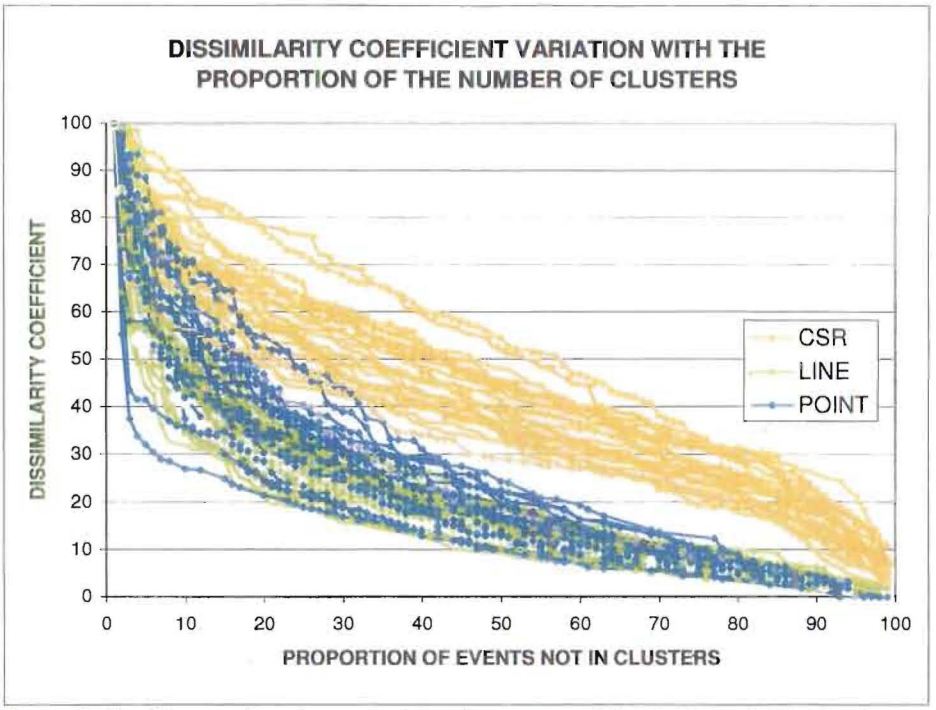


Figure 7.21a: The envelope for each of the three types of distributions obtained using single-linkage technique

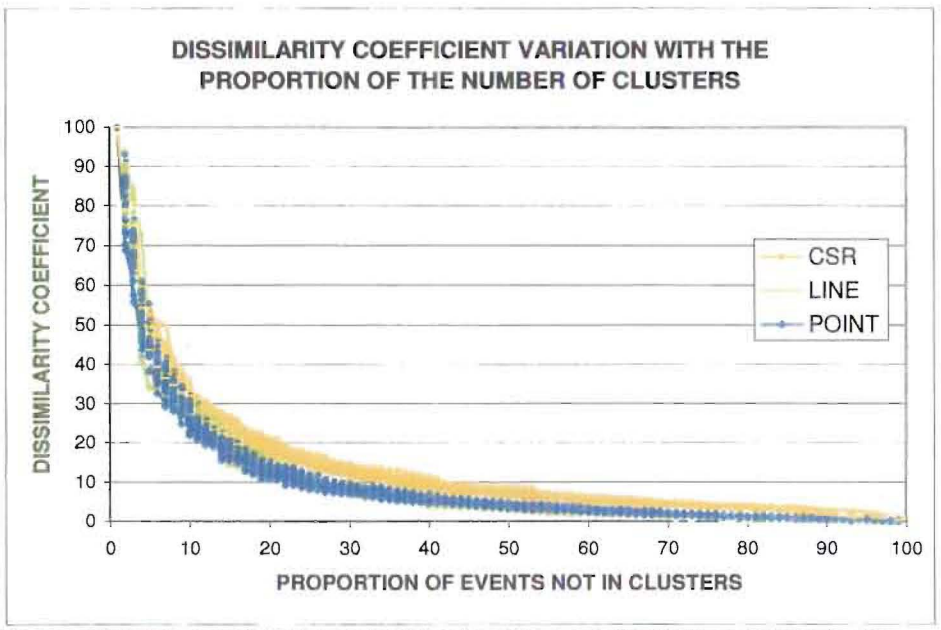
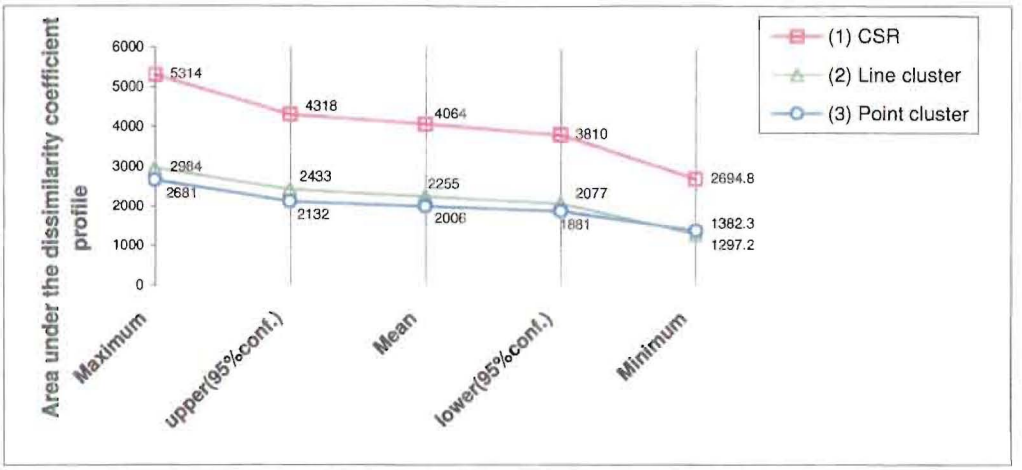
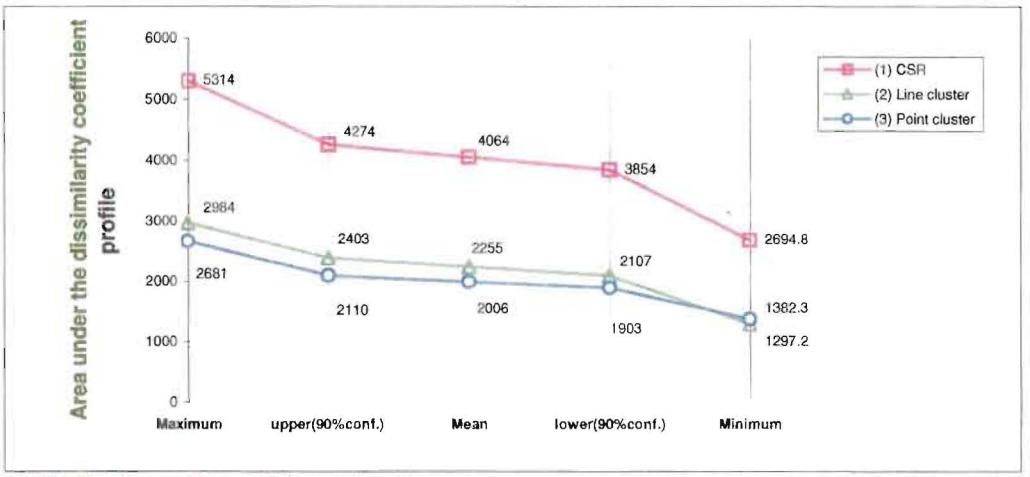


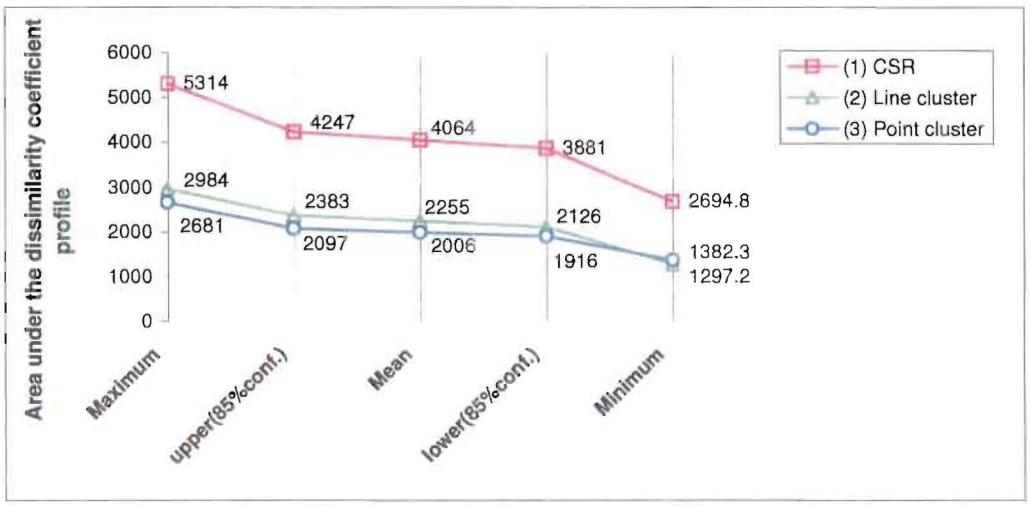
Figure 7.21b: The envelope for each of the three types of distributions obtained using complete-linkage technique



(a) Indicating 95% confidence level



(b) Indicating 90% confidence level



(c) Indicating 85% confidence level

Figure 7.22: The confidence band for the area under the dissimilarity coefficient profile for 25 examples of basic distribution types

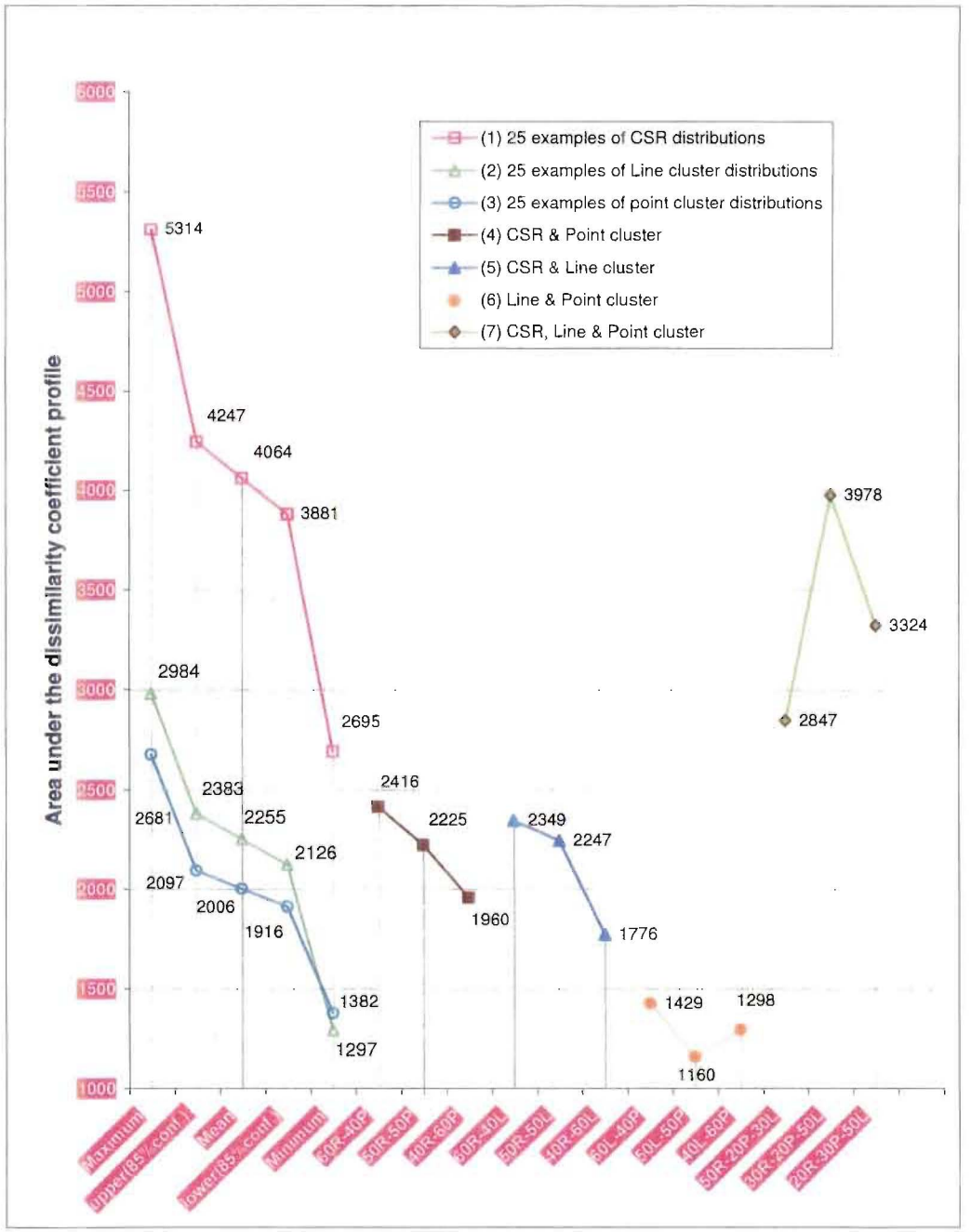


Figure 7.23: Computed area under the dissimilarity coefficient profile using single-linkage method for 25 examples of basic distribution types and one example of each mixture distribution

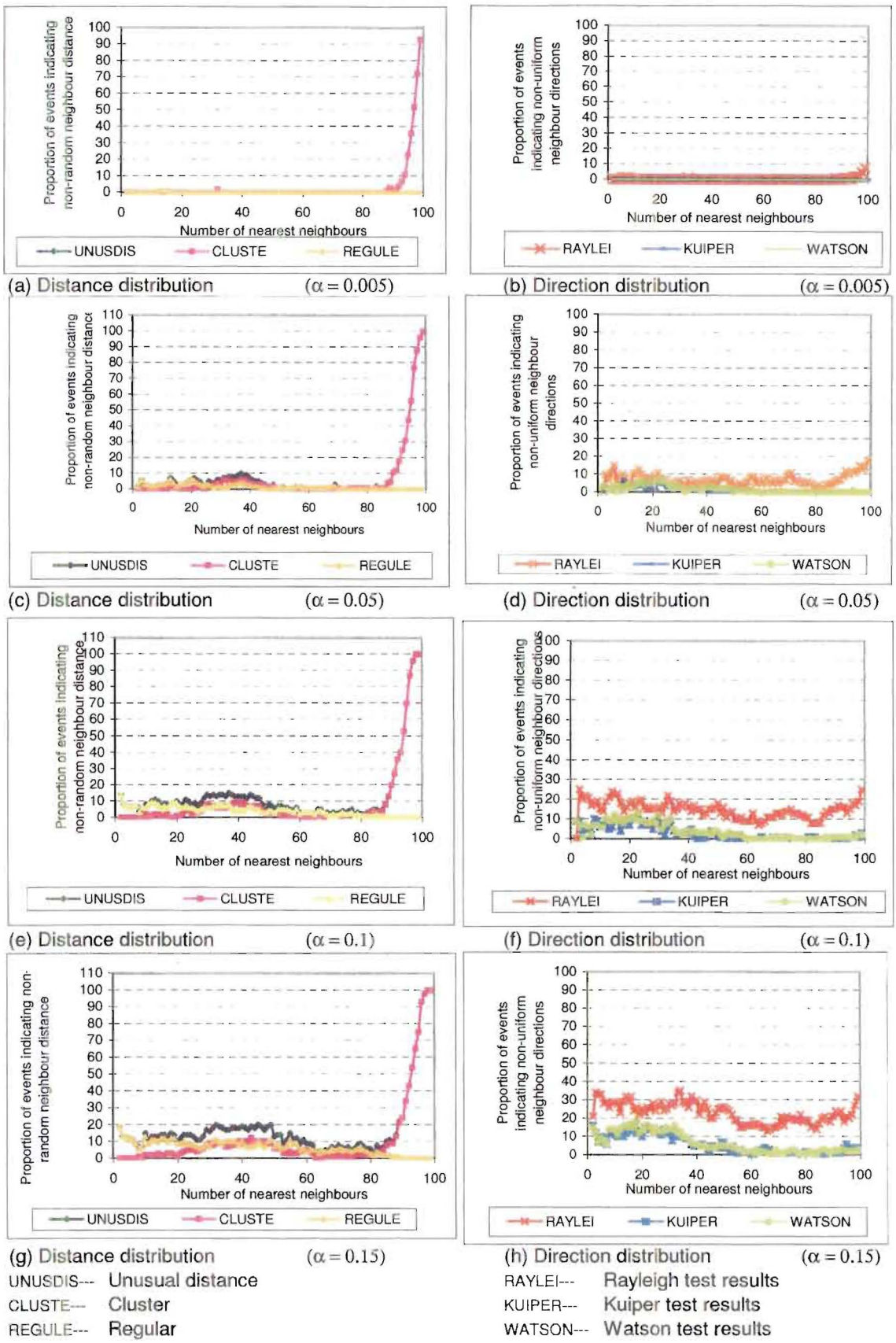


Figure 7.24: Nearest-neighbour analysis results for completely spatially random distribution (Fig. 7.01).

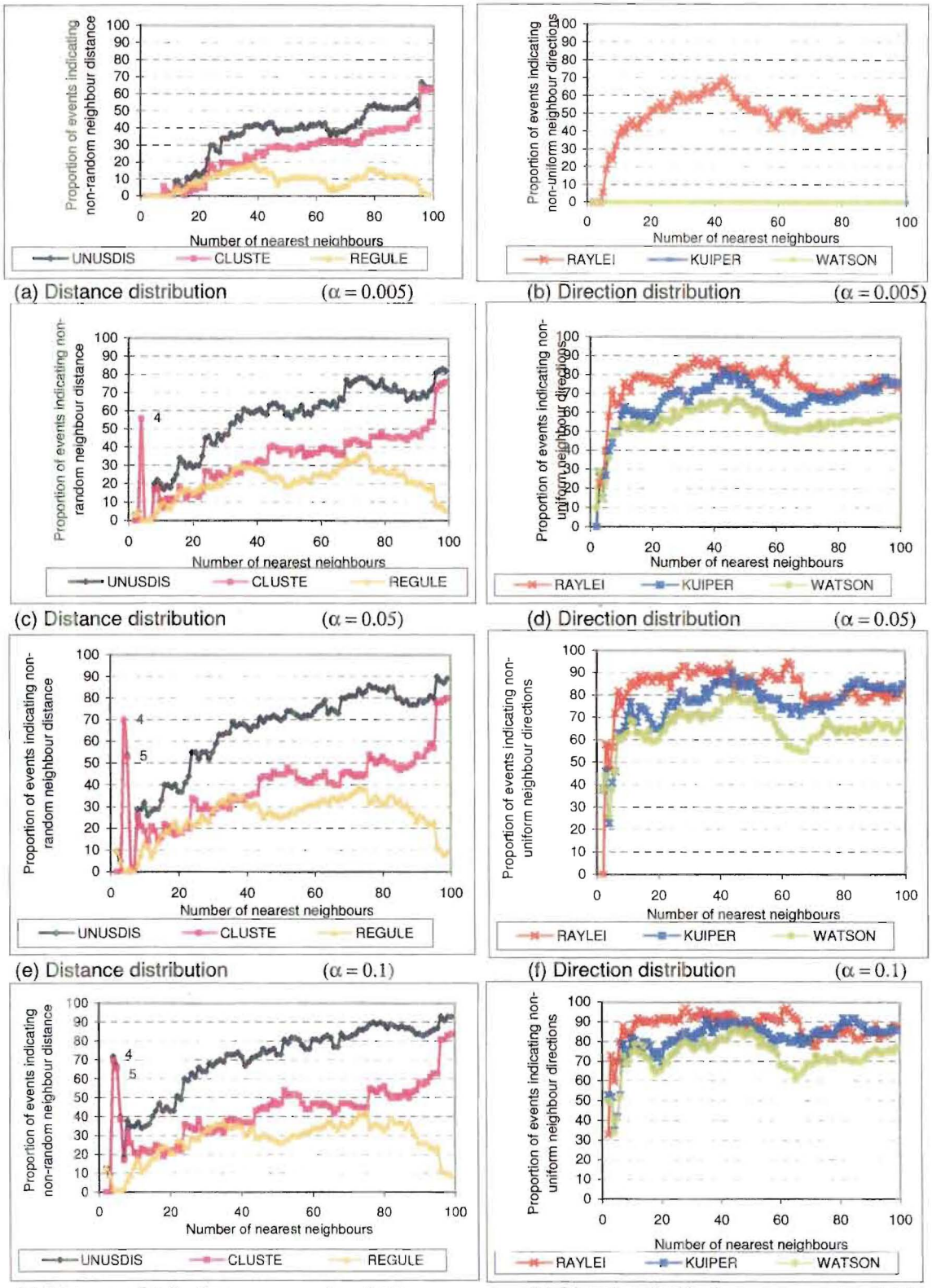
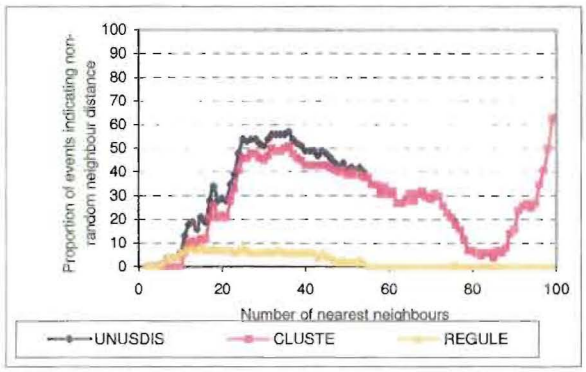
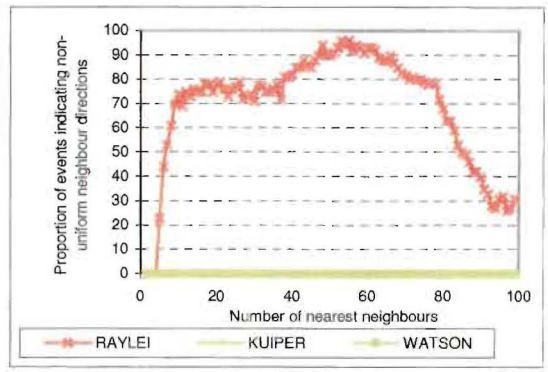


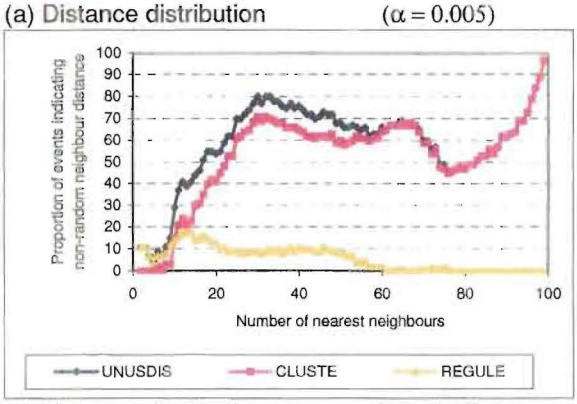
Figure 7.25: Nearest-neighbour analysis results for point cluster distribution (Fig. 7.03).



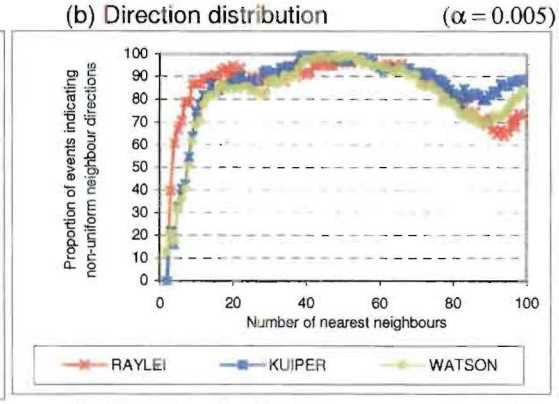
(a) Distance distribution ($\alpha = 0.005$)



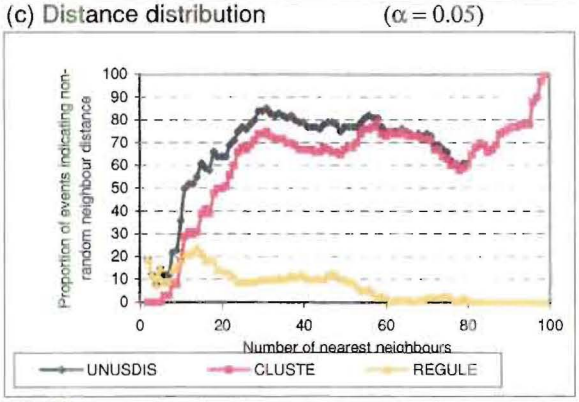
(b) Direction distribution ($\alpha = 0.005$)



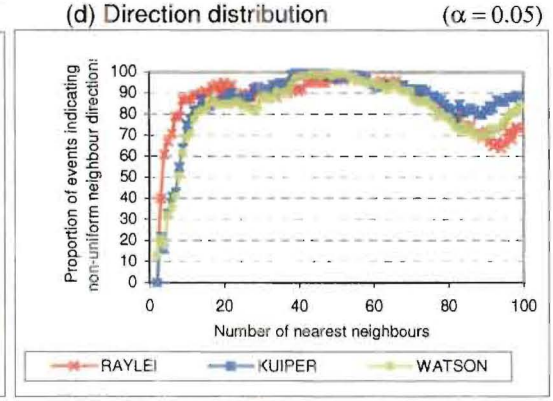
(c) Distance distribution ($\alpha = 0.05$)



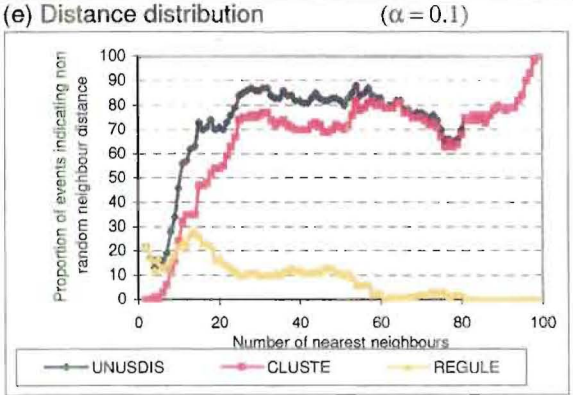
(d) Direction distribution ($\alpha = 0.05$)



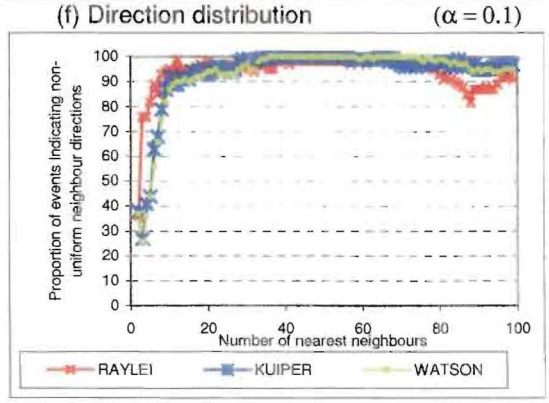
(e) Distance distribution ($\alpha = 0.1$)



(f) Direction distribution ($\alpha = 0.1$)



(g) Distance distribution ($\alpha = 0.15$)



(h) Direction distribution ($\alpha = 0.15$)

UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure 7.26: Nearest-neightbour analysis results for line cluster distribution (Fig. 7.02).

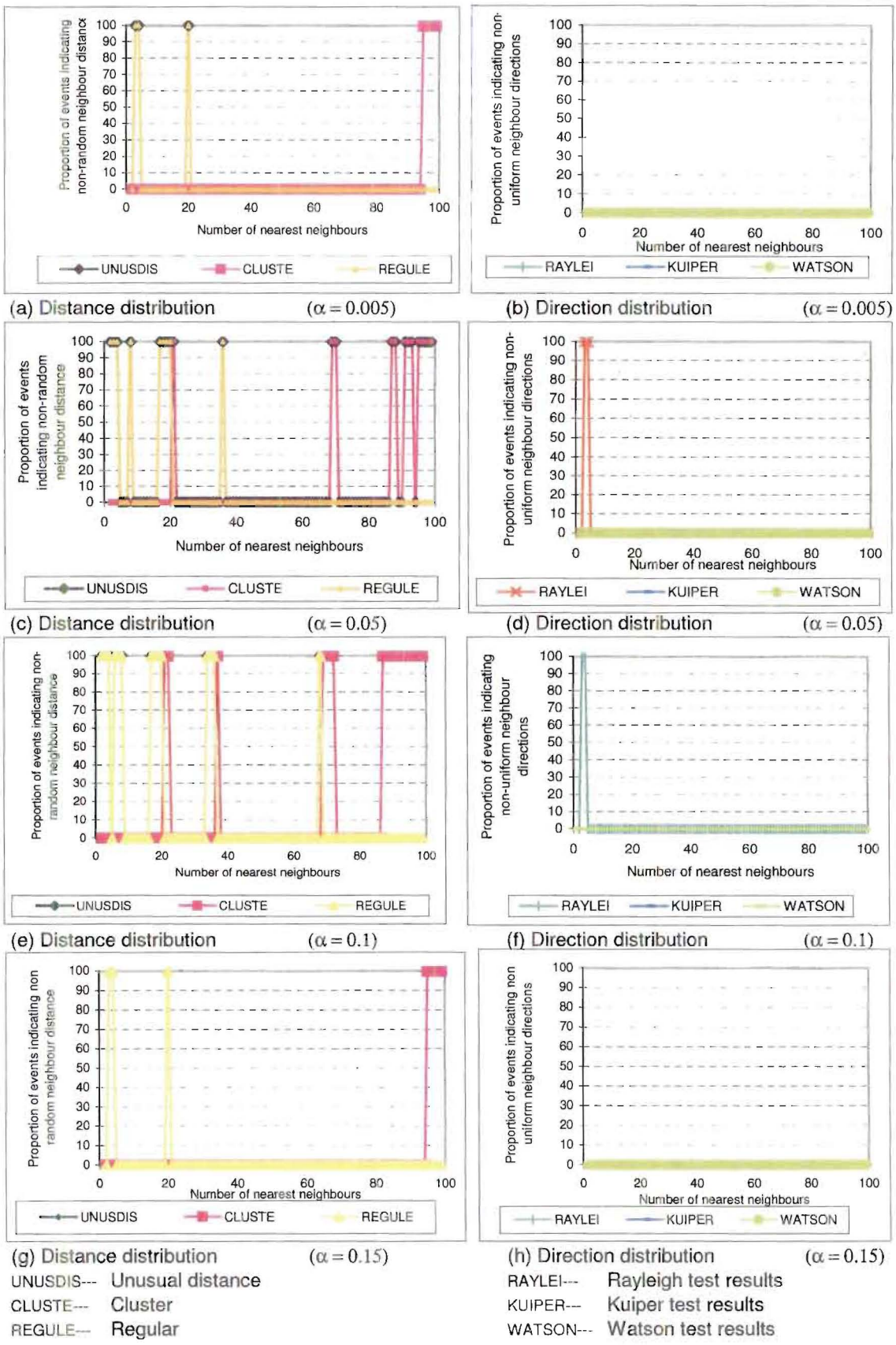
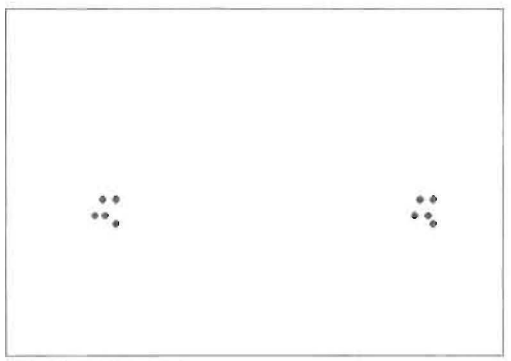
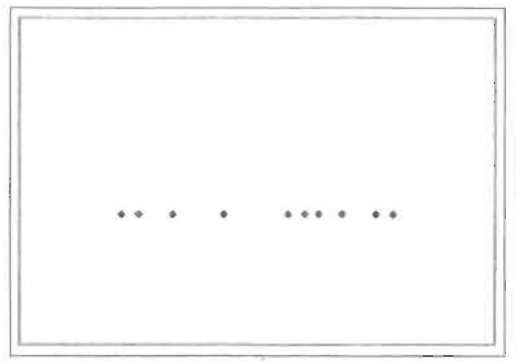


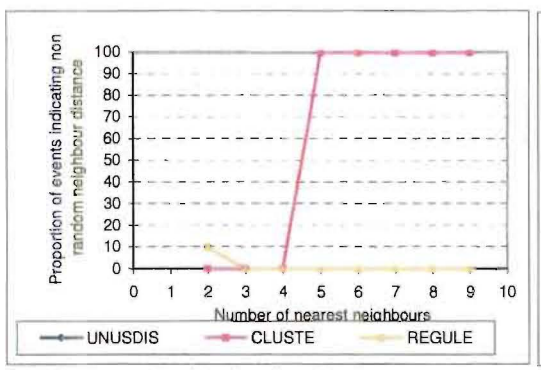
Figure 7.27: Nearest-neighbour analysis results for regular distribution (Fig. 7.04).



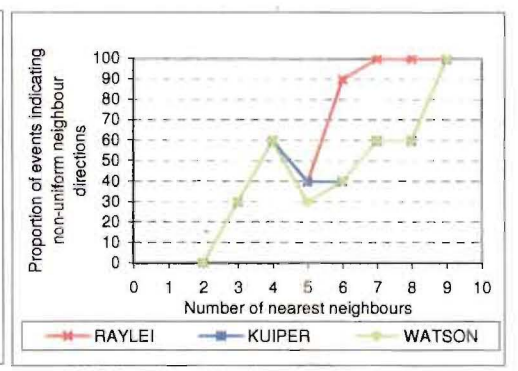
(a) Point clusters



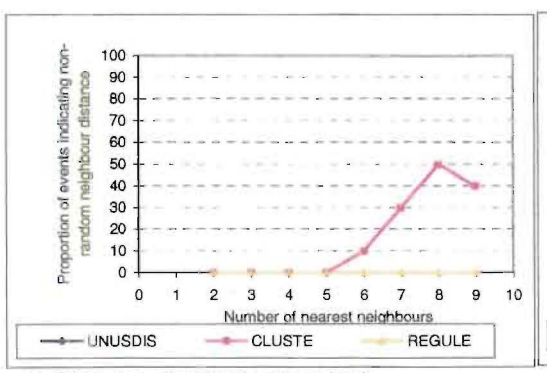
(b) A line cluster



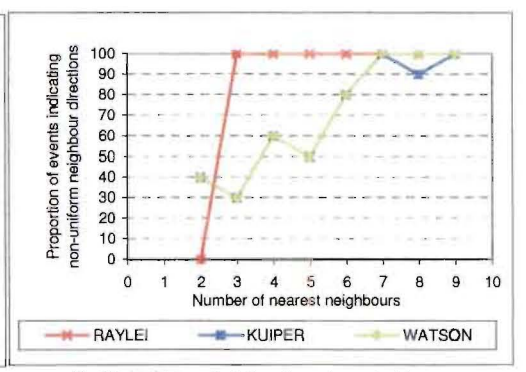
(c) Distance distribution for plot a



(d) Direction distribution for plot a



(e) Distance distribution for plot b



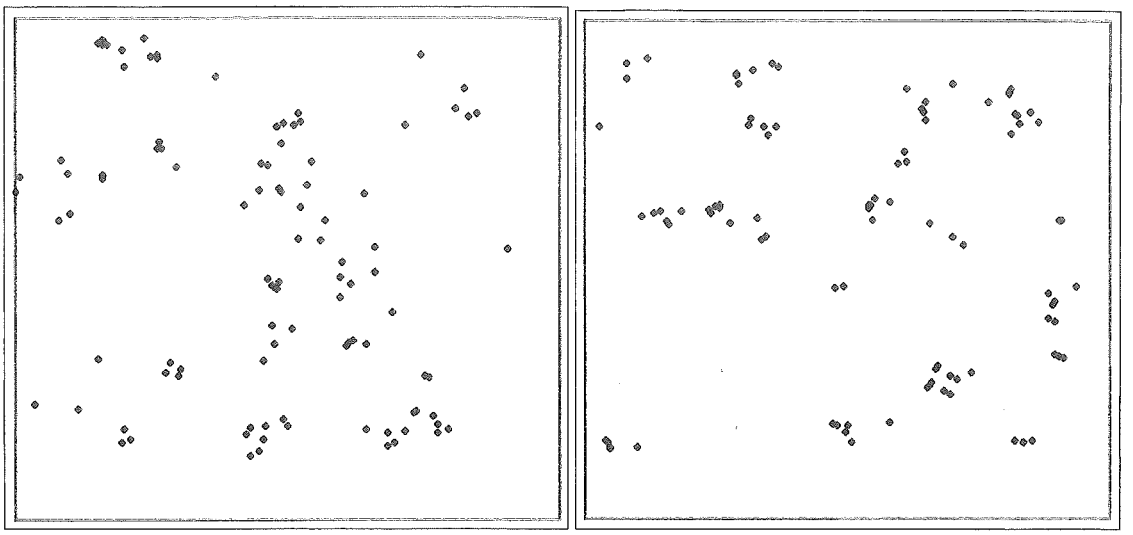
(f) Direction distribution for plot b

UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

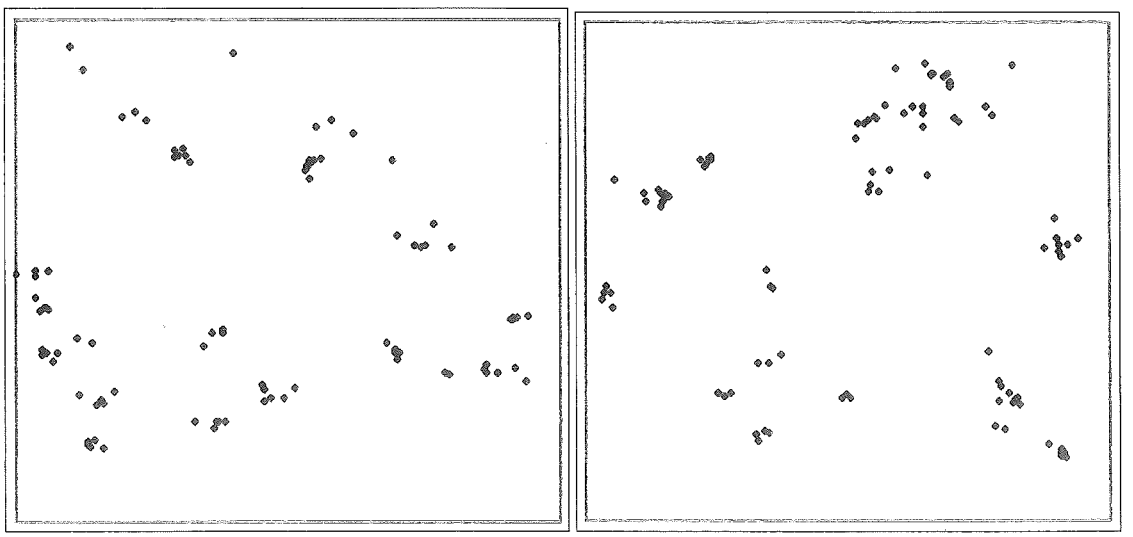
Significance level 0.1 used for distance and direction analysis

Figure 7.28: Nearest-neighbour analysis results for the plots a and b.



(a) Distribution - 1

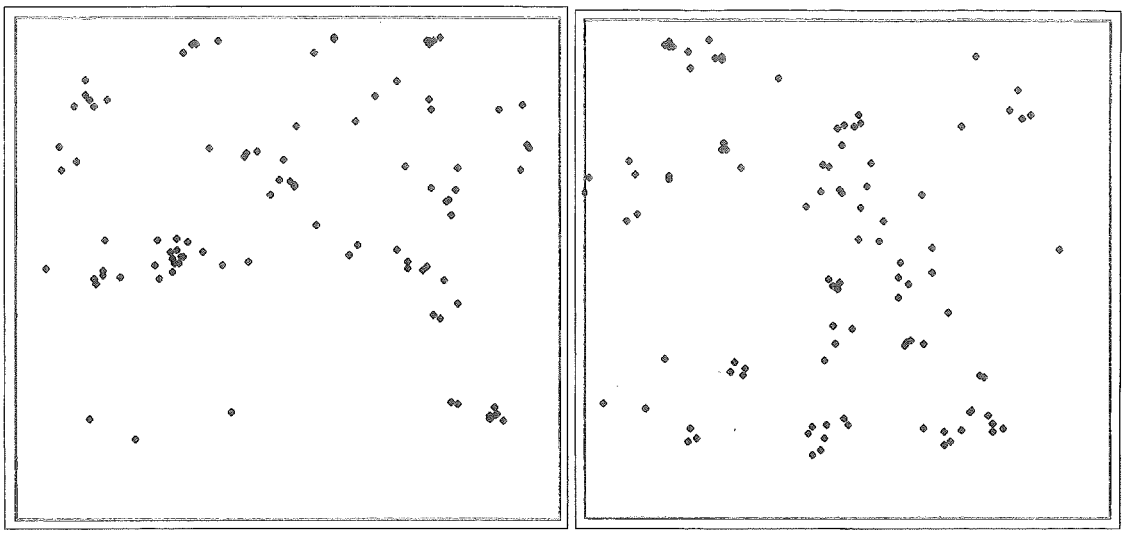
(b) Distribution - 2



(c) Distribution - 3

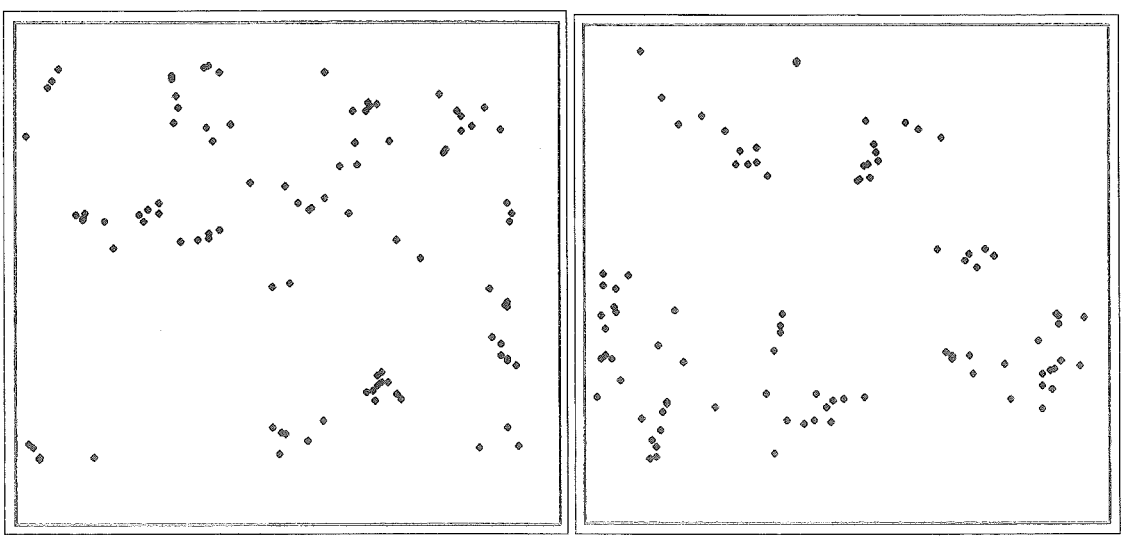
(d) Distribution - 4

Figure 7.29: Location plot for four point cluster (dense events in cluster) distributions



(a) Distribution - 1

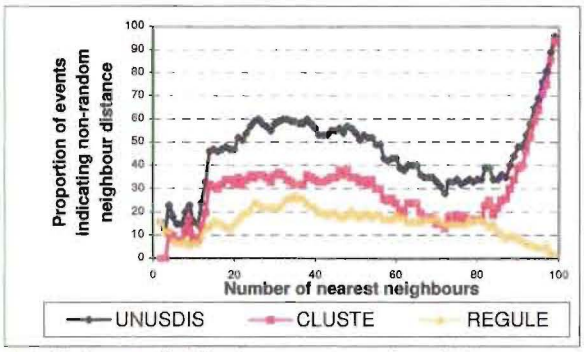
(b) Distribution - 2



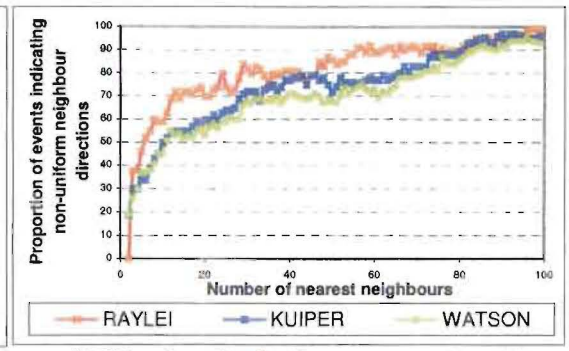
(c) Distribution - 3

(d) Distribution - 4

Figure 7.30: Location plot for four point cluster (sparse events in cluster) distributions

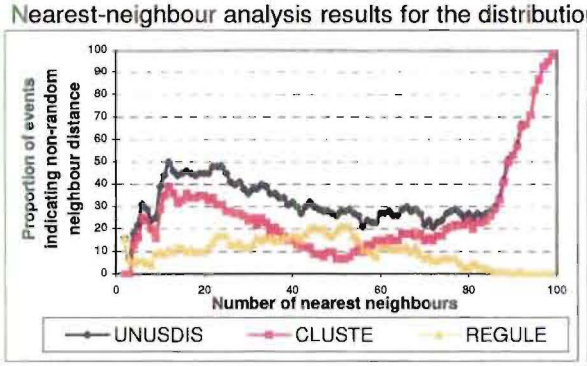


(a) Distance distribution ($\alpha = 0.1$)

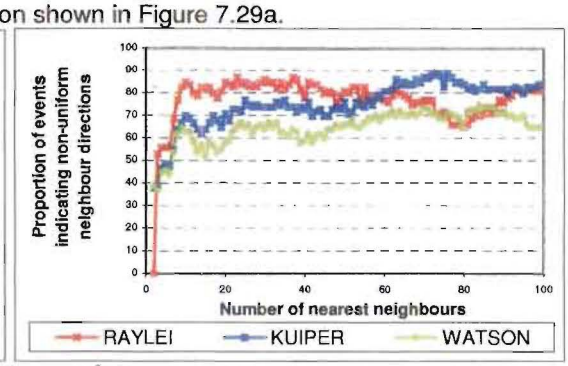


(b) Direction distribution ($\alpha = 0.1$)

Nearest-neighbour analysis results for the distribution shown in Figure 7.29a.

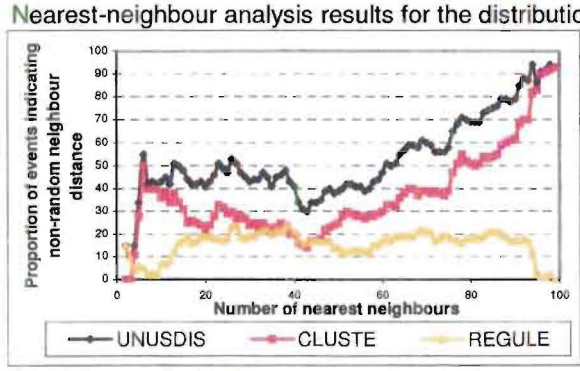


(c) Distance distribution ($\alpha = 0.1$)

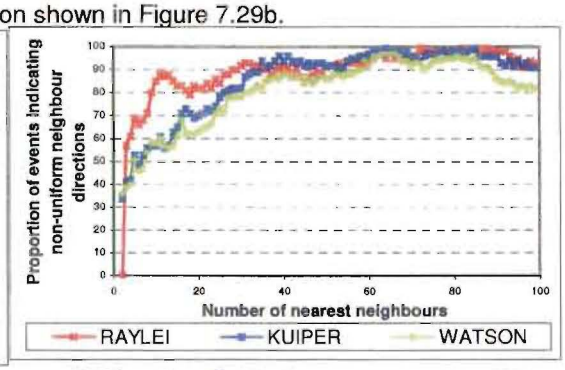


(d) Direction distribution ($\alpha = 0.1$)

Nearest-neighbour analysis results for the distribution shown in Figure 7.29b.

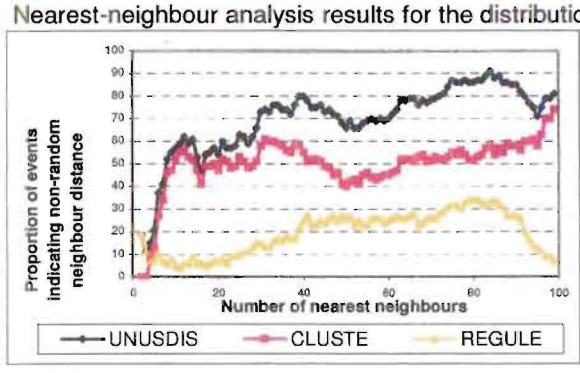


(e) Distance distribution ($\alpha = 0.1$)

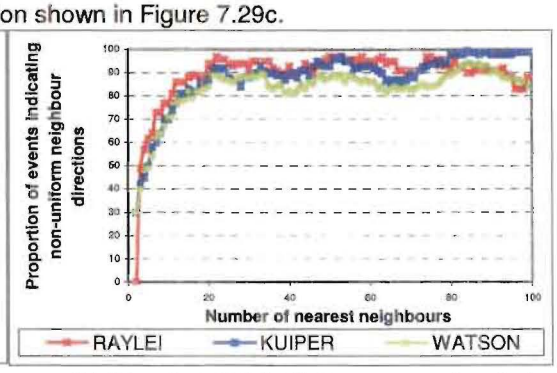


(f) Direction distribution ($\alpha = 0.1$)

Nearest-neighbour analysis results for the distribution shown in Figure 7.29c.



(g) Distance distribution ($\alpha = 0.1$)

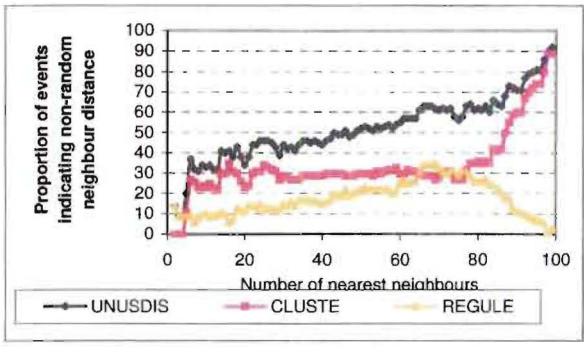


(h) Direction distribution ($\alpha = 0.1$)

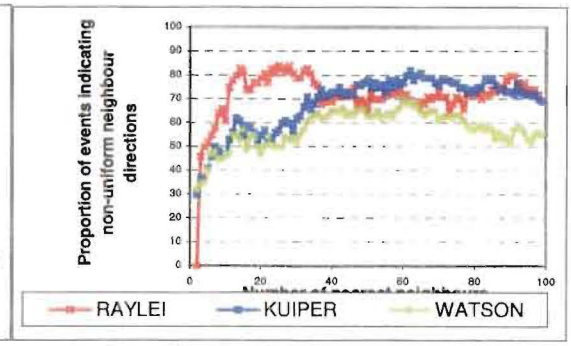
Nearest-neighbour analysis results for the distribution shown in Figure 7.29d.

UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular
RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure 7.31: Four point cluster (dense locations) distributions

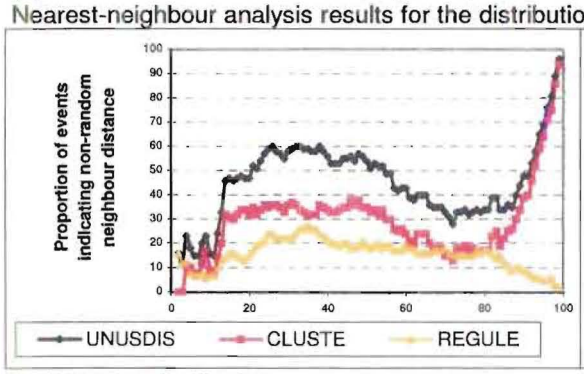


(a) Distance distribution ($\alpha = 0.1$)

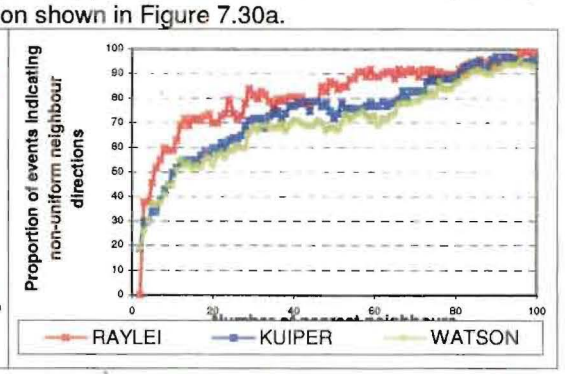


(b) Direction distribution ($\alpha = 0.1$)

Nearest-neighbour analysis results for the distribution shown in Figure 7.30a.

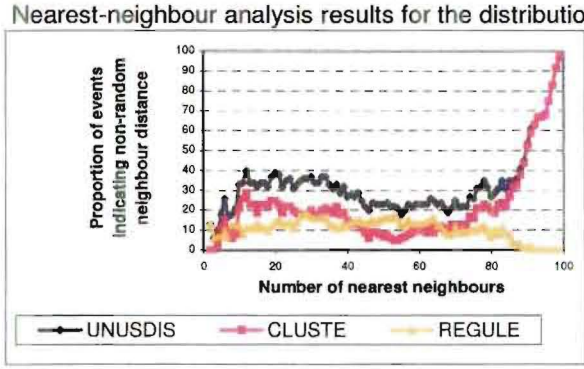


(c) Distance distribution ($\alpha = 0.1$)

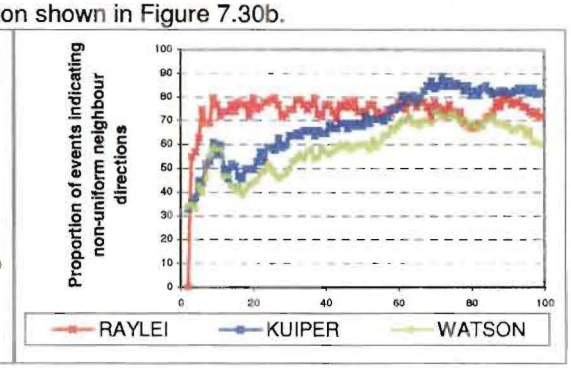


(d) Direction distribution ($\alpha = 0.1$)

Nearest-neighbour analysis results for the distribution shown in Figure 7.30b.

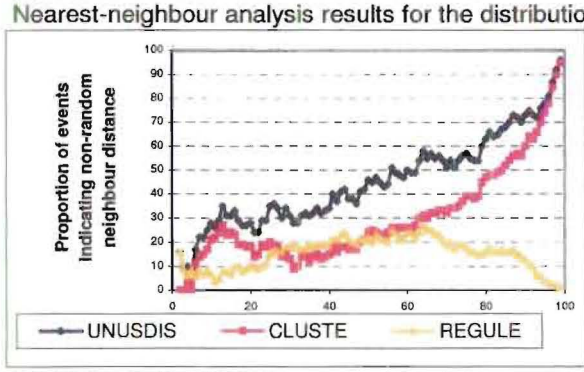


(e) Distance distribution ($\alpha = 0.1$)

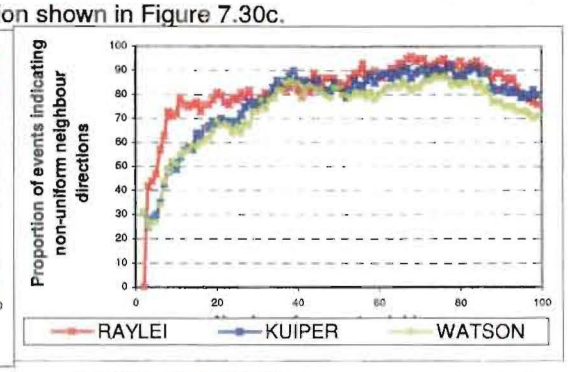


(f) Direction distribution ($\alpha = 0.1$)

Nearest-neighbour analysis results for the distribution shown in Figure 7.30c.



(g) Distance distribution ($\alpha = 0.1$)

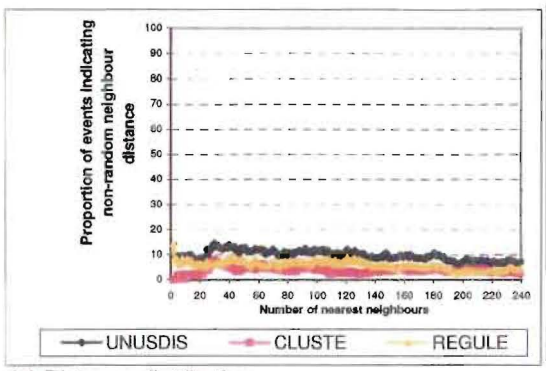


(h) Direction distribution ($\alpha = 0.1$)

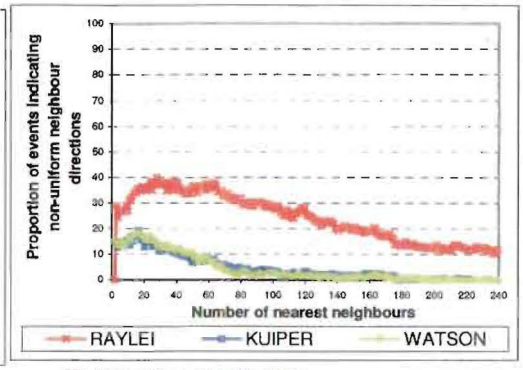
Nearest-neighbour analysis results for the distribution shown in Figure 7.30d.

UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular
RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure 7.32: Four point cluster (sparse locations) distributions

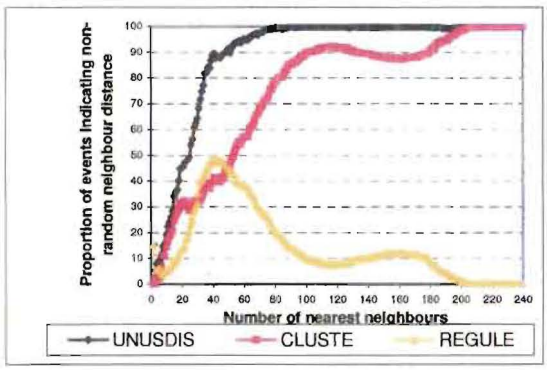


(a) Distance distribution

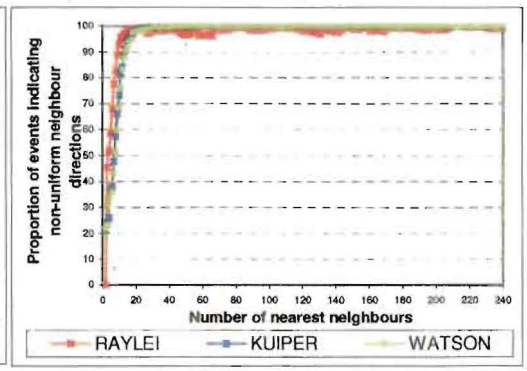


(b) Direction distribution

Nearest-neighbour analysis results for completely spatially random distribution (Fig. 7.05).

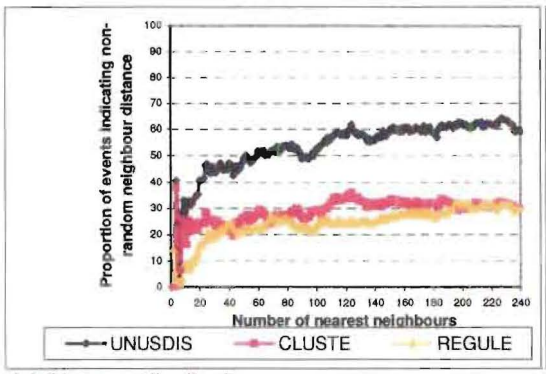


(c) Distance distribution

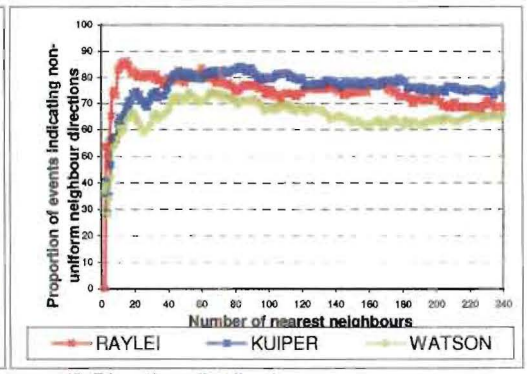


(d) Direction distribution

Nearest-neightbour analysis results for line cluster distribution (Fig. 7.07).



(e) Distance distribution

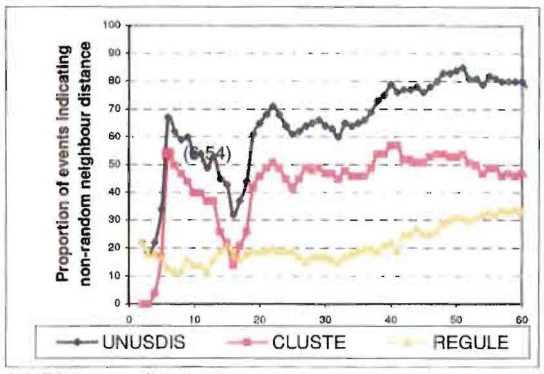


(f) Direction distribution

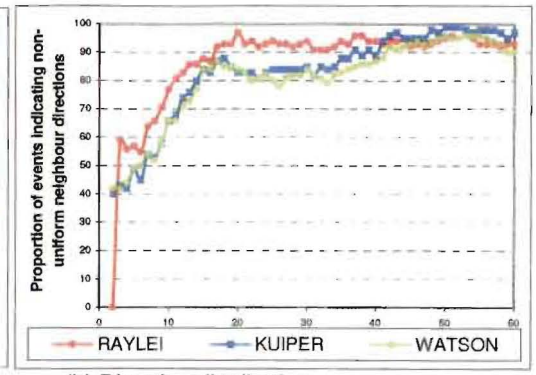
Nearest-neighbour analysis results for point cluster distribution (Fig. 7.06).

UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular
RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure 7.33: Nearest-neighbour analysis results for CSR, Line and Point cluster distributions (Figures 7.05, 7.06 and 7.07)

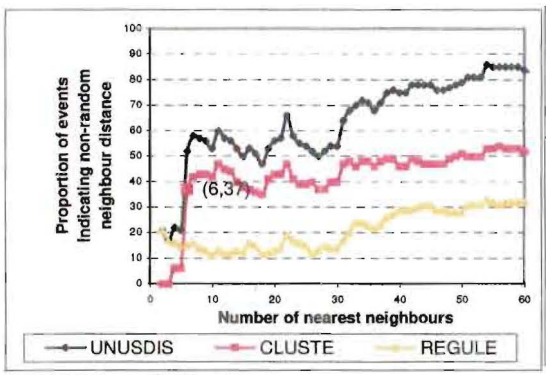


(a) Distance distribution

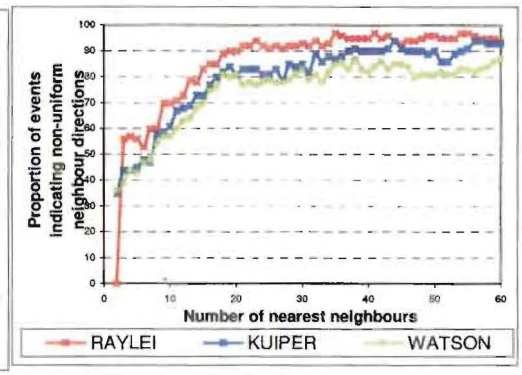


(b) Direction distribution

(40 events from CSR and 60 events from point cluster distributions)

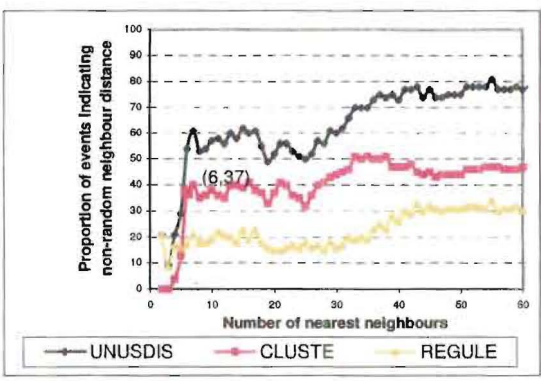


(c) Distance distribution

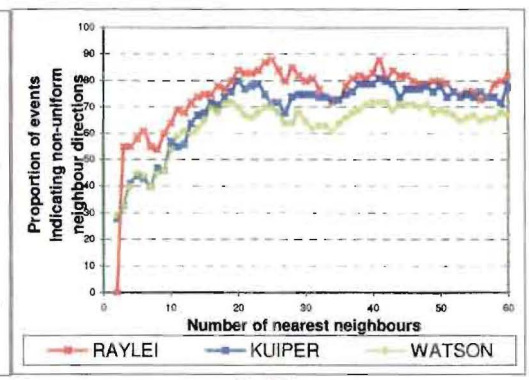


(d) Direction distribution

(50 events from CSR and 50 events from pointcluster distributions)



(e) Distance distribution



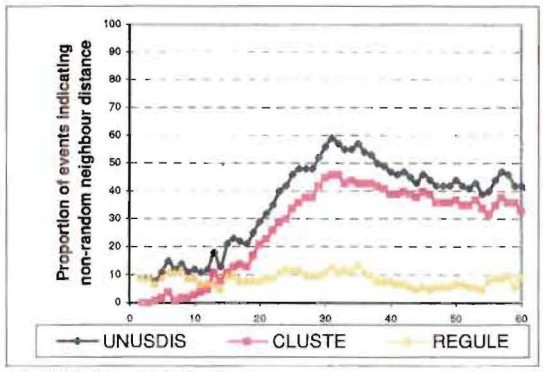
(f) Direction distribution

(60 events from CSR and 40 events from point cluster distributions)

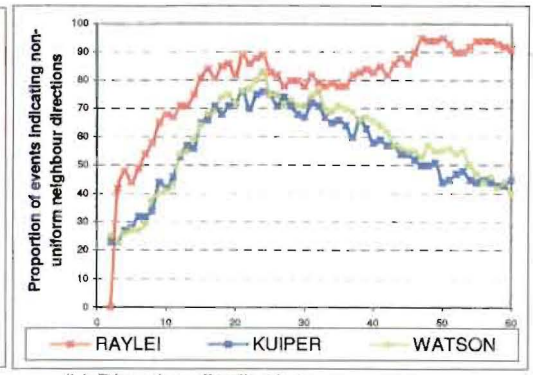
UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure 7.34: Nearest-neighbour analysis results for mixed distribution (CSR and point cluster distributions are mixed)

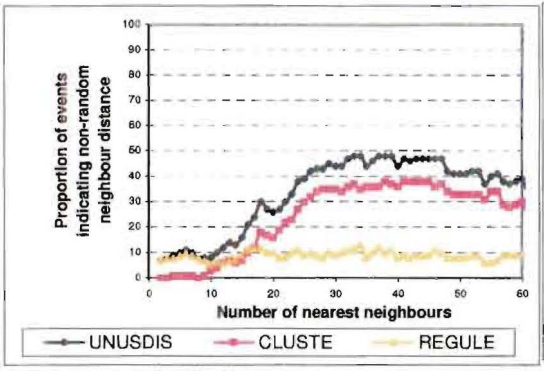


(a) Distance distribution

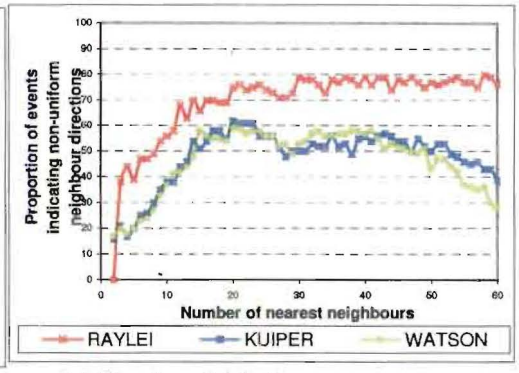


(b) Direction distribution

(40 events from CSR and 60 events from line cluster distributions)

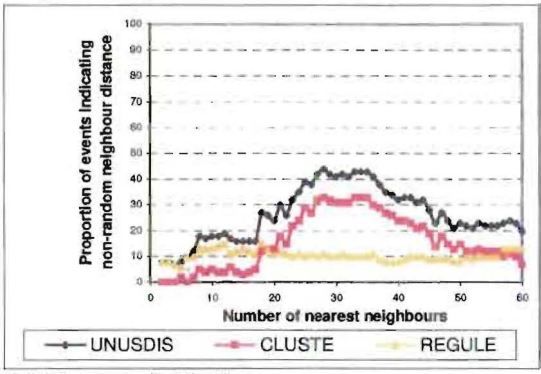


(c) Distance distribution

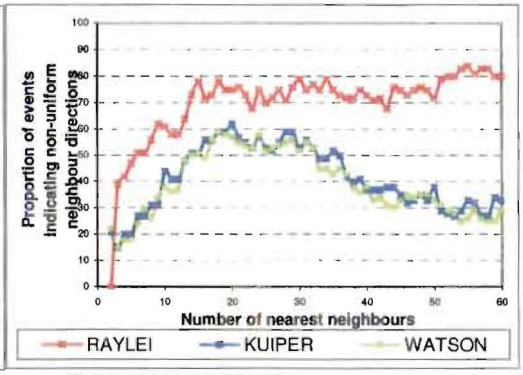


(d) Direction distribution

(50 events from CSR and 50 events from line cluster distributions)



(e) Distance distribution



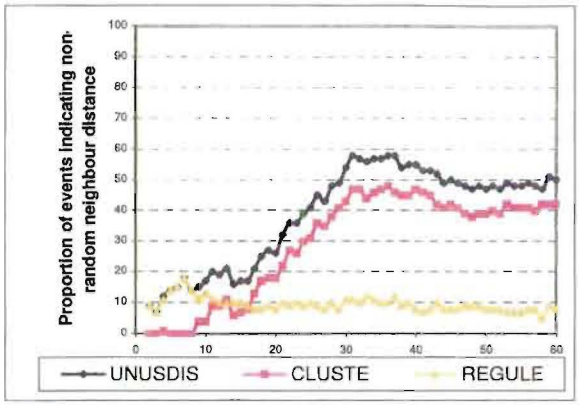
(f) Direction distribution

(60 events from CSR and 40 events from line cluster distributions)

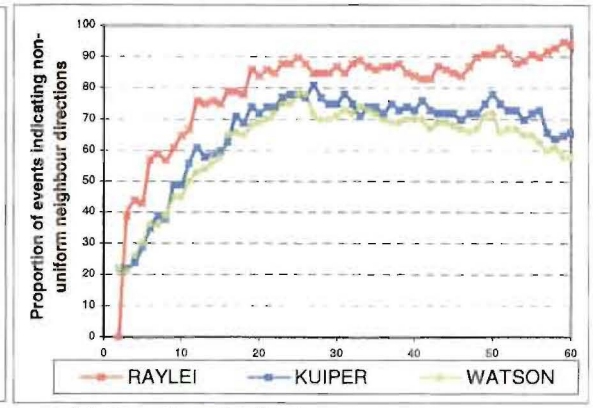
UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure 7.35: Nearest-neighbour analysis results for mixed distribution (CSR and line cluster distributions are mixed)

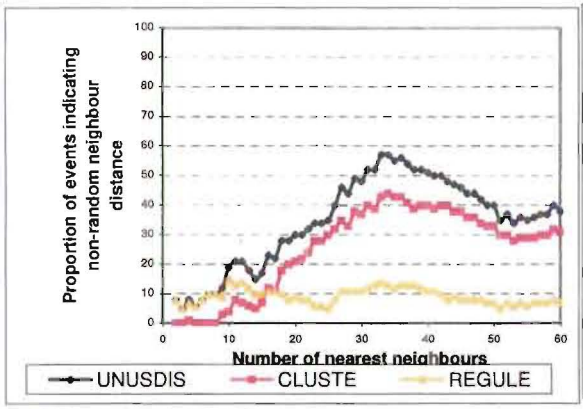


(a) Distance distribution

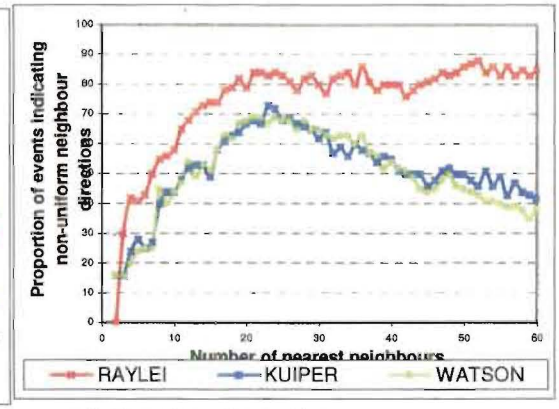


(b) Direction distribution

(20 events from CSR,30 events from point cluster distribution and 50 events from line cluster distributions)

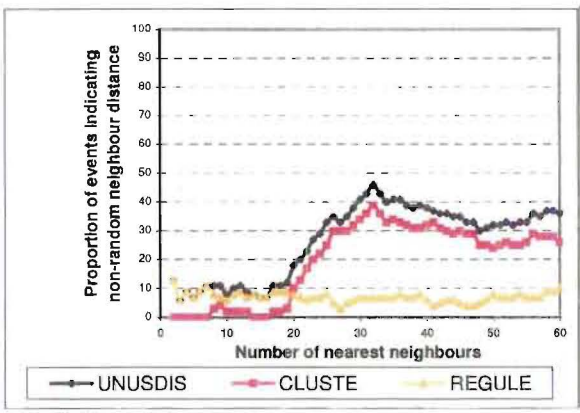


(c) Distance distribution

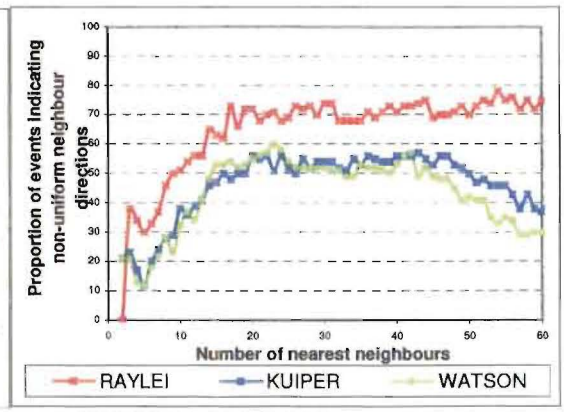


(d) Direction distribution

(30 events from CSR 20 events from point cluster distribution and 50 events from line cluster distributions)



(e) Distance distribution



(f) Direction distribution

(50 events from CSR, 20 events from point cluster and 30 events from line cluster distributions)

UNUSDIS--- Unusual distance

CLUSTE--- Cluster

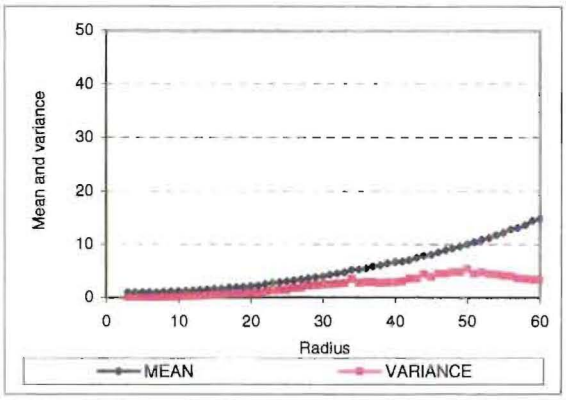
REGULE--- Regular

RAYLEI--- Rayleigh test results

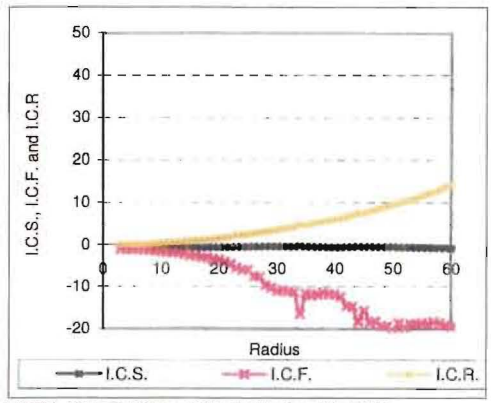
KUIPER--- Kuiper test results

WATSON--- Watson test results

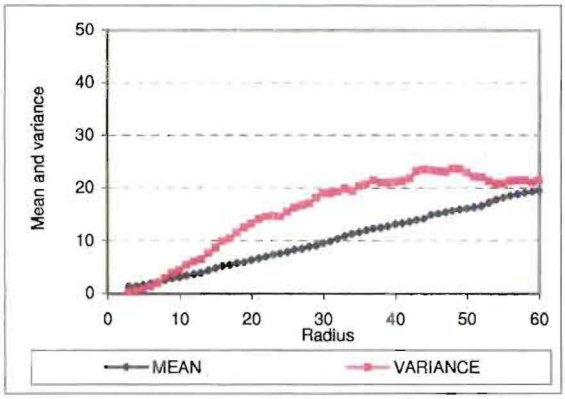
Figure 7.36: Nearest-neighbour analysis results for mixed distribution (CSR, point and line cluster distributions are mixed)



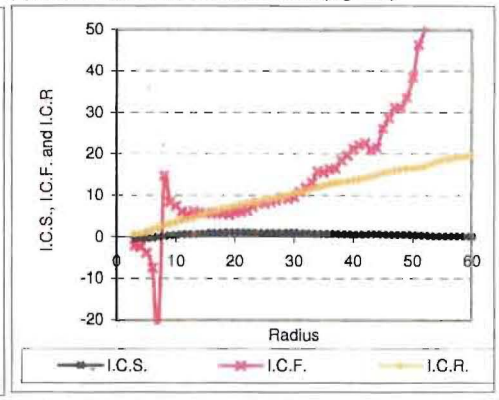
(a) Mean and variance for CSR distribution (Fig 7.01)



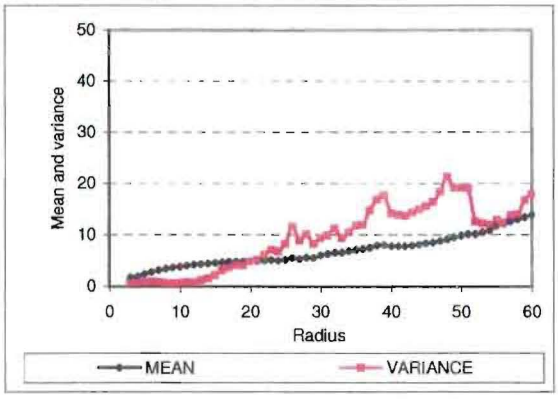
(b) ICS, ICFand ICR for CSR distribution (Fig 7.01).



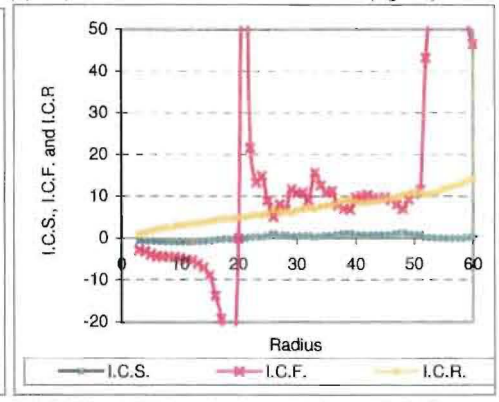
(c) Mean and variance for line cluster distribution (Fig. 7.02)



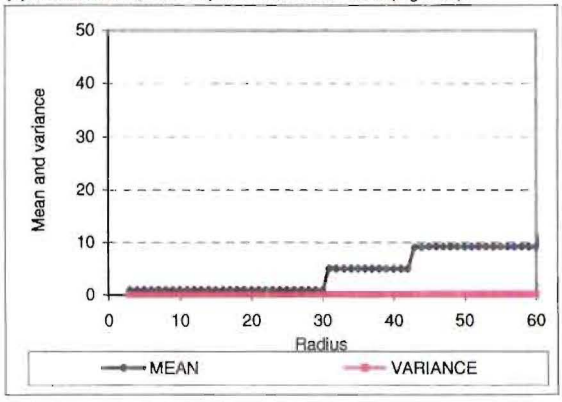
(d) ICS, ICFand ICR for line cluster distribution (Fig7.02).



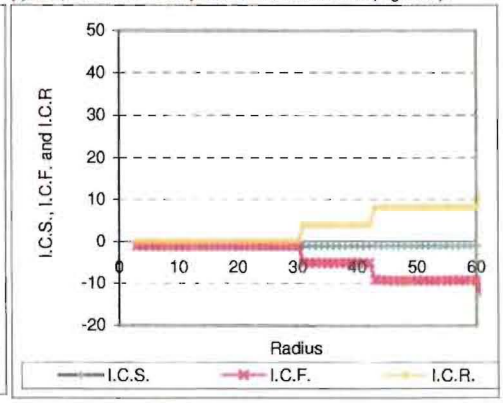
(e) Mean and variance for point cluster distribution(Fig.7.03).



(f) ICS, ICFand ICR for point cluster distribution(Fig.7.03).

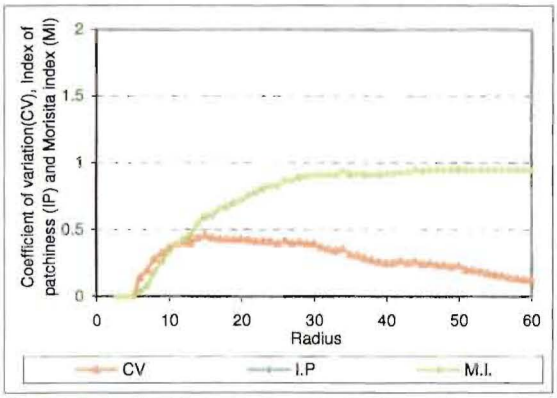


(g) Mean and variance for regular distribution (Fig. 7.04).

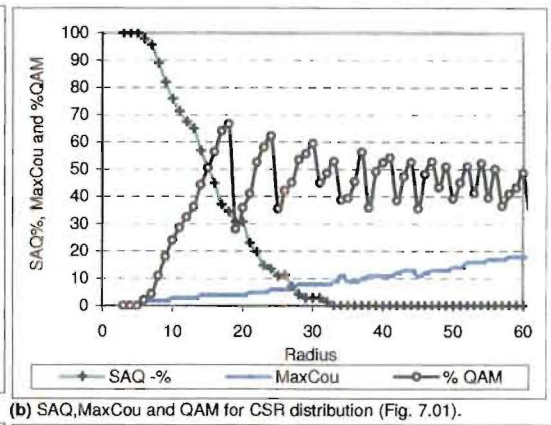


(h) ICS, ICFand ICR for regular distribution (Fig. 7.04).

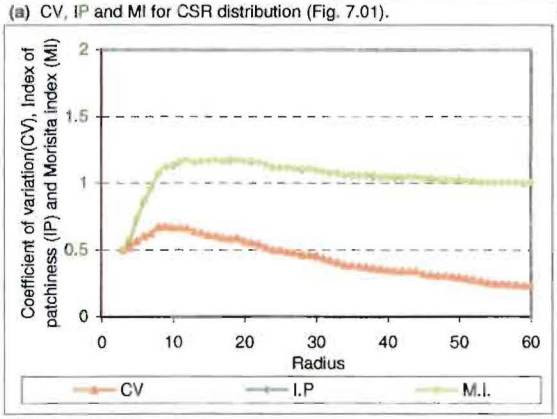
Fig. 7.37: Variation of mean, variance, ICS, ICF and ICR calculated for the four basic distributions (Figures 7.01, 7.02, 7.03 and 7.04)



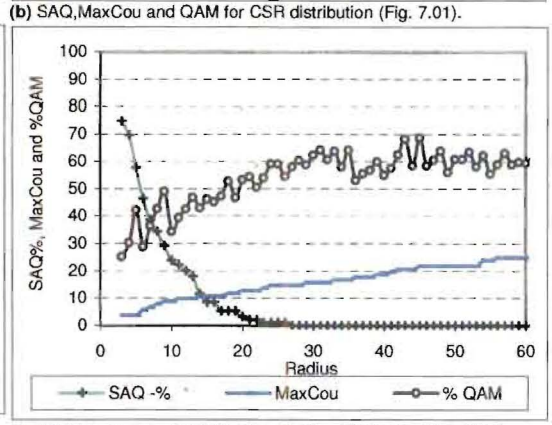
(a) CV, IP and MI for CSR distribution (Fig. 7.01).



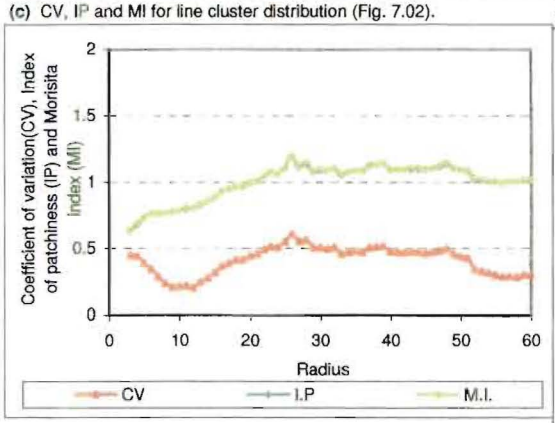
(b) SAQ,MaxCou and QAM for CSR distribution (Fig. 7.01).



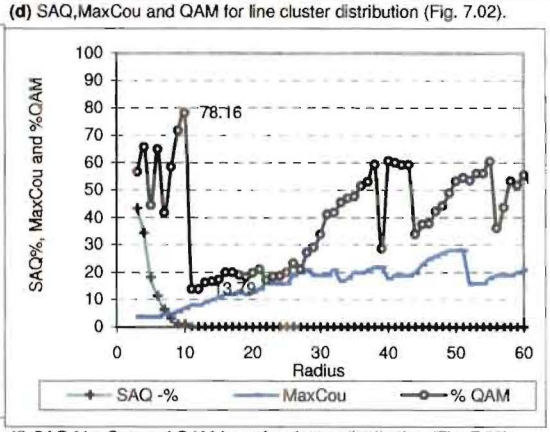
(c) CV, IP and MI for line cluster distribution (Fig. 7.02).



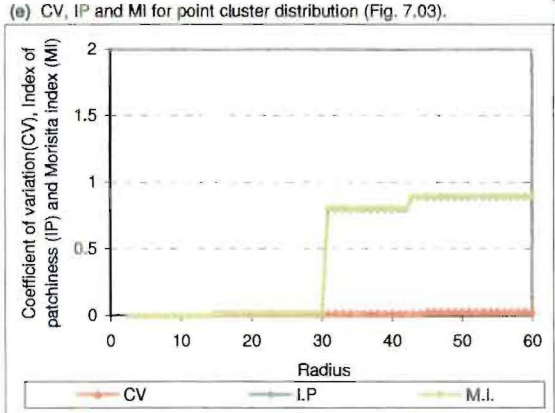
(d) SAQ,MaxCou and QAM for line cluster distribution (Fig. 7.02).



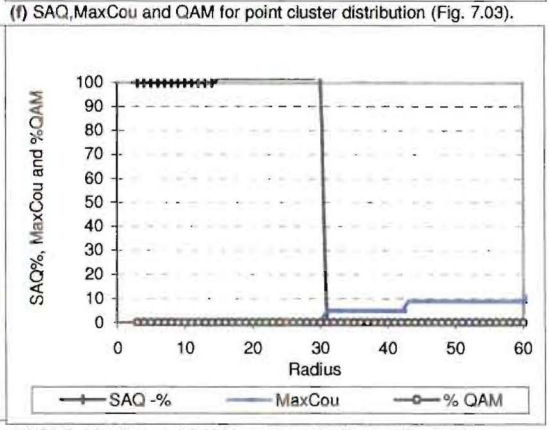
(e) CV, IP and MI for point cluster distribution (Fig. 7.03).



(f) SAQ,MaxCou and QAM for point cluster distribution (Fig. 7.03).

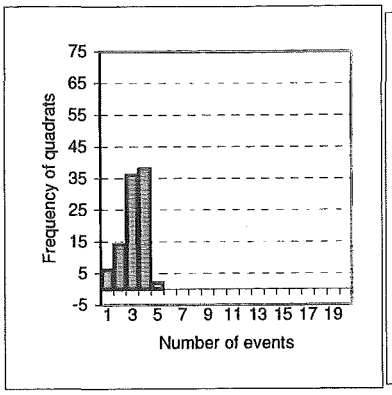


(g) CV, IP and MI for reguar distribution (Fig. 7.04).

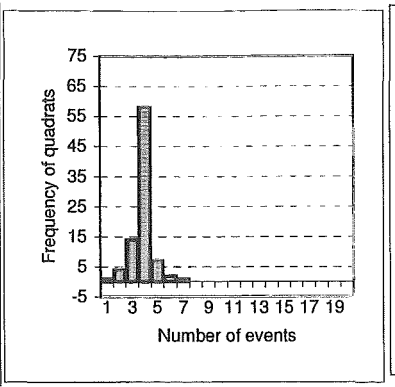


(h) SAQ, MaxCou and QAM for reguar distribution (Fig. 7.04).

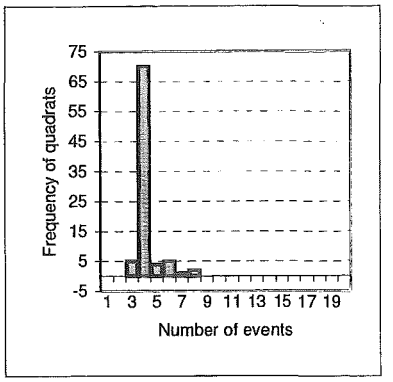
Fig. 7.38: Variation of CV, IP, MI, SAQ%, MaxCou and %QAM calculated for the four basic distributions (Figures 7.01, 7.02, 7.03 and 7.04)



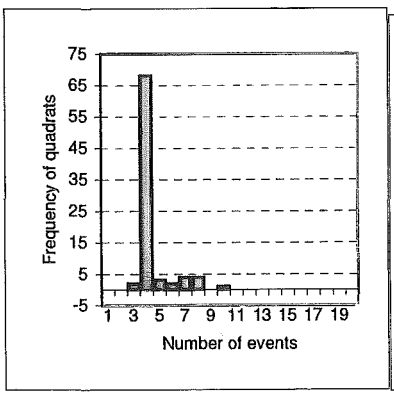
(a) Quadrat radius of 7 units, KUR = 2.1 and SKW = -0.52 (negatively skewed)



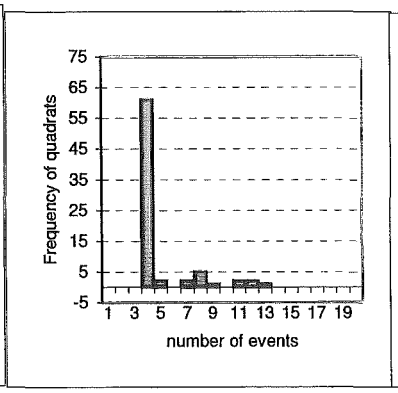
(b) Quadrat radius of 10 units, KUR = 3.18 and SKW = 0.01 (skewness is zero)



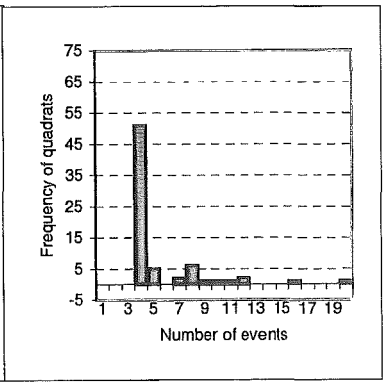
(c) Quadrat radius of 12 units, KUR = 6.11 and SKW = 1.72 (positively skewed)



(d) Quadrat radius of 14 units, KUR = 8.49 and SKW = 2.47 (positively skewed)

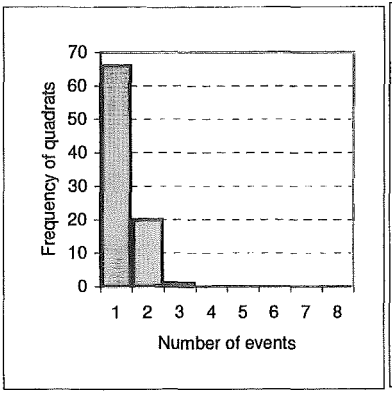


(e) Quadrat radius of 20 units, KUR = 7.16 and SKW = 2.3 (positively skewed)

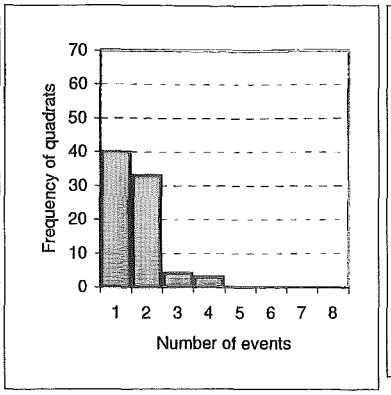


(f) Quadrat radius of 27 units, KUR = 11.6 and SKW = 2.81 (positively skewed)

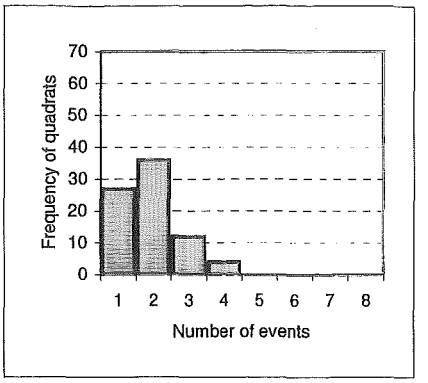
Fig. 7.39: Frequency polygon for different radii of quadrats for a point cluster distribution (Figure 7.03)



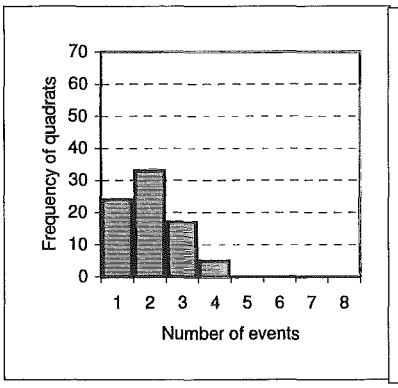
(a) Quadrat radius of 10 units, KUR = 0.18 and SKW = 0.15 (positively skewed)



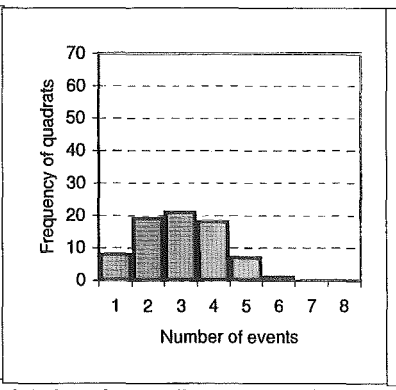
(b) Quadrat radius of 15 units, KUR = 1.45 and SKW = 0.51 (positively skewed)



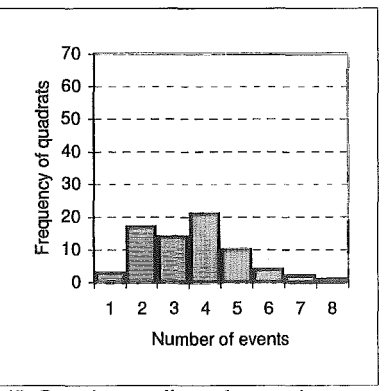
(c) Quadrat radius of 18 units, KUR = 1.41 and SKW = 0.36 (positively skewed)



(d) Quadrat radius of 19 units, KUR = 1.48 and SKW = 0.28 (positively skewed)

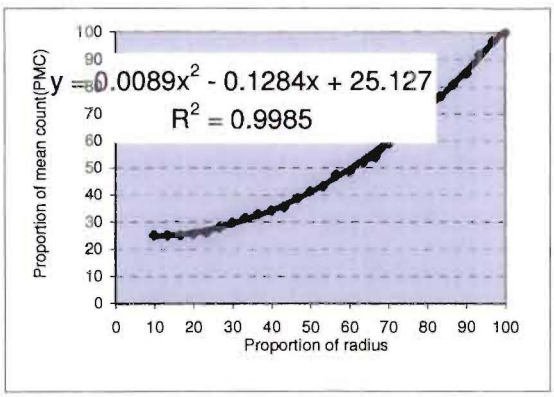


(e) Quadrat radius of 25 units, KUR = 0.04 and SKW = 2.28 (positively skewed)

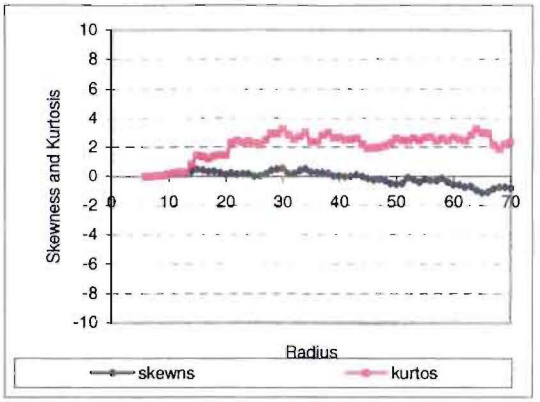


(f) Quadrat radius of 28 units, KUR = 2.94 and SKW = 0.42 (positively skewed)

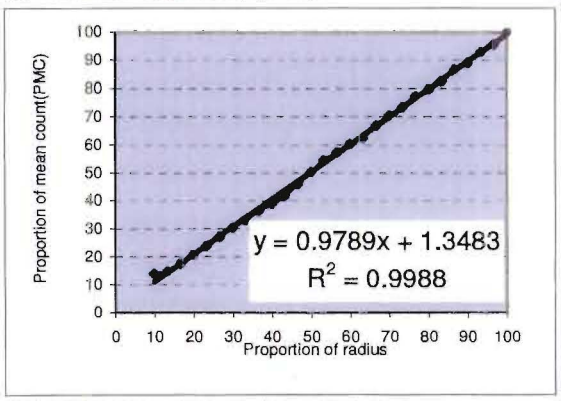
Fig. 7.40: Frequency polygon for different radii of quadrats for a CSR distribution (Figure 7.01)



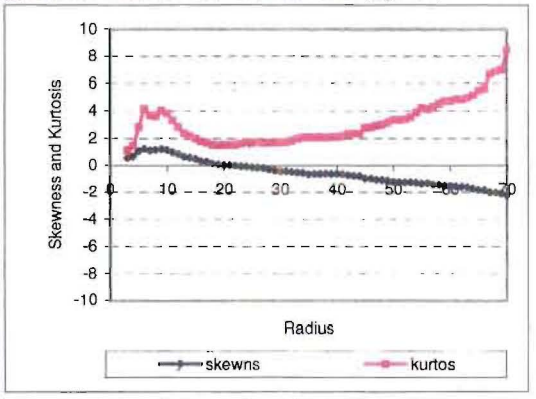
(a) PMC for CSR distribution (Fig. 7.01).



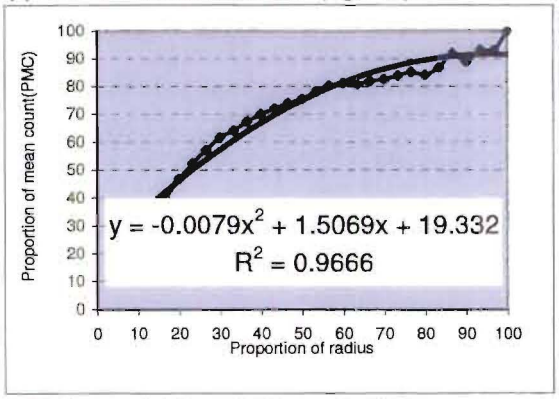
(b) Skewness and kurtosis for CSR distribution (Fig. 7.01)



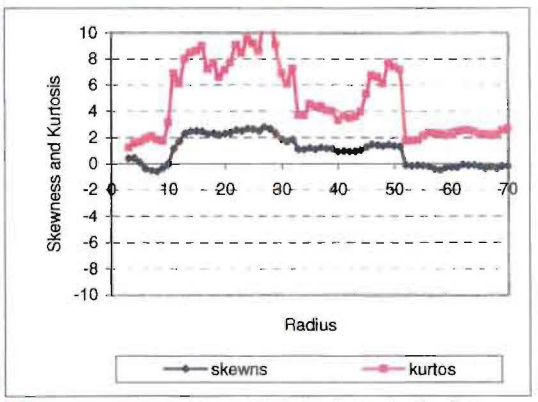
(c) PMC for line cluster distribution (Fig. 7.02)



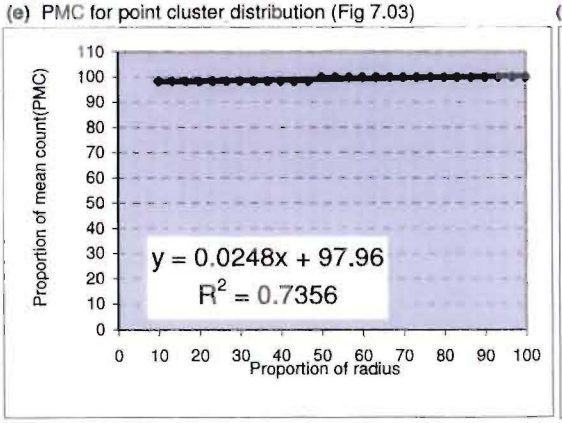
(d) Skewness and kurtosis for line cluster distribution (Fig. 7.02)



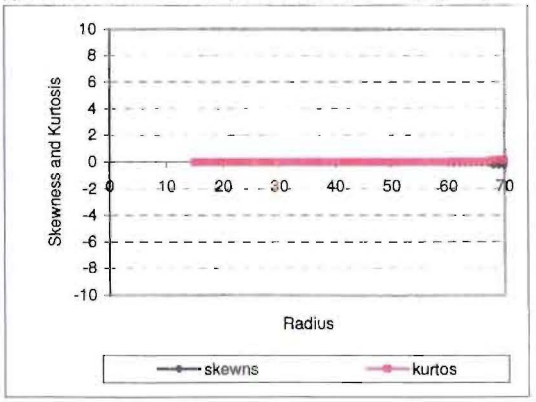
(e) PMC for point cluster distribution (Fig 7.03)



(f) Skewness and kurtosis for point cluster distribution (Fig 7.03)

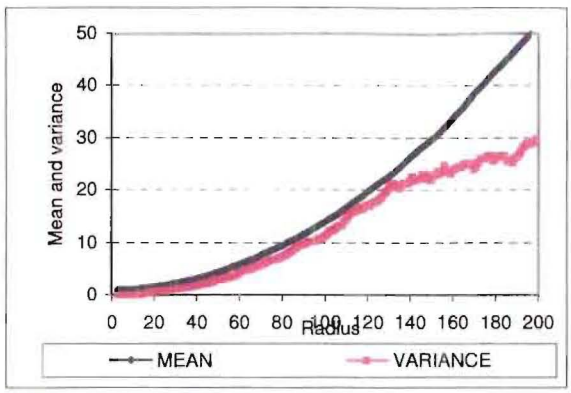


(g) PMC for regular distribution (Fig. 7.04)

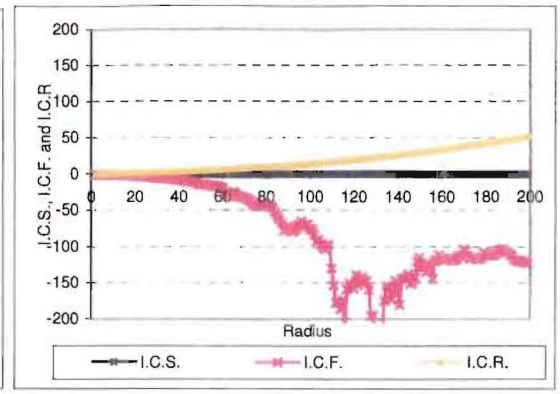


(h) Skewness and kurtosis for regular distribution (Fig. 7.04)

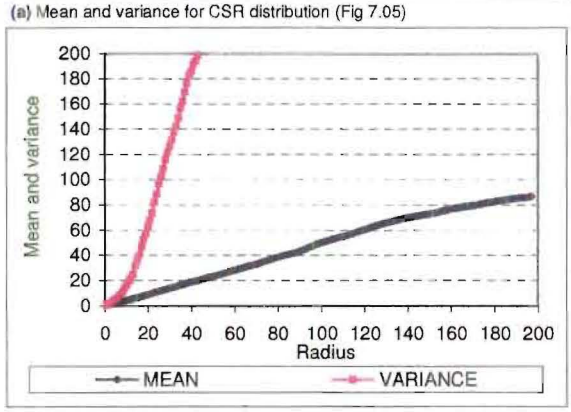
Fig. 7.41: Variation of PMC, skewness and kurtosis calculated for the four basic distributions (Figures 7.01, 7.02, 7.03 and 7.04)



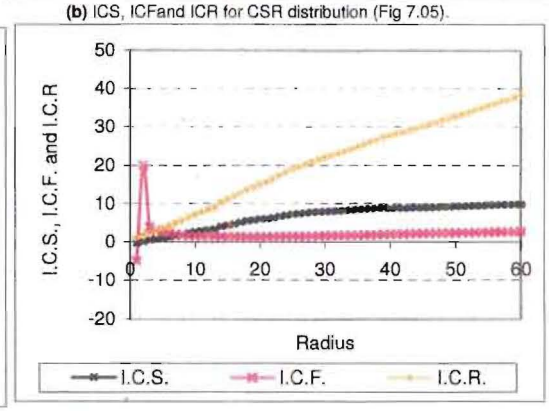
(a) Mean and variance for CSR distribution (Fig 7.05)



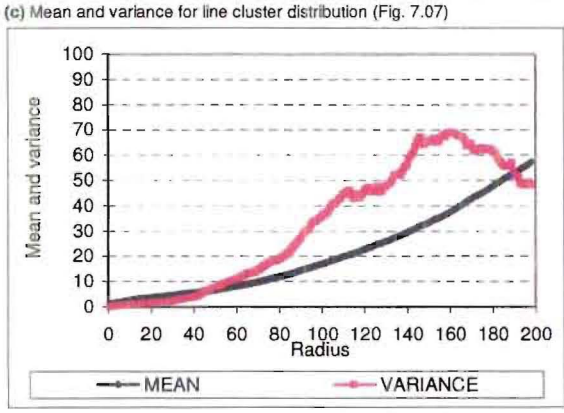
(b) ICS, ICFand ICR for CSR distribution (Fig 7.05).



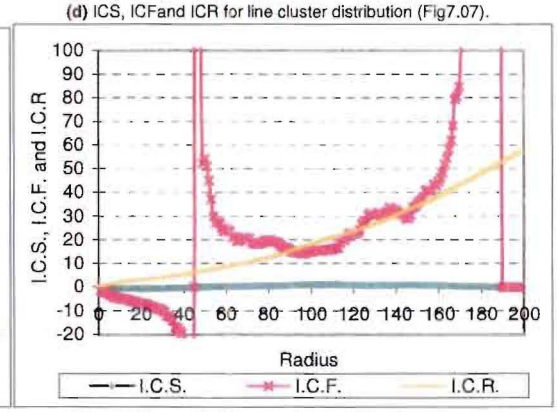
(c) Mean and variance for line cluster distribution (Fig. 7.07)



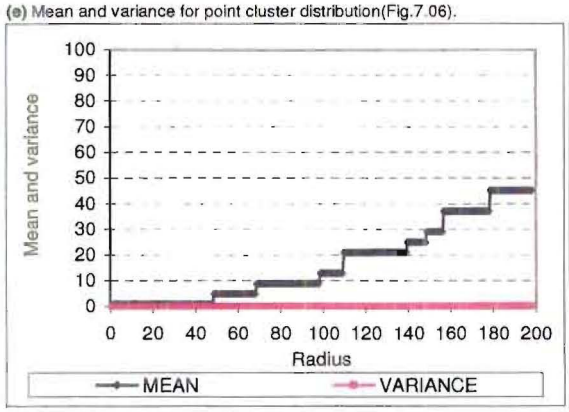
(d) ICS, ICFand ICR for line cluster distribution (Fig7.07).



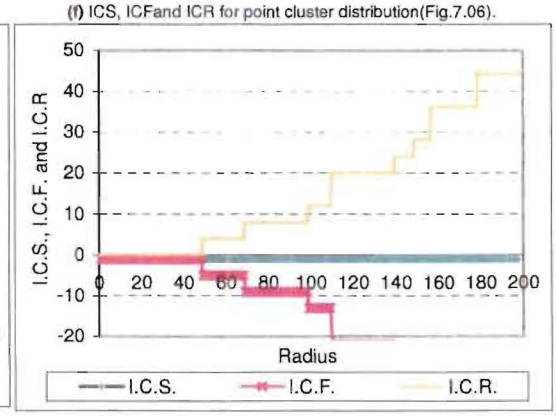
(e) Mean and variance for point cluster distribution(Fig.7.06).



(f) ICS, ICFand ICR for point cluster distribution(Fig.7.06).

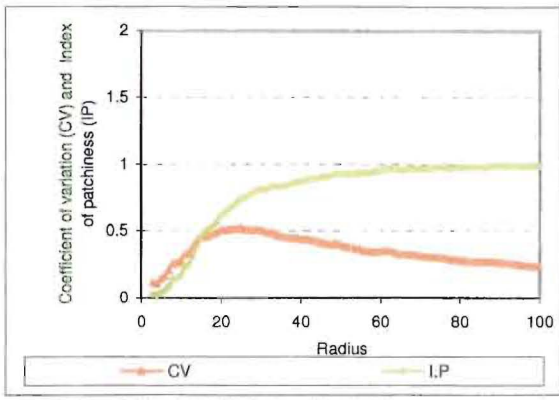


(g) Mean and variance for regular distribution (Fig. 7.08).

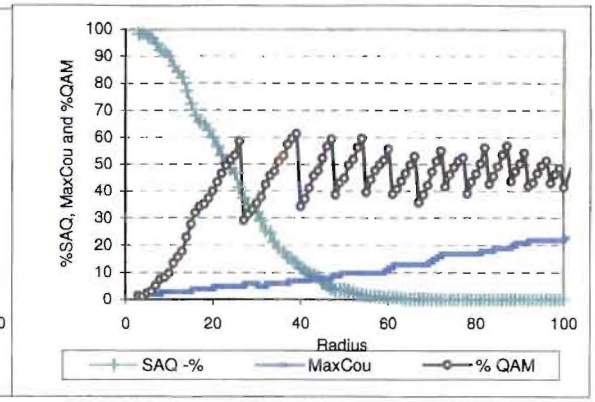


(h) ICS, ICFand ICR for regular distribution (Fig. 7.08).

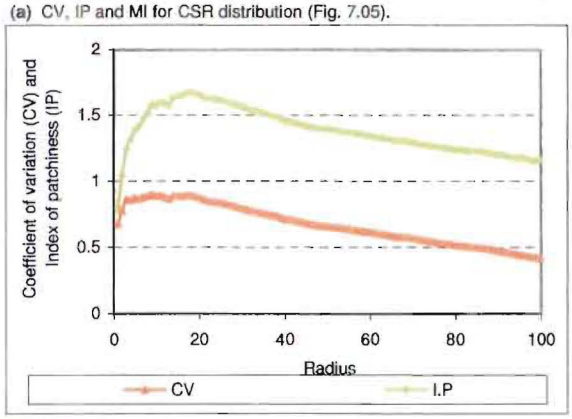
Figure 7.42: Variation of mean, variance, ICS, ICF and ICR calculated for the four distributions (Figures 7.05, 7.07, 7.06 & 7.08)



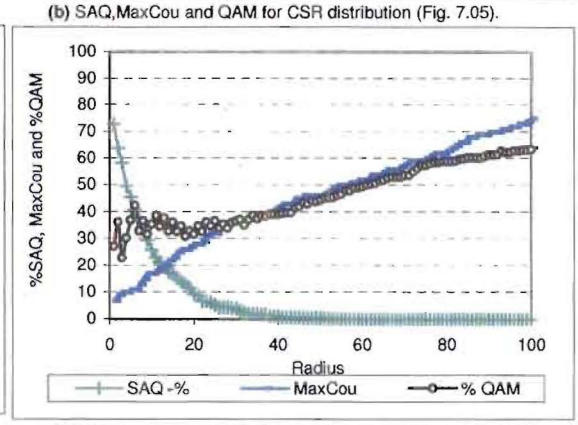
(a) CV, IP and MI for CSR distribution (Fig. 7.05).



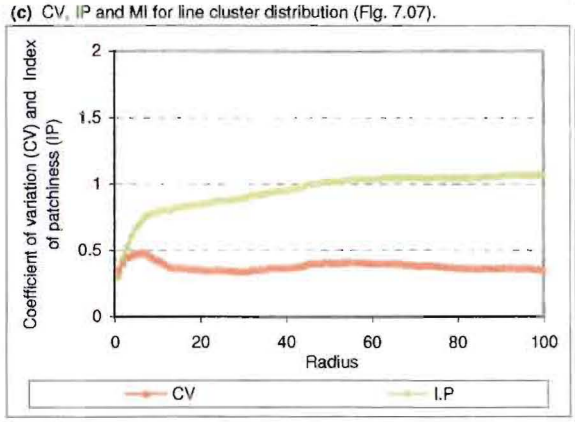
(b) SAQ,MaxCou and QAM for CSR distribution (Fig. 7.05).



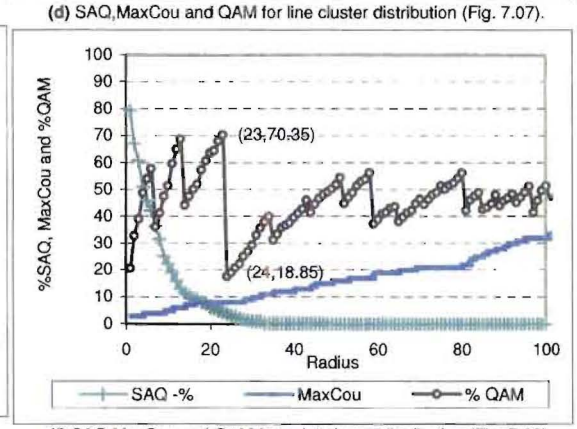
(c) CV, IP and MI for line cluster distribution (Flg. 7.07).



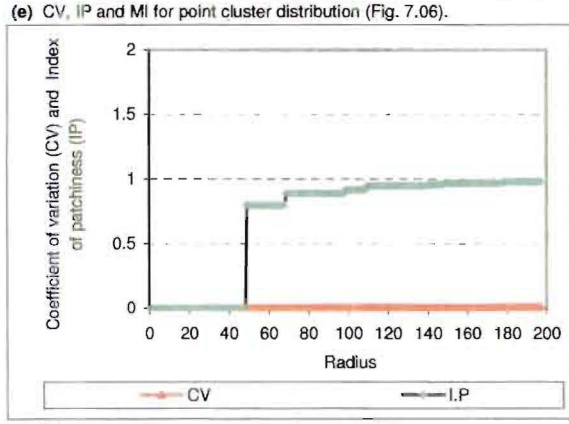
(d) SAQ,MaxCou and QAM for line cluster distribution (Fig. 7.07).



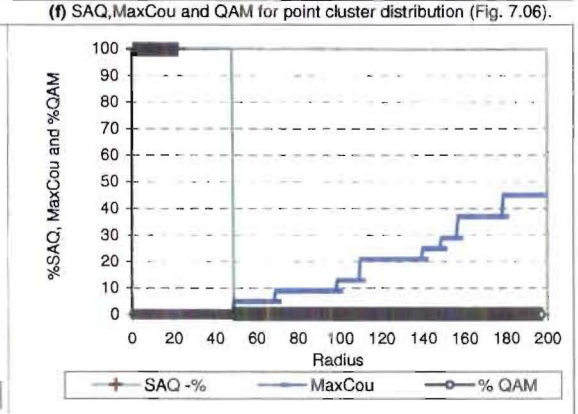
(e) CV, IP and MI for point cluster distribution (Fig. 7.06).



(f) SAQ,MaxCou and QAM for point cluster distribution (Fig. 7.06).

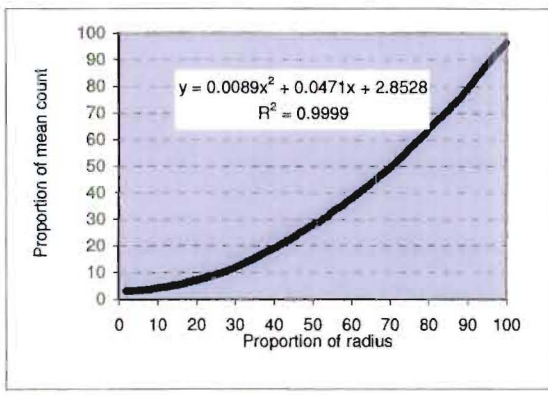


(g) CV, IP and MI for reguar distribution (Fig. 7.0.8).

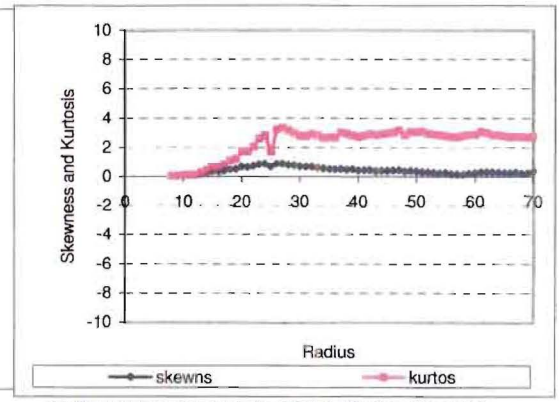


(h) SAQ, MaxCou and QAM for reguar distribution (Fig. 7.0.8).

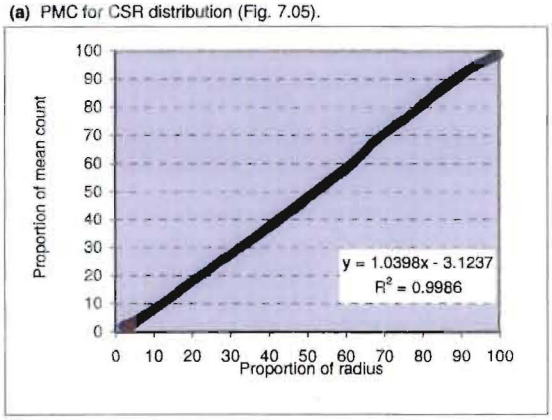
Figure 7.43: Variation of CV, IP, SAQ%, MaxCou and %QAM calculated for the four basic distributions (Figures 7.05, 7.07, 7.06 & 7.08)



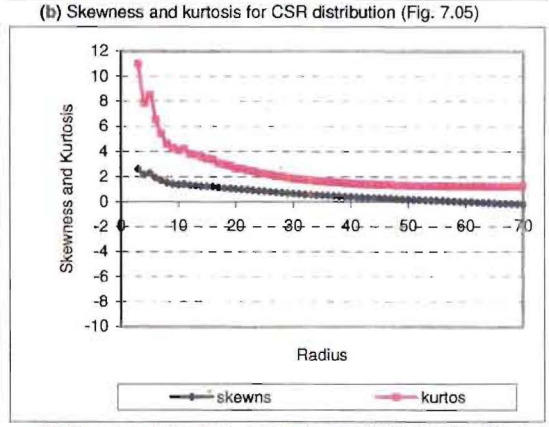
(a) PMC for CSR distribution (Fig. 7.05).



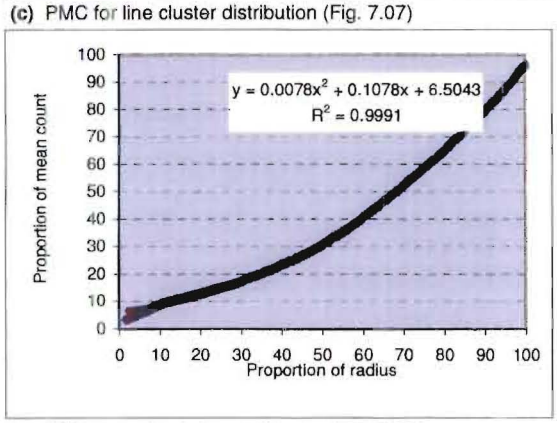
(b) Skewness and kurtosis for CSR distribution (Fig. 7.05)



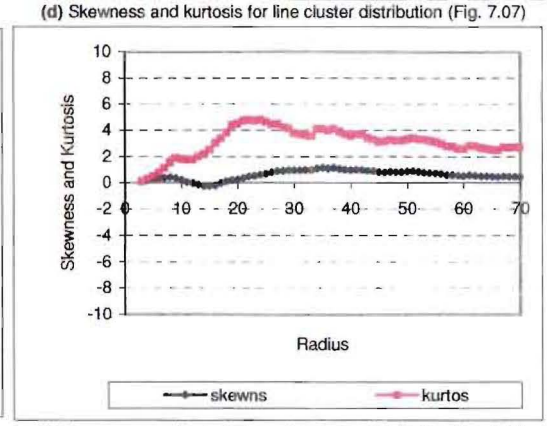
(c) PMC for line cluster distribution (Fig. 7.07)



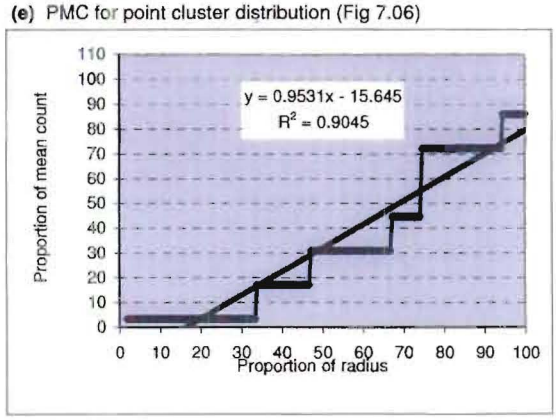
(d) Skewness and kurtosis for line cluster distribution (Fig. 7.07)



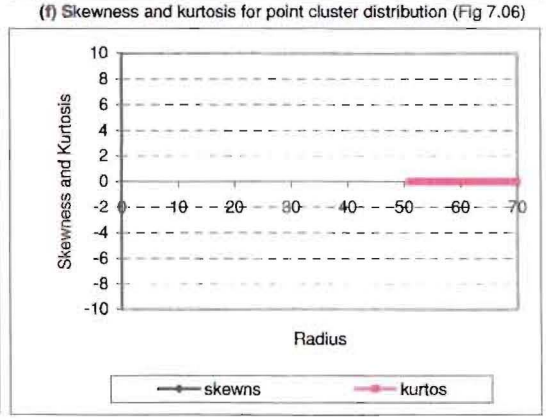
(e) PMC for point cluster distribution (Fig 7.06)



(f) Skewness and kurtosis for point cluster distribution (Fig 7.06)

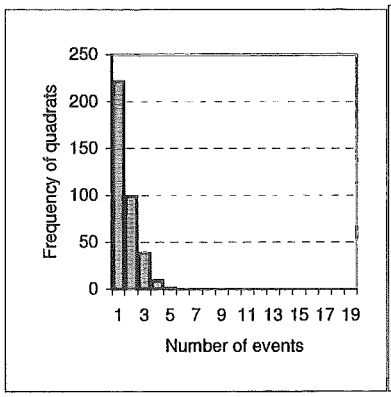


(g) PMC for regular distribution (Fig. 7.08)

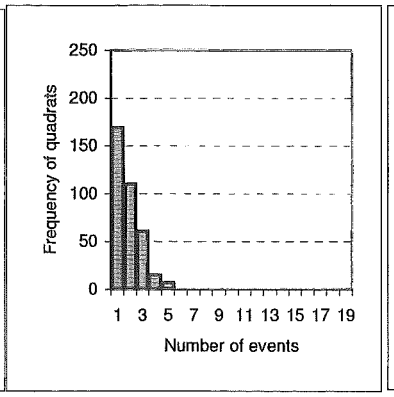


(h) Skewness and kurtosis for regular distribution (Fig. 7.08)

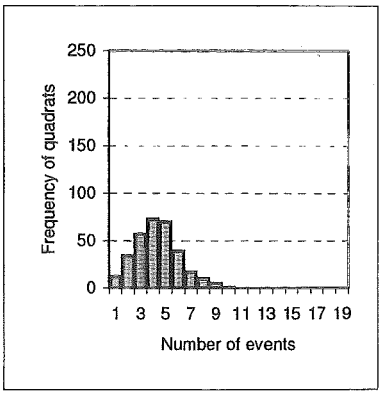
Figure 7.44: Variation of PMC, skewness and kurtosis calculated for the four basic distributions (Figures 7.05, 7.06, 7.07 & 7.08)



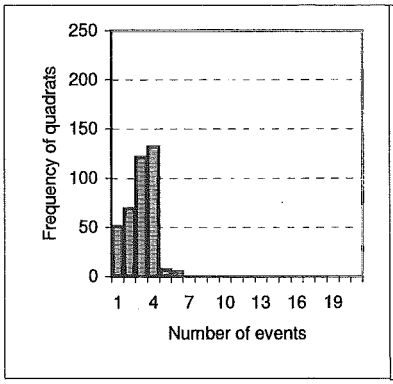
(a) Quadrat radius of 20 units, KUR = 1.77 and SKW = 0.69 (positively skewed) for CSR distribution



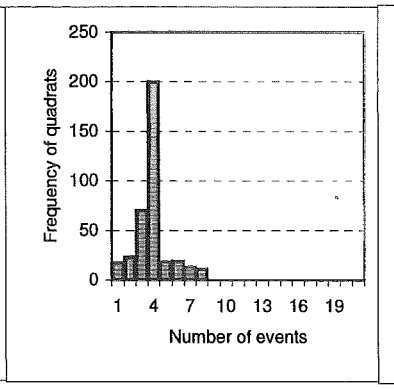
(b) Quadrat radius of 25 units, KUR = 1.77 and SKW = 0.69 (positively skewed) for CSR distribution



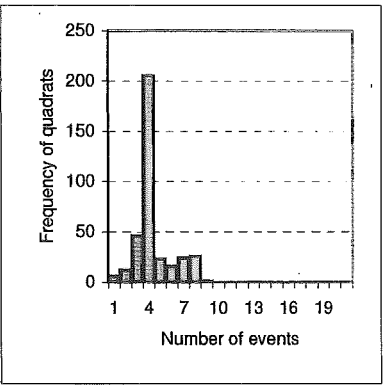
(c) Quadrat radius of 50 units, KUR = 3.06 and SKW = 0.37 (positively skewed) for CSR distribution



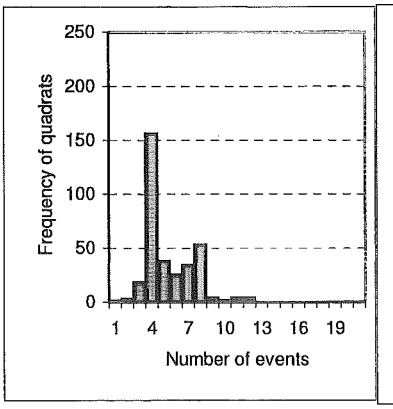
(d) Quadrat radius of 15 units, KUR = 2.21 and SKW = -0.22 (negatively skewed) for point cluster distribution



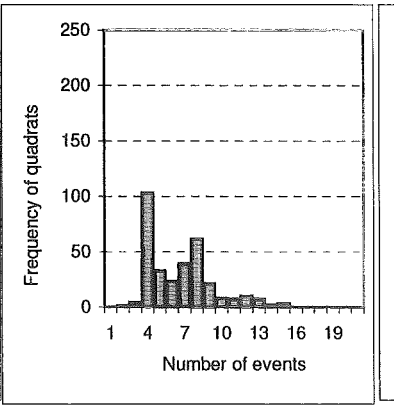
(e) Quadrat radius of 25 units, KUR = 4.67 and SKW = 0.7 (positively skewed) for point cluster distribution



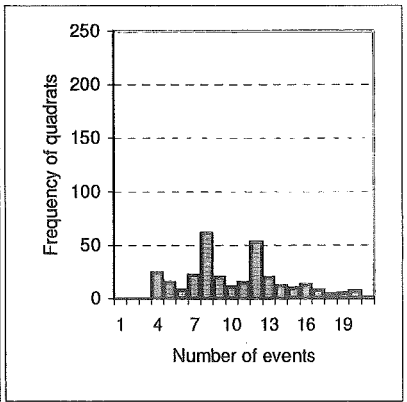
(f) Quadrat radius of 30 units, KUR = 3.81 and SKW = 0.97 (positively skewed) for point cluster distribution



(g) Quadrat radius of 40 units, KUR = 3.59 and SKW = 1.0 (positively skewed) for point cluster distribution

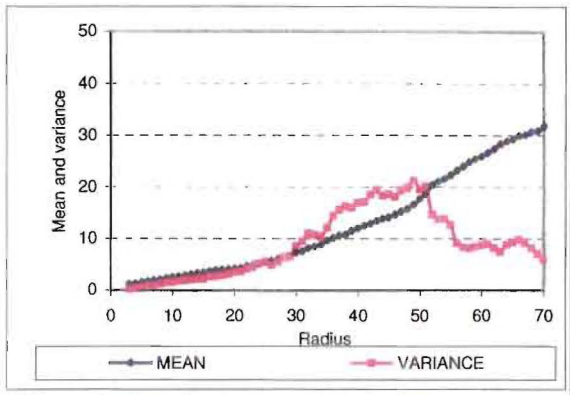


(h) Quadrat radius of 50 units, KUR = 3.35 and SKW = 0.89 (positively skewed) for point cluster distribution

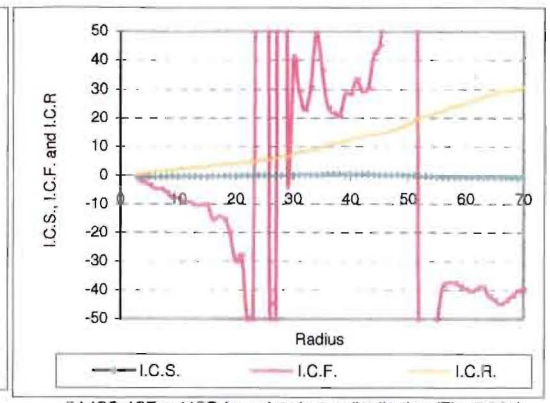


(i) Quadrat radius of 75 units, KUR = 2.6 and SKW = 0.4 (positively skewed) for point cluster distribution

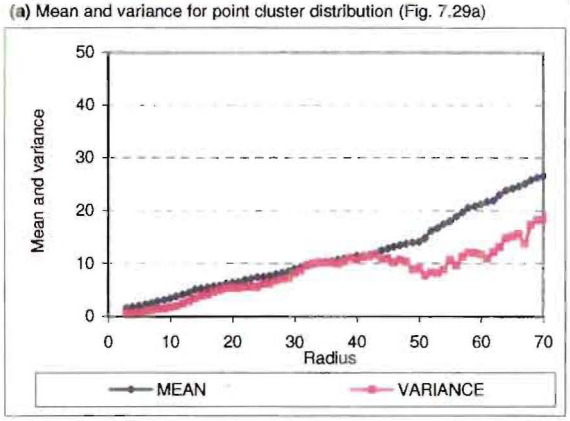
Figure 7.45: Frequency polygons for different radii of quadrats for the CSR and the point cluster distributions (Figures 7.05 and 7.06)



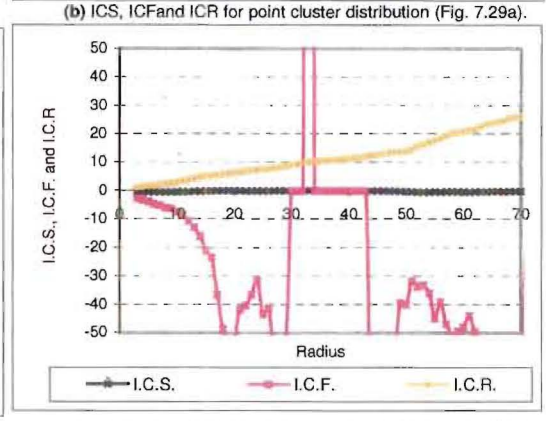
(a) Mean and variance for point cluster distribution (Fig. 7.29a)



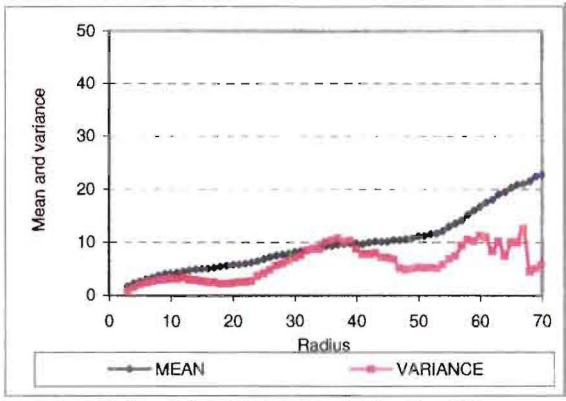
(b) ICS, ICFand ICR for point cluster distribution (Fig. 7.29a).



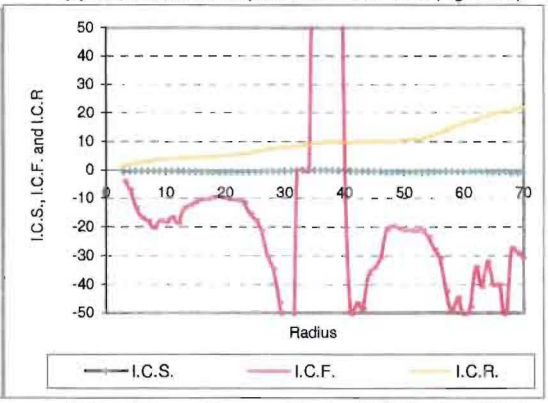
(c) Mean and variance for point cluster distribution (Fig. 7.29b)



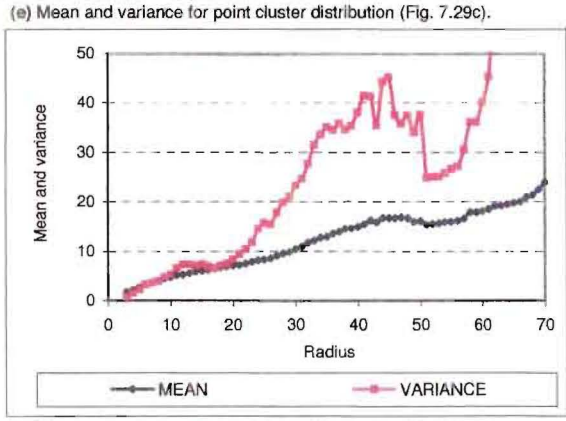
(d) ICS, ICFand ICR for point cluster distribution (Fig. 7.29b).



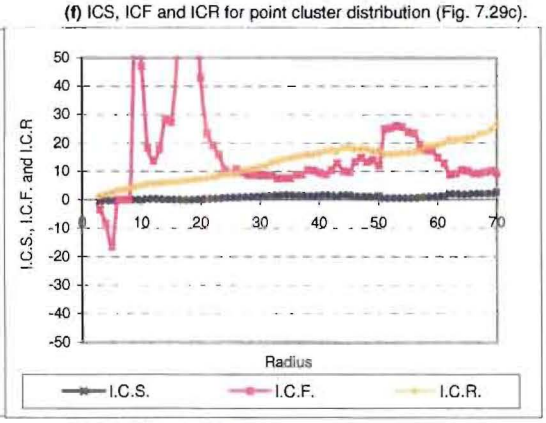
(e) Mean and variance for point cluster distribution (Fig. 7.29c).



(f) ICS, ICF and ICR for point cluster distribution (Fig. 7.29c).

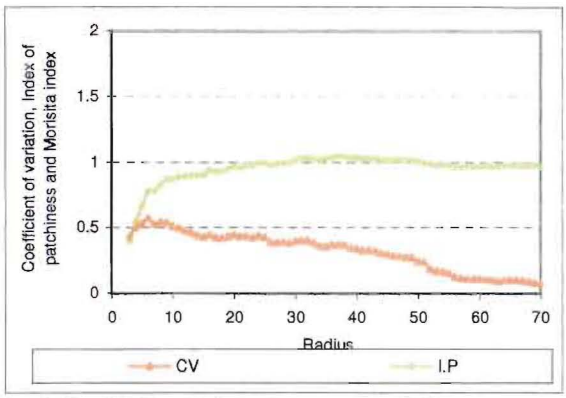


(g) Mean and variance for point cluster distribution (Fig. 7.29d).

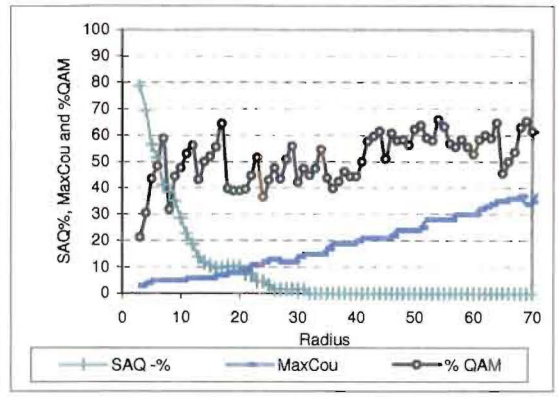


(h) ICS, ICFand ICR for point cluster distribution (Fig. 7.29d).

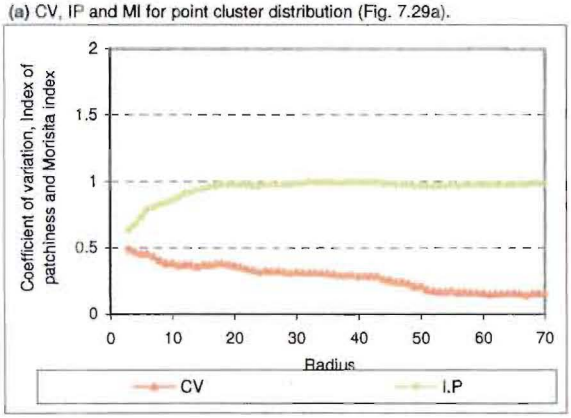
Fig. 7.46: Variation of mean, variance, ICS, ICF and ICR calculated for the four distributions (Figure 7.29)



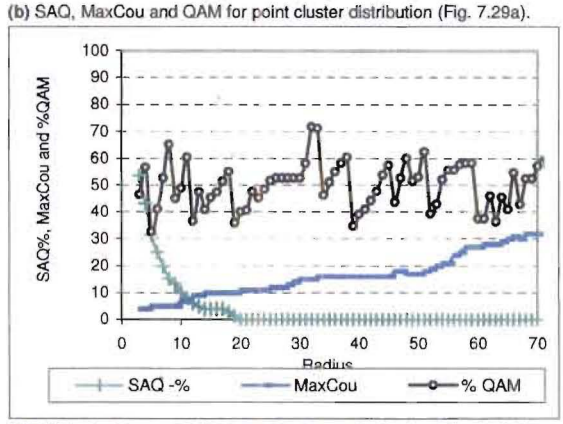
(a) CV, IP and MI for point cluster distribution (Fig. 7.29a).



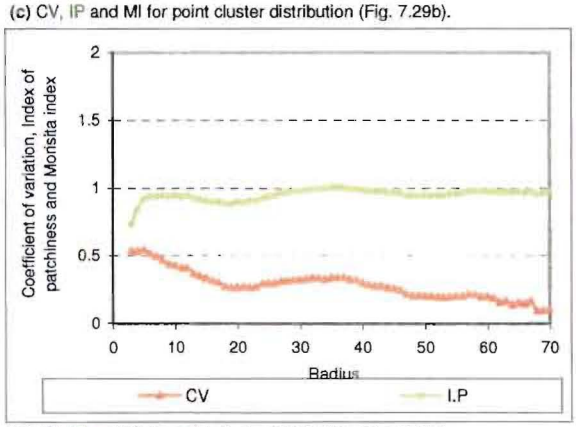
(b) SAQ, MaxCou and QAM for point cluster distribution (Fig. 7.29a).



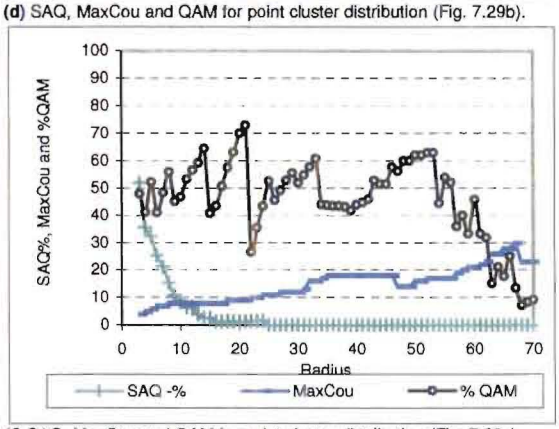
(c) CV, IP and MI for point cluster distribution (Fig. 7.29b).



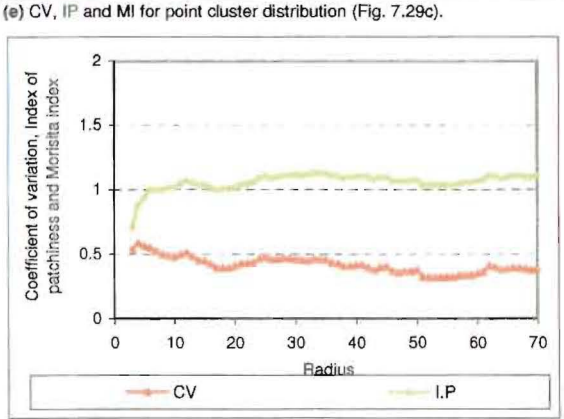
(d) SAQ, MaxCou and QAM for point cluster distribution (Fig. 7.29b).



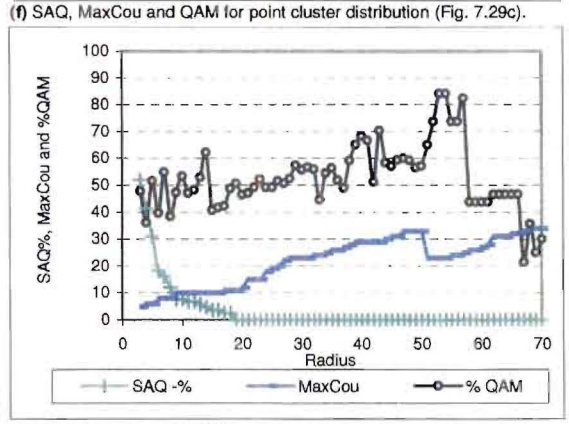
(e) CV, IP and MI for point cluster distribution (Fig. 7.29c).



(f) SAQ, MaxCou and QAM for point cluster distribution (Fig. 7.29c).

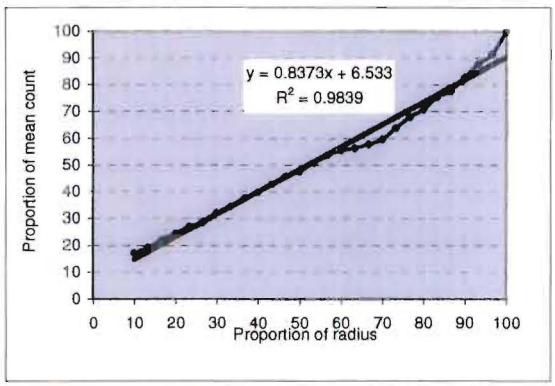


(g) CV, IP and MI for point cluster distribution (Fig. 7.29d).

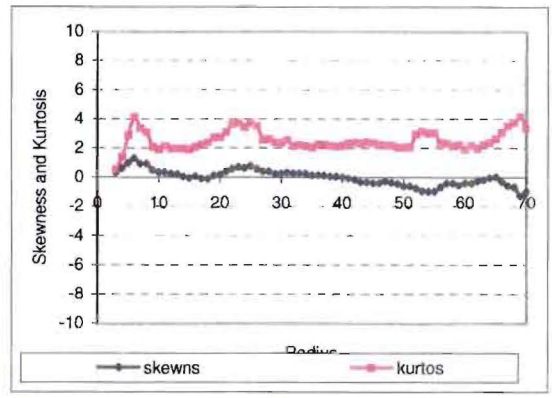


(h) SAQ, MaxCou and QAM for point cluster distribution (Fig. 7.29d).

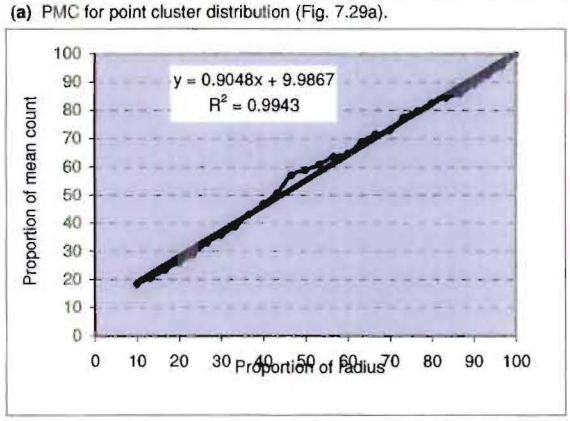
Fig. 7.47: Variation of CV, IP, MI, SAQ%, MaxCou and %QAM calculated for the four distributions (Figure 7.29)



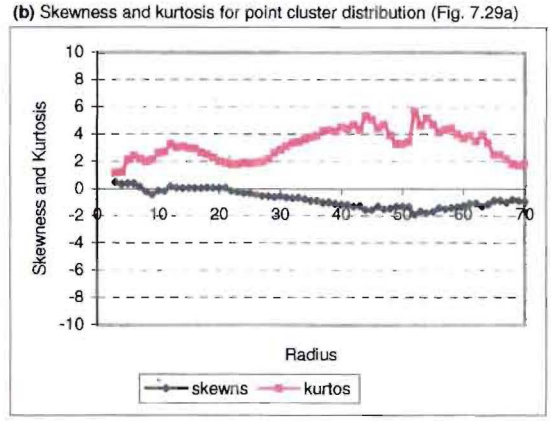
(a) PMC for point cluster distribution (Fig. 7.29a).



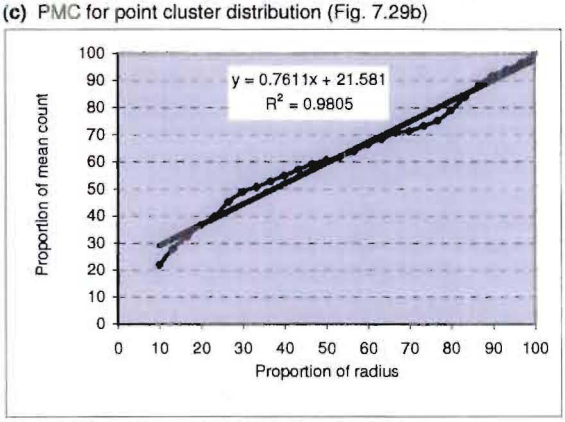
(b) Skewness and kurtosis for point cluster distribution (Fig. 7.29a)



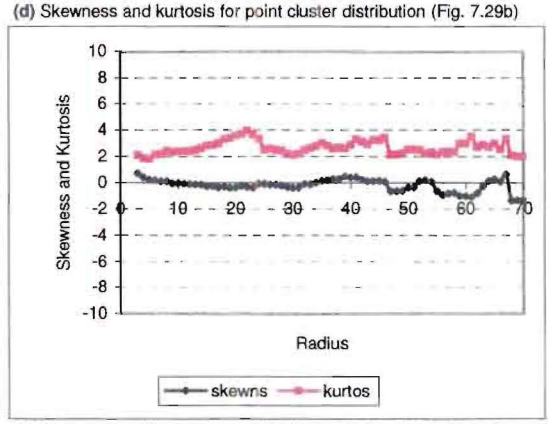
(c) PMC for point cluster distribution (Fig. 7.29b)



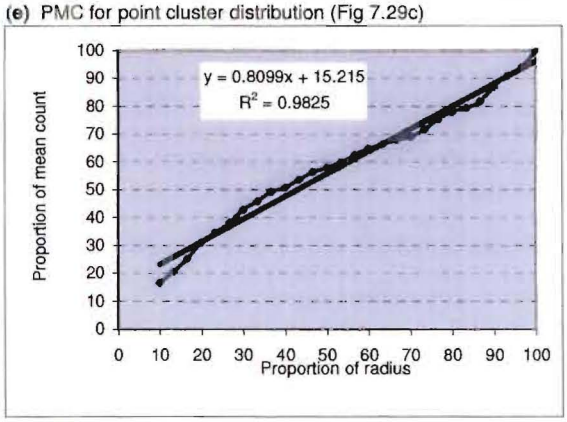
(d) Skewness and kurtosis for point cluster distribution (Fig. 7.29b)



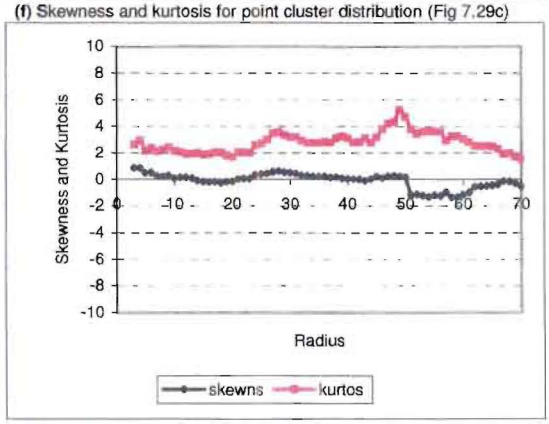
(e) PMC for point cluster distribution (Fig 7.29c)



(f) Skewness and kurtosis for point cluster distribution (Fig 7.29c)

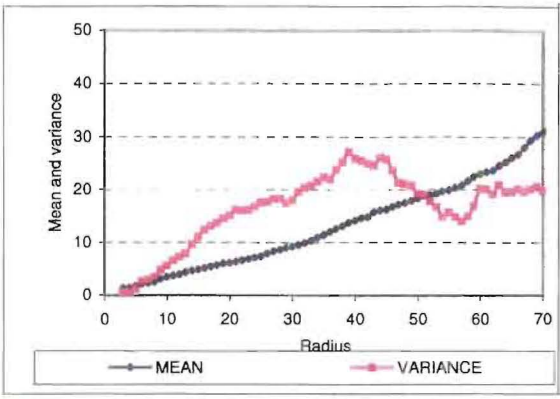


(g) PMC for point cluster distribution (Fig. 7.29d)

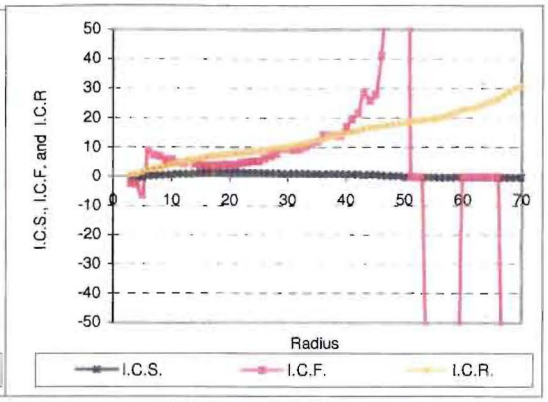


(h) Skewness and kurtosis for point cluster distribution (Fig. 7.29d)

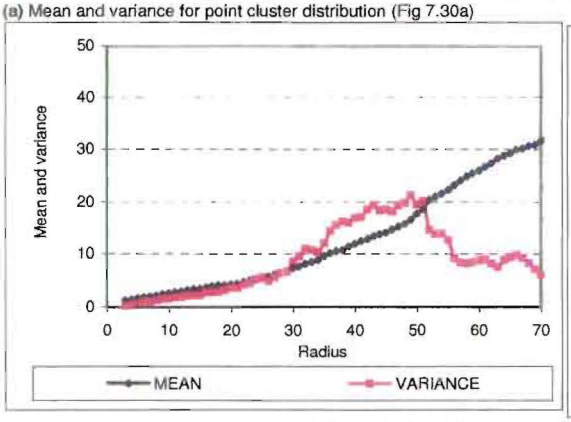
Fig. 7.48: Variation of PMC, skewness and kurtosis calculated for the four distributions (Figure 7.29)



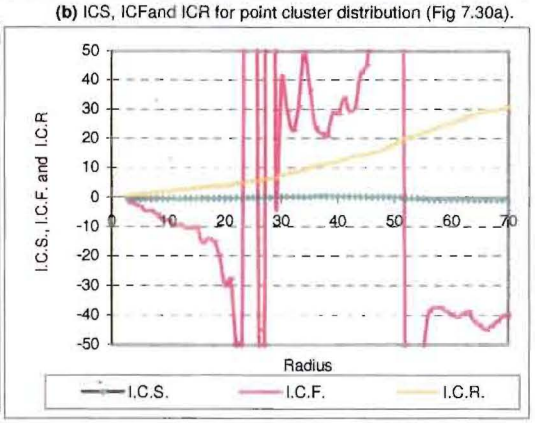
(a) Mean and variance for point cluster distribution (Fig 7.30a)



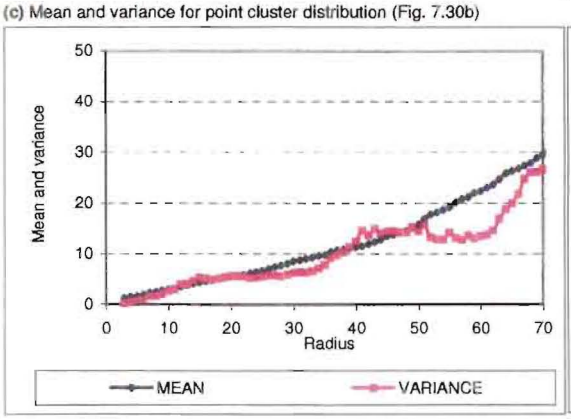
(b) ICS, ICFand ICR for point cluster distribution (Fig 7.30a).



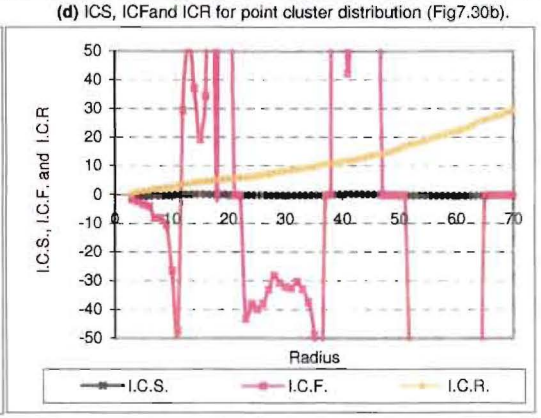
(c) Mean and variance for point cluster distribution (Fig. 7.30b)



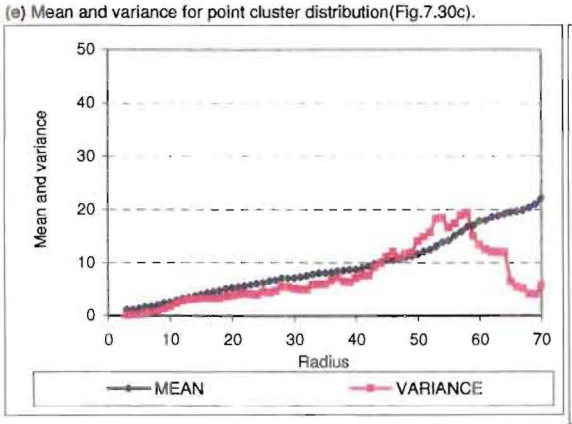
(d) ICS, ICFand ICR for point cluster distribution (Fig7.30b).



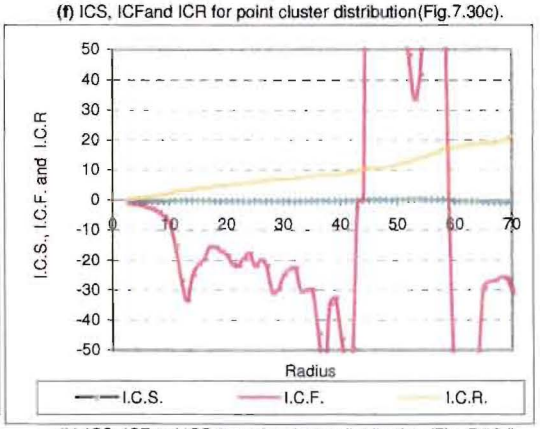
(e) Mean and variance for point cluster distribution(Fig.7.30c).



(f) ICS, ICFand ICR for point cluster distribution(Fig.7.30c).

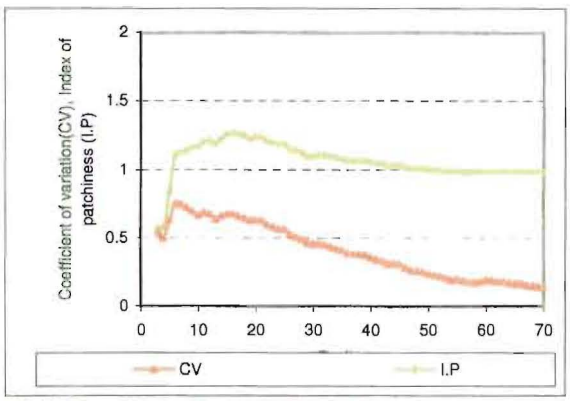


(g) Mean and variance for point cluster distribution (Fig. 7.30d).

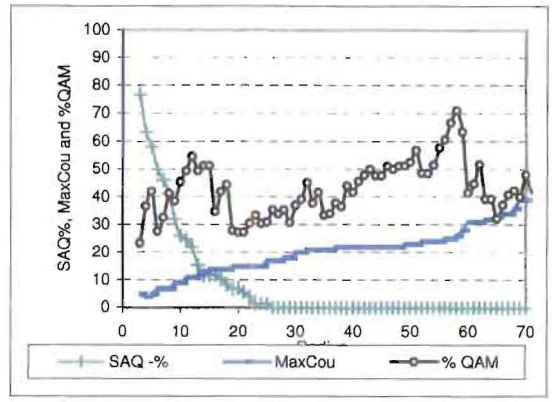


(h) ICS, ICFand ICR for point cluster distribution (Fig. 7.30d).

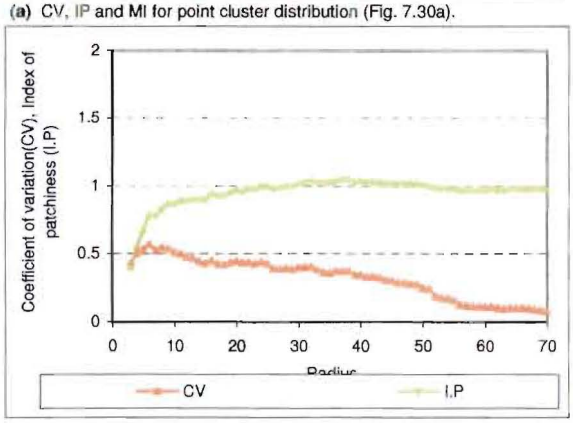
Fig. 7.49: Variation of mean, variance, ICS, ICF and ICR calculated for the four distributions (Figure 7.30)



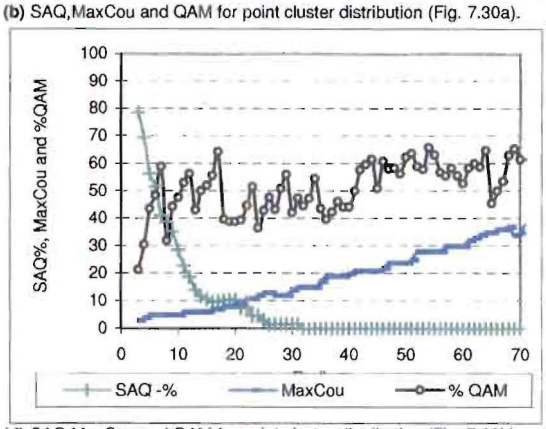
(a) CV, IP and MI for point cluster distribution (Fig. 7.30a).



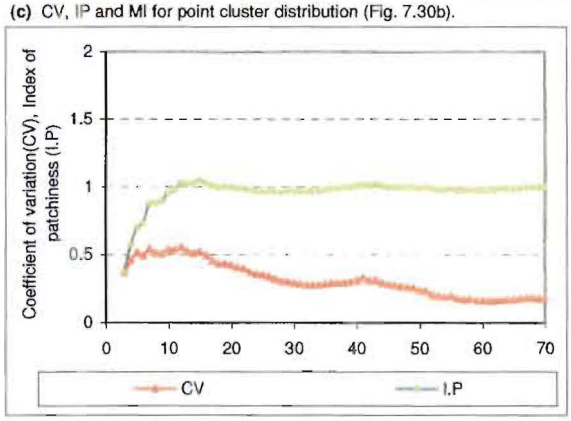
(b) SAQ,MaxCou and QAM for point cluster distribution (Fig. 7.30a).



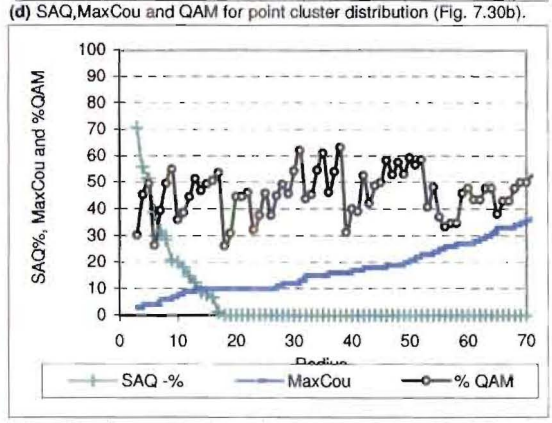
(c) CV, IP and MI for point cluster distribution (Fig. 7.30b).



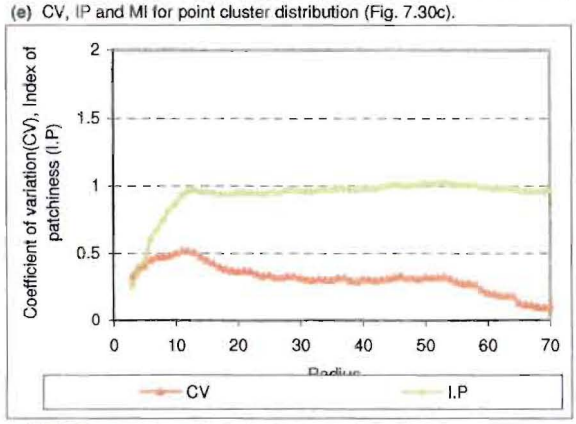
(d) SAQ,MaxCou and QAM for point cluster distribution (Fig. 7.30b).



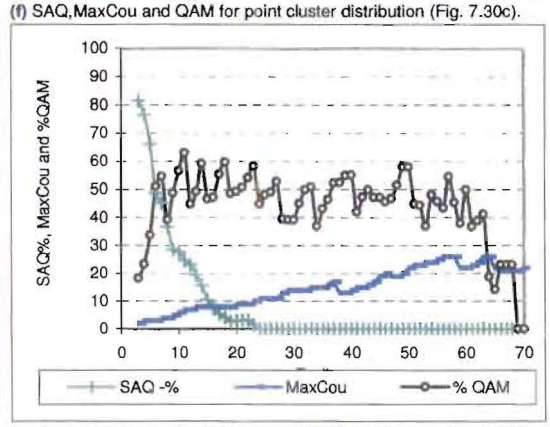
(e) CV, IP and MI for point cluster distribution (Fig. 7.30c).



(f) SAQ,MaxCou and QAM for point cluster distribution (Fig. 7.30c).

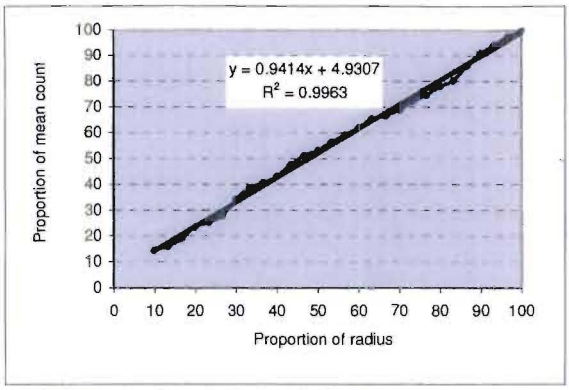


(g) CV, IP and MI for point cluster distribution (Fig. 7.30d).

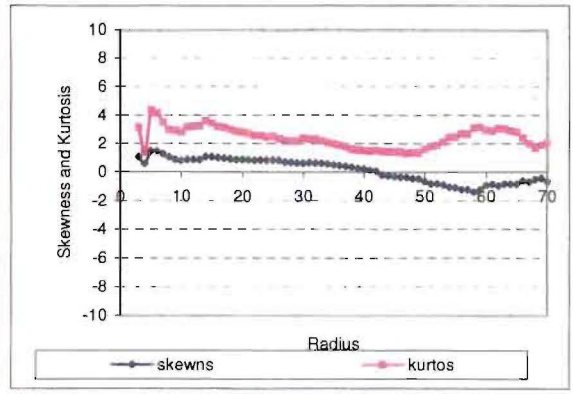


(h) SAQ, MaxCou and QAM for point cluster distribution (Fig. 7.30d).

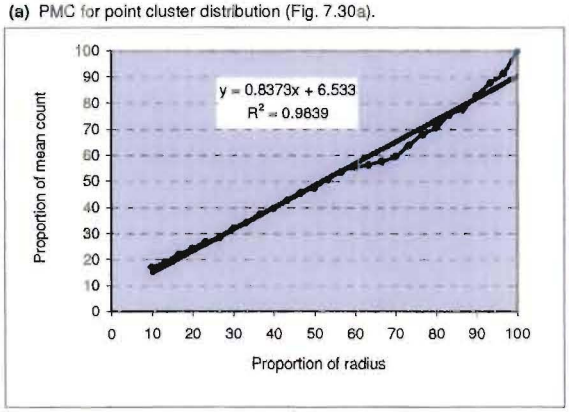
Fig. 7.50: Variation of CV, IP, MI, SAQ%, MaxCou and %QAM calculated for the four distributions (Figures 7.30)



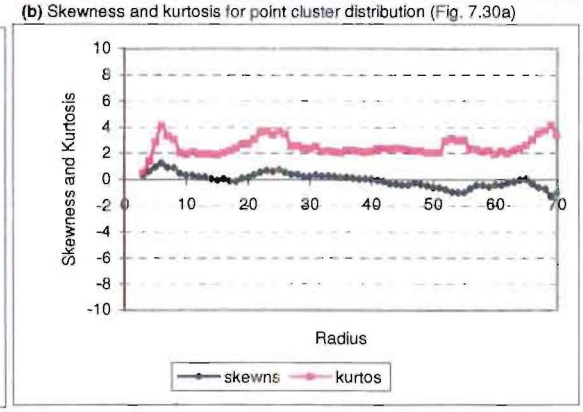
(a) PMC for point cluster distribution (Fig. 7.30a).



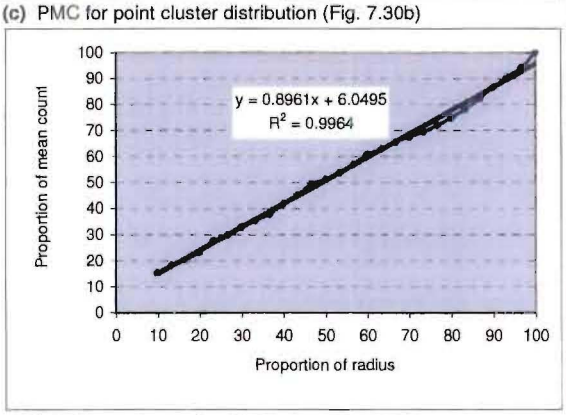
(b) Skewness and kurtosis for point cluster distribution (Fig. 7.30a)



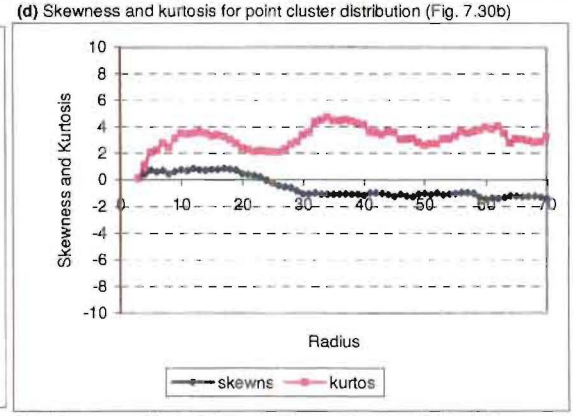
(c) PMC for point cluster distribution (Fig. 7.30b)



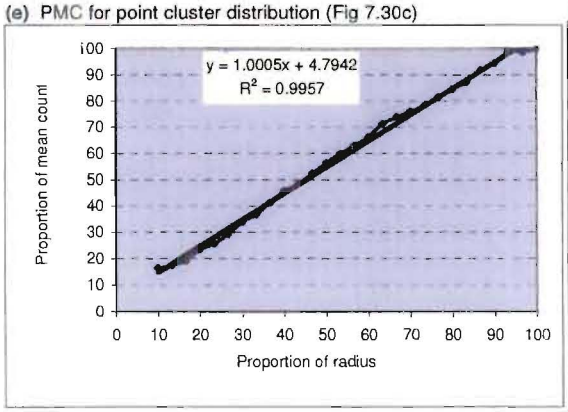
(d) Skewness and kurtosis for point cluster distribution (Fig. 7.30b)



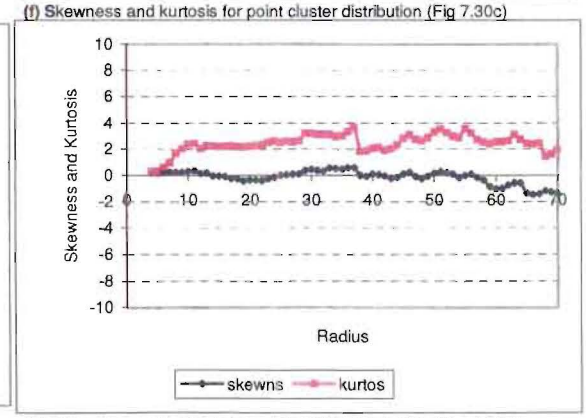
(e) PMC for point cluster distribution (Fig 7.30c)



(f) Skewness and kurtosis for point cluster distribution (Fig 7.30c)

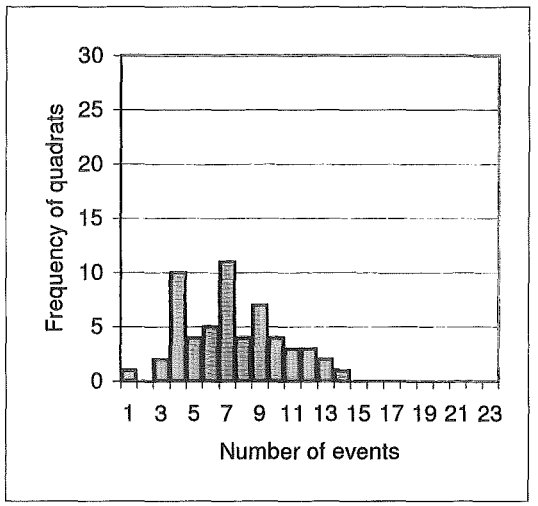


(g) PMC for point cluster distribution (Fig. 7.30d)

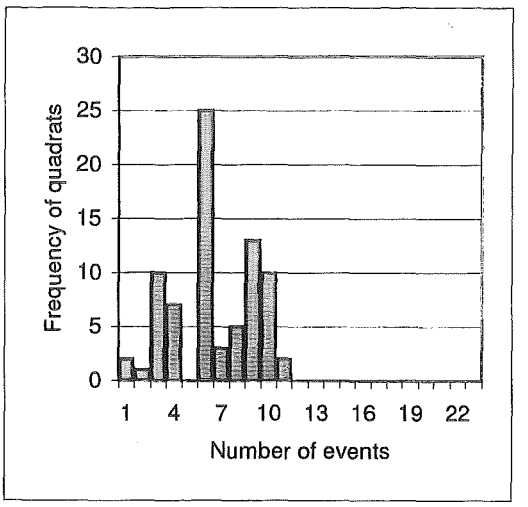


(h) Skewness and kurtosis for point cluster distribution (Fig. 7.30d)

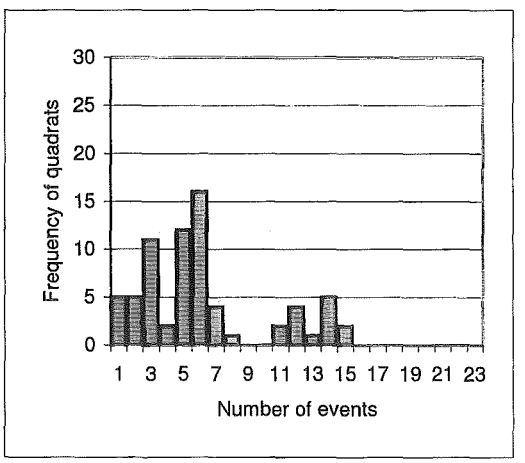
Fig. 7.51: Variation of PMC, skewness and kurtosis calculated for the four distributions (Figure 7.30)



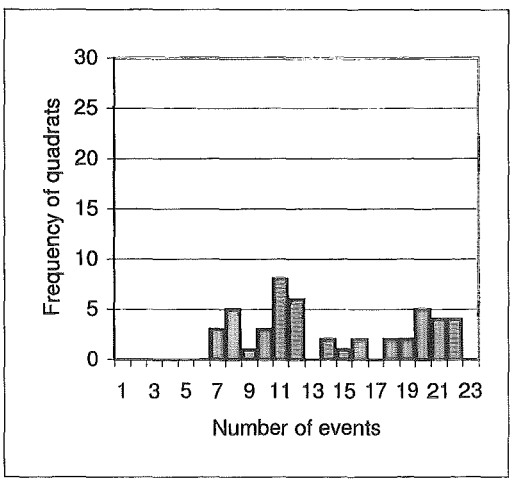
(a) Plotted from the distribution shown in Fig. 7.29a
usin quadrat radius of 30 units.



(a) Plotted from the distribution shown in Fig. 7.29d
usin quadrat radius of 17 units.

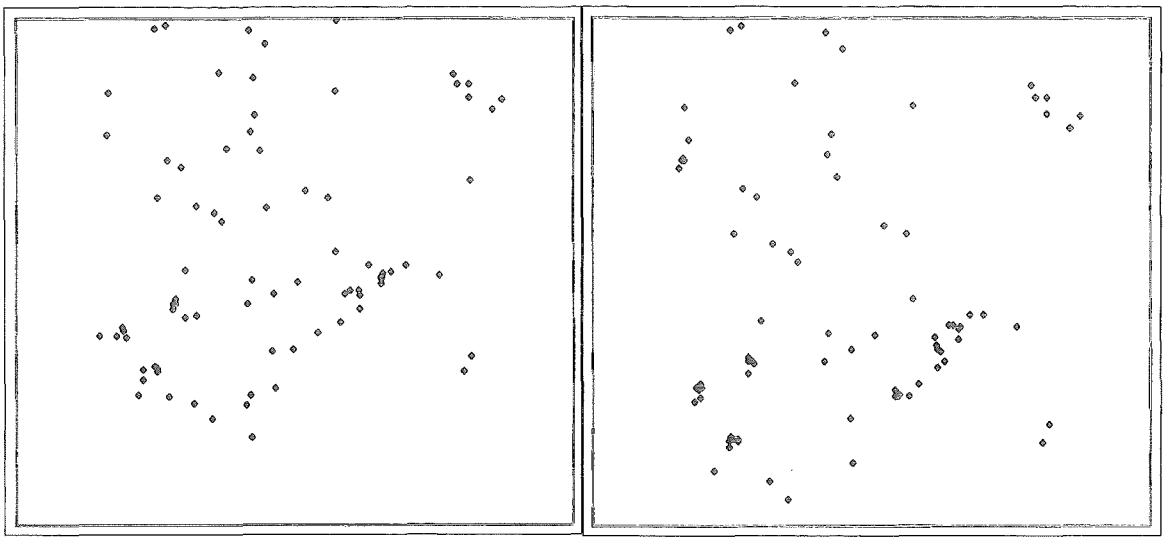


(c) Plotted from the distribution shown in Fig. 7.30a
usin quadrat radius of 20 units.



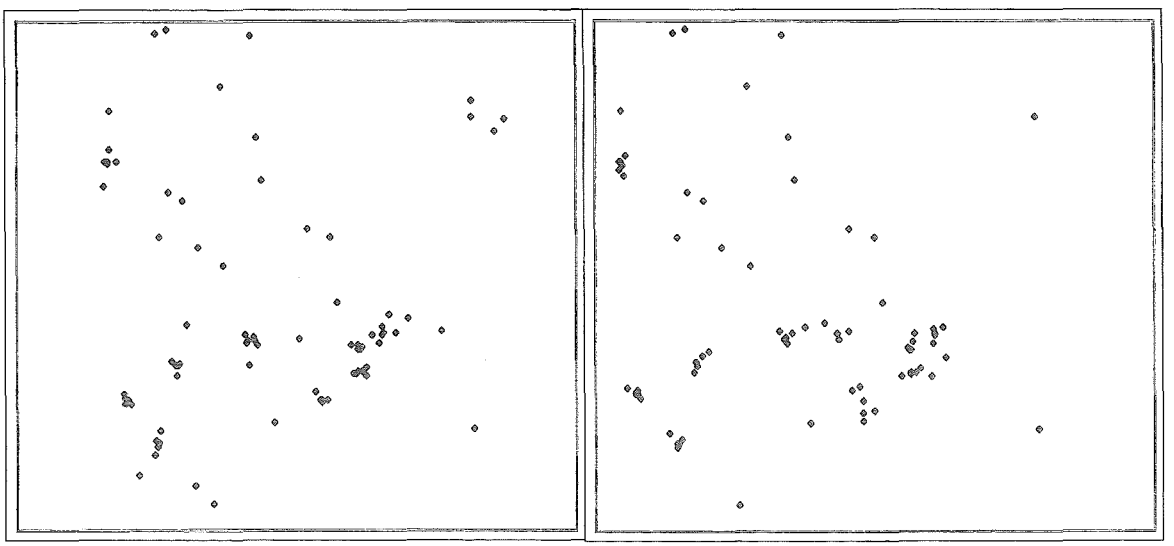
(d) Plotted from the distribution shown in Fig. 7.30a
usin quadrat radius of 40 units.

Figure 7.52: Frequency polygons from different quadrat radii.



(a) Distribution - 1
(70 events from CSR and 30 events
from point cluster distributions)

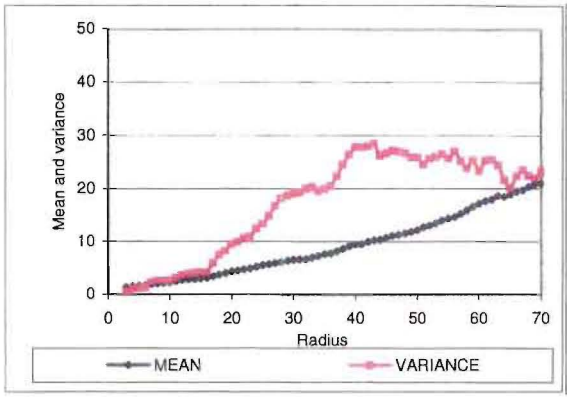
(b) Distribution - 2
(60 events from CSR and 40 events
from point cluster distributions)



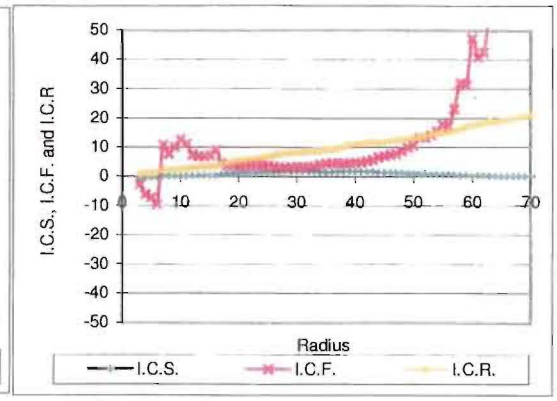
(c) Distribution - 3
(50 events from CSR and 50 events
from point cluster distributions)

(d) Distribution - 4
(40 events from CSR and 60 events
from point cluster distributions)

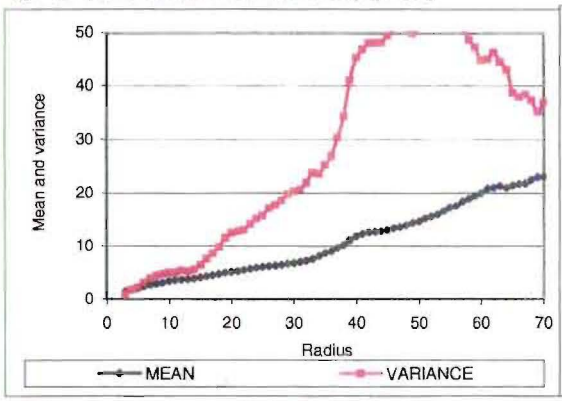
Figure 7.53: Location plot for four mixed distributions (CSR and point cluster are mixed).



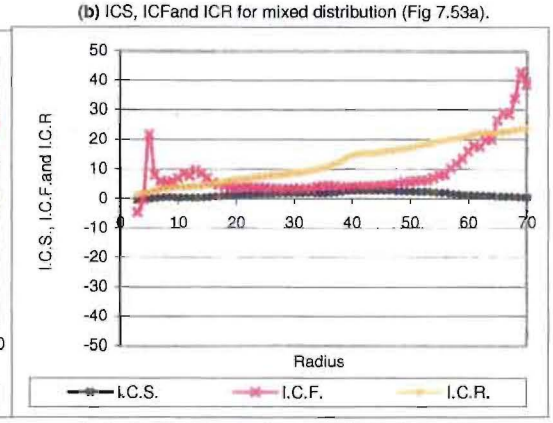
(a) Mean and variance for mixed distribution (Fig 7.53a)



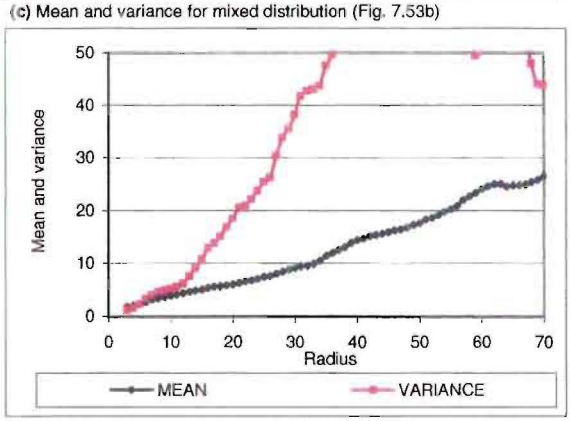
(b) ICS, ICFand ICR for mixed distribution (Fig 7.53a).



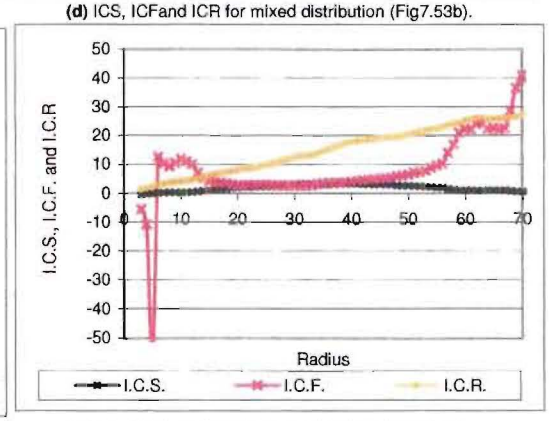
(c) Mean and variance for mixed distribution (Fig. 7.53b)



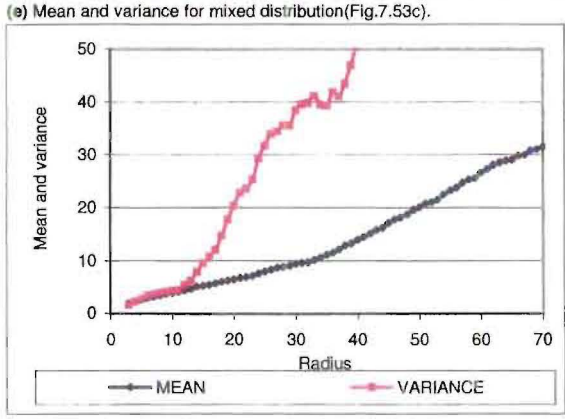
(d) ICS, ICFand ICR for mixed distribution (Fig7.53b).



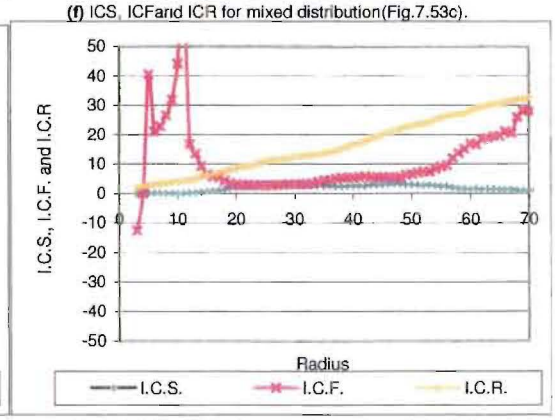
(e) Mean and variance for mixed distribution(Fig.7.53c).



(f) ICS, ICFand ICR for mixed distribution(Fig.7.53c).

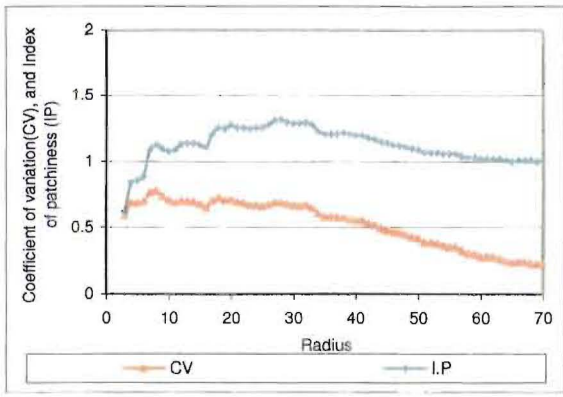


(g) Mean and variance for mixed distribution (Fig. 7.53d).

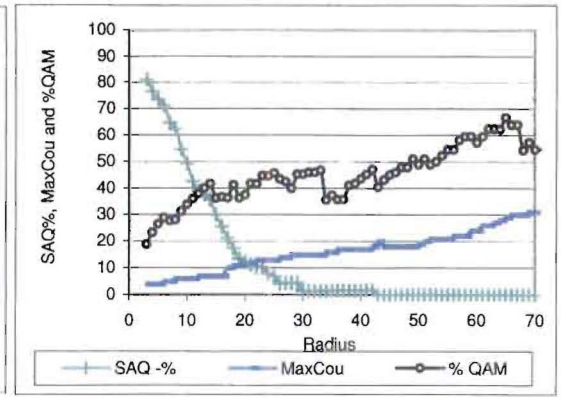


(h) ICS, ICFand ICR for mixed distribution (Fig. 7.53d).

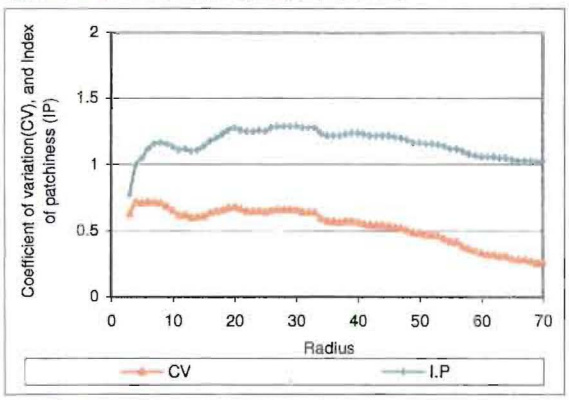
Fig. 7.54: Variation of mean, variance, ICS, ICF and ICR calculated for the four distributions (Figure 7.53)



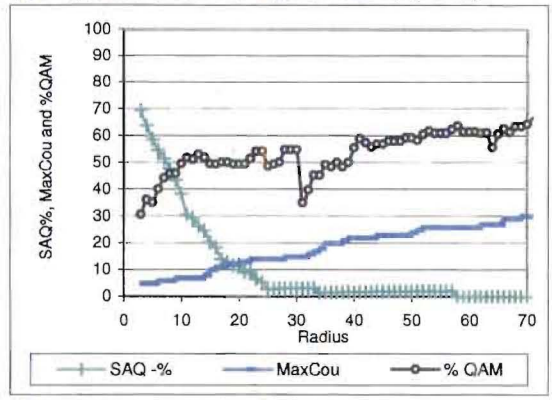
(a) CV, IP and MI for mixed distribution (Fig. 7.53a).



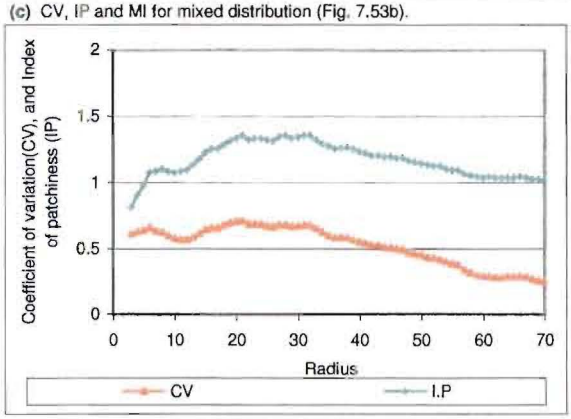
(b) SAQ,MaxCou and QAM for mixed distribution (Fig. 7.53a).



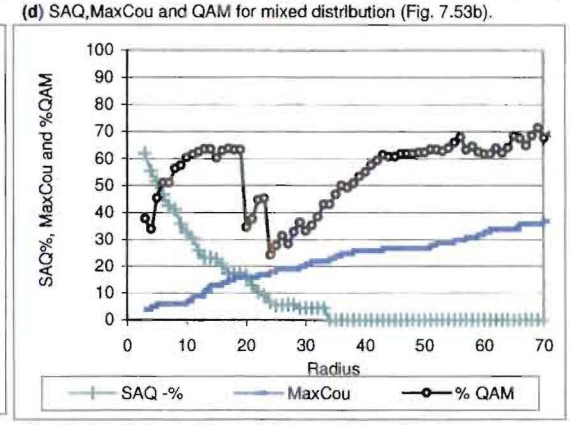
(c) CV, IP and MI for mixed distribution (Fig. 7.53b).



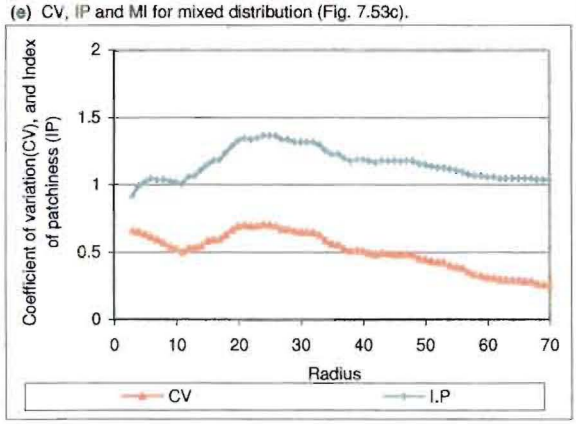
(d) SAQ,MaxCou and QAM for mixed distribution (Fig. 7.53b).



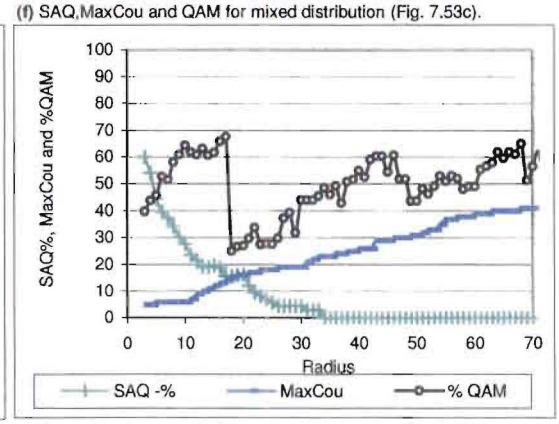
(e) CV, IP and MI for mixed distribution (Fig. 7.53c).



(f) SAQ,MaxCou and QAM for mixed distribution (Fig. 7.53c).

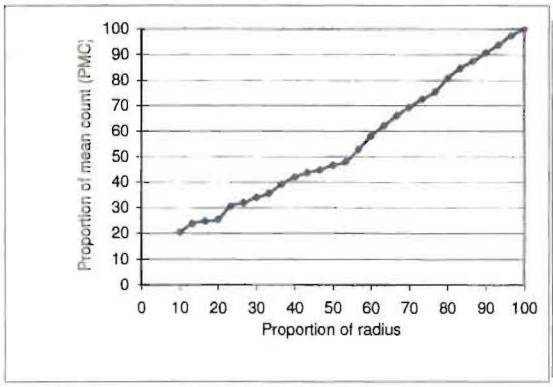


(g) CV, IP and MI for mixed distribution (Fig. 7.53d).

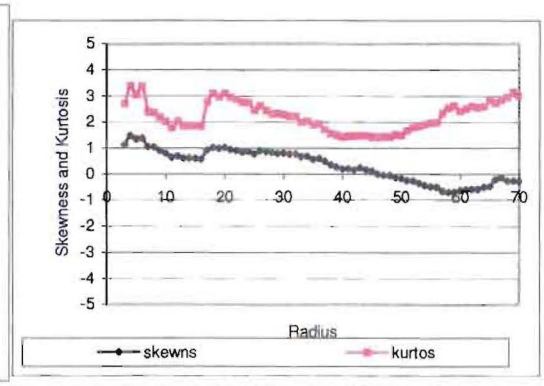


(h) SAQ, MaxCou and QAM for mixed distribution (Fig. 7.53d).

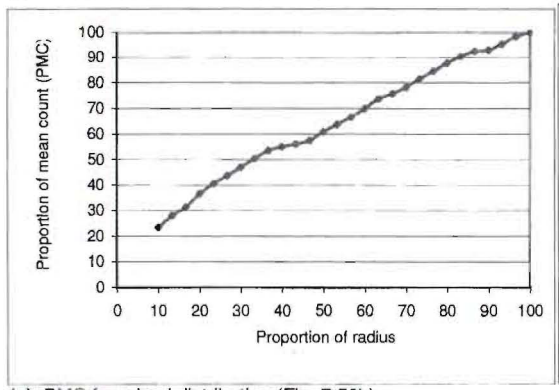
Fig. 7.55: Variation of CV, IP, MI, SAQ%, MaxCou and %QAM calculated for the four distributions (Figures 7.53)



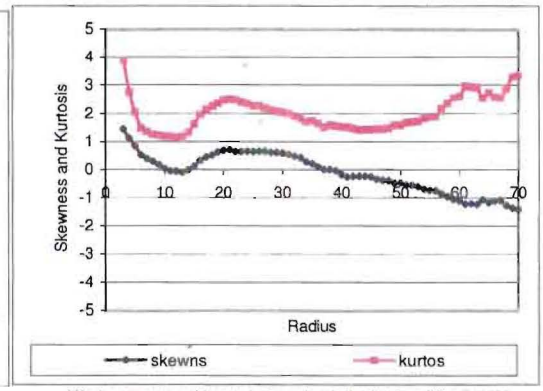
(a) PMC for mixed distribution (Fig. 7.53a).



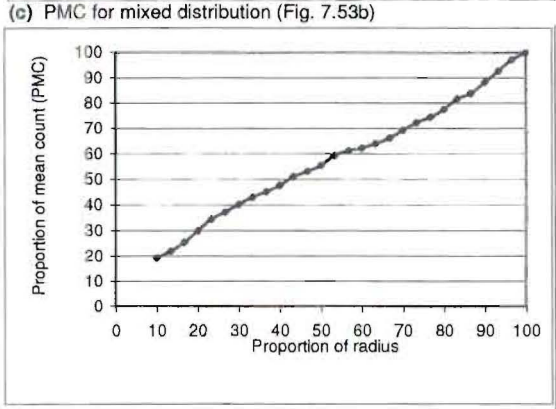
(b) skewness and kurtosis for mixed distribution (Fig. 7.53a)



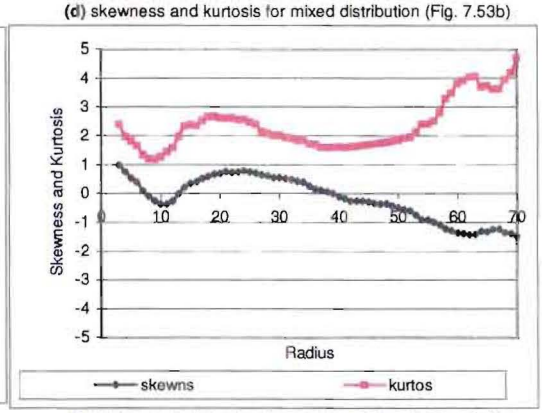
(c) PMC for mixed distribution (Fig. 7.53b)



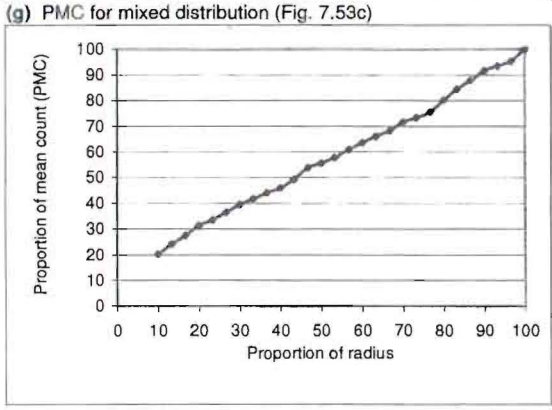
(d) skewness and kurtosis for mixed distribution (Fig. 7.53b)



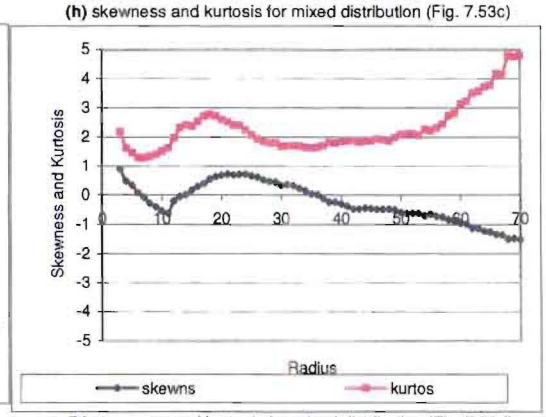
(g) PMC for mixed distribution (Fig. 7.53c)



(h) skewness and kurtosis for mixed distribution (Fig. 7.53c)

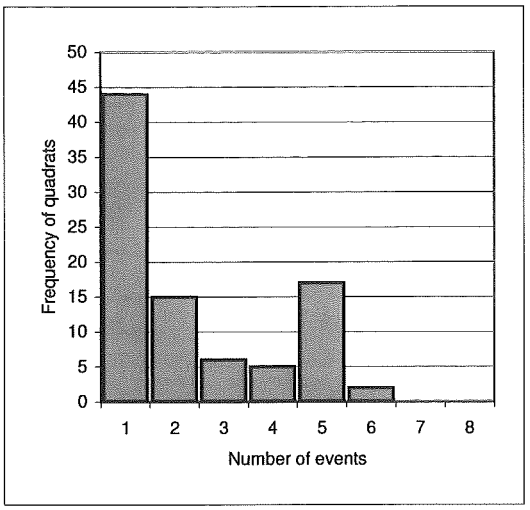


(g) PMC for mixed distribution (Fig. 7.53d)

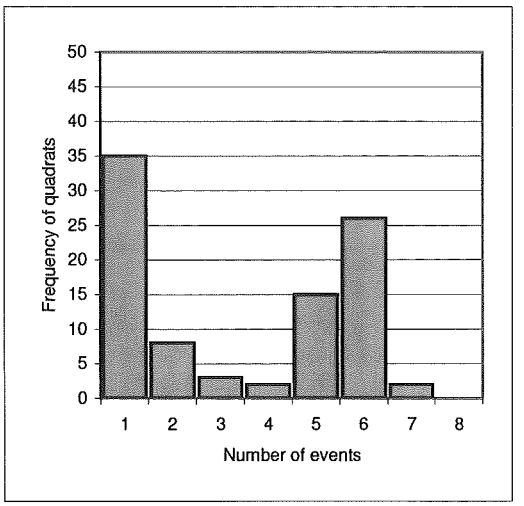


(h) skewness and kurtosis for mixed distribution (Fig. 7.53d)

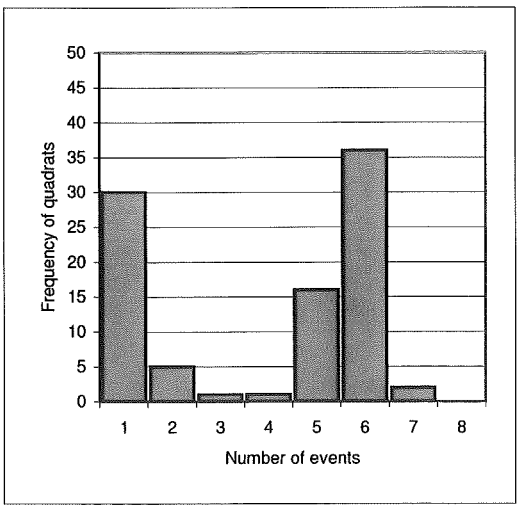
Fig. 7.56: Variation of PMC, skewness and kurtosis calculated for the four distributions (Figure 7.53)



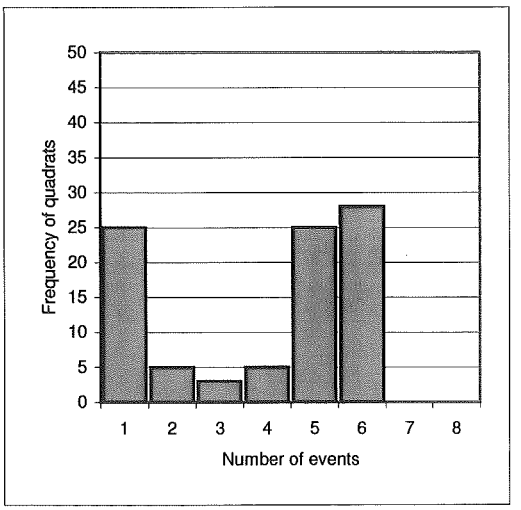
(a) Plotted from the distribution shown in Fig. 7.53a



(b) Plotted from the distribution shown in Fig. 7.53b



(c) Plotted from the distribution shown in Fig. 7.53c



(d) Plotted from the distribution shown in Fig. 7.53d

Fig. 7.57: Frequency polygon for the 10 units radius of quadrat counts for the mixed distributions (Fig. 7.53).

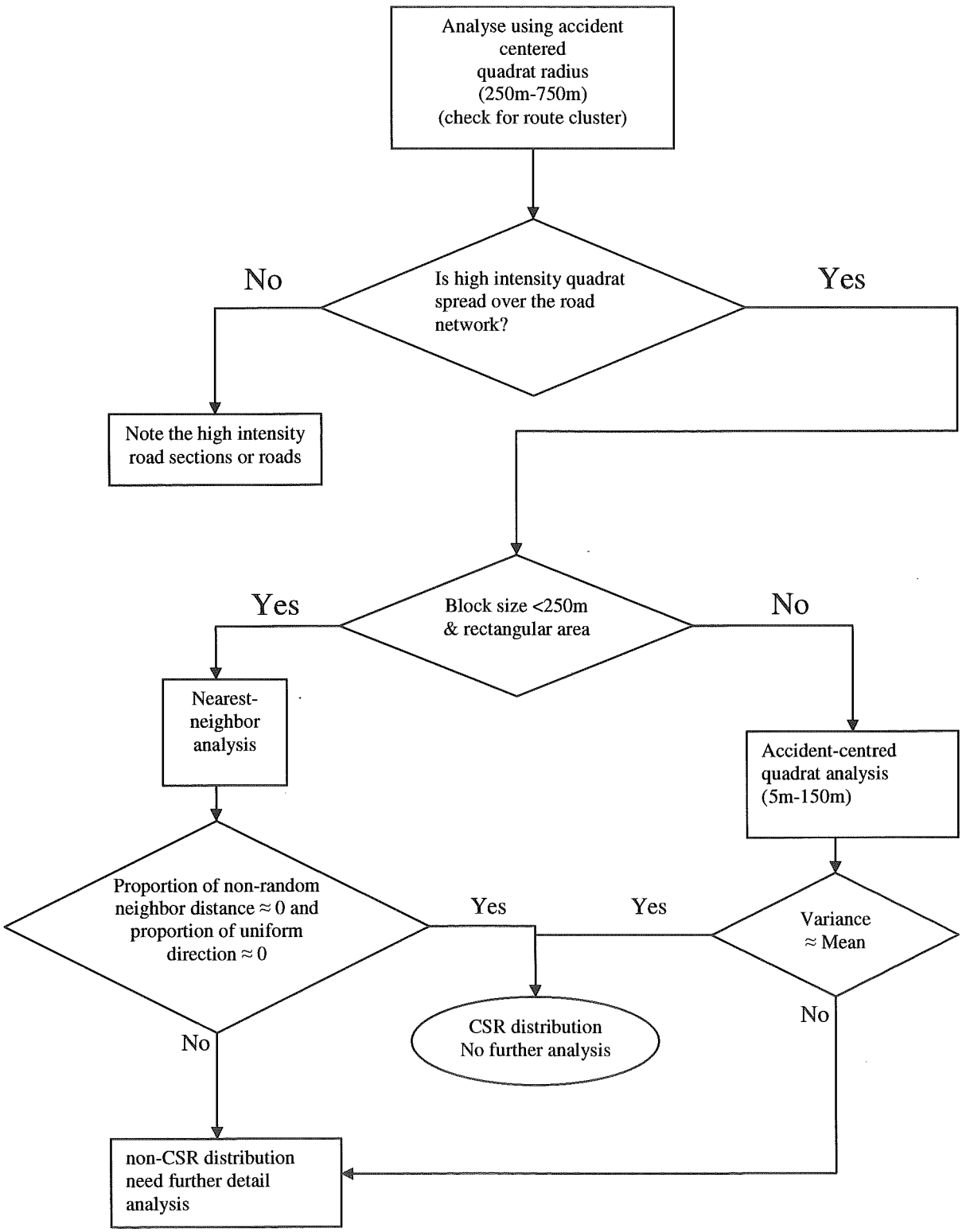


Figure 7.58: Flowchart to identify accident distributions

Chapter 8

ANALYSIS RESULTS FOR ACTUAL ACCIDENT DISTRIBUTIONS

8.1 Description of data

Two areas of the road network in Christchurch (New Zealand) were selected for analysis. They were the Central Business District (CBD) and the suburb of Riccarton. Although the CBD area and the Riccarton areas are adjacent parts of the network, it is important to consider these separately as the two areas have different traffic road environment characteristics. If they are analyses together, the CBD area may appear as an area cluster, as discussed in Section 2.2.2. The road network in the CBD is more regular and dense than the Riccarton network. The two road networks are shown in Figures 8.01a and b. The two selected areas are rectangular, so that each can be surrounded by identical patterns or a buffer zone (i.e. the edge correction methods explained in Chapter 2 can be used).

In these two road networks, the four five-year periods 1982-1986, 1987-1991, 1992-1996 and 1997-2001 were selected for analysis. The reason for selecting the five-year periods was to minimise the effect of short-term temporal variations in accident occurrence and to obtain adequate data to produce statistically reliable results. The four five-year periods will help to identify the progress of changes in accident clustering in the 20-year period. The accident data, obtained from the Land Transport Safety Authority, were for fatal and injury accidents, which normally cost more than non-injury accidents. The accident location plots for the two road networks are shown in Figures 8.02 and 8.03.

8.2 Data scanning

The accident data collected from the LTSA were scanned for simple errors, such as identifying accident locations outside the selected area, using the maximum and minimum of the coordinates of the accident data (also used for edge correction), and the distance between the accident location (on the first street) and the second street, where the first street means the street on which the accident is located and the second street means the cross road

nearest to that accident location. A program was developed to do this scanning in Microsoft Excel. In addition manual scanning was also done, as explained below.

The first street or road names were used to identify the high intensity road or road sections. It was necessary to modify the accident data file for the CBD area, because instead of the street names a road number had been used. The two streets Barbadoes Street and Madras Street are identified as SH74, because these two one-way roads are part of State Highway 74. When a common identifier is used for these two roads, then the quadrat count will be counted from both roads regardless of the quadrat-centre, if the two road sections come within the quadrat. If this happened then the analysis results might indicate SH74 as having a high number of accidents, because the count for the quadrat-centred on one roads might include accidents on the other road. Therefore, the data were modified by entering the road name instead of the road number.

8.3 Computer programs for analysis

The accident data were analysed using three computer programs, based on the techniques described in Chapter 7.

The first program uses nearest-neighbour techniques and is called Nearest Neighbour Analysis Techniques (NNAT). This program was used in Chapter 7 and the technique was explained in Chapter 5. The program NNAT assumes that the selected area is surrounded by eight identical accident distributions, and is one of the traditional edge correction methods. NNAT can be used only for a rectangular road network, which is relatively dense (block size < 100m), but is of limited reliability in the range 100 to 250m, and is unreliable for block size > 250m.

The second program uses the buffer zone method for the edge correction. The road network can be any shape and the road network can be dense or sparse. The program consists of two parts: the Black Spot Analysis Techniques Using Quadrat Method (BSATUQM), which is the same program used in Chapter 7, and the other part is a modified version of BSATUQM called Black Route Analysis Techniques Using Quadrat Method (BRATUQM). The computer programs BSATUQM and BRATUQM were developed to analyse the traffic accident data using the accident-centred quadrat method, as explained in Chapter 6.

The program NNAT uses the input data as a two dimensional matrix of accident location co-ordinates, as explained in Chapter 5. The program BSATUQM also uses the same input data for analysis. The program BRATUQM needs additional information. They are the first and second street names of each accident location and the distance between the second street and the accident location. There are provisions available in BSATUQM and BRATUQM to analyse risk and cost density (as discussed in Chapter 2). For the analysis of risk and cost density, additional data (such as the cost of each accident and the traffic volume at each accident location) are necessary, but these are not included in this study.

The characteristic of a black route was discussed in Chapter 1. As mentioned in IHT [1986] a road having more than the average number of accidents indicates a black route for that class of road. The program BRATUQM gives an output file called "HighDensityRoad.out", which is developed from the input data file. This output file consists of the coordinates of accidents in the road section which has the highest accident count and the file helps to identify roads or road sections which have a high number of accidents within a road network. If the same roads or section of roads have a high number of accidents for different quadrat radii (in the range from 250m to 2000m say) and the roads or road sections which have high number of accidents are not spread over the road network, then those roads are indicated as line clusters.

Another output file from the program (BRATUQM) is called "RdClust.dto". This file contains six columns:

- the first column shows the quadrat radius;
- the second column shows the information about the highest accident count quadrat, such as the serial number of the accident where the quadrat centre was located, the intensity (defined as 100 × the number of accidents per metre of quadrat diameter) and the first street name;
- the third, fourth, fifth and sixth columns show the information (i.e. quadrat centre, intensity, first street name) about the second, third, fourth and fifth highest accident count quadrat respectively.

The file "RdClust.dto" is used to investigate the five highest accident count quadrat centres and to check whether they are along a single road or spread over several roads. This analysis was repeated for several quadrat radii. If the highest accident count quadrat centres are

spread along one or more (user defined number) of road sections, but are not spread over the road network, then this is an indication of "user defined number" of line clusters. An example of an output file is given in Section 8.4 with a more detailed explanation.

If high intensity quadrats are spread over several roads which are close to each other, then it indicates that the part of the road network is an area cluster. If high intensity quadrats are spread over several roads or road sections but are not within a single part of the road network, then the network may have point clusters or an area cluster. It should be verified using the following steps:

(1) analysing the accident data using the program BSATUQM or NNAT will help to identify a point cluster distribution;

(2) analysing the accident data using BRATUQM for extended boundaries of the previously analysed road network area (the previously analysed road network area is part of an area cluster), to help to confirm that there is no line cluster.

BRATUQM was developed for quadrat radii greater than 150 m, and is not suitable for black spot analysis. A small quadrat is helpful for identifying a CSR distribution. The results from BSATUQM for a CSR distribution were discussed in Sections 7.4.1 and 7.4.4. The BRATUQM was developed for large quadrats and is not sufficiently accurate to identify a CSR distribution, as was discussed in Section 7.4.4. The quadrat count distribution obtained from BRATUQM is not suitable for analysing the mean and the variance count.

The programs NNAT or BSATUQM help to identify CSR, point cluster or regular distributions, as described in Chapter 7. If the accidents are clustered at a single location, or over a short length of road with an above average number of accidents, then these sites will be black spots. The program NNAT indicates the proportion of events having cluster, regular and random neighbour distance distribution. The program BSATUQM indicates the variance and the mean quadrat count for different quadrat radii. This helps to distinguish between CSR, point cluster or regular accident distributions. Other indices, which are helpful for more detailed analysis, were discussed in Chapter 7.

The program must be selected according to the road network characteristics (eg. rectangular or dense). NNAT is reliable if the roads are dense (block size < 100m), is of limited reliability in the range 100 to 250m, and is unreliable for block size > 250m. The program

BSATUQM selects quadrats centred on accidents, and hence the space between roads will not affect the analysis results, and the road network can be dense or sparse.

Figure 8.04 is a flowchart, which shows the necessary steps and the name of the program to be used at each step. There are two numbered lines in Figure 8.04. The portion of the flowchart enclosed within the line number 1 is described further in Sections 8.4 and 8.5 using the two case studies (regular and irregular road network) mentioned in Section 8.1. The portion of the flowchart enclosed within the line number 2 is discussed in this Section.

8.4 CBD accident data analysis results

Accident data from the CBD area were analysed using the program BRATUQM. The output file "RdClust.dto", for quadrat radius from 205m to 505m, is tabulated in Table 8.01. From the table, for quadrat radius of 245m centred on Manchester Street, the accident intensity is 7.76 accidents per 100m for the 5 years period (1997-2001) i.e. 15.52 accidents per km per year. Table 8.01 shows that high intensity quadrats are spread along sections of Manchester Street and Colombo Street.

The output file "HighAccRd.out" from the program BRATUQM gives the accident locations contributing to the high accident counts. Figure 8.05 shows four sets of data, for the periods 1982-1986, 1987-1991, 1992-1996 and 1997-2001. Figure 8.05 indicates route clustering but still it is necessary to confirm that the CBD accident data is not a CSR distribution. The procedure is shown in the flowchart (Figure 8.04)

The selected CBD road network shown in Figure 8.01a is a square area. Nicholson [1995] noted that " the typical block size for the road network in the Christchurch (NZ) Central Business District is 110m by 220m". This block size is less than 250m, which is small enough that we can use the program NNAT.

AREA=CITYCENTRE1997-2001

radius	cent-# AcINT	cent-# AcINT	cent-# AcINT	cent-# AcINT	cent-# AcINT
205.	111. 6.83	134. 6.83	330. 6.34	331. 6.34	333. 6.34
	COLOMBO S	COLOMBO S	MANCHESTE	MANCHESTE	MANCHESTE
215.	111. 6.51	134. 6.51	129. 6.28	93. 6.05	135. 6.05
	COLOMBO S	COLOMBO S	COLOMBO S	COLOMBO S	COLOMBO S
225.	110. 6.67	111. 6.22	128. 6.22	129. 6.22	134. 6.22
	COLOMBO S	COLOMBO S	COLOMBO S	COLOMBO S	COLOMBO S
235.	110. 6.60	111. 6.38	135. 6.17	136. 6.17	137. 6.17
	COLOMBO S	COLOMBO S	COLOMBO S	COLOMBO S	COLOMBO S
245.	354. 7.76	355. 7.76	112. 6.94	113. 6.94	135. 6.53
	MANCHESTE	MANCHESTE	COLOMBO S	COLOMBO S	COLOMBO S
255.	354. 7.45	355. 7.45	112. 6.67	113. 6.67	131. 6.27
	MANCHESTE	MANCHESTE	COLOMBO S	COLOMBO S	COLOMBO S
265.	353. 7.17	354. 7.17	355. 7.17	110. 6.42	112. 6.42
	MANCHESTE	MANCHESTE	MANCHESTE	COLOMBO S	COLOMBO S
275.	353. 6.91	354. 6.91	355. 6.91	327. 6.73	110. 6.18
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	COLOMBO S
285.	327. 6.67	353. 6.67	354. 6.67	355. 6.67	110. 6.32
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	COLOMBO S
295.	353. 6.44	354. 6.44	355. 6.44	110. 6.10	111. 6.10
	MANCHESTE	MANCHESTE	MANCHESTE	COLOMBO S	COLOMBO S
305.	330. 7.05	331. 7.05	134. 6.23	353. 6.23	354. 6.23
	MANCHESTE	MANCHESTE	COLOMBO S	MANCHESTE	MANCHESTE
315.	330. 6.98	331. 6.98	134. 6.03	333. 6.03	353. 6.03
	MANCHESTE	MANCHESTE	COLOMBO S	MANCHESTE	MANCHESTE
325.	330. 6.77	331. 6.77	111. 5.85	134. 5.85	333. 5.85
	MANCHESTE	MANCHESTE	COLOMBO S	COLOMBO S	MANCHESTE
335.	330. 6.57	331. 6.57	353. 5.97	111. 5.67	129. 5.67
	MANCHESTE	MANCHESTE	MANCHESTE	COLOMBO S	COLOMBO S
345.	330. 6.38	331. 6.38	353. 6.23	387. 5.65	110. 5.51
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	COLOMBO S
355.	330. 6.20	331. 6.20	353. 6.20	372. 5.92	354. 5.63
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
365.	334. 6.85	335. 6.85	330. 6.03	331. 6.03	353. 6.03
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
375.	334. 6.67	335. 6.67	330. 5.87	331. 5.87	353. 5.87
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
385.	334. 6.49	335. 6.49	330. 5.84	331. 5.84	353. 5.71
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
395.	334. 6.33	335. 6.33	330. 5.70	331. 5.70	353. 5.57
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
405.	334. 6.17	335. 6.17	330. 5.56	331. 5.56	135. 5.43
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	COLOMBO S
415.	333. 6.02	334. 6.02	335. 6.02	330. 5.42	331. 5.42
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
425.	330. 5.88	331. 5.88	333. 5.88	334. 5.88	335. 5.88
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
435.	333. 5.86	330. 5.75	331. 5.75	334. 5.75	335. 5.75
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
445.	333. 5.84	360. 5.73	361. 5.73	330. 5.62	331. 5.62
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
455.	333. 5.71	360. 5.60	361. 5.60	330. 5.49	331. 5.49
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
465.	345. 5.81	333. 5.59	360. 5.48	361. 5.48	330. 5.38
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
475.	344. 5.68	345. 5.68	333. 5.47	353. 5.37	360. 5.37
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
485.	333. 5.57	344. 5.57	345. 5.57	346. 5.57	347. 5.57
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
495.	328. 5.56	346. 5.56	347. 5.56	348. 5.56	349. 5.56
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE
505.	346. 5.54	347. 5.54	348. 5.54	349. 5.54	350. 5.54
	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE	MANCHESTE

Cent-# -Serial number for the accident location where the quadrat centre located
AcINT --100Xnumber of accidents per meter of quadrat diameter

Table 8.01: CBD top accident count quadrats for various quadrat radii

Accident data from the CBD area were analysed using the program nearest-neighbour analysis method (NNAT) and the accident centred quadrat method. The nearest-neighbour distance analysis results are plotted in Figures 8.06 a, c, e and g for the periods 1982-1986, 1987-1991, 1992-1996 and 1997-2001 respectively. These plots indicate that there is no evidence of a CSR distribution for the four sets of five-year periods, and that there is no evidence of a high proportion of equal sized clusters. If the accident distribution was CSR, then the plot should be very similar to the plot shown in Figure 7.24e or 7.33a. If the accident distribution had a high proportion of constant size clusters, then the plot should be very similar to the peak-and-trough pattern shown in Figure 7.25e.

The nearest-neighbour direction analysis results are plotted in Figures 8.06 b, d, f and h for the periods 1982-1986, 1987-1991, 1992-1996 and 1997-2001 respectively. These four plots indicate that the distributions of directions to nearest-neighbours are not uniform. The nearest-neighbour results therefore confirm that there is no indication of CSR distribution.

Figures 8.06a, c, e and g show that about 95% of events have a non-random nearest-neighbour distance distribution, about 65% of events have clustered distance distributions, and about 30% of events have a regular distance distribution. Figures 8.06b, d, f and h indicate about 95% of events have a non-uniform nearest-neighbour direction distribution. The differences between the four figures are not substantial (i.e. the spatial distributions for the four periods are fairly similar).

The accident data were further analysed using the accident-centred quadrat method. The mean and variance calculated from the quadrat counts were plotted against the quadrat radius and are shown in Figures 8.07 a, c, e and g for the periods 1982-1986, 1987-1991, 1992-1996 and 1997-2001 respectively. On average the highest difference between the mean and the variance are noted for the period 1982-1986, the second highest is for the period 1992-1996, the third highest is for the period 1987-1991 and the lowest difference is for the period 1997-2001. In all these four plots the variance is greater than the mean. If the distribution is CSR then the mean is expected to be only slightly greater than the variance, as shown in Figure 7.37a and Figure 7.42a. The results from the quadrat method therefore do not indicate a CSR distribution.

Figure 8.07g indicates that when the quadrat radius is in the range 70m to 85m, the difference between the mean and the variance is relatively small, with the variance being slightly greater than the mean. As the variance becomes smaller the variation in counts decreases (i.e., difference between quadrats counts becomes smaller).

The proportion of single accident quadrats, the maximum count, and the proportion of quadrat counts above the mean, are plotted against quadrat radius in Figures 8.07 b, d, f and h respectively for the periods 1982-1986, 1987-1991, 1992-1996 and 1997-2001. Figure 8.07h shows that about 45 % of quadrat counts are above the mean. Figure 8.07g shows that the mean quadrat count is 8 for the quadrat radius of 70m. The maximum count for a 70m quadrat radius has been reduced from 52 to 16 during the period 1982-2001, and the proportion of single-accident quadrats has increased from 2.46 to 8.12. These facts, along with the ratio between the mean and the variance, indicate route clustering or area clustering rather than point clustering. The result from the program (BRATUQM) already indicated route clustering and that route action plans are appropriate.

8.5 Riccarton suburb accident data analysis results

Some of the block sizes in the Riccarton suburb are more than 250m and NNAT was therefore not used. The accident data for the Riccarton suburb were analysed using BRATUQM and BSATUQM. The locations of the accidents are shown in Figure 8.03. The BRATUQM output files RdClust.dto and "HighAccRd.out" are shown in Table 8.02 and Figure 8.08 respectively. From Table 8.02, for a quadrat radius of 265m centred on Riccarton Road, the accident intensity was 3.58 accidents per 100m for the 5 year period (1997-2001) i.e. 7.16 accidents / km / year.

The analysis results from BRATUQM shows route clusters. The centres of the quadrats with high accident count were located on sections of Riccarton Road. The locations of accident within these quadrats are shown in Figure 8.08. A quadrat centre located in Clarence Road had the 5th highest accident count during the period 1992 - 1996 and for this reason both the Riccarton and Clarence Roads are highlighted in Figure 8.08c. The highest accident count quadrat centres were located in Matipo Street and Riccarton Road during the period 1997 – 2001 (see Figure 8.08d).

AREA=Riccarton1997-2001

radius	cent-# AcINT	cent-# AcINT	cent-# AcINT	cent-# AcINT	cent-# AcINT
205.	181. 3.17	158. 2.93	159. 2.93	160. 2.93	161. 2.93
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
215.	181. 3.02	24. 2.79	25. 2.79	26. 2.79	158. 2.79
	RICCARTON	CLARENCE	CLARENCE	CLARENCE	RICCARTON
225.	181. 3.11	24. 2.67	25. 2.67	26. 2.67	156. 2.67
	RICCARTON	CLARENCE	CLARENCE	CLARENCE	RICCARTON
235.	181. 3.19	116. 2.98	117. 2.98	118. 2.98	157. 2.77
	RICCARTON	MATIPO ST	MATIPO ST	MATIPO ST	RICCARTON
245.	116. 3.27	117. 3.27	118. 3.27	181. 3.06	24. 2.86
	MATIPO ST	MATIPO ST	MATIPO ST	RICCARTON	CLARENCE
255.	114. 3.14	116. 3.14	117. 3.14	118. 3.14	181. 2.94
	MATIPO ST	MATIPO ST	MATIPO ST	MATIPO ST	RICCARTON
265.	143. 3.58	116. 3.21	117. 3.21	118. 3.21	114. 3.02
	RICCARTON	MATIPO ST	MATIPO ST	MATIPO ST	MATIPO ST
275.	143. 3.45	116. 3.09	117. 3.09	118. 3.09	114. 2.91
	RICCARTON	MATIPO ST	MATIPO ST	MATIPO ST	MATIPO ST
285.	143. 3.33	114. 2.98	116. 2.98	117. 2.98	118. 2.98
	RICCARTON	MATIPO ST	MATIPO ST	MATIPO ST	MATIPO ST
295.	143. 3.22	114. 2.88	116. 2.88	117. 2.88	118. 2.88
	RICCARTON	MATIPO ST	MATIPO ST	MATIPO ST	MATIPO ST
305.	143. 3.11	114. 2.79	116. 2.79	117. 2.79	118. 2.79
	RICCARTON	MATIPO ST	MATIPO ST	MATIPO ST	MATIPO ST
315.	142. 3.17	143. 3.02	114. 2.70	116. 2.70	117. 2.70
	RICCARTON	RICCARTON	MATIPO ST	MATIPO ST	MATIPO ST
325.	142. 3.08	143. 2.92	114. 2.77	116. 2.62	117. 2.62
	RICCARTON	RICCARTON	MATIPO ST	MATIPO ST	MATIPO ST
335.	142. 2.99	143. 2.84	156. 2.84	157. 2.84	114. 2.69
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	MATIPO ST
345.	142. 2.90	143. 2.75	156. 2.75	157. 2.75	114. 2.61
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	MATIPO ST
355.	158. 2.96	159. 2.96	160. 2.96	161. 2.96	162. 2.96
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
365.	157. 2.88	158. 2.88	159. 2.88	160. 2.88	161. 2.88
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
375.	156. 2.93	175. 2.93	176. 2.93	142. 2.80	157. 2.80
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
385.	175. 2.99	176. 2.99	156. 2.86	157. 2.86	142. 2.73
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
395.	175. 2.91	176. 2.91	177. 2.91	156. 2.78	157. 2.78
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
405.	175. 2.84	176. 2.84	177. 2.84	156. 2.72	157. 2.72
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
415.	175. 2.77	176. 2.77	177. 2.77	142. 2.65	156. 2.65
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
425.	142. 2.71	175. 2.71	176. 2.71	177. 2.71	156. 2.59
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
435.	142. 2.64	175. 2.64	176. 2.64	177. 2.64	156. 2.53
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
445.	142. 2.58	175. 2.58	176. 2.58	177. 2.58	156. 2.47
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
455.	142. 2.53	175. 2.53	176. 2.53	177. 2.53	156. 2.42
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
465.	142. 2.47	143. 2.47	175. 2.47	176. 2.47	177. 2.47
	RICCARTON	RICCARTON	RICCARTON	RICCARTON	RICCARTON
475.	177. 2.63	120. 2.53	121. 2.53	143. 2.53	142. 2.42
	RICCARTON	MATIPO ST	MATIPO ST	RICCARTON	RICCARTON
485.	177. 2.58	120. 2.47	121. 2.47	143. 2.47	142. 2.37
	RICCARTON	MATIPO ST	MATIPO ST	RICCARTON	RICCARTON
495.	157. 2.53	177. 2.53	120. 2.42	121. 2.42	143. 2.42
	RICCARTON	RICCARTON	MATIPO ST	MATIPO ST	RICCARTON
505.	120. 2.48	121. 2.48	143. 2.48	156. 2.48	157. 2.48
	MATIPO ST	MATIPO ST	RICCARTON	RICCARTON	RICCARTON

Cent-# -Serial number for the accident location where the quadrat centre located
AcINT --100X number of accidents per meter of quadrat diameter

Table 8.02: Riccarton top accident count quadrats for varius quadrat radii

During the periods 1982 - 1986 and 1987 - 1991 almost the same section of Riccarton Road had high accident intensity but during the periods 1992 - 1996 and 1997 - 2001 a different section had high accident intensity. The reasons are:

- during 1982 - 1986 the number of accidents at the Riccarton Road / Matipo Street intersection was less compared to the period 1997 - 2001 and
- during the period 1992 - 1996 the number of accidents at the Riccarton Road/Mandeville Street intersection was greater than the number of accidents during 1997 - 2001.

The analysis results from BSATUQM are plotted in Figure 8.09. The variance to mean ratio decreases from Figures 8.09c, e to g. The results show that the accident distributions for the four periods were not CSR, further confirming the existence of route clustering.

8.6 Visually examining the accident plots

Visually examining the plotted maps (Figures 8.02 and 8.03) is a difficult task because some spots that are marked as a single dot in these figures may contain more than one accident. For example in Figure 8.3d a single dot (i.e., coordinate 2477315m North and 5741848m East) at the Matipo Street and Riccarton Road intersection represents seven accidents, but another single dot (i.e., coordinate 2477823m North and 5741905m East) at the Riccarton Road and Clarence Street intersection represents a single accident. Therefore, visual examination of these types of plots may result in misjudgement.

There are other styles of location plot, such as the plot of rainfall data in Figure 8.10, where the locations are marked as circles, with the diameter of the circle being proportional to the number of accidents at that location. The size (and shading) of the circles can be used to indicate the intensity of accidents. Small circles can be marked on top of big circles, so that small circles are not hidden by big circles. With this method, however, it would still be difficult to identify the appropriate accident reduction program.

These types of plots (Figure 8.10) are used in the Crash Analysis System (CAS) used by the LTSA. For example Figure 1.01 was plotted using CAS and shows the locations of injury crashes in Christchurch, including both the CBD and the Riccarton suburb. In the figure the enlarged portion in the upper right hand corner illustrates the location of injury accidents in the CBD area, and illustrate the difficulty of visual examination of these types of plots.

From visual examination one may conclude that an area action plan is suitable, but the analysis result in Figure 8.05 leads to a different conclusion. Visual examination of Figure 1.01 might well result in the conclusion that the distribution is CSR. The conclusion might be different from person to person attempting to identify whether there is a pattern in the spatial distribution, and the nature of that pattern.

8.7 Christchurch CBD accident distribution for the period (1966-1996)

Comprehensive data for the period (1966-1996) are not available. Douglass [2000] has the plots of accident locations for the CBD for 1966, 1984 and 1996. Figure 8.11a gives the location plot for the 530 injury crashes in 1966, which was before a type of area action plan (i.e. a one-way streets system) was implemented. Figure 8.11b gives the location plot for the 212 injury crashes in 1984, which was after the one way system was implemented. Figure 8.11c gives the location plot for the 148 injury crashes in 1996. The accident distribution has substantially changed over the period (1966-1996). The report noted that "overall the plan has been successful in safety terms, and it has been calculated that the benefit/cost ratio from crashes alone was 30:1".

After the change of some two-way streets to one-way streets, the movement function on those streets was improved and the movement function on the other streets was reduced. This is a type of area traffic plan which will reduce area clustering, and accident reduction was one of the goals, the others being to improve the amenity of the central area by shifting the traffic out onto designated arterial routes. The analysis result in Section 8.4 indicates that the current need is to adopt a route action plan. This might appear contrary to the suggestion by Nicholson [1989 and 1990] that there is a natural progression from site, to route and then to area action plans, but confirms the comment of Nicholson that it is important to examine the spatial distribution of accidents, which can change as a result of accident reduction plans, and to choose an accident reduction plan appropriate to the spatial distribution.

8.8 Conclusions

Two areas in Christchurch, New Zealand were used to demonstrate the application of the nearest-neighbour method and accident-centred quadrat method. The analysis results

indicate the existence of route clusters, but not CSR or point cluster distributions, and hence skewness, kurtosis and frequency polygons were not plotted.

The visual examination of accident plots suggests that the CBD area might initially have been an area cluster, while route clusters appear to have been dominant in the Riccarton suburb throughout. However, the CBD area analysis clearly shows that the accident distribution has changed during the twenty years (1982-2001), with route clusters becoming more obvious. Hence, route action plans appear to be the appropriate approach for both areas.

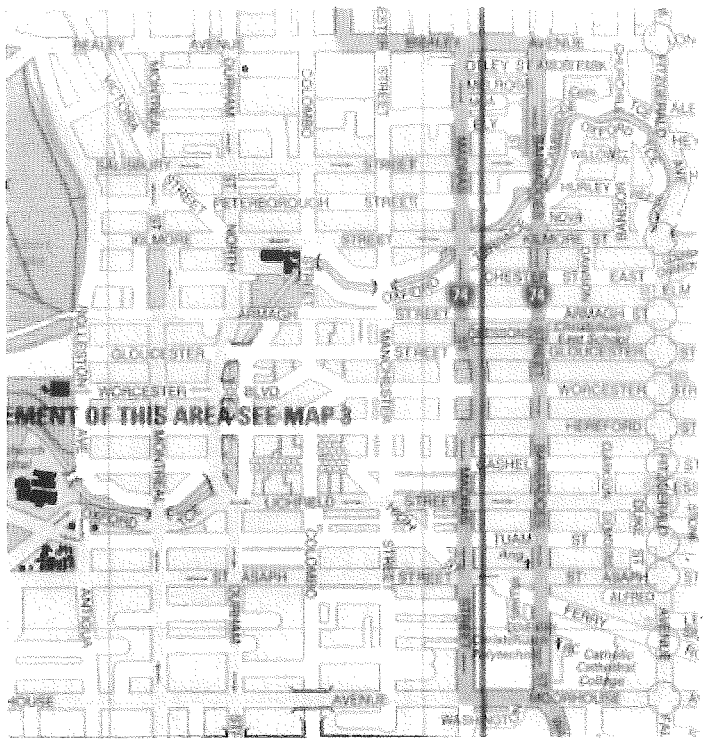
The top two route clusters during the analysis period (1997 – 2001) are:

1. Manchester Street, between Oxford Tce and Tuam Street, for the CBD area;
2. Riccarton Road, between Puriri and Clarence Streets, for the Riccarton suburb.

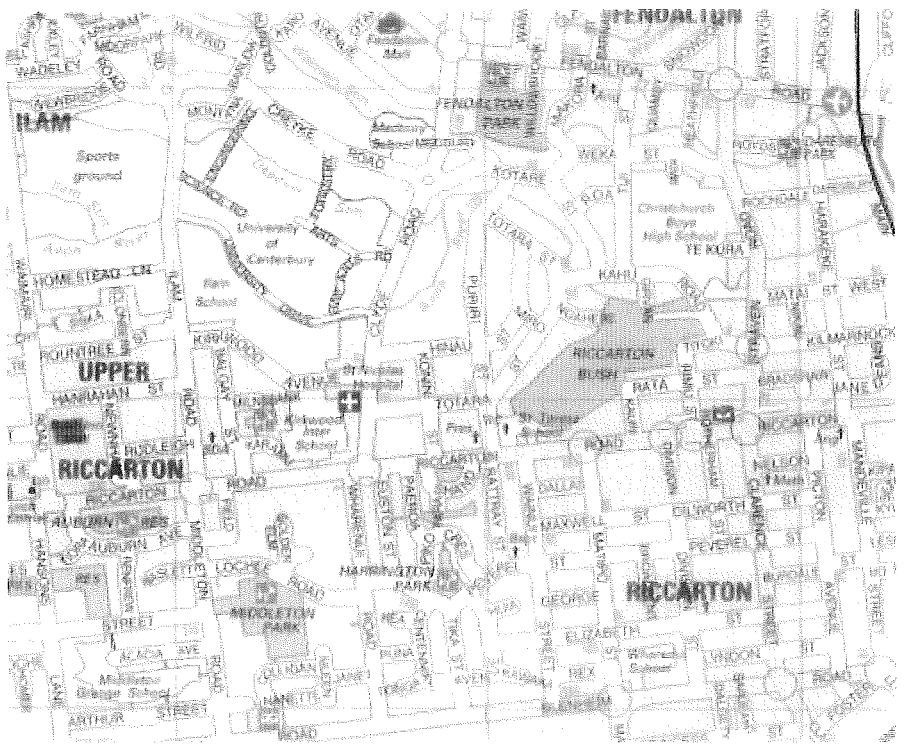
The analysis indicates that the next route cluster is Colombo Street for the CBD area, but for the Riccarton suburb, no other route section was identified as a route cluster.

Douglass [2000] suggested that accidents on Manchester Street were reduced by the one-way system (i.e. Madras and Barbadoes street). However, the analysis results indicate part of the Manchester Street was a route cluster for the analysis period (1997-2001) which was after the introduction of the one way system.

The main problem in the Riccarton route cluster appears to be road functionality. These routes service both the movement and access functions. Detailed analysis of the accidents will help to identify the common factors in the accidents and solutions for accident reduction work.

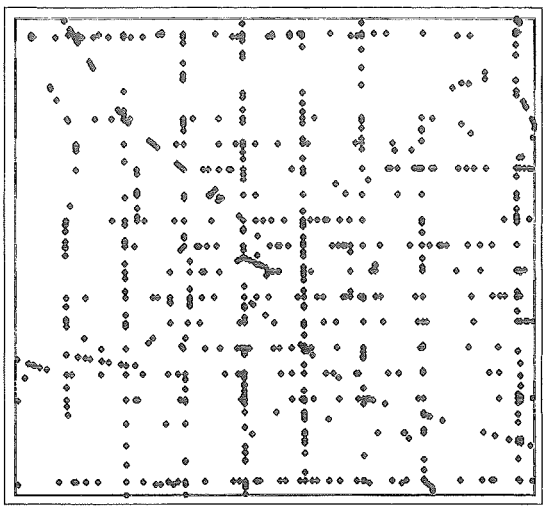


(a) CBD area

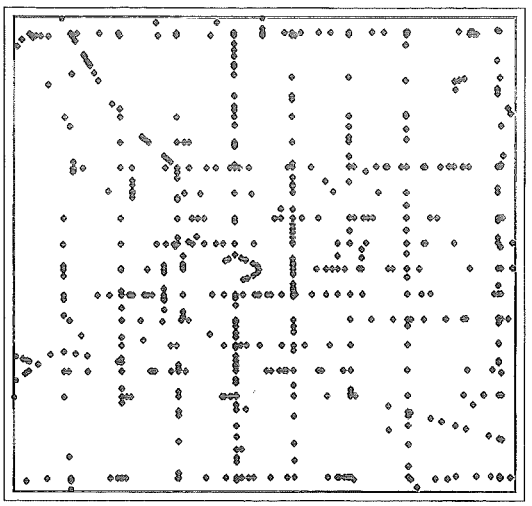


(b) Riccarton suburb area

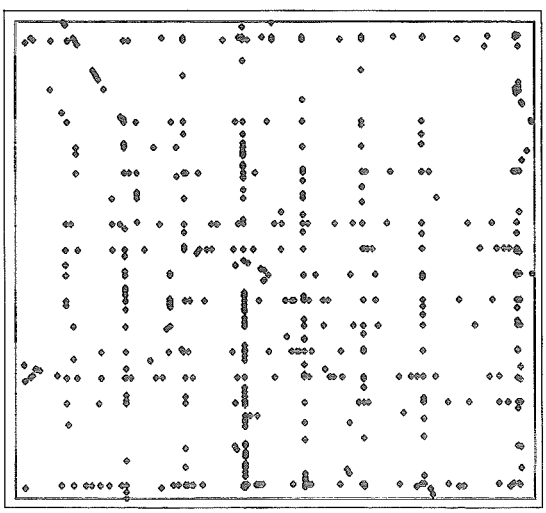
Figure 8.01: Two selected road networks in Christchurch



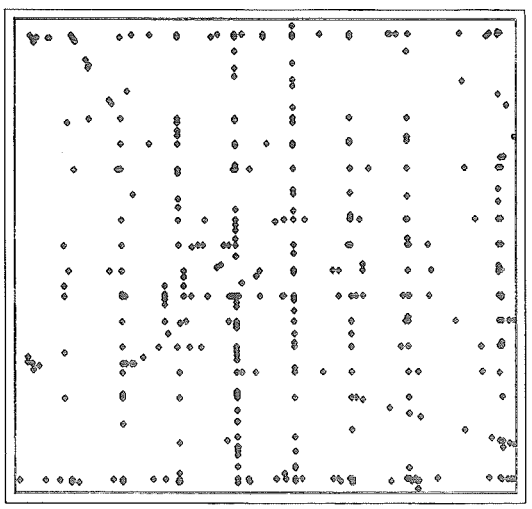
(a) Accident distribution for 1982-1986



(b) Accident distribution for 1987-1991

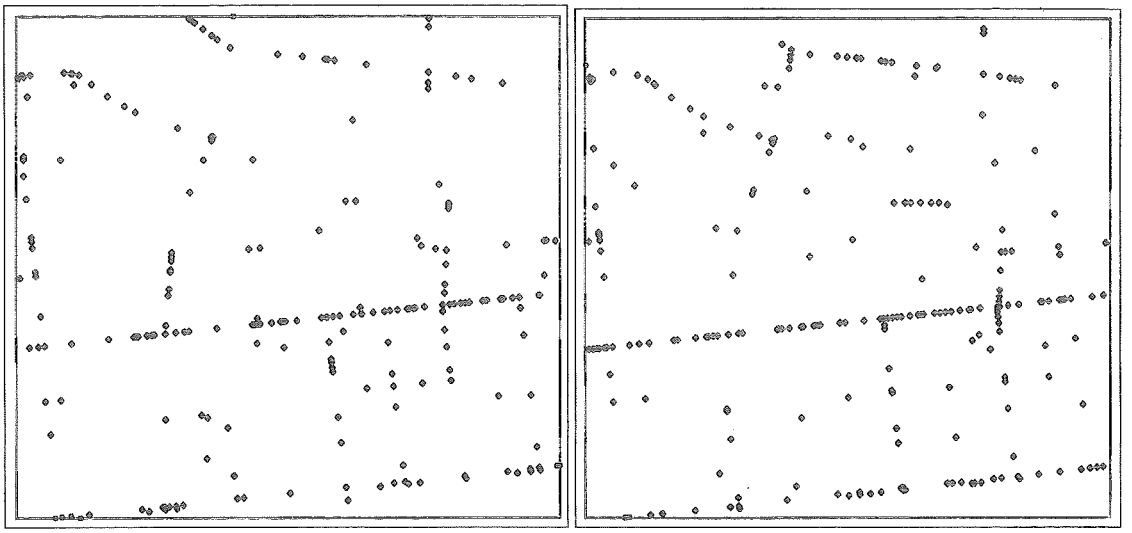


(c) Accident distribution for 1992-1996



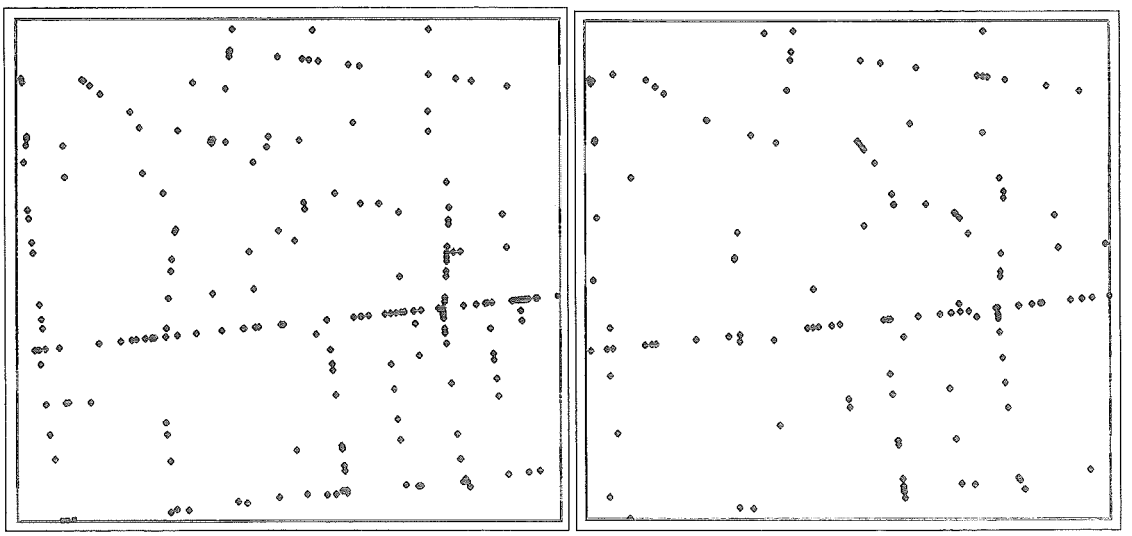
(d) Accident distribution for 1997-2001

Figure 8.02: Locations of accidents in the Christchurch CBD for 1982-2001 (fatal and injury crashes)



(a) Accident distribution for 1982-1986

(b) Accident distribution for 1987-1991



(c) Accident distribution for 1992-1996

(d) Accident distribution for 1997-2001

Figure 8.03: Location of accidents in the Riccarton suburb in Christchurch during 1982 - 2001. (fatal and injury crashes)

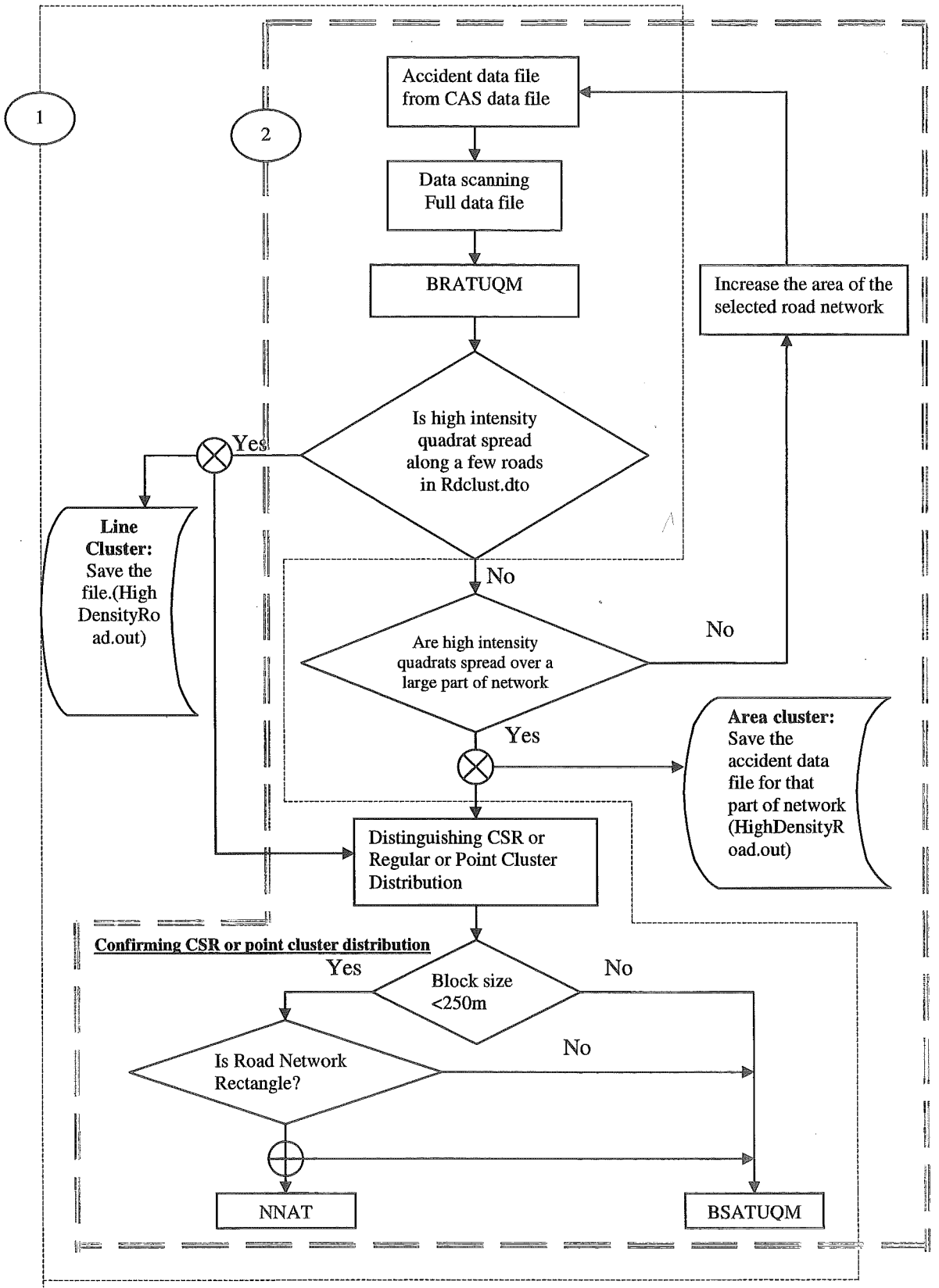
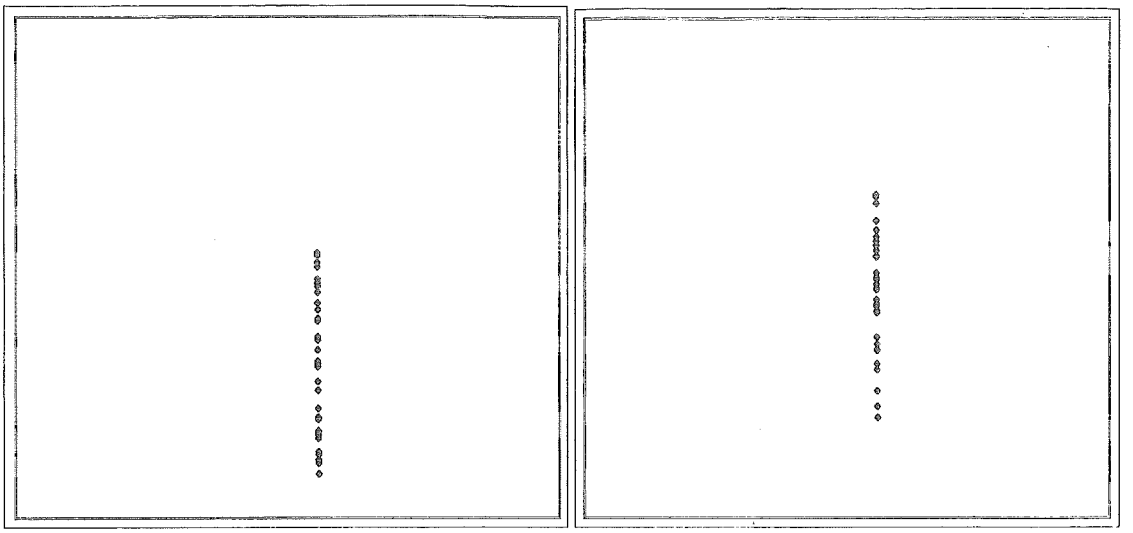
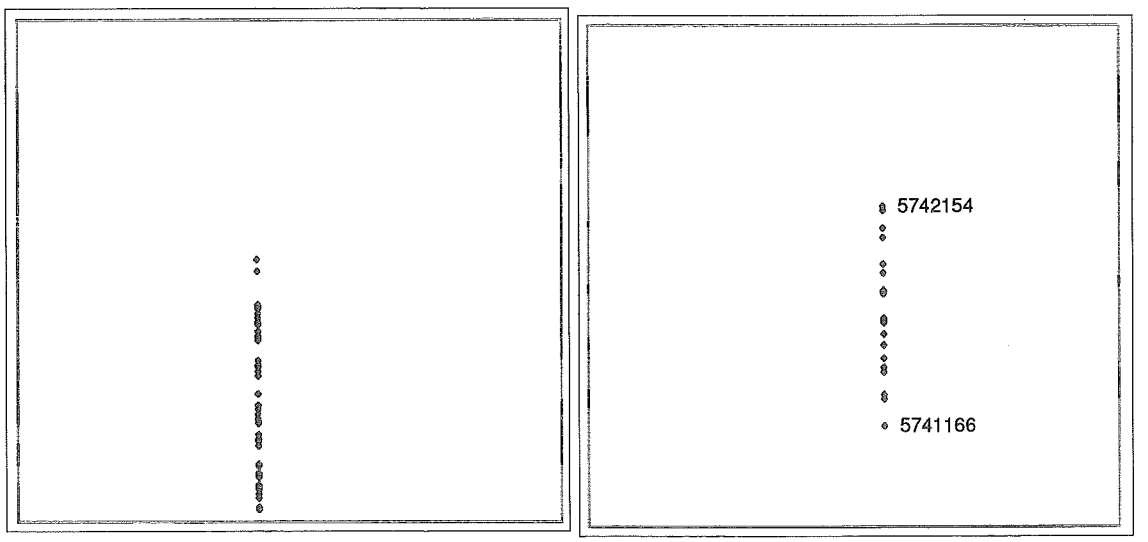


Figure 8.04: Flowchart showing the steps to choose programs



(a) High accident intensity for 1982-1986 (Manchester St.)

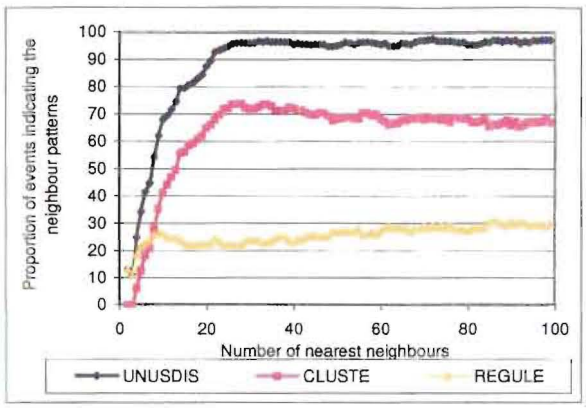
(b) High accident intensity for 1987-1991 (Manchester St.)



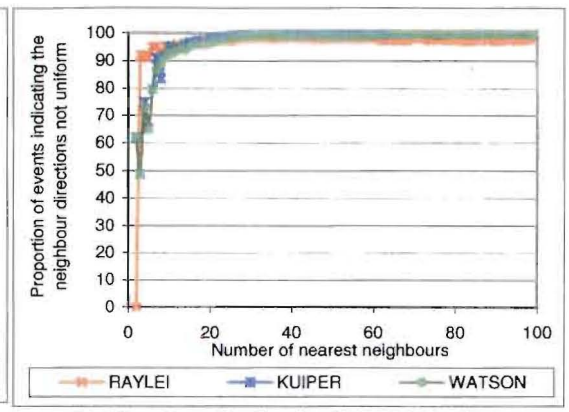
(c) High accident intensity for 1992-1996 (Colombo St.)

(d) High accident intensity for 1997-2001 (Manchester St.)

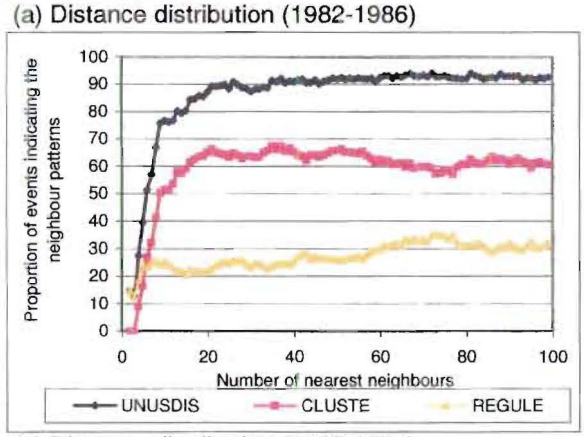
Figure 8.05: High accident (fatal and injury crashes) intensity locations shown for central city in Christchurch (quadrat radius 505m)



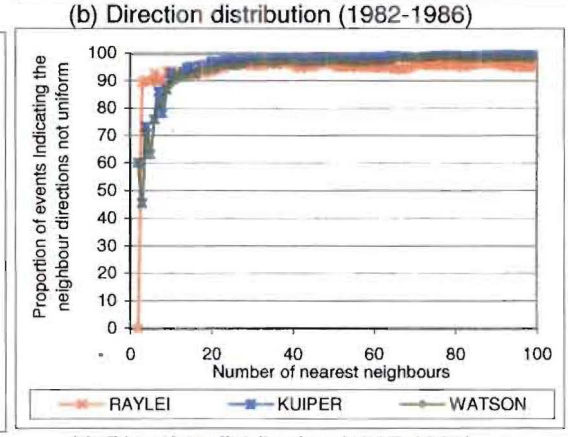
(a) Distance distribution (1982-1986)



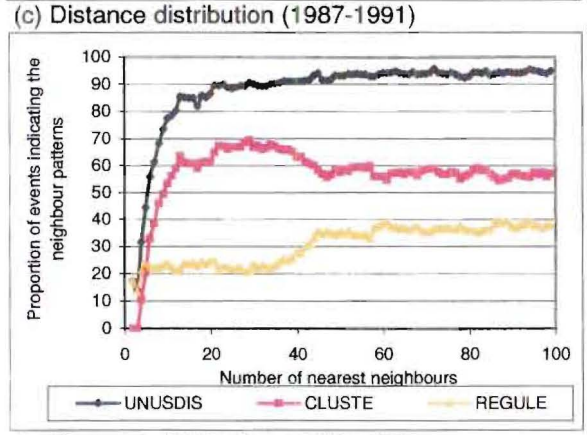
(b) Direction distribution (1982-1986)



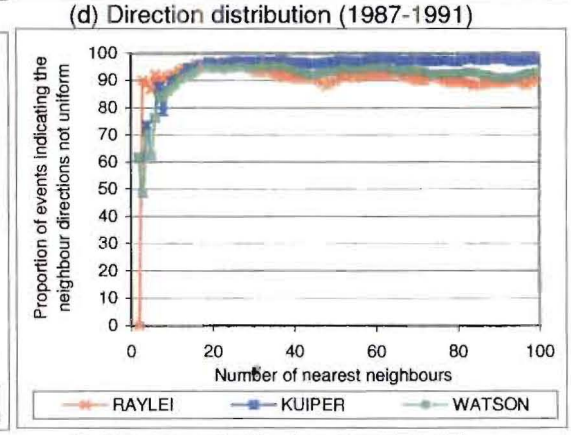
(c) Distance distribution (1987-1991)



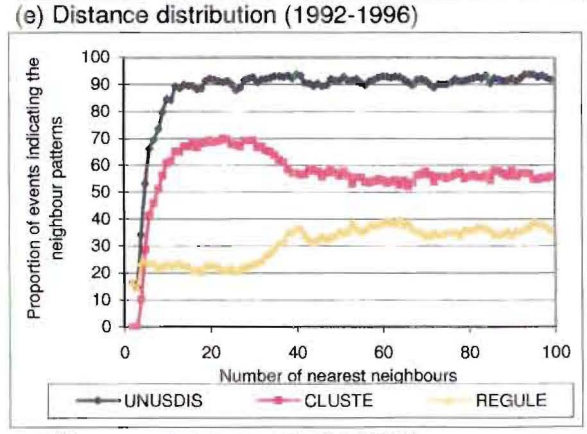
(d) Direction distribution (1987-1991)



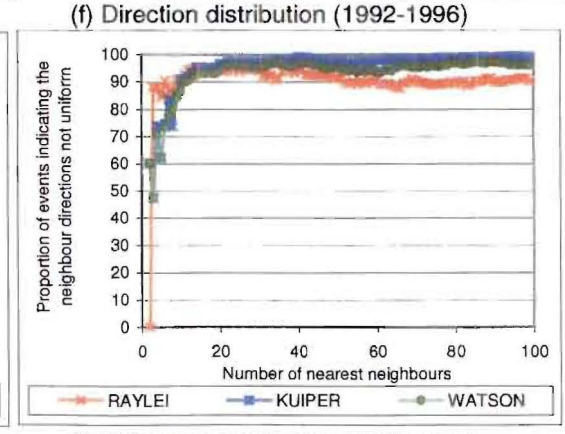
(e) Distance distribution (1992-1996)



(f) Direction distribution (1992-1996)

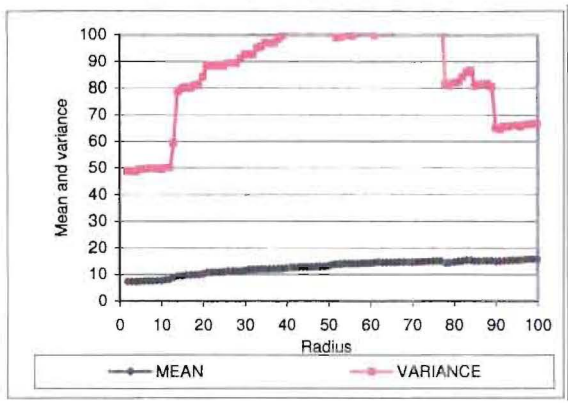


(g) Distance distribution (1997-2001)

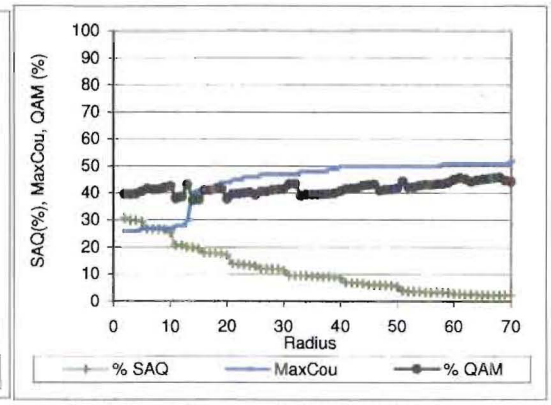


(h) Direction distribution (1997-2001)

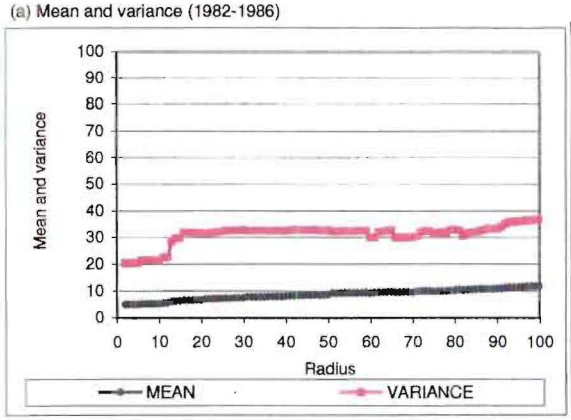
Figure 8.06: Nearest-neighbour distance and direction distributions (CBD Christchurch)



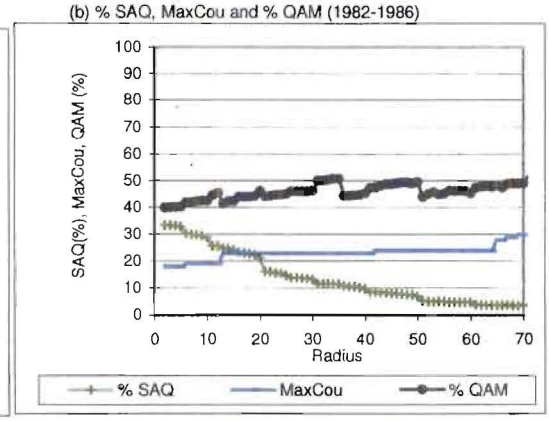
(a) Mean and variance (1982-1986)



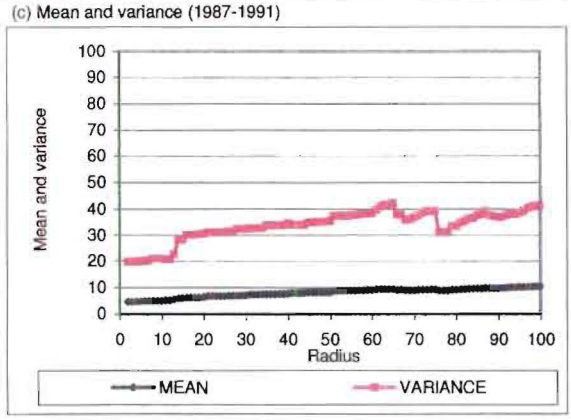
(b) % SAQ, MaxCou and % QAM (1982-1986)



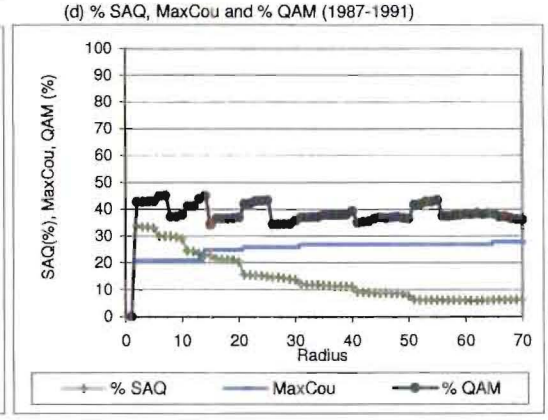
(c) Mean and variance (1987-1991)



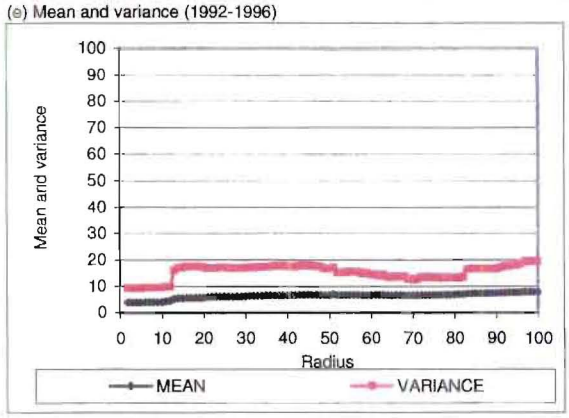
(d) % SAQ, MaxCou and % QAM (1987-1991)



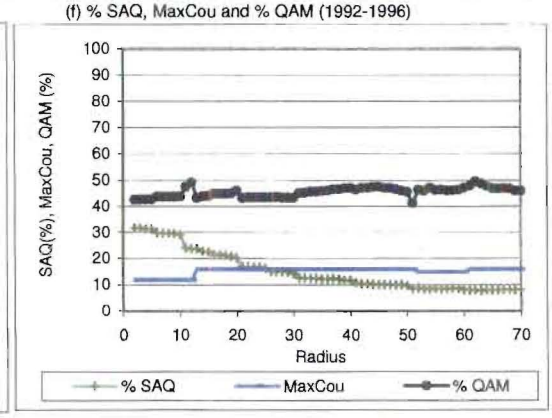
(e) Mean and variance (1992-1996)



(f) % SAQ, MaxCou and % QAM (1992-1996)



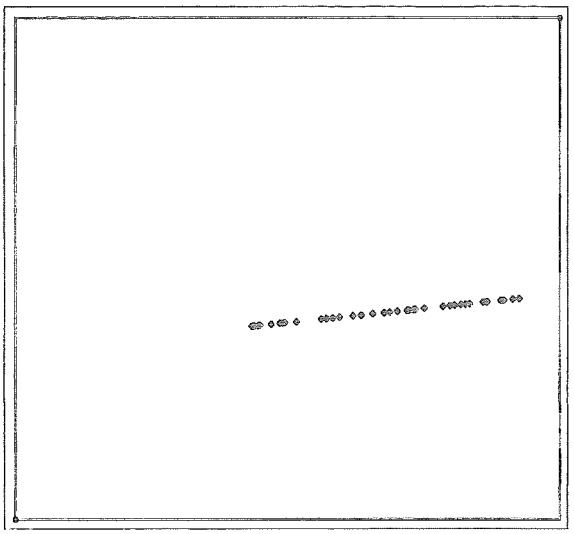
(g) Mean and variance (1997-2001)



(h) % SAQ, MaxCou and % QAM (1997-2001)

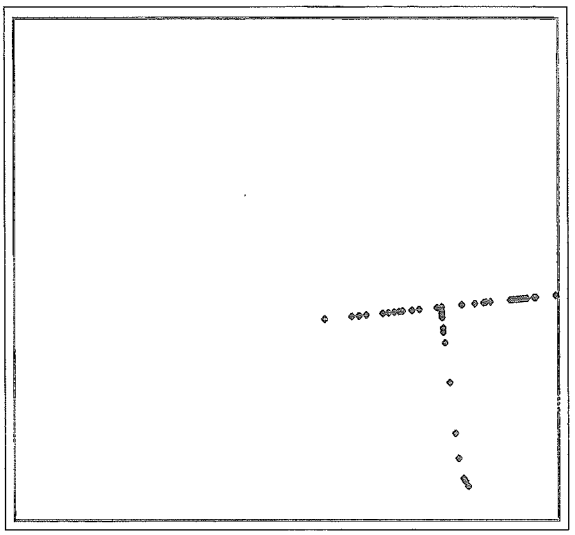
% SAQ - percentage of single accident quadrats MaxCou - maximum count % QAM - percentage of quadrat count above mean

Figure 8.07: Variation of mean, variance, %SAQ, MaxCou and %QAM with increasing quadrat radius (Christchurch CBD).

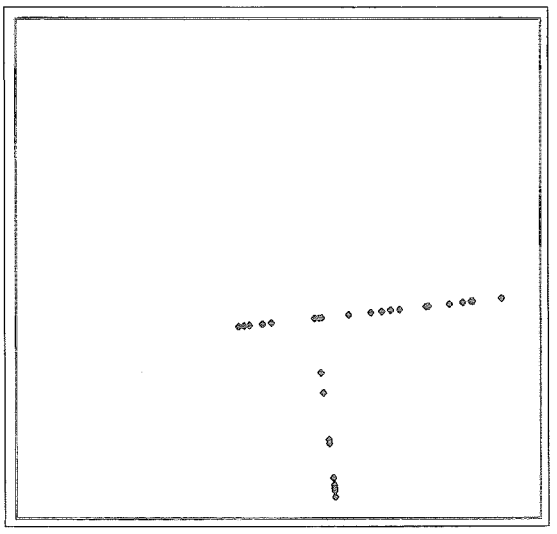


(a) High accident intensity during 1982-1986
(Riccarton Rd.)

(b)High accident intensity during 1987-1991
(Riccarton Rd.)

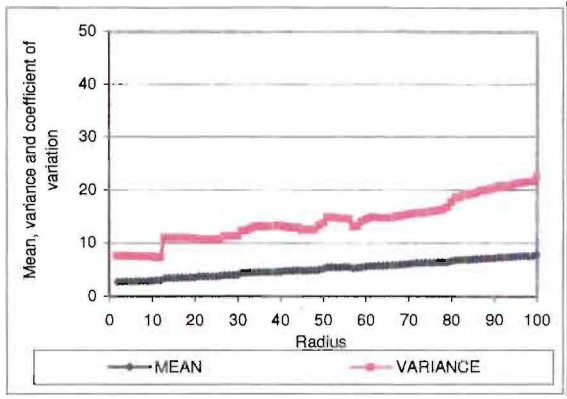


(c) High accident intensity during 1992-1996
(Riccarton Rd. & Clarence St.)

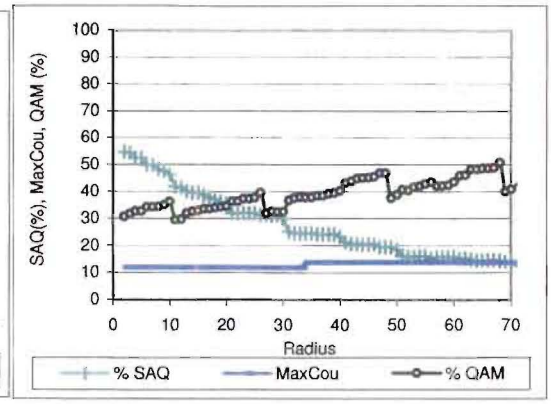


(d) High accident intensity during 1997-2001
(Riccarton Rd. & Matipo St.)

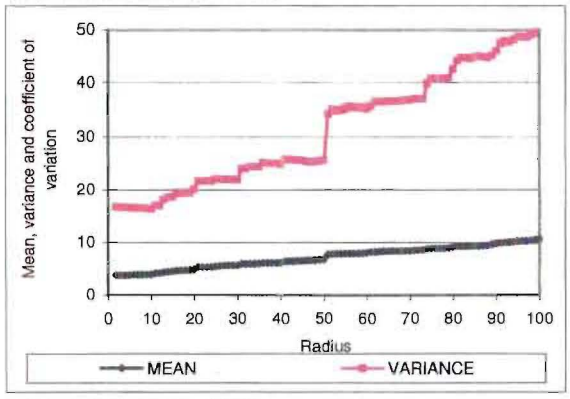
Figure 8.08: High accident (fatal and injury crashes only) intensity locations shown for Riccarton suburb in Christchurch (quadrat radius 505m)



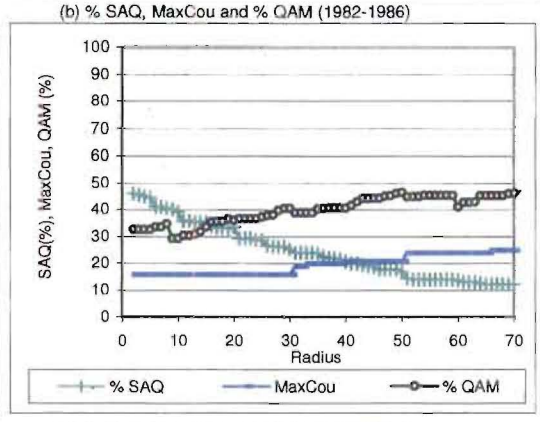
(a) Mean and variance (1982-1986)



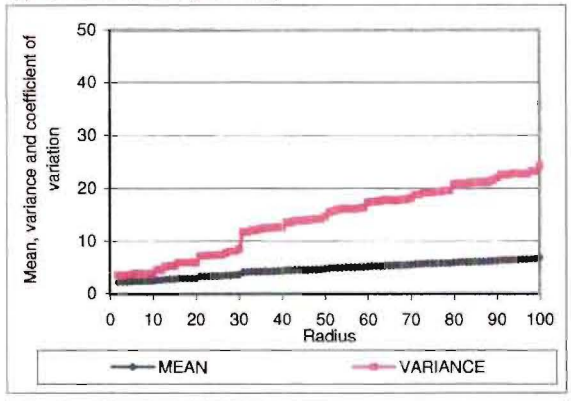
(b) % SAQ, MaxCou and % QAM (1982-1986)



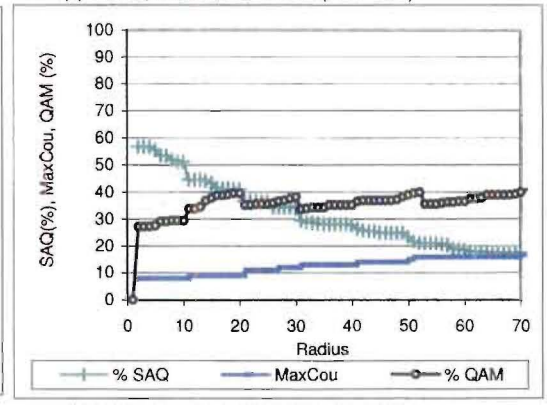
(c) Mean and variance (1987-1991)



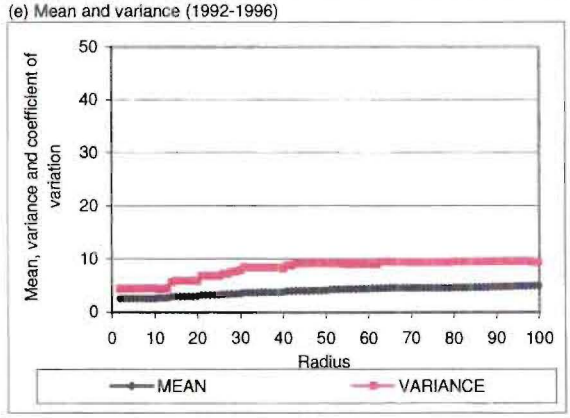
(d) % SAQ, MaxCou and % QAM (1987-1991)



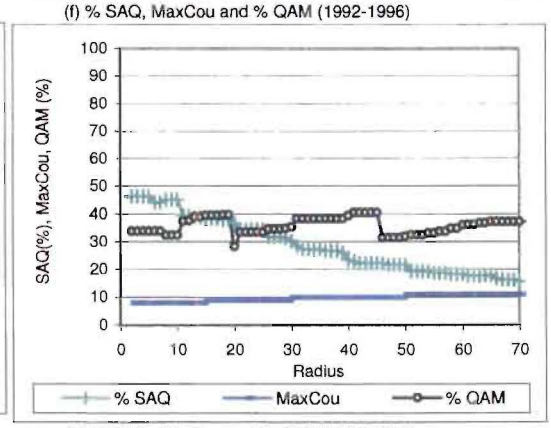
(e) Mean and variance (1992-1996)



(f) % SAQ, MaxCou and % QAM (1992-1996)



(g) Mean and variance (1997-2001)



(h) % SAQ, MaxCou and % QAM (1997-2001)

% SAQ - percentage of single accident quadrats MaxCou - maximum count % QAM - percentage of quadrat count above mean

Figure 8.09: Variation of mean, variance, %SAQ, MaxCou and %QAM with increasing quadrat radius (Riccarton suburb in Christchurch).

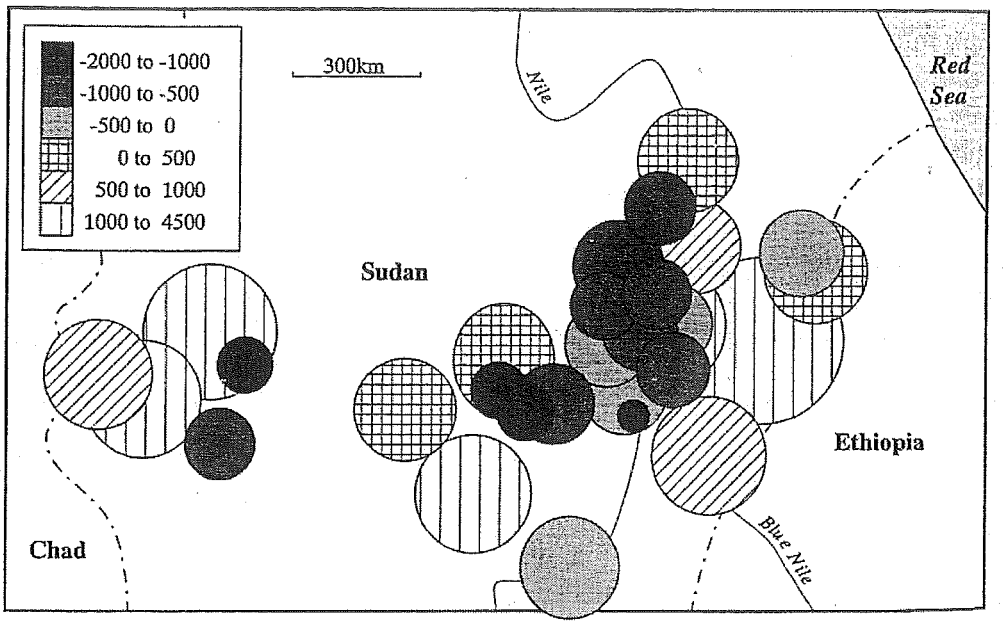
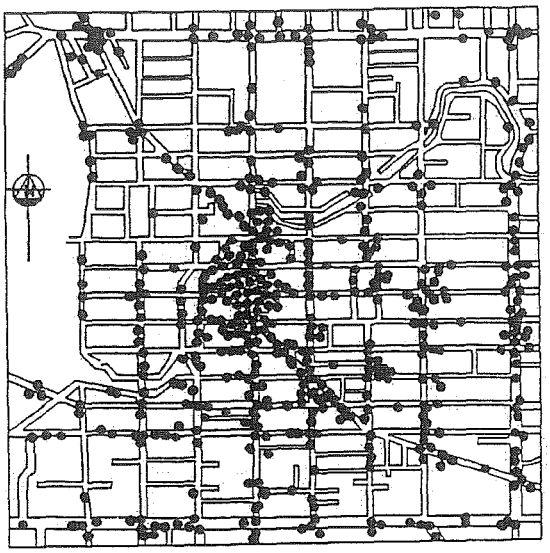
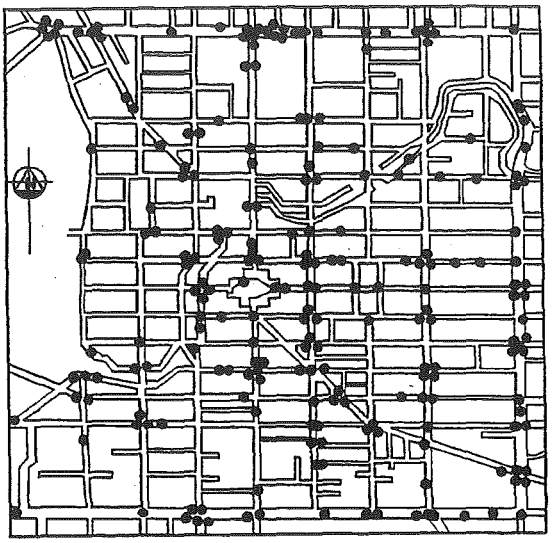


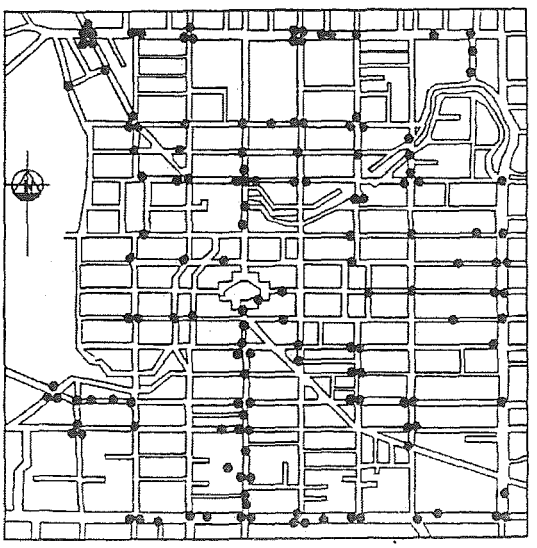
Figure 8.10: Sudan rainfall plot, 1982
(Extracted from Bailey and Gatrell [1995])



a. Injury crashes 1966- Total 530
(Prior to one-way Streets)



b. Injury crashes 1984-Total 212
(Following one-way Streets)



b. Injury crashes 1996-Total 148

Figure 8.11: Accident plots in three different years in Christchurch (CBD area)

[Figure extracted from Douglass [October 2000]]

Chapter 9

DISCUSSION AND CONCLUSIONS

9.1 Discussion

In this Chapter outcomes of each of the three statistical techniques are briefly discussed and the methods suitable for accident analysis are suggested. The tentative benefit-cost ratio for the application of the identified method is estimated. Future research on statistical techniques for analysing spatial distribution of accidents is discussed. The final discussion is about the research contribution towards the year 2010 NZ target of 300 or less traffic accident deaths per year.

9.1.1 *Cluster analysis*

The analysis of hypothetical distributions in Chapter 7 indicated that the cluster analysis method is not helpful for accident analysis. The complete-linkage method is not helpful in identifying the point or line clusters. Neither the single-linkage method nor the complete-linkage method is helpful in identifying the type of spatial distribution when the distribution is a mixture of point and line clusters, or a mixture of point cluster, line cluster and CSR distributions. The single-linkage method was expected to be useful in identifying line clusters distributions, but it does not work well if CSR or point cluster distributions are also present. Generally accident distributions will not consist of a single type of distribution, and therefore the cluster analysis method, using the area under the dissimilarity coefficient profile, is not useful for accident analysis.

9.1.2 *Nearest-neighbour analysis*

The nearest-neighbour distance method identifies the accident distribution very well when it is CSR. This is an important test before deciding on an accident reduction plan. The reason is that if the distribution is CSR then the accident distribution is not stable, and can change without any accident reduction treatment. Because of this instability, an accident reduction plan may not work well.

The nearest-neighbour method is capable of identifying a spatial distribution which has a high proportion of equal sized clusters. The nearest-neighbour direction analysis is useful in identifying the distributions with non-uniform directions, but the method does not distinguish between the point cluster distribution and the line cluster distribution. Therefore, other methods, in addition to nearest-neighbour analysis, must be used to investigate accident distributions.

The nearest-neighbour analysis method is suitable for a regular dense road network, but when the road network is sparse then the technique does not work well. The nearest-neighbour analysis method considers the road network as a continuum, which is valid only when the road network is relatively dense (e.g., the lengths of road between intersections, that is road links, are small). Nicholson [1995] pointed out that when the length of the road links is less than 100m, the network can be treated as a continuum. Nicholson noted that the error in the nearest-neighbour analysis result may be neglected up to a block size of 250m. In the case of the Christchurch CBD the blocks are 110m by 220m, and therefore the continuum approximation is reasonable for nearest-neighbour analysis. If the length of the road links is more than 250m then some other techniques need to be used for accident analysis.

9.1.3 *Accident-centred quadrat analysis*

This method examines a number of indices that evaluate certain properties related to the spatial distribution, using the locations of individual accidents in a road network. It was intended to get more output from the analysis. For example, if the accidents are dispersed, the percentage of single accident quadrats is obtained from %SAQ. If the accident locations are clustered, then the cluster size is obtained from the number of events within the quadrat and the quadrat radius. How this cluster size relates to the overall pattern is decided using the mean and variance of the quadrat counts. Spatial distributions in which events are locally dense or locally sparse can be identified using the quadrat count variance.

This analysis technique does not appear to be affected by whether the network is dense or sparse. The analyses of the two case study areas (the Christchurch CBD and the Riccarton suburb in Christchurch) indicate dispersion of accident clusters during each five-year period.

The five indices (i.e., ICS, ICF, ICR, IP and IM) are not useful, but the mean, the variance, the maximum count, the percentage of single accident quadrats, and the percentage of quadrat counts above the mean are useful indicators when analysing the accident distribution. The ratio of the variance to the mean is a useful indicator for identifying excessive clustering or regularity, and may be noted from mean and variance plots. The mean and variance plots clearly indicate how the accident distributions have changed from 1981 - 2001 in the two case study areas.

The accident-centred quadrat method identifies route clusters well compared to other statistical techniques. The advantage of this method is the ability to identify the high accident count sections of the road. In this analysis the top five accident intensity quadrats for each quadrat radius were defined as high accident intensity quadrats, but the user can define how many of the top accident intensity quadrats will be analysed. If the high intensity quadrat centres are not spread throughout an area but are spread throughout a road or road section, then it is an indication that there is a route cluster rather than an area cluster. These were discussed in Chapter 8 with the help of the two case study areas (Riccarton suburb and CBD area of Christchurch).

Accident data from different parts of the area may indicate different types of clusters (see Figure 3.20). If the high intensity quadrat centres are spread throughout a large part of a road network but not throughout a few roads or road sections, then it is an indication that there is an area cluster in that part of the network. If the high intensity quadrat centres are spread throughout the entire road network rather than on a few roads or road sections, then it is an indication that there is an area cluster rather than route clusters. This should be confirmed by analysing a larger area of the road network. The indication of an area cluster needs to be confirmed with further analysis using the nearest-neighbour or quadrat method (with a small quadrat radius, i.e. 5 to 70m, depending on the road network) to make sure that the accident distribution is CSR or regular and not point clusters.

9.2 Overall performance of the proposed method

The three programs NNAT, BSATUQM and BRATUQM work well in terms of spatial analysis of accident distributions. The output files from the three programs were used to plot

the results as line graphs, histograms and location plots. The plotted results are useful for confirming the accident distribution.

Research into spatial data analysis in a variety of fields has been conducted for more than two decades. The availability of a number of dependent and independent variables makes it difficult for researchers to identify a simple index system, which can describe fully and accurately the nature of the spatial distribution. Some researchers (Nicholson [1990], Thomas [1995], Nicholson [1999]) identified the distribution with a single graph which did not fully describe the spatial distribution for different cases (e.g., point cluster, line cluster, CSR distributions).

Several dependent variables are involved in the spatial analysis of accidents (e.g. the number of nearest-neighbours and the quadrat radius). The results depend on the characteristics of the clusters, and we need to choose the values of the variables carefully. The problem is that the best values depend upon the characteristic of the clusters, which is not known until the analysis results are seen. Hence, we need to analyse the results for a range of values for these variables. Therefore plots are needed rather than index values.

The accidents are located within the boundaries of roads or car parks. The area enclosed by roads (e.g. block size) may influence the results of the analysis. The difficulty in finding an independent variable, which is not related to the space between roads, makes it difficult to find a simple index to represent the accident distribution characteristics. The spatial accident data analysis results need a number of graphs to identify the accident distribution (i.e., CSR, black spots, black route, black area and mixture of these).

9.3 Accident data requirements

The reliability of statistical analysis depends on the number of accidents in the data set. If the total number of accidents is small, then the analysis results might not be reliable. This is one of the reasons why five years of accident data were selected for analysis in Chapter 8. It is possible that with a shorter time period (e.g., six months) of accident data, the analysis results will point towards an inappropriate accident reduction plan. For example, a particular road may be shown as a route cluster in one year, but in the following year it may not be, but

instead another road may appear as a route cluster. In general whenever possible, it is prudent to select a reasonably long period (such as five years) for accident analysis.

There is a possibility that changes (e.g. the start of a new traffic management plan or changes to road geometry in one of the five year period) might have occurred during the five year period, and can change the accident distribution. In this type of situation careful consideration (e.g. analysis of different five year periods, the progress of clustering during the periods) is necessary before deciding on the appropriate accident reduction plan.

In order to reduce the effect of randomness in the accident data and obtain statistically reliable results, it is necessary to include as many accidents in the data as possible. Fatal accident data are more reliable (better reported) than non-injury data. However, since the number of fatal accidents is much less, if we analyse only the fatal accidents, then the effect of randomness on the results of the analysis may be very high. This is why both fatal and injury accident data are included in the analysis.

On some rural roads, traffic flows are very small, and hence the fatal and injury accident data for analysis are inadequate. In such cases we may need to include non-injury accidents. Caution is necessary when including non-injury accident data for analysis. It is likely that the reported fatal and injury accidents are more reliable in terms of spatial location than non-injury accidents. The reason is that the police may not visit some non-injury accident sites and the reporting rate also changes from place to place. The cost resulting from accidents is considerably less for non-injury accidents compared to fatal or injury accidents. Therefore an accident reduction plan based on non-injury accident data may not have a high benefit-cost ratio. The accident data can be selected from roads having approximately equal traffic flow. For example, motorway accident data can be analysed with data for other highways or motorways, but not with distributor roads or arterial roads. The distributor and arterial road accident locations can be analysed together.

There are some continuous roads with sections called by different names, for example, Memorial Ave and Fendalton Road in Christchurch. If the road network selected for accident data analysis includes the two roads (i.e., for example a section of Memorial Ave and a section of Fendalton Road) then it is better to use a common identifier (i.e. the first street name should be the same, and either Memorial Ave or Fendalton Road) for the route

analysis. This is because a route is analysed using a large quadrat (ie., the diameter range is from 500m to 2000m). If a quadrat is centred on one of the two roads (e.g., Fendalton Road) and a section of the second road (Memorial Ave) is also within that quadrat, then the accidents on the second road will not be counted and this will lead to an error. Therefore, the two different road names must be changed to a single name for black route analysis.

There is a similar difficulty in the analysis of the CBD accident data. The roads Sherbourne St, part of Bealey Avenue, Madras Street, Barbadoes Street and part of Moorhouse Ave are recoded as SH 74 in the LTSA accident data. For example, the accident road names were recoded as chainage distances (eg., 74/0/8.592, 74/0/8.581) in the LTSA accident data file. The input data file was changed according to the selected road name (i.e. aggregating location name into one route name) for route analysis. These types of data scanning and modification are necessary before starting route cluster analysis.

9.4 Cost-benefit analysis

This section investigates the potential benefits and costs associated with the further development and eventual implementation of this research. As previously discussed the main benefits of this research are the identification of site, route and area wide safety issues that may then be addressed as part of a safety improvement programme. The research will also assist in better targeting of safety improvements, by defining the "boundaries" to improvement areas, beyond which increased investment will provide decreasing returns. These techniques may be used to define the scope of physical works, and to direct enforcement resources or even education programmes, to provide the maximum return on the road safety resources available.

The LTSA [2003a] Overal Results of Crash Reduction Monitoring System has shown that in a period of 14 years, accident reduction works in NZ have resulted in an estimated saving of approximately $ 3.0 billion in the social cost injury accidents at sites where low cost engineering measures have been implemented. These costs are based on the levels of crash reductions achieved, disaggregated by severity, and the social cost of each crash type. The analysis [LTSA 2003a] focussed on reductions in reported injury crashes and the resulting benefit calculations were based on two extreme assumptions, that:

1. all injury crashes were avoided, but the number of non-injury crashes remained unchanged;

2. all injury crashes were reduced to non-injury crashes.

The estimated social cost savings using the first assumption was $2970 million and using the second assumption was $2960 million (based on June 2002 price).

LTSA [2004b] noted that based on the LTSA [2003a] analyses, the mean annual savings associated with sites active in the last ten calendar years (1994-2003) is $203 million per annum. It further noted that,

- "These social cost estimates take into account crashes occurring during the 'after' monitoring period only. There may be additional benefits from after the monitoring period has finished. In this respect, these estimates are conservative.

- These estimates assume that the crash savings were constant over the 'after' monitoring period. This is likely to be a reasonable assumption in most cases, but it is possible that the effect of some interventions may have decreased over the (typically five-year) 'after' period.

- These estimates assume that all injury crashes were avoided. A previous report ... found a 0.5% decrease in the estimated social cost saving under the assumption that all injury crashes were reduced to non-injury crashes".

In this cost-benefit calculation, a mean annual savings associated with sites active in the ten year period (1994-2003), equal to $203 million per annum, is used.

It is acknowledged that over time the expected return on minor safety works may be expected to decrease, as increasingly more high return projects are completed. Therefore conservatively, it is assumed that the average return on low cost safety improvements in the future may be only 50% of $203 million, i.e. $101.5 million. It is assumed that no additional funds would be used, either to prioritise and target the safety improvements under the accident reduction programme, using the analyses techniques developed as part of this research, or to implement the safety improvements. The analytical technique developed will enable more targeted and effective implementation of safety improvements using existing funds. If the implementation of this research were to improve targeting, leading to an

additional benefit from the accident reduction programme of 2%, then the annual benefits could be assessed to be $2.03 million per annum.

If the estimated benefit from applying the outcome of this thesis is $ 2.03 million per annum over a 6 year evaluation period (5 year plus 1 year to start the application) and using a discount rate of 10%,

the present value of the savings is $= \$ 2.03 \text{ million} \times (4.57-0.954)$

$= \$ 7.34 \text{ million}.$

where the present worth factor for 6 years and 1 year were obtained from the project evaluation manual [Transfund, 2000].

If the probability of the research results not being implemented successfully is 0.1 (i.e. the probability of the research results being implemented successfully is 0.9) then the expected benefit of the research is

$$\$ 7.34 \text{ million} \times 0.9 = \$ 6.6 \text{ million}$$

The costs are those associated with the refinement of the system, the development of the software modules NNAT, BSATUQM and BRATUQM into a user-friendly system, the development of user manuals and training materials. The costs associated with these items are expected to be roughly $300,000. In addition users of the system would also incur staff training costs, estimated at about $60,000. This would be a recurring cost possibly every 5 years as new staff need to be trained.

So if the project development costs are $300,000 with a further $60,000 for training, assuming the commitment for these costs is made in the first year, the net present value of the costs is $0.36 million.

Hence the expected benefit/cost ratio is $= \$6.6 \text{ million} / \$ 0.36 \text{ million}$

≈ 18

The indicative benefit / cost analysis shows that there are significant benefits in developing and applying the analytical techniques developed in this research, to prioritise and target actions under the accident reduction programme.

9.5 Future research

Potential areas for future research are discussed below.

1. The risk and cost density were discussed in Section 2.3.1. The risk is a measure of cost per person-kilometre travelled. The cost-density is a measure of the annual social cost of crashes per kilometre of road. To calculate these two indices the volume of traffic at each accident location and the cost of each accident are needed. In future research the volume of traffic can be entered into the accident data file and the cost of each accident can be calculated from the accident data file.

   A new input data file, which contains the movement category (e.g. head on, hit object and overtaking), speed limit (e.g. 50km/h, 70km/h and 100km/h), accident severity (fatal, serious, minor injury and non-injury) and accident sites (e.g., bridge and railway crossings) used in the accident data file will help to calculate the accident cost. The CAS program uses a similar data file to estimate accident costs. We may import this file or use the data for each accident cost needed for analysis. If the cost of each accident is available then the program BRATUQM can estimate the two indices (risk and cost-density) for quadrats.

   The program BRATUQM is set up for counting the accidents within the quadrats on the road where the quadrat centre is located. Instead of counting the accidents within the quadrat the program could estimate the total accident cost. The accident cost divided by traffic volume per unit road length would be a measure of the risk (cents/vehicle-km), and the accident cost divided by the quadrat diameter will be a measure of the cost-density (cents/km). Using these two indices one could identify the high-risk locations (road or road section) and the high cost-density locations (road or road section).

   These two indices may be used to analyse the spatial distribution of risk and cost-density. The mean risk or mean cost-density for a certain quadrat radius (e.g. 70m or 500m) could be used to compare the regions or the sub-areas in the region.

   The advantages of extending the research in this way are:

- maximising efficiency (i.e. producing high benefit-cost ratios)
- increasing equity for road users.

This will be a useful method for utilising road safety resources in terms of cost-density and risk.

2. In this thesis, the number of high accident intensity quadrats were selected as a user-defined value (say five or ten top accident intensity quadrats for each quadrat radius) and used for investigating route or area clusters. Further research is necessary to investigate how to choose this user-defined value. This user-defined value could be selected from the quadrat count distributions for each quadrat radius. The selection could be either the top 5, 10 or 15 percent of accident intensity quadrats or by investigating the quadrat count frequency polygon. The lower limit of the high accident intensity could be decided from the quadrat accident intensity distribution.

3. Assessing the variation in the individual visual examination results by:
   a. using accident plots from the Crash Analysis System (CAS), an example of which is shown in Figure 1.01, where the locations are marked with circles of radius proportional to the number of accidents;
   b. using the accident plot from a different software called Accident Information Management System (AIMS), an example of which is a three dimensional plot marked with stacks of circles or squares, as shown in Figure 9.01.

   The analysis results from the three programs (NNAT, BSATUQM and BRATUQM) may be compared with the visual examination results.

   To assess visual examination results and the statistical analysis results the following road network characteristics must be included.

   i. Block size less than 250m.

   ii. Block size greater than 250m

   iii. Spider road net work (e.g. Dunedin CBD road network)

   iv. Regular trapezoidal (e.g. Christchurch CBD road network)

v. Irregular road network

vi. Rural road network.

4. Sequences of mixtures of

(i) Point clusters and CSR distributions

(ii) Line cluster and CSR distributions

were discussed in Chapter 7 and the statistical results are given in Appendices. The location plots may be shown to practitioners (working in the traffic accident analysis area) for visual examination and their ordering of the sequence of mixtures recorded. Next, the statistical results need to be supplied with each of the mixed spatial distribution plots and the practitioner's new ordering of the sequence of mixtures need to be recorded. This will assess how well the statistical measures compare with visual examination. This assessment could be done in future research.

5. The research presented in this thesis will help to analyse the accident distribution on roads with less than 70km/h speed zone (urban road network) or on roads with more than 70km/h (rural road network) including Motorways. For practical reasons short sections with different speed zones (e.g. 30km/h or 60km/h) cannot be analysed separately.

Transfund [1997] noted that "there have been marked differences between the accident trends in 50km/h areas compared with 70km/h and above areas, and different factors are used to modify the accident numbers for the different posted speed limit areas". Using those adjustment factors stipulated in Transfund the quadrat count could be adjusted according to the speed zone. These adjusted counts might influence the quadrat count variance. Hence, identifying a CSR distribution may not be realistic but identifying line cluster or area cluster could be acceptable. Further research is necessary to identify the influence on analysis result from accommodating the factor introduced in Transfund.

6. The additional data (first and second street names, as explained in Chapter 8) were used to identify line clusters or area clusters for accident-centred quadrat method. It is unreasonable to say that the accident-centred quadrat is better than the other two methods (cluster analysis and nearest neighbour analysis) because the accident-centred quadrat method identifies line and area clusters. This is because the quadrat method has been extended to include additional data. It may be possible to extend the other methods, so they also use the additional data. Further research on how to use the additional data in nearest-neighbour and cluster analysis methods would be worthwhile. The most suitable method for accident analysis may then be identified.

   Without the additional data, the single-linkage cluster analysis results show different dissimilarity coefficient profiles for the four basic spatial distributions. On this basis the area under the dissimilarity coefficient profile was chosen for further analysis. However, it was found this did not work very well. There is a possibility that quite different dissimilarity coefficient profiles, can have equal area under those dissimilarity coefficient profiles but the first or second moments could be different, and could be used to identify accident distribution. Further research on this might improve the capability to distinguish accident distributions without additional data such as first and second street names.

7. In a line cluster, the events will increase in only one dimension (along the line) but for point cluster, CSR or regular distributions, events will increase in two dimensions. Hence the optimum radius of quadrats for each of the four basic distributions (CSR, line, point cluster or regular distribution) might be different. The difference in the mean count per quadrat may not be considerable for CSR, line cluster, point cluster or regular distributions, when we use the appropriate quadrat radius for each spatial distribution. However, the mean count may differ when we use the same quadrat radius for each of the four (CSR, line, point cluster or regular distribution). In Section 6.3.1 the relationship between proportion of radius (h/R) and proportion of mean count (M_h / M_R) was discussed, but was investigated for a particular range of radii. The relationship may be investigated for different range of radii.

The range of the quadrat radius selected to obtain the profiles shown in Figures 7.41a, c, e and g and the Figure 7.44a, c, e and g may well have influenced the results. A possible reason is that the relationship between M_h / M_R and h/R may similar to point or line or CSR distributions for narrow range of radii. For a larger range of radii this profile may be different.

Further research is necessary to investigate the relationship between the proportion of radius (h/R) and the proportion of mean count (M_h / M_R) for the wider range of quadrat radii for each of the four basic distributions.

8. Line cluster may increase in length or increase in the number of accidents within a particular length, or both. This results in the following question: If there was a route cluster within a road network for the past ten years and the single site plan was applied for a few sites along the route, but black route treatment was not applied, then what will happen to the line cluster? A study of the effect of applying the inappropriate treatment type (e.g. site treatment when the problem is route clustering) would be worthwhile.

9. In Section 8.5 as an example Riccarton Road section was indicated as a route cluster but single site plan was applied in the past. We can investigate the effectiveness of line cluster when applying the single site plan using the program "BRATUQM", which will help to find the length of the line cluster, intensity (number of accidents per km) and location.

In this section nine research topics have been proposed and discussed. They would enable development of the spatial distribution analysis approach presented in this thesis, so that they may be more easily and effectively applied in practice.

9.6 Conclusion

In this thesis three statistical analysis techniques (cluster analysis, nearest-neighbour and quadrat analysis) were tested with hypothetical distributions. The nearest-neighbour and quadrat were identified as suitable techniques for accident analysis. The traditional quadrat

analysis method was modified to analyse road accident data using accident-centred quadrat method. The nearest-neighbour and the accident-centred quadrat methods were selected for accident analysis because these helped to distinguish accident distributions for the following circumstances.

- The nearest-neighbour analysis and accident-centred quadrat method identified the CSR distributions among the four basic spatial distributions.
- The nearest-neighbour analysis results indicated the proportion of events having non-random nearest-neighbour distance distributions.
- The accident-centred quadrat method performed well when analysing accident data for dense (block size <250m) or sparse (block size >250m), regular or irregular road networks. Nearest-neighbour method performed well for dense networks only.
- The accident-centred quadrat method was able to identify line clusters, and indicate changes in the intensity (number of accidents per km) and the length of line clusters, as well as identify area clusters.

The two methods (the nearest-neighbour and the accident-centred quadrat methods) were used to analyse twenty years of accident data for the CBD and the Riccarton suburb in Christchurch, New Zealand. The CBD area is a dense road network, while the Riccarton suburb is a sparse network and the roads are not as regularly spaced as they are in the CBD area. The analysis results show that the accident distribution has changed during the twenty years (1982-2001) and that it is now appropriate to switch to route action plans for the two areas.

There are several ways to reduce social cost (e.g. road rule enforcement, construction of safer roads, enforcement of speed management). For example we can reduce accident severity by reducing the speed limit and the police enforcement, but the effectiveness depends on maintaining the police enforcement, which means on-going cost. Identifying unsafe locations (black spots, black routes, black areas) and improving the locations will be a fixed cost, and may be cheaper in the longer term. Identifying the appropriate accident reduction plan for accident reduction treatment is therefore an important matter.

The New Zealand Road Safety Strategy 2010 [LTSA, 2000 b] noted that if we do not take appropriate action then the annual social cost of injury crashes will increase from $3.1 billion in 1998 to $4.6 billion in 2010. The techniques for identifying accident distributions and applying appropriate accident reduction plans, as explained in this research, should be used. The cost-benefit analysis indicates that it would be well worth implementing the outcome of the research, to help reduce the future social cost of accidents (i.e. loss of life, permanent injury, non injury or property loss caused by traffic accidents).

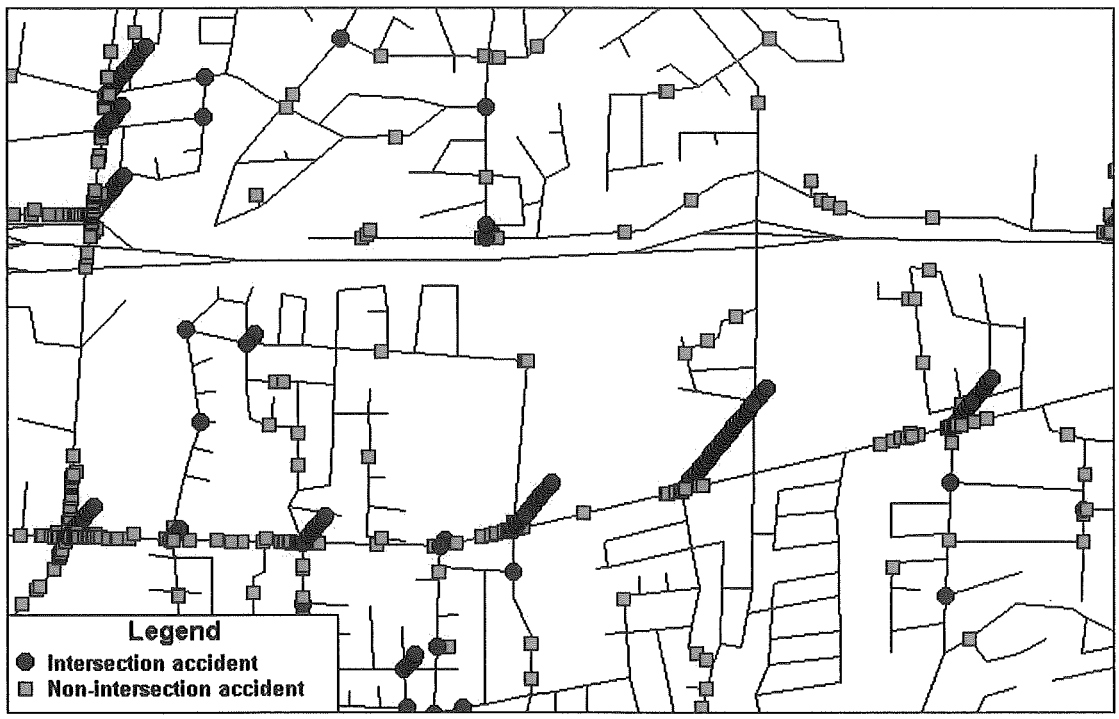


Figure 9.01: Accident plotted on exact location on a road map using Accident Information Management System software.

(Figures extracted from web site: WWW.jmwengineering.com)

References:

Alley, W.M., 1993. Regional Ground-Water Quality. Van Nostrand Reinhold, New York.

American Association of State Highway Officials, 1946. A Policy on Geometric Design of Rural Highways. Washington, USA.

Anderberg, M.R., 1973. Cluster Analysis for Applications, Academic Press Inc, New York.

Anujah, R., 1997. Analysis of Spatial Distribution of Accidents. Master of Engineering Report, Department of Civil Engineering, University of Canterbury, Christchurch, NZ.

Austin, K., 1995. The Identification of Mistakes in Road Accident Records: Part 1, Locational Variables. Accident Analysis and Prevention, 27(02), 261-276.

Bailey T.C., and Gatrell A.C., 1995. Interactive Spatial Data Analysis, John Wiley & Sons, New York.

Baldwin, D.M., 1946. The Relation of Highway Design to Traffic Accident Experience. Procs American Association of State Highway Officials Convention Group Meeting. Washington D.C., USA.

Beenstock, M., Gafnia, D., and Goldin, E., 2001. The Effect on Traffic Policing on Road Safety in Israel. Accident Analysis and Prevention, 33(01), 73-80.

Benjamin, J.R., and Cornell, C.A., 1970. Probability, Statistics, and Decision for Civil Engineers, McGraw-Hill, New York.

Bennett, G.T., and Marland, J., 1978. Road Accidents in Traditionally Designed Residential Estates. Transport and Road Research Laboratory Research Report SR394, Crowthorne, UK.

Boyle, A.J., and Wright, C.C., 1984. Accident Migration After Remedial Treatment at Accident Black Spots. Traffic Engineering and Control, 25(05), 260-267.

Brindle, R.E., 1983. Local Street Traffic and Safety: A Perspective. In Australian Road Research Board Report ARR 129 (Local Street Traffic and Safety: Workshop Papers and Discussions. Brindle, R.E., and Sharp, K.G.), Melbourne, Australia.

Cameron, A.C., and Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University press, UK.

Chapman, R.A., 1967. Traffic Collision Exposure. Proceedings of New Zealand National Roads Board Roading Symposium, Wellington, 184-198.

Chapman, R.A., 1973. The Concept of Exposure. Accident Analysis and Prevention, 5(02), 95-109.

Charlesworth, G., and Coburn, T.M., 1957. The Influence of Road Layout on Speeds and Accidents in Rural Areas. Journal of Institution of Municipal Engineers, 83(07), 221-240.

Colgate, M., and Tanner, J.C., 1967. Accidents at Rural Three-way Junctions. Road Research Laboratory Report LR 87.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.

Dalby, E., 1979. Area-wide Measures in Urban Road Safety. Transport and Road Research Laboratory Report SR517, Crowthorne, UK.

Dalby, E., 1987. The Spatial Distribution of Urban Road Accidents. Centre for Transport Studies, University of London Centre for Transport Studies.

Dale M. R.T., 1999. Spatial Pattern Analysis in Plant Ecology. Cambridge, UK.

Department of Transport (UK) 1986. Accident Investigation Manual.

Diggle, P.J., 1983. Statistical Analysis of Spatial Point Patterns. Academic Press, Inc. New York.

Douglass, M., October, 2000. Christchurch City Centre 40 Years of Change, Traffic, Planning 1959 – 1999, Christchurch City Council, Christchurch, NZ.

Elvik, R., 1995. The Safety Value of Guardrails and Crash Cushions: A Meta-Analysis of Evidence from Evaluation Studies. Accident Analysis and Prevention, 27(04), 523-549.

Elvik, R., 2001. Area-wide Traffic Calming Schemes: a Meta-analysis of Safety Effects. Accident analysis and Prevention, 33(03), 327-336.

Evans, L., 1985. Traffic Safety and the Driver. Van Nostrand Reinhold, New York.

Everitt, B.S., 1974. Cluster Analysis. John Wiley & Sons, Inc., New York.

Finch, D.J., Kompfner, P., Lockwood, C.R., and Maycock, G., 1994. Speed, Speed Limits and Accidents. Transport Research Laboratory Project Report 58, Crowthorne, UK.

Fischer, M., Scholten, H.J., and Unwin, D., 1996. Spatial Analytical Perspectives on GIS. Taylor & Francis Ltd, U.K.

Forrest, M., Council and Julie, A. Cirillo., 1988. Current Status Research and Implementation. Highway Safety: at Cross Roads Conference, San Antonio 25-51.

Fortheringham, S., and Rogerson, P., 1994. Spatial Analysis and GIS. Taylor and Francis Ltd, London.

Frank, H., and Althoen, S.C., 1994. Statistics Concepts and Applications, Cambridge University press, Cambridge.

Gaber, N.J., and Gadirau, R., 1988. Speed Variance and its Influence on Accidents. AAA Foundation for Traffic Safety, Washington, USA.

Glennon, J.C., 1987. Effect of Sight Distance on High way Safety. In Relationship Between Safety and Key Highways Features, State of the Art Report 6, Transportation Research Board, Washington, USA.

Goh, P.C., 1993. Traffic Accident Analysis Using Geoprocessing Techniques. Road and Transport Research, 2(02), 76-85.

Hakkert, A.S., and Mahalel, D., 1982. Estimating the Number of Accidents at Intersections from a Knowledge of the Traffic Flows on the Approaches. Accident Analysis and Prevention, 14(05), 359-369.

Harwood, D.W., and Hoban, C.J., 1987. Low Cost Operational and Safety Improvements for Two Lane Roads. Department of Transportation Report FHWA-IP-82-2, Washington D.C., USA .

Hauer, E., 1982. Traffic Conflicts and Exposure. Accident Analysis and Prevention, 14(05), 359-369.

Hauer, E., 1995. On Exposure and Accident Rate. Traffic Engineering Control, 36(03), 134-138.

Hedman, K.O., 1990. Road Design and Safety. VTI Report 351 A, 225-238, Swedish Road and Traffic Research Institute, Linkoping, Sweden.

Hoban, C.J., 1982. The Two-and-a-Half Lane Road. Procs 11th Australian road Research Board Conference, 11(04), 59-67.

Hoque, M.M., and Andreassen, D.C., 1986. Pedestrian Accident: An Examination by Road Class, with Reference to Accident Cluster. Traffic Engineering Control, 27(7/8), 391-395.

IHT (Institution of Highways and Transportation), 1990. Highway Safety: Guidelines for Accident Reduction and Prevention (2nd Edition). I.H.T., London.

Jacobs, G.D., Sayer, I.A., and Downing, A.J., 1981. A Preliminary Study of Road User Behaviour in Developing Countries. Transport Road Research Laboratory Report SR646, Crowthorne, UK.

Jain A.K., and Dubes R.C., 1988. Algorithms for Clustering Data. Prentice-Hall, UK.

Jorgenson and Associates, 1978. Cost and Safety Effectiveness of Highway Design Elements. National Cooperative Highway Research Program Report 1997, Transportation Research Board, Washington, USA.

Keall, D.M., Povey, L.J., and Frith, W.J., 2001. The Relative Effectiveness of a Hidden Versus a Visible Speed Camera Programme. Accident Analysis and Prevention, 33(02), 277-284.

Kihlberg, J.A., and Tharp, K.J., 1968. Accident Rate as Related to Design Elements of Rural Highways. National Cooperative Highway Research Program Report 47, Highways Research Board, Washington, USA.

Krammes, R.A., et al, 1995. Horizontal Alignment Design Consistency for Rural Two-Lane Highways. Federal Highway Administration Report FHWA-RD-94-034. Washington D.C., USA.

Lalani, N., and Walker, D., 1981. Correlating Accidents and Volumes at Intersections and on Urban Arterial Street Segments, Traffic Engineering and Control, 22(06), 359-363.

Lamm R., Psarianos B., and Mailaender T., 1999. Highway Design and Traffic Safety Engineering Handbook. McGraw-Hill, New York.

Land Transport Safety Authority, 1994. Accident Investigation System Manual. LTSA, Wellington, NZ.

Land Transport Safety Authority, January, 1998. Overall Results of Crash Investigation Monitoring. LTSA, Wellington, NZ.

Land Transport Safety Authority, 1999. Accident Investigation Monitoring Analysis. LTSA, Wellington, NZ.

Land Transport Safety Authority, March, 2000a. Accident Investigation Monitoring Analysis. LTSA, Wellington, NZ.

Land Transport Safety Authority, October, 2000b. Road Safety Strategy 2010 A consultation Document. LTSA, Wellington, NZ.

Land Transport Safety Authority, February, 2002. Accident Investigation Monitoring Analysis. LTSA, Wellington, NZ.

Land Transport Safety Authority, March, 2003a. Overall Results of Crash Reduction Study Safety Improvements. LTSA, Wellington, NZ.

Land Transport Safety Authority, October, 2003b. Road Safety to 2010. LTSA, Wellington, NZ.

Land Transport Safety Authority, October, 2004. Overall Results of Crash Reduction Study Safety Improvements. LTSA, Wellington, NZ.

LaScala, E.A., Gerber, D., and Gruenewald, P.J., 2000. Demographic and Environmental Correlates of Pedestrian Injury Collisions: A Spatial Analysis. Accident Analysis and Prevention, 32(05), 651-658.

Lay, M.G., 1986. Handbook of Road Technology (Vol. 2). (Gordon and Breach, London).

Layfield, R.E., Summersgill, I., Hall, R.D., and Chatterjee, K., 1996. Accidents at Urban Priority Crossroads and Staggered Junctions. Transport Research Laboratory Report 185, Crowthorne, UK.

Levine, N., Kim, K.E., and Hitz, L.H., 1995. Spatial Analysis of Honolulu Motor Vehicle Crashes: I. Spatial Patterns. Accident Analysis and Prevention, 27(05), 663-674.

Loveday, J., 1989. Spatial Analysis of Road Accident Data and the Accident Migration Hypothesis 21st Annual Conference of The Universities Transport Studies Group, Napier University, Edinburgh.

Loveday, J., 1991. Spatial Autocorrelation and Road Accident Migration. 23rd Annual Conference of the Universities Transport Studies Group, University of Nottingham.

Lynam, D.A., Mackie, A.M., and Davies, C.H., 1988. Urban Safety Project: Design and Implementation of Schemes. Transport and Road Research Laboratory Report RR153.

Mahalel, D., 1986. A Note on Accident Risk. Transportation Research Record, 1068, 85-86.

Maher, M.J., 1987. Accident Migration: A Statistical Explanation? Traffic Engineering Control, 28(09), 480-483.

Maher, M.J., and Mountain, L.J., 1988. The Identification of Accident Blackspots: A Comparison of current Methods. Accident Analysis and Prevention, 20(02), 143-151.

Maycock, G., and Hall, R.D., 1984. Accident at Four Arm Roundabouts. Transport and Road Research Laboratory Report LR1120, Crowthorne, UK.

McBean, P.A., 1982. The Influence of Road Geometry at a Sample of Accident Sites. Transport and Road Research Laboratory Report LR1083, Crowthorne, UK.

McLean, J.R., 1996. Review of Accident and Rural Cross-Section Elements Including Roadsides. ARRB Transport Research Report ARR297, Melbourne, Australia.

Michael, R.A., 1973. Cluster Analysis for Applications. Academic Press Inc, New York.

Michie, J.D., and Bronstad., 1971. M.E. Location Selection and Maintenance of Highway Traffic Barriers. National Cooperative Highway Research Program Repeort 118, Transportation Research Board, Washington D.C., USA

Mood,M.A., Graybill, F.A., and Boes, D.C., 1974. Introduction to the Theory of Statistics, McGraw-Hill, New York.

Mountain, L., Fawaz, B. and Jarrett, D., 1996. Accident Prediction Models for Roads with Minor Junctions. Accident Analysis and Prevention, 28(6), 695-707.

Mullian, B.F.K., and Keese,C.J., 1961. Freeway Traffic Accident Analysis and Safety Study. Highway Research Board Bulletin 291, HRB, Washington D.C., USA.

Navin, F., Zein, S. and Felipe, E., 2000. Road Safety Engineering: an Effective Tool in the Fight Against Whiplash Injuries. Accident Analysis and Prevention, 32 (02), 271-275.

New stead, S.V., Cameron, M.H., Mark, L., and Leggett, W., 2001. The Crash Reduction Effectiveness of a Network-wide Traffic Police Deployment System. Accident Analysis and Prevention, 33(03), 393-406.

Nicholson,A.J., 1986. Estimation of the Underlying True Accident Rate: A New Procedure. Procs 13th Australian Road Research Board Conference, 13(09), 1-11.

Nicholson, A.J., 1987. The Estimate of Accident Rate and Countermeasure Effectiveness. Traffic Engineering and Control, 28(10), 518-523.

Nicholson, A.J., 1989. Accident Clustering; Some Simple Measures. Traffic Engineering Control, 30(05), 241-246.

Nicholson, A.J., 1990. Measure of Accident Clustering. In: M.Koshi (Ed), Transportation and Traffic Theory, Elsevier, New York.

Nicholson, A.J., 1995. Road Accidents and Spatial Point Processes. 7th World conference on Transport Research, Sydney.

Nicholson, A.J., 1998. Selection of the Appropriate Accident Reduction Plan Type. REAAA Conference, Wellington, (Vol. 1), 308-314.

Nicholson, A.J., 1999. Analysis of Spatial Distribution of Accidents. Safety Science (in print) Elsevier, New York.

OECD, 1999. Organisation for Economic Cooperation and Development. Safety Strategies for Rural Roads, OECD, Paris.

Ogden, K.W., 1996. Safer Roads: A Guide to Road Safety Engineering. Avebury Technical, Aldershot, UK.

Olmstead, T., 2001. Freeway Management System and Motor Vehicle Crashes: A Case Study of Phoenix, Arizona. Accident Analysis and Prevention, 33(04), 433-447.

O'Loughlin, J., Flint,C., and Anselin, L., 1994. The Political Geography of the Nazi Vote: Context, Confession and Class in the 1930 Reichstage Election, Annals, Association of American Geographers, 84, 351-80.

Paniati, J.F., and Council, F.M., 1991. The Highway Safety Information System: Application and Future Direction. Public Roads, 54(4), Federal Highway Administration, Washington D.C., USA.

Pearson, E.S., and Hartley, H.O., 1972. Biometrika Tables for Statisticians, (Vol. 2) Cambridge University Press, Cambridge. Peled, A., and Hakkert, A.S., 1993. A PC-Oriented GIS Application for Road Safety Analysis and Management. Traffic Engineering Control, 34(7/8), 355-361.

Porter, B. E., Berry, D. T., 2001. A Nation Wide Survey of Self-reported Red Light Running: Measuring Prevalence, Predictors, and Perceived Consequences. Accident Analysis and Prevention, 33(06), 735-741.

Press, H. W., Teukolsky, S.A., Vetterling, W. T., and Flannery, B. P., 1992. Numerical Recipes in Fortran, University of Cambridge, Cambridge University Press.

Raff, M.S., 1953. Interstate Highway Accident Study. Highway Research Bulletin, 74, 18-456.

Ripley, B.D., 1981. Spatial Statistics. Wiley, New York.

Road research Laboratory, 1965. Research on Road Traffic. HMSO, London.

Satterthwaite, S.P., 1981. A Survey of Research into Relationship Between Traffic Accidents and Traffic Volumes. Transport and Research Laboratory Report SR682.

Shaikh, N.S., 1990. An Investigation of Accident Clustering. ME Thesis, Department of Civil Engineering, University of Canterbury, Christchurch, NZ.

Silcock, D.T., and Smyth, A.W., 1984. The Method Used by British Highway Authority to Identify Accident Black spots. Traffic Engineering Control, 25(11), 542-545.

Silcock,D.T., and Worsey, G.M., 1982 Relationships Between Accident Rates, Road Characteristics and Traffic on Two Urban Routes. Transport Operations Research Group Report No. 40, University of Newcastle-upon-Tyne, Newcastle-upon-Tyne, UK.

Solomon, D., 1988. Historical Changes in Highway Safety: 1966-1987.Highway Safety: at Cross Roads Conference, San Antonio 10-24.

Stephens, M.A., 1969. Test for Randomness of Directions Against Two Circular Alternatives. Journal of American Statistical Association, 64(325), 280-289.

Summersgill, I., and Layfield, R.E., 1996. Non-junction Accidents on Urban Single-Carriageway Roads. Transport Research Laboratory Report 183

Tanner, J.C., 1953. Accident at Rural 3-way Junctions. Journal of Institute of Highway Engineers, 2(11), 56-67.

Taylor, M., 2000. A Closer Look at Speed and Accident Frequency. Traffic Engineering Control, 41(04), 103-131.

Thomas, I., 1996. Spatial Data Aggregation: Exploratory Analysis of Road Accidents. Accident Analysis and Prevention,. 28 (02), 251-264.

Transfund New Zealand, 1997. Project Evaluation Manual. Transfund New Zealand, Wellington, New Zealand.

Turner, S., and Nicholson, A.J., 1998. Intersection Accident Estimation: The Role of Intersection Location and Non-Collision flows. Accident Analysis and Prevention, 30(04), 505-5017.

Upton, G.J.G., and Fingleton, B., 1989. Spatial Data Analysis by Example (Vol. 2) Categorical and Directional Data. Wiley, New York.

Vey, A.H., 1937. Relationship Between Daily Traffic and Accident Rates. American City, 52(19), 119.

Walz, F.H., Hogfliger, M., and Fehlmann, W., 1983. Speed Limit Reduction from 60 to 50 km/h and Pedestrian Injuries. Procs of 27th Stapp Car Crash Conference, 311-318, Society of Automotive Engineering, Warrendale, USA.

Wilkie, D., 1983. Releigh Test for Randomness of Circular Data. Applied Statistics, 32, 311-312.

Wright, P.H., and Robertson, L.S., 1976 Priorities for Roadside Hazard Modification. Traffic Engineering, 46(08), 24-30.

Zegeer, C.V., 1982. Highway Accident Analysis Systems. Transportation Research Board, National Cooperative Highway Research Program, Synthesis of Highway Practice 91.

Zegreer, C.V., Hummer, J., Reinfurt, D., Herf, L., and Hunter, W., 1987. Safety Effect of Cross-section Design for Two-lane Roads. Federal Highway Administration Report FHWARD-87-008, Washington D.C., USA.

Appendix -A

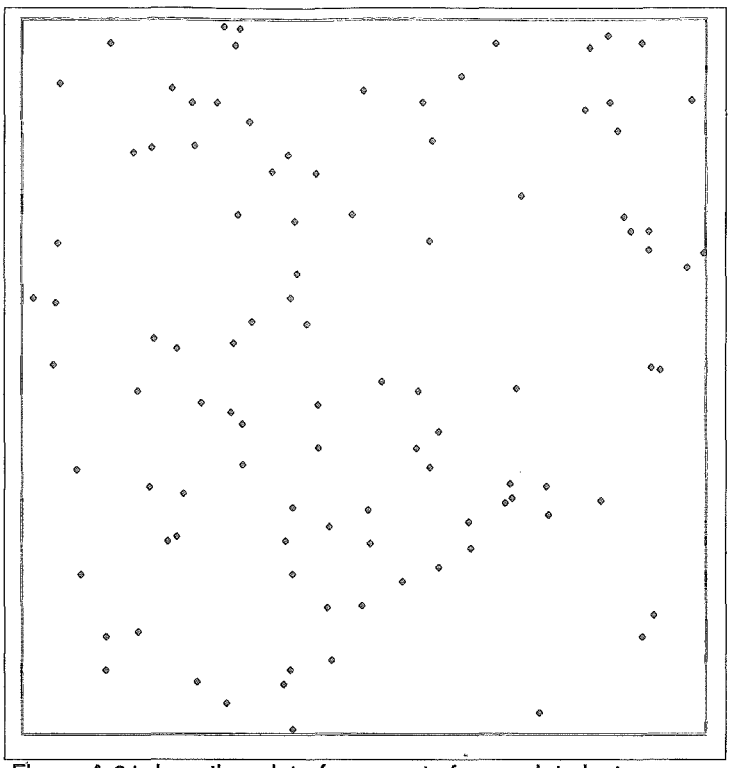


Figure A.01: Location plot of no events from point cluster and 100 events from CSR distribution

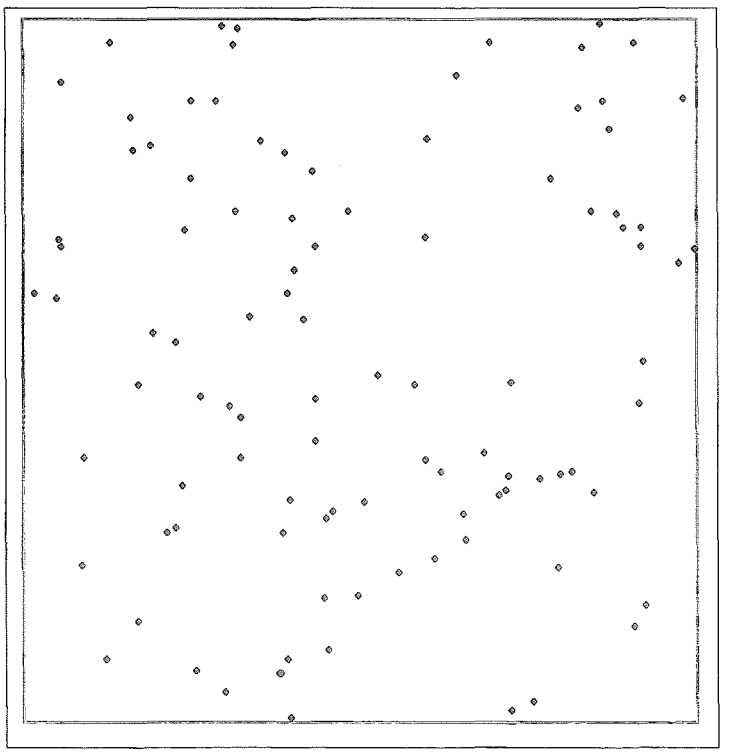


Figure A.02: Location plot of 20 events from point cluster and 80 events from CSR distribution

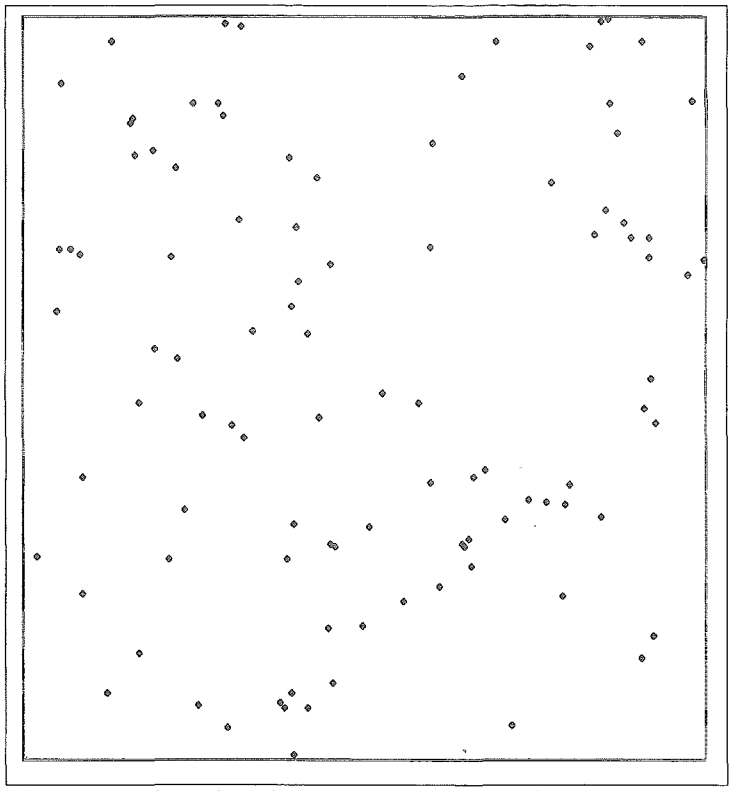


Figure A.03: Location plot of 30 events from point cluster and 70 events from CSR distribution

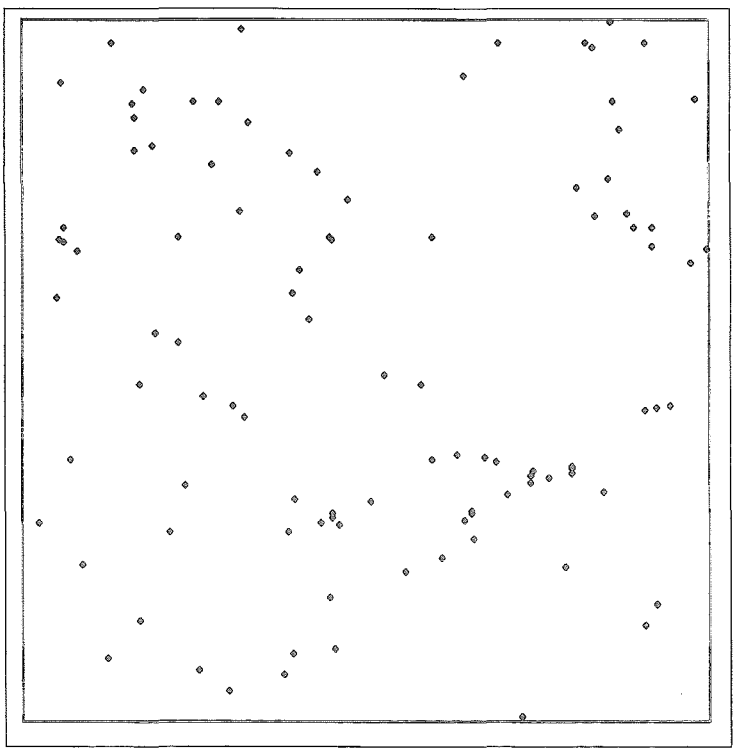


Figure A.04: Location plot of 40 events from point cluster and 60 events from CSR distribution

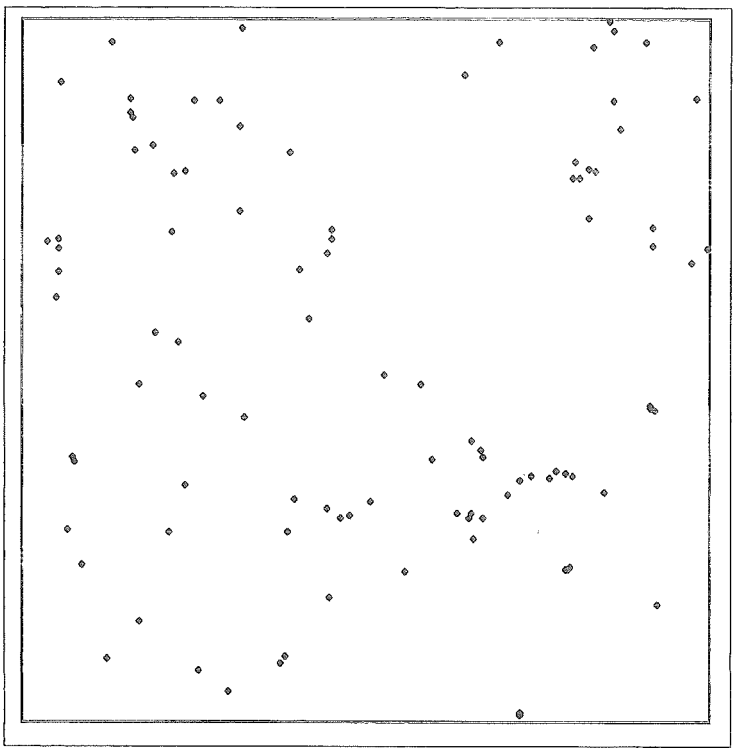


Figure A.05: Location plot of 50 events from point cluster and 50 events from CSR distribution

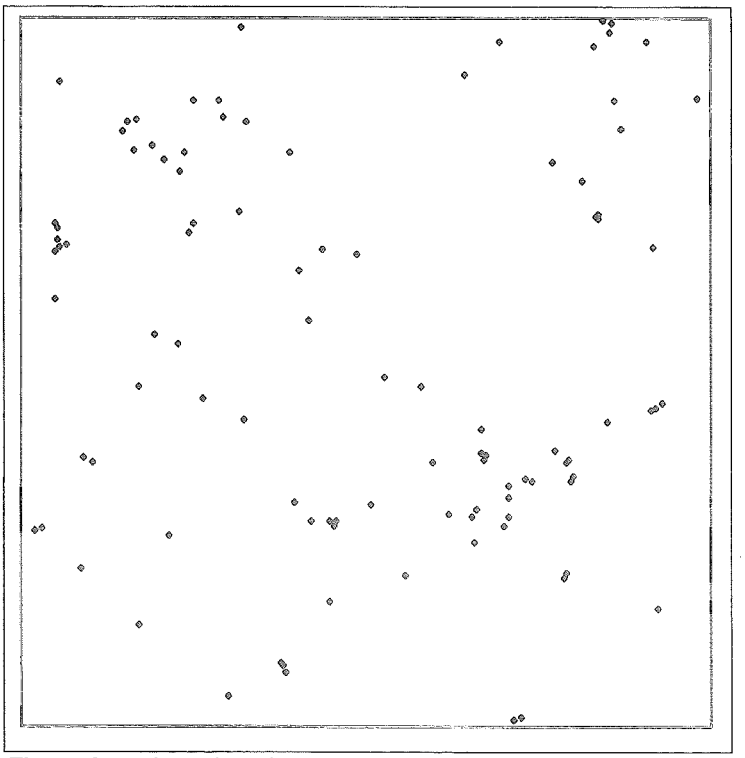


Figure A.06: Location plot of 60 events from point cluster and 40 events from CSR distribution

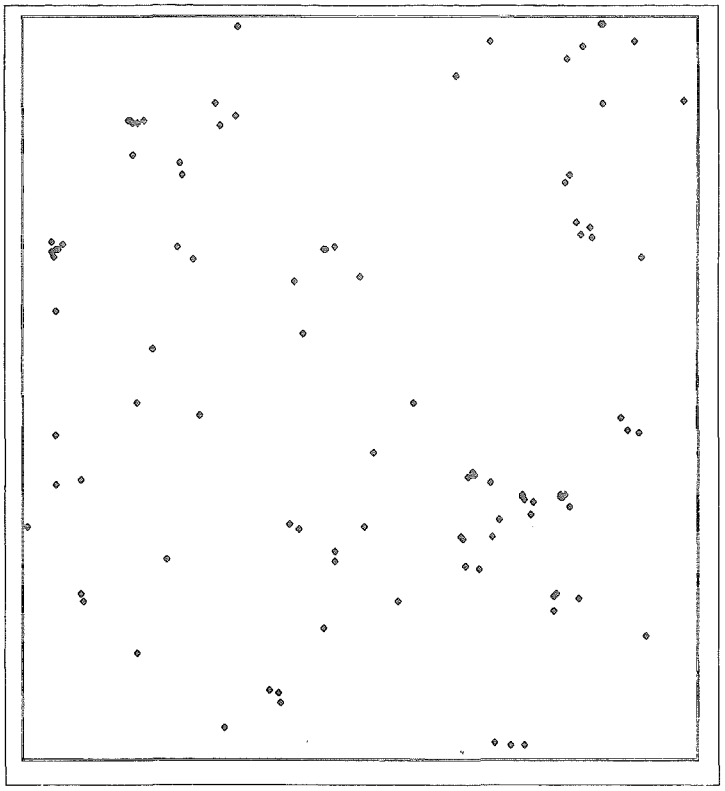


Figure A.07: Location plot of 70 events from point cluster and 30 events from CSR distribution

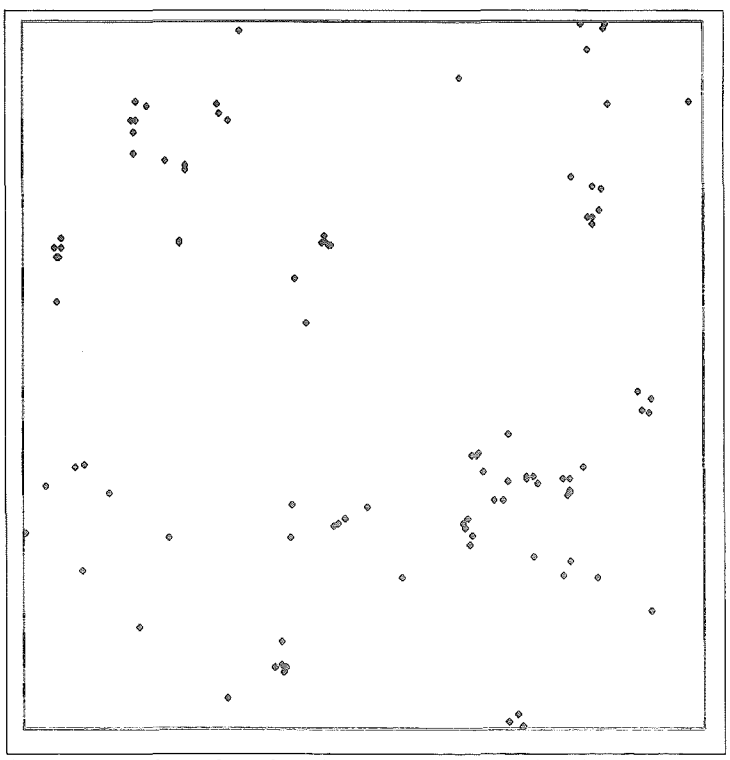


Figure A.08: Location plot of 80 events from point cluster and 20 events from CSR distribution

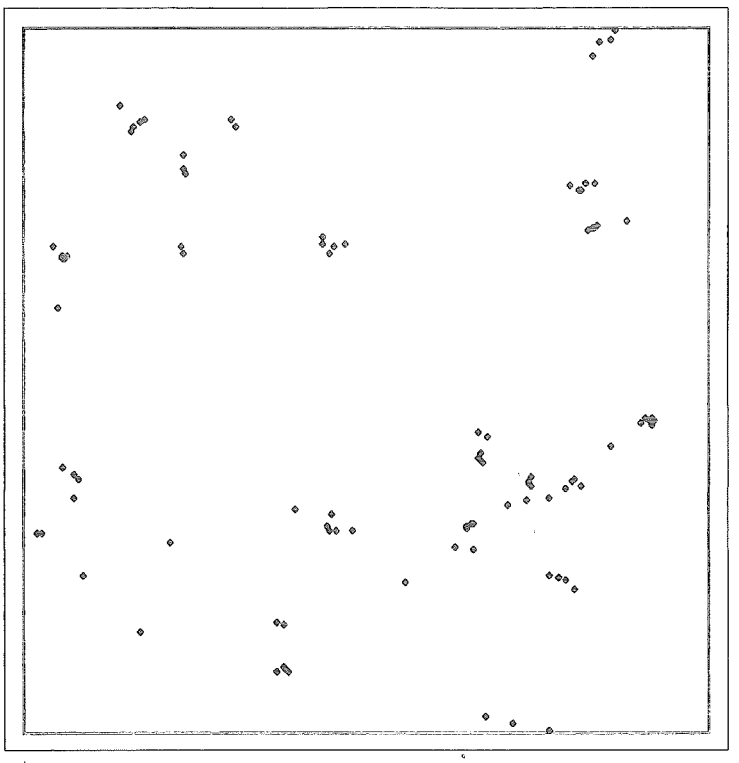


Figure A.09: Location plot of 90 events from point cluster and 10 events from CSR distribution

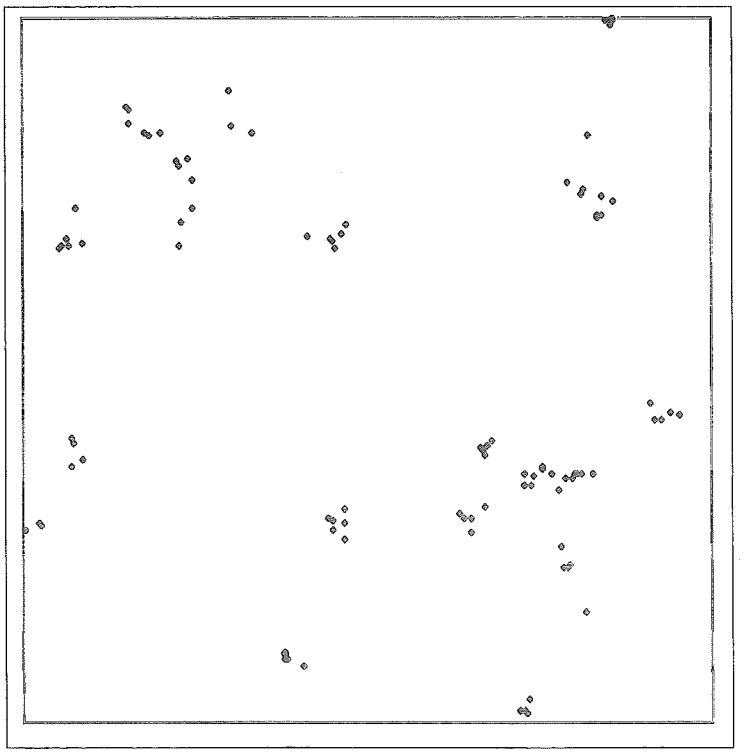
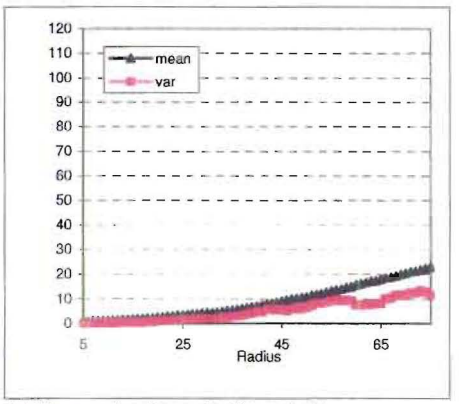
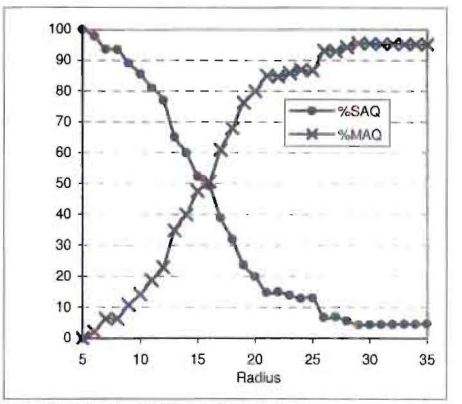


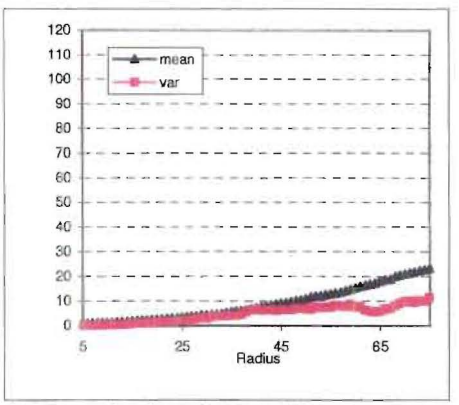
Figure A.10: Location plot of 100 events from point cluster and no event from CSR distribution



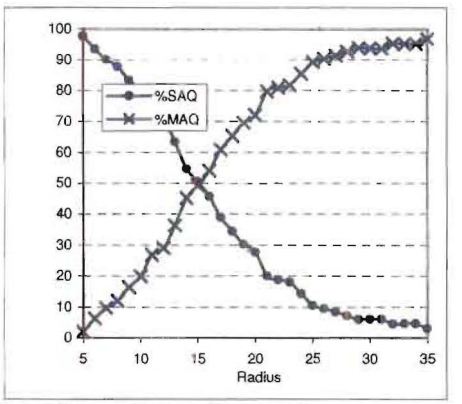
(a) Mean and variance for Figure A.01
(no events from point cluster and
100 events from CSR distribution)



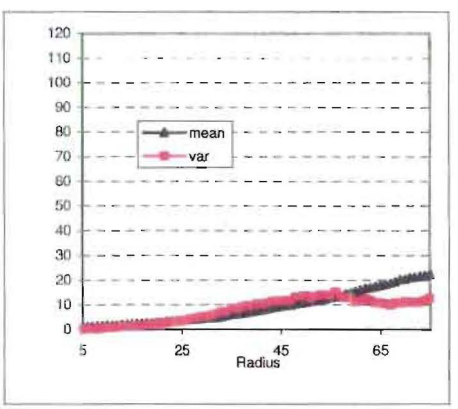
(b) %SAQ and %MAQ for Figure A.01
(no events from point cluster and
100 events from CSR distribution)



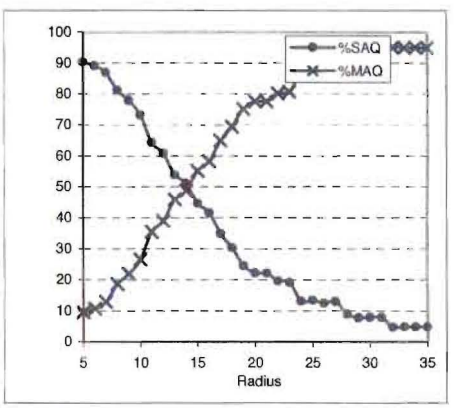
(c) Mean and variance for Figure A.02
(20 events from point cluster and
80 events from CSR distribution)



(d) %SAQ and %MAQ for Figure A.02
(20 events from point cluster and
80 events from CSR distribution)



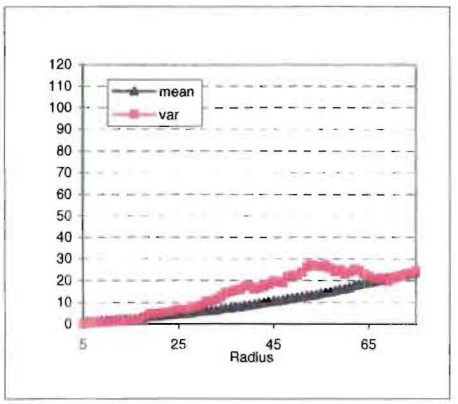
(e) Mean and variance for Figure A.03
(30 events from point cluster and
70 events from CSR distribution)



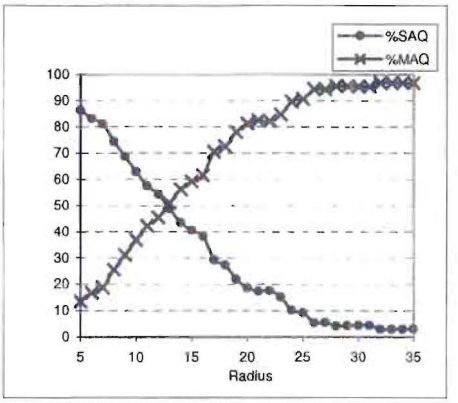
(f) %SAQ and %MAQ for Figure A.03
(30 events from point cluster and
70 events from CSR distribution)

%SAQ - Percentage of single accident quadrats %MAQ - Percentage of multiple accident quadrats

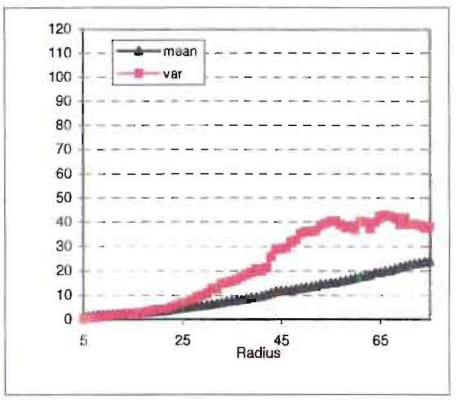
Figure A.11: Variation of mean, variance, %SAQ and %MAQ with increasing quadrat radius for the distributions (Figures A.01, A.02 and A.03).



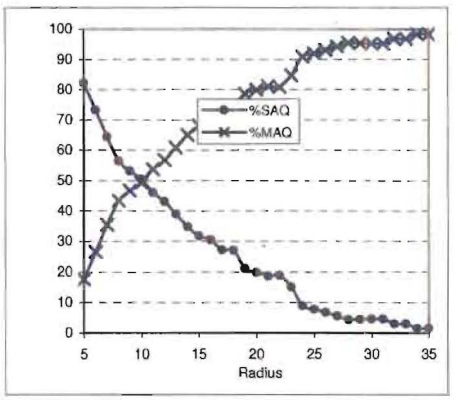
(a) Mean and variance for Figure A.04
(40 events from point cluster and
60 events from CSR distribution)



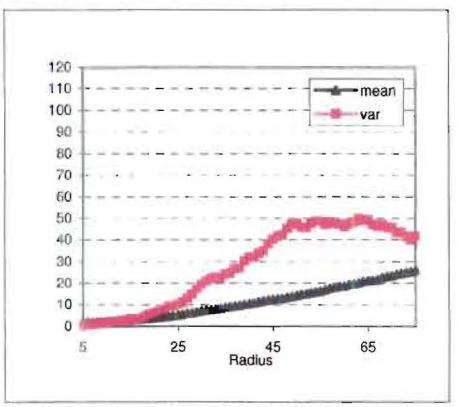
(b) %SAQ and %MAQ for Figure A.04
(40 events from point cluster and
60 events from CSR distribution)



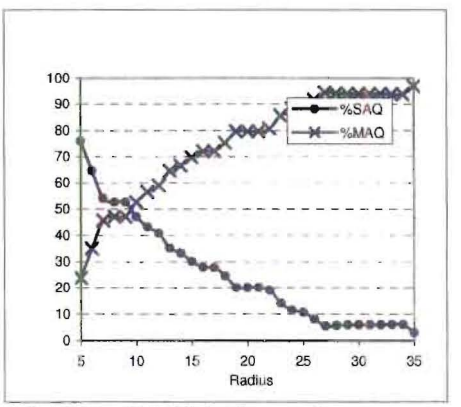
(c) Mean and variance for Figure A.05
(50 events from point cluster and
50 events from CSR distribution)



(d) %SAQ and %MAQ for Figure A.05
(50 events from point cluster and
50 events from CSR distribution)



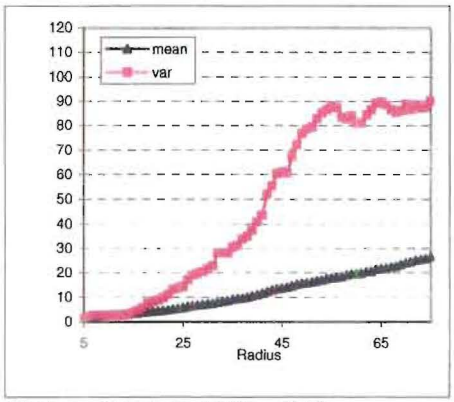
(c) Mean and variance for Figure A.06
(60 events from point cluster and
40 events from CSR distribution)



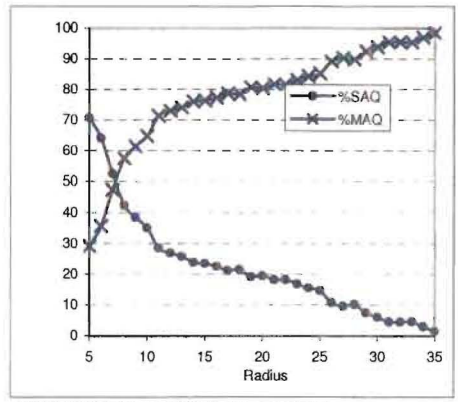
(d) %SAQ and %MAQ for Figure A.06
(60 events from point cluster and
40 events from CSR distribution)

%SAQ - Percentage of single accident quadrats %MAQ - Percentage of multiple accident quadrats

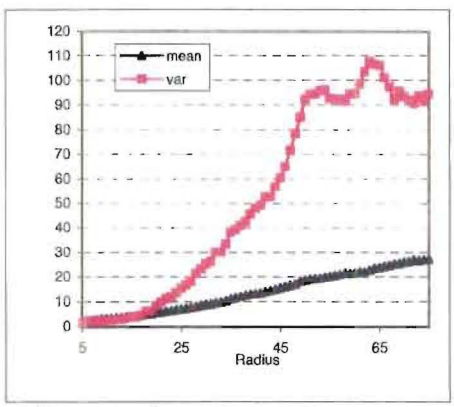
Figure A.12: Variation of mean, variance, %SAQ and %MAQ with increasing quadrat radius for the distributions (Figures A.04, A.05 and A.06).



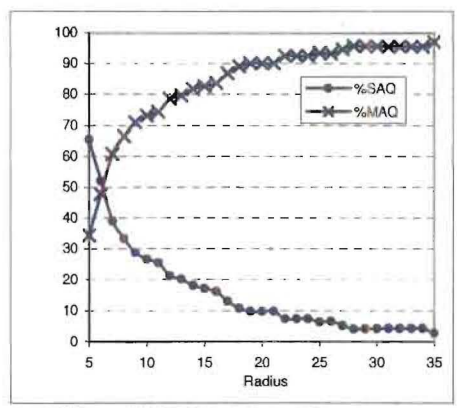
(a) Mean and variance for Figure A.07
(70 events from point cluster and
30 events from CSR distribution)



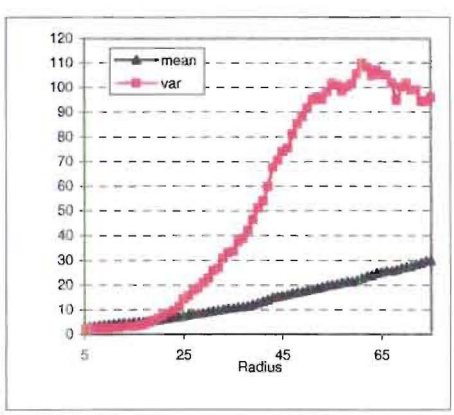
(b) %SAQ and %MAQ for Figure A.07
(70 events from point cluster and
30 events from CSR distribution)



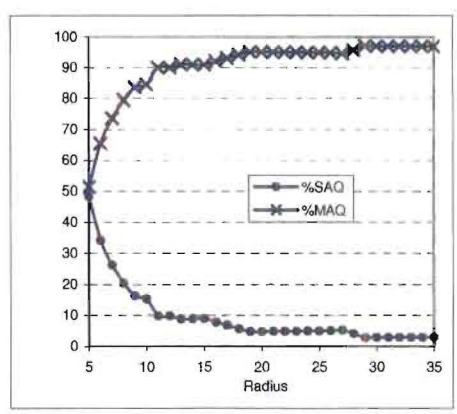
(c) Mean and variance for Figure A.08
(80 events from point cluster and
20 events from CSR distribution)



(d) %SAQ and %MAQ for Figure A.08
(80 events from point cluster and
20 events from CSR distribution)



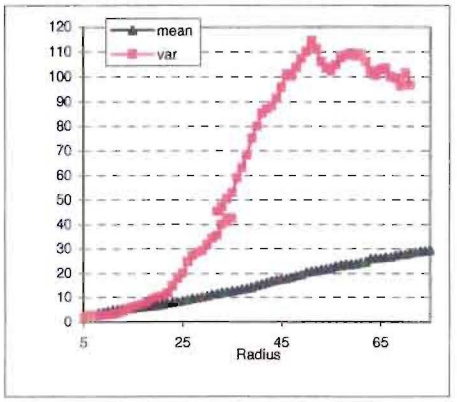
(e) Mean and variance for Figure A.09
(90 events from point cluster and
10 events from CSR distribution)



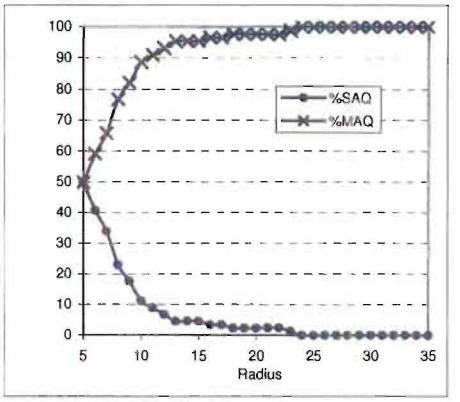
(f) %SAQ and %MAQ for Figure A.09
(90 events from point cluster and
10 events from CSR distribution)

%SAQ - Percentage of single accident quadrats %MAQ - Percentage of multiple accident quadrats

Figure A.13: Variation of mean, variance, %SAQ and %MAQ with increasing quadrat radius for the distributions (Figures A.07, A.08 and A.09).



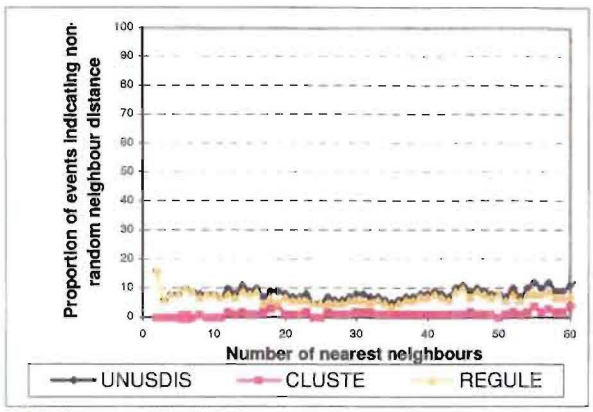
(a) Mean and variance for Figure A.10 (100 events from point cluster and no events from CSR distribution)



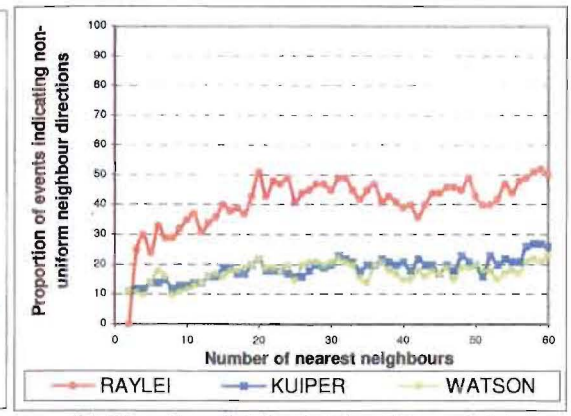
(b) %SAQ and %MAQ for Figure A.10 (100 events from point cluster and no events from CSR distribution)

%SAQ - Percentage of single accident quadrats %MAQ - Percentage of multiple accident quadrats

Figure A.14: Variation of mean, variance, %SAQ and %MAQ with increasing quadrat radius for the distributions (Figure A.10).

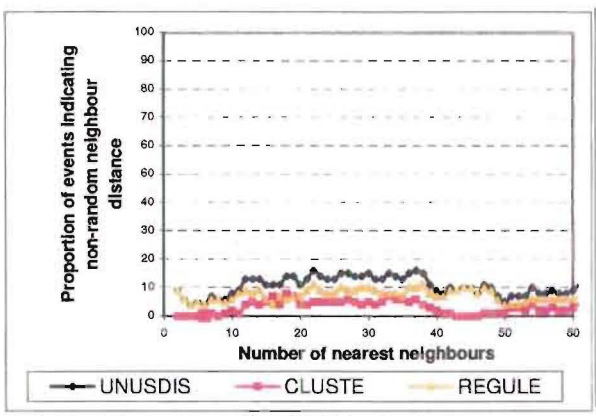


(a) Distance distribution for Figure A.01

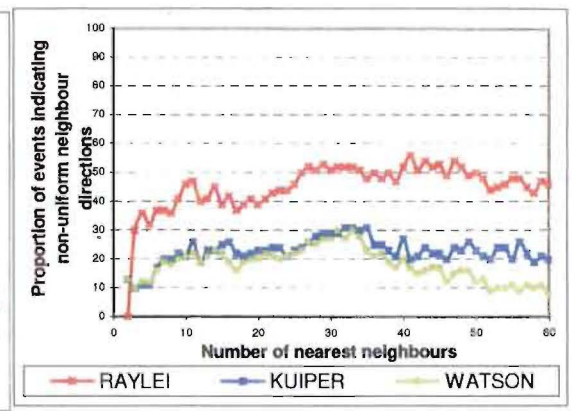


(b) Direction distribution for Figure A.01

(100 events from CSR and no events from point cluster distributions)

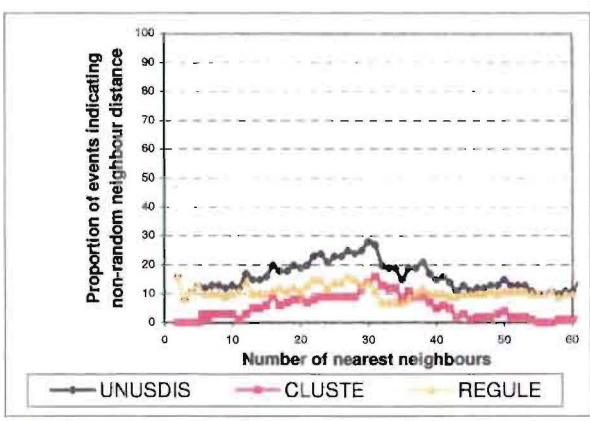


(c) Distance distribution for Figure A.02

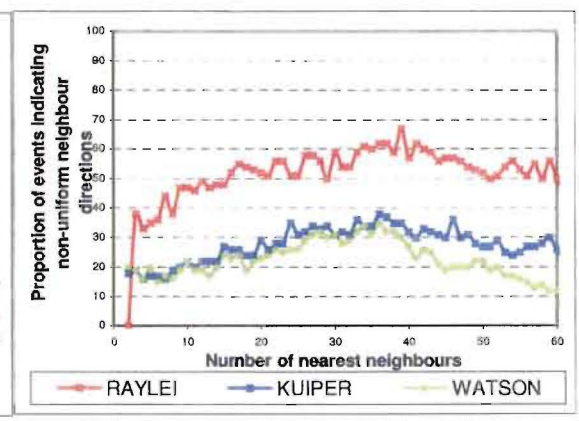


(d) Direction distribution for Figure A.02

(80 events from CSR and 20 events from pointcluster distributions)



(e) Distance distribution for Figure A.03



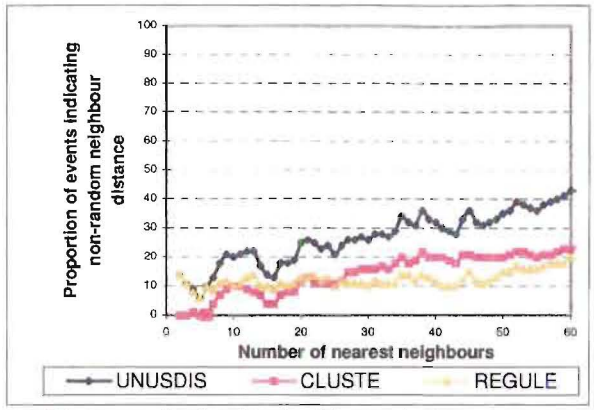
(f) Direction distribution for Figure A.03

(70 events from CSR and 30 events from point cluster distributions)

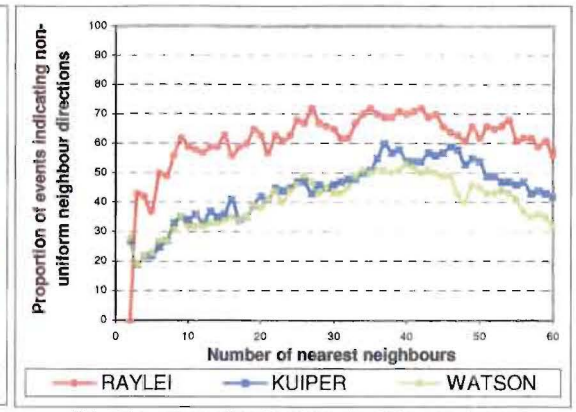
UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure A.15: Nearest-neighbour analysis results for mixed distributions (Figures A.01, A.02 and A.03).

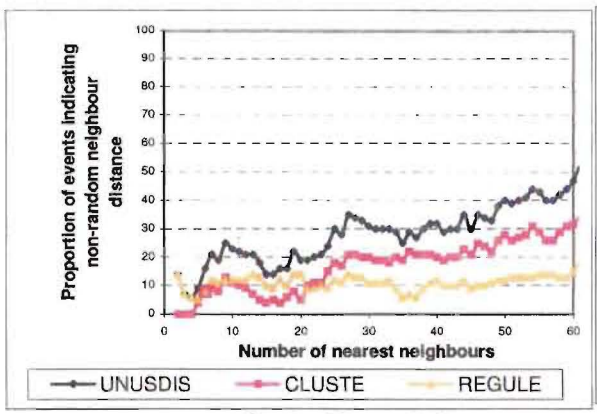


(a) Distance distribution for Figure A.04

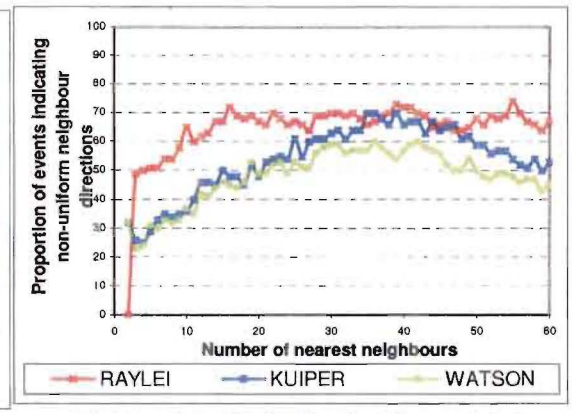


(b) Direction distribution for Figure A.04

(60 events from CSR and 40 events from point cluster distributions)

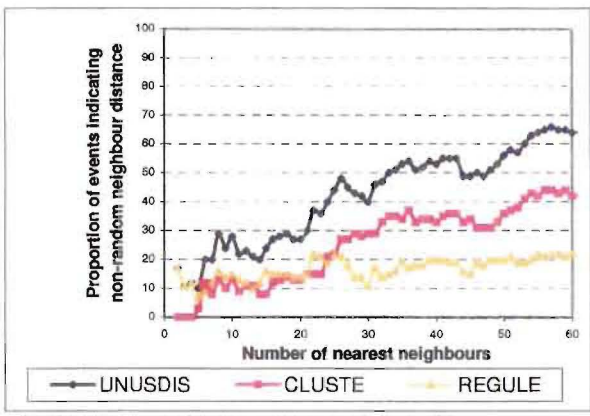


(c) Distance distribution for Figure A.05

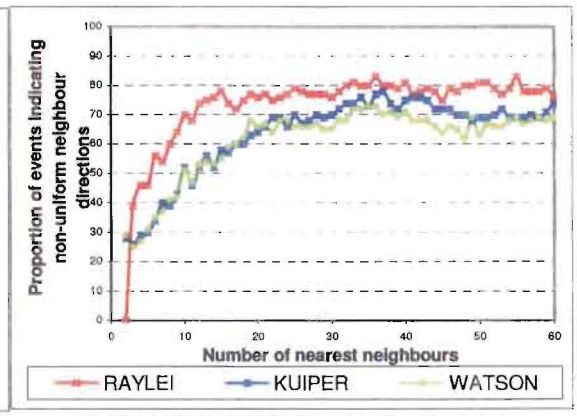


(d) Direction distribution for Figure A.05

(50 events from CSR and 50 events from pointcluster distributions)



(e) Distance distribution for Figure A.06



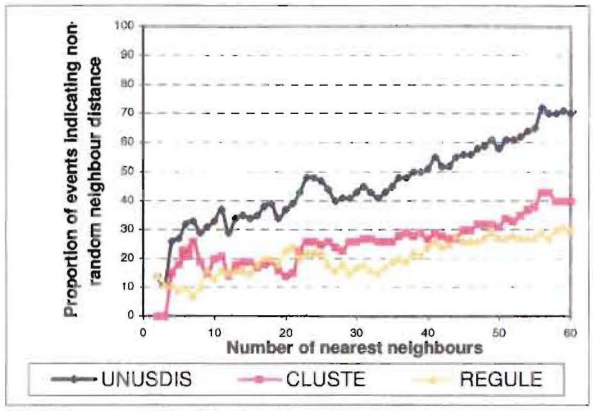
(f) Direction distribution for Figure A.06

(40 events from CSR and 60 events from point cluster distributions)

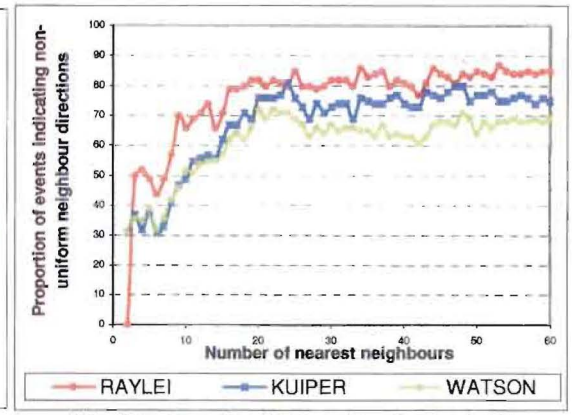
UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure A.16: Nearest-neighbour analysis results for mixed distributions (Figures A.04, A.05 and A.06).

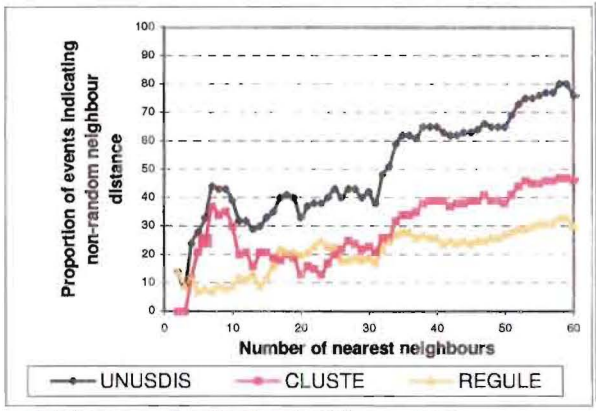


(a) Distance distribution for Figure A.07

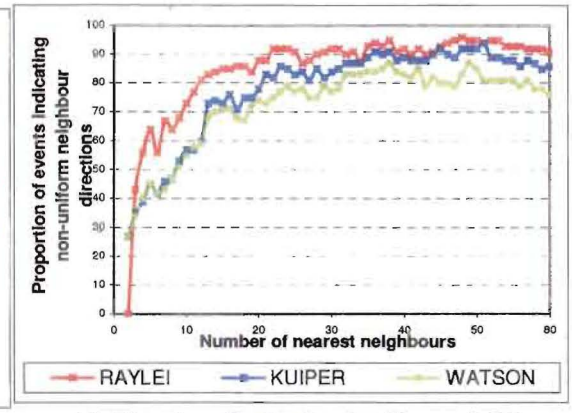


(b) Direction distribution for Figure A.07

(30 events from CSR and 70 events from point cluster distributions)

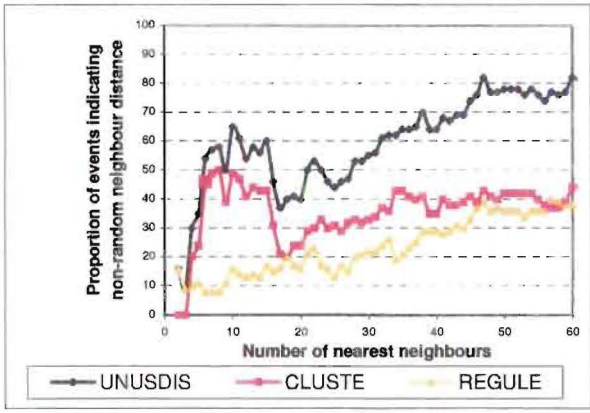


(c) Distance distribution for Figure A.08

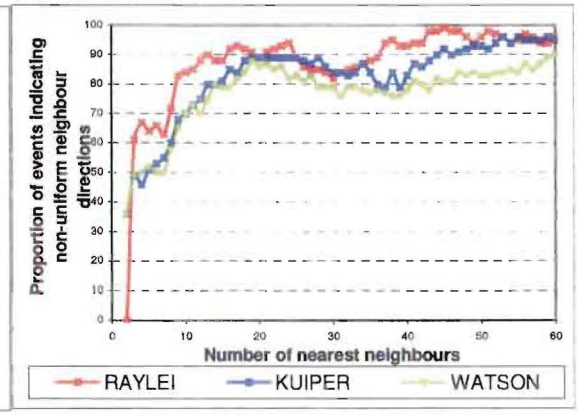


(d) Direction distribution for Figure A.08

(20 events from CSR and 80 events from pointcluster distributions)



(e) Distance distribution for Figure A.09



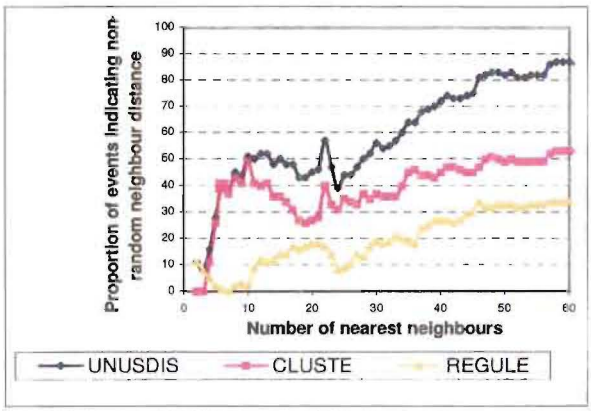
(f) Direction distribution for Figure A.09

(10 events from CSR and 90 events from point cluster distributions)

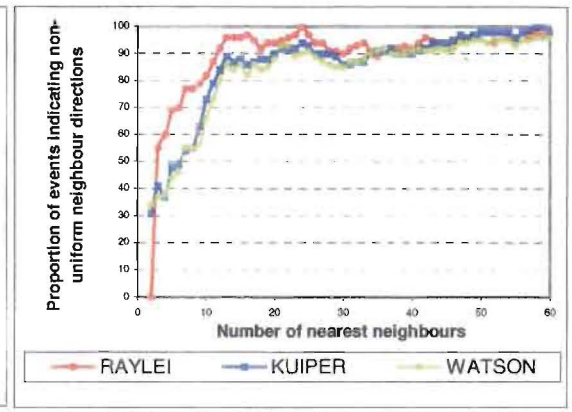
UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure A.17: Nearest-neighbour analysis results for mixed distributions (Figures A.07, A.08 and A.09).



(a) Distance distribution for Figure A.10



(b) Direction distribution for Figure A.10

(no events from CSR and 100 events from point cluster distributions)

UNUSDIS---	Unusual distance	RAYLEI---	Rayleigh test results
CLUSTE---	Cluster	KUIPER---	Kuiper test results
REGULE---	Regular	WATSON---	Watson test results

Figure A.18: Nearest-neighbour analysis results for mixed distribution (Figure A.10).

Appendix - B

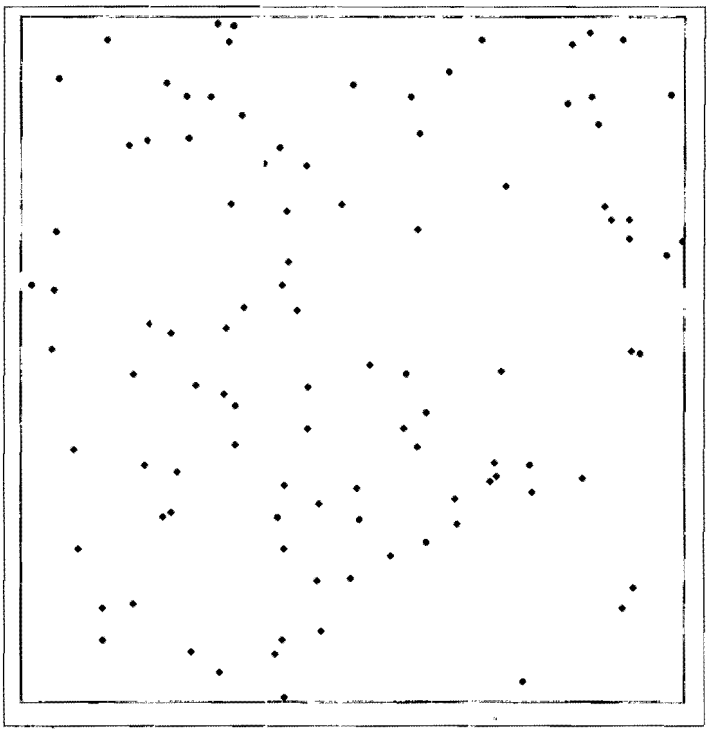


Figure B.01: Location plot of no events from line cluster and 100 events from CSR distribution

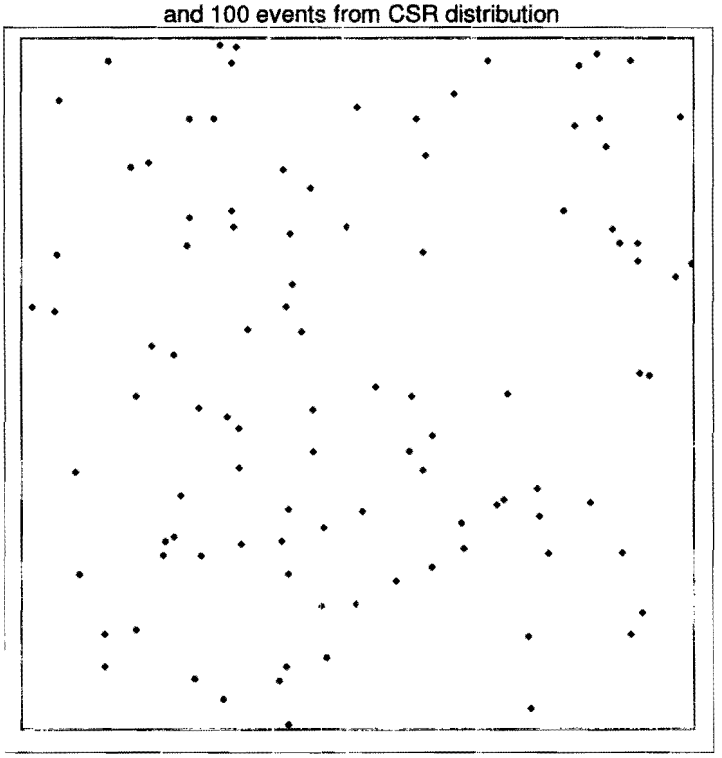


Figure B.02: Location plot of 10 events from line cluster and 90 event from CSR distribution

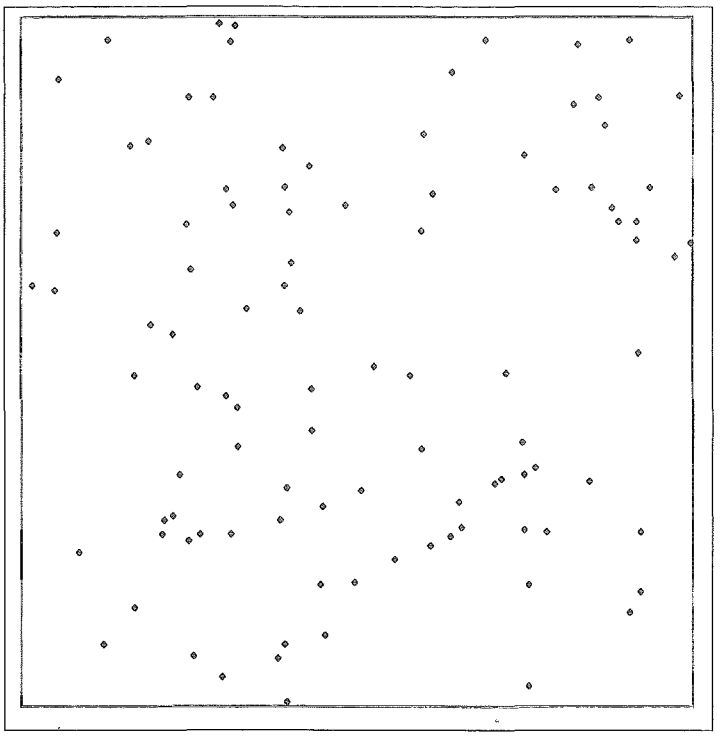


Figure B.03: Location plot of 20 events from line cluster and 80 event from CSR distribution

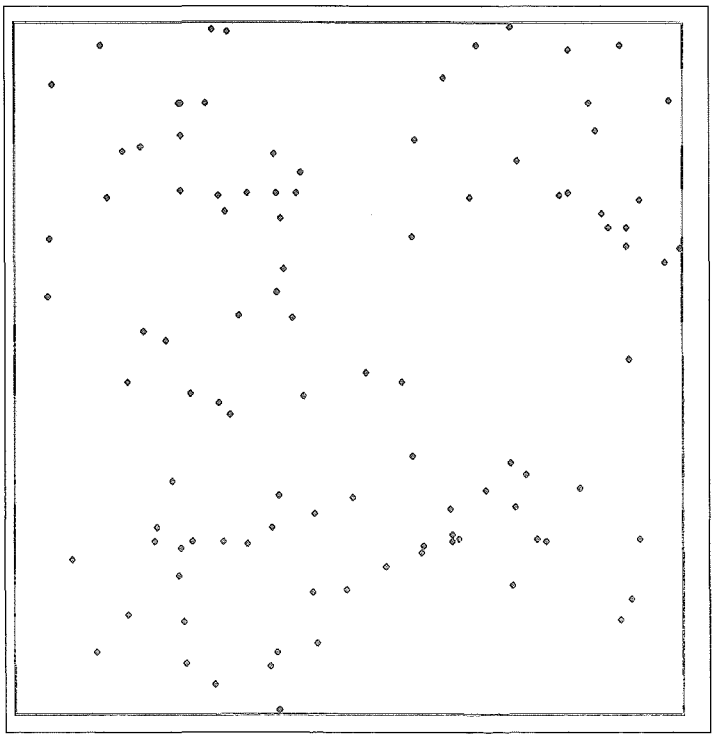


Figure B.04: Location plot of 30 events from line cluster and 70 events from CSR distribution

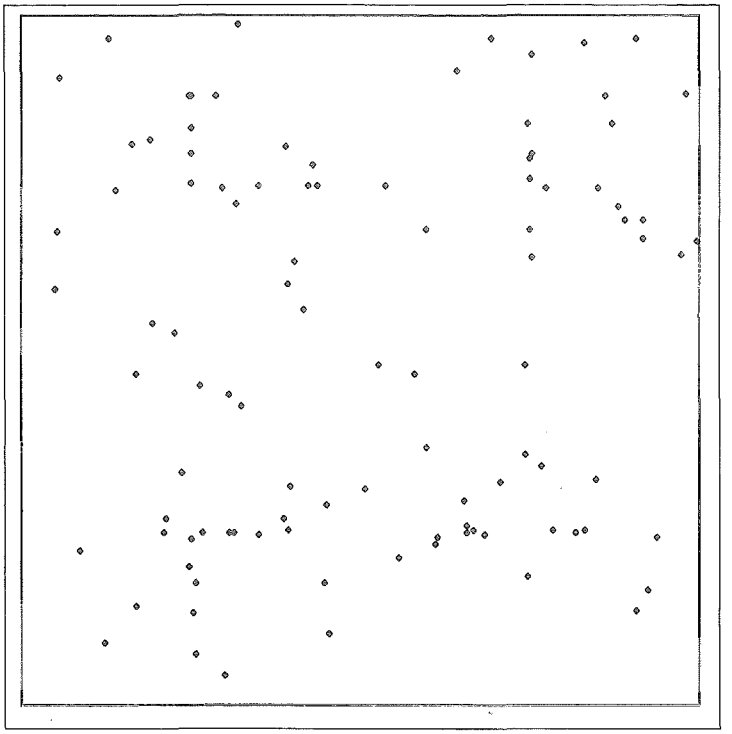


Figure B.05: Location plot of 40 events from line cluster and 60 events from CSR distribution

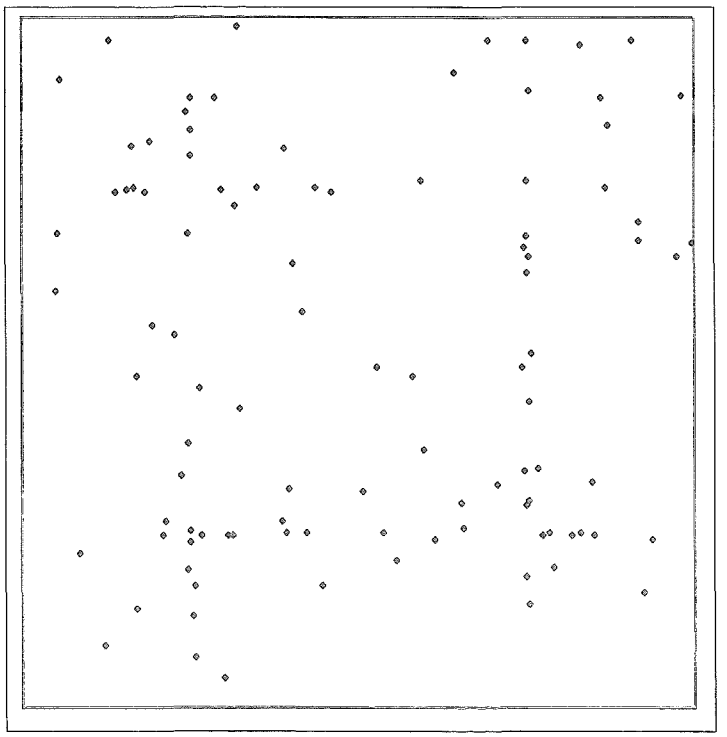


Figure B.06: Location plot of 50 events from line cluster and 50 events from CSR distribution

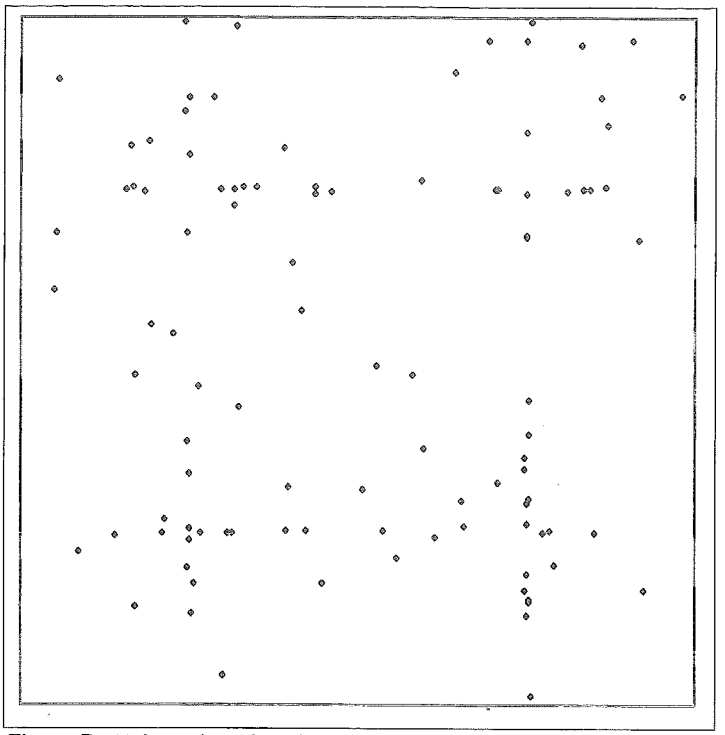


Figure B.07: Location plot of 60 events from line cluster and 40 event from CSR distribution

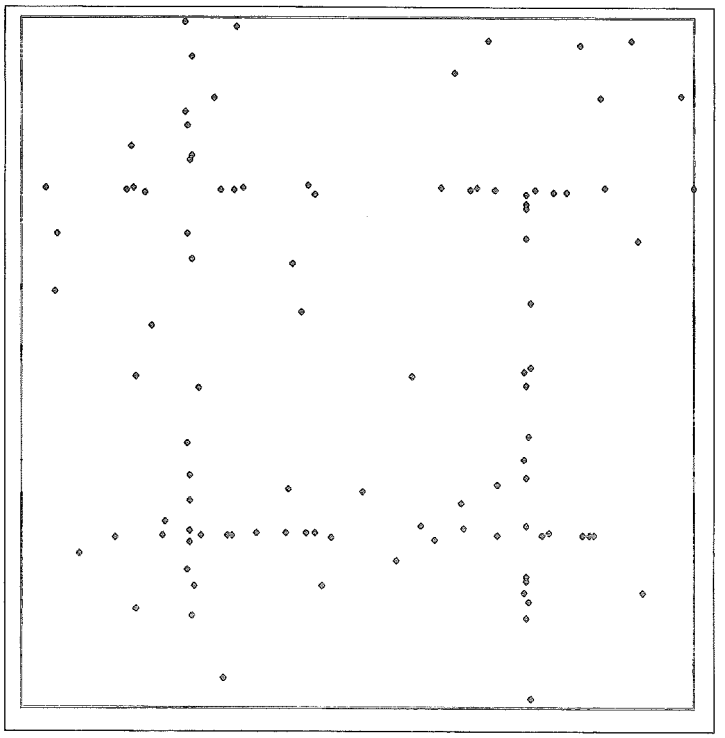


Figure B.08: Location plot of 70 events from line cluster and 30 events from CSR distribution

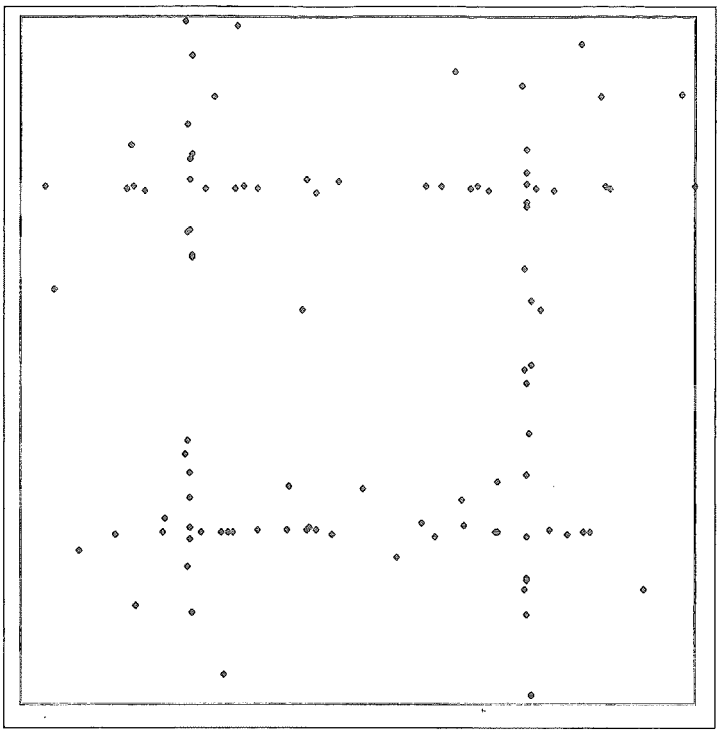


Figure B.09: Location plot of 80 events from line cluster and 20 events from CSR distribution

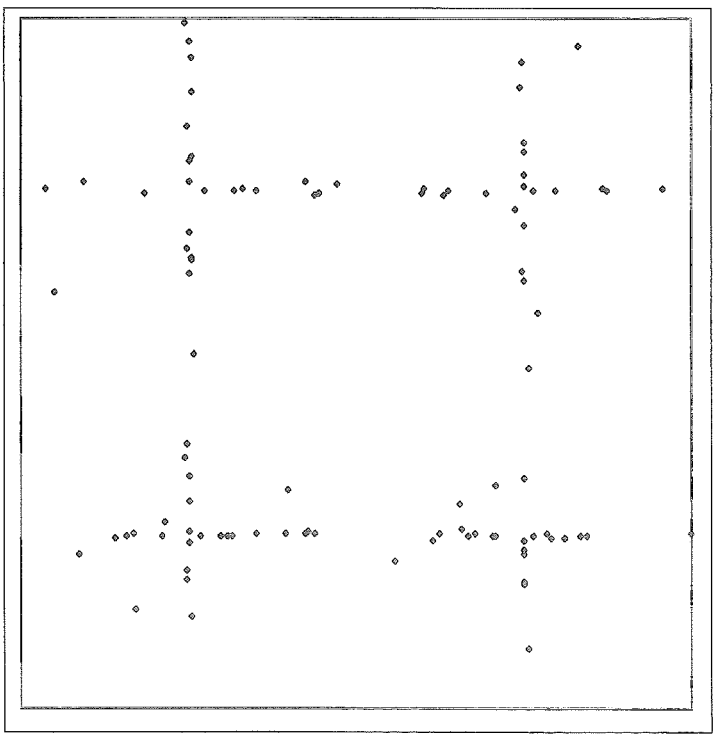


Figure B.10: Location plot of 90 events from line cluster and 10 events from CSR distribution

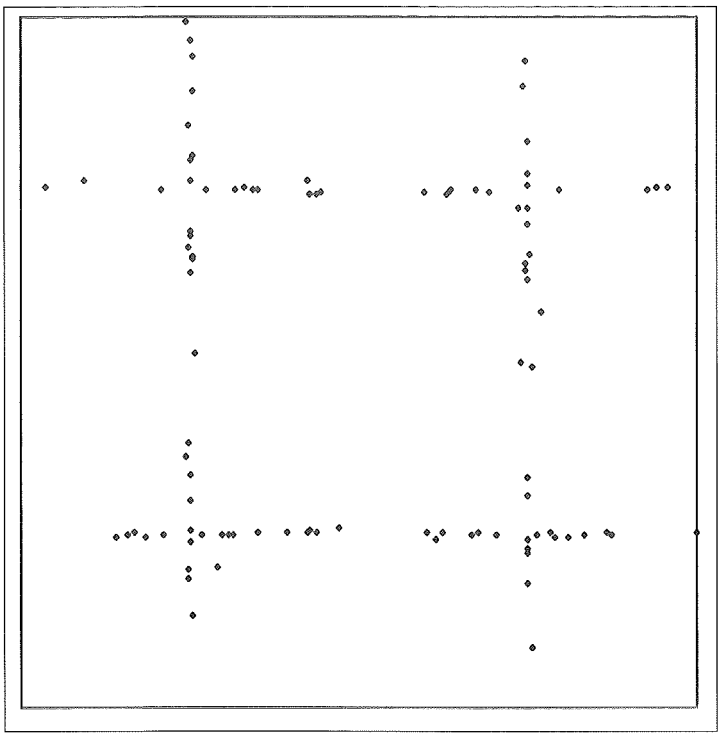
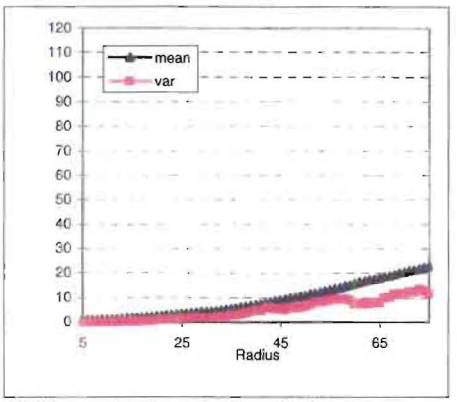
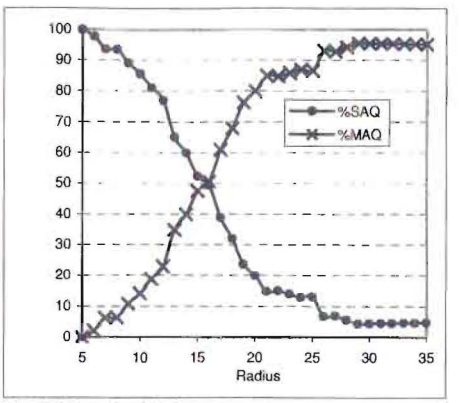


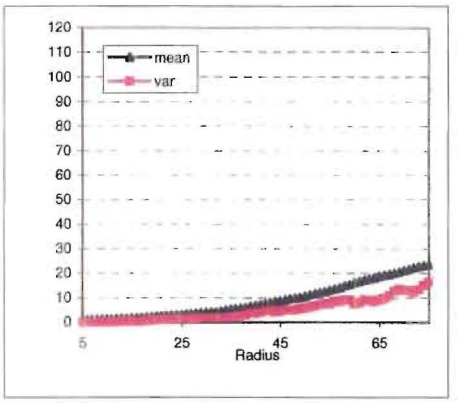
Figure B.11: Location plot of 100 events from line cluster and no event from CSR distribution



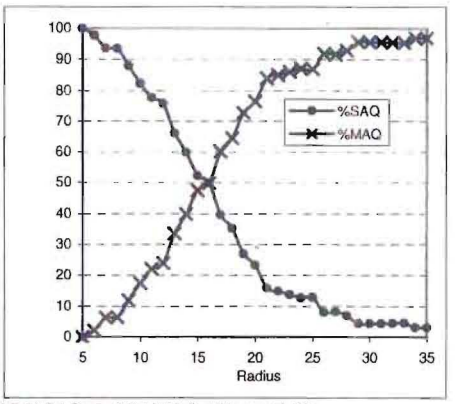
(a) Mean and variance for Figure B.01
(no events from line cluster and
100 events from CSR distribution)



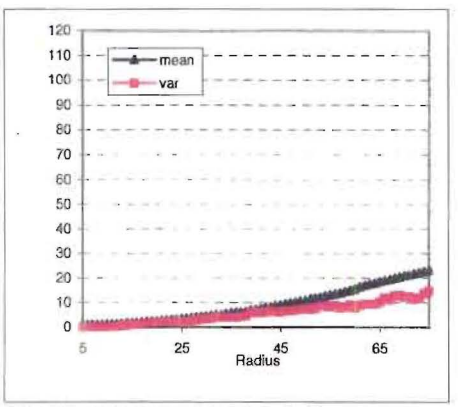
(b) %SAQ and %MAQ for Figure B.01
(no events from line cluster and
100 events from CSR distribution)



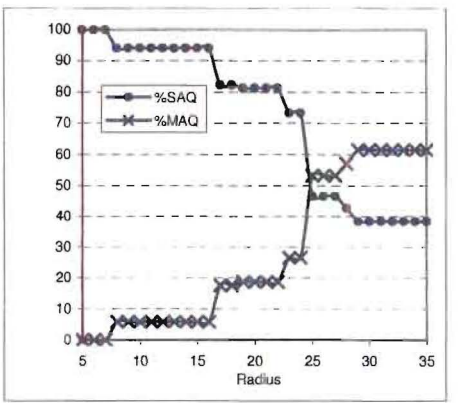
(c) Mean and variance for Figure B.02
(10 events from line cluster and
100 events from CSR distribution)



(d) %SAQ and %MAQ for Figure B.02
(10 events from line cluster and
100 events from CSR distribution)



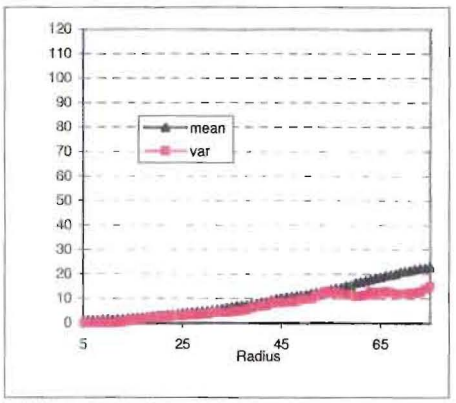
(e) Mean and variance for Figure B.03
(20 events from line cluster and
80 events from CSR distribution)



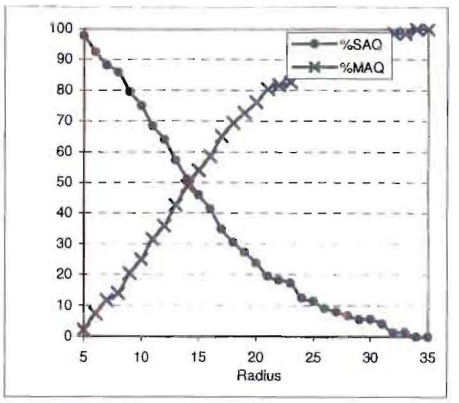
(f) %SAQ and %MAQ for Figure B.03
(20 events from line cluster and
80 events from CSR distribution)

%SAQ - Percentage of single accident quadrats %MAQ - Percentage of multiple accident quadrats

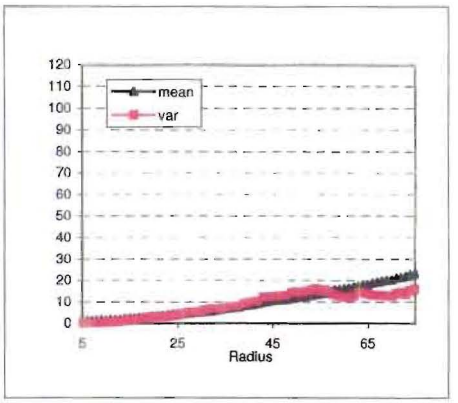
Figure B.12: Variation of mean, variance, %SAQ and %MAQ with increasing quadrat radius for the distributions (Figure B.01, B.02 and B.03).



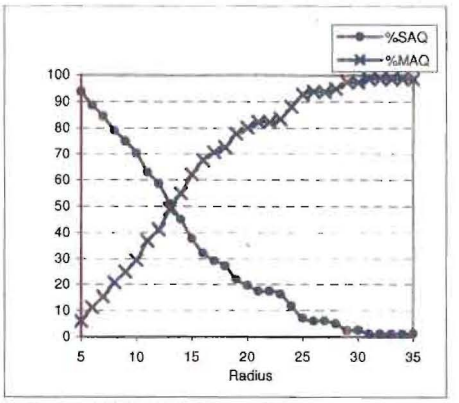
(a) Mean and variance for Figure B.04
(30 events from line cluster and
70 events from CSR distribution)



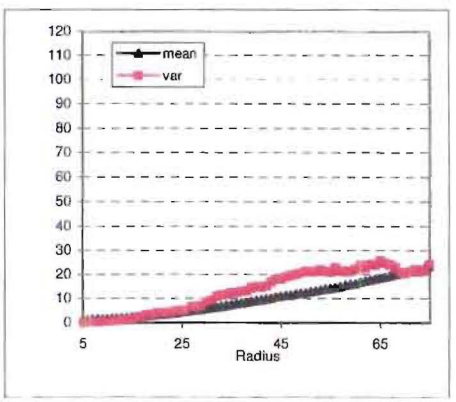
(b) %SAQ and %MAQ for Figure B.04
(30 events from line cluster and
70 events from CSR distribution)



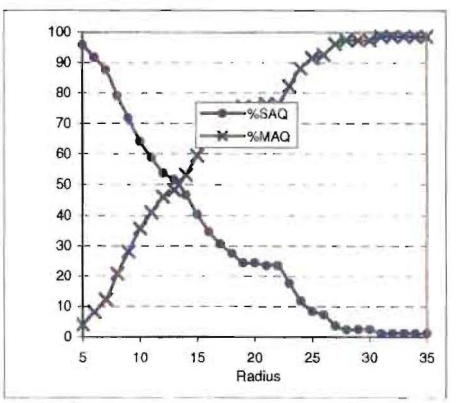
(c) Mean and variance for Figure B.05
(40 events from line cluster and
60 events from CSR distribution)



(d) %SAQ and %MAQ forFigure B.05
(40 events from line cluster and
60 events from CSR distribution)



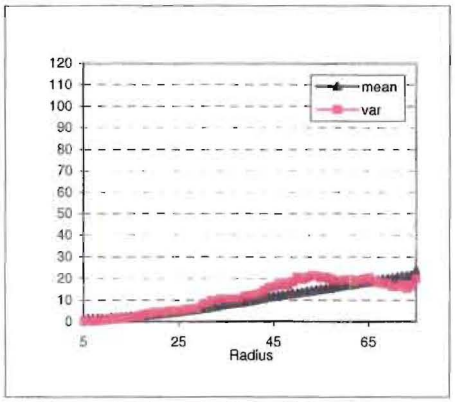
(e) Mean and variance for Figure B.06
(50 events from line cluster and
50 events from CSR distribution)



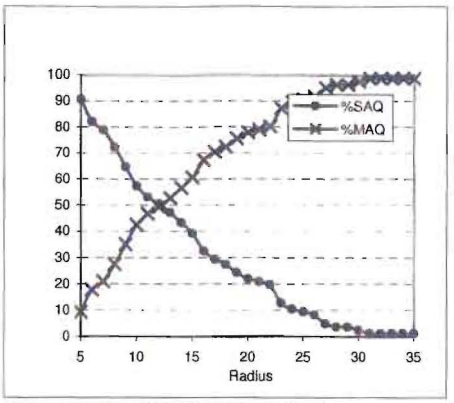
(f) %SAQ and %MAQ for Figure B.06
(50 events from line cluster and
50 events from CSR distribution)

%SAQ - Percentage of single accident quadrats %MAQ - Percentage of multiple accident quadrats

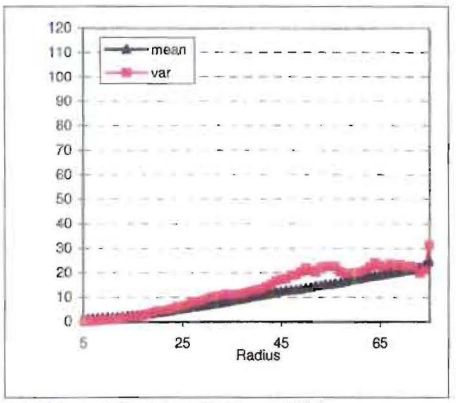
Figure B.13: Variation of mean, variance, %SAQ and %MAQ with increasing quadrat radius for the distributions (Figures B.04, B.05 and B.06).



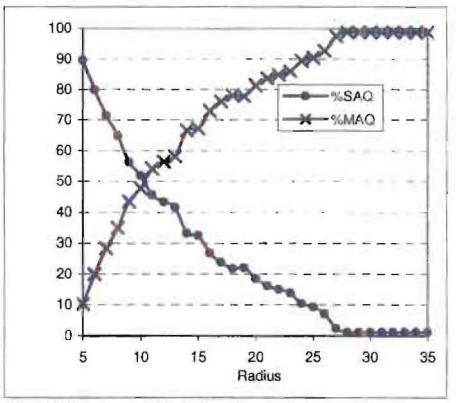
(a) Mean and variance for Figure B.07 (60 events from line cluster and 40 events from CSR distribution)



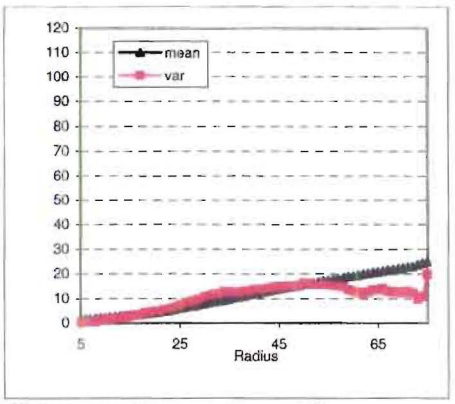
(b) %SAQ and %MAQ for Figure B.07 (60 events from line cluster and 40 events from CSR distribution)



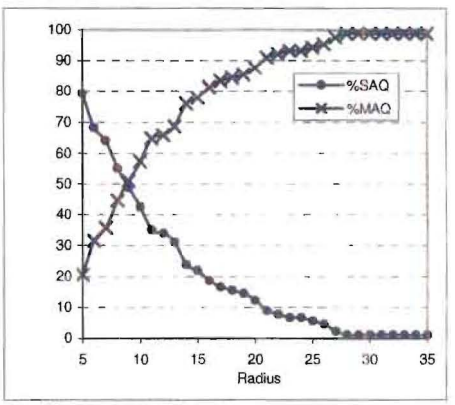
(c) Mean and variance for Figure B.08 (70 events from line cluster and 30 events from CSR distribution)



(d) %SAQ and %MAQ for Figure B.08 (70 events from line cluster and 30 events from CSR distribution)



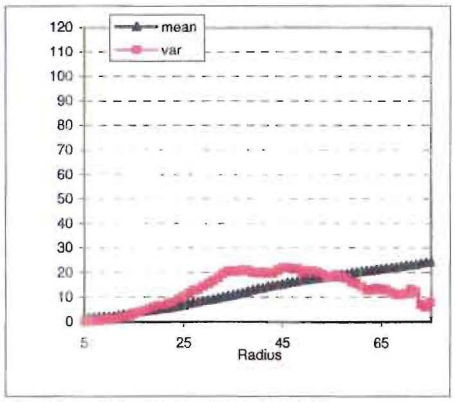
(e) Mean and variance for Figure B.09 (80 events from line cluster and 20 events from CSR distribution)



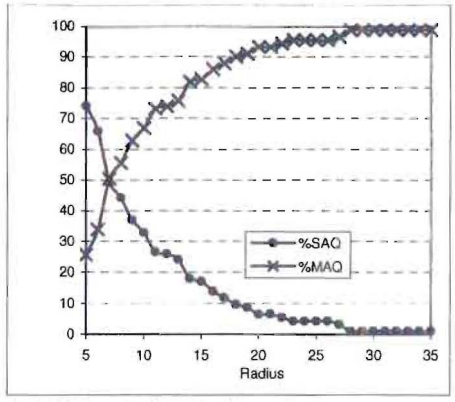
(f) %SAQ and %MAQ for Figure B.09 (80 events from line cluster and 20 events from CSR distribution)

%SAQ - Percentage of single accident quadrats %MAQ - Percentage of multiple accident quadrats

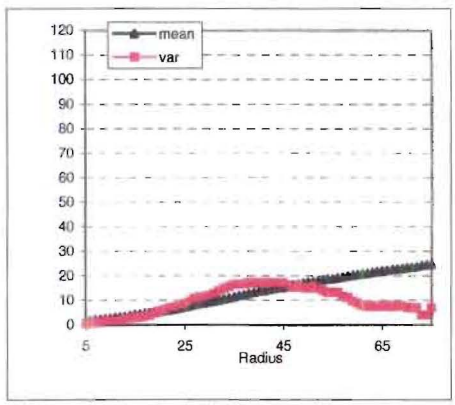
Figure B.14: Variation of mean, variance, %SAQ and %MAQ with increasing quadrat radius for the distributions (Figures B.07, B.08 and B.09).



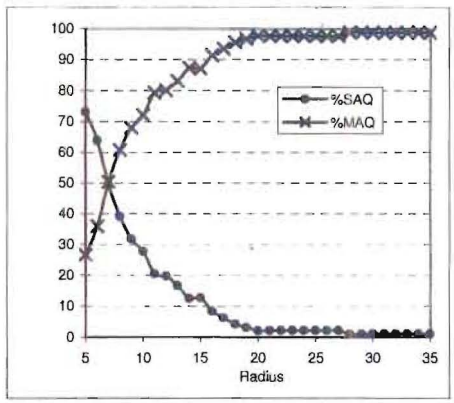
(a) Mean and variance for Figure B.10
(90 events from line cluster and
10 events from CSR distribution)



(b) %SAQ and %MAQ for Figure B.10
(90 events from line cluster and
10 events from CSR distribution)



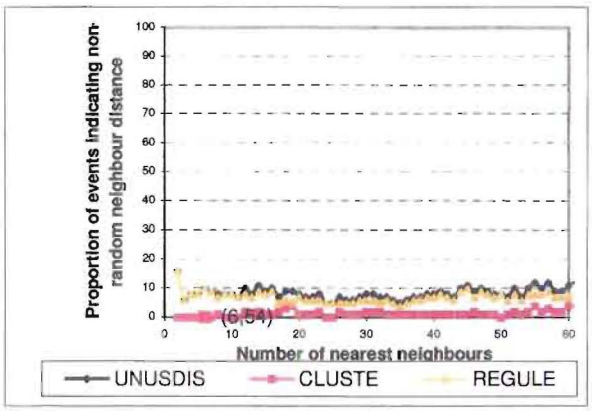
(b) Mean and variance for Figure B.11
(100 events from line cluster and
no events from CSR distribution)



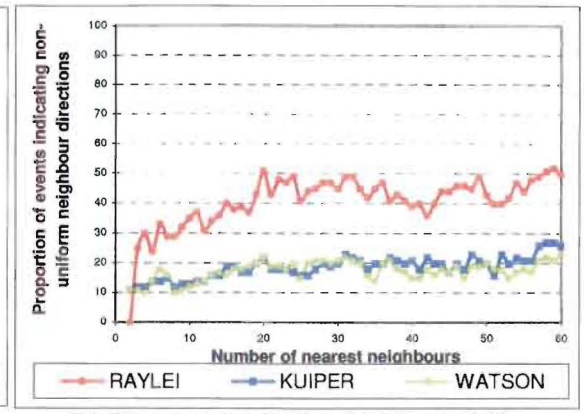
(c) %SAQ and %MAQ for Figure B.11
(100 events from line cluster and
no events from CSR distribution)

%SAQ - Percentage of single accident quadrats %MAQ - Percentage of multiple accident quadrats

Figure B.15: Variation of mean, variance, %SAQ and %MAQ with increasing quadrat radius for the distributions (Figures B.10 and B.11).

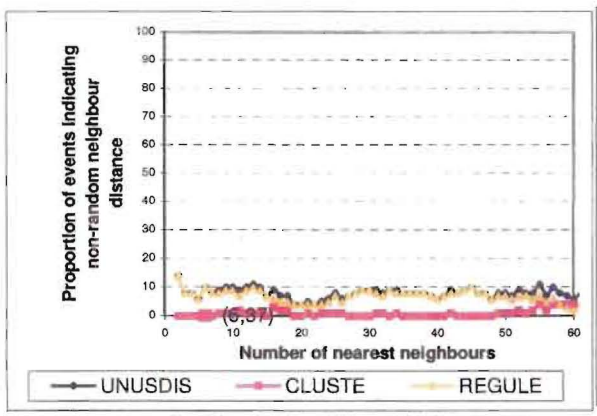


(a) Distance distribution for Figure B.01

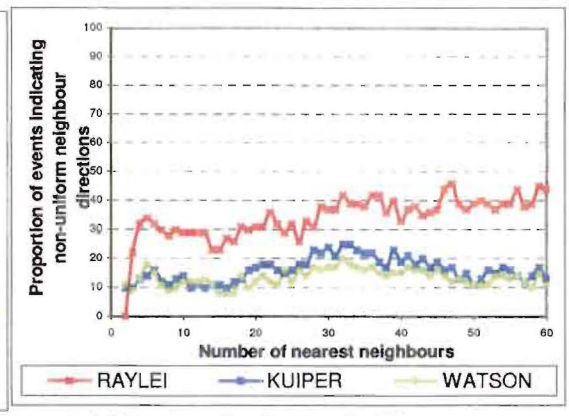


(b) Direction distribution for Figure B.01

(100 events from CSR and no events from line cluster distributions)

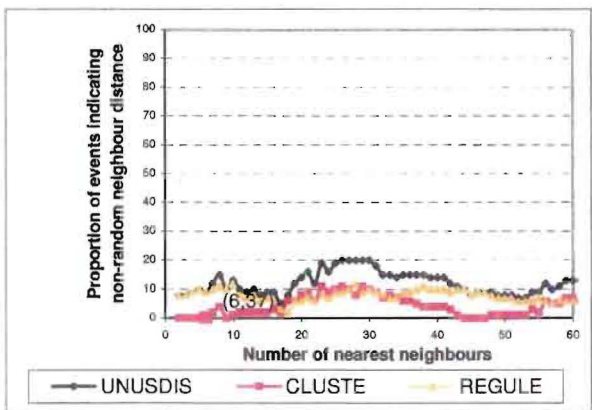


(c) Distance distribution for Figure B.02

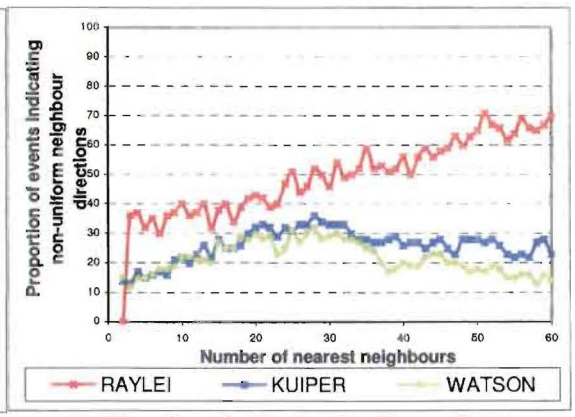


(d) Direction distribution for Figure B.02

(90 events from CSR and 10 events from line cluster distributions)



(e) Distance distribution for Figure B.03



(f) Direction distribution for Figure B.03

(80 events from CSR and 20 events from line cluster distributions)

UNUSDIS--- Unusual distance

CLUSTE--- Cluster

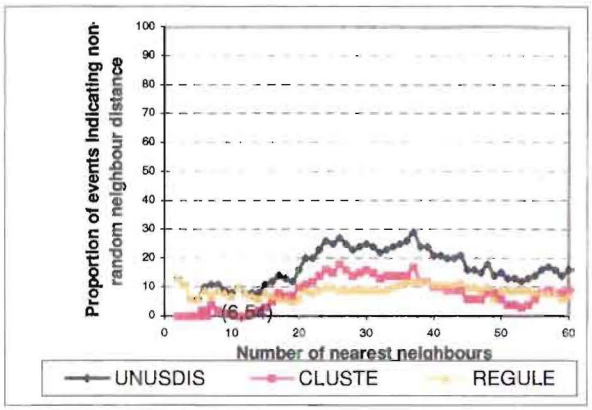
REGULE--- Regular

RAYLEI--- Rayleigh test results

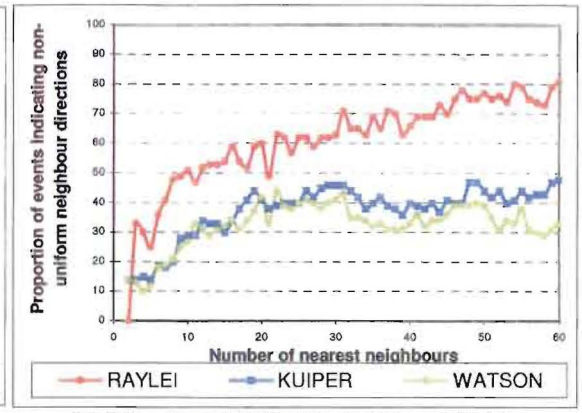
KUIPER--- Kuiper test results

WATSON--- Watson test results

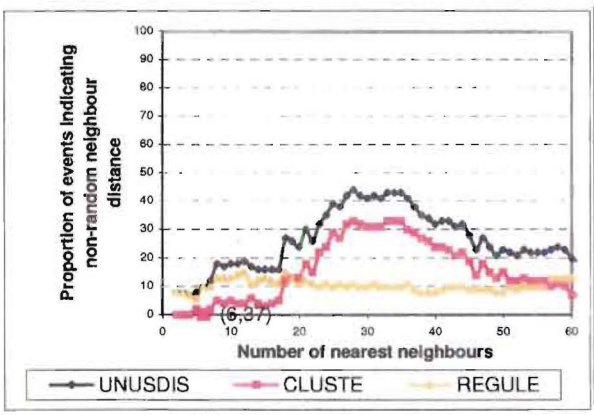
Figure B.16: Nearest-neighbour analysis results for mixed distributions (Figures B.01, B.02 and B.03)



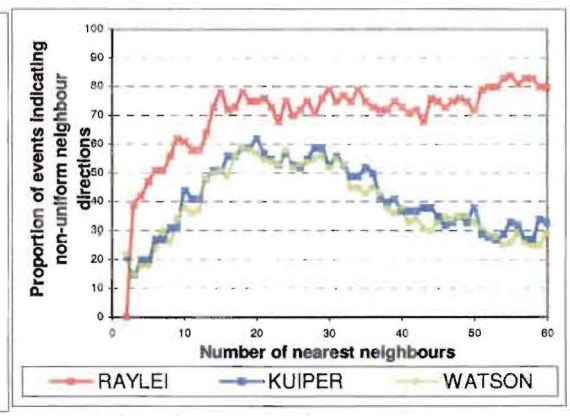
(a) Distance distribution for Figure B.04



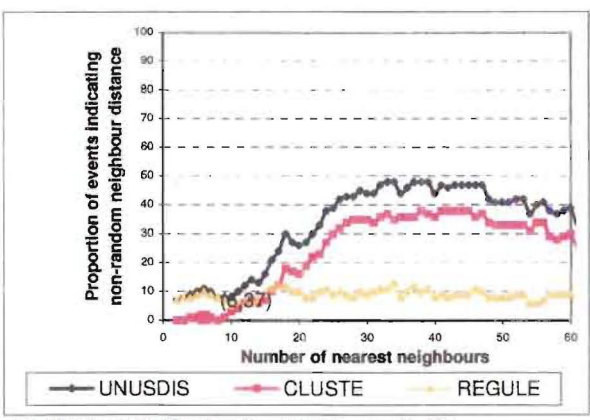
(b) Direction distribution for Figure B.04
(70 events from CSR and 30 events from line cluster distributions)



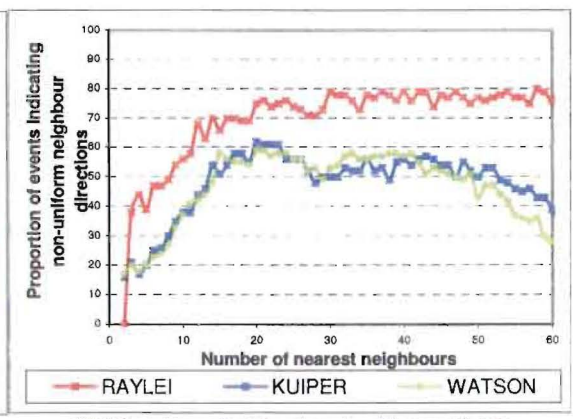
(c) Distance distribution for Figure B.05



(d) Direction distribution for Figure B.05
(60 events from CSR and 40 events from line cluster distributions)



(e) Distance distribution for Figure B.06

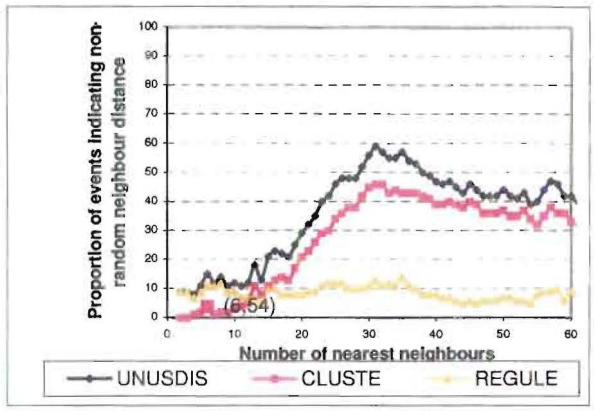


(f) Direction distribution for Figure B.06
(50 events from CSR and 50 events from line cluster distributions)

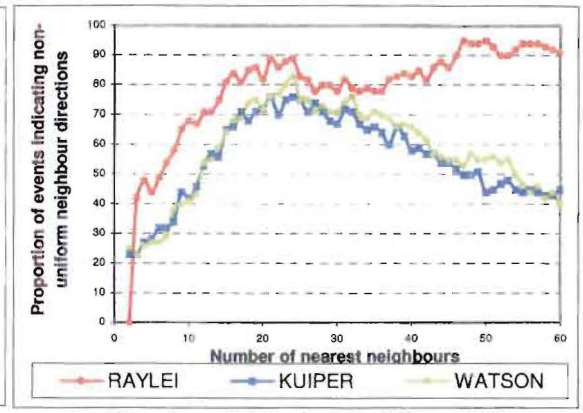
UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure B.17: Nearest-neighbour analysis results for mixed distributions (Figures B.04, B.05 and B.06)

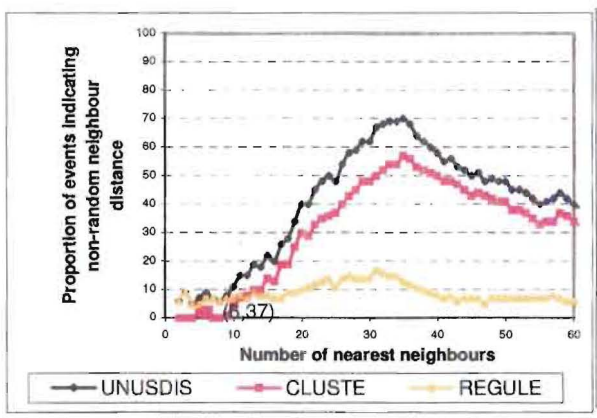


(a) Distance distribution for Figure B.07

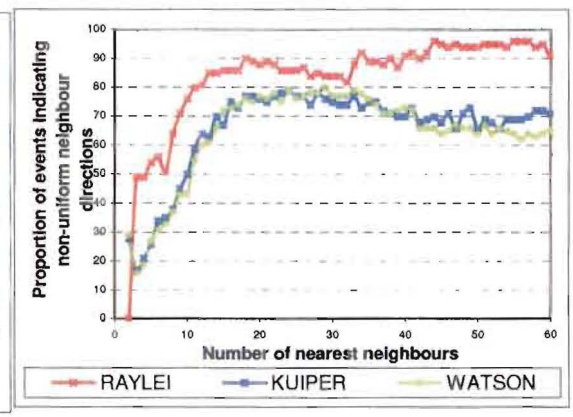


(b) Direction distribution for Figure B.07

(40 events from CSR and 60 events from line cluster distributions)

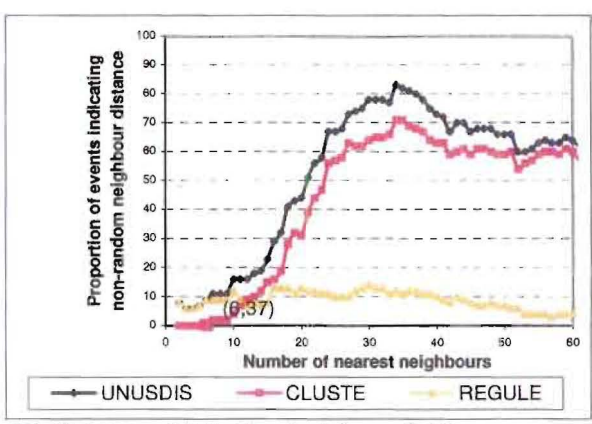


(c) Distance distribution for Figure B.08

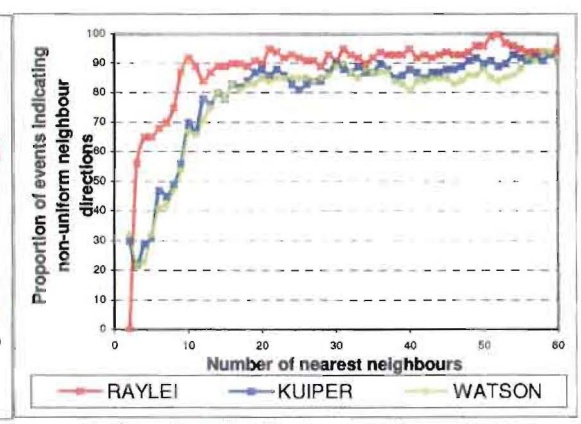


(d) Direction distribution for Figure B.08

(30 events from CSR and 70 events from line cluster distributions)



(e) Distance distribution for Figure B.09



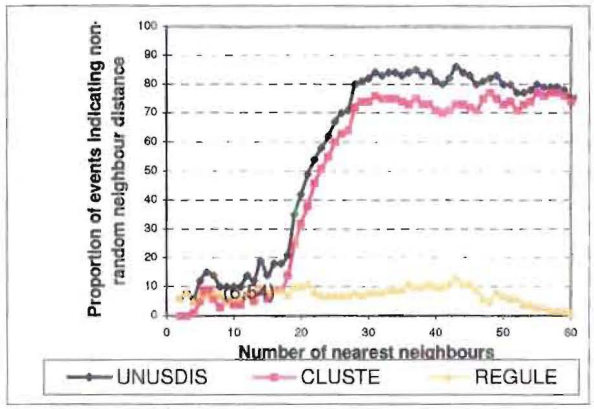
(f) Direction distribution for Figure B.09

(20 events from CSR and 80 events from line cluster distributions)

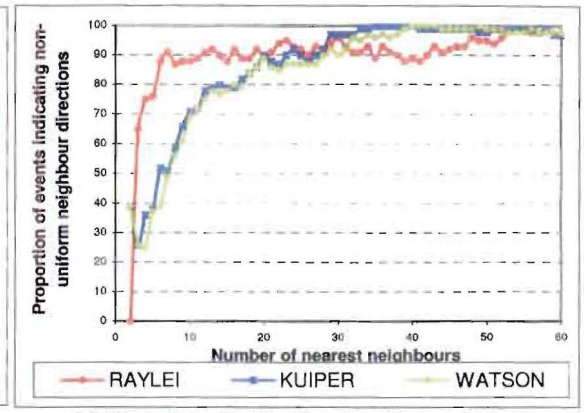
UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure B.18: Nearest-neighbour analysis results for mixed distributions (Figures B.07, B.08 and B.09)

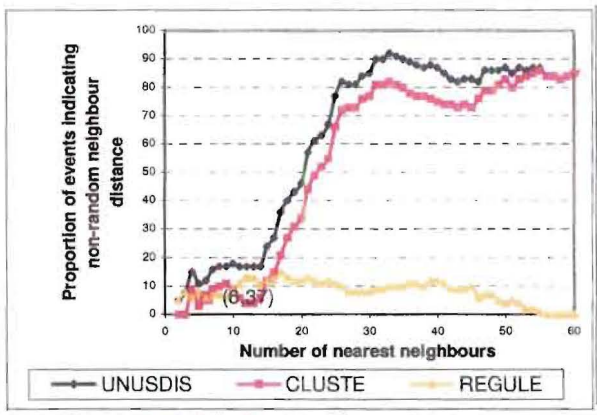


(a) Distance distribution for Figure B.10

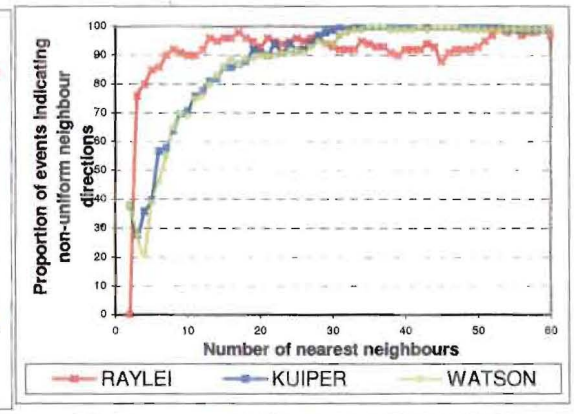


(b) Direction distribution for Figure B.10

(10 events from CSR and 90 events from line cluster distributions)



(c) Distance distribution for Figure B.11



(d) Direction distribution for Figure B.11

(no events from CSR and 100 events from line cluster distributions)

UNUSDIS--- Unusual distance
CLUSTE--- Cluster
REGULE--- Regular

RAYLEI--- Rayleigh test results
KUIPER--- Kuiper test results
WATSON--- Watson test results

Figure B.19: Nearest-neighbour analysis results for mixed distributions (Figures B.10 and B.11)