METHODS FOR THE INVESTIGATION OF SPATIAL CLUSTERING, WITH EPIDEMIOLOGICAL APPLICATIONS

Niall Hay Anderson B.Sc. F.S.S.

A Dissertation Submitted to the
University of Glasgow
for the Degree of
Doctor of Philosophy

Department of Statistics
October, 1992

ABSTRACT

When analysing spatial data, it is often of interest to investigate whether or not the events under consideration show any tendency to form small aggregations, or clusters, that are unlikely to be the result of random variation. For example, the events might be the coordinates of the address at diagnosis of cases of a malignant disease, such as acute lymphoblastic leukaemia or non-Hodgkin's lymphoma. This thesis considers the usefulness of methods employing nonparametric kernel density estimation for the detection of clustering, as defined above, so that specific, and sometimes limiting, alternative hypotheses are not required, and the continuous spatial context of the problem is maintained. Two approaches, in particular, are considered; first, a generalisation of the Scan Statistic to two dimensions, with a correction for spatial heterogeneity under the null hypothesis, and secondly, a statistic measuring the squared difference between kernel estimates of the probability density functions of the principal events and a sample of controls.

Chapter 1 establishes the background for this work, and identifies four different families of techniques that have been proposed, previously, for the study of clustering. Problems inherent in typical applications are discussed, and then used to motivate the approach taken subsequently. Chapter 2 describes the Scan Statistic for a one-dimensional problem, assuming that the distribution of events under the null hypothesis is uniform. A number of approximations to the statistic's distribution and methods of calculating critical values are compared, to enable significance testing to be carried out with minimum effort. A statistic based on the supremum of a kernel density estimate is also suggested, but an empirical study demonstrates that this has lower power than the Scan Statistic.

Chapter 3 generalises the Scan Statistic to two dimensions and demonstrates empirically that existing bounds for the upper tail probability are not sufficiently sharp for significance testing purposes. As an aside, the chapter also describes a problem that can occur when a single pseudo-random number generator is used to produce parallel streams of uniform deviates. Chapter 4 investigates a method, suggested by Weinstock (1981), of correcting for a known, non-uniform null distribution when using the Scan Statistic in one dimension, and proposes that a kernel estimator replace the exact density, the estimate being calculated from a second set of (control) observations. The approach is generalised to two dimensions, and approximations are developed to simplify the computation required. However, simulation results indicate that the

accuracy of these approximations is often poor, so an alternative implementation is suggested.

For the case where two samples of observations are available, the events of interest and a group of control locations, Chapter 5 suggests the use of the integrated squared difference between the corresponding kernel density estimates as a measure of the departure of the events from null expectation. By exploiting its similarity to the integrated square error of a k.d.e., the statistic is shown to be asymptotically normal; the proof generalises a central limit theorem of Hall (1984) to the two-sample case. However, simulation results suggest that significance testing should use the bootstrap, since the exact distribution of the statistic appears to be noticeably skewed. A modified statistic, with the smoothing parameters of the two k.d.e.'s constrained to be equal and non-random, is also discussed, and shown, both asymptotically and empirically, to have greater power than the original.

In Chapter 6, the two techniques are applied to the geographical distribution of cases of laryngeal cancer in South Lancashire for the period 1974 to 1983. The results are similar, for the most part, to a previous analysis of the data, described by Diggle (1990) and Diggle *et al* (1990). The differences in the two analyses appear to be attributable to the bias or variability of the k.d.e.'s required to calculate the integrated squared difference statistic, and the inaccuracy of the approximations used by the corrected Scan Statistic. Chapter 7 summarises the results obtained in the preceding sections, and considers the implications for further research of the observations made in Chapter 6 regarding the weaknesses of the two statistics. It also suggests extensions to the basic methodology presented here that would increase the range of problems to which the two methods could be applied.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to record my gratitude to my supervisor, Professor D. M. Titterington, for all his advice, encouragement and patience during the three years leading to the submission of this thesis. Funding for this period of research was provided by the U.K. Science and Engineering Research Council, in the form of a Research Studentship.

I would also like to thank the following people for their assistance, and acknowledge the valuable contribution they have made to my research:

Dr. F.E. Alexander, of the L.R.F. Centre for Clinical Epidemiology at the University of Southampton, for her interest in this project, and for suggesting that I become involved;

Professor P.J. Diggle and Dr. A.C. Gatrell, of the University of Lancaster, for kindly providing the data analysed in Chapter 6;

Professor P.G. Hall, of the Australian National University and C.S.I.R.O., for the opportunity to collaborate on some interesting research problems relevant to parts of this thesis;

and Dr. A.W. Bowman and Dr. J.F. Bithell, of the Universities of Glasgow and Oxford, respectively, for a number of very useful discussions.

My final acknowledgement is reserved for my parents, for all their help and support, not just for the duration of my research, but in whatever I have attempted to do. Thank you for everything.

DECLARATION

Section 3.6 has been submitted for publication to the journal *Statistics and Computing*.

The material discussed in Sections 5.4 and 5.5 has been presented previously in Anderson *et al* (1992a).

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Many different types of scientific investigation may lead to data sets which contain the location of objects, events or individuals in some two-dimensional planar region. For example, a botanist might be interested in the spread of a particular species of moss within a small area of forest or an archaeologist in the distribution of fragments of pottery in the soil stratum corresponding to a certain period in the history of an Iron-Age settlement. In general, considering spatial patterns of this sort may contribute to the understanding of the process generating the coordinates.

One particular type of behaviour which may be important is the aggregation of points into one or more groupings that could be described as clusters. Providing a rigorous definition of this latter term that would be suitable for all applications is extremely difficult. Approximately, however, a 'cluster' could be described as a collection of points distributed more densely than would seem to be typical, when judged by the geographical spread of the full data set over the entire region of interest.

Diggle (1983, p.2) describes an example of a data set that displays this feature very strongly. The locations of 62 redwood seedlings inside a square area of ground appear to fall into six distinct groups. Such a clear pattern suggested that there was some underlying mechanism influencing where the plants grew. Further investigation revealed that the seedlings were clustered around redwood stumps, the positions of which had not been recorded initially. Clearly, this explained the observed pattern.

It is also true, however, that apparent patterns of aggregated points may be produced by nothing more than the operation of chance. Investigations of geographical distribution are only concerned with clusters that may have been caused by some form of generating process other than randomness, so the above *ad hoc* definition must be extended to say that the character of the group of points must be such that it would be very unlikely for the cluster to have been created randomly. Deciding whether or not a cluster is

genuine, *i.e.* non-random, is then a statistical problem and this thesis investigates ways in which the problem may be tackled.

Studies of spatial clustering have been employed frequently in epidemiological investigations of malignant disease, especially varieties of leukaemia in children. In an overview of the epidemiology of childhood leukaemias, Doll (1989) discusses factors that are known to cause different forms of the disease. The list includes genetic susceptibility, exposure to ionising radiation (either *in utero*, from radiographic examination of the mother during pregnancy, or after birth, through radiotherapy) and the development of acute myeloid leukaemia following chemotherapy, which is an inevitable result of the treatment but only forms a tiny minority of the total number of cases of all varieties. Factors suggested, but not established, as causes include parental exposure to certain chemicals, viral infection or environmental factors such as natural radiation or proximity to a possible, man-made source of environmental pollution. The most controversial examples of the last of these are nuclear installations, especially the reprocessing plant at Sellafield in West Cumbria.

Most attention has been paid to investigating whether or not sources of environmental pollution can be responsible for increased incidence of leukaemia, although the viral infection hypothesis has also been studied; see, for example, Kinlen (1988). Some form of spatial clustering around such locations would suggest that the risk of developing the disease was greater in that particular area than in the surrounding region. Further investigations could then be carried out to determine whether the installation was indeed responsible or whether there was some other factor peculiar to the local area which would increase the risk. The benefits of this work would be seen both in terms of prevention, *i.e.* by removing the leukaemogenic agent, and, more generally, by improving knowledge of the disease's aetiology.

These considerations have prompted a great deal of research effort over the last ten years and the results have been reviewed in Gardner (1989) and Wakeford *et al* (1989). Considerable evidence suggests that there have been more cases of childhood leukaemia near the Sellafield plant than would have been expected from national incidence rates. Weaker evidence suggests the same for the area surrounding the United Kingdom's other nuclear waste reprocessing facility at Dounreay on the north coast of Scotland. Research in areas near other establishments has been inconclusive. As far as the second stage of investigation is concerned, determining the cause of childhood leukaemia clusters, little progress has been made.

The particular context of leukaemia clustering stimulated the work reported here and also motivated much of the literature reviewed in the next section. For this reason, later discussion will often make use of the following terminology, which is typical of that used in the study of disease in human populations:

domain: the geographical region of interest, within which the investigation is carried out;

zones: small, administrative sub-regions within the domain, used to calculate estimates of the population at risk; *e.g.* civil wards, parishes, counties, Census Enumeration Districts (EDs) *etc*;

centroid: a central reference point for a zone, equivalent to the centre of gravity of a physical object, which is used to represent its location in space, since boundaries in digital form are usually not available;

cases: the point locations of the units or events of primary interest;

controls: the point locations of a second type of event which provides ancillary information on the domain.

It is intended, however, that the research discussed in this thesis should be relevant to a wider range of applications.

1.2 Review of Existing Methods of Detecting Spatial Clustering

The literature relating to spatial data analysis contains a wide range of techniques for detecting spatial clustering. It is possible to stratify these methods into three families, an approach taken by Besag and Newell (1991) and the review paper of Marshall (1991). The groups are differentiated by the type of question that the members are designed to answer. The first consists of techniques for analysing the pattern of cases in the region surrounding a particular, fixed point, the second evaluates the general pattern of cases in the domain without reference to a specific location and the third assesses whether or not groups of cases are larger than would be expected by chance and hence form likely clusters.

1.2.1 Detecting Clustering Around a Fixed Point

A number of methods have been proposed for studying the pattern of disease incidence relative to a given point in the region of interest. This point is taken to represent the location of, for example, a source of environmental pollution and the purpose of the investigation is to decide whether or not proximity to the source is associated with an increase in risk. A recent, comprehensive review of such techniques can be found in Hills and Alexander (1989).

The most straightforward approach used by Black (1984), amongst others, is to compare observed and expected numbers of cases within some pre-specified distance of the source, by assuming that the number of cases has a Poisson distribution under the null hypothesis of no clustering. For the United Kingdom, population information for all the EDs contained within the domain is available through the decennial Census and, when combined with national or regional incidence rates, this allows the calculation of expected numbers. Hills and Alexander (1989) point out that it may be difficult to choose incidence rates for the most suitable geographical scale, difficult to choose the size of the region to be considered and that the Poisson distribution may not adequately describe the observed variability. The last problem is often termed "extra-Poisson variation" in the literature.

To achieve a more sensitive test procedure, it is necessary to work with numbers of cases in geographical units at a much finer scale, *e.g.* EDs within the region of interest. However, each of the smaller areas may well contain different levels of population, so even in the absence of any increased risk due to pollution, the observed spatial distribution of cases may well display marked variability. Methods discussed subsequently try to correct for this behaviour.

Black *et al* (1991), for example, describe an algorithm which groups zones into larger areas of approximately equal population size or expected numbers of cases. The algorithm's target population count or expectation is set to be that of the zone containing the point source, and a goodness of fit test is used to compare the observed pattern to a Poisson distribution. Unfortunately, different aggregations of zones tend to give different results, and the power of the test is low for small numbers of zones (Hills and Alexander, 1989).

An alternative to accumulating small areas is the Density Equalised Map Projection method of Schulman *et al* (1988). This magnifies or shrinks each zone until its area is proportional to the population contained within, whilst maintaining the same total area.

The resulting map should have a uniform distribution of cases if there is no clustering effect in operation. Tests can be constructed by considering statistics such as the average, minimum or maximum distance between cases and source of pollution, although these are not measured in terms of the original units.

A second type of approach is suggested in Stone (1988), who develops a test of trend in risk with distance from the fixed point of concern. The zones are ranked by distance from centroid to source, and observed and expected numbers of cases, denoted by O_i and E_i for the i th closest zone, are found for each. The null hypothesis that the number of cases in the i th area has a Poisson distribution with mean E_i is compared to an alternative representing a decrease in risk with increasing separation from the source. The same distribution is assumed but now with mean $\lambda_i E_i$, where

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq \ldots,$$

using either a likelihood ratio test or a test based on $\hat{\lambda}_1$.

Stone (1988) also outlines a test typical of a third type of analytic method for this problem. This does not make use of small area information on population or cases but assumes that the spatial locations of the latter can be obtained very accurately, *e.g.* through address at time of diagnosis. The role of the former quantities is played by a sample of control locations: individuals who do not possess the defining characteristic of the case group but are similar in other respects. If δ_i is an indicator variable with value 1 if the i th point closest to the source is a case and value 0 otherwise, Stone (1988) suggests the test statistic

$$T = \max_{1 \leq n \leq N} \sum_{i=1}^{n} \frac{\delta_i}{n},$$

where N is the total number of cases and controls.

Another example of this third group is provided by Diggle (1990), who uses the theory of spatial point processes to model semiparametrically the intensity of disease as a function of distance from the fixed point at coordinate $\underline{x}_o$. The intensity function, $\lambda(\underline{x})$, of the inhomogeneous Poisson process from which the cases are assumed to have been sampled is decomposed as

$$\lambda(\underline{x}) = \rho \lambda_c(\underline{x}) f(\underline{x} - \underline{x}_o; \underline{\theta}),$$

where ρ represents the overall, average intensity, $\lambda_c(\cdot)$ the variation in intensity due to variations in population density, $f(\cdot,\cdot)$ the contribution from proximity to the source, and $\underline{\theta}$ is a vector of parameters. The control sample is used to estimate $\lambda_c(\cdot)$ by a kernel method (see, for example, Silverman (1986)) and $f(\cdot,\cdot)$ is chosen to represent some

sensible model of risk decaying with increasing separation. Maximum likelihood methods are used to estimate $\underline{\theta}$ and test a null hypothesis of no association between intensity and $\underline{x}_o$.

The use of a control sample was a feature of a relatively early method proposed by Lyon *et al* (1981). The test of clustering suggested therein compares the numbers of the two types of point within increasing radii of the source using a χ^2 procedure. The method has two undesirable features. First, a separate test is applied at each distance, but there is no correction for multiple testing. Secondly, the cases and controls used in tests for the shorter distances are included in each of the tests for the longer distances in a cumulative fashion. Hence, the tests are not independent.

1.2.2 Detecting Clustering Behaviour

The second family of techniques dispenses with the fixed point that was the focus for an analysis using one of the methods described above. Instead, the spatial distribution of cases is investigated for any evidence of a pattern that does not appear to be induced by the underlying spatial distribution of the population. Most of the techniques search for regions where cases occur much closer together than would seem likely by chance, given the available population information. Within this general definition, there is a further division between methods that take account of the variability in population density through zone totals and those that employ a sample of controls.

Whittemore *et al* (1987) is an example of the first subdivision. The geographical location of each case is only ascertained to the scale of the zones in use; hence, each case is assumed to be positioned at the centroid of the subregion in which it falls. The proposed test statistic is the mean distance between all pairs of cases, with cases at the same centroid having a separation of zero. The population sizes of each zone are used to calculate the mean and variance of the test statistic, and an asymptotic normality result allows significance testing. The method also permits the stratification of observations into different risk categories. Although there may be some reduction in the researcher's workload by simply assigning to each case the coordinate of a centroid, the test seems to be partially investigating the (discrete) geographical distribution of the zones, which is not of interest. If Euclidean distance is to be used at all in an analysis, it would be better to define coordinates as accurately as possible and place the problem in a more natural, continuous spatial context.

Black *et al* (1991) employ an adaptation of their original method, discussed in the previous section, which searches for evidence of extra-Poisson variation. Aggregations

are constructed for a range of pre-specified expected numbers of cases, and the Poisson indices of dispersion for each sub-region, the ratio of the sample variance to sample mean of the observed number of cases, are calculated and compared to a normal distribution, with a correction made for multiple testing. The power of this test is thought to be greater than that of the χ^2 statistic used previously, but the *caveat* regarding the sensitivity to different aggregations still applies.

A qualitatively similar aggregation procedure is used in Turnbull *et al* (1990) to provide a set of overlapping areal units of equal population size. For each of the original zones in turn, a neigbourhood is created by adding surrounding zones in whole or part until a target population size, denoted by R, is reached. The total number of cases in each neighbourhood is found by adding the same proportion of cases from each contributing zone as was required to achieve R. The suggested test statistic is the maximum incidence rate, *i.e.* maximum number of cases over all neighbourhoods divided by R, with significance assessed by a Monte Carlo implementation of a randomisation test. It seems likely that the results of this approach will be sensitive to different aggregations for the same value of R and, of course, to different values of R itself. However, a range of target sizes and a multiple testing procedure might address this latter problem, as in Black *et al* (1991).

Many other measures of spatial clustering based on case and population data in small administrative areas have been proposed in the literature. Alexander (1991), for example, describes three possibilities, namely the NNA test, the Potthoff-Whittinghill test and a modification of a method due to Barnes *et al* (1987). The first of these is adapted from Besag and Newell (1991), which is discussed in Section 1.2.3, and the second is a test for extra-Poisson variation based on the numbers of pairs of cases in each zone.

The second group of techniques, those which use a sample of controls to represent the spatial heterogeneity of the population, is exemplified by Cuzick and Edwards (1990). Asymptotic normality is demonstrated for a test statistic which is the sum of score functions from each case, where the score is the number of cases amongst its k nearest neighbours, for some integer k. As k is arbitrary and its value affects the results of the test procedure considerably, a statistic is also proposed that is a linear combination of the original test statistics for different values of k, in an attempt to avoid the problem.

On a less formal level, Bithell (1990) describes a graphical method of displaying a spatial relative risk function for the entire domain. This is constructed using the quantity

$$\hat{\rho}(\underline{x}) = \frac{\hat{f}_{cases}(\underline{x}) + c}{\hat{f}_{controls}(\underline{x}) + c},$$

where $\hat{f}_{\bullet}(\cdot)$ is an adaptive kernel density estimate calculated from the relevant sample and c is a constant included to ensure that $\hat{\rho}(\underline{x}) \to 1$ in areas that have few data points. The transformation

$$P(\underline{x}) = \frac{\hat{\rho}(\underline{x})}{1 + \hat{\rho}(\underline{x})}$$

is applied to ensure plotted values remain within the interval (0,1). The surface $P(\underline{x})$ may be examined for peaks that might indicate an unusually high local risk. Although bandwidths must be specified for the case and control kernel density estimates and the value of c may also affect the smoothness of the final estimate, the method performs well in its intended, exploratory role.

A final example of this type of technique is Diggle and Chetwynd (1991), which models the spatial distribution of two types of event using inhomogeneous Poisson processes. The test is based on

$$D(s) = K_{11}(s) - K_{22}(s),$$

where

$$K_{ij}(s) = \lambda_j^{-1}\, \mathbf{E}[\#\ \text{type } j \text{ events within distance } s \text{ of arbitrary type } i \text{ event}]$$

and λ_j is the intensity of type j events, for j = 1, 2. Large, positive values of $D(s)$ suggest spatial clustering of type 1 events over and above that related to the behaviour of type 2 events. Values of $D(s)$ at a discrete set of distances can be plotted against pointwise tolerance limits or combined to give a test statistic, for which significance is assessed by simulation.

1.2.3 Investigating Possible Clusters

The third, and smallest, family of methods has an underlying rationale different from those of the two groups discussed above. The intention now is to provide exploratory techniques which, rather than formally assessing the pattern over the whole domain or near a particular point, search for small areas that are likely candidates for the status of

clusters, without assigning any causal explanation to regions so classified. Subsequent investigations would be used to decide whether the clusters were due to chance or to some aetiological factor. The methods have been proposed as being particularly suitable for the routine, possibly automatic, surveillance of a large area, allowing effort to be targeted on the most promising putative clusters.

The Geographical Analysis Machine (GAM) of Openshaw *et al* (1987, 1988) is a graphical method designed to highlight areas where there is a mismatch between the observed number of cases and the associated population size. A large set of possible locations in the domain are tested to see whether or not each could be the centre of a cluster. This is accomplished by overlaying the domain with a grid, each point on which acts as the centre of a circle of a given radius. The grid size is chosen to ensure that the circles always overlap neighbouring ones. The number of cases falling within each circle is counted and the relevant population at risk calculated from each zone that has its centroid inside the circle. A Monte Carlo test based on 500 replications is used to assess significance at each grid point. This process is repeated for different radii and all circles significant at the 0.002 level are plotted on a map of the domain. Possible clusters are indicated by dense groups of circles of various radii.

A number of weaknesses in this procedure have been discussed by different authors in the literature. Openshaw (1990) summarises some of these and Besag and Newell (1991) explore some points in more detail. There is a considerable multiple testing problem, both at the scale of the whole domain and at that of a given 'cluster' of circles. The interpretation of the latter is very difficult, since many different circles may be due to the same cases and are therefore not independent. The GAM is also highly computationally intensive. Openshaw (1990), however, also describes improvements to the basic algorithm, some of which address these difficulties. For example, methods for assessing the significance of circle aggregations are outlined and algorithms are described that make use of statistics and associated tests such as those of Stone (1988) and Besag and Newell (1991) in place of the count of observed number of cases plus Monte Carlo test which was used in Openshaw *et al* (1987).

The method of Besag and Newell (1991) was proposed as both an alternative inferential procedure to that employed by the GAM and as a cluster detector in its own right. In its former guise, it was intended to provide a statistical basis for the type of techniques used in Openshaw *et al* (1987). Cases are placed at the centroid of the zone in which they are located and then examined in turn. A test is carried out to determine whether or not the location of the reference case forms the centre of a cluster of size k+1 (the

reference case plus k others), where k is a pre-specified integer. The procedure is as follows: all the zones are ranked in ascending order of distance from the reference centroid; cases are accumulated from neighbouring zones in this order until k (or more) other cases have been found; the corresponding population at risk is calculated simultaneously from the totals for each included zone, and a Poisson (approximate) tail probability is calculated for the number of zones required to find at least k cases. The distance from the reference centroid to the centroid of the last zone to be included can be used as the radius of a circle, centred on the former point, which is plotted on a map of the domain if the significance of the test exceeds a given level. This is analogous to the GAM. In addition, tests are outlined for checking whether or not there are more clusters than would be expected and whether or not there is evidence of some form of clustering behaviour, the type of test considered in Section 1.2.2.

1.2.4 Tests of Space-Time Clustering

A fourth family of techniques, although not strictly concerned with spatial clustering, is relevant to the discussion here. Tests of space-time clustering, also known as space-time interaction tests, first appeared in the literature much earlier than the methods described above and have been cited frequently. They were developed primarily to detect aggregations of cancers that had been recorded at approximately the same time, possibly indicating that a viral infection was causing the spread of the disease.

Knox (1964a,b) categorises all possible pairs of cases according to whether they are less than or greater than some critical distance apart in space and time. Interest centres on the cell of the resulting 2×2 table containing pairs that are close on both criteria, with a large count indicating some form of clustering. Mantel (1967) generalises this approach and presents a measure of closeness defined to be the reciprocal of the pair's separation in space or time. Both techniques require the specification of arbitrary constants capable of affecting the results of the analysis: critical distances for the former and additive constants to prevent division by zero for the latter, should discreteness in time or space coordinates permit identical case locations.

A second type of approach makes use of locational data in the form of counts of cases in cells formed by distinct geographical areas and time periods. Ederer *et al* (1964) use the sum over all spatial regions of the maximum number of cases for the range of temporal intervals as an index of clustering, while Raubertas (1988) uses a generalised linear model for cell probabilities with spatial and temporal main effects and a space-time interaction term. By examining contrasts, within sub-regions of the whole area, of parameter estimates for the spatial main effect, it is possible to derive a test for spatial

clustering. However, both the size of the sub-regions and the weights for the contrasts are arbitrary and it would appear to be possible for the tests to reflect nothing more than different numbers of cases arising from variable population sizes in the individual cells.

McAuliffe and Afifi (1984) construct a test statistic from the distance between a reference case and its nearest neighbour from a previous time period, for a specified lag, by minimising, over all lags, the sum of standardised distances. Nearest neighbour distances are also used by Ross and Davis (1990), to investigate the clustering of Hodgkin's disease, given full residential information for each case, during particular time periods hypothesised to be aetiologically relevant. Using a permutation test, the pattern of cases is compared to that of a sample of controls, so that the behaviour of observed nearest neighbour distances is compared with what would be expected in the absence of any contagion effect.

If interest is primarily in an investigation of a source of environmental pollution, the choice of a space-time interaction test would seem to be unsuitable. Any clustering effect under this model would be unlikely to manifest itself more visibly in one time period than in any other and, thus, the power of the above methods would be much reduced. For the small number of cases typical of a study of malignant disease, Wartenberg and Greenberg (1990) demonstrate that the tests due to Ederer *et al* (1964) and Mantel (1967) have low power against two simple alternative hypotheses of locally elevated risk due to a point source.

1.3 Problems Inherent in Spatial Clustering Studies

In the general area of epidemiology, and in leukaemia studies in particular, a number of different authors have considered solely, or in part, the difficulties inherent in studies of spatial clustering. Examples of this are provided by the papers of Besag and Newell (1991), Bithell and Stone (1989), Gardner (1989), Hills and Alexander (1989), Wakeford (1990) and Wakeford *et al* (1989). The following discussion is set in the same context and some points are specific to that; however, many will apply whatever the circumstances of the study.

When working with data on populations of human beings, positive or negative results may be simply artefacts of inaccuracies in numerator (case) or denominator (population at risk) information. The latter situation is a primary concern of the contributions of Besag *et al* and Openshaw and Craft to Draper (1991). In the United Kingdom, information on the human population is usually based on figures from the Census of

Population, which is taken once every ten years. This interval is long enough to encompass considerable demographic change, especially change in the numbers of young children, due to fluctuations in birth rates or large scale migration into or out of the domain, for example. If a population were to increase rapidly after a Census, there would be a corresponding increase in the number of cases expected within the domain. Subsequently, an analysis based on the original Census estimate of population at risk would be susceptible to the detection of spurious clusters, because there would be too many cases for the older population figures to explain. A post-Census decrease might lead to the dilution, or even concealment, of evidence of real clustering.

Numerator data may also be at fault. Most methods for detecting clustering assume, implicitly, that case ascertainment is complete and free from duplications, or at least that it is at a uniform level throughout the domain. In practical terms, this may be unlikely. Additionally, classifications or diagnoses may have been made incorrectly or based on criteria that have changed during the course of the study. Leukaemia is a good example of this, as the distinctions between some types of lymphoid leukaemia and non-Hodgkin's lymphoma have been altered over the last twenty years; see, for example, Draper (1991, Chapter 2). Migration of cases into or out of the domain could affect results in a similar way to that of members of the population at risk. Cases generated by a genuine clustering effect may be lost to the investigation when they move out of the domain, leading to a loss of power. There would seem to be a particular risk of this sort of error in studies of cancer, because of the long latent periods associated with malignant diseases. Conversely, cases that originated outside the domain may be included when, in fact, they are little more than a confounding influence. To avoid these problems, a number of authors have proposed that cases should be located at their address at birth rather than that of diagnosis, although in general this information is much harder to obtain.

A number of problems may be introduced through the conduct of investigations and the techniques of statistical inference used in their analysis. The spatial scale at which the study is carried out may be extremely important, in terms of the size of both the domain and any administrative zones in use. If these are too large, a small scale or local clustering effect may be swamped. Equally, if clustering is present at a scale which is much greater than the chosen areas, then it is unlikely that it could be detected, a situation analogous to one in the analysis of time series, where a short term trend is revealed to be part of a long term cycle by looking at a longer series. It is also possible that evidence of local clustering derives from nothing more than the spatial heterogeneity of incidence over a much larger region. The method of Stone (1988) is

an attempt to reduce the importance of the initial choice of one variable, the size of domain, but in general it would appear to be necessary to accept these limitations, unless there is some prior information on the scale of any possible clustering effect that can be incorporated into the study.

The power of an investigation into spatial clustering will be reduced if the number of cases expected in the domain is small. This will certainly be true for types of rare event such as cases of malignant disease. The effect on power of out-migration has already been mentioned, and it is likely that a complex aetiology, *i.e.* the interaction of different causative agents, would similarly make clustering much harder to detect. This last factor explains why it is often very difficult to propose a plausible model of clustering in a spatial problem and, thus, provide alternative hypotheses that could be tested by the most appropriate and powerful methods.

When a technique employs administrative zones or subregions, the continuous spatial distribution of population in the domain is transformed into a set of counts at discrete zone centroid locations. Although U.K. Census EDs, for example, are drawn up to include approximately the same number of households where possible, very often other considerations intervene. The Office of Population Censuses and Surveys and the General Register Office for Scotland, the Civil Service organisations responsible for administering each Census, have a duty to maintain confidentiality, *i.e.* to ensure that no individual or particular region can be identified from the Small Area Statistics that they produce. This may mean that ED boundaries are altered to split a small village between different zones or to include a much larger area than usual, so that the population within the zone reaches an acceptable level. Therefore, the zones may not reflect the true distribution of population very accurately. As with the method of Black *et al* (1991), the results of techniques which are based on zone population counts are likely to be sensitive to different partitionings. This is an undesirable feature and it may be better to work within a continuous frame of reference, if this can be accommodated in the study.

Hypothesis testing is designed to be used within the scientific paradigm, in which preliminary experimentation or prior knowledge leads to the specification of theories and then hypotheses, which are confirmed or rejected by further, independent experiments. The independence of the confirmatory tests is crucial to this process: interpretation of significance becomes complex if the hypothesis under consideration is influenced by knowledge of the data to be used at the second stage. One of the dangers in studies of spatial clustering is that this type of knowledge might cause the adjustment of boundaries to be used in the analysis; for instance, the size and position of the

domain, the number and type of zones, the time period for which data are collected, the sub-types of event to be considered or the specification of risk sub-groups based on confounding variables such as age or sex. It might be possible to choose such boundaries, either consciously or unconsciously, so that the significance of any apparent clustering was inflated or so that a cluster was created from a random pattern. This would also introduce an implicit multiple testing problem, since the set of boundaries would have been selected from a range of combinations that could have been pre-defined. Correcting for this effect would be very difficult since the comparisons would be statistically dependent and (usually) unspecified in number. As Wakeford (1990) comments, it may be permissible to define boundaries *post hoc* in an hypothesis-generating or exploratory study, provided significance levels are used for guidance only. However, fresh data would be required for valid hypothesis testing, by examining either a new time period for the same domain or the same time period for other, comparable domains, although this raises the question of whether or not hidden or confounding factors would make it possible to define such regions.

The study of the geographical distribution of leukaemia near sources of environmental pollution is prone to *post hoc* hypothesis formulation, because investigations are often initiated following public concern over an apparent excess number of cases that seem to be related by proximity to the site. Hills and Alexander (1989) suggest that overall tests of clustering, discussed in Section 1.2.2, may be useful for reducing the magnitude of the problem, since the selection bias induced by focussing on the spatial pattern around the source is removed. Instead, the results are placed in the context of the whole domain.

1.4 Aims and Outline

The aim of this thesis is to investigate methods of detecting the spatial clustering of point locations that is not attributable to chance. The type of technique that will be of interest is one that may be used in the early stages of a study, at a relatively small scale or local level, to assess the overall pattern in the domain. For this reason, the existence of a coordinate representing, say, a source of pollution is not assumed, although this would be a common motivation for such an investigation. This may also serve to moderate the effects of selection, as described at the end of Section 1.3. The low numbers of cases typical of applications requiring this methodology suggest that smoothing techniques may be appropriate, so that estimates at any given point may incorporate neighbouring information. Arbitrary administrative zones will not be used because of their possible influence on results and the deficiencies in Census figures.

Instead, estimates of the population at risk will be derived from a sample of controls, making the (perhaps not negligible) assumption that a sampling frame exists for the domain under consideration that will allow the drawing of representative observations without excessive difficulty. Due to the absence of convincing models of clustering, and hence alternative hypotheses, in applications such as leukaemia studies, methods which are nonparametric in character will be favoured.

Chapter 2 considers the Scan Statistic, which provides a test for clustering in one dimension, *e.g.* time. The assessment of significance by approximating critical values is considered, as is the relative power of different methods. Chapter 3 explores the generalisation of the Scan Statistic to two dimensions and Chapter 4 introduces a correction for a non-uniform distribution of events under the null hypothesis, which allows the Scan Statistic to be used, for example, with human populations. The behaviour (both asymptotic and simulated) of an integrated squared difference statistic based on kernel density estimates is described in Chapter 5 and a power study indicates the (asymptotic) advantage of using a fixed bandwidth. Chapter 6 applies the two methods to a data set consisting of all cases of laryngeal cancer diagnosed in South Lancashire between 1974 and 1983. Chapter 7 closes the thesis with some discussion of results obtained.

CHAPTER 2

DETECTING CLUSTERING IN ONE DIMENSION

2.1 Introduction

A study of techniques for detecting the clustering of events in a single dimension can serve two purposes. First, the methods may be of interest in themselves, for applications that are concerned only with clustering in, say, time, and secondly, generalisation to two dimensions may be aided by the exploratory work carried out in one.

This chapter considers two conceptually straightforward methods, the Scan Statistic and a test based on the supremum of a kernel density estimate. The former test is discussed in Section 2.2, in which different types of scanning procedure are evaluated and a number of approximations to critical values are outlined. Section 2.3 explores the kernel estimate based statistic and Section 2.4 reports the results of a power simulation study, which compares the two tests.

Chapters 2 and 3 assume that events in the region of interest have a uniform distribution, in the absence of any clustering. As has already been indicated, this assumption may be unrealistic when dealing with human populations. However, to reduce the complexity of the problem, it is convenient to work with a null hypothesis of this form in the first instance. A way of correcting for a non-uniform null distribution is discussed in Chapter 4.

2.2 The Scan Statistic

The Scan Statistic is a particularly simple method of examining the distribution of a sample of data points for evidence suggesting that it may be the result of something other than random chance. It was originally proposed as a measure of clustering in time by Naus (1965a, 1966a) and a recent review was provided by Naus (1988). Taking the region of interest to be the interval (0,1], without loss of generality, and assuming that N events are distributed within it, the Scan Statistic, S_d, is defined to be the maximum number of points that can be included in a sub-interval $[x - d, x]$, where x is allowed to

vary over $(d,1]$ and d is a prespecified constant, such that $d < 1$. This process represents a window 'scanning' over the unit interval in a continuous fashion, with S_d being the largest, instantaneous count of points so 'framed'. Clearly, large values of this statistic would indicate some form of clustering.

Two observations simplify the development of an algorithm for calculating the Scan Statistic that can be implemented in a programming language such as FORTRAN. First, since the statistic is defined to be the maximum count of events within a window, regions of the interval (0,1] that contain no events need not be considered. Hence, the coverage of the unit interval achieved by the algorithm will not necessarily be complete; the scanning process will concentrate on event locations, rather than all possible locations. Secondly, if the direction of scan is assumed to be from left to right, the total number of events contained within the window may only increase if the right-hand boundary reaches the location of a new event, and will decrease if the left-hand boundary moves to the right of an event that was previously inside the window. We are interested in large counts and, therefore, principally in events that enter the window from the right. Hence, if the ith event, e_i, forms the upper boundary of a scanning interval, $[e_i\text{-}d,e_i]$, of length d and c_{id} = count of events in $[e_i\text{-}d,e_i]$, then the Scan Statistic is equal to

$$\max_{i\in 1,\ldots,N} c_{id}.$$

The programming task is then reduced to counting the numbers of events in N distinct intervals of the form $[e\text{-}d,e]$.

Distributional results for the Scan Statistic, under a null hypothesis of no clustering, are derived assuming that events are uniformly distributed over (0,1]. For some integer n, denote $\Pr(S_d \geq n \mid N\ ;\ d)$ by $P(n|N;d)$. Naus (1965a) provides an exact result for $P(n|N;d)$, using a combinatorial argument, as follows:

$$P(n|N;d) = \begin{cases} C(n|N;d) - R(n|N;d), & \text{for } d \geq \frac{1}{2}, n > (N+1)/2, \\ C(n|N;d), & \text{for } d \leq \frac{1}{2}, n > N/2, \end{cases} \tag{2.1}$$

where

$$C(n|N;d) = (N-n+1)\left\{\sum_{i=n-1}^{N} b(i|N;d) + \sum_{i=n+1}^{N} b(i|N;d)\right\} - 2(N-n)\sum_{i=n}^{N} b(i|N;d),$$

$$R(n|N;d) = \sum_{i=n}^{N} b(i|N;d)\left\{\sum_{j=0}^{N-n} b(j|i;(1-d)/d)\right\} + H(n|N;d)b(n|N;d),$$

$$H(n|N;d) = \{n(1-d)/d\}\sum_{i=0}^{N-n} b(i|(n-1);(1-d)/d) - (N-n+1)\sum_{i=o}^{N-n+1} b(i|n;(1-d)/d)$$

and

$$b(k|M;p) = \binom{M}{k} p^k (1-p)^{M-k}.$$

For $d \geq 1/2$ and $n \leq (N+1)/2$, $P(n|N;d)$ is identically equal to 1; however, the case for which $d \leq 1/2$ and $n \leq N/2$ is not covered by the argument. The latter range of d and n could be of particular interest, in practice, because detecting a small scale clustering effect presumably requires a small scanning interval, if the effect is not to be concealed by the larger scale pattern of events. The associated value of the Scan Statistic, therefore, would be quite small, possibly much less than $N/2$.

An expression for $P(n|N;d)$ with better coverage of the range of d and n is described in Naus (1966a). For $d = 1/k$, where k is an integer, $k \geq 2$ and $2 \leq n \leq N$,

$$P(n|N;1/k) = 1 - N!k^{-N}\sum \det\left(1/c_{ij}!\right),$$

where

$$c_{ij} = \begin{cases} (j-i)n - \sum_{r=i}^{j-1} n_r + n_i, & i < j, \\ (j-i)n + \sum_{r=j}^{i} n_r, & i \geq j, \end{cases}$$

n_i denotes the number of points in the interval $((i\text{-}1)/k, i/k)$, $1/c_{ij}!$ is defined as zero when $c_{ij} > N$ or $c_{ij} < 0$, and the summation is over all partitions of N into k positive integers, each less than n. Huntington and Naus (1975) and Hwang (1977) derive expressions for $P(n|N;d)$ that are somewhat simpler and not limited to interval widths of the form $1/k$, k an integer. Both are qualitatively similar to Naus (1966a), in that they involve the summation of the determinants of possibly large matrices, over a set of partitions of N. Although the relevant matrices are smaller and the sets of partitions have fewer elements, the two alternatives are still difficult to compute. This objection also applies to Wallenstein and Naus (1974), who express $P(n|N;1/k)$ in terms of multiple intersections of events, and to Naus (1982), who gives an approximation for the same probability. The accuracy of the latter result is best when the analysis does not condition on the total number of events, N, *e.g.* when the data in (0,1] form part of a

longer stream of information. In retrospective studies, however, which are common in areas of application such as epidemiology, it is usual to take N as fixed, so the approximation will be less useful.

Three possible approaches to overcoming the complexity of the Scan Statistic's distributional form are considered here. The first is to search for a different scanning procedure with comparable results but simpler behaviour. The second is to find direct approximations to the statistic's critical values, since these are the only quantities needed for significance testing. The third approach is to calculate critical values from approximations to the distribution of the Scan Statistic that are easier to use than the ones mentioned above. These options are discussed in turn in the following sections.

2.2.1 An Alternative Scanning Procedure

Instead of a window sweeping continuously over the unit interval, consider $k \geq 2$ disjoint intervals, each of length $d = 1/k$, covering (0,1] completely. If the number of events in the j th interval is denoted by N_j, then the joint distribution of N_1, ..., N_k is multinomial, with each cell probability equal to $1/k$. If the analogue of the Scan Statistic for this situation is

$$\max_{1 \leq j \leq k} N_j,$$

then its distribution is that of the largest cell frequency from a multinomial distribution.

The covariance between N_i and N_j, for some i and j, is $-N/k^2$, which may be small enough to be neglected if k is large. This would allow the $\{N_j\}$ to be regarded as a sequence of independent and identically distributed binomial$(N,1/k)$ random variables, since each N_j is marginally binomial, which could be approximated by normal$(N/k,(N/k)(1\text{-}1/k))$ random variables, or normal$(N/k,N/k)$, since k is assumed to be large. If M is the maximum of the standardised versions of these normal random variables, then

$$\sqrt{2\log k}\,(M - l_k)$$

has a Gumbel distribution (David, 1981), where

$$l_k = \sqrt{2\log k} - \tfrac{1}{2}(\log\log k + \log 4\pi) / \sqrt{2\log k}\,.$$

To assess whether or not a disjoint scan approach is an adequate approximation to the continuous Scan Statistic, the empirical null distribution functions of both methods were simulated for a range of N and k. For comparison, a maximum count statistic was also simulated for discrete, overlapping intervals. The percentage of an interval's length covered by a neighbour was chosen to be 10%, 50% or 90%. It is interesting to note that this procedure bears some resemblance to the time component of a space-time interaction test proposed by Ederer *et al* (1964) and also to part of the algorithm of the Geographical Analysis Machine of Openshaw *et al* (1987, 1988).

Table 1 compares the 5% critical values obtained from 1000 simulations of each scanning procedure for different combinations of N and k, generating the N events by sampling from a uniform(0,1] distribution. As might be expected, there is an obvious trend upwards over the different methods for a given N and k, from the disjoint procedure, through the overlapping intervals in order of overlap proportion, to the continuous procedure. The results for the disjoint procedure are consistently too low, although the error is moderated slightly for the largest values of k; these are of more interest, since the distributional result associated with this procedure requires small between-interval correlations. It would appear, however, that the disjoint intervals method is not sufficiently accurate to replace the continuous Scan Statistic and, therefore, that the multinomial extreme value argument suggested above cannot be employed. A further weakness of the method is described in Naus (1966b), which proves that the power of the disjoint procedure is less than that of the continuous Scan Statistic for all alternative hypotheses that specify continuous p.d.f's, when d is sufficiently small.

				Overlapping Intervals		
N	*k*	Cont.	Disjoint	10%	50%	90%
50	4	23	20	20	21	22
	8	15	13	14	14	15
	16	11	9	10	10	10
	32	8	7	7	8	8
	64	6	6	6	6	6
100	4	39	36	36	37	39
	8	25	22	23	23	24
	16	17	15	15	16	16
	32	12	10	10	11	11
	64	9	8	8	8	9
200	4	70	64	64	66	69
	8	42	38	38	40	41
	16	26	24	24	25	26
	32	17	16	15	16	17
	64	12	11	11	11	12
400	4	126	119	119	122	125
	8	73	68	68	70	72
	16	44	39	40	41	43
	32	28	25	25	26	27
	64	18	16	16	17	18
800	4	238	227	228	230	234
	8	132	124	125	127	130
	16	76	71	71	72	74
	32	45	42	42	43	44
	64	29	26	26	27	28

Table 1: Empirical 5% critical values for the Scan Statistic calculated from 1000 simulations using the continuous, disjoint and overlapping intervals procedures.

2.2.2 Direct Approximation of Critical Values

A second approach to the problem of assessing the significance of the Scan Statistic is to find some method of approximating critical values simply and directly, without necessarily making reference to a distributional result. Figure 1 plots 5% critical values, for sample size N = 50, against the relevant interval widths, d = 2/3, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64.

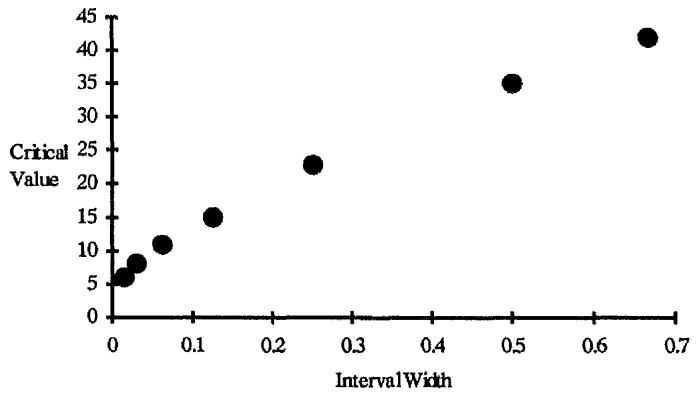


Figure 1: Plot of simulated 5% critical values for the Scan Statistic against seven interval widths, for $N = 50$. Based on 1000 replications.

The critical values were obtained by simulation of a continuous scanning process, using 1000 independent replications for each interval size. Some of the results appear in numeric form in Table 1 in the previous section. The monotonic increase of critical value with d, observed in Figure 1, is typical of the pattern for all sample sizes. If the 5% critical value, given N and d, is denoted by $\omega_{N,d}$, then a linear model of the form

$$\frac{\omega_{N,d}}{N} = \alpha + \beta.d \qquad (2.2)$$

might be plausible, where α and β are, respectively, intercept and slope parameters, and the standardisation by N is an attempt to make the model independent of sample size.

However, the curvature apparent at small values of d in Figure 1 is not removed by simple transformations, such as $\log_e$ or square root. This may indicate that a quadratic model of the form

$$\frac{\omega_{N,d}}{N} = \alpha + \beta.d + \gamma.d^2 \qquad (2.3)$$

would be more suitable.

Parameter estimates were obtained for (2.2) and (2.3) by least squares from simulated critical values for seven choices of each of N and d and these are displayed in Table 2. This is intended to be an *ad hoc* approach, rather than rigorous statistical modelling, so there is no assumption of a particular error structure, for example.

	Coefficients				
	Linear Model, (2.2)		Quadratic Model, (2.3)		
N	α	β	α	β	γ
20	0.270	1.096	0.221	1.824	-1.106
50	0.143	1.089	0.112	1.557	-0.711
100	0.098	1.076	0.080	1.338	-0.398
150	0.078	1.069	0.060	1.329	-0.395
200	0.062	1.067	0.046	1.297	-0.348
400	0.042	1.046	0.032	1.191	-0.222
800	0.029	1.032	0.021	1.149	-0.177

Table 2: Parameter estimates for models (2.2) and (2.3), for simulated critical values of the Scan Statistic regressed on seven choices of interval width.

The linear model (2.2) is usually much less accurate than the quadratic model (2.3), over the range of N considered here. For example, Table 3 compares the 5% critical values obtained by simulation with those calculated from (2.2) and (2.3), for N = 50. The quadratic model correctly estimates five of the values, whereas the linear model can only manage two. The accuracy of both models degrades as N increases, but (2.3) maintains its superior performance.

d	Continuous	Linear (2.2)	Quadratic (2.3)
2/3	42	44	42
1/2	35	35	36
1/4	23	21	23
1/8	15	14	15
1/16	11	11	11
1/32	8	9	8
1/64	6	9	7

Table 3: Comparison of 5% critical values for $N = 50$, calculated by simulation, (2.2) and (2.3).

Returning to Table 2, we see that the coefficients for the quadratic model appear to be converging, with increasing N, to 0, 1 and 0 for α, β and γ, respectively. This suggests that it might be possible to improve the accuracy of (2.3) by fitting linear models to the three parameters, with N as an explanatory variable, and then substituting the resulting expressions into the original model. After transformations to the logarithmic scale, the following models for the parameters were obtained:

$$\log \alpha = 0.331 - 0.629 \log N,$$

$$\log (\beta - 1) = 1.202 - 0.469 \log N$$

and

$$\log (-\gamma) = 1.576 - 0.505 \log N.$$

Back-substitution produced a new model for $\omega_{N,d}$:

$$\frac{\omega_{N,d}}{N} = 1.392 N^{-0.629} + \left(1 + 3.328 N^{-0.469}\right) d - 4.835 N^{-0.505} d^2. \qquad (2.4)$$

A second direct method of obtaining critical values is derived by generalising an argument used in the proof of Theorem IV in Naus (1966b), which holds for $d = 1/2$ and uses approximations to $P(n|N;d)$ for $n > (N+1)/2$. Consider now the case of $d \leq 1/2$ and, from (2.1), observe that for $n > (N+1)/2$

$$P(n|N;d) = (N-n+1)\left\{2\sum_{i=n}^{N} b(i|N;d) + b(n-1|N;d) - b(n|N;d)\right\} - 2(N-n)\sum_{i=n}^{N} b(i|N;d)$$

$$= 2\sum_{i=n}^{N} b(i|N;d) + (N-n+1)\{b(n-1|N;d) - b(n|N;d)\}$$

$$= 2\sum_{i=n}^{N} b(i|N;d) + \left(\frac{n}{d} - N - 1\right) b(n|N;d). \qquad (2.5)$$

Assume that N is large and employ a normal approximation to the binomial probabilities, so that

$$P(n|N;d) \approx \frac{2}{\sqrt{2\pi Nd(1-d)}} \int_{n-\frac{1}{2}}^{\infty} \exp\left\{-\frac{(x-Nd)^2}{2Nd(1-d)}\right\} dx$$

$$+ \frac{(n/d - N - 1)}{\sqrt{2\pi Nd(1-d)}} \int_{n-\frac{1}{2}}^{n+\frac{1}{2}} \exp\left\{-\frac{(x-Nd)^2}{2Nd(1-d)}\right\} dx.$$

By making the substitutions $y = (x - Nd)/\{Nd(1 - d)\}^{½}$ and $n = (N + k\sqrt{N})d$, and neglecting the first continuity correction, it can be demonstrated that

$$P(n|N;d) \approx 2\{1 - \phi(\theta k)\} + \frac{k \exp\left(-\frac{1}{2}\theta^2 k^2\right)}{\sqrt{2\pi d(1-d)}},$$

where $\theta = \{d/(1-d)\}^{½}$ and $\phi(\cdot)$ is the normal c.d.f.. Hence, for a significance level of α, choose $k = k_\alpha$, so that

$$2\{1 - \phi(\theta k_\alpha)\} + \frac{k_\alpha \exp\left(-\frac{1}{2}\theta^2 k_\alpha^2\right)}{\sqrt{2\pi d(1-d)}} \approx \alpha;$$

the corresponding approximate critical value is, therefore,

$$\hat{\omega}_{N,d} = \left(N + k_\alpha \sqrt{N}\right) d. \qquad (2.6)$$

For $d > 1/2$ or $n < (N+1)/2$, (2.6) is still applied, since expression (2.5) is then an approximation to the true value of $P(n|N;d)$; this approximation is derived by Wallenstein and Neff (1987) and forms one of the methods to be discussed in the next section.

2.2.3 Approximations to P(n|N;d)

The third approach to simplifying significance testing for the Scan Statistic is to use an approximation for $P(n|N;d)$ that is not computationally intensive. When $d < 1/2$, Glaz (1989) derives a sequence of approximations, for each of $L = 1, ..., N-n$, so that

$$P(n|N;d) \approx 1 - \hat{Q}_{N-n+1}^{(L+1)},$$

where

$$\hat{Q}_{N-n+1}^{(L+1)} = Q_L \left(Q_{L+1}/Q_L\right)^{N-n-L+1},$$

$$Q_i = Q_1 - \sum_{j=2}^{i} Q_j^*,$$

$$Q_1 = \Pr\left(A_0^c\right),$$

$$Q_j^* = \Pr\left\{A_0 \cap \left(\bigcap_{k=1}^{j-1} A_k^c\right)\right\},$$

A_i is the event $\{X_{(n+i-1)} - X_{(i)} \leq d\}$, a superscript 'c' denotes a set complement and $X_{(j)}$ is the j th order statistic from the sample. The terms Q_1 and Q_j^* can be written as weighted sums of binomial probabilities, which are straightforward to calculate.

The accuracy of the approximation increases with L, but so does the number of calculations. Glaz (1989) suggests $L = 3$ as a suitable compromise and, also, that the approximation is best when the average number of points in a scanning interval is low, *e.g.* $N \times d \leq 5$. When the latter condition does not hold, a suggested alternative is the approximation of Wallenstein and Neff (1987), which justifies the use of (2.5) as an approximation to $P(n|N;d)$ for all n and d, by derivation for the case where $d = 1/L$, L an integer, and then by arguing that the result should hold for more general d.

2.2.4 Comparison of Direct Methods and Approximations to P(n|N;d)

Following the model provided by Section 2.2.1, the proposals of Sections 2.2.2 and 2.2.3 are compared by their ability to produce accurate 5% critical values. Other criteria might be more appropriate if the overall behaviour of the different methods was of interest, but for significance testing purposes it is only the upper tail that is important. The benchmark values are those obtained by simulation (with 1000 replications), which were used previously in Table 1.

Table 4 presents the results for the quadratic model (2.3), the quadratic model incorporating a trend in the parameters (2.4), the approximate formula (2.6), and the approximations of Glaz (1989) and Wallenstein and Neff (1987). The two quadratic models, (2.3) and (2.4), are at best moderately accurate and, in fact, there seems to be little or no advantage to be gained from including the linear components for the convergence of the parameters to zero or one. The method generalised from Naus (1966b) is quite accurate for large d, *e.g.* $d \geq 1/4$, but poorer for small d. When the window is narrow, typical values of the Scan Statistic will be quite small, relative to N, so it is likely that $P(n|N;d)$ would have to be evaluated for $n < N/2$, whereas the expression used to derive (2.6) is valid only for $n > N/2$. Hence the tail probability under calculation is already an approximation, before the further approximating steps are taken. In addition, the derivation requires a binomial probability at a single value, $b(n|N;d)$, to be replaced by a normal approximation, which may not be adequate.

The results calculated from Glaz (1989) and Wallenstein and Neff (1987) are the best of the group, with most of the simulated critical values being reproduced by both methods. The latter technique would seem to have advantages over the former, since it does not require $d < 1/2$ and demands less computational effort. It would also appear that the guideline regarding a large average number of points in a scanning interval, suggested by Glaz (1989), has some justification. The inaccurate critical values estimated using this method occur for large d and N, which might imply a large mean event count, whereas the Wallenstein and Neff (1987) method gives more precise values at the same locations. For small d, however, it would appear that the methods have about the same success in approximating tail probabilities for significance testing.

N	d	Cont[1]	Quad[2]	(2.4)	(2.6)	Glaz	WN[3]
20	2/3	19	19	19	20	-	19
	1/2	17	18	18	17	-	17
	1/4	12	13	12	11	12	12
	1/8	9	9	9	8	9	9
	1/16	7	7	7	6	7	7
	1/32	5	6	6	4	5	5
	1/64	5	5	5	3	5	5
50	2/3	42	42	43	43	-	42
	1/2	35	36	36	35	-	35
	1/4	23	23	24	22	23	23
	1/8	15	15	15	14	15	15
	1/16	11	11	11	10	11	11
	1/32	8	8	9	7	8	8
	1/64	6	7	8	5	6	6
100	2/3	80	80	79	80	-	80
	1/2	64	65	66	64	-	64
	1/4	39	39	40	39	39	39
	1/8	25	25	25	24	25	25
	1/16	17	17	17	15	17	17
	1/32	12	13	12	10	12	12
	1/64	9	11	10	7	9	9
150	2/3	116	116	116	116	-	116
	1/2	93	94	94	93	-	93
	1/4	55	56	55	54	55	55
	1/8	34	33	33	33	34	34
	1/16	22	22	22	20	22	22
	1/32	15	16	16	13	15	15
	1/64	11	13	13	9	11	11

1. Simulated 5% critical values, 1000 replications. 2. Quadratic model (2.3). 3. Wallenstein and Neff (1987) approximation.

Table 4: Comparison of 5% critical values for the Scan Statistic, obtained by five different approximate methods, with those calculated from simulations. (Continued overleaf).

N	d	Cont[1]	Quad[2]	(2.4)	(2.6)	Glaz	WN[3]
200	2/3	152	152	151	152	-	151
	1/2	120	122	122	120	-	120
	1/4	70	70	70	69	70	70
	1/8	42	41	41	41	42	42
	1/16	26	26	26	25	26	26
	1/32	17	18	18	16	17	17
	1/64	12	14	14	10	12	12
400	2/3	292	292	292	293	-	292
	1/2	228	230	230	228	-	228
	1/4	126	127	128	127	129	127
	1/8	73	72	72	72	74	73
	1/16	44	43	43	42	44	44
	1/32	28	28	28	26	28	28
	1/64	18	21	21	16	18	18
800	2/3	568	567	569	570	-	569
	1/2	439	442	442	440	-	440
	1/4	238	239	238	238	241	238
	1/8	132	130	130	131	133	132
	1/16	76	75	74	74	76	76
	1/32	45	46	46	44	46	45
	1/64	29	32	31	27	29	29

1. Simulated 5% critical values, 1000 replications. 2. Quadratic model (2.3). 3. Wallenstein and Neff (1987) approximation.

Table 4: (Continued) Comparison of 5% critical values for the Scan Statistic, obtained by five different approximate methods, with those calculated from simulations.

2.3 Density Estimate Supremum Statistic

The number of points in a scanning interval is controlled by the true underlying probability density function in the corresponding region of (0,1], and is therefore related to the height of the density function in that area. This suggests that finding the Scan Statistic, which is the maximum number of points included in an interval of fixed length, is analogous to searching for the maximum height of some estimate of the true p.d.f., which has been calculated from the sample.

A useful approach to the estimation of a density function from a sample of data is nonparametric kernel density estimation, which has been described, for example, in the monograph of Silverman (1986). If the true p.d.f. is denoted by $f(x)$, from which a sample $\{X_1, ..., X_N\}$, of size N, has been drawn, then the corresponding estimate is

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right);$$

h is a smoothing parameter satisfying $h \to 0$ as $N \to \infty$ and $Nh \to \infty$ as $N \to \infty$, and $K(\cdot)$ is the kernel function, which satisfies the following conditions:

(a) $K(t) \geq 0, \forall t \in (-\infty, \infty)$,

(b) $\int_{-\infty}^{\infty} K(t)\,dt = 1$,

(c) $\int_{-\infty}^{\infty} tK(t)\,dt = 0$.

The usual choice for K is a symmetric probability density function, such as the standard normal (Gaussian) density.

Rosenblatt (1971) derives the asymptotic distribution of the maximum of a standardised kernel density estimate (k.d.e.) that is defined for the domain [0,1]. Although the conditions required in the proof are very strong (described in the paper as "unpleasant and completely impractical"), it is thought that the result is useful under much weaker conditions. If

$$\lambda = \int K^2(t)\,dt,$$

$$A = \log\left(\frac{B^{\frac{1}{2}}}{2\pi}\right)$$

and

$$B = -\frac{2}{\lambda}\frac{d^2}{dt^2}\left\{\int K(u)K(u+t)\,du\right\}\Big|_{t=0},$$

then

$$\Pr\left[\max_{0\le x\le 1}\left\{\frac{Nh}{\lambda f(x)}\right\}^{\frac{1}{2}}\left\{\hat{f}(x) - \mathbf{E}\hat{f}(x)\right\} \le \sqrt{2\log h^{-1}} + \frac{A+z}{\sqrt{2\log h^{-1}}}\right] \to \exp\{-\exp(-z)\}, \quad (2.7)$$

as $N \to \infty$, given conditions on $f(\cdot)$, $K(\cdot)$ and h.

In the current application, $f(x) = 1$, $\forall\, x \in [0,1]$. If $K(\cdot)$ is chosen to be the Gaussian density,

$$\lambda = (4\pi)^{-\frac{1}{2}} \quad \text{and} \quad \int K(u)K(u+t)\,du = (4\pi)^{-\frac{1}{2}}\exp\left(-t^2/4\right).$$

Therefore, $B = 1$ and $A = \log(1/2\pi)$. It can be demonstrated, using a Taylor expansion, that

$$\mathbf{E}\hat{f}(x) = f(x) + \tfrac{1}{2}h^2\left\{\int t^2K(t)\,dt\right\}f''(x) + o(h^2) \approx 1,$$

for $x \in [0,1]$. Finally, for a critical value at the $\alpha\%$ level, the value of z must be chosen so that

$$\exp\{-\exp(-z)\} = 1-\alpha \quad \Rightarrow \quad z = -\log\log\left\{(1-\alpha)^{-1}\right\}.$$

Substituting these quantities into (2.7) gives the result that, for large N,

$$\Pr\left[\max_{0\le x\le 1}\hat{f}(x) \ge 1 + \left(2Nh\sqrt{\pi}\right)^{-\frac{1}{2}}\left\{\sqrt{2\log h^{-1}} - \frac{\log\left\{2\pi\log(1-\alpha)^{-1}\right\}}{\sqrt{2\log h^{-1}}}\right\}\right] = \alpha, \quad (2.8)$$

and, hence, an expression for an approximate $\alpha\%$ critical value.

In Table 5, 5% critical values obtained from (2.8), for four different bandwidths (see below) and seven different sample sizes, are compared to empirical values, estimated by simulation. The latter results were based on 1000 sets of data, each one of which was generated by sampling from U(0,1]. The k.d.e. of the true p.d.f. for each set was calculated using the Fast Fourier Transform (FFT) algorithm of Silverman (1982), with the addition of the corrections suggested by Jones and Lotwick (1984). The algorithm requires the simulated data to be discretised to a grid of, say, 1024 points, by splitting

an observation with value x into two weights, $(z-x)/(z-y)$ and $(x-y)/(z-y)$, which are assigned to the grid points with values y and z respectively, where y (z) is the point immediately to the left (right) of x. The FFT of the grid, calculated with wrap-around edge conditions, is multiplied by the Fourier transform (FT) of the Gaussian kernel, to give the discrete FT of the k.d.e.. An inverse transformation provides values of the density estimate at 1024 distinct points. The maximum of these was taken to be a reasonable approximation to the supremum of the k.d.e. over [0,1].

		Bandwidth, c =			
N	Method	0.5	1.0	1.5	2.0
20	Sim	2.107	1.526	1.261	1.108
	(2.8)	2.162	1.748	1.575	1.414
50	Sim	1.745	1.387	1.224	1.122
	(2.8)	1.824	1.532	1.410	1.294
100	Sim	1.600	1.321	1.197	1.120
	(2.8)	1.635	1.411	1.317	1.228
150	Sim	1.514	1.280	1.183	1.118
	(2.8)	1.545	1.353	1.273	1.196
200	Sim	1.444	1.241	1.162	1.108
	(2.8)	1.489	1.317	1.245	1.176
400	Sim	1.350	1.194	1.130	1.091
	(2.8)	1.376	1.245	1.189	1.136
800	Sim	1.264	1.159	1.112	1.083
	(2.8)	1.289	1.189	1.146	1.105

Table 5: Comparison of 5% critical values for the density estimate supremum statistic, found by simulation and the asymptotic result (2.8).

The same smoothing parameters were used for both the simulations and the approximations from (2.8). The usual formula for an optimal value of h, obtained by minimising approximate mean integrated squared error, is

$$h_{opt} = \left\{\int t^2 K(t)\,dt\right\}^{-\frac{2}{5}} \left\{\int K^2(t)\,dt\right\}^{\frac{1}{5}} \left\{\int f''(x)^2\,dx\right\}^{-\frac{1}{5}} N^{-\frac{1}{5}}; \qquad (2.9)$$

e.g. if the observations are drawn from a normal distribution with variance σ^2, say,

$$h_{opt} = 1.06\sigma N^{-\frac{1}{5}}.$$

The uniform p.d.f. has zero second derivative, preventing (2.9) from being used directly. It does, however, suggest that bandwidths of the form

$$h = c\sigma N^{-\frac{1}{5}} \qquad (2.10)$$

would be sensible, where σ is set equal to the standard deviation of the uniform distribution, *i.e.* $\sigma = 12^{-\frac{1}{2}}$. In Table 5, the value of the constant c was chosen from the set {0.5, 1.0, 1.5, 2.0}.

When smoothing examples of the simulated data sets, the bandwidth with c = 0.5 in (2.10) gave the best results to the eye; it also achieved two significant figure accuracy over most of the range of N, which was better than the other three choices. The overall accuracy of approximation (2.8) is poor, but improves for the smaller bandwidths and for larger sample sizes. The simulated values, it should be noted, may themselves lack precision. First, they are based on the empirical distribution of maxima that are found by a grid search over [0,1], and are, therefore, likely to be underestimates. Secondly, the empirical critical value is effectively discretised, being the 951st order statistic from a sample of 1000 simulated maxima, whereas the true critical value is measured on a continuous scale. Hence, accuracy to a large number of decimal places would not be expected.

2.4 Power Comparison

In Sections 2.2 and 2.3, two statistics were described that were designed to detect the clustering of points along the real line. Methods of approximating critical values for significance testing were also investigated. This section attempts to compare the Scan Statistic and the density estimate supremum statistic (DESUP) on the basis of their ability to identify data that has been sampled from a distribution representing a small cluster within a larger scale random (uniform) pattern.

The best method for calculating critical values for the Scan Statistic proved to be the approximation to $P(n|N;d)$ due to Wallenstein and Neff (1987), hereafter referred to as WN. Results for the approximation of Glaz (1989) were almost identical, but required

greater computational effort. The least accurate of the methods considered were the quadratic models, (2.3) and (2.4), of which the former, denoted by Q, seemed to have a slight advantage in that it was marginally simpler. Approximation (2.6) fell somewhere in between. To indicate the range of figures for power that would be expected for the Scan Statistic, when using these methods to obtain critical values, only WN and Q are included in the comparison.

The study was carried out by simulation, for tests at the 5% level. One thousand sets of artificial data were generated for each technique, sample size and, in the case of WN and Q, interval length, by sampling from the mixture distribution

$$p\ U(0,1] + (1-p)\ U(0.5-\varepsilon,0.5+\varepsilon], \qquad (2.11)$$

where $U(a,b]$ is the uniform p.d.f. on $(a,b]$, p is a mixture parameter chosen from the set $\{0.8, 0.9, 0.95, 1.0\}$ and $\varepsilon = 0.025$. The significance level of each test was taken to be 0.05 and the value $p = 1.0$ was included to check empirically the level accuracy of the three methods. Critical values for DESUP were obtained from (2.8), using a bandwidth of form (2.10) with $c = 0.5$.

Table 6 summarises the results for $p = 0.8, 0.9, 0.95$. The power of all three methods shows the expected increase with sample size and decrease with p. The mixture parameter has the interpretation of one minus the proportion of events that might be expected in the cluster, so $p = 0.8$ corresponds to an average of 20%, for example. The three methods are clearly more powerful for larger clusters. There is considerable variability with interval length for WN and Q, with smaller intervals generally giving greater power. There seems to be evidence indicating that power increases to a maximum at around $d = 1/16$, decreasing thereafter. The width of the cluster generated by (2.11) is 0.05, so it would seem that the power of the Scan Statistic is maximised when the length of the scanning interval is approximately the same as that of the interval within which the clustering effect is operating. This observation confirms a similar simulation result in the discussion of Wallenstein and Neff (1987).

WN is more powerful than Q for $p = 0.8$ and for large N when $p = 0.9$, but for a small cluster there is little difference between them. One feature of Q that is apparent over the entire range of N and p is the often sharp decrease in power for very small d. Examination of earlier results suggests that Q almost always overestimates the relevant critical value and, hence, that a test based on Q for $d = 1/64$ is very conservative. DESUP is usually less powerful than the best results of either WN or Q for a given choice of N and p.

		$p = 0.8$			$p = 0.9$			$p = 0.95$		
N	d	WN	Q	D^1	WN	Q	D^1	WN	Q	D^1
20	2/3	0.101	0.101		0.062	0.062		0.053	0.053	
	1/2	0.114	0.024		0.046	0.006		0.026	0.004	
	1/4	0.180	0.092		0.050	0.013		0.028	0.005	
	1/8	0.200	0.221	0.267	0.045	0.043	0.070	0.019	0.016	0.044
	1/16	0.237	0.284		0.040	0.056		0.013	0.014	
	1/32	0.408	0.189		0.102	0.024		0.057	0.010	
	1/64	0.093	0.123		0.090	0.027		0.005	0.005	
50	2/3	0.298	0.314		0.128	0.148		0.074	0.086	
	1/2	0.424	0.311		0.181	0.092		0.084	0.032	
	1/4	0.542	0.534		0.135	0.148		0.047	0.061	
	1/8	0.750	0.744	0.664	0.219	0.243	0.175	0.068	0.079	0.049
	1/16	0.809	0.808		0.212	0.247		0.053	0.055	
	1/32	0.806	0.797		0.231	0.265		0.068	0.063	
	1/64	0.715	0.457		0.221	0.062		0.079	0.090	
100	2/3	0.380	0.412		0.118	0.132		0.064	0.057	
	1/2	0.744	0.632		0.285	0.215		0.124	0.083	
	1/4	0.904	0.909		0.352	0.362		0.097	0.116	
	1/8	0.959	0.963	0.964	0.429	0.438	0.435	0.089	0.118	0.092
	1/16	0.986	0.988		0.534	0.550		0.093	0.110	
	1/32	0.981	0.948		0.481	0.351		0.084	0.039	
	1/64	0.936	0.647		0.289	0.057		0.053	0.007	
150	2/3	0.609	0.611		0.202	0.211		0.081	0.095	
	1/2	0.859	0.798		0.332	0.244		0.112	0.078	
	1/4	0.979	0.965		0.480	0.414		0.093	0.112	
	1/8	0.998	0.999	0.999	0.659	0.719	0.639	0.136	0.203	0.145
	1/16	1.000	0.999		0.785	0.791		0.184	0.189	
	1/32	1.000	0.999		0.734	0.639		0.160	0.112	
	1/64	0.988	0.927		0.504	0.191		0.070	0.013	

1. DESUP statistic of Section 2.3. Note that values of d apply only to WN and Q columns.

Table 6: Empirical power of three tests of one-dimensional clustering, using 1000 replications, against alternatives with three sizes of cluster. (Continued overleaf).

		$p = 0.8$			$p = 0.9$			$p = 0.95$		
N	d	WN	Q	D^I	WN	Q	D^I	WN	Q	D^I
200	2/3	0.802	0.726		0.303	0.260		0.128	0.102	
	1/2	0.967	0.916		0.491	0.351		0.176	0.102	
	1/4	0.997	0.997		0.642	0.613		0.165	0.166	
	1/8	1.000	1.000	0.999	0.820	0.853	0.818	0.226	0.308	0.232
	1/16	1.000	1.000		0.944	0.930		0.362	0.376	
	1/32	1.000	1.000		0.925	0.847		0.349	0.240	
	1/64	1.000	0.999		0.798	0.466		0.218	0.043	
400	2/3	0.978	0.974		0.505	0.516		0.201	0.226	
	1/2	1.000	1.000		0.786	0.680		0.287	0.248	
	1/4	1.000	1.000		0.944	0.946		0.380	0.390	
	1/8	1.000	1.000	1.000	0.990	0.993	0.999	0.546	0.618	0.565
	1/16	1.000	1.000		0.999	0.999		0.682	0.738	
	1/32	1.000	1.000		0.998	0.994		0.604	0.604	
	1/64	1.000	1.000		0.992	0.890		0.476	0.135	
800	2/3	1.000	1.000		0.818	0.858		0.331	0.408	
	1/2	1.000	1.000		0.980	0.942		0.484	0.406	
	1/4	1.000	1.000		1.000	1.000		0.709	0.670	
	1/8	1.000	1.000	1.000	1.000	1.000	1.000	0.890	0.911	0.921
	1/16	1.000	1.000		1.000	1.000		0.967	0.970	
	1/32	1.000	1.000		1.000	1.000		0.961	0.936	
	1/64	1.000	1.000		1.000	0.999		0.823	0.511	

1. DESUP statistic of Section 2.3. Note that values of d apply only to WN and Q columns.

Table 6: (Continued) Empirical power of three tests of one-dimensional clustering, using 1000 replications, against alternatives with three sizes of cluster.

The level accuracy of all three methods is moderate and the tests are usually quite conservative. For example, empirical significance levels are displayed in Table 7 for $N = 50$; a full range of interval lengths is included for WN and Q. Conservatism in the Scan Statistic may result from a discretisation effect; *i.e.* it may not be possible to achieve a level of 0.05 in practice, since the statistic can only take integer values. Table 7 also contains columns of empirical levels for WN and Q that were calculated using critical values that had been reduced by one. For WN, all the levels are greater than 0.05, while five out of seven are greater than 0.05 for Q, suggesting that discretisation may be a plausible explanation, and that Q's level accuracy may be less than that of WN. Q is especially poor for $d = 1/64$, for the reasons noted above.

d	WN	WN - reduced[1]	Q	Q - reduced[1]	DESUP[2]
2/3	0.045	0.101	0.046	0.116	
1/2	0.043	0.096	0.015	0.037	
1/4	0.029	0.066	0.039	0.071	
1/8	0.036	0.090	0.039	0.118	0.022
1/16	0.022	0.074	0.015	0.082	
1/32	0.029	0.125	0.022	0.115	
1/64	0.042	0.232	0.002	0.041	

1. Empirical level calculated using critical values reduced by one. 2. Independent of d.

Table 7: Empirical significance levels for the density estimate supremum statistic and two critical value approximation methods for the Scan Statistic, for $N = 50$. Simulations based on 1000 replications.

Power and level accuracy results have not been included in Tables 6 or 7 for significance testing by simulation, *i.e.* by comparing the Scan Statistic for each artificial data set sampled from (2.11) to the appropriate simulated critical value from the first column of Table 4. The missing entries are almost identical to those of WN, because of the close agreement between the two sets of critical values (see Section 2.2.4), and, hence, the comments regarding the performance and level accuracy of WN apply to the

simulation technique. Alternatively, one could use a Monte Carlo significance test (Barnard, 1963; Hope, 1968), in place of the empirical critical values, at each replication; however, the results should be the same, if allowance is made for sampling variability.

2.5 Conclusion

The Scan Statistic and density estimate supremum statistic are reasonably effective methods of detecting clustering in one dimension. For the former technique, the use of an approximation to $P(n|N;d)$ that may be calculated easily, such as that of Wallenstein and Neff (1987), is more satisfactory than estimating critical values directly, or using scanning intervals that are disjoint or discrete and overlapping. Significance testing for the k.d.e. based statistic may employ an asymptotic result, but the accuracy of the critical values obtained can be poor. Simulation provides an alternative approach to the problem that is quite attractive for both methods.

For the simple alternative hypotheses considered, the Scan Statistic was more powerful than the k.d.e. maximum, provided the size of the scanning interval was chosen to reflect the scale at which clustering was occurring. If this parameter was mis-specified, the supremum statistic performed very favourably, although the results were similarly dependent on an arbitrary constant, *i.e.* the bandwidth, h. The Scan Statistic could also be regarded as a smoothing technique, *e.g.* a type of moving average procedure, and thus both the Scan Statistic and the density estimate supremum statistic can be seen to share a problem common to all smoothing methods, *i.e.* dependence on the particular values of a smoothing parameter.

CHAPTER 3

THE SCAN STATISTIC IN TWO DIMENSIONS

3.1 Introduction

In one dimension, the Scan Statistic is calculated from counts of points within a window that moves in a continuous fashion over the unit interval. If we wish to consider an analogous method in two dimensions, it is natural to generalise the unit interval to the unit square, $[0,1] \times [0,1]$, and the scanning interval to a scanning rectangle, with sides of length u and v, parallel to the x and y axes, respectively. This formulation of the two-dimensional statistic was first proposed by Naus (1965b).

Let N events be distributed within the unit square and let

$$n_{xy} = \text{the number of points within the rectangle } [x - u,x] \times [y - v,y].$$

The two-dimensional version of the Scan Statistic is defined to be

$$S_{uv} = \max_{\substack{x \in [u,1] \\ y \in [v,1]}} n_{xy}.$$

The decision to refer to a particular rectangle by its top right-hand corner is arbitrary, but convenient for computational purposes. A two-dimensional version of the algorithm for calculating the Scan Statistic that was discussed in Section 2.2 can be developed easily, since the observations made previously regarding sufficient coverage of the region of interest generalise to two dimensions. Each event location is used in turn as a reference point for a one-dimensional scan, with the first count being made in the square with top right-hand corner located at the reference event, and subsequent counts being made when new events enter the square as it scans horizontally; the scan finishes when the current reference point is dropped by the window. The next event in the sequence is then selected as the reference point, and the process is repeated. This algorithm calculates the Scan Statistic correctly, but, as in the case of the one-dimensional version, does not necessarily scan all possible locations.

Using notation similar to that of Section 2.2, the tail probability

$$\Pr(S_{uv} \geq n | N;u,v)$$

may be abbreviated to

$$P(n|N;u,v).$$

Under the relevant null hypothesis, which is that there is no process operating in the unit square that would generate clusters, the N events are assumed to be sampled from a uniform distribution that is defined on $[0,1] \times [0,1]$.

Conover *et al* (1979) describe an application of the theory of one-dimensional Scan Statistics to the positions on a two-dimensional map of anomalous radioactivity counts from bismuth-214, which have been measured along parallel flight lines from aerial reconnaissance aircraft. A finite number of discrete horizontal scans is performed along the paths of these samples, with the height of the scanning window chosen so that a given number of lines is included at one time, *e.g.* three paths. For this application, the number of points inside a window at a specific location has the same distribution as the analagous count of points in the one-dimensional case, allowing the exact results of Naus (1966a), for example, or approximate results for $P(n|N;d)$ to be employed. Those areas achieving a count exceeding a pre-specified critical value are shaded on a map of the observations. Further investigation is to be targeted subsequently on areas with a high density of shading. The algorithm is qualitatively similar to Openshaw *et al* (1987, 1988), and may therefore share the deficiencies of that method, which were discussed in Section 1.2.3. One of the main weaknesses is the dependence between overlapping, shaded areas, since different passes through the data may share flight lines.

An initial investigation of the behaviour of the two-dimensional Scan Statistic is made by simulation in Section 3.2. Naus (1965b) presents upper and lower bounds for $P(n|N;u,v)$, which are discussed further in Section 3.3. The evaluation of empirical power is undertaken by simulation in Section 3.4, and Section 3.5 summarises the conclusions of the preceding work. Section 3.6 is in the spirit of an aside; it considers a problem with multiplicative linear congruential pseudo-random number generators that was encountered while carrying out the simulations reported in Section 3.2.

3.2 Simulation of Critical Values

Simulation of the Scan Statistic under the null hypothesis stated in Section 3.1 can provide useful information about the expected behaviour of the test procedure, and provides a reference point for later work. We will concentrate on the case of a square

scanning window, a decision made partly for convenience, but also prompted by two observations that suggest a square may be the optimal shape, in some sense. First, a comment in Naus (1965b) suggests that, for a given, fixed area of scanning window, a square gives a higher probability of seeing a large cluster than would a rectangle, in the sense of maximising $P(N|N;u,v)$ subject to $uv = A$, A a constant. Secondly, a rectangle may be more appropriate for situations where there is prior information to suggest that one co-ordinate direction may be more important than the other for measuring clustering. In the type of examples used in this investigation, there is no such anisotropic effect; therefore, subsequent work assumes $u = v = d$.

N	d						
	2/3	1/2	1/4	1/8	1/16	1/32	1/64
50	34 (0.048)	24 (0.046)	12 (0.014)	7 (0.018)	5 (0.007)	4 (0.003)	3 (0.014)
100	62 (0.034)	41 (0.043)	18 (0.039)	10 (0.009)	6 (0.032)	4 (0.049)	4 (0.003)
200	113 (0.038)	73 (0.048)	28 (0.050)	14 (0.022)	8 (0.008)	5 (0.034)	4 (0.013)
400	211 (0.049)	132 (0.044)	47 (0.036)	20 (0.033)	11 (0.014)	7 (0.006)	5 (0.007)

Table 8: Empirical 5% critical values for the two-dimensional Scan Statistic, calculated from 1000 simulations. Exact significance levels in parentheses.

Table 8 displays some empirical 5% critical values and the simulated upper tail probability to which each corresponds. The results are derived from the empirical frequency distributions of 1000 replications, with the co-ordinates of events being

simulated by sampling from the uniform distribution on the unit square. Some of the tail probabilities are much smaller than 0.05, especially for small d. It would appear that the Scan Statistic suffers from a discretisation problem in two dimensions, just as it does in one, *i.e.* it may not be possible to attain a significance level of precisely 0.05, because of the restriction of the statistic to integer values.

The following example helps to explain why there is considerable variability in the empirical levels in Table 8 for small d. The empirical distribution in the above simulations for $N = 50$ and $d = 1/2$ was

S	15	16	17	18	19	20	21	22	23	24	25	26	27
F	10	35	97	163	217	166	123	94	49	24	15	3	4

where S represents values of the statistic and F the frequency, whereas the empirical distribution for $d = 1/32$ was

Statistic	1	2	3	4
Frequency	2	876	119	3

The range of values taken by the Scan Statistic for small d is much narrower than that for larger scanning squares. There is a greater concentration of weight on a few central values and the tails are shorter. This can lead to a greater difference in the estimates of $P(n|N;d,d)$ for adjacent values of n than would be observed for larger choices of d.

A point of interest is that, for the same value of N, the critical values in Table 8 for a square of side d are of the same order of magnitude as those in Table 4, for the continuous one-dimensional Scan Statistic for an interval of length d^2. This seems plausible, since the two windows cover the same proportion of their respective regions of interest.

3.3 Bounds for Tail Probabilities

Naus (1965b) presents upper and lower bounds for $P(n|N;u,v)$ that have the same limiting form, as u and v tend to zero. If

$$P'(1{:}n|N;u,v) = \Pr[\text{One and only one set of } n \text{ points is contained within a } u \times v \text{ rectangle; no rectangle with the same dimensions contains more than } n \text{ points}],$$

then it is demonstrated that

$$L \le P(n|N;u,v) \le U,$$

where

$$L = \max\{P(n|N;u,1)P(n|n;1,v);\ P(n|N;1,v)P(n|n;u,1)\}$$

and

$$U = P(n|N;u,1)P(n|N;1,v) - \left\{1 - \binom{N}{n}^{-1}\right\} P'(1{:}n|N;u,1)P'(1{:}n|N;1,v).$$

It is also noted in Naus (1965b) that if either u or v is equal to 1, $P(n|N;u,v)$ and $P'(1{:}n|N;u,v)$ reduce to the one-dimensional Scan Statistic tail probabilities for intervals of size equal to the measurement that does not have the value of unity. Hence, the bounds for a $d \times d$, square scanning window can be rewritten in the form

$$L = P(n|N;d)P(n|n;d) \tag{3.1}$$

and

$$U = \{P(n|N;d)\}^2 - \left\{1 - \binom{N}{n}^{-1}\right\} \{P'(1{:}n|N;d)\}^2, \tag{3.2}$$

where $P'(1{:}n|N;d)$ is the one-dimensional equivalent of $P'(1{:}n|N;d,d)$, in an obvious notation.

Calculation of the bounds to $P(n|N;d,d)$ from (3.1) and (3.2) requires evaluation of three probabilities, namely $P(n|N;d)$, $P(n|n;d)$ and $P'(1{:}n|N;d)$. The first two are standard one-dimensional tail probabilities for the Scan Statistic, for which different exact and approximate results were discussed in Chapter 2. It is difficult to derive the third quantity analytically, but it can be estimated empirically by the proportion of those Scan Statistics equal to n and found at only one location in, say, 1000 simulations, where N artificial events for each replication are sampled from the uniform distribution on (0,1].

The best approximation to $P(n|N;d)$ in Chapter 2 was the result of Wallenstein and Neff (1987), denoted by WN, which was both accurate and computationally undemanding when used to calculate 5% critical values. Table 9 displays some specimen values of (3.1) and (3.2) for a scanning square of side d and a sample size of 20, which were produced using a combination of WN and simulation. The values of n included in the

table were chosen to bracket the appropriate two-dimensional 5% critical values obtained by simulation in the previous section.

Five of the values for the upper bound are greater than one, which is a worrying feature; since (3.2) is defined to be the difference of a pair of probabilities, the calculation should produce a result that is less than one. This behaviour is a consequence of a weakness in WN that was described in Glaz (1989). When the true value is large, *i.e.* when n is small with respect to N, WN may generate an approximation to $P(n|N;d)$ that is greater than one. In addition, the magnitude of the error increases as $\mu = Nd$ decreases. The approximations to $P(n|N;d)$ corresponding to the entries in Table 9 are shown in Table 10. The pattern of WN results that are greater than one matches that of the upper bounds in Table 9.

WN can be applied successfully to the estimation of tail probabilities for significance testing in the context of a single dimension, since the relevant quantities are usually in the range 0.01 to 0.1. Probabilities of this size ensure that n is quite large relative to N, so the result is very accurate. In two dimensions, however, values of n that give tail probabilities in the above range are much smaller than the corresponding values in the one-dimensional case. The two-dimensional bounds require the calculation of one-dimensional tail probabilities for the same choice of n, and so the WN approximation has to be evaluated in a region in which its performance is poorer. Table 10 indicates clearly that WN is unsuitable for calculating (3.1) and (3.2).

A different approach to evaluating the bounds for $P(n|N;d,d)$ is to make use of the tables of exact values for $P(n|N;d)$ that have been included in, for example, Naus (1966a) and Wallenstein and Naus (1974). Results from the latter reference cover sample sizes of up to 100 for a limited range of interval sizes and ordinates, n. These have been used in the construction of Table 11, which compares the lower bound (3.1), the upper bound (3.2) and simulated values of $P(n|N;d,d)$, based on 10000 replications. $P'(1{:}n|N;d)$ was also estimated from 10000 simulations, and it is straightforward to calculate $P(n|n;d)$ from the exact probability statement (2.5) in Chapter 2. The choices of n for which exact $P(n|N;d)$ exist are not ideal for a comparison directed towards significance testing, since they are generally larger than the corresponding two-dimensional 5% critical values (see Table 8). This also means that many of the associated tail probabilities are too small to be estimated from 10 000 simulations. However, the available parameter combinations may be sufficient to indicate suitable conditions for obtaining good performance from (3.1) and (3.2).

d	n	Lower Bound, (3.1)	Upper Bound, (3.2)
1/4	6	5.898×10^{-3}	1.6167
	7	1.632×10^{-3}	1.4761
	8	3.332×10^{-4}	0.7381
1/16	3	0.0310	7.6197
	4	1.115×10^{-3}	1.3789
	5	2.499×10^{-5}	0.0864
1/64	2	0.1183	14.5560
	3	4.176×10^{-4}	0.2309
	4	8.026×10^{-7}	9.841×10^{-4}

Table 9: Examples of the lower and upper bounds, (3.1) and (3.2), to $P(n|N;d,d)$ for $N = 20$ events. Calculated using the Wallenstein and Neff (1987) approximation and simulation.

d	n	WN approximation to $P(n|N;d)$
1/4	6	1.2715
	7	1.2153
	8	0.8734
1/16	3	2.7604
	4	1.1970
	5	0.3447
1/64	2	3.8152
	3	0.5761
	4	0.0532

Table 10: Sample values of the Wallenstein and Neff (1987) approximation to $P(n|N;d)$ for $N = 20$, for comparison with Table 9.

The upper and lower bounds show little evidence of convergence to some estimate of $P(n|N;d,d)$ over the range of parameters considered, although the absolute difference between L and U decreases as n and d increase and decrease, respectively. It is possible that a tighter bounding interval might be achieved for much smaller values of d. The upper bound is often much larger than the simulated value of $P(n|N;d,d)$, although, again, the error is less for larger n and smaller d. However, noting the above comment on the available range of n, the greater differences for small n between (3.1), (3.2) and the simulated values of $P(n|N;d,d)$ suggest that the bounds are not particularly sharp for choices of n close to the relevant 5% critical value for the two-dimensional Scan Statistic.

N	n	d	Lower Bound	Upper Bound	Simulation
10	4	1/4	5.012×10^{-2}	0.9675	0.2594
		1/6	1.248×10^{-2}	0.5040	0.0419
		1/8	3.639×10^{-3}	0.1786	0.0098
	5	1/4	1.050×10^{-2}	0.3702	0.0361
		1/6	9.295×10^{-4}	4.548×10^{-2}	0.0018
		1/8	1.351×10^{-4}	7.097×10^{-3}	0.0003
20	7	1/4	1.328×10^{-3}	0.9765	0.0399
		1/6	7.523×10^{-5}	0.2985	0.0008
		1/8	5.770×10^{-6}	4.226×10^{-2}	*[2]
	8	1/4	3.120×10^{-4}	0.6404	0.0056
		1/6	6.054×10^{-6}	4.703×10^{-2}	*
		1/8	2.344×10^{-7}	2.903×10^{-3}	*
	9	1/4	5.170×10^{-5}	0.2036	0.0008
		1/6	3.656×10^{-7}	4.542×10^{-3}	*
		1/8	7.153×10^{-9}	1.084×10^{-4}	*
30	11	1/4	6.890×10^{-6}	0.7069	0.0007
		1/6	2.748×10^{-8}	2.562×10^{-2}	*
		1/8	2.724×10^{-10}	7.012×10^{-4}	*
	12	1/4	1.286×10^{-6}	0.3168	0.0002
		1/6	1.821×10^{-9}	2.972×10^{-3}	*
		1/8	9.895×10^{-12}	4.375×10^{-5}	*
	13	1/4	1.931×10^{-7}	9.274×10^{-2}	*
		1/6	1.061×10^{-10}	2.970×10^{-4}	*
		1/8	3.347×10^{-13}	3.360×10^{-6}	*
40	14	1/4	1.368×10^{-7}	0.7175	0.0001
		1/6	1.078×10^{-10}	1.167×10^{-2}	*
		1/8	2.701×10^{-13}	9.776×10^{-5}	*
	15	1/4	2.682×10^{-8}	0.3729	*
		1/6	7.435×10^{-12}	1.592×10^{-3}	*
		1/8	9.038×10^{-15}	6.750×10^{-6}	*
	16	1/4	4.427×10^{-9}	0.1379	*
		1/6	4.594×10^{-13}	1.695×10^{-4}	*
		1/8	-[1]	-	*

1. No value of $P(n|N;d)$ available.
2. Simulation estimates tail probability to be less than 1.0×10^{-4}

Table 11: Comparison of (3.1), (3.2) and simulated values of $P(n|N;d,d)$ based on 10000 replications, using exact values of $P(n|N;d)$ from Wallenstein and Naus (1974).

3.4 Investigation of Empirical Power

The level accuracy of a test based on the two-dimensional Scan Statistic is reduced by the restriction of the statistic to integer values (see Section 3.2). To counteract this problem, significance testing in the following investigation of power is carried out with the addition of a Randomised Rule (Gibbons, 1986). If the Scan Statistic calculated using a square of side d, denoted by S_d, is compared to a critical value c, which corresponds to an actual significance level of α_1, where $\alpha_1 < \alpha$, and α is the intended significance level, then the standard hypothesis test rejects H0 at the α_1 level if

$$S_d \geq c. \tag{3.3}$$

A test at level α may be constructed by using the following decision rule instead of (3.3):

reject H0 with probability 1 if $S_d \geq c$;
reject H0 with probability ϕ if $S_d = (c - 1)$;
do not reject H0 if $S_d < (c - 1)$. (3.4)

The overall Type I Error for this scheme is

$$\alpha_1 + \phi(\alpha_2 - \alpha_1),$$

where α_2 is the significance level corresponding to a critical value of $(c - 1)$, such that $\alpha_2 > \alpha$. The probability ϕ is chosen to ensure that

$$\alpha_1 + \phi(\alpha_2 - \alpha_1) = \alpha \quad \Rightarrow \quad \phi = \frac{\alpha - \alpha_1}{\alpha_2 - \alpha_1}. \tag{3.5}$$

The power of the Scan Statistic to detect a single cluster alternative was investigated by simulation. One thousand sets of artificial data were generated by sampling from a bivariate version of (2.11), the mixture used to evaluate the power of the one-dimensional test in Section 2.4. Specifically, the p.d.f. of the locations of the events was

$$p\underline{U}(0,1] + (1 - p)\underline{U}(k - \varepsilon, k + \varepsilon], \tag{3.6}$$

where the mixture parameter, p, was selected from {0.8, 0.9, 0.95}, $\varepsilon = 0.025$ and $k = 0.5$, 0.75, giving two possible cluster locations. $\underline{U}(a,b]$ represents the bivariate uniform p.d.f. on the square with lower left-hand corner (a,a) and top right-hand corner (b,b). Table 12 shows the empirical power of the Scan Statistic in two dimensions against alternative (3.6), with each test employing simulated critical values from Table 8 and the Randomised Rule (3.4), in order to ensure that an exact significance level of 0.05 was attained. Values of α_1 and α_2 in (3.5) were estimated from the empirical null distributions calculated for Table 8.

		k = 0.5			k = 0.75		
N	d	p = 0.8	p = 0.9	p = 0.95	p = 0.8	p = 0.9	p = 0.95
50	2/3	0.521	0.217	0.121	0.390	0.136	0.076
	1/2	0.836	0.363	0.159	0.584	0.183	0.071
	1/4	0.943	0.504	0.208	0.949	0.509	0.199
	1/8	0.971	0.561	0.203	0.969	0.566	0.195
	1/16	0.990	0.699	0.278	0.988	0.693	0.259
	1/32	0.972	0.636	0.259	0.972	0.635	0.258
	1/64	0.887	0.400	0.141	0.889	0.407	0.135
100	2/3	0.842	0.321	0.134	0.697	0.215	0.091
	1/2	0.983	0.645	0.288	0.916	0.354	0.139
	1/4	0.999	0.763	0.270	0.999	0.776	0.241
	1/8	1.000	0.926	0.438	1.000	0.930	0.436
	1/16	1.000	0.975	0.566	1.000	0.976	0.576
	1/32	1.000	0.970	0.595	1.000	0.969	0.591
	1/64	0.998	0.783	0.346	0.998	0.785	0.334
200	2/3	0.995	0.602	0.232	0.955	0.419	0.160
	1/2	1.000	0.886	0.397	0.997	0.647	0.177
	1/4	1.000	0.979	0.541	1.000	0.979	0.546
	1/8	1.000	0.998	0.794	1.000	0.997	0.759
	1/16	1.000	1.000	0.915	1.000	1.000	0.916
	1/32	1.000	0.999	0.887	1.000	1.000	0.888
	1/64	1.000	0.988	0.608	1.000	0.989	0.600
400	2/3	1.000	0.863	0.356	1.000	0.720	0.246
	1/2	1.000	0.997	0.693	1.000	0.933	0.384
	1/4	1.000	1.000	0.884	1.000	1.000	0.900
	1/8	1.000	1.000	0.989	1.000	1.000	0.986
	1/16	1.000	1.000	1.000	1.000	1.000	0.998
	1/32	1.000	1.000	0.994	1.000	1.000	0.995
	1/64	1.000	1.000	0.915	1.000	1.000	0.920

Table 12: Empirical power of the two-dimensional Scan Statistic, with a square scanning window, against alternative (3.6), with three sizes of cluster (p) and two locations (k). Based on 1000 simulations.

The pattern of results for the two-dimensional statistic in Table 12 is very similar to that of the one-dimensional statistic; *c.f.* Table 6. Power is clearly greater for smaller values of p, which correspond to larger expected proportions of events in the cluster. Smaller scanning squares are more powerful, and the best performance is achieved for $d = 1/16$. As the cluster defined by (3.6) has approximate dimensions 0.05×0.05 and the 'best' window is 0.0625×0.0625, it is apparent that the two-dimensional statistic has greatest power if the scanning square is chosen to be roughly the same size as the area covered by the events that have been produced by the clustering effect. Changing the location of the cluster from the centre of the unit square ($k = 0.5$ in (3.6)) to the top right-hand corner ($k = 0.75$) reduces power for the test when $d = 2/3$ or $1/2$. However, the Scan Statistic appears to be insensitive to cluster location for smaller scanning windows.

3.5 Discussion

The upper and lower bounds to $P(n|N;u,v)$ of Naus (1965b) do not appear to be particularly useful for significance testing. They lack sharpness overall, and the trend in increasing separation of (3.1) and (3.2) with decreasing n, observed for the particular choices of n, N and d employed in Section 3.3, suggests a particular weakness for values of n near the true 5% critical values. The bounds depend on two different one-dimensional tail probabilities, one of which is not amenable to analytic derivation and must, therefore, be simulated, and a second, for which the best approximation method from Chapter 2 does not produce reliable results. These objections suggest that some form of Monte Carlo method will provide a better approach to assessing the significance of the Scan Statistic.

Section 3.4 indicates that the power of the two-dimensional statistic is low for small numbers of events, although it should be noted that the alternative (3.6) is demanding, since a given realisation may contain only a few (or perhaps no) events that are due to the clustering component. The results from both Chapters 2 and 3 support the observation of Wallenstein and Neff (1987), that the power of the Scan Statistic is greatest when the geographical extent of the cluster is matched by the size of the scanning window. This suggests a method of choosing the arbitrary constant, d, before an analysis is undertaken; *i.e.* to select a window that will be of approximately the same magnitude as the supposed cluster. However, as this information will be rarely, if ever, known in practice, this guideline appears to be of limited benefit and applicability. A better approach may be to carry out several analyses using a range of square dimensions, and then to apply a suitable multiple comparisons correction to the results.

3.6 A Problem Encountered with Pseudo-Random Number Generators

3.6.1 Introduction

Simulation of the Scan Statistic in two dimensions requires the production of a large number of artificial spatial point patterns. The basic assumption of the theory associated with the statistic is that events are uniformly distributed under the null hypothesis of no clustering. Therefore, the x and y co-ordinates of each point must be sampled from independent sequences of uniform (0,1] pseudo-random deviates, assuming that the unit square is taken to be the region of interest.

The sampling procedure is normally undertaken by employing a multiplicative linear congruential generator, of the form

$$x_{i+1} = \frac{(aMx_i) \bmod M}{M}, \quad i = 0,1,\ldots, \tag{3.7}$$

with the sequence of random deviates, $\{x_i;\ i = 1, 2, \ldots\}$, being started by the "seed", $x_0 = s/M$, where s is specified by the user. A generator of the form of (3.7) produces only a finite set of distinct values of x_i. This set can contain no more than $(M - 1)$ elements, since

$$x_i = 0 \Rightarrow x_j = 0, \forall j \geq i+1.$$

At some point, the sequence $\{x_i:\ i = 1, 2, \ldots\}$ must reach a repeated value and so will cycle through the same set of values, with a given period. The choices of a and M govern the theoretical maximal period of the generator, which may be less than $(M - 1)$, and whether or not the actual period, p, attains this upper bound. For example, if $M = 2^{\beta} \geq 16$, for some integer β, then (3.7) has a maximal period of only $M/4$, which is attained if and only if $a \bmod 8 = 3$ or 5 (Ripley, 1987, Section 2.2). Familiar versions of (3.7) include the generator DURAND within the I.B.M. Engineering and Scientific Subroutine Library (ESSL) package, in which $a = 7^5$ and $M = 2^{31} - 1$, and G05CAF from the NAg FORTRAN Subroutine Library (NAg, 1990), in which $a = 13^{13}$ and $M = 2^{59}$.

These commercial generators have been designed to achieve long periods, good marginal uniformity and insignificant serial correlation, as described in Ripley (1987, Section 2.2), for instance. Suppose, however, two sequences $\{x_i;\ i = 1, 2, \ldots, N\}$ and $\{y_i;\ i = 1, 2, \ldots, N\}$, corresponding to the x and y co-ordinates of events in the simulated data sets, are generated according to (3.7). For some choices of initial seeds $x_0 = s_1/M$ and $y_0 = s_2/M$, the standard Pearson product moment correlation between $\{x_i\}$ and $\{y_i\}$

is very high. For example, with $N = 100$, $a = 7^5$ and $M = 2^{31} - 1$, the correlations of a set of 1000 sequentially generated pairs of pseudo-random sequences had mean 0.49264 and standard deviation 0.09950, with $s_1 = 5$ and $s_2 = 10$; similarly, for $s_1 = 20$ and $s_2 = 40$, the mean correlation was 0.49122, with a standard deviation of 0.09688. Further experimentation suggested that the magnitude of the average correlation was independent of both N and the number of sequence pairs examined.

These results are wholly unacceptable for an application requiring independent sets of co-ordinates. The following sections identify the source of the problem and the necessary solutions.

3.6.2 Cross-Correlations of Pseudo-Random Deviates

Examination of point patterns $\{(x_i,y_i); i = 1, ..., N\}$, generated from (3.7) with different seeds (x_0,y_0), suggested that the strength of the relationship between $\{x_i\}$ and $\{y_i\}$ depends on the ratio y_0/x_0. Table 13 lists the means and standard deviations of 1000 correlations calculated from simultaneously generated pairs of random deviate streams of length 100, where y_0 was chosen to be an integer multiple of x_0. The particular generator in use was DURAND, *i.e.* $a = 7^5$ and $M = 2^{31} - 1$. As the ratio of the two seeds increases, the strength of the association exhibited declines.

Ratio	s_1	s_2	Mean Correlation	Standard Deviation
2	1	2	0.49585	0.06734
	64	128	0.50106	0.07093
	4096	8192	0.49568	0.07093
3	1	3	0.33227	0.08419
	64	192	0.33543	0.08540
	4096	12288	0.32735	0.08596
4	1	4	0.24514	0.08983
	64	256	0.24821	0.09490
	4096	16384	0.24346	0.08975

Table 13: Means and standard deviations of samples of 1000 cross-correlations between pseudo-random deviate streams of length 100, generated by DURAND, with $x_0 = s_1/M$ and $y_0 = s_2/M$, where $M = 2^{31} - 1$.

Figure 2 plots four examples of point patterns, each of 100 events that were generated with seeds in different proportions, revealing the structure underlying the evidence of dependence provided by the cross-correlations in Table 13. For Figures 2(a) to 2(c), all the observations lie on a small number of parallel lines in the x-y plane. If

$$\frac{y_0}{x_0} = \frac{n_1}{n_2},$$

where n_1 and n_2 are integers with greatest common divisor 1, then the diagrams suggest that there are $(n_1 + n_2 - 1)$ lines, each with gradient n_1/n_2, at vertical intervals of $1/n_2$. Figure 2(d), apparently, has more acceptable behaviour; the ratio y_0/x_0, in this case, is non-integer, whereas (a) to (c) have integral ratios of 2, 3 and 4. However when a greater number of points (*e.g.* 10000) are included, it becomes clear that here, too, there is a structural relationship between the $\{x_i\}$ and $\{y_i\}$ streams (see Figure 3).

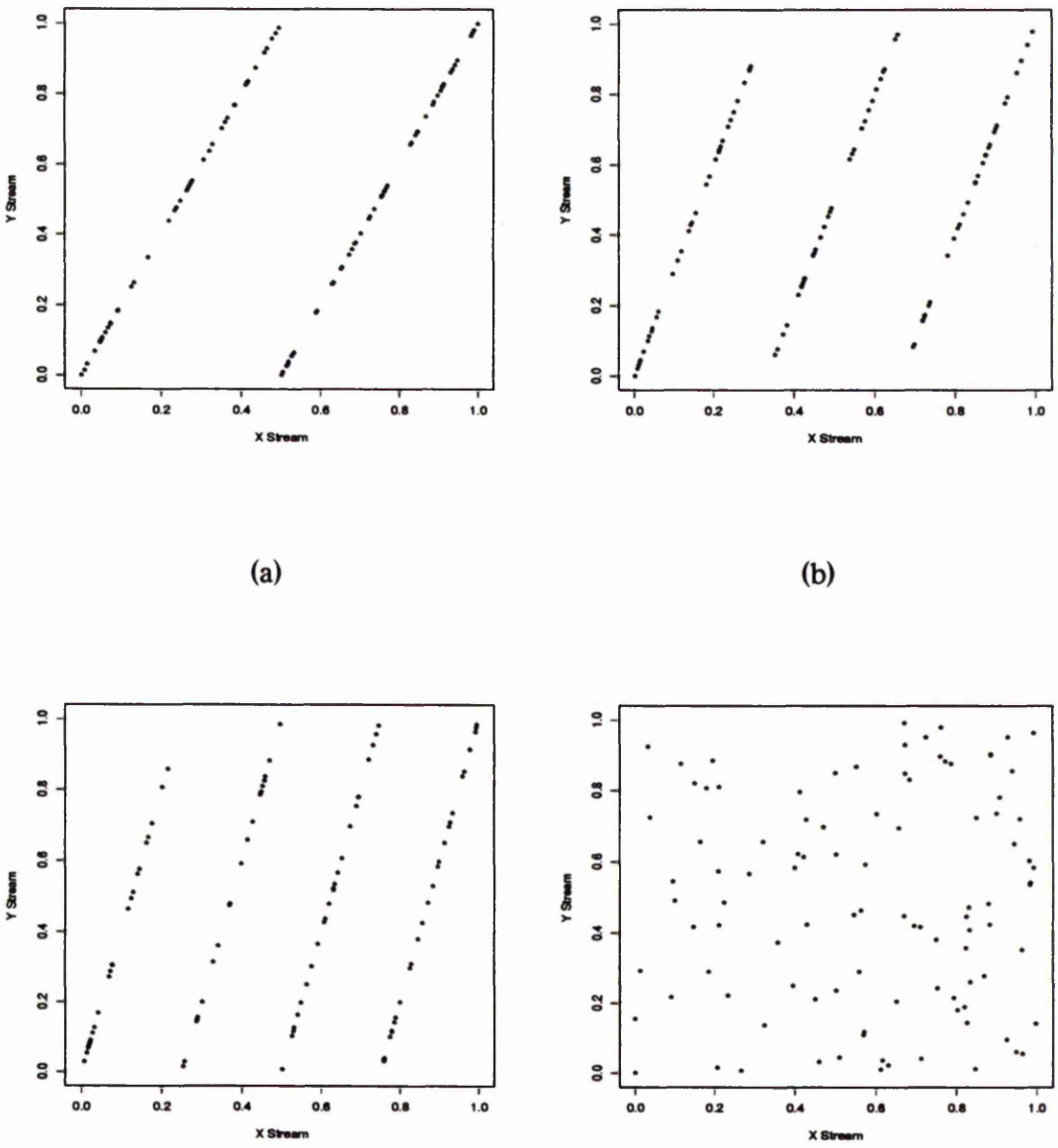


Figure 2: Point patterns of 100 events. X and Y streams generated in parallel from (3.7) with $a = 7^5$ and $M = 2^{31} - 1$, using seeds s_1 and s_2 as follows: (a) 1 and 2; (b) 64 and 192; (c) 4096 and 16384; (d) 13 and 64.

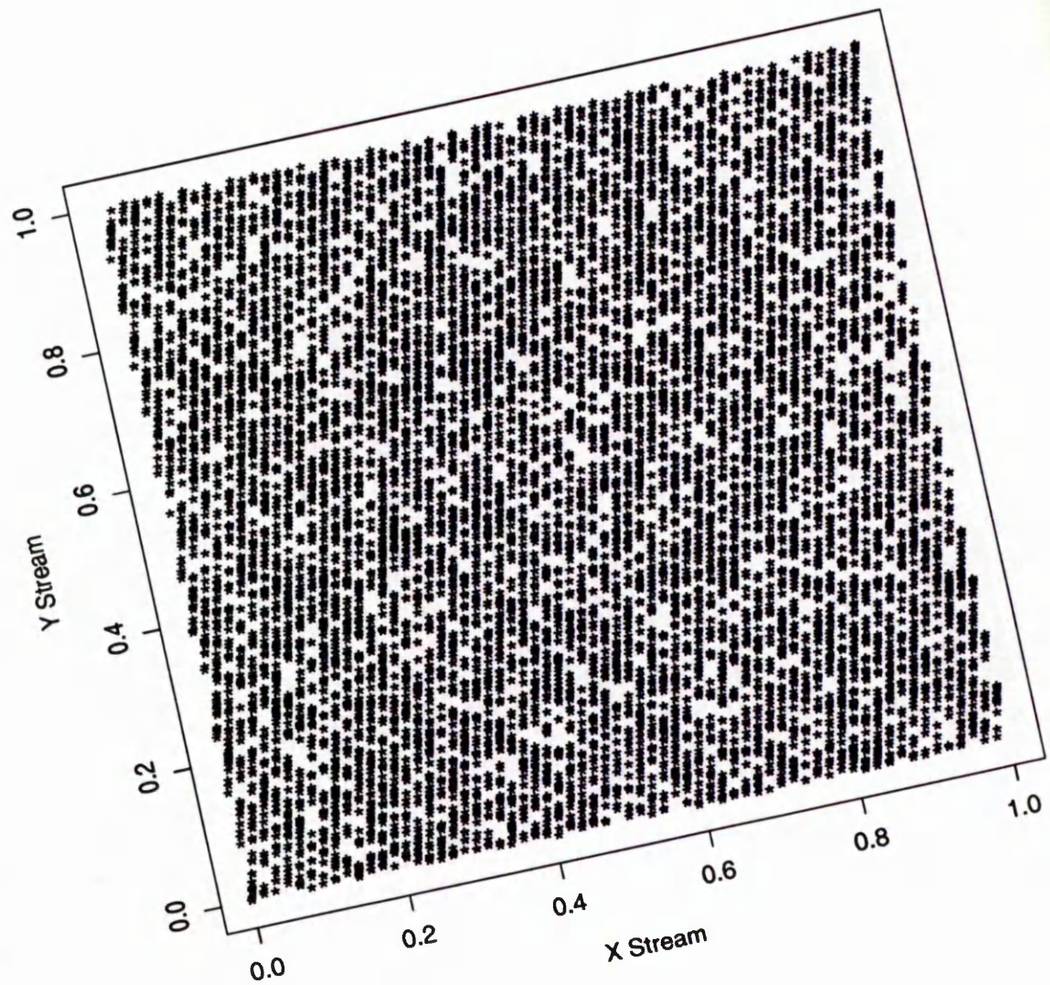


Figure 3: Point pattern of 10 000 events. X and Y streams generated in parallel from (3.7) with $a = 7^5$ and $M = 2^{31} - 1$, using seeds $s_1 = 13$ and $s_2 = 64$.

By the method outlined in Section 3.6.5, it is possible to demonstrate that the following relationship must hold between $\{x_i\}$ and $\{y_i\}$.

Theorem 3.1. Suppose sequences $\{x_i;\ i = 0, 1, ...\}$ and $\{y_i;\ i = 0, 1, ...\}$ are generated according to (3.7) and that

$$y_0 = \frac{n_1}{n_2} x_0, \tag{3.8}$$

where n_1 and n_2 are positive, relatively prime integers. Then, for all $i = 0, 1, ...,$

$$y_i = \frac{n_1}{n_2} x_i + \frac{r_i}{n_2}, \tag{3.9}$$

for *some* integer $r_i \in \{-(n_1 - 1), \ldots, (n_2 - 1)\}$.

This result explains the appearance of sets of parallel lines in Figures 2 and 3. It is also possible to present an approximate argument, based on random variables jointly distributed on a degenerate sample space but marginally uniform, that indicates the form of the association between the choice of seeds, x_0 and y_0, and the cross-correlation coefficient of the sequence $\{(x_i, y_i);\ i = 1, 2, ...\}$, generated by calls to (3.7). This sequence is, of course, deterministic, so it is only strictly valid to consider its full-period correlation coefficient, C, where

$$C = \frac{p\left\{\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)\right\}}{\left\{p\sum x_i^2 - \left(\sum x_i\right)^2\right\}^{\frac{1}{2}}\left\{p\sum y_i^2 - \left(\sum y_i\right)^2\right\}^{\frac{1}{2}}}, \tag{3.10}$$

the summations are for $i = 0$ to $i = p - 1$, and the generator has period p. Exact calculations are often possible for the serial correlation coefficient of a single sequence from a linear congruential generator, in terms of generalised Dedekind sums, as discussed in Section 3.3.3 of Knuth (1981). We failed to construct an analogous exact calculation of C as given by (3.10). As Knuth (1981, p.88) points out, however, such exact computations are indeed difficult, but, if p is large, useful approximations are available by averaging over all real values, rather than the discrete set realised by the generators. The exercises on p.88 of Knuth (1981) show how effective such approximations can be, and Theorem 3.2 represents a further manifestation of the usefulness of this approximate approach.

Theorem 3.2. Suppose (X,Y) are a pair of random variables with probability density function

$$f(x,y)\begin{cases}>0, & (x,y)\in S,\\ =0, & \text{otherwise,}\end{cases} \tag{3.11}$$

where S is the sample space made up of the intersection of the unit square $\{0 \le x \le 1\} \times \{0 \le y \le 1\}$ and the set of parallel lines $\{y = (n_1/n_2)\, x + (r/n_2);\ r = -(n_1 - 1), \ldots, (n_2 - 1)\}$. Suppose also $X \sim U(0,1]$, $Y \sim U(0,1]$ and ϱ is the correlation coefficient between X and Y. Then

$$\varrho = (n_1 n_2)^{-1}.$$

The proof of Theorem 3.2 is also given in Section 3.6.5. The result suggests that the correlation between streams $\{x_i: i = 1, 2, \ldots\}$ and $\{y_i: i = 1, 2, \ldots\}$, calculated from seeds in the ratios 2:1, 3:1 and 4:1, should be approximately 1/2, 1/3 and 1/4, which corresponds closely to the empirical results of Table 13.

3.6.3 Remarks About Existing, Related Literature

The two-stream random number generation process described in Section 3.6.1 can be expressed as a single, matrix generator of the form

$$\mathbf{x}_{i+1} = \mathbf{A}\mathbf{x}_i \ (\mathrm{mod}\, M), \tag{3.12}$$

where $\mathbf{x}_i^T = (x_i, y_i)$, in our previous notation, $\mathbf{A} = a\ \mathbf{I}_2$, where $\mathbf{I}_2$ is the 2 x 2 identity matrix, and $\mathbf{x}_0^T = (x_o, y_o)$ is used as a two-dimensional seed. The generator defined by (3.12), but for more general $\mathbf{A}$, has a substantial literature, some of which identifies potentially worrisome dependences, although none of them is precisely of the form encountered in Section 3.6.2.

Grothe (1987) considers both period length and a method for constructing a maximal period matrix generator. Eichenauer-Herrmann et al (1989) consider the case where the modulus $M = p^{\alpha}$ for p prime and $\alpha \ge 2$. They derive the maximal period length and show how to construct a generator that achieves it. In the review paper of L'Ecuyer (1990), the discussion is written in terms of matrix generators, treating the scalar version simply as a special case.

Afflerbach and Grothe (1988) consider the properties of

$$\mathbf{y}_i^T = \left(\mathbf{x}_i^T, \mathbf{x}_{i+1}^T, \ldots, \mathbf{x}_{i+n-1}^T\right), \quad i = 1,2,3,\ldots,$$

where the $\{\mathbf{x}_i\}$ making up $\mathbf{y}_i$ are a set of n r-dimensional vectors ($r = 2$ in our context). They provide a description of the lattice structure of a matrix generator in $n \times r$ dimensional space. They also comment that this information is relevant to assessing the independence of a sequence of n pseudo-random vectors and not just the distribution of the $\mathbf{y}_i$. A basis for the lattice on which the vectors in the latter situation lie, with r and A chosen to correspond to those of our context, is formed from the $2n$ column vectors of

$$\mathbf{B} = \begin{pmatrix} \mathbf{I}_2 & & & \mathbf{0} & \\ a\mathbf{I}_2 & M\mathbf{I}_2 & & & \\ a^2\mathbf{I}_2 & & M\mathbf{I}_2 & & \\ \vdots & & & \ddots & \\ a^{n-1}\mathbf{I}_2 & \mathbf{0} & & & M\mathbf{I}_2 \end{pmatrix}.$$

However, this does not seem to be the cause of the patterns in our case, which are related to the starting values $\mathbf{x}_0$ when the multiplier is of the form $\mathbf{A} = a\,\mathbf{I}_2$.

Of greater relevance is Durst (1989), who suggests two methods of generating parallel streams of random deviates: either a single sequence, split at random locations, or many different sequences, each using a different multiplier, a. The inference from Durst (1989) is that using the same multiplier for each sequence would not give satisfctory results. In the case of a modulus of the form $M = 2^e$, for some e, Durst (1989) states that two maximal period multipliers, a_1 and a_2, will satisfy

$$a_1 = a_2^j, \text{ for some odd } j, \tag{3.13}$$

and that, if

$$j \equiv 1 \pmod{2^{e-r}}, \quad r > 2, \tag{3.14}$$

then the pairs (a_1^n, a_2^n) lie on at most 2^{r-2} lines. Since $X_n = \{a^n \bmod M\}\ X_0 \pmod M$ for each sequence, a^n is equivalent to X_n in some sense. Therefore, a small value of j in (3.13) would lead to strong dependences. If $a_1 = a_2$ then $j = 1$ in (3.13) and (3.14) is satisfied for all $2 < r < e$ and for $r = 2$ as well, although this latter case is not allowed by the result. If $r = 2$ could be regarded as a pathological case, we have (a_1^n, a_2^n) lying on a single line and, therefore, the worst case of dependence. The results of the previous section suggest that this observation may extend to multipliers other than $M = 2^e$, such as those of the form $M = 2^f - 1$, e.g. DURAND with $f = 31$.

The paper of Fishman and Moore (1982) is also of interest here. They consider the behaviour of, amongst others, the generator with multiplier 7^5 and modulus $2^{31} - 1$ when used to produce non-overlapping t-tuples for $t = 1, 2, 3$. Its overall behaviour is found to be unsatisfactory by testing the uniformity of 100 sequences of t-tuples, each sequence being of length n/t for $t = 1$, 2 or $(n - 2)/t$ for $t = 3$, where $n = 200000$. However, the vectors were all generated from within a single sequence and so the paper does not address the problem of dependence due to seed choice.

3.6.4 Discussion

Theorem 3.1 demonstrates that a very strong structure is inherent in the parallel generation, from different starting points, of two sequences of pseudo-random deviates by any simple multiplicative linear congruential generator. The potential for error is great in any application that may require, or receive accidentally, two or more initialisations, especially when one seed is an integer multiple of the other. The pattern discussed here would clearly carry over to any transformed versions of the two sequences, *e.g.* the generation of deviates from a distribution other than the uniform by an inverse transform method (Rubinstein, 1981).

The problem can, of course, be dissipated easily, in practical terms, by generating the two sequences in series from a single seed. In the case of the generator DURAND, however, the results of Fishman and Moore (1982), described in Section 3.6.3, indicate that the independence of 2- and 3-vectors produced in this way may be suspect. Other solutions include following the recommendations of Durst (1989), regarding the random splitting of a single sequence, or the use of a more sophisticated generator, such as RANMAR or RCARRY, described by James (1990).

A final comment relates to the NAg subroutine G05CAF. Use of this generator is normally preceded by a call to an initialising routine, G05CBF, with the intended seed, s, as an argument. G05CBF carries out a preliminary transformation of s to

$$s' = 2s + 1,$$

so that two sequences generated from seeds s_1 and s_2 would have, effectively, the seeds $2s_1 + 1$ and $2s_2 + 1$. A plot of $\{x_i\}$ versus $\{y_i\}$ would therefore show, for example, 7 "lines" of gradient 5/3, for $s_1 = 1$ and $s_2 = 2$, rather than 2 "lines", each with gradient 2, as might otherwise be expected.

3.6.5 Proofs

Proof of Theorem 3.1

We prove (3.9) by induction. The case $i = 0$ is true by definition, from expression (3.8). Therefore, assume (3.9) is true for $i = k$ and demonstrate that it holds for $i = k + 1$.

First, note that

$$y_{k+1} = \frac{(aMy_k) \bmod M}{M} = \frac{\left(\frac{an_1Mx_k}{n_2} + \frac{aMr_k}{n_2}\right) \bmod M}{M}.$$

(In the following argument, K_1, K_2, K_3, K_4, L_1, L_2, L_3, L_4 and L_5 are integer constants.) Let

$$an_1Mx_k + aMr_k = K_1M + L_1 + aMr_k = K_2M + L_1 \tag{3.15}$$

and let

$$\frac{an_1Mx_k}{n_2} + \frac{aMr_k}{n_2} = K_3M + L_2. \tag{3.16}$$

Then, from (3.16),

$$an_1Mx_k + aMr_k = n_2K_3M + n_2L_2. \tag{3.17}$$

Since $n_2 \in Z$, $n_2L_2 \in Z$ so we can fix an integer p such that

$$pM < n_2L_2 < (p + 1)M$$

for $p \in \{0, ..., (n_2 - 1)\}$ and write $n_2L_2 = pM + L_3$ and therefore, from (3.17),

$$an_1Mx_k + aMr_k = (n_2K_3 + p)M + L_3.$$

Compare this with (3.15) to see that $L_3 = L_1$ and hence

$$L_2 = \frac{L_1 + pM}{n_2}.$$

From (3.16),

$$\left(\frac{an_1Mx_k}{n_2} + \frac{aMr_k}{n_2}\right) \bmod M = L_2 = \frac{L_1 + pM}{n_2}$$

$$= \frac{(an_1Mx_k) \bmod M}{n_2} + \frac{pM}{n_2}. \tag{3.18}$$

From the derivation of (3.15) it is clear that

$$an_1Mx_k = K_1M + L_1. \tag{3.19}$$

Let $aMx_k = K_4M + L_4$. Then

$$an_1Mx_k = n_1K_4M + n_1L_4. \tag{3.20}$$

Now fix q such that

$$qM < n_1L_4 < (q + 1)M,$$

for $q \in \{0, ..., (n_1 - 1)\}$, and write $n_1L_4 = qM + L_5$. Comparing (3.19) with (3.20) suggests that $L_5 = L_1$ and, hence, that

$$L_1 = n_1L_4 - qM,$$

or that

$$(an_1Mx_k) \bmod M = n_1\{(aMx_k) \bmod M\} - qM.$$

Substitute this into (3.18) to obtain

$$y_{k+1} = \frac{n_1}{n_2}\frac{(aMx_k) \bmod M}{M} + \frac{p-q}{n_2}$$

$$= \frac{n_1}{n_2}x_{k+1} + \frac{r_{k+1}}{n_2},$$

where $r_{k+1} = p - q$, with p such that

$$pM < n_2\left\{\left(\frac{an_1Mx_k}{n_2} + \frac{aMr_k}{n_2}\right) \bmod M\right\} < (p+1)M,$$

for $p \in \{0, ..., (n_2 - 1)\}$, and q such that

$$qM < n_1\{(aMx_k) \bmod M\} < (q + 1)M,$$

for $q \in \{0, ..., (n_1 - 1)\}$.

Proof of Theorem 3.2

Let X and Y be two uniform random deviates, related by the expression

$$Y = \frac{n_1}{n_2} X + \frac{R}{n_2}, \tag{3.21}$$

for $R \in \{-(n_1 - 1), ..., (n_2 - 1)\}$. Then

$$XY = \frac{n_1}{n_2} X^2 + \frac{R}{n_2} X$$

and so

$$\mathbb{E}(XY|X) = \frac{n_1}{n_2} X^2 + \frac{X}{n_2} \mathbb{E}(R|X). \tag{3.22}$$

Further progress requires information about the distribution of R, conditional on X. Asymptotically, this distribution is a discrete uniform distribution, as can be justified by considering the (joint) distribution of X and Y. The joint sample space, S, for X and Y consists of a set of parallel line-segments in the unit square, as shown for example in Figure 4 for the case $n_1 = 3$ and $n_2 = 5$. Thus the joint density of X and Y is degenerate, concentrated on the above degenerate (in a two-dimensional sense) sample space. Suppose we consider an arbitrary subset of S, of length $\delta > 0$. It is clear that, in order that both X and Y have marginal uniform distributions, the probability of (X,Y) falling in that subset must be proportional to δ. In other words, (X,Y) are jointly uniformly distributed on their sample space. (The $f(x,y)$ defined in (3.11) is constant on S.) From this it follows that, given X, R (which is then equivalent to Y) is uniformly distributed on its finite sample space.

To follow up the implications of this for $\mathbb{E}(R|X)$ it helps again to refer to the exemplar provided by Figure 4. For example, at A the values of R which are available are 0, 1, 2, ..., $(n_2 - 1)$ and so

$$\mathbb{E}(R|X = A) = (n_2 - 1)/2.$$

Similarly, for $X = B$ the range of R is -1, 0, 1, ..., $(n_2 - 2)$, which implies

$$\mathbb{E}(R|X = B) = \{(n_2 - 1)/2\} - 1.$$

Hence, in general,

$$\mathbb{E}(R|X) = \{(n_2 - 1)/2\} - i, \quad \text{for } i/n_1 < X \le (i+1)/n_1,$$

for $i = 0, 1, ..., (n_1 - 1)$. Using (3.22), we have

$$\mathbb{E}(XY) = \mathbb{E}\{\mathbb{E}(XY|X)\}$$

$$= \frac{n_1}{n_2}\mathbb{E}(X^2) + \frac{1}{n_2}\int_0^1 x\,\mathbb{E}(R|X)\,dx$$

$$= \frac{n_1}{3n_2} + \frac{1}{n_2}\left[\frac{n_2-1}{2}\int_0^1 x\,dx - \sum_{i=0}^{n_1-1}\int_{i/n_1}^{(i+1)/n_1} ix\,dx\right]$$

$$= \frac{n_1}{3n_2} + \frac{1}{n_2}\left[\frac{n_2-1}{4} - \frac{1}{n_1^2}\sum_{i=1}^{n_1-1}\frac{i(2i+1)}{2}\right]$$

$$= \frac{n_1}{3n_2} + \frac{1}{4} - \frac{1}{4n_2} - \frac{1}{2n_1^2 n_2}\left[\frac{2n_1(n_1-1)(2n_1-1)+3n_1(n_1-1)}{6}\right]$$

$$= \frac{1}{4} + \frac{1}{12n_1n_2}.$$

Hence

$$\rho = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

$$= \frac{\frac{1}{4} + \frac{1}{12n_1n_2} - \frac{1}{2}\cdot\frac{1}{2}}{\sqrt{\frac{1}{12}\cdot\frac{1}{12}}}$$

$$= (n_1n_2)^{-1},$$

since X and Y are marginally uniform (0,1].

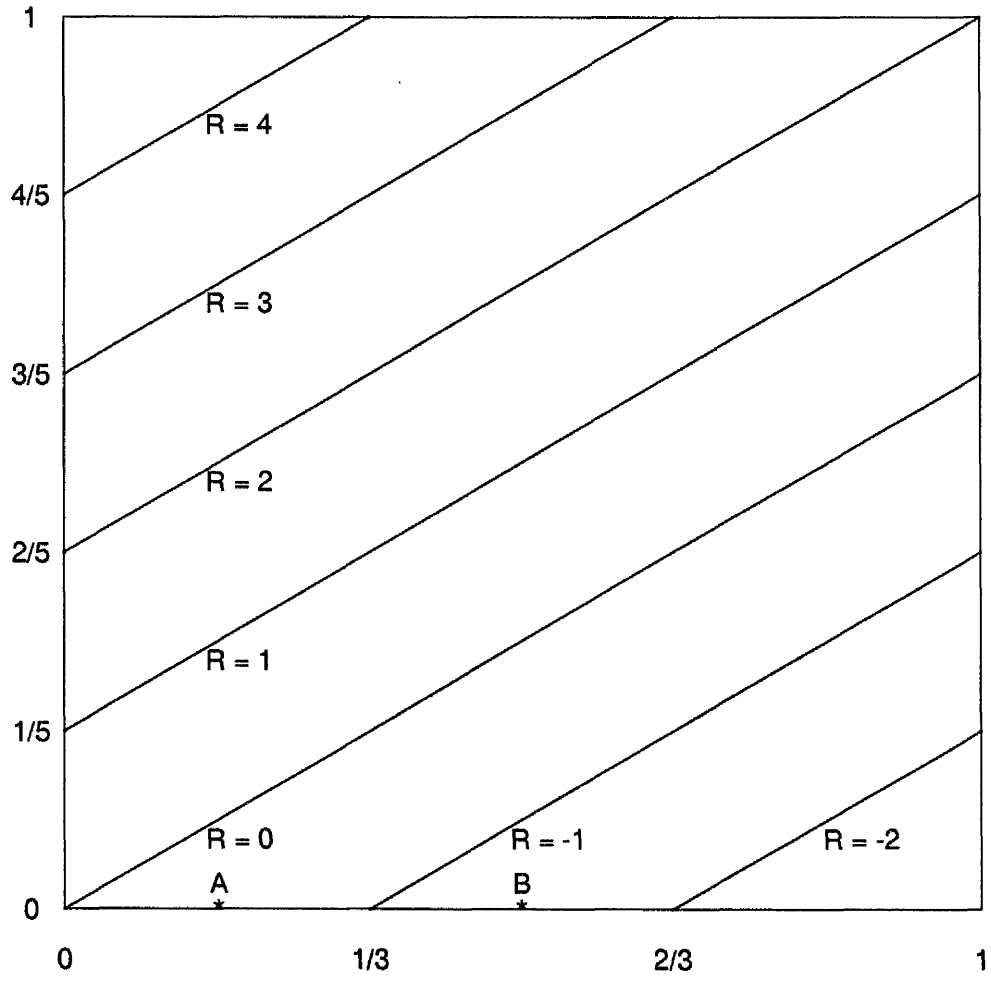


Figure 4: Diagram of the joint sample space for uniform random variables X and Y, which are related by expression (3.21), representing parallel pseudo-random deviate generation with $n_1 = 3$ and $n_2 = 5$.

CHAPTER 4

CORRECTING FOR A NON-UNIFORM NULL DISTRIBUTION

4.1 Introduction

The previous two chapters assumed that, in the absence of any clustering effect in the region of interest, the distribution of the type of events under consideration would be uniform. In practice, however, geographical features, natural variation or some other factor may ensure that the distribution under the null hypothesis is markedly non-uniform (Bithell, 1990; Diggle, 1990). For example, human populations aggregate into high density (urban) areas surrounded by low density (rural) ones. Thus, it would be reasonable to expect that cases of a particular disease would occur more frequently in the former regions than in the latter, and that a map of cases displayed without reference to the population baseline would exhibit apparently significant clustering in the areas of high population density. This type of clustering is not of interest, in general; it is necessary, therefore, to correct for this effect, and then to search for aggregation that may have been caused by a factor with more aetiological relevance.

A method of allowing for the type of problem described above in analyses employing the one-dimensional Scan Statistic was described by Weinstock (1981). Instead of using a scanning interval of a given magnitude, d, the technique allows the length of the interval to depend on the null probability density function of events. If the region of interest is the interval $[0, A]$, in which N events are distributed with p.d.f. $f(\cdot)$, then the length of the scanning interval with right hand boundary at $x \in [0, A]$ is denoted by δ, where δ satisfies

$$\int_{x-\delta}^{x} f(t)\,dt = k, \tag{4.1}$$

and k is a pre-specified constant, which takes the place of the constant d in the standard analysis. If $[0, A]$ represents time, and the events are cases of a human disease, k has the interpretation of the (constant) number of person-years at risk occurring inside the (variable length) scanning interval.

This correction ensures that the expected number of cases in the window under the null hypothesis remains the same for all locations. Thus, if $k = d/A$, say, and the Scan Statistic calculated using this procedure is n, then the corresponding tail probability is obtained from the theory for the original statistic, as discussed previously; *i.e.*

$$\Pr(\text{Scan Statistic} \geq n \mid N, k = d/A) = P(n|N;d),$$

where $P(n|N;d)$ is the quantity defined in Section 2.2. If $f(\cdot)$ in (4.1) is the uniform p.d.f., $\delta = k$, $\forall x$, and we recover the method of Chapter 2.

The true p.d.f., $f(\cdot)$, will not usually be known, so that practical implementation of (4.1) will require its replacement by some estimate $\hat{f}(\cdot)$. The use of a kernel density estimate (Silverman, 1986), which could be calculated from a sample of controls representing the population at risk, to replace $f(\cdot)$ in (4.1) is investigated in Section 4.2, as well as ways of simplifying the computational process. Section 4.3 generalises the method to two dimensions and Section 4.4 discusses the results obtained and suggests a simpler and more robust implementation of the technique.

4.2 Investigation in One Dimension

The aim of this section is to present a method of calculating δ from (4.1) that is straightforward to employ in practice. For clarity, it is presented first for the one-dimensional case, then generalised to two dimensions in the next section, although the results of this section may be useful for certain applications in their own right, such as the detection of clustering in time.

4.2.1 Computational Considerations

Assuming for the moment that the null distribution, $f(\cdot)$, is known, the first step in solving (4.1) is to approximate the integral by some simple numerical rule, such as Simpson's Rule or the Trapezoidal Method. Table 14 compares the accuracy of the two techniques for small numbers of knots at a range of coordinates within the unit interval. Each coordinate, x, forms the upper boundary of a scanning interval, the magnitude of which is estimated by solving

$$\int_{x-\delta}^{x} f(t)\,dt = 0.25 \qquad (4.2)$$

for δ, using the Bisection Method (Press *et al*, 1989), where $f(\cdot)$ is taken to be the p.d.f. of the Be(3,3) distribution. Subsequently, values of the integral

$$\int_{x-\delta}^{x} f(t)\,dt$$

are approximated using each of the numerical methods and the interval sizes, δ, obtained from (4.2). These are listed in Table 14. The best balance between simplicity and accuracy is achieved by Simpson's Rule with three knots, leading to the approximation

$$\int_{x-\delta}^{x} f(t)\,dt \approx \frac{\delta}{6}\{f(x-\delta)+4f(x-\tfrac{1}{2}\delta)+f(x)\}, \tag{4.3}$$

with the magnitude of the associated error term being

$$|E| \le \delta^5 \frac{M}{2880}, \quad \text{where } M = \max_{t\in(x-\delta,x)} \left|f^{(4)}(t)\right|.$$

Coordinate	Trapezoidal Rule			Simpson's Rule	
(x)	2 knots	3 knots	5 knots	3 knots	5 knots
0.3835	0.259803	0.252201	0.250537	0.249897	0.249993
0.5194	0.256813	0.251587	0.250390	0.249898	0.249999
0.5328	0.256702	0.251564	0.250385	0.249988	0.249999
0.6789	0.257539	0.251751	0.250430	0.249982	0.249999
0.7556	0.259137	0.252082	0.250509	0.249961	0.249997

Table 14: Comparison of different numerical integration rules over a range of coordinates in the unit interval. Integral is area under Be(3,3) p.d.f. bounded by a scanning window with length estimated from (4.2) and right-hand boundary at x.

Rather than evaluating the function, $f(\cdot)$, at three different coordinates, the second step in this implementation of (4.1) is to replace $f(x - \delta)$ and $f(x - \tfrac{1}{2}\delta)$ in (4.3) by their Taylor expansions. By discarding terms after the first derivative, (4.1) may be approximated by the quadratic equation in δ

$$-\tfrac{1}{2}f'(x)\delta^2 + f(x)\delta - k = 0, \tag{4.4}$$

or by the cubic

$$\tfrac{1}{6}f''(x)\delta^3 - \tfrac{1}{2}f'(x)\delta^2 + f(x)\delta - k = 0, \tag{4.5}$$

if terms after the second derivative are neglected. The remainder terms are of the form $c_1\ \delta^2$ for the two first order expansions that lead to (4.4), and $c_2\ \delta^3$ for the second order expansions. Hence, the accuracy of each approximating step is proportional to a power of δ.

Since a polynomial of degree p will have p zeros, some of which may be complex, an heuristic rule for choosing the correct solution of (4.4) or (4.5) is to let δ equal the smallest of the real zeros within the allowable range, $[0, x]$, at coordinate x. We are not aware of any rigorous justification for this observation. If x is small enough, so that

$$\int_0^x f(t)\,dt < k,$$

then no solution of (4.4) or (4.5) will be permissible. However, any length of scanning interval greater than x can in fact be used safely at the lower edge of $[0, A]$, since the count of cases falling inside it will remain the same, whatever value is used for δ.

The null distribution of events is again assumed to be Be(3,3), so that $A = 1$, for Table 15, which compares interval sizes calculated from (4.4) and (4.5) with an 'exact' solution, found by the Bisection Method, at four coordinates, x, for five different choices of $k = d$. Expression (4.5) is reasonably accurate for the range of coordinates and choices of d examined, while (4.4) is less effective, particularly for larger d, although its accuracy is still adequate. Both approximations are weak in the upper tail of the distribution, where values of the p.d.f. are small and, thus, numerical instability in (4.4) and (4.5) is more likely. In a repeated analysis, with $f(\cdot)$ chosen to be Be(2,2), accuracy in the tails improved for both expressions. The Be(2,2) distribution is 'flatter' and 'wider' than Be(3,3), so values of the former p.d.f. tend to be larger than those of the latter at the same coordinates in the tails of the densities.

x	Interval From	d 1/4	1/8	1/16	1/32	1/64
0.3943	(4.1)	0.194639	0.079900	0.037931	0.018583	0.009208
	(4.4)	0.172394	0.078500	0.037787	0.018567	0.009206
	(4.5)	0.205260	0.080008	0.037935	0.018584	0.009208
0.5971	(4.1)	0.135901	0.069020	0.035114	0.017758	0.008936
	(4.4)	0.130285	0.068235	0.035007	0.017744	0.008935
	(4.5)	0.135519	0.068990	0.035112	0.017758	0.008936
0.6735	(4.1)	0.146698	0.077772	0.040632	0.020873	0.010595
	(4.4)	0.140967	0.076843	0.040493	0.020854	0.010593
	(4.5)	0.145653	0.077673	0.040623	0.020872	0.010595
0.9171	(4.1)	0.279656	0.190731	0.129036	0.085071	0.053892
	(4.4)	0.319723	0.214730	0.141287	0.090429	0.055849
	(4.5)	0.251269	0.180291	0.125507	0.084044	0.053650

Table 15: Scanning interval sizes at coordinates x, calculated from (4.1), with $k = d$, and the two approximate polynomials, (4.4) and (4.5); $f(\cdot)$ is assumed to be Be(3,3).

4.2.2 Introduction of a Kernel Estimator

Naturally, in real applications of the Scan Statistic, the true distribution of events under a null hypothesis of no clustering will be unknown. Therefore, implementation of a correction based on (4.1) will require the replacement of $f(\cdot)$ by some estimate $\hat{f}(\cdot)$. If a control sample from the population from which events are assumed to derive is available,

$$\{Y_1,\ldots,Y_{N_c}\},$$

say, then a suitable estimate of $f(\cdot)$ may be calculated using non-parametric kernel density estimation. This technique was previously discussed in Section 2.3 and is studied in detail by Silverman (1986). Given a kernel function, $K(\cdot)$, and a smoothing parameter, h, $f(\cdot)$ is estimated from the controls by

$$\hat{f}(x) = \frac{1}{N_c h}\sum_{i=1}^{N_c} K\left(\frac{x - Y_i}{h}\right).$$

This is a particularly convenient estimator for our purposes, since its mth derivative is readily available, *viz*

$$\hat{f}^{(m)}(x) = \frac{1}{N_c h^{m+1}} \sum_{i=1}^{N_c} K^{(m)}\left(\frac{x - Y_i}{h}\right),$$

providing $K(\cdot)$ is sufficiently differentiable, and where a superscript "(m)" indicates an mth order derivative. Therefore, the estimate, $\hat{f}(\cdot)$, and all the derivatives required by (4.4) or (4.5) can be calculated on a single pass through the data.

Table 16 provides an example of the procedure, by comparing simulated 5% critical values obtained for the standard Scan Statistic of Chapter 2 to values obtained for a statistic applying (4.4), using control samples of size 50, 100 or 200. The observations (events and controls) were sampled, under a null hypothesis of no clustering, from the mixture distribution

$$¼\ \mathrm{U}(0,1] + ¾\ \mathrm{Be}(3,3), \tag{4.6}$$

which has greater probability mass in the tails than the Be(3,3) distribution used previously, making the estimation of interval length easier and more robust. Since the tails are important in the calculation, the null density was estimated using an adaptive kernel method (Silverman, 1986, pp. 100 - 102). In this approach, a rough pilot estimate is computed initially, to obtain information about the locations of high and low density areas. A second estimate uses the pilot's guidance to add more weight to low density regions. A sensitivity parameter, α, controls the influence of the pilot on the final estimate; Table 16 was calculated with $\alpha = ½$.

Both the initial and final estimates used Gaussian kernels and the smoothing parameter

$$h = 0.9 A N_c^{-\frac{1}{5}}, \tag{4.7}$$

where

$$A = \min(s, r/1.34),$$

s is the sample standard deviation and r the interquartile range of $\{Y_1,\ldots,Y_{N_c}\}$. The bandwidth (4.7) was suggested as a widely applicable smoothing parameter for use with a Gaussian kernel, by Silverman (1986, pp. 47 - 48), to increase robustness to skew or bimodal distributions. It is obtained by modifying the bandwidth that minimises approximate mean integrated square error in an intuitively sensible way. The results in Table 16 are based on 1000 replications.

N	d	Uniform Null	Mixture Null Distribution, (4.6) Numbers of Controls		
			50	100	200
20	1/4	12	16	13	14
	1/8	9	11	10	9
	1/16	7	8	7	7
	1/32	5	6	6	5
	1/64	5	5	5	5
50	1/4	23	31	30	30
	1/8	15	22	20	18
	1/16	11	15	13	12
	1/32	8	10	9	8
	1/64	6	8	7	7
100	1/4	39	53	58	51
	1/8	25	33	32	27
	1/16	17	20	18	18
	1/32	12	13	12	12
	1/64	9	10	9	9
150	1/4	55	105	82	74
	1/8	34	56	57	39
	1/16	22	31	29	23
	1/32	15	19	18	15
	1/64	11	13	13	11

Table 16: Simulated 5% critical values for the Scan Statistic, assuming a null distribution that is either uniform or of form (4.6), but using correction (4.4) with an adaptive kernel method in the latter case. Based on 1000 replications.

The agreement beween the empirical critical values for a uniform null distribution and those based on (4.6) improves as the number of controls increases, since a larger N_c suggests an improved estimate of $f(\cdot)$. However, the accuracy of the correction for $d = 1/4, 1/8$ is still relatively poor, even for $N_c = 200$. With smaller values of d, both the numerical integrations and the Taylor expansions will be more accurate. As previously indicated, the error of each approximation depends on δ^l, for various positive integers, l, and a small d implies a correspondingly small value of δ.

To investigate the source of error in Table 16 in more detail, the same type of simulation study was carried out again, first by assuming that the null distribution of events was known and equal to the mixture distribution (4.6), and secondly, by generating control samples from U(0,1]. The former analysis was included to assess the contribution of the kernel density estimation procedure to inaccuracies in the correction method, and the latter to assess whether or not the shape of the null distribution was actually affecting the empirical critical values obtained, contrary to theoretical expectation. Table 17 compares the results of the above two simulations with the uniform null distribution results of Chapter 2, with each entry being calculated from the empirical distribution of 1000 replications. Use of an exact null distribution allowed $f(x)$ and $f'(x)$ to be evaluated precisely for the solution of (4.4) for δ, whereas the second analysis described above required the calculation of kernel density estimates from samples of 100 controls, using the adaptive procedure referred to earlier, with a bandwidth, h, calculated from (4.7). A modification included to improve the estimation of the uniform distribution in the latter case was to employ wrap-around edge conditions, *i.e.* to add the probability mass of the estimate that fell below the lower boundary of [0,1], due to individual kernels overlapping the edge of the interval, to the upper tail of the density estimate, and *vice versa*. This was accomplished by using estimates of the form

$$\hat{f}_W(x) = \frac{1}{N_c h} \sum_{i=1}^{N_c} \left\{ K\left(\frac{x-1-Y_i}{h}\right) + K\left(\frac{x-Y_i}{h}\right) + K\left(\frac{x+1-Y_i}{h}\right) \right\}. \qquad (4.8)$$

It should be noted that, although (4.8) is apparently of the wrong form for the estimator of a p.d.f. (since it will integrate to the value three on $(-\infty,\infty)$), we are interested only in an estimate of a density for the interval [0,1], on which the ordinary kernel estimator, $\hat{f}(\cdot)$, will integrate to a value usually less than one. On the same domain, $\hat{f}_W(\cdot)$ will better approximate a probability density function in terms of its integral. However, this may still fail to achieve a value of one, particularly if the kernel chosen has infinite support, as the following, simple example demonstrates.

Example

With only $N_c = 2$ data points, $Y_1 = 0.1$ and $Y_2 = 0.9$, Gaussian kernels and a smoothing parameter of $h = 1$,

$$\hat{f}(x) = \frac{1}{2\sqrt{2\pi}} \sum_{i=1}^{2} \exp\left\{-\tfrac{1}{2}(x - Y_i)^2\right\}$$

and

$$\hat{f}_W(x) = \frac{1}{2\sqrt{2\pi}} \sum_{i=1}^{2} \left[\exp\left[-\tfrac{1}{2}\{x - (Y_i + 1)\}^2\right] + \exp\left\{-\tfrac{1}{2}(x - Y_i)^2\right\} + \exp\left[-\tfrac{1}{2}\{x - (Y_i - 1)\}^2\right] \right].$$

Then,

$$\int_{-\infty}^{\infty} \hat{f}(x)\,dx = 1,$$

clearly, but

$$\int_0^1 \hat{f}(x)\,dx = 0.3557$$

and

$$\int_0^1 \hat{f}_W(x)\,dx = 0.8356.$$

The substitution of an exact distribution in place of the k.d.e. does not improve the accuracy of the empirical critical values. In fact, for $d \leq 1/8$, accuracy usually decreases, while for small d, the uniform and exact results are very similar. This suggests that something other than the kernel estimation procedure is mainly responsible for most of the error in Table 16, although, as that table demonstrates, improving the precision of the estimate clearly does improve accuracy. This latter observation also indicates that the correction method will be sensitive to the value of the smoothing parameter, h, which also affects the accuracy of $\hat{f}(\cdot)$. The second set of results in Table 17 is very similar to that obtained previously for 200 controls; therefore, the shape (or non-uniformity) of the null distribution does not explain the magnitude of the inaccuracy for large d. Hence, it seems likely that the main constraint on the implementation of (4.1) described in this section is the precision of the numerical integration rule, (4.3), and Taylor expansions, (4.4) or (4.5), rather than any other factor.

N	*d*	Uniform Null	Exact Mixture Null, (4.6)	Sample from U(0,1] + (4.8)
20	1/4	12	17	13
	1/8	9	9	9
	1/16	7	6	7
	1/32	5	5	5
	1/64	5	4	5
50	1/4	23	44	28
	1/8	15	19	17
	1/16	11	11	11
	1/32	8	8	8
	1/64	6	6	6
100	1/4	39	95	53
	1/8	25	45	29
	1/16	17	17	18
	1/32	12	11	12
	1/64	9	8	9
150	1/4	55	148	78
	1/8	34	71	40
	1/16	22	24	24
	1/32	15	15	16
	1/64	11	10	11

Table 17: Simulated 5% critical values for the Scan Statistic, comparing uniform null results to those calculated using correction (4.4), assuming the exact distribution (4.6), or using an adaptive k.d.e., calculated from 100 observations drawn from U(0,1]. Based on 1000 replications.

4.3 Generalisation to Two Dimensions

Working by analogy, it is possible to extend the methods of Section 4.2 to two dimensions. The region under consideration is taken to be $[0, A] \times [0, B]$, defined on which is a two-dimensional p.d.f. representing the distribution of events under a null hypothesis of no clustering. In Chapter 3, this density was the two-dimensional uniform p.d.f.

$$f(x,y) = \begin{cases} (AB)^{-1}; & 0 \le x \le A,\ 0 \le y \le B, \\ 0; & \text{elsewhere,} \end{cases} \tag{4.9}$$

with $A = B = 1$. A square, with sides of constant length d, scanning over $[0, A] \times [0, B]$ in a continuous fashion would always bound a fixed volume under (4.9) of d^2/AB. Therefore, if $f(\cdot,\cdot)$ is some non-uniform p.d.f., then the length of the sides of the scanning square with top right-hand corner at coordinates (x,y) is δ, where δ is chosen to satisfy

$$\int_{x-\delta}^{x}\int_{y-\delta}^{y} f(s,t)\,ds\,dt - \frac{d^2}{AB} = 0. \tag{4.10}$$

4.3.1 Numerical Procedures

To reduce (4.10) to a simpler equation for δ, it is possible to use a numerical integration rule and Taylor expansions of $f(\cdot,\cdot)$ at coordinates other than (x,y), by applying them to each coordinate direction in turn. To demonstrate this, the following example uses Simpson's Rule with 3 knots and expands terms of the form $f(x - \delta, y)$ and $f(x - \tfrac{1}{2}\delta, y)$ to third order.

Example Let $I = \int_{x-\delta}^{x}\int_{y-\delta}^{y} f(s,t)\,ds\,dt$.

$$\begin{aligned}
\int_{y-\delta}^{y} f(s,t)\,dt &\approx \frac{\delta}{6}\left\{f(s,y-\delta) + 4f(s,y-\tfrac{1}{2}\delta) + f(s,y)\right\} \\
&\approx \frac{\delta}{6}\Big[f(s,y) - \delta f_t(s,y) + \tfrac{1}{2}\delta^2 f_{tt}(s,y) - \tfrac{1}{6}\delta^3 f_{ttt}(s,y) \\
&\quad + 4\left\{f(s,y) - \tfrac{1}{2}\delta f_t(s,y) + \tfrac{1}{8}\delta^2 f_{tt}(s,y) - \tfrac{1}{48}\delta^3 f_{ttt}(s,y)\right\} \\
&\quad + f(s,y)\Big] \\
&= f(s,y)\delta - \tfrac{1}{2} f_t(s,y)\delta^2 + \tfrac{1}{6} f_{tt}(s,y)\delta^3 - \tfrac{1}{24} f_{ttt}(s,y)\delta^4 \\
&= g(s,y), \text{ say,}
\end{aligned}$$

where a subscript s or t represents a partial derivative with respect to the first or second coordinate direction, as appropriate, and multiple subscripts represent higher order derivatives. Then,

$$\begin{aligned} I &\approx \int_{x-\delta}^{x} g(s,y)\,ds \\ &\approx g(x,y)\delta - \tfrac{1}{2}g_s(x,y)\delta^2 + \tfrac{1}{6}g_{ss}(x,y)\delta^3 - \tfrac{1}{24}g_{sss}(x,y)\delta^4 \\ &= I_{\text{approx}}, \end{aligned}$$

where

$$\begin{aligned} I_{\text{approx}} = {} & f(x,y)\delta^2 - \tfrac{1}{2}\{f_s(x,y)+f_t(x,y)\}\delta^3 + \{\tfrac{1}{6}f_{ss}(x,y)+\tfrac{1}{4}f_{st}(x,y)+\tfrac{1}{6}f_{tt}(x,y)\}\delta^4 \\ & -\{\tfrac{1}{24}f_{sss}(x,y)+\tfrac{1}{12}f_{sst}(x,y)+\tfrac{1}{12}f_{stt}(x,y)+\tfrac{1}{24}f_{ttt}(x,y)\}\delta^5 \\ & +\{\tfrac{1}{48}f_{ssst}(x,y)+\tfrac{1}{36}f_{sstt}(x,y)+\tfrac{1}{48}f_{sttt}(x,y)\}\delta^6 \\ & -\tfrac{1}{144}\{f_{ssstt}(x,y)+f_{ssttt}(x,y)\}\delta^7 + \tfrac{1}{576}f_{sssttt}(x,y)\delta^8. \end{aligned} \tag{4.11}$$

By discarding terms of the appropriate order in the argument above, it is clear that the equations for δ that approximate (4.10), and that are based on first, second and third order Taylor expansions, are respectively

$$\tfrac{1}{4}f_{st}(x,y)\delta^4 - \tfrac{1}{2}\{f_s(x,y)+f_t(x,y)\}\delta^3 + f(x,y)\delta^2 - \frac{d^2}{AB} = 0, \tag{4.12}$$

$$\begin{aligned} & \tfrac{1}{36}f_{sstt}(x,y)\delta^6 - \tfrac{1}{12}\{f_{sst}(x,y)+f_{stt}(x,y)\}\delta^5 + \{\tfrac{1}{6}f_{ss}(x,y)+\tfrac{1}{4}f_{st}(x,y)+\tfrac{1}{6}f_{tt}(x,y)\}\delta^4 \\ & -\tfrac{1}{2}\{f_s(x,y)+f_t(x,y)\}\delta^3 + f(x,y)\delta^2 - \frac{d^2}{AB} = 0 \end{aligned} \tag{4.13}$$

and

$$I_{\text{approx}} - \frac{d^2}{AB} = 0, \tag{4.14}$$

where I_{approx} is as defined in (4.11).

The zeros of these polynomials in δ may be found numerically to a good degree of accuracy by, for example, the routine C02AGF from the NAg FORTRAN Subroutine Library (NAg, 1990). The algorithm used by this procedure was proposed by Smith (1967) as a modification of the iterative scheme called Laguerre's Method.

If we assume that $f(\cdot,\cdot)$ is known and integrable, it is possible to compare (4.12), (4.13) and (4.14) with the exact result (4.10), which may be expressed as a polynomial equation for δ by evaluating the double integral analytically. If $f(\cdot,\cdot)$ is taken to be the two dimensional equivalent of (4.6), *i.e.*

$$f(x,y) = \tfrac{1}{4}\, u(x,y) + \tfrac{3}{4}\, b(x,y), \tag{4.15}$$

where $u(x,y)$ is the p.d.f. of the two-dimensional uniform distribution on the unit square and $b(x,y)$ is the p.d.f. with coordinates that are independently Be(3,3), then (4.10) is a polynomial of degree 10, the zeros of which may be found by a numerical procedure, as discussed above. The same heuristic rule that was applied in Section 4.2, regarding the selection of the smallest real, positive zero, may be used here.

N	Method	*d*				
		1/4	1/8	1/16	1/32	1/64
20	(4.10)	7	5	4	3	3
	(4.12)	8	5	4	3	3
	(4.13)	9	5	4	3	3
	(4.14)	8	5	4	3	3
50	(4.10)	12	7	5	4	3
	(4.12)	18	7	5	4	3
	(4.13)	15	8	5	4	3
	(4.14)	11	7	5	4	3
100	(4.10)	19	10	6	5	3
	(4.12)	34	10	6	4	4
	(4.13)	26	12	6	4	4
	(4.14)	18	10	6	4	4
150	(4.10)	23	12	7	5	4
	(4.12)	52	14	7	5	4
	(4.13)	36	17	7	5	4
	(4.14)	24	11	7	5	4

Table 18: Simulated 5% Scan Statistic critical values (100 replications) with the exact null distribution (4.15) and the exact, (4.10), or approximate, (4.12) to (4.14), correction methods.

A simulation study was carried out to compare the accuracy of (4.12), (4.13) and (4.14) to (4.10). For different choices of N and d, events were drawn from the density (4.15), and, throughout, the null distribution was assumed to be of this form also. For each of 100 replications, the Scan Statistic was calculated, using either the exact correction or one of the three above approximations; given (4.15), it is straightforward to calculate the partial derivatives required by the latter methods. Empirical 5% critical values derived from these simulations are displayed in Table 18.

As would be expected, the results for (4.10) correspond almost exactly to the simulated critical values for the original two-dimensional Scan Statistic; *cf* Table 8 in Chapter 3, which is based on 1000 replications. The differences, for $N = 100$, are minor and attributable to sampling variability and the smaller number of replications. The accuracy of the approximations is very good for small d, and for the whole range of constants for (4.14), but (4.12) and (4.13) are much poorer for $d = 1/4, 1/8$. As in the one-dimensional case, this may be explained by the role of δ in the error and remainder terms of Simpson's Rule and the Taylor expansion.

However, a weakness of the approximations is that, at some locations in the unit square, the resulting polynomials may have no real zeros, or the smallest real and positive zero may be larger than unity. This was also observed in Section 4.2. Near the boundary, the problem may arise because there is, in fact, no solution to (4.10); that is, for (x,y) near the left-hand and lower edges of $[0,1] \times [0,1]$, the largest possible δ that ensures the scanning square remains completely contained within that region, *i.e.* $\delta = \min(x,y)$, is such that

$$\int_{x-\delta}^{x} \int_{y-\delta}^{y} f(s,t)\, ds\, dt < d^2.$$

Away from the boundary, the approximation to the exact equation for δ may simply break down. In the simulations reported above, and later in this section, the occurrence of either error was flagged. If all zeros returned by the numerical root finding procedure were complex, a value of zero was returned for δ; if δ was real but larger than one, the value was left unchanged. Table 19 contains some examples of the total numbers of the two types of error counted over each set of 100 replications used to form Table 18. The increased frequency for larger N is an artefact of the algorithm used to calculate the Scan Statistic. Each event in turn forms the focus of a scan of its vicinity, so that the larger the number of events, the larger the number of calculations of δ that are required. The number of errors decreases with d and decreases as the order of the Taylor expansion used in the correction increases; both factors will tend to improve

the approximation to (4.10), so this observation is intuitively reasonable. Neither type of error was recorded for the simulations implementing (4.14).

N	Method	d		
		1/4	1/16	1/64
20	(4.12)	663 / 183	53 / 16	0 / 0
	(4.13)	226 / 59	0 / 0	0 / 0
	(4.14)	0 / 0	0 / 0	0 / 0
150	(4.12)	23412 / 4597	963 / 227	0 / 0
	(4.13)	10576 / 1915	0 / 0	0 / 0
	(4.14)	0 / 0	0 / 0	0 / 0

Table 19: Number of errors occurring in the estimation of δ in the simulations used to produce Table 18. First figure equals the number of intervals set to zero; second is the number of intervals greater than one.

4.3.2 Generalisation of the Kernel Method

The results of Section 4.2 suggest that, under certain conditions, an adaptive kernel density estimate could provide an acceptable method of estimating the null distribution of events, which is unknown in practice, for inclusion in the correction method. The adaptive k.d.e. generalises easily to two dimensions (Silverman, 1986), and so it is possible to replace $f(\cdot,\cdot)$ in (4.10) by a suitable $\hat{f}(\cdot,\cdot)$. To reduce the computational workload inherent in calculating a kernel density estimate, it is desirable to use kernels that have bounded support; *i.e.* kernels that are only non-zero at coordinates in a finite interval in their domains. The two-dimensional Epanechnikov kernel,

$$K_e(\underline{x}) = \begin{cases} 2\pi^{-1}\left(1-\underline{x}^T\underline{x}\right), & \underline{x}^T\underline{x} < 1, \\ 0, & \text{otherwise}, \end{cases}$$

is ideal for the pilot estimate, because it is simple to calculate, while still having good theoretical properties. The second-stage, or adaptive, estimator requires greater differentiability to allow the implementation of (4.12), (4.13) or (4.14), as these involve partial derivatives of $\hat{f}(\cdot,\cdot)$ to increasing orders. A suitable kernel for the first two approximations is

$$K_2(\underline{x}) = \begin{cases} 3\pi^{-1}\left(1-\underline{x}^T\underline{x}\right)^2, & \underline{x}^T\underline{x} < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The third approximation, (4.14), requires an estimate with more derivatives, so the kernel $K_3(\cdot)$, where

$$K_3(\underline{x}) = \begin{cases} 4\pi^{-1}\left(1-\underline{x}^T\underline{x}\right)^3, & \underline{x}^T\underline{x} < 1, \\ 0, & \text{otherwise,} \end{cases}$$

is a better choice. Both $K_2(\cdot)$ and $K_3(\cdot)$ were suggested by Silverman (1986).

The above kernels were used to investigate, by simulation, the effect of inserting a kernel density estimate into (4.12), (4.13) and (4.14). The null distribution of events was assumed to be the mixture p.d.f., (4.15), and 100 controls and a given number of events were sampled from this distribution on each of 100 replications. The smoothing parameter was chosen to minimise the approximate mean integrated square error of the estimator, the usual criterion for an 'optimal' bandwidth. For a general two-dimensional problem, the formula for this optimal h is

$$h_{\text{opt}} = \left[2\beta\alpha^{-2}\left\{\int\left(\nabla^2 f\right)^2\right\}^{-1}\right]^{\frac{1}{6}} N_c^{-\frac{1}{6}}, \tag{4.16}$$

where

$$\alpha = \iint s^2 K(s,t)\,ds\,dt,$$

$$\beta = \iint K^2(s,t)\,ds\,dt,$$

N_c is the number of (control) observations in the sample and $K(\cdot)$ is the particular kernel in use. For $f(\cdot,\cdot)$ equal to (4.15), $\int\left(\nabla^2 f\right)^2$ can be shown to be 2148.98 (to two decimal places). For the K_2 kernel described above, the optimal smoothing parameter, given (4.15), is therefore

$$h_2 = 0.56970 N_c^{-\frac{1}{6}};$$

similarly, use of the K_3 kernel leads to

$$h_3 = 0.63937 N_c^{-\frac{1}{6}}.$$

Critical values at the 5% level for the Scan Statistic, when simulated using the three different approximations to (4.10) and the above parameters, are displayed in Table 20. For reference, the corresponding critical values for the uncorrected statistic (assuming a uniform null distribution of events) are also included. Overall, the accuracy of the three approximations is poor, although there is a trend towards the correct critical values as the number of terms retained in the Taylor expansion increases. For (4.14), the values for small d, $d < 1/16$, are close to the target uniform results, if allowance is made for sampling variability and the different numbers of replications: the latter figures were generated from 1000 simulations, rather than 100 for the rest of Table 20.

By increasing the precision of the estimate of $f(\cdot,\cdot)$, however, some improvement can be made in the accuracy of this type of correction procedure. Table 21 employs (4.14) only and, using the same null distribution, kernels, bandwidth and number of replications as Table 20, estimates 5% critical values for the Scan Statistic by simulation, with control sample sizes of 200, 400 and 800. Accuracy increases with the number of controls, so that with $N_c \geq 400$, the correct values are quite closely achieved for $d \leq 1/16$. For larger d, however, (4.14) is still very poor, with most of the results much greater than the corresponding true values.

As in the previous section, it is clear that the smoothing parameter, h, of the kernel density estimate will also affect the accuracy of the approximations. It is well known that the precision of a k.d.e. depends on h, so the bandwidth will have considerable influence on the results, just as the sample size does in Table 21.

N	Method	*d*				
		1/4	1/8	1/16	1/32	1/64
20	Uniform	7	5	4	3	3
	(4.12)	9	6	4	3	3
	(4.13)	14	6	4	3	3
	(4.14)	13	5	4	3	3
50	Uniform	12	7	5	4	3
	(4.12)	30	12	6	4	3
	(4.13)	48	12	6	4	3
	(4.14)	40	9	5	4	3
100	Uniform	18	10	6	4	4
	(4.12)	89	19	10	5	4
	(4.13)	100	23	8	5	4
	(4.14)	90	82	7	5	4
150	Uniform	23	12	7	5	4
	(4.12)	139	32	16	6	4
	(4.13)	150	106	9	6	4
	(4.14)	132	117	8	6	4

Table 20: Simulated 5% Scan Statistic critical values (100 replications) for null distribution (4.15), using three approximate corrections based on adaptive k.d.e.'s calculated from 100 controls, compared to uniform null values.

N	No. of Controls	d 1/4	1/8	1/16	1/32	1/64
20	100	13	5	4	3	3
	200	10	5	4	3	3
	400	10	5	4	3	3
	800	9	5	4	3	3
	Uniform	7	5	4	3	3
50	100	40	9	5	4	3
	200	26	8	5	4	3
	400	22	7	5	4	3
	800	19	7	5	4	3
	Uniform	12	7	5	4	3
100	100	90	82	7	5	4
	200	56	51	6	4	4
	400	46	25	6	5	4
	800	39	13	7	5	4
	Uniform	18	10	6	4	4
150	100	132	117	8	6	4
	200	103	77	8	5	4
	400	74	47	7	5	4
	800	54	29	7	5	4
	Uniform	23	12	7	5	4

Table 21: Simulated 5% Scan Statistic critical values (100 replications) for null distribution (4.15), using (4.14) based on adaptive k.d.e.'s with different numbers of controls, compared to uniform null values.

4.4 Discussion

The results of both the one and two-dimensional investigations provide a consistent assessment of the implementation of a correction based on Weinstock (1981) that makes use of numerical integration and Taylor expansions. The performance of the method is dependent on the number of controls, N_c, the smoothing parameter, h, and the 'square size' constant, d. N_c and h govern the precision of the estimate of the null distribution of events; increasing the number of controls, for example, clearly improves accuracy when estimating the Scan Statistic and, hence, when assessing significance. The constant d affects the magnitude of, respectively, the error and remainder terms in the numerical integration and Taylor expansions of $f(\cdot)$ or $f(\cdot,\cdot)$. With smaller values of d, these terms are reduced and the correction is more successful. In the two-dimensional case, if $N_c = 800$ then the correction is adequate for choices of d less than or equal to 1/16. If the conditions of the problem are below this optimum, then the correction seems to overestimate the Scan Statistic, sometimes to a very great extent, which would lead to a very conservative test with correspondingly low power. In practice, h and d are quantities that must be specified by the user and, therefore, introduce an element of subjectivity to the procedure.

At the left-hand and lower boundaries of the region of interest, the polynomial for δ may have no real zeros; see the discussion of this in Section 4.3. If a square of either side 0 or l, where l is the maximum size of square that may be completely contained within the domain at the current point, is substituted, then it is possible for the Scan Statistic to be underestimated, since part of the total area has not been scanned properly. A similar comment may be applied to points nearer to the centre, at which no solution to the approximation to (4.10) is found. A more serious problem is suggested by the simulation results that indicate the approximations may provide values of δ that are much larger than the appropriate true size of square, even to the extent of returning $\delta > 1$, when the region of interest is the unit square. Although this behaviour is moderated by larger numbers of controls or smaller choices of d, it could lead to an increased risk of Type I errors.

A useful feature of the correction described in this chapter is the simplicity of tests of significance associated with it. The construction of the method is such that a null hypothesis of no clustering may be treated as being equivalent to one of a uniform distribution of events on the region of interest. Therefore, a Monte Carlo significance test, suggested as the most effective inferential method for the Scan Statistic in Chapter 3, may be implemented by sampling a number of artificial events, equal to the sample

size of the original data set, independently from a uniform distribution on the unit square, then calculating the associated Scan Statistic. If the observed statistic was calculated with a constant $k = d^2/AB$ in (4.10), then the simulations are carried out using scanning squares with sides of length d. With this substitution, there is no need to sample artificial controls. The process is repeated independently R times. If the number of simulated statistics greater than or equal to the observed value is denoted by r, then an approximate p-value for the data set in question is

$$p = (r + 1) / (R + 1). \qquad (4.17)$$

When coded in FORTRAN, for example, this algorithm can be executed very quickly, even when R is large.

Instead of the procedure based on Weinstock (1981), it might be possible to correct for a non-uniform null distribution in other ways, *e.g.* by using a cartogram or a bootstrap significance test. A cartogram is a transformed version of an ordinary geographical map, with the property that the area of any sub-region is scaled to be proportional to the null distribution of events in that sub-region. Thus, the distribution over the whole region becomes uniform, which would allow the Scan Statistic of Chapter 3 to be applied without further modification. Schulman *et al* (1988) employed a cartogram with a different measure of clustering in their Density Equalised Map Projection method, which was discussed in Chapter 1.

An alternative approach would be to calculate the Scan Statistic for the observed set of events, with no corrections or transformations and a scanning window of a fixed size. A test of the 'no clustering' null hypothesis could be carried out by using a bootstrap significance test (Hinkley, 1988), based on a set of controls. The observed statistic would be compared to a set of simulated statistics calculated from a large number, *e.g.* 99 or 999, of artificial data sets of the same size as the original. Each simulated set of data could be obtained by independently sampling with replacement from the controls. Calculation of the p-value would then follow the model of (4.17) above. The second data set would be necessary so that resampling was carried out under the null hypothesis (Hall and Wilson, 1991); *i.e.* so that the bootstrap samples represented the distribution of events in the absence of any clustering component. It is likely, however, that this method would have very low power to detect clusters in areas for which the null density took very small values, within a domain that mixed regions of both high and low density. In this situation, the number of events in the cluster could be large for the immediately surrounding area, but small relative to the number found in the high density regions. It would be reasonable to expect the observed Scan Statistic to be found in one of the latter areas, since the scanning window remains fixed, and the same

would probably be true for the Scan Statistic of each bootstrap sample also. Therefore, the observed and bootstrapped statistics would have very similar values, and hence, the associated p-value would be quite large, even if some genuine clustering effect was present in the domain.

The use of a second set of data to provide an estimate of the null distribution of events has been suggested previously for other methods of detecting spatial clustering. In the context of malignant disease in human populations, for example, a number of the techniques discussed in Chapter 1, such as Lyon *et al* (1981), Cuzick and Edwards (1990) and Diggle and Chetwynd (1991), make use of samples of the non-diseased population as some form of baseline for the null hypothesis. However, as some of these authors note, it may be very difficult to obtain a representative sample of controls. The proposals contained in this thesis assume that a suitable sampling frame is available and that controls may be drawn from it without excessive difficulty. The results of the preceding sections suggest that as many controls as possible should be obtained, so that the estimate of the null distribution achieves maximum accuracy. Therefore, the sampling procedure must be capable of generating a large number of controls, without a disproportionate penalty in terms of cost. If human populations are of interest, it would often seem to be desirable to sample controls from the decennial Census; this has been suggested by a number of authors, *e.g.* in the contributions of Mantel and Clayton and Yandell to the discussion of Cuzick and Edwards (1990). However, it is possible that the limitations of Census information, such as insensitivity to migration and a discretised coordinate system (discussed further in Chapter 1), would carry over to controls obtained in this way. The use of an ancillary data set was proposed partly as a means of avoiding such problems, so there would appear to be disadvantages to the Census sampling approach, despite its relative simplicity.

The correction method of Weinstock (1981), when generalised to two dimensions, is a sensible and intuitive way of allowing the Scan Statistic to be used when the distribution of events under the null hypothesis is non-uniform. However, the particular implementation described here may be unreliable under certain circumstances. Therefore, it would be desirable to find a more robust algorithm for use in a real application. One possibility is based on a piecewise constant approximation to the density estimate of $f(\cdot,\cdot)$ in (4.10). Values of the kernel density estimate, denoted by e_{ij}, are calculated for each point (i,j) of a very fine grid overlaying the region of interest. If the grid points are separated by a distance, g, along each axis, then the value of the estimate within the $g \times g$ square region centred on (i,j) is approximated by e_{ij}. The volume under the k.d.e. bounded by a general square, W, of any size may then be

approximated by summing the contributions $g^2\ e_{ij}$ for each grid point contained within W. Hence, the length of the side of a scanning square at a particular location can be set to $(m\text{-}1)g$, where m is the maximum number of grid points in each axis direction allowing an approximate volume, as defined above, that is less than or equal to d^2/AB. The use of a Fast Fourier Transform algorithm to calculate the kernel density estimate at a grid of distinct values would decrease significantly the computational workload of this method.

CHAPTER 5

AN INTEGRATED SQUARED DIFFERENCE STATISTIC

5.1 Introduction

The general principle of a test of spatial clustering of the type examined here is to compare the distribution of the kind of events under consideration to the distribution that would be expected if there was no clustering effect. In the preceding chapters, this was mainly accomplished by using the Scan Statistic to examine the pattern of events and a kernel density estimate, calculated from a sample of controls, to represent the null distribution. However, Chapter 2 briefly investigated the assessment of clustering by calculating the maximum of a k.d.e. for the probability density of events, when the null distribution is assumed to be uniform. Chapter 5 combines the two approaches, by considering a method of comparing two kernel estimates, one of which is calculated from the events and the other from the controls.

A number of different measures could be defined for the purpose of comparing two kernel density estimates. For example, Ahmad (1980) describes an affinity measure, λ, between two probability density functions that is estimated by replacing the true p.d.f's by kernel estimates; *i.e.*

$$\hat{\lambda} = \frac{\int \hat{f}(x)\, dG_n(x) + \int \hat{g}(x)\, dF_n(x)}{\int \hat{f}^2(x)\, dx + \int \hat{g}^2(x)\, dx},$$

where $\hat{f}(\cdot)$ and $\hat{g}(\cdot)$ are the kernel estimates, and $F_n(\cdot)$ and $G_n(\cdot)$ the empirical distribution functions, calculated from the two sets of data. An alternative would be the ratio of the two kernel estimators, as suggested by Bithell (1990). The measure to be investigated here is the integrated squared difference (ISD) between $\hat{f}_1(\cdot)$ and $\hat{f}_2(\cdot)$, *i.e.*

$$T_{h_1 h_2} = \int \left\{ \hat{f}_1(x) - \hat{f}_2(x) \right\}^2 dx, \tag{5.1}$$

where $\hat{f}_1(\cdot)$ is the kernel estimate of the p.d.f. of events, calculated with smoothing parameter h_1, and $\hat{f}_2(\cdot)$ is the controls estimate, which has smoothing parameter h_2.

Expression (5.1) would seem to be a natural test statistic to consider in the kernel density estimation context, because of its similarity to the integrated square error (ISE) of a k.d.e.,

$$\text{ISE} = \int \left\{ \hat{f}(x) - f(x) \right\}^2 dx, \tag{5.2}$$

which is used frequently to measure the global performance of the kernel estimator.

Hall (1984) provides a central limit theorem for (5.2), using martingale and U-statistic theory. The results, and methods of the proof, are employed in Section 5.2 to demonstrate the asymptotic normality of (5.1). Simulation results regarding the small sample behaviour of (5.1) are described in Section 5.3, and Section 5.4 considers some modifications to $T_{h_1 h_2}$ that improve the theoretical properties of the statistic. Power against a certain class of alternatives is investigated by simulation in Section 5.5. The chapter closes with discussion in Section 5.6 and, in Section 5.7, some intermediate results required for the proof in Section 5.2.

5.2 Asymptotic Behaviour of the ISD Statistic

5.2.1 Notation

Let $\{X_1, \ldots, X_{n_1}\}$ and $\{Y_1, \ldots, Y_{n_2}\}$ be two independent samples from the p-dimensional p.d.f. $f(\cdot)$. Given two bandwidths or smoothing parameters, h_1 and h_2, the kernel density estimates of $f(\cdot)$ from the X and Y samples are

$$\hat{f}_1(x) = \left(n_1 h_1^p\right)^{-1} \sum_{i=1}^{n_1} K\left(\frac{x - X_i}{h_1}\right)$$

and

$$\hat{f}_2(x) = \left(n_2 h_2^p\right)^{-1} \sum_{i=1}^{n_2} K\left(\frac{x - Y_i}{h_2}\right),$$

respectively, where $K(\cdot)$ is a p.d.f. that satisfies the conditions

$$\int K(z)\,dz = 1, \quad \int z_i K(z)\,dz = 0 \quad \text{and} \quad \int z_i z_j K(z)\,dz = \delta_{ij}\alpha,$$

where α is a constant and δ_{ij} is the Kronecker delta.

The following quantities, included here for ease of reference, are used to simplify some complex expressions in the statement and proof of the Theorem below:

$$\sigma_A^2 = \int \left\{\nabla^2 f(x)\right\}^2 f(x)\,dx - \left[\int \left\{\nabla^2 f(x)\right\} f(x)\,dx\right]^2,$$

$$\sigma_B^2 = \int f^2(x)\,dx \left[\int \left\{\int K(u)K(u+v)\,du\right\}^2 dv\right],$$

$$\sigma_{hB}^2 = \int f^2(x)\,dx \left[\int \left\{\int K(u)K(u+v)\,du\right\}\left\{\int K(u)K(u+h_1 v/h_2)\,du\right\} dv\right],$$

$$\sigma_C^2 = \int \left\{\nabla^2 f(x)\right\} f(x)\,dx \int K^2(u)\,du,$$

$$\sigma_D^2 = \int \left\{\nabla^2 f(x)\right\}^2 f(x)\,dx$$

and

$$H(X_i, Y_j) = \int \left\{K\left(\frac{x-X_i}{h_1}\right) - \mathbb{E}K\left(\frac{x-X_i}{h_1}\right)\right\}\left\{K\left(\frac{x-Y_j}{h_2}\right) - \mathbb{E}K\left(\frac{x-Y_j}{h_2}\right)\right\} dx.$$

5.2.2 Theorem

Let λ_1, λ_2, λ_3, λ_4 and λ_5 be positive, finite constants. If, for $i = 1,2$,

$$h_i \xrightarrow{n_i \to \infty} 0,\ n_i h_i^p \xrightarrow{n_i \to \infty} \infty \text{ and } h_1/h_2 = O(1),$$

then $T_{h_1 h_2}$ is asymptotically normal with mean

$$\left\{\left(n_1 h_1^p\right)^{-1} + \left(n_2 h_2^p\right)^{-1}\right\} \int K^2(u)\,du + \tfrac{1}{4}\alpha^2\left(h_1^2 - h_2^2\right)^2 \int \left\{\nabla^2 f(x)\right\}^2 dx \qquad (5.3)$$

and variance

$$\Omega_1 + \Omega_2 + \Omega_3 - (\Omega_4 + \Omega_5) \qquad (5.4)$$

where

$$\Omega_i = \begin{cases} n_i^{-1} h_i^4 \alpha^2 \sigma_A^2, & \text{if } n_i h_i^{p+4} \to \infty, \\ 2 n_i^{-2} h_i^{-p} \sigma_B^2, & \text{if } n_i h_i^{p+4} \to 0, \\ n_i^{-\frac{p+8}{p+4}} \left(\alpha^2 \lambda_i^{\frac{4}{p+4}} \sigma_A^2 + 2\lambda_i^{-\frac{p}{p+4}} \sigma_B^2\right), & \text{if } n_i h_i^{p+4} \to \lambda_i, \end{cases}$$

for $i = 1,2$,

$$\Omega_3 = \begin{cases} \left(n_1^{-1}h_2^4 + n_2^{-1}h_1^4\right)\alpha^2\sigma_A^2, & \text{if } n_2h_2^{p+4} \to \infty \text{ or } n_1h_1^4h_2^p \to \infty, \\ 4\left(n_1n_2h_2^p\right)^{-1}\sigma_{hB}^2, & \text{if } n_2h_2^{p+4} \to 0 \text{ and } n_1h_1^4h_2^p \to 0, \\ n_1^{-1}n_2^{-\frac{4}{p+4}}\lambda_2^{-\frac{p}{p+4}}\left\{(\lambda_2+\lambda_3)\alpha^2\sigma_A^2 + 4\sigma_{hB}^2\right\}, & \text{if } n_2h_2^{p+4} \to \lambda_2 \text{ and } n_1h_1^4h_2^p \to \lambda_3, \end{cases}$$

$$\Omega_4 = \begin{cases} n_1^{-1}h_1^2h_2^2\alpha^2\sigma_D^2, & \text{if } n_1h_1^{p+2} \to \infty, \\ n_1^{-2}h_1^{-p}h_2^2\alpha\sigma_C^2, & \text{if } n_1h_1^{p+2} \to 0, \\ n_1^{-\frac{p+4}{p+2}}\lambda_4^{\frac{2}{p+2}}h_2^2\alpha\left(\lambda_4^{-1}\sigma_C^2 + \alpha\sigma_D^2\right), & \text{if } n_1h_1^{p+2} \to \lambda_4, \end{cases}$$

and

$$\Omega_5 = \begin{cases} n_2^{-1}h_1^2h_2^2\alpha^2\sigma_D^2, & \text{if } n_2h_2^{p+2} \to \infty, \\ n_2^{-2}h_1^2h_2^{-p}\alpha\sigma_C^2, & \text{if } n_2h_2^{p+2} \to 0, \\ n_2^{-\frac{p+4}{p+2}}\lambda_5^{\frac{2}{p+2}}h_1^2\alpha\left(\lambda_5^{-1}\sigma_C^2 + \alpha\sigma_D^2\right), & \text{if } n_2h_2^{p+2} \to \lambda_5. \end{cases}$$

5.2.3 Proof

Write $T_{h_1h_2}$ as

$$\int\left\{\hat{f}_1(x) - f(x)\right\}^2 dx + \int\left\{\hat{f}_2(x) - f(x)\right\}^2 dx - 2\int\left\{\hat{f}_1(x) - f(x)\right\}\left\{\hat{f}_2(x) - f(x)\right\}dx.$$

(1) From Hall (1984), $\int\left(\hat{f}_i - f\right)^2$, for $i = 1,2$, is asymptotically normal with mean

$$\left(n_ih_i^p\right)^{-1}\int K^2(u)\,du + \tfrac{1}{4}\alpha^2h_i^4\int\left\{\nabla^2 f(x)\right\}^2 dx$$

and variance V_i, where

$$V_i = \begin{cases} n_i^{-1}h_i^4\alpha^2\sigma_A^2, & \text{if } n_ih_i^{p+4} \to \infty, \\ 2n_i^{-2}h_i^{-p}\sigma_B^2, & \text{if } n_ih_i^{p+4} \to 0, \\ n_i^{-\frac{p+8}{p+4}}\left(\alpha^2\lambda_i^{\frac{4}{p+4}}\sigma_A^2 + 2\lambda_i^{-\frac{p}{p+4}}\sigma_B^2\right), & \text{if } n_ih_i^{p+4} \to \lambda_i. \end{cases}$$

(2) Following the method of Hall (1984), write $\int(\hat{f}_1 - f)(\hat{f}_2 - f)$ as

$$\int(\hat{f}_1 - \mathbb{E}\hat{f}_1)(\mathbb{E}\hat{f}_2 - f) + \int(\hat{f}_2 - \mathbb{E}\hat{f}_2)(\mathbb{E}\hat{f}_1 - f)$$
$$+\int(\hat{f}_1 - \mathbb{E}\hat{f}_1)(\hat{f}_2 - \mathbb{E}\hat{f}_2) + \int(\mathbb{E}\hat{f}_1 - f)(\mathbb{E}\hat{f}_2 - f), \tag{5.5}$$

and denote the first three terms in (5.5) by I_{11}, I_{12} and I_2, respectively. The fourth term is non-stochastic in nature.

(a) Let

$$I_{11} = (n_1 h_1^p)^{-1} \sum_{i=1}^{n_1} z_i,$$

where

$$z_i = \int\left[K\{(x - X_i)/h_1\} - \mathbb{E}K\{(x - X_i)/h_1\}\right]\{\mathbb{E}\hat{f}_2(x) - f(x)\}dx.$$

As in the proof of Lemma 1 from Hall (1984), let

$$t_i = \int K\{(x - X_i)/h_1\}\{\mathbb{E}\hat{f}_2(x) - f(x)\}dx.$$

Then

$$\mathbb{E}(t_i) = h_1^p \int\left\{\int K(z) f(x - zh_1)dz\right\}\{\mathbb{E}\hat{f}_2(x) - f(x)\}dx$$
$$= \tfrac{1}{2} h_1^p h_2^2 \alpha \int\{\nabla^2 f(x)\} f(x)\,dx + O(h_1^p h_2^2),$$
$$\mathbb{E}(t_i^2) = h_1^{2p} \int\int\left\{\int K(z)K(z+u) f(x - zh_1)dz\right\}\{\mathbb{E}\hat{f}_2(x) - f(x)\}$$
$$\{\mathbb{E}\hat{f}_2(x + uh_1) - f(x + uh_1)\}du\,dx$$
$$= \tfrac{1}{4} h_1^{2p} h_2^4 \alpha^2 \sigma_D^2 + O(h_1^{2p} h_2^4)$$

and

$$\mathbb{E}(t_i^k) \le c_k h_1^{kp} h_2^{2k},$$

for some constant c_k. Hence,

$$\mathbb{E}(z_i^2) \approx \tfrac{1}{4} h_1^{2p} h_2^4 \alpha^2 \sigma_A^2$$

and

$$\mathbb{E}(z_i^4) \le c_4 h_1^{4p} h_2^8.$$

Using the Lindeberg-Feller Theorem (Heyde, 1983) for this context,

$$\sum_{i=1}^{n_1} z_i$$

is asymptotically normal if, $\forall\ \varepsilon > 0$,

$$s^{-2}\sum_{i=1}^{n_1} \mathbf{E}\left\{z_i^2 I\left(|z_i| > \varepsilon s\right)\right\} \xrightarrow{n_1 \to \infty} 0, \tag{5.6}$$

where

$$s^2 = \sum_{i=1}^{n_1} \mathbf{E}\left(z_i^2\right).$$

If $g(\cdot)$ represents the p.d.f. of z_i, then noting that

$$\begin{aligned}
\mathbf{E}\left\{z_i^2 I\left(|z_i| > \varepsilon s\right)\right\} &= \int z^2 I\left(|z_i| > \varepsilon s\right) g(z)\, dz \\
&= \int_{|z| > \varepsilon s} z^2 g(z)\, dz \\
&\le (\varepsilon s)^{-2} \int_{|z| > \varepsilon s} z^4 g(z)\, dz \\
&\le (\varepsilon s)^{-2} \int z^4 g(z)\, dz \\
&= (\varepsilon s)^{-2}\, \mathbf{E}\left(z^4\right),
\end{aligned}$$

the left-hand side of (5.6) becomes

$$\begin{aligned}
\varepsilon^{-2} s^{-4} \sum_{i=1}^{n_1} \mathbf{E}\left(z^4\right) &\le \varepsilon^{-2}\left(\tfrac{1}{4} n_1 h_1^{2p} h_2^4 \alpha^2 \sigma_A^2\right)^{-2} n_1 c_4 h_1^{4p} h_2^8 \\
&= O\left(n_1^{-1}\right) \xrightarrow{n_1 \to \infty} 0.
\end{aligned}$$

Therefore, I_{11} is asymptotically normal with mean zero and variance

$$\tfrac{1}{4} n_1^{-1} h_2^4 \alpha^2 \sigma_A^2.$$

By symmetry, I_{12} is also asymptotically normal, with mean zero and variance

$$\tfrac{1}{4} n_2^{-1} h_1^4 \alpha^2 \sigma_A^2.$$

(b) Let

$$I_2 = \left(n_1 n_2 h_1^p h_2^p\right)^{-1} U,$$

where

$$U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} H\left(X_i, Y_j\right).$$

Note that

$$\mathbb{E}\{H(X,Y)\} = 0, \ \mathbb{E}\{H(X,Y)|X\} = 0 \text{ and } \mathbb{E}\{H(X,Y)|Y\} = 0.$$

Hence, U is a degenerate, two-sample U-statistic. Theorem 2 of Khashimov (1988) demonstrates that U is asymptotically normal if, for $i = 1,2$,

$$\text{(i)} \ \ n_i^{-\frac{1}{2}} \sigma^{-3} \mathbb{E}\left\{\left|H(X,Y)\right|^3\right\} \xrightarrow{n_i \to \infty} 0,$$

and

$$\text{(ii)} \ \ \sigma^{-4} \mathbb{E}\left[\left\{\mathbb{E}\left(H(X_1,Y_2)H(X_1,Y_3)|Y_2,Y_3\right)\right\}^2\right] \xrightarrow{n_i \to \infty} 0,$$

where

$$\sigma^2 = \mathbb{E}\left\{H^2(X,Y)\right\}.$$

For condition (i),

$$\begin{aligned}\mathbb{E}\left\{\left|H(X,Y)\right|^3\right\} &\le \left[\mathbb{E}\left\{\left|H(X,Y)\right|^6\right\}\right]^{\frac{1}{2}} \\ &\le \left[\mathbb{E}\left\{H^2(X,Y)\right\}\right]^{\frac{3}{2}}.\end{aligned}$$

Therefore, using Result 1 of Section 5.7,

$$\begin{aligned}n_i^{-\frac{1}{2}}\left[\mathbb{E}\left\{H^2(X,Y)\right\}\right]^{-\frac{3}{2}} \mathbb{E}\left\{\left|H(X,Y)\right|^3\right\} &\le O\left\{n_i^{-\frac{1}{2}}\left(h_1^{2p}h_2^p\right)^{-\frac{3}{2}}\left(h_1^{2p}h_2^p\right)^{\frac{3}{2}}\right\} \\ &= O\left(n_i^{-\frac{1}{2}}\right) \xrightarrow{n_i \to \infty} 0,\end{aligned}$$

for $i = 1,2$. For condition (ii), Result 2 of Section 5.7 demonstrates that

$$\mathbb{E}\left[\left\{\mathbb{E}\left(H(X_1,Y_2)H(X_1,Y_3)|Y_2,Y_3\right)\right\}^2\right] = O\left(h_1^{4p}h_2^{3p}\right),$$

so that

$$\left[\mathbb{E}\left\{H^2(X,Y)\right\}\right]^{-2}\mathbb{E}\left[\left\{\mathbb{E}\left(H(X_1,Y_2)H(X_1,Y_3)\middle|Y_2,Y_3\right)\right\}^2\right]=O\left(h_1^{-4p}h_2^{-2p}h_1^{4p}h_2^{3p}\right)$$
$$=O\left(h_2^p\right)\xrightarrow{n_2\to\infty}0.$$

Thus, U is asymptotically normal with mean zero and variance $n_1n_2\,\mathbb{E}\left\{H^2(X,Y)\right\}$, which can be shown to equal $n_1n_2h_1^{2p}h_2^p\sigma_{hB}^2$ using Result 1 of Section 5.7. Hence, I_2 is asymptotically normal with mean zero and variance

$$\left(n_1n_2h_2^p\right)^{-1}\sigma_{hB}^2.$$

(c) I_{11}, I_{12} and I_2 are uncorrelated and their joint distribution is asymptotically normal. Therefore, using (a) and (b), we find that $I_{11}+I_{12}+I_2$ is asymptotically normal with mean zero and variance V, where

$$V=\tfrac{1}{4}\left(n_1^{-1}h_2^4+n_2^{-1}h_1^4\right)\alpha^2\sigma_A^2+\left(n_1n_2h_2^p\right)^{-1}\sigma_{hB}^2; \tag{5.7}$$

this can be rewritten as

$$V=\begin{cases}\tfrac{1}{4}\left(n_1^{-1}h_2^4+n_2^{-1}h_1^4\right)\alpha^2\sigma_A^2, & \text{if } n_2h_2^{p+4}\to\infty \text{ or } n_1h_1^4h_2^p\to\infty,\\ \left(n_1n_2h_2^p\right)^{-1}\sigma_{hB}^2, & \text{if } n_2h_2^{p+4}\to 0 \text{ and } n_1h_1^4h_2^p\to 0,\\ \tfrac{1}{4}n_1^{-1}n_2^{-\frac{4}{p+4}}\lambda_2^{-\frac{p}{p+4}}\left\{(\lambda_2+\lambda_3)\alpha^2\sigma_A^2+4\sigma_{hB}^2\right\}, & \text{if } n_2h_2^{p+4}\to\lambda_2 \text{ and } n_1h_1^4h_2^p\to\lambda_3,\end{cases}$$

by examining the order of the ratio of the two terms on the right-hand side of (5.7).

(3) Parts (1) and (2) demonstrate that the three components of $T_{h_1h_2}$ are jointly asymptotically normal. $T_{h_1h_2}$ is a linear combination of these terms, so it must also be asymptotically normal (Miller, 1964, pp. 22 - 24). The appropriate mean and variance are calculated as follows:

(a) $$\mathbb{E}\left(T_{h_1h_2}\right)=\mathbb{E}\left\{\int\left(\hat{f}_1-f\right)^2\right\}+\mathbb{E}\left\{\int\left(\hat{f}_2-f\right)^2\right\}-2\,\mathbb{E}\left\{\int\left(\hat{f}_1-f\right)\left(\hat{f}_2-f\right)\right\}$$
$$=\int Var_1+\int Var_2+\int\left(B_1-B_2\right)^2$$

$$\approx \left\{\left(n_1 h_1^p\right)^{-1} + \left(n_2 h_2^p\right)^{-1}\right\} \int K^2 + \tfrac{1}{4}\alpha^2 \left(h_1^2 - h_2^2\right)^2 \int \left\{\nabla^2 f\right\}^2,$$

where Var_i is the variance, and B_i the bias, of the kernel density estimate based on the ith data set, for $i = 1,2$.

(b) $$Var\left(T_{h_1 h_2}\right) = Var\left\{\int \left(\hat{f}_1 - f\right)^2\right\} + Var\left\{\int \left(\hat{f}_2 - f\right)^2\right\} + 4\,Var\left\{\int \left(\hat{f}_1 - f\right)\left(\hat{f}_2 - f\right)\right\}$$

$$-2\,Cov\left\{\int \left(\hat{f}_1 - f\right)^2, \int \left(\hat{f}_1 - f\right)\left(\hat{f}_2 - f\right)\right\} - 2\,Cov\left\{\int \left(\hat{f}_2 - f\right)^2, \int \left(\hat{f}_1 - f\right)\left(\hat{f}_2 - f\right)\right\}.$$

The covariance terms are evaluated in Result 3 of Section 5.7. By considering the ratio of n_i and h_i terms in each component, the first covariance may be rewritten as

$$\begin{cases} \frac{1}{2} n_1^{-1} h_1^2 h_2^2 \alpha^2 \sigma_D^2, & \text{if } n_1 h_1^{p+2} \to \infty, \\ \frac{1}{2} n_1^{-2} h_1^{-p} h_2^2 \alpha \sigma_C^2, & \text{if } n_1 h_1^{p+2} \to 0, \\ \frac{1}{2} n_1^{-\frac{p+4}{p+2}} \lambda_4^{\frac{2}{p+2}} h_2^2 \alpha \left(\lambda_4^{-1} \sigma_C^2 + \alpha \sigma_D^2\right), & \text{if } n_1 h_1^{p+2} \to \lambda_4, \end{cases}$$

and the second covariance similarly, by substituting n_2 for n_1, h_2 for h_1 and λ_5 for λ_4. This proves the Theorem of Section 5.2.2.

5.3 Finite Sample Behaviour

The theorem in Section 5.2.2 considers the limiting behaviour of $T_{h_1 h_2}$, making use of the central limit theorem for the ISE of a kernel density estimate from Hall (1984). Real applications, however, will involve the analysis of data sets that are finite and possibly quite small, *e.g.* 50 or 100 events. It is necessary, therefore, to have some information about the effectiveness of the asymptotic result as an approximation to the exact distribution of both ISE and the test statistic (5.1).

Tables 22 and 23 contain specimen values of the mean and variance of the asymptotic normal distributions of ISE and $T_{h_1 h_2}$, respectively. The former values were calculated from Hall (1984) and the latter from Section 5.2.2. The parameters were obtained for three different univariate p.d.f's, $f(\cdot)$, and three choices of sample size, n or n_1. In the case of Table 23, the value of n_2 was taken to be three times that of n_1, a ratio of events to controls that might be commonly chosen in a practical example. The bandwidth or

bandwidths were calculated from the usual formula for a smoothing parameter that minimises approximate mean integrated square error (see expression (2.9) in Section 2.3 or Silverman, 1986), assuming the use of a Gaussian kernel throughout. Once the required functionals of $f(\cdot)$ and $K(\cdot)$ had been evaluated analytically, the results quoted in the two tables were calculated by a short FORTRAN programme, using double precision arithmetic. It should be noted that the same calculations performed in single precision FORTRAN or on a pocket calculator gave different answers, especially for the case of the standard normal probability density function.

ISE and $T_{h_1 h_2}$ are the integrals of squared expressions and must be, therefore, entirely non-negative. Hence, if the distribution of each is to be adequately approximated by a normal density, it would be reasonable to expect that the quantities

$$\text{mean} \pm k \text{ standard deviations}, \tag{5.8}$$

when calculated from Hall (1984) or the Theorem above, should be contained entirely within the non-negative part of the real line, where k takes all of the values $\{1, 2, ..., l\}$, and l is at least three. However, in both cases, the standard deviations in the respective tables are usually more than half the corresponding mean, so that $k = 1$ is the best that can be achieved in (5.8). This suggests that the two densities may be skewed for the range of sample sizes considered here, and that there may be some practical difficulties in trying to use a normal distribution for significance testing with $T_{h_1 h_2}$ or as a measure of the behaviour of ISE.

Support for these observations is obtained from simulations of the distribution of $T_{h_1 h_2}$ under the null hypothesis. Using 100 replications, events and controls were sampled independently from the three one-dimensional probability density functions used in the production of Tables 22 and 23. Smoothing parameters were calculated as described above, again employing Gaussian kernels. Empirical density functions for each choice of $f(\cdot)$ and sample size pair are displayed in Figures 5, 6 and 7, and the mean and standard deviation of each set of simulated statistics corresponding to these histograms are listed in Table 24.

Density	Sample Size	Mean	Standard Deviation
N(0,1)	50	0.01456	0.01177
	250	0.00402	0.00276
	1250	0.00111	0.00065
Ga(2,1)	50	0.02077	0.01265
	250	0.00573	0.00297
	1250	0.00158	0.00070
½N(-1,1) + ½N(1,1)	50	0.01048	0.00786
	250	0.00289	0.00185
	1250	0.00080	0.00043

Table 22: Specimen means and standard deviations of the asymptotic normal distribution of ISE, from Hall (1984). Kernel density estimate uses a Gaussian kernel and optimal bandwidth.

Density	n_1	n_2	Mean	Standard Deviation
N(0,1)	50	150	0.01685	0.01443
	250	750	0.00465	0.00337
	1250	3750	0.00128	0.00079
Ga(2,1)	50	150	0.02404	0.01603
	250	750	0.00663	0.00374
	1250	3750	0.00183	0.00088
½N(-1,1) + ½N(1,1)	50	150	0.01213	0.00989
	250	750	0.00335	0.00231
	1250	3750	0.00092	0.00054

Table 23: Specimen means and standard deviations for the asymptotic normal distribution of $T_{h_1 h_2}$ from (5.3) and (5.4). Kernel density estimates use Gaussian kernels and optimal bandwidths.

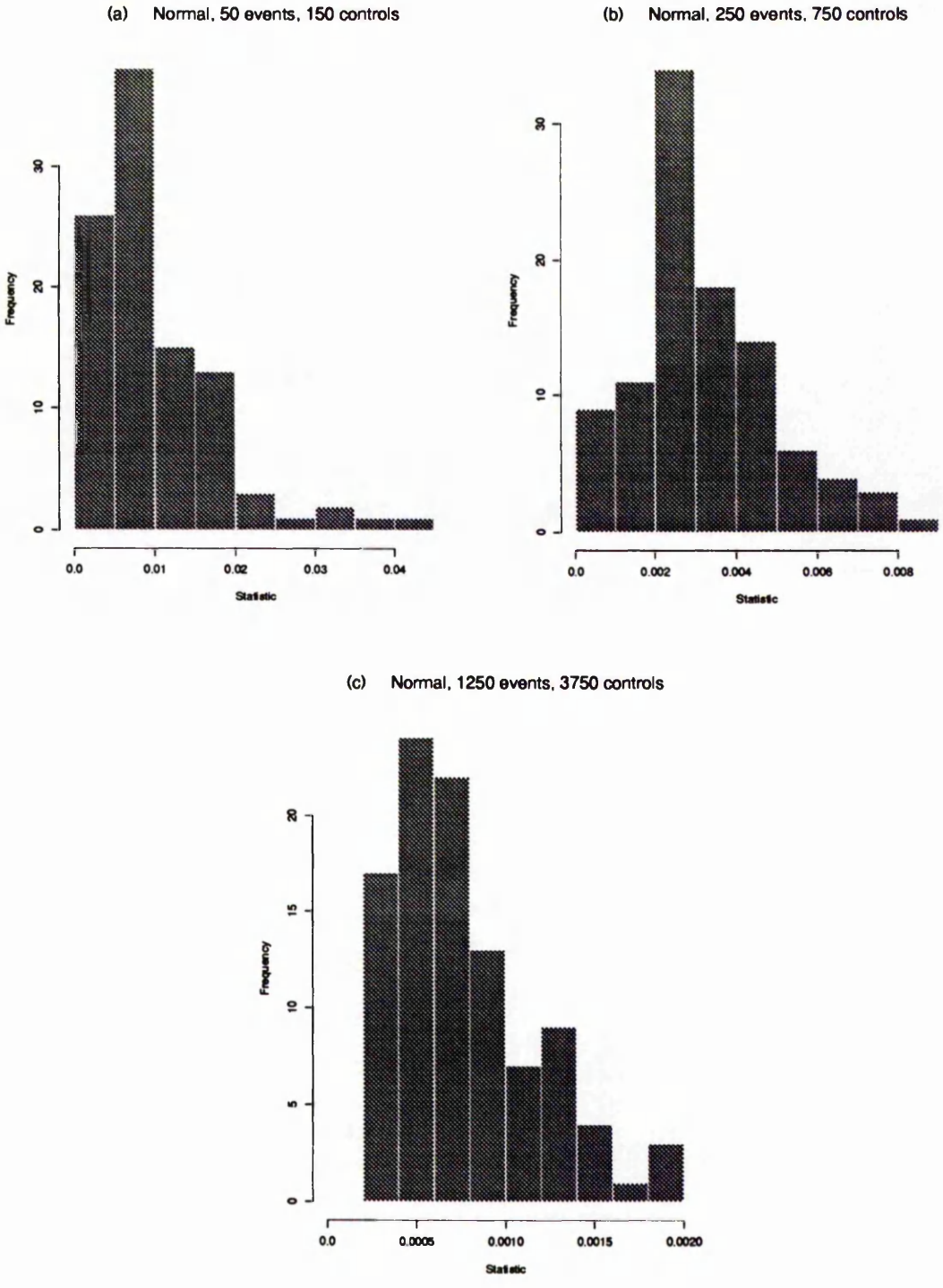


Figure 5: Empirical distribution of $T_{h_1 h_2}$ under the null hypothesis for N(0,1) density, with (a) $n_1 = 50$, $n_2 = 150$, (b) $n_1 = 250$, $n_2 = 750$ and (c) $n_1 = 1250$, $n_2 = 3750$. Calculated from 100 replications, with Gaussian kernels and optimal bandwidths.

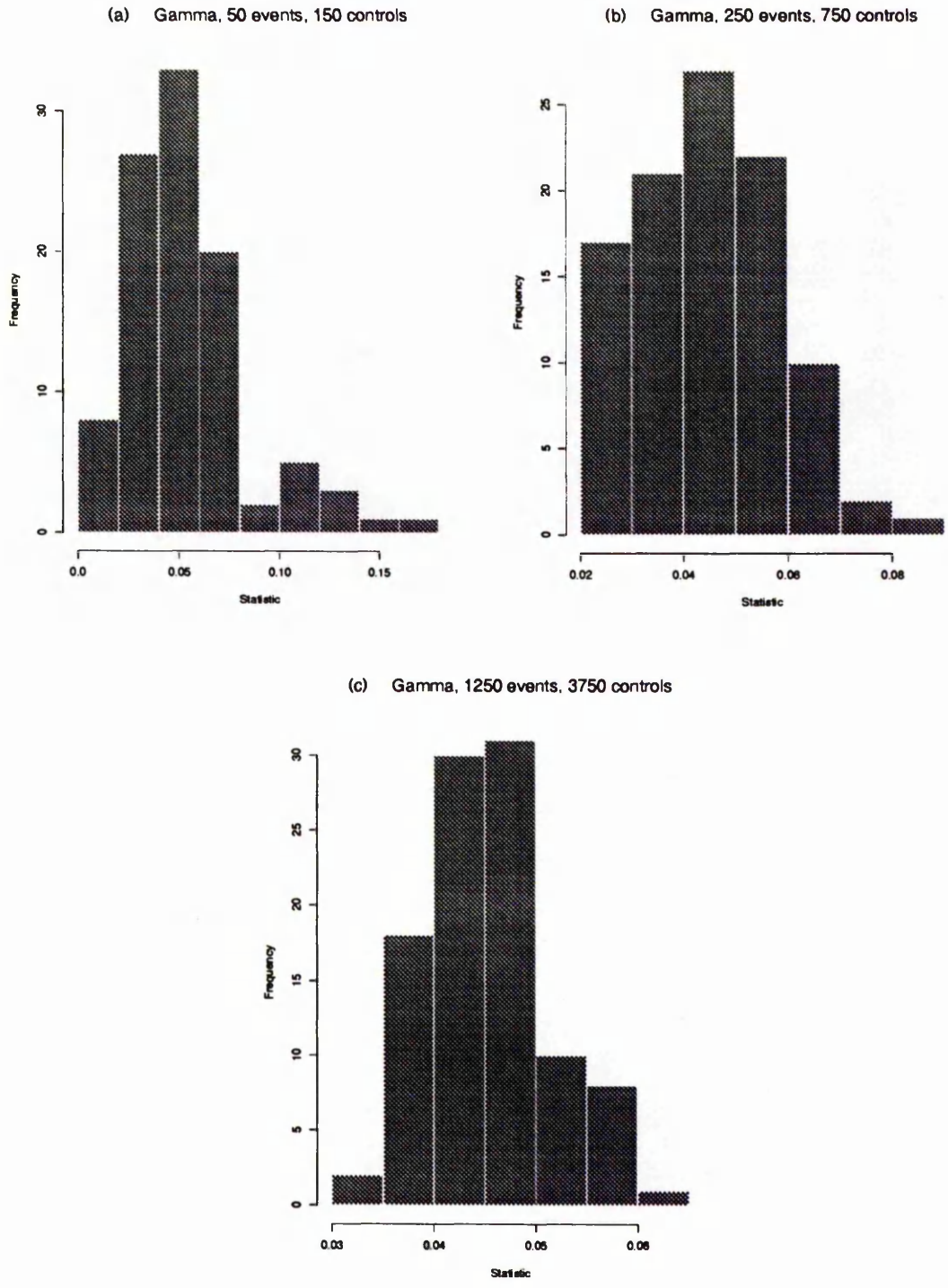


Figure 6: Empirical distribution of $T_{h_1 h_2}$ under the null hypothesis for Ga(2,1) density, with (a) $n_1 = 50$, $n_2 = 150$, (b) $n_1 = 250$, $n_2 = 750$ and (c) $n_1 = 1250$, $n_2 = 3750$. Calculated from 100 replications, with Gaussian kernels and optimal bandwidths.

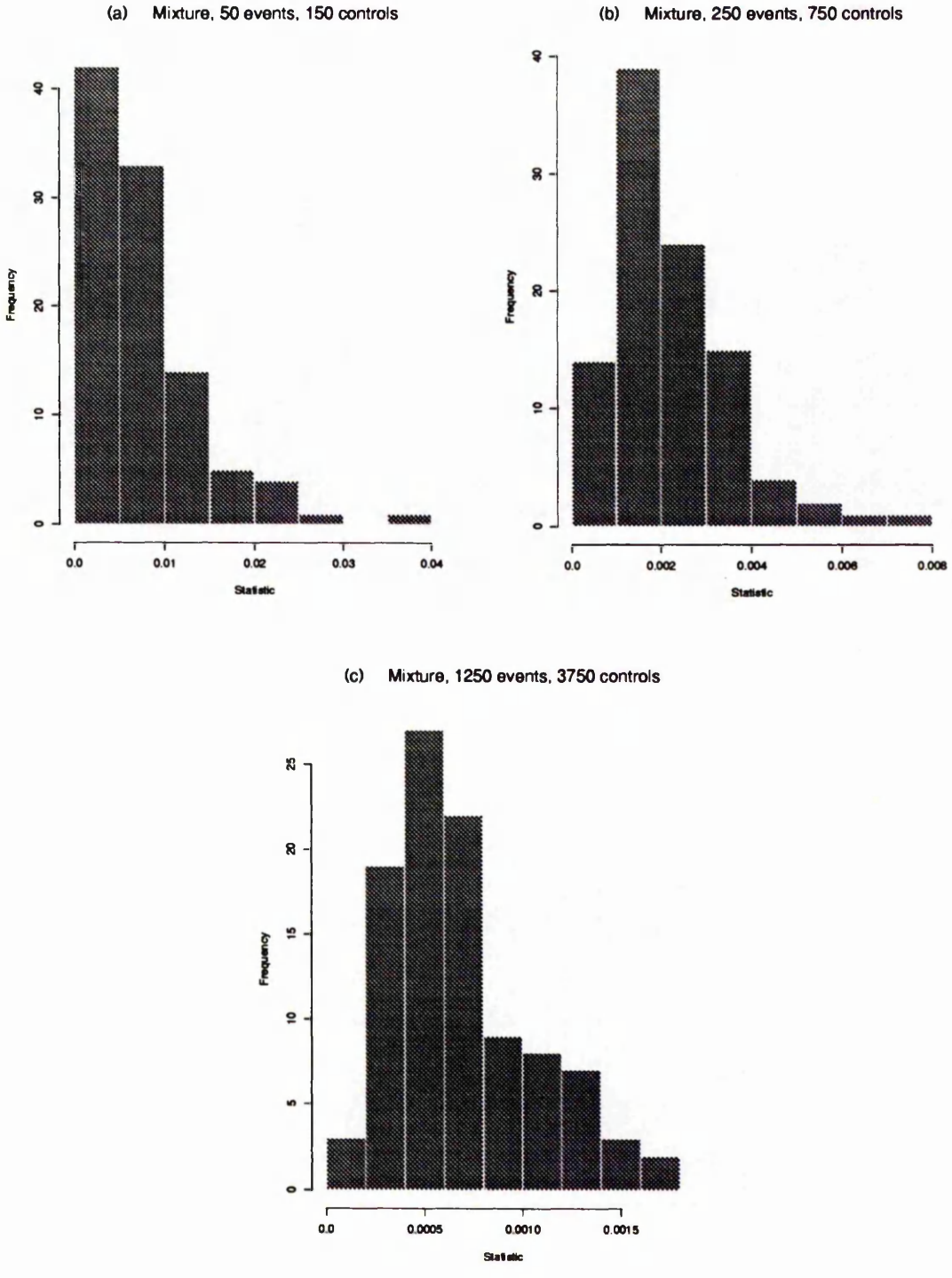


Figure 7: Empirical distribution of $T_{h_1 h_2}$ under the null hypothesis for mixture density, with (a) $n_1 = 50$, $n_2 = 150$, (b) $n_1 = 250$, $n_2 = 750$ and (c) $n_1 = 1250$, $n_2 = 3750$. Calculated from 100 replications, with Gaussian kernels and optimal bandwidths.

The empirical distributions in Figures 5 to 7 are positively skewed for all sample sizes and choices of $f(\cdot)$, which agrees with the analysis of the specimen means and standard deviations in Table 23. The simulated parameter values in Table 24 do not agree closely with those predicted from the Theorem, being between approximately 50% and 100% of the latter values for the Gaussian and mixture p.d.f's and approximately double for the Gamma (2,1) density. Since the available evidence suggests that the asymptotic result is a poor approximation to the small sample distribution, it would be unrealistic to expect the corresponding estimates of mean and variance to be substantially better. However, the ratio of mean to standard deviation, at least, does appear to be similar for the two tables.

Density	n_1	n_2	Mean	Standard Deviation
N(0,1)	50	150	0.00978	0.00761
	250	750	0.00314	0.00169
	1250	3750	0.00076	0.00040
Ga(2,1)	50	150	0.05408	0.03089
	250	750	0.04518	0.01291
	1250	3750	0.04507	0.00604
½N(-1,1) + ½N(1,1)	50	150	0.00758	0.00628
	250	750	0.00223	0.00131
	1250	3750	0.00068	0.00037

Table 24: Simulated means and standard deviations for $T_{h_1h_2}$, under the null hypothesis, corresponding to the empirical distributions in Figures 5 to 7. Based on 100 replications.

5.4 Modifications to the ISD Statistic

A number of alterations can be made to the form of $T_{h_1h_2}$ that improve its theoretical performance with respect to bias and the minimum distance between alternatives it is possible for the statistic to detect. A fuller account of the material in this section and in 5.5 is presented in Anderson *et al* (1992a).

If the standard normal distribution was to be used for hypothesis testing with the ISD statistic, (5.1) would have to be centred and standardised. From (5.3), it can be seen that the expectation of $T_{h_1h_2}$ is composed of the sum of two functionals, one involving $K^2(\cdot)$ and the other $\nabla^2 f(\cdot)$. The former can be calculated analytically, but the latter is an unknown and, therefore, would have to be estimated in practice. However, by setting $h_1 = h_2 = h$ (and writing $T_{h_1h_2}$ as T_h), the term involving $\nabla^2 f(\cdot)$ will be canceled out, leaving a much simpler expression for the mean. Essentially, using the same bandwidth for $\hat{f}_1(\cdot)$ and $\hat{f}_2(\cdot)$ reduces the bias inherent in the use of kernel methods to estimate the ISD between two probability density functions. Although this might lead to sub-optimal estimates of the individual densities, interest here resides in a comparison of the two p.d.f's and the quality of the estimates is not of particular concern, so the importance of the bandwidth is reduced.

Under the alternative hypothesis for the test considered here, the two samples of data are drawn from different p.d.f's, f_1 and f_2. Anderson *et al* (1992a) consider the minimum separation at which an hypothesis test based on T_h can discriminate between f_1 and f_2, when the alternative is true. This is investigated for a class of alternatives of the form

$$f_1 = f \quad \text{and} \quad f_2 = f + \delta g, \tag{5.9}$$

where f is a given p.d.f. and g is a function integrating to zero. Clearly, (5.9) is not a restrictive definition, since any two p.d.f's may be included in the class by rewriting f_2 as

$$f_2 = f_1 + \delta\left(\frac{f_2 - f_1}{\delta}\right).$$

If δ is chosen to be of the form

$$\delta = cn^{-\frac{1}{2}}h^{-\frac{p}{2}}, \tag{5.10}$$

for some non-zero constant, c, $n = n_1 + n_2$ and n_1/n_2 bounded away from zero and infinity, then T_h can be shown to be asymptotically normal when the alternative hypothesis, (5.9) plus (5.10), is true. The argument follows that of Section 5.2.3 or Hall (1984), by demonstrating that the first two components of the expansion

$$T_h = \int \left\{ \hat{f}_1 - \hat{f}_2 - \mathbf{E}_{H_1}\left(\hat{f}_1 - \hat{f}_2\right) \right\}^2 + 2\int \left\{ \hat{f}_1 - \hat{f}_2 - \mathbf{E}_{H_1}\left(\hat{f}_1 - \hat{f}_2\right) \right\} \mathbf{E}_{H_1}\left(\hat{f}_1 - \hat{f}_2\right)$$
$$+ \int \left\{ \mathbf{E}_{H_1}\left(\hat{f}_1 - \hat{f}_2\right) \right\}^2$$

are asymptotically normal and that

$$\int \left\{ \mathbf{E}_{H_1}\left(\hat{f}_1 - \hat{f}_2\right) \right\}^2 \sim \delta^2 \int g^2 .$$

As a result, the variance of the statistic is an increasing function of c, so that the power of T_h against the alternative hypothesis tends to one as $|c| \to \infty$. Thus, the minimum separation permitting T_h to discriminate between f_1 and f_2 is

$$O\left(n^{-\frac{1}{2}} h^{-\frac{p}{2}}\right).$$

As $h \to 0$, $n^{-\frac{1}{2}} h^{-\frac{p}{2}} \to \infty$ for a given n, so that, in general, if we wish to discriminate between distributions separated by $O(n^{-1/2})$, we must fix the smoothing parameter, h. Setting $h = 1$ is convenient, since this simplifies the calculations for T_h, although any value that does not depend on n will suffice. Hereafter, denote the test statistic with a fixed smoothing parameter equal to one by T. A minimum discrimination distance for T of $O(n^{-1/2})$ compares favourably with many tests in a parametric setting (Hall and Hart, 1990).

Under the null hypothesis, T is no longer asymptotically normal, since the fixed bandwidth contravenes one of the conditions of the Theorem in Section 5.2.2. If

$$a(x) = \int K(x-y) f(y) dy ,$$

$$J_1 = \int K^2 - \int a^2$$

and

$$M(x_1, x_2) = \int \{K(x-x_1) - a(x)\}\{K(x-x_2) - a(x)\} dx ,$$

then Anderson *et al* (1992a) demonstrate that

$$n\left\{T-\left(n_1^{-1}+n_2^{-1}\right)J_1\right\}\rightarrow S=\sum_{k=1}^{\infty}\lambda_k\left\{\left(\rho_1^{-\frac{1}{2}}Z_{1k}-\rho_2^{-\frac{1}{2}}Z_{2k}\right)^2-\left(\rho_1^{-1}+\rho_2^{-1}\right)\right\}$$

in distribution. Of the constituent parts of S, above, the terms ρ_i, $i = 1, 2$, are constants such that $n_i = \rho_i n$, for $0 < \rho_i < \infty$, the Z_{ij}, $i = 1, 2$, $j = 1, 2, \ldots$, are independent standard normal random variables and the λ_j, $j = 1, 2, \ldots$, are the coefficients of the orthogonal expansion of $M(\cdot,\cdot)$ in its eigenfunctions with respect to the weight $f(\cdot)$; *i.e.*

$$M(x_1,x_2)=\sum_{k=1}^{\infty}\lambda_k\omega_k(x_1)\omega_k(x_2),$$

where

$$\int M(x_1,x_2)\omega_k(x_1)f(x_1)dx_1=\lambda_k\omega_k(x_2)$$

and

$$\int\omega_k(x)\omega_l(x)f(x)dx=\delta_{kl},$$

where δ_{kl} is the Kronecker delta. This distribution is quite complex and depends upon the unknowns λ_1, λ_2, ..., so that significance testing may be more practical if the bootstrap (Hinkley, 1988) is used instead.

Given the original data, denoted by D, two artificial data sets , $D_1^{*(1)}$ and $D_2^{*(1)}$, of respective sizes n_1 and n_2, are created by sampling independently with replacement from D. The fixed bandwidth test statistic is calculated for these two samples, giving a value $T^{*(1)}$. This process is repeated a further (R - 1) times, *e.g.* where R = 100 or 1000, giving a sample of simulated (bootstrapped) test statistics,

$$\left\{T^{*(1)},\ldots,T^{*(R)}\right\}.$$

A critical value at the $100\alpha\%$ level, t_α, is estimated from this sample by setting $\hat{t}_\alpha$ equal to the bootstrapped statistic that ensures

$$\Pr\left(T^*>\hat{t}_\alpha|D\right)=\alpha.$$

The null hypothesis is rejected subsequently at the $100\alpha\%$ level if T, the observed value of the test statistic, is greater than $\hat{t}_\alpha$.

A further consideration is the composition of the "original data", D, in the preceding paragraph. Since the null hypothesis assumes that the two observed samples,

$$\{X_1,\ldots,X_{n_1}\} \text{ and } \{Y_1,\ldots,Y_{n_2}\},$$

are drawn from the same distribution, D might be taken to equal either sample individually or the pooled sample

$$\{X_1,\ldots,X_{n_1},Y_1,\ldots,Y_{n_2}\}.$$

The latter choice would naturally contain more data points, and should, therefore, have better level accuracy when the null hypothesis is true. However, in a context where one sample represents controls or some baseline distribution, an approach more intuitively connected to the detection of clustering might be to resample from the control data only. The set of simulated statistics thus obtained would be more typical of those expected in the absence of clustering, rather than simply the values expected when the two original samples were drawn from the same distribution. This condition is, of course, all that is required for the bootstrap test to be carried out correctly, so that, in practice, the choice of resampling scheme is arbitrary.

5.5 Power Comparison

The ability of the test statistic T to detect differences of smaller order than T_h, as indicated in the previous section, suggests that the former statistic may have greater power, both in general and for alternatives that differ from the null by only a small amount. To investigate this empirically, the powers of the two statistics were compared for a particular form of alternative by simulation.

Events were sampled independently on each of 1000 replications from two different probability density functions,

$$f_1(x) = (1-\gamma)\phi(x;0,1)+\gamma\phi\left(x;0,\sigma^2\right) \tag{5.11}$$

or

$$f_2(x) = \phi(x;0,1), \tag{5.12}$$

where $\phi\left(x;\mu,\sigma^2\right)$ denotes the density function of the univariate normal distribution with mean μ and variance σ^2. In (5.11), the variance σ^2 was chosen to be 2 or 4 and the mixture parameter, γ, was defined to be

$$\gamma = c(n_1+n_2)^{-\frac{1}{2}}, \tag{5.13}$$

where n_1 and n_2 were the sizes of the samples to be drawn from f_1 and f_2, respectively, and c was a constant, taking the values 1, 2, 4 or 6. The kernel estimates for T and T_h were calculated using Epanechnikov kernels (Silverman, 1986, p. 42), and in the case of

the latter statistic, the bandwidth was chosen to minimise approximate mean integrated square error for (5.12) given the total sample size, *i.e.* formula (2.9) with $K(\cdot)$ equal to the Epanechnikov kernel, $f(x) = \phi(x;0,1)$ and $N = n_1 + n_2$.

At each replication, the significance of the simulated test statistic was assessed by a bootstrap hypothesis test (Hinkley, 1988). However, rather than following the method presented at the end of Section 5.4, each test followed the scheme of the sequential Monte Carlo tests proposed by Besag and Clifford (1991), the main benefit of which was the reduction of the computational load that the procedure permitted. For the sequential method, R bootstrap samples are generated as normal, but the test is declared to be significant at the $100\alpha\%$ level if

$$\frac{r+1}{R+1} \leq \alpha,$$

where r is the number of bootstrapped test statistics greater than or equal to the observed value. However, if after generating the ith bootstrap sample,

$$r_i + 1 > \alpha(R+1),$$

where r_i is the number of bootstrapped values matching or exceeding the observed test statistic at stage i, then the bootstrap procedure is halted, since the test could not then attain significance at the $100\alpha\%$ level. In the simulations reported here, $\alpha = 0.05$ and $R = 199$. In the case of T_h, the original bandwidth was used to calculate each bootstrapped value of the test statistic.

Table 25 displays the results of the simulations for (a) $\sigma^2 = 2$ and (b) $\sigma^2 = 4$. Both cases are demanding so far as testing is concerned, and (a) in particular, and this is reflected in the low magnitude of the figures overall. As would be expected, power is greater for case (b) and increases with c. The sample sizes were chosen for consistency with previous work, but do not affect the results substantially, since the alternative hypothesis is corrected for n_1 and n_2 through (5.13): as the total number of observations increases, γ decreases and $f_1(\cdot) \rightarrow f_2(\cdot)$. When $c \geq 2$, the superiority of T is fairly clear, especially for case (b), which is consistent with the theoretical analysis of Section 5.4. The results for $c = 1$ suggest that this particular alternative is rather too demanding and that the significance of individual tests is simply attributable to chance and sampling variability, which means that the comparison of T_h and T is more difficult in this case.

(a) $\sigma^2 = 2$

n_1	n_2	Statistic	c			
			1.0	2.0	4.0	6.0
20	60	T_h	0.076	0.053	0.079	0.131
		T	0.066	0.082	0.107	0.172
50	150	T_h	0.066	0.072	0.063	0.130
		T	0.054	0.075	0.087	0.132
100	300	T_h	0.047	0.064	0.071	0.079
		T	0.048	0.054	0.104	0.136

(b) $\sigma^2 = 4$

n_1	n_2	Statistic	c			
			1.0	2.0	4.0	6.0
20	60	T_h	0.068	0.109	0.166	0.334
		T	0.076	0.115	0.259	0.498
50	150	T_h	0.056	0.064	0.162	0.279
		T	0.058	0.087	0.235	0.480
100	300	T_h	0.065	0.066	0.107	0.233
		T	0.048	0.094	0.241	0.451

Table 25: Empirical power of T_h and T against the alternative (5.11), (5.12) and (5.13), with (a) $\sigma^2 = 2$ and (b) $\sigma^2 = 4$. Based on 1000 replications, with sequential bootstrap significance tests using a maximum of 199 resamples.

5.6 Discussion

The ISD statistic considered in this chapter is defined to be a measure of the difference between two probability density functions. As noted previously, it is an attractive and intuitive choice for analyses in the context of kernel density estimation. However, it is not dedicated to detecting an excess of one particular type of events over another. Therefore, $T_{h_1h_2}$, T_h or T will be sensitive to both excesses and deficits of, for example, cases compared to controls, in the case of applications involving incidence of disease, or what might be termed positive and negative clustering, respectively (Marshall, 1991). A benefit of this is that the test statistics may be useful in a range of problems that is broader than might otherwise be the case, *i.e.* as varieties of two-sample goodness of fit statistics. However, a disadvantage is that, for most types of investigations of spatial clustering, the power of an ISD statistic may be lower than that of a measure that detects differences in only one direction. A further consideration is that the value of the test statistic will not discriminate between the two situations in which a significant result is due to positive or negative clustering, although comparing three-dimensional plots of

$$\hat{f}_1 - \hat{f}_2$$

and

$$\left(\hat{f}_1 - \hat{f}_2\right)^2$$

may be informative, since the former quantity will indicate the sign, and the latter the relative magnitude, of peaks due to large, local differences in the values of the two kernel estimates.

The conditions imposed on the smoothing parameters, h_1 and h_2, by the Theorem in Section 5.2.2 should be reasonable in most practical applications. The bandwidths for the two samples would normally be chosen by the same method, *e.g.* by reference to a standard distribution (Silverman, 1986) or by least squares cross-validation (Bowman, 1984). These and similar methods would normally ensure that h_1 and h_2 were of order

$$O\left(n^{-\frac{1}{4+p}}\right),$$

where p is the dimensionality of the problem. Thus, both smoothing parameters would tend to zero with increasing n and would be of the same magnitude, satisfying the conditions in the Theorem.

However, the sample parameter values and simulation results of Section 5.3 suggest that the Theorem may be of limited use for significance testing. The apparent skewness of the small sample distribution of $T_{h_1h_2}$ implies that the use of a normal approximation may be restricted to very large samples, which will probably be too large to be achieved in practice. It is also clear that the application of the results of Hall (1984) for purposes other than theoretical analysis must be treated with great care. Usually, therefore, hypothesis testing for any of the statistics discussed in this chapter will have to be accomplished by a bootstrap approach, as described in Section 5.4.

The alterations to $T_{h_1h_2}$ outlined in Section 5.4 provide a way of extending the statistic based on (5.1) so that its theoretical properties, such as power, are improved. The simulation results of Section 5.5 provide some corroboration of this. T may prove to be an attractive statistic in practice, since it does not require the estimation of a suitable bandwidth, the usual difficulty encountered with implementing nonparametric smoothing methods. It is also qualitatively similar to a test proposed by Hall and Hart (1990) for comparing two nonparametric regression curves. The statistic proposed therein is based on the quantity

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}(x_i)-\hat{g}(x_i)\right\}^2,$$

where $\{x_1,..., x_n\}$ are the design points and $\hat{f}(\cdot)$ and $\hat{g}(\cdot)$ are the regression function kernel estimators, which have a common, fixed smoothing parameter. Asymptotically, T_n behaves in a similar way to T, since its distribution is non-normal unless the bandwidth is permitted to tend to zero as $n \to \infty$. In addition, T_n is capable of discriminating between functions that are $O(n^{-1/2})$ apart. However, difficulties may arise when using T if n_1 and n_2 are of very different magnitudes. Use of the same smoothing parameter for both of the estimators in these circumstances could lead to drastic over- or under-smoothing of one kernel density estimate. This might introduce bias, in the case of over-smoothing, or variability, otherwise, sufficient to make the results of the test unreliable.

A more general investigation of the use of orthogonal expansions to represent the distributions of certain test statistics is carried out by Anderson *et al* (1991), who examine Edgeworth expansions for sums of independent p-vectors, where p increases with sample size. Practical application of the results is limited in many circumstances, since individual terms in the Edgeworth expansion may be very difficult to compute.

However, knowledge that an expansion exists as a series in given powers of the sample size, n, can give an accurate interpolation rule for use with existing tables of a distribution (Anderson *et al*, 1992b). In the two-sample problem, as considered here, some progress can be made by using a Gram-Schmidt procedure, to provide empirical approximations to the required orthogonal functions, and a bootstrap approximation to the distribution of the test statistic, which is of the form

$$\hat{T} = n_2^{-1} \sum_{j=1}^{\infty} \lambda_j^2 \left\{ \sum_{i=1}^{n_2} \hat{\omega}_j(X_{2i}) \right\},$$

where

$$\{\lambda_j, j = 1,2,\ldots\}$$

is a sequence of positive constants that converges to zero,

$$\{\hat{\omega}_j(\cdot), j = 1,2,\ldots\}$$

is a sequence of orthonormal polynomials of degree j, estimated from the first sample of data, and

$$\{X_{21},\ldots,X_{2n_2}\}$$

is the second sample of data. By appropriate choices of the constants λ_j, the power of the statistic to detect a difference in one particular characteristic of the distributions, *e.g.* location or scale, can be increased.

5.7 Intermediate Results for Section 5.2.3

Result 1

$$\mathbb{E}\{H^2(X,Y)\} = \iint \mathbb{E}\left[\left\{K\left(\frac{x-X}{h_1}\right) - \mathbb{E}K\left(\frac{x-X}{h_1}\right)\right\}\left\{K\left(\frac{y-X}{h_1}\right) - \mathbb{E}K\left(\frac{y-X}{h_1}\right)\right\}\right]$$

$$\times \mathbb{E}\left[\left\{K\left(\frac{x-Y}{h_2}\right) - \mathbb{E}K\left(\frac{x-Y}{h_2}\right)\right\}\left\{K\left(\frac{y-Y}{h_2}\right) - \mathbb{E}K\left(\frac{y-Y}{h_2}\right)\right\}\right] dx\,dy$$

$$= \iint \left[\int K\left(\frac{x-X}{h_1}\right)K\left(\frac{y-X}{h_1}\right)f(X)dX - \int K\left(\frac{x-W}{h_1}\right)f(W)dW \int K\left(\frac{y-V}{h_1}\right)f(V)dV\right]$$

$$\times\left[\int K\left(\frac{x-Y}{h_2}\right)K\left(\frac{y-Y}{h_2}\right)f(Y)dY - \int K\left(\frac{x-W}{h_2}\right)f(W)dW \int K\left(\frac{y-V}{h_2}\right)f(V)dV\right] dx\,dy$$

$$= h_1^p h_2^p \iint AB\,dx\,dy, \tag{5.14}$$

where

$$A = \int K(z_1)K\left(\frac{y-x}{h_1} + z_1\right)f(x - z_1h_1)dz_1$$

$$-h_1^p \int K(z_2)f(x - z_2h_1)dz_2 \int K(z_3)f(x - z_3h_1)dz_3,$$

and B is as above, with h_2 replacing h_1 throughout. Then, (5.14) equals

$$h_1^{2p}h_2^p \iint \left\{\int K(z_1)K(z_1+u)f(x)dz_1 + O\left(h_1^p\right)\right\}$$

$$\times\left\{\int K(z_2)K\left(z_2 + \frac{h_1}{h_2}u\right)f(x)dz_2 + O\left(h_2^p\right)\right\}dx\,du$$

$$\approx h_1^{2p}h_2^p\sigma_{hB}^2$$

$$= O\left(h_1^{2p}h_2^p\right), \text{ if } h_1/h_2 = O(1).$$

Result 2

Let $E_{23} = \mathbb{E}\{H(X_1,Y_2)H(X_1,Y_3)|Y_2,Y_3\}$. Then

$$E_{23} = \iint \left\{K\left(\frac{x-Y_2}{h_2}\right) - \mathbb{E}K\left(\frac{x-Y_2}{h_2}\right)\right\}\left\{K\left(\frac{y-Y_3}{h_2}\right) - \mathbb{E}K\left(\frac{y-Y_3}{h_2}\right)\right\}$$

$$\times \mathbf{E}\left[\left\{K\left(\frac{x-X_1}{h_1}\right) - \mathbb{E}K\left(\frac{x-X_1}{h_1}\right)\right\}\left\{K\left(\frac{y-X_1}{h_1}\right) - \mathbb{E}K\left(\frac{y-X_1}{h_1}\right)\right\}\right] dx\,dy$$

$$= \iint \left\{K\left(\frac{x-Y_2}{h_2}\right) - \mathbb{E}K\left(\frac{x-Y_2}{h_2}\right)\right\}\left\{K\left(\frac{y-Y_3}{h_2}\right) - \mathbf{E}K\left(\frac{y-Y_3}{h_2}\right)\right\}$$

$$\times h_1^p \left\{\int K(z)K\left(\frac{y-x}{h_1}+z\right) f(x-zh_1)\,dz\right.$$

$$\left. -h_1^p \int K(z_1)f(x-z_1h_1)\,dz_1 \int K(z_2)f(y-z_2h_1)\,dz_2\right\} dx\,dy$$

$$\approx h_1^{2p} \iint \left\{K\left(\frac{x-Y_2}{h_2}\right) - \mathbf{E}K\left(\frac{x-Y_2}{h_2}\right)\right\}\left\{K\left(\frac{x-Y_3}{h_2}+\frac{h_1}{h_2}u\right) - \mathbf{E}K\left(\frac{x-Y_3}{h_2}+\frac{h_1}{h_2}u\right)\right\}$$

$$\times f(x)\left\{\int K(z)K(z+u)\,dz\right\} dx\,du.$$

$$\mathbb{E}\left(E_{23}^2\right) = h_1^{4p} \iint\iint f(x_1)f(x_2)\left\{\int K(z_1)K(z_1+u_1)\,dz_1\right\}\left\{\int K(z_2)K(z_2+u_2)\,dz_2\right\}$$

$$\times \mathbf{E}\left[\left\{K\left(\frac{x_1-Y_2}{h_2}\right) - \mathbb{E}K\left(\frac{x_1-Y_2}{h_2}\right)\right\}\left\{K\left(\frac{x_2-Y_2}{h_2}\right) - \mathbf{E}K\left(\frac{x_2-Y_2}{h_2}\right)\right\}\right]$$

$$\times \mathbf{E}\left[\left\{K\left(\frac{x_1-Y_3}{h_2}+\frac{h_1}{h_2}u_1\right) - \mathbb{E}K\left(\frac{x_1-Y_3}{h_2}+\frac{h_1}{h_2}u_1\right)\right\}\right.$$

$$\left.\times\left\{K\left(\frac{x_2-Y_3}{h_2}+\frac{h_1}{h_2}u_2\right) - \mathbb{E}K\left(\frac{x_2-Y_3}{h_2}+\frac{h_1}{h_2}u_2\right)\right\}\right] dx_1\,du_1\,dx_2\,du_2. \quad (5.15)$$

The first subsidiary expectation in (5.15) is approximately equal to

$$h_2^p \int K(a) K\left(a + \frac{x_2 - x_1}{h_2}\right) f(x_1 - a h_2)\, da + O\left(h_2^{2p}\right)$$

$$\approx f(x_1) h_2^p \int K(a) K\left(a + \frac{x_2 - x_1}{h_2}\right) da,$$

and the second to

$$h_2^p \int K\left(b + \frac{h_1}{h_2} u_1\right) K\left(b + \frac{x_2 - x_1}{h_2} + \frac{h_1}{h_2} u_2\right) f(x_1 - b h_2)\, db + O\left(h_2^{2p}\right)$$

$$\approx f(x_1) h_2^p \int K\left(b + \frac{h_1}{h_2} u_1\right) K\left(b + \frac{x_2 - x_1}{h_2} + \frac{h_1}{h_2} u_2\right) db.$$

Therefore,

$$\mathbb{E}\left(E_{23}^2\right) = h_1^{4p} h_2^{2p} \iint\!\!\iint f^3(x_1) f(x_2) \left\{\int K(z_1) K(z_1 + u_1)\, dz_1\right\}$$

$$\times \left\{\int K(z_2) K(z_2 + u_2)\, dz_2\right\} \left\{\int K(a) K\left(a + \frac{x_2 - x_1}{h_2}\right) da\right\}$$

$$\times \left\{\int K\left(b + \frac{h_1}{h_2} u_1\right) K\left(b + \frac{x_2 - x_1}{h_2} + \frac{h_1}{h_2} u_2\right) db\right\} dx_1\, du_1\, dx_2\, du_2$$

$$\approx h_1^{4p} h_2^{3p} \int f^4(x)\, dx \iint \left\{\int K(z_1) K(z_1 + u_1)\, dz_1\right\} \left\{\int K(z_2) K(z_2 + u_2)\, dz_2\right\}$$

$$\times \left[\int \left\{\int K(a) K(a + c)\, da\right\} \left\{\int K\left(b + \frac{h_1}{h_2} u_1\right) K\left(b + c + \frac{h_1}{h_2} u_2\right) db\right\} dc\right] du_1\, du_2,$$

by making the substitution $c = (x_2 - x_1)/h_2$. Therefore if $h_1/h_2 = O(1)$,

$$\mathbb{E}\left[\left\{\mathbb{E}\left\{H(X_1, Y_2) H(X_1, Y_3) \middle| Y_2, Y_3\right\}\right\}^2\right] = O\left(h_1^{4p} h_2^{3p}\right).$$

Result 3

To evaluate

$$Cov\left\{\int\left(\hat{f}_1-f\right)^2,\int\left(\hat{f}_1-f\right)\left(\hat{f}_2-f\right)\right\} \tag{5.16}$$

write $\int\left(\hat{f}_1-f\right)^2$ as

$$\int\left(\hat{f}_1-\mathbb{E}\hat{f}_1\right)^2+2\int\left(\hat{f}_1-\mathbb{E}\hat{f}_1\right)\left(\mathbb{E}\hat{f}_1-f\right)+\int\left(\mathbb{E}\hat{f}_1-f\right)^2,$$

and write $\int\left(\hat{f}_1-f\right)\left(\hat{f}_2-f\right)$ as

$$\int\left(\hat{f}_1-\mathbb{E}\hat{f}_1\right)\left(\hat{f}_2-\mathbb{E}\hat{f}_2\right)+\int\left(\hat{f}_1-\mathbb{E}\hat{f}_1\right)\left(\mathbb{E}\hat{f}_2-f\right)$$
$$+\int\left(\hat{f}_2-\mathbb{E}\hat{f}_2\right)\left(\mathbb{E}\hat{f}_1-f\right)+\int\left(\mathbb{E}\hat{f}_1-f\right)\left(\mathbb{E}\hat{f}_2-f\right).$$

Most of the resulting covariances are zero, leaving (5.16) equal to

$$Cov\left\{\int\left(\hat{f}_1-\mathbb{E}\hat{f}_1\right)^2,\int\left(\hat{f}_1-\mathbb{E}\hat{f}_1\right)\left(\mathbb{E}\hat{f}_2-f\right)\right\}$$
$$+2Cov\left\{\int\left(\hat{f}_1-\mathbb{E}\hat{f}_1\right)\left(\mathbb{E}\hat{f}_1-f\right),\int\left(\hat{f}_1-\mathbb{E}\hat{f}_1\right)\left(\mathbb{E}\hat{f}_2-f\right)\right\}. \tag{5.17}$$

The first covariance term in (5.17) equals

$$\left(n_1h_1^p\right)^{-3}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{l=1}^{n_1}Cov\left[\int\left\{K\left(\frac{x-X_i}{h_1}\right)-\mathbb{E}K\left(\frac{x-X_i}{h_1}\right)\right\}\left\{K\left(\frac{x-X_j}{h_1}\right)-\mathbb{E}K\left(\frac{x-X_j}{h_1}\right)\right\}dx,\right.$$

$$\left.\int\left\{K\left(\frac{x-X_l}{h_1}\right)-\mathbb{E}K\left(\frac{x-X_l}{h_1}\right)\right\}\left\{\mathbb{E}\hat{f}_2(x)-f(x)\right\}dx\right]$$

$$\approx\left(n_1h_1^p\right)^{-3}\sum_{i=1}^{n_1}\iint\tfrac{1}{2}h_2^2\alpha\left\{\nabla^2f(x)\right\}h_1^{2p}\left\{\int K^2(z)K(z+t)f(x)dz\right\}dt\,dx$$

$$=\tfrac{1}{2}n_1^{-2}h_1^{-p}h_2^2\alpha\sigma_C^2,$$

by ignoring all the terms in the summation other than those where $i=j=l$ and making the necessary substitutions. Using similar steps we find that the second covariance term in (5.17) equals

$$\left(n_1 h_1^p\right)^{-2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} Cov\left[\int \left\{ K\left(\frac{x-X_i}{h_1}\right) - \mathbb{E}K\left(\frac{x-X_i}{h_1}\right) \right\} \left\{ \mathbb{E}\hat{f}_1(x) - f(x) \right\} dx, \right.$$

$$\left. \int \left\{ K\left(\frac{x-X_j}{h_1}\right) - \mathbb{E}K\left(\frac{x-X_j}{h_1}\right) \right\} \left\{ \mathbb{E}\hat{f}_2(x) - f(x) \right\} dx \right]$$

$$\approx \tfrac{1}{4} n_1^{-1} h_1^2 h_2^2 \alpha^2 \sigma_D^2 \iint K(z) K(z+u)\, dz\, du$$

$$= \tfrac{1}{4} n_1^{-1} h_1^2 h_2^2 \alpha^2 \sigma_D^2,$$

since the double integral is a convolution (of a symmetric function, $K(\cdot)$), integrated over its entire range. Similarly,

$$Cov\left\{ \int \left(\hat{f}_2 - f\right)^2, \int \left(\hat{f}_1 - f\right)\left(\hat{f}_2 - f\right) \right\} \approx \tfrac{1}{2} n_2^{-2} h_1^2 h_2^{-p} \alpha \sigma_C^2 + \tfrac{1}{2} n_2^{-1} h_1^2 h_2^2 \alpha^2 \sigma_D^2.$$

CHAPTER 6

APPLICATION TO LARYNGEAL CANCER DATA

6.1 Introduction

The purpose of this penultimate chapter is to provide an example of the use of the Scan and ISD statistics that were examined on a theoretical basis in the preceding sections. The two techniques are used to analyse a real data set, which was provided by Professor P.J. Diggle and Doctor A.C. Gatrell of the University of Lancaster. The data are concerned with the geographical distribution of cases of cancer of the larynx in South Lancashire, and were previously analysed in Diggle (1990) and Diggle *et al* (1990).

The remainder of this section describes the data in more detail and outlines the method of analysis developed in Diggle (1990), together with the results of its application to the South Lancashire data. Section 6.2 carries out an investigation with the ISD statistic of Chapter 5 and compares the results to exploratory plots of the type proposed by Bithell (1990). Section 6.3 considers the application of a null-density-corrected Scan Statistic, as suggested in Chapter 4, and discusses the problems involved in its implementation. The results of the analyses are discussed in Section 6.4.

6.1.1 The South Lancashire Data

The data set consists of two groups of point locations in the Chorley and South Ribble Health Authority district in Lancashire, England. The first sample, containing 58 observations, provides the Ordnance Survey map reference of each case of laryngeal cancer recorded in the above area between 1974 and 1983, the reference being obtained from the Postcode of the address at diagnosis (Diggle *et al*, 1990). These coordinates are plotted in Figure 8. The second sample acts as a control group, and consists of the grid references of all 978 cases of lung cancer presenting in the same area and time period. These controls are shown, for comparison, in Figure 9. Addresses with the same Postcode were assigned the same grid reference, so the controls data set has a number of coordinate points at which there are two or more observations.

The small group of four cases of laryngeal cancer near the coordinates (35600,41400) in Figure 8 seemed anomalous when the two samples were compared, because the

density of lung cancer in that area appeared to be much lower than that of laryngeal cancer. This gave rise to some concern, since an industrial waste incinerator, located very close to the four cases, had been operating between 1972 and 1980 (Diggle *et al*, 1990). The aim of the original analysis in Diggle (1990) was to ascertain whether or not there could be a link between this possible source of environmental pollution and the apparent local increase in risk of cancer of the larynx.

The selection of controls from cases of a disease other than the one of interest has been suggested previously by, for example, Lyon *et al* (1981), but is quite unusual. Diggle *et al* (1990) suggest that the approach has two advantages. First, it models simultaneously the overall distribution of population and the less obvious variation in the composition of risk groups within the population that are defined by, for example, age and sex. Secondly, it controls for the effect of smoking, which is a risk factor of considerable importance in the aetiologies of both cancers. However, it is necessary to assume that the control malignancy is not associated with the mechanism responsible for the increased case incidence, if there is such a factor. Since the larynx and lungs are both part of the human respiratory system, an association with airborne pollution for the former site might cast doubt on the assumption that cancer of the latter site would not be affected similarly. Diggle *et al* (1990) point out, however, that the consequence of 'overmatching' in this way would be for the power of the method to be reduced, so that it would become important only if no association between source and cases was found.

6.1.2 Methods and Results of Diggle (1990)

Diggle (1990) analysed the Lancashire data by developing a technique for modelling the intensity function, $\lambda(\mathbf{x})$, of the inhomogeneous Poisson process from which cases of laryngeal cancer are assumed to derive, in terms of the distance from the point location representing the source. The method was one of those discussed in Section 1.2.1. A semiparametric, multiplicative model is assumed for $\lambda(\mathbf{x})$, with three separate terms to represent the different sources of variation, *i.e.*

$$\lambda(\mathbf{x}) = \rho\lambda_c(\mathbf{x})f(\mathbf{x}-\mathbf{x}_0;\alpha,\beta). \tag{6.1}$$

The first component of (6.1), ρ, represents the overall intensity or number of events per unit area. The second term, $\lambda_c(\mathbf{x})$, represents the variation in intensity due to the spatial heterogeneity of the population at risk. It is estimated by a kernel method from the control sample; given a suitable kernel function, $K(\cdot)$, and smoothing parameter, h, the estimator is

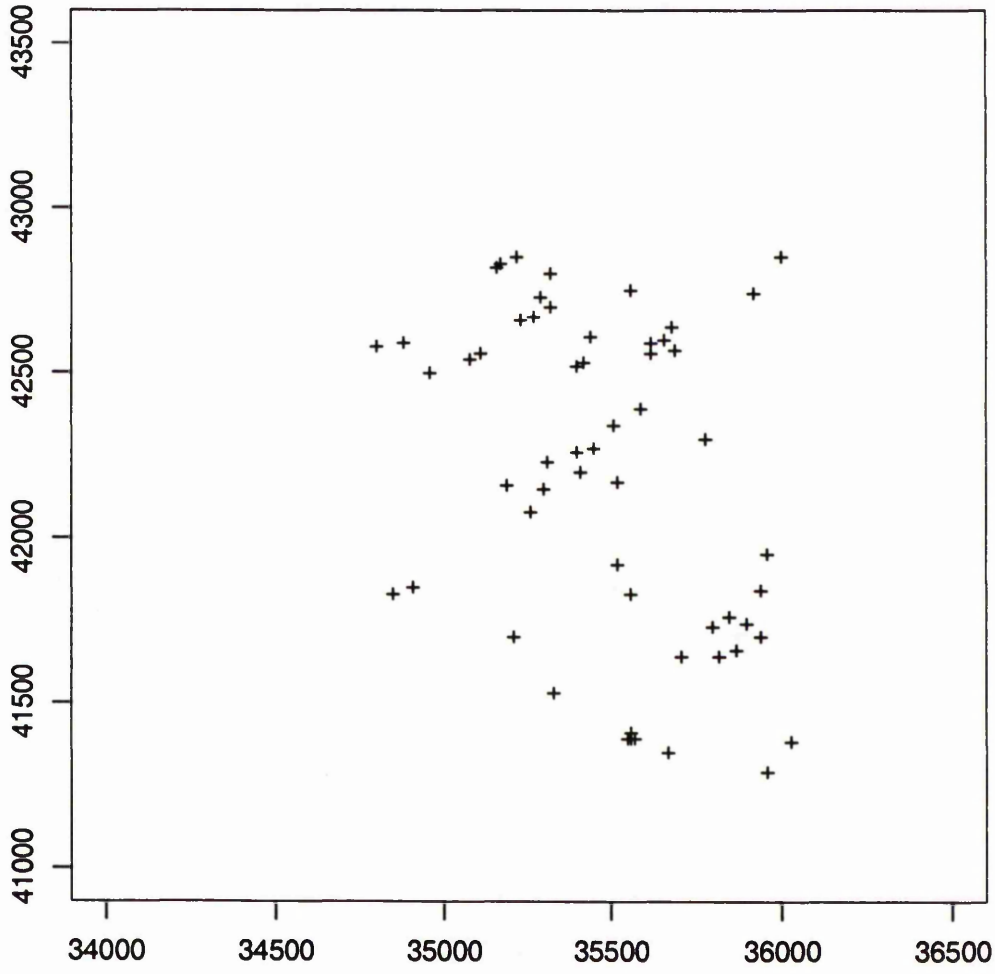


Figure 8: Geographical distribution of 58 cases of cancer of the larynx, from the South Lancashire data. Coordinates are Ordnance Survey map references.

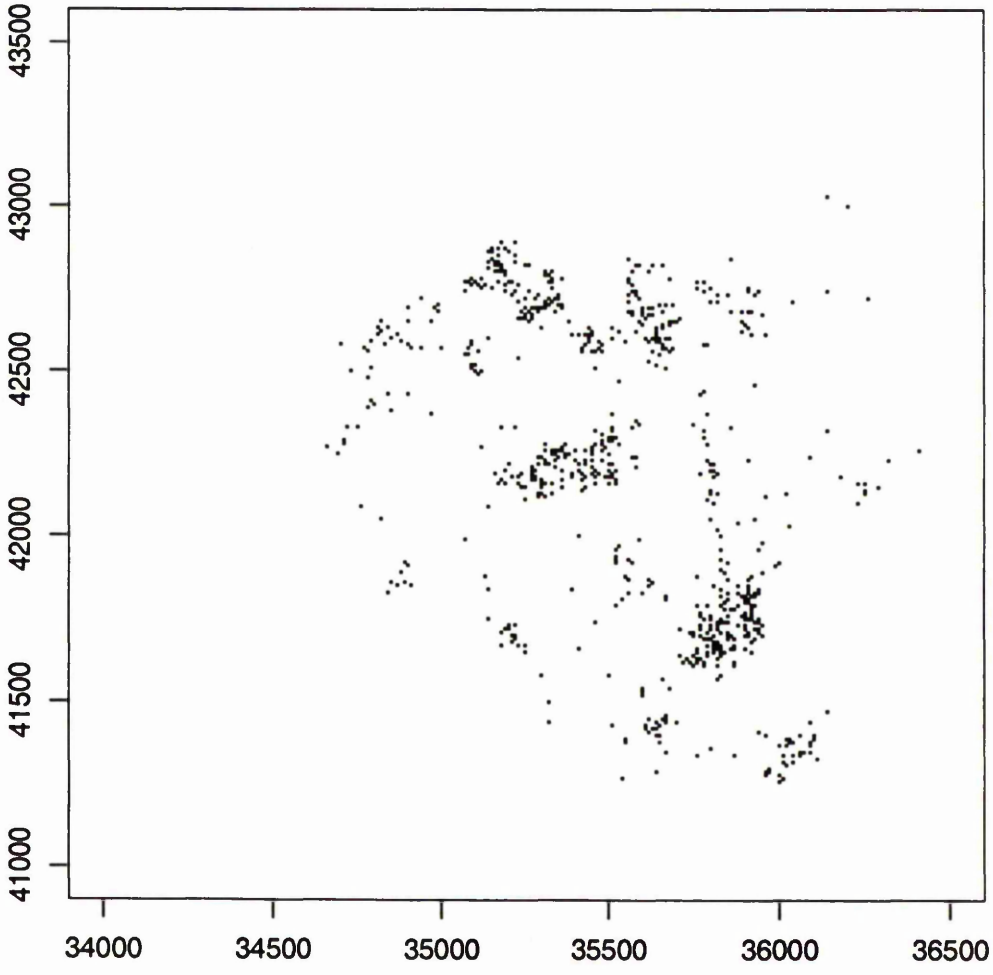


Figure 9: Geographical distribution of 978 cases of cancer of the lung, from the South Lancashire data. Coordinates are Ordnance Survey map references.

$$\hat{\lambda}_c(\mathbf{x}) = h^{-2} \sum_{i=1}^{n_2} K\{(\mathbf{x} - \mathbf{Y}_i)/h\}, \tag{6.2}$$

where $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}\}$ represents the set of n_2 controls. The general form of (6.2) is very similar to that of a k.d.e.; $K(\cdot)$ must satisfy the same conditions as those stated in Section 2.3, and the value of h plays a role in controlling the quality of the final estimate similar to that of the smoothing parameter in the density context. It should be noted, however, that there is a factor of n_2^{-1} missing from (6.2) that is present in the general expression for a kernel density estimator. The smoothing parameter is estimated by the method of Diggle (1985), with an auxiliary result from Berman and Diggle (1989), which selects the value of h that minimises an estimate of the mean square error of $\hat{\lambda}_c(\cdot)$, assuming that the data are a realisation of a Cox process (Diggle, 1983), so that $\lambda_c(\cdot)$ is also a stochastic quantity.

The third component of (6.1) models the dependence of intensity on distance from the point source at coordinates $\mathbf{x}_0$. The particular formulation of $f(\cdot,\cdot)$ determines the type of model being fitted, although the most likely requirement is for intensity to be greatest at the source and to decay exponentially in all directions as distance from $\mathbf{x}_0$ increases, when there is a genuine clustering effect, or for $f(\cdot,\cdot)$ to be equal to unity, under the null hypothesis. Diggle (1990) chooses

$$f(\mathbf{t};\alpha,\beta) = 1 + \alpha \exp\left(-\beta \mathbf{t}^T\mathbf{t}\right), \tag{6.3}$$

which is straightforward and satisfies the above requirements. Maximum likelihood estimation is used to fit the model, with a test of the null hypothesis, *i.e.* $f(\cdot,\cdot) = 1$, being provided by comparing the deviance statistic to χ^2_2.

Using a Gaussian kernel in (6.2), with a smoothing parameter of value $h = 0.15$ km (the region plotted in Figures 8 and 9 is a 25 × 25 km square), the maximum likelihood estimates of α and β in (6.3) for the South Lancashire data were 23.67 and 0.91, respectively, with corresponding estimated standard errors of 24.69 and 0.60 and a correlation of 0.83. The deviance statistic had an associated p-value of 0.008, a figure supported by the results of a Monte Carlo significance test, which suggested that the intensity of events within the region of interest was dependent on distance from the waste incinerator. Thus, the four cases previously identified could have been connected, in some way, to the operation of the plant. The scale of the clustering effect

appeared to be quite small, however, as it was sensitive to the removal of one or two of the four cases in the group. The deviance p-value for model (6.1), with (6.2) and (6.3), when one case was deleted at random from the "cluster", was 0.054. Randomly removing a second case increased the p-value again to

$$0.2231 < p < 0.2346.$$

6.2 ISD Statistic

6.2.1 Calculating the Test Statistic

Our first analysis of the Lancashire data employs the statistic $T_{h_1h_2}$ from (5.1), rather than T_h or T as considered in Section 5.4. The decision to use different, random bandwidths for the kernel estimates of the two samples was taken because of the great difference between the number of cases (58) and the number of controls (978). As a result of this disparity in sample sizes, using the same smoothing parameter for both samples could produce one very poor kernel estimate. For example, it is likely that a bandwidth calculated for the lung cancer data would be too small to be used for the cases of laryngeal cancer as well, because it would give a very undersmoothed, or noisy, estimate of the p.d.f. from which the latter observations were sampled. The noise might increase the apparent difference between the two kernel estimates and, thus, increase the probability of a Type I error.

Diggle and Marron (1988) demonstrate that the value of the smoothing parameter chosen for the kernel estimation of the intensity function of an inhomogeneous Poisson process by the method of Diggle (1985) is the same as that selected for kernel estimation of the density function by least squares cross-validation (Bowman, 1984), hereafter abbreviated to LSCV. This is a useful result, since LSCV is not straightforward for the controls data. The repeated coordinates, discussed in Section 6.1.1, cause the LSCV score function,

$$M_1(h) = n^{-2}h^{-1}\sum_i\sum_j K^*\left(\frac{X_i - X_j}{h}\right) + 2n^{-1}h^{-1}K(0), \qquad (6.4)$$

where

$$K^*(t) = K^{(2)}(t) - 2K(t),$$

and $K^{(2)}(t)$ is the convolution of the kernel function with itself, to approach minus infinity as h tends to zero, forcing the degenerate choice of bandwidth, $h = 0$. As Silverman (1986, pp. 51 - 52) points out, a large number of off-diagonal terms that are

non-stochastic, *i.e.* where $X_i = X_j$ for $i \neq j$, may dominate (6.4), because $K^*(0)$ takes a constant, negative value. Silverman (1986) suggests the condition

$$m / n > \beta,$$

where m = number of pairs with $i < j$ for which $X_i = X_j$ in a sample of size n, and

$$\beta = \tfrac{1}{2}K^{(2)}(0) \Big/ \left\{2K(0) - K^{(2)}(0)\right\},$$

to determine that $M_1(h) \to -\infty$ as $h \to 0$. For the lung cancer locations, $m = 442$, $n = 978$ and $\beta = 1/6$, so that there is clearly a discretisation effect with the controls data set. It is possible to obtain a smoothing parameter by LSCV directly, if repeated values are perturbed, *e.g.* by a random displacement distributed as bivariate $N(0,\sigma^2)$, with zero correlation coefficient. This is an unsatisfactory solution, however, because the final choice of h depends on the standard deviation of the noise, as can be seen from Table 26.

σ	0.05	0.075	0.1	0.125
LSCV smoothing parameter (km)	0.107	0.136	0.159	0.176

Table 26: Smoothing parameters obtained from least squares cross-validation of South Lancashire controls with Gaussian kernels. Repeated coordinates perturbed by random deviates from $N(0,\sigma^2)$.

The bandwidth selection procedure of Diggle (1985) was derived under the assumption that a uniform kernel is to be used, *i.e.*

$$K(\mathbf{x}) = \begin{cases} \pi^{-1}, & \mathbf{x}^T\mathbf{x} \leq 1, \\ 0, & \text{otherwise}, \end{cases}$$

for the two-dimensional case. With this approach, Diggle (1990) obtained a smoothing parameter for the controls data of $h = 0.3$ km. However, the argument of Diggle (1985) cannot be used to derive a corresponding result for use with a general kernel. Instead, it is necessary to calculate the smoothing parameter, h_u, for a uniform $K(\cdot)$, and then to

employ a correction to standardise the old bandwidth by the expected squared radial distance of the new kernel; this may be accomplished by using the formula

$$h_{new} = (2c)^{-\frac{1}{2}} h_u, \tag{6.5}$$

where c is the quantity $\mathbb{E}(\mathbf{X}^T\mathbf{X})$ for non-uniform kernel. For example, the bivariate uniform kernel has $c = 1/2$, so that we recover $h_{new} = h_u$; $c = 2$ for a bivariate Gaussian kernel, giving $h_{new} = h_u/2$, as in Diggle (1990), and $c = 1/6$ for the bivariate Epanechnikov kernel.

Since the bandwidth for the controls, h_2, calculated by minimising the approximate mean square error of the intensity estimator, is equivalent to that found by LSCV for the density estimator, h_2 will be of the form

$$h_2 = k n_2^{-\frac{1}{6}}, \tag{6.6}$$

where $n_2 = 978$ and k is some constant. The cases and controls are sampled from the same density under the null hypothesis, and the corresponding k.d.e.'s will, usually, be calculated using the same type of kernel, so it would be reasonable to employ (6.6) for the cases kernel estimate, with a correction to allow for the substantially different number of observations. Thus,

$$h_1 = h_2 n_2^{\frac{1}{6}} n_1^{-\frac{1}{6}}, \tag{6.7}$$

where $n_1 = 58$, would be a suitable choice of smoothing parameter.

Data can only be collected on a finite, bounded region, A, outwith which one has no information on the behaviour of the probability density, $f(\cdot)$, of the events under consideration. Clearly, a sudden change in the character of $f(\cdot)$ just outside the boundary of A would affect the estimate, $\hat{f}(\cdot)$, calculated by some smoothing technique on a larger region, B, that completely contained A. However, the estimate calculated from A itself would not reflect this feature. Therefore, the degree of uncertainty about the quality of a density estimate increases near the boundary, and its accuracy may be, in fact, quite poor if there is some such "edge effect" present. One approach to minimising the influence of the area outside A is to use a kernel of bounded support, *e.g.* the bivariate Epanechnikov kernel (see Section 4.3.2). A second is to generalise the "wrap-around" method employed in one dimension in Section 4.2.2, by identifying the top edge of a rectangle enclosing A with the bottom edge, and the left edge with the right, to form a torus. Implementing this in practice, however, may be difficult if A has

a very irregular boundary, as seems to be the case for the South Lancashire data, and so this correction was not applied in the following analyses.

It is possible that the ISD statistic will be less sensitive to edge effects than, for example, the Scan Statistic. Under the null hypothesis, the two types of event are drawn from the same $f(\cdot)$, so the inaccuracy in one estimate near the boundary should also be reflected in the second, with the same sign and magnitude. Hence, the measurement of discrepancy between the two samples by $T_{h_1 h_2}$ should remain unaffected.

If an Epanechnikov kernel is used to reduce the influence of the boundary, the limits of integration within $T_{h_1 h_2}$ become quite complex, which, in turn, makes the analytic evaluation of the statistic very difficult. In addition, experimentation with the South Lancashire data suggested that the integrand for this problem was not sufficiently well behaved for a numerical integration procedure to be successful. Therefore, the analyses reported below obtained values of the test statistic by Monte Carlo integration (Rubinstein, 1981, pp. 115 - 121). Let

$$I(\mathbf{x}) = \left\{ \hat{f}_1(\mathbf{x}) - \hat{f}_2(\mathbf{x}) \right\}^2, \tag{6.8}$$

the integrand within $T_{h_1 h_2}$. The technique assumes that a bounding region in the (x, y) plane, B, is available, such that

$$I(\mathbf{x}) = 0, \quad \forall \mathbf{x} \notin B,$$

and that

$$0 \leq I(\mathbf{x}) \leq b, \quad \forall \mathbf{x} \in B.$$

For convenience, B is normally chosen to be a rectangle, so that $I(\cdot)$ is completely contained within a cuboid (usually referred to as an "envelope") of volume

$$V = A_B \, b,$$

where A_B is the area of B. A large number, *e.g.* $R = 10^6$, of random vectors $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, R$, are generated from the bivariate uniform distribution on B for the $\{\mathbf{x}_i\}$, and, independently, U(0,b] for the $\{y_i\}$. For each vector, a successful trial is recorded if

$$y_i \leq I(\mathbf{x}_i), \tag{6.9}$$

and then the value of the integral (interpreted as the volume under the surface $I(\mathbf{x})$) is estimated by rV/R, where r is the number of successes. For each vector, (6.9)

represents a Bernoulli trial, in which the probability of success is estimated by r/R, so an approximate 95% confidence interval for the value of the integral is

$$\frac{rV}{R} \pm 1.96 \frac{V\{r(R-r)\}^{\frac{1}{2}}}{R^{\frac{3}{2}}}. \tag{6.10}$$

6.2.2 Analysis

As noted in Section 6.2.1, Diggle (1990) obtained a bandwidth of $h_2 = 0.3$ km for the lung cancer data using a bivariate uniform kernel function. From (6.5), this is equivalent to a smoothing parameter of

$$h_2 = 0.520, \tag{6.11}$$

to three decimal places, for a bivariate Epanechnikov kernel. Correcting for the smaller sample size of the laryngeal cancer cases, (6.7) gives a corresponding bandwidth of

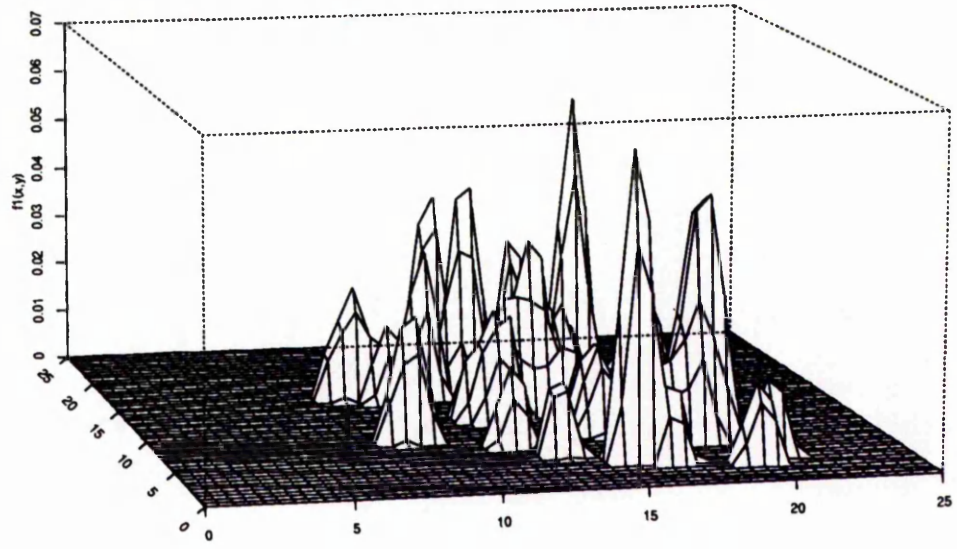
$$h_1 = 0.832. \tag{6.12}$$

Perspective plots of the kernel density estimates of the cases and controls are displayed in Figures 10(a) and 10(b), respectively, and the squared difference between the two surfaces, *i.e.* a plot of $I(\mathbf{x})$ in (6.8), is shown in Figure 11(a). The three diagrams were produced with the appropriate smoothing parameters from (6.11) and (6.12) and are displayed from the point of view of an observer to the south west of the region of interest. The surfaces were evaluated over a 51×51 grid and the x and y axes have been converted to kilometres. The case and control k.d.e.'s seem to be undersmoothed, and so the squared difference surface in Figure 11(a) is noisy. There are two pronounced spikes visible, one at approximately the location of the four cases mentioned in Section 6.1.1, and the other just to the north east of this position. If Figure 11(a) is compared to Figure 11(b), which represents the surface

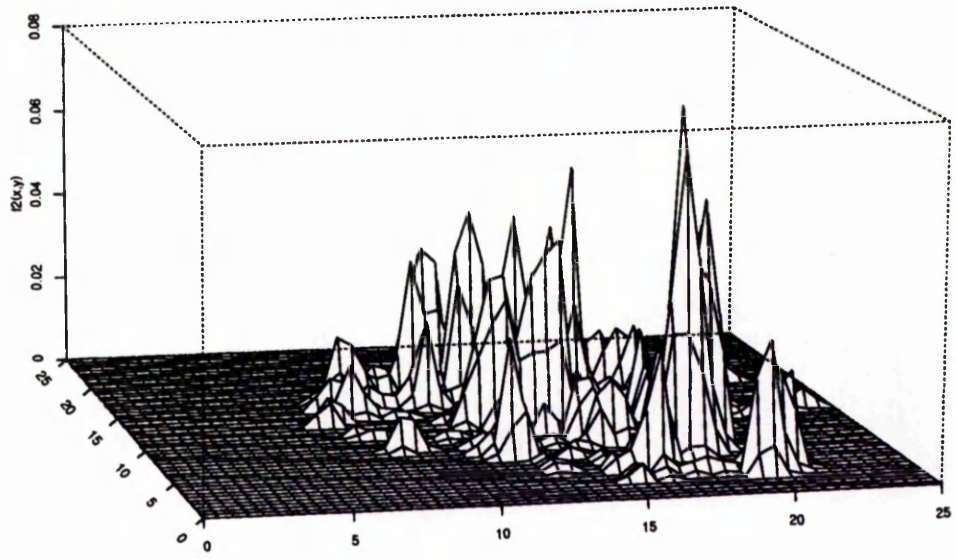
$$\left\{\hat{f}_1(\mathbf{x}) - \hat{f}_2(\mathbf{x})\right\},$$

it becomes clear that the former peak is due to a relatively greater density of cases in that area, whereas the latter is attributable to an excess of controls. With the dimensions of a suitable envelope estimated from the data used to produce Figure 11(a), the corresponding Monte Carlo estimate of $T_{h_1 h_2}$, using a sample of $R = 10^6$ random vectors, was 0.012025, with an approximate 95% confidence interval from (6.10) of

$$(0.011617, 0.012433).$$

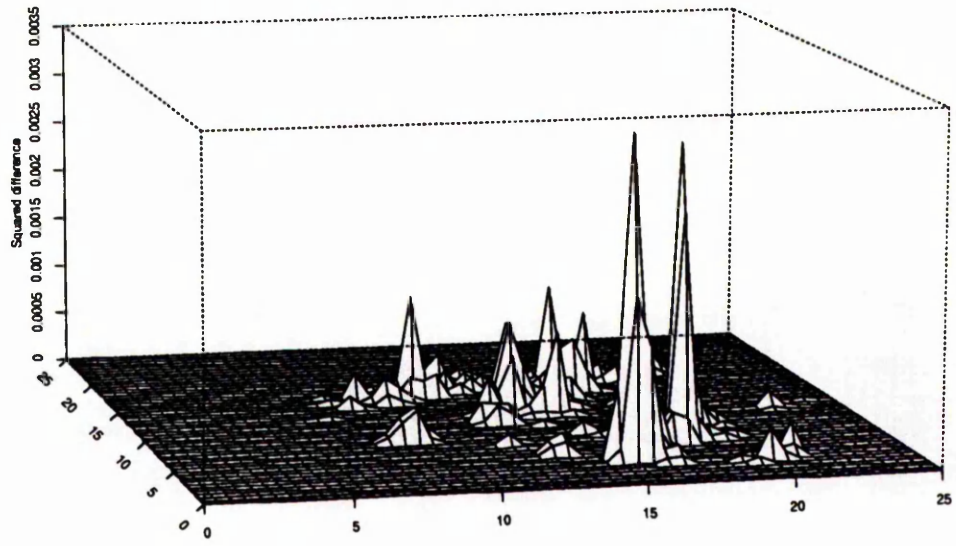


(a)

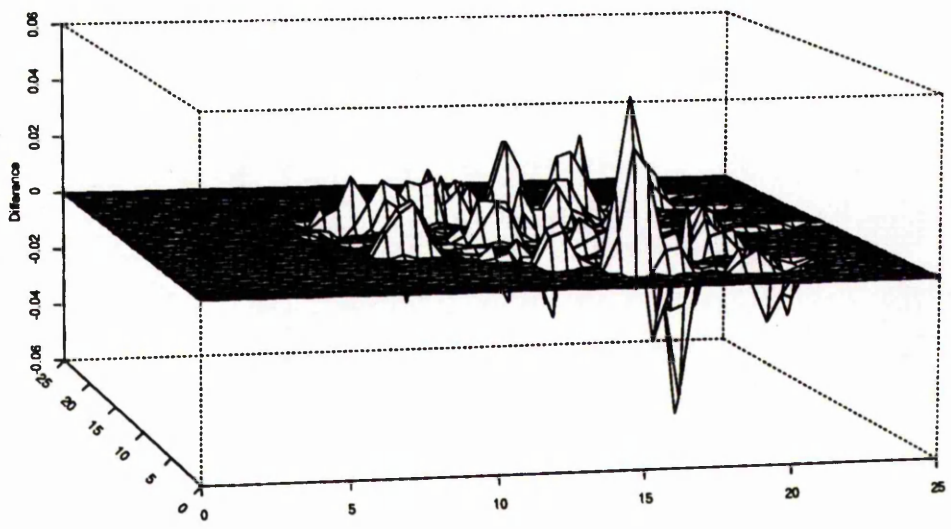


(b)

Figure 10: Kernel density estimates of (a) cases, and (b) controls, with smoothing parameters (6.12) and (6.11), respectively. Viewed from the south west.



(a)



(b)

Figure 11: (a) Squared, and (b) {case - control}, difference surfaces, with smoothing parameters from (6.11) and (6.12). Viewed from the south west.

The significance of this observed value of the test statistic was assessed by a Monte Carlo test implemented with the smoothed bootstrap, a technique outlined (for more general contexts) in Silverman (1986, pp. 144 - 147) and discussed in more detail by Silverman and Young (1987). The principle of the smoothed bootstrap is that, instead of sampling from the observed data with replacement, new observations are created by sampling from the relevant kernel estimate. If the simulation is to be based on a data set, D, containing n observations, then sampling from the k.d.e. calculated with smoothing parameter h is accomplished by creating the bootstrap data set

$$\{\mathbf{Z}_i + h\varepsilon_i; i = 1,\ldots,n_b\},$$

where the $\{\mathbf{Z}_i\}$ are sampled with replacement from D, the $\{\varepsilon_i\}$ are independent random deviates sampled from the kernel p.d.f., $K(\cdot)$, and n_b represents the size of bootstrap sample required. For the South Lancashire data, n_b is 58 or 978.

This procedure was preferred to the standard bootstrap, because, in general, the "with replacement" resampling scheme of the latter technique produces simulated data sets containing repeated values. The set of original lung cancer coordinates contains duplicate points before any selection is carried out, as discussed in Section 6.1.1. Therefore, it is possible that the standard method would increase the frequency of identical observations, making the kernel estimates more noisy.

The issue of whether to replace D, above, by the pooled sample or one of the individual samples for an analysis with the ISD statistic was discussed in Section 5.4. The sample of laryngeal cancer locations is too small to be used on its own, so the choice lies, effectively, between the 978 controls and the pooled sample of 1036 observations. There may be a small advantage in terms of power if resampling is carried out from the controls only, since this may better approximate the null hypothesis of no clustering. Therefore, the significance tests reported below were performed with this choice of D.

The Monte Carlo test with the smoothed bootstrap is similar in other repects to the type of procedure described previously: a test statistic is calculated for each set of 58 simulated cases and 978 simulated controls, and a p-value for the observed statistic is based on its rank when it is added to the set of bootstrapped values. For $T_{h_1 h_2} = 0.012025$, as given above, the corresponding p-value was 0.68, when 99 bootstrap replications were employed. In this test, and in subsequent analyses, each bootstrapped statistic was calculated with the same bandwidths as the observed value.

It is interesting to compare Figures 11(a) and 11(b) to Figure 12, which is a plot of the type proposed by Bithell (1990), previously discussed in Section 1.2.2. The surface drawn is

$$P(\mathbf{x}) = \frac{\hat{\rho}(\mathbf{x})}{1+\hat{\rho}(\mathbf{x})}, \tag{6.13}$$

where

$$\hat{\rho}(\mathbf{x}) = \frac{\hat{f}_1(\mathbf{x})+c}{\hat{f}_2(\mathbf{x})+c},$$

and $\hat{f}_1(\cdot)$ and $\hat{f}_2(\cdot)$ are kernel estimates for the cases and controls, respectively. The constant c is included to prevent division by zero and to ensure that $\hat{\rho}(\cdot) \to 1$ and $P(\cdot) \to \frac{1}{2}$ outside the area in which data were collected; in Bithell (1990), it was chosen to be 0.1 × the height at the origin of the kernel in use. This formula gives a value of $c = 0.064$ for the bivariate Epanechnikov kernel employed here, which has an origin height of $2\pi^{-1}$. The surface in Figure 12 is noisy, due to the variability of the two kernel estimates that were used in its construction; however, the two strongest features seem to correspond with those observed in Figures 11(a) and 11(b), with an excess of laryngeal cancer at the location of the four case "cluster", and a control excess to the north east.

To investigate sensitivity to the values of the smoothing parameters, the significance test was repeated for kernel estimates of the cases and controls that were calculated with bandwidths of the form

$$h_i^{(new)} = \mu h_i, \tag{6.14}$$

for i = 1, 2 and μ = 1.5 (0.5) 5.5. The h_i terms in (6.14) are the original smoothing parameters from (6.11) and (6.12). Figures 13, 14 and 15 represent the laryngeal and lung cancer k.d.e.'s, together with the corresponding squared difference surfaces and plots of (6.13), for μ = 2, 3 and 4. Comparison of the two kernel estimates suggests that the density of laryngeal cancer near the four case group of particular interest is greater in relation to that of the surrounding area than the density of controls in the same region. However, as μ increases, the prominence of the peak at that location in the squared difference surface decreases, until, by μ = 4, the most significant feature is the control excess spike to the north east. This behaviour is echoed by the sequence of plots of (6.13). The relative height of the peak due to the "cluster" is reduced as μ increases, whereas a deficit of cases nearby is suggested quite strongly.

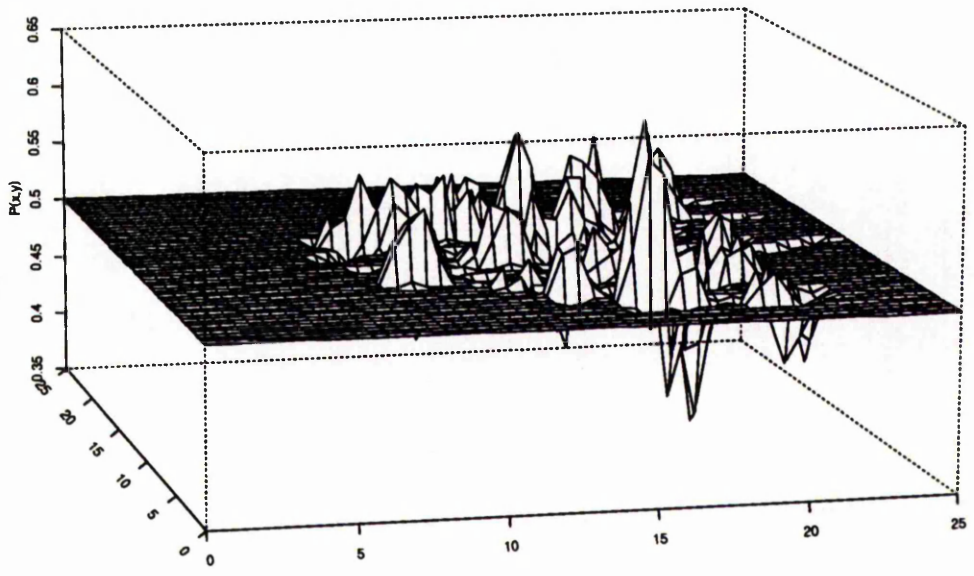
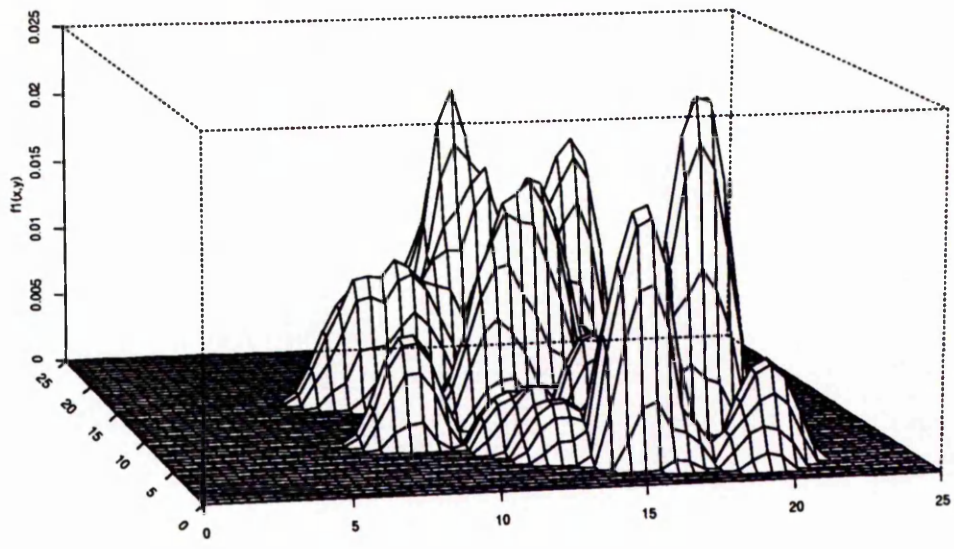
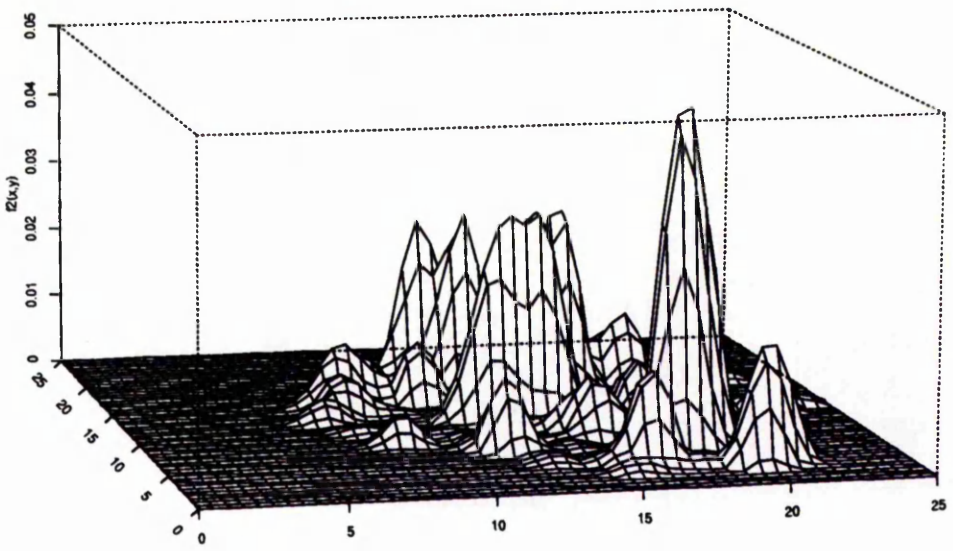


Figure 12: Plot of (6.13), due to Bithell (1990), with smoothing parameters from (6.11) and (6.12). Ratio constant, c, equals 0.064. Viewed from the south west.

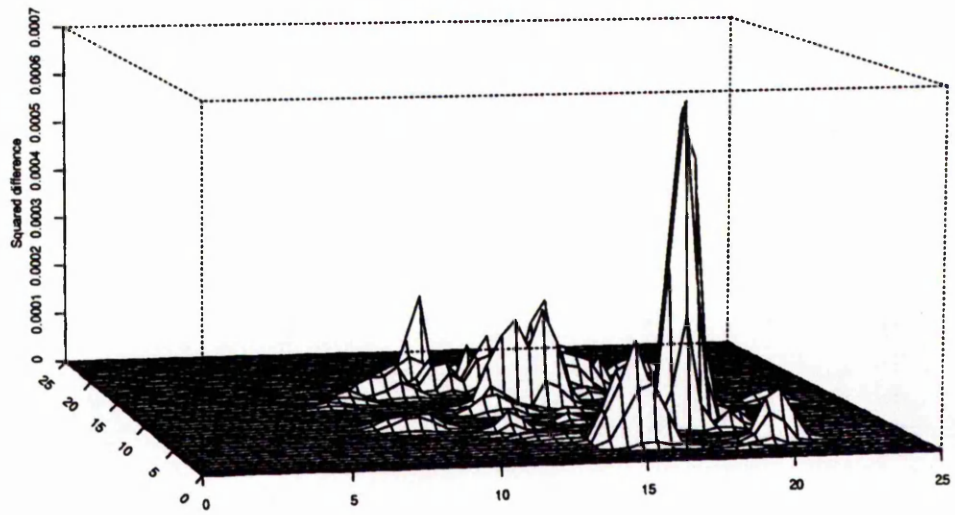


(a)

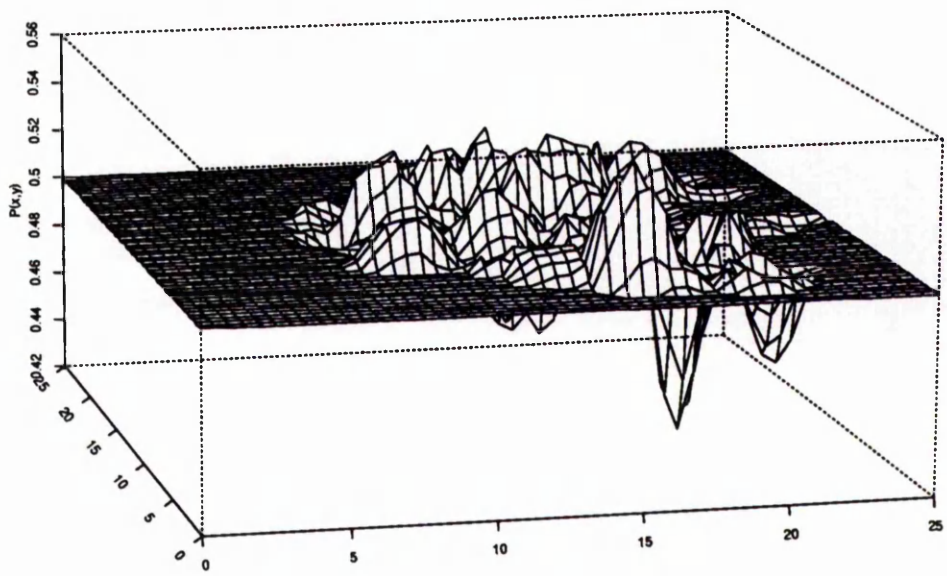


(b)

Figure 13: Kernel density estimates of (a) cases, and (b) controls for South Lancashire data. Bandwidths of form (6.14), with $\mu = 2$. (Continued overleaf)

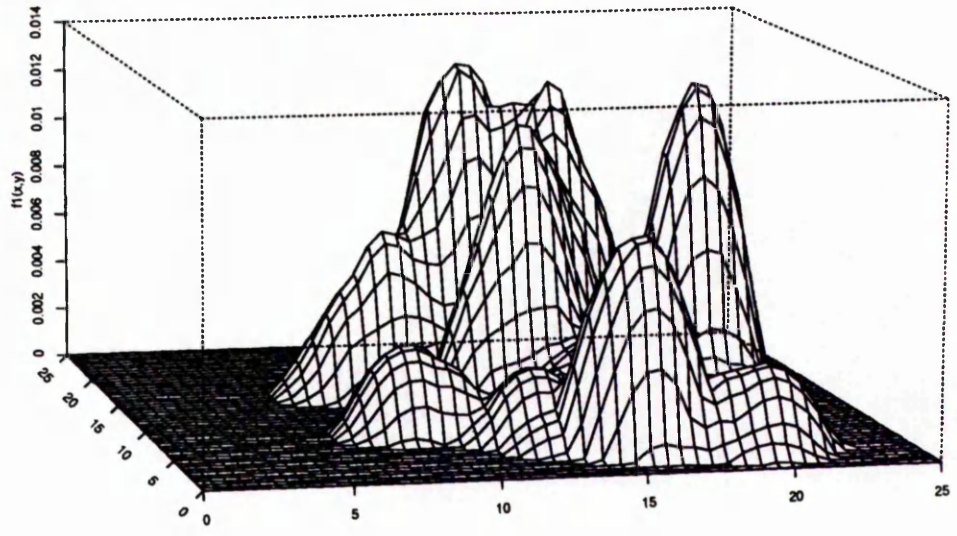


(c)

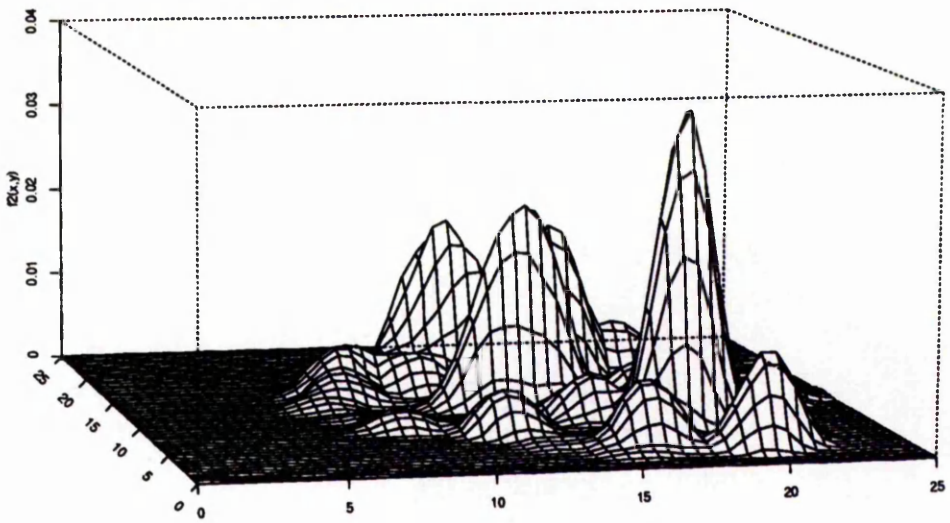


(d)

Figure 13: (Continued) (c) Squared difference surface, and (d) plot of (6.13), for South Lancashire data. Bandwidths of form (6.14), with $\mu = 2$.

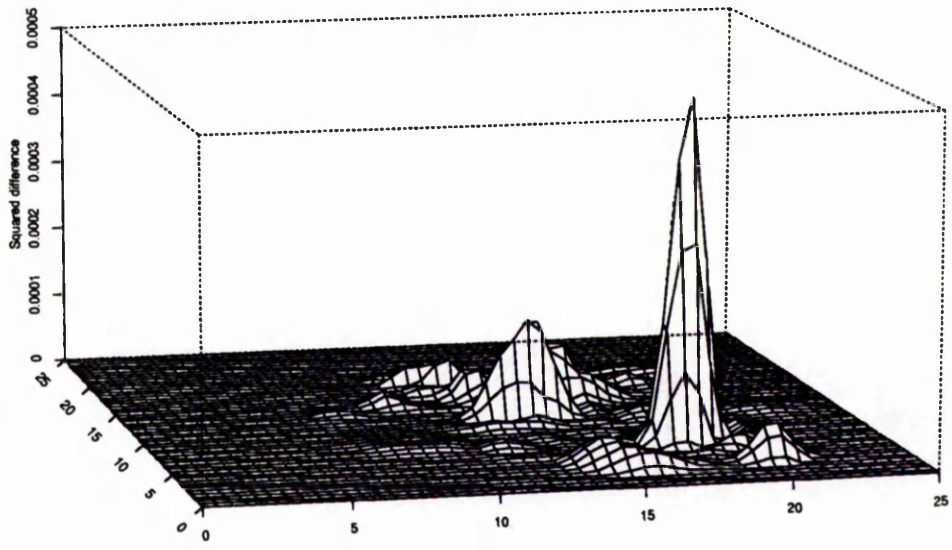


(a)

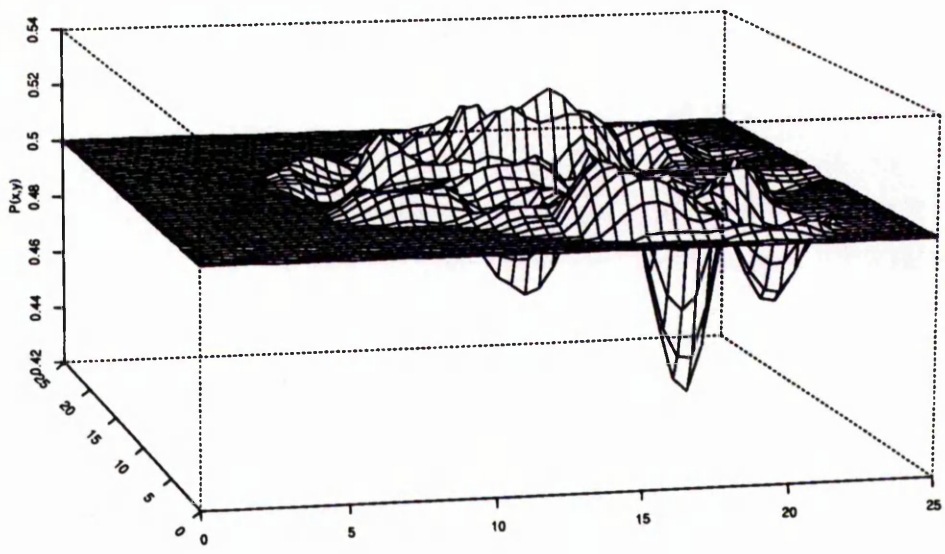


(b)

Figure 14: Kernel density estimates of (a) cases, and (b) controls for South Lancashire data. Bandwidths of form (6.14), with $\mu = 3$. (Continued overleaf)

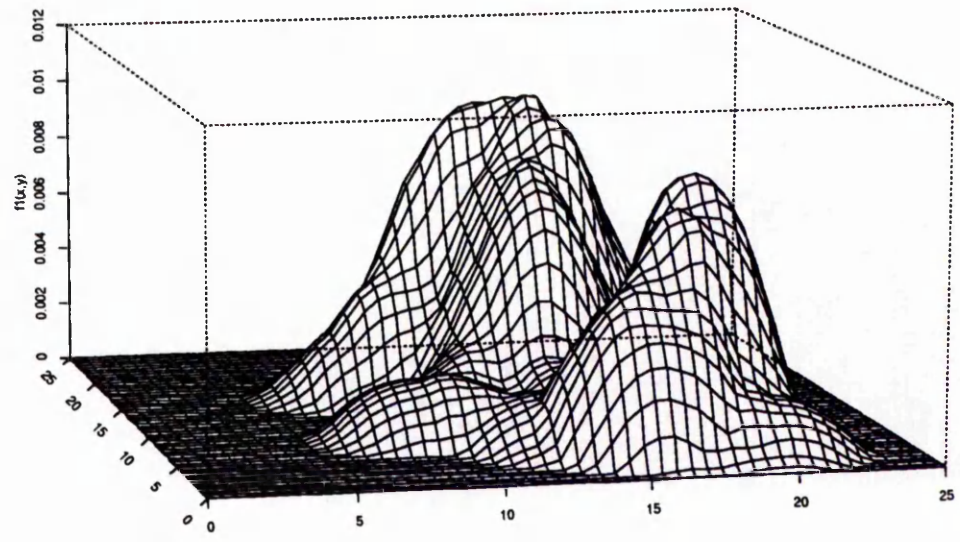


(c)

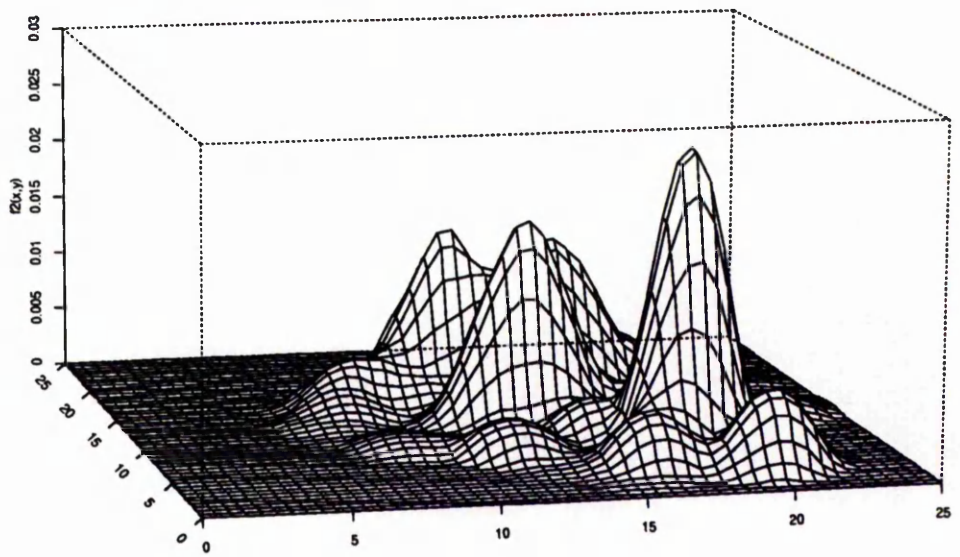


(d)

Figure 14: (Continued) (c) Squared difference surface, and (d) plot of (6.13), for South Lancashire data. Bandwidths of form (6.14), with $\mu = 3$.

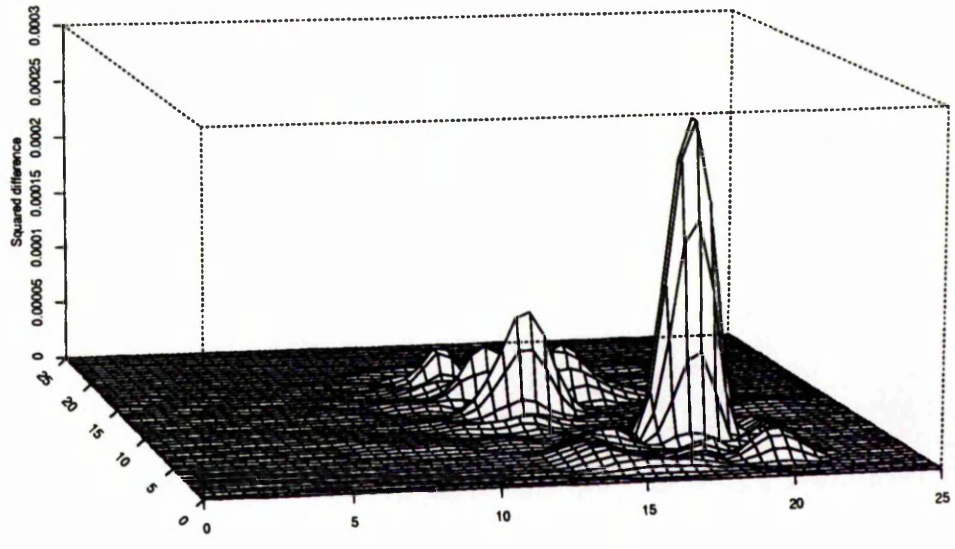


(a)

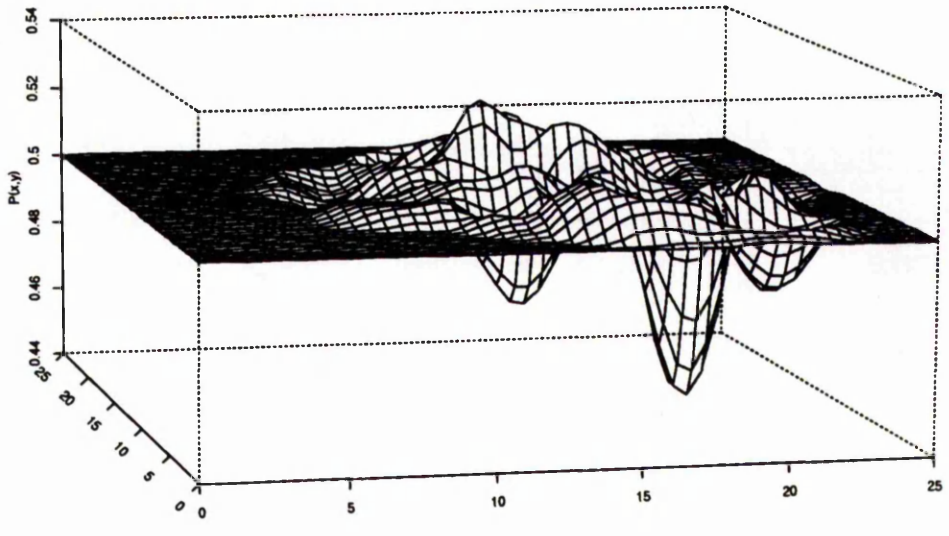


(b)

Figure 15: Kernel density estimates of (a) cases, and (b) controls for South Lancashire data. Bandwidths of form (6.14), with $\mu = 4$. (Continued overleaf)



(c)



(d)

Figure 15: (Continued) (c) Squared difference surface, and (d) plot of (6.13), for South Lancashire data. Bandwidths of form (6.14), with $\mu = 4$.

Table 27 lists the p-values for the tests corresponding to each value of μ, with each test comparing the observed statistic to 99 simulated values that were obtained by the smoothed bootstrap approach described above. A trend towards significant p-values with increasing μ is evident, such that for $\mu \geq 4$, the distribution of cases appears to be substantively different to that of the controls. The sample plots of Figures 13 to 15 seem to indicate that these results are due, mainly, to a lack of cases to the north east of the group that originally gave rise to concern; *i.e.* there seems to be some evidence of negative clustering (as defined in Marshall (1991) and Section 5.6).

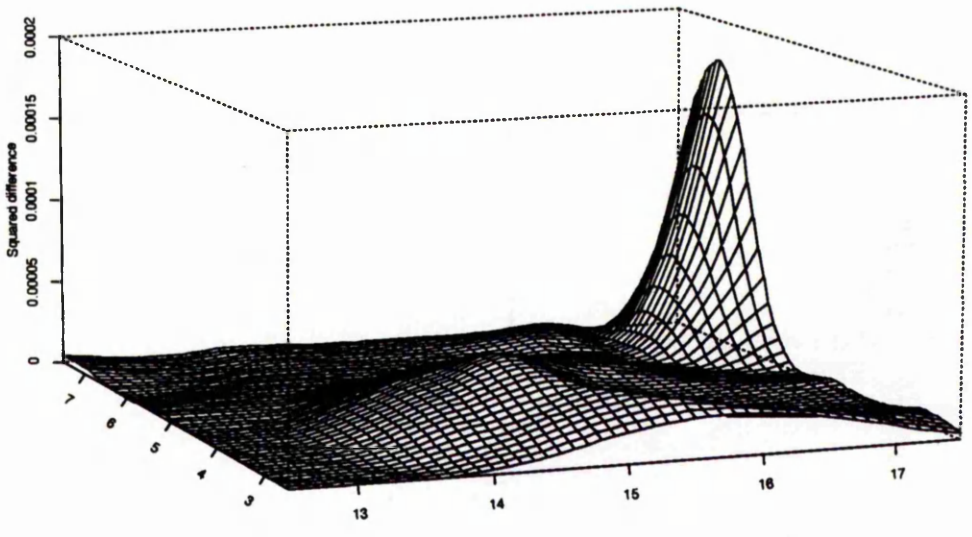
μ	1	1.5	2	2.5	3
p-value	0.68	0.72	0.57	0.25	0.22
μ	3.5	4	4.5	5	5.5
p-value	0.06	0.01	0.03	0.02	0.05

Table 27: Results for smoothed bootstrap significance tests of the South Lancashire data, using bandwidths from (6.14) with different values of μ. Monte Carlo test based on 99 replications.

6.2.3 Further Investigation

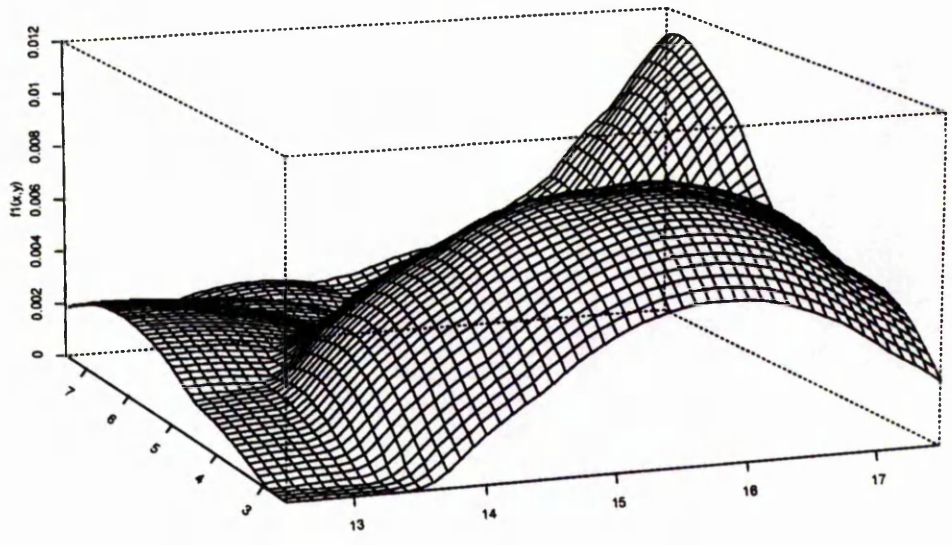
The results of Diggle *et al* (1990) suggest that the scale of association with the incinerator is quite small, extending over a distance of only about three to four kilometres from the facility. The ISD statistic measures the difference between the case and control densities over the whole 25 × 25 km region, and, thus, may have lower power against alternatives with a highly localised clustering effect. Therefore, it may be necessary for $T_{h_1 h_2}$ to be focussed more closely on the region in which clustering is occurring for the effect to be detected. To illustrate this, the integrated squared difference between the two kernel estimates, based on the full data with $\mu = 3$, was calculated for a reduced area of 5 × 5 km around the group of four cases, namely [35250, 35750] × [41250, 41750] or, on the kilometre scale, [12.5, 17.5] × [2.5, 7.5].

The relevant sections of the two k.d.e.'s and the squared difference surface are shown in Figure 16. With the same type of significance test as before, the original p-value of 0.22 was reduced to 0.08, which is much closer to being conventionally significant. Therefore, concentrating the ISD analysis on the particular region of concern does provide more evidence of an anomalous distribution of cases, although the most important contribution to this result may still be from a lower than expected density of laryngeal cancer in the north east of this area. It should be noted, of course, that the *post hoc* selection of a sub-region in this way is not valid inferentially, but it is useful for the purposes of an example.

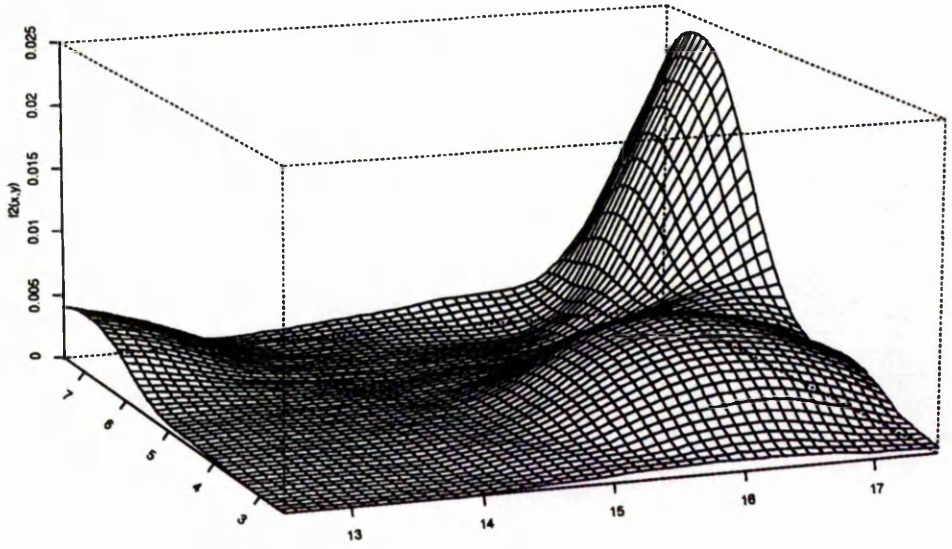


(a)

Figure 16: (a) Section of squared difference surface for South Lancashire data within 5 × 5 km sub-region. Bandwidths from (6.14), with $\mu = 3$. (Continued overleaf)



(b)



(c)

Figure 16: (Continued) Sections of (b) case, and (c) control, k.d.e.'s for South Lancashire data within 5 × 5 km sub-region. Bandwidths from (6.14), with $\mu = 3$.

6.3 Scan Statistic

To demonstrate the application of the Scan Statistic to a real set of data, the geographical distribution of laryngeal cancer for the South Lancashire area between 1974 and 1983 was analysed, with the correction for a non-uniform null density based on a kernel estimate of the p.d.f. of cancer of the lung in the same region. The approximation of the exact equation, (4.10), for the square dimension, δ, was attempted, as in Chapter 4, by numerical integration and Taylor expansions of first, second and third orders, *i.e.* by using the polynomials (4.12), (4.13) and (4.14). The smoothing parameter for the k.d.e. was chosen to be

$$h_2 = 1.819,$$

which is equivalent to $\mu = 3.5$ in (6.14). It was hoped that a larger bandwidth, resulting in a smoother estimate, would improve the reliability of the square-size calculation.

Section 4.4 describes a Monte Carlo procedure for assessing the significance of an observed value of the Scan Statistic. For the South Lancashire data, transformed to a kilometre scale, each replication of the simulation requires 58 events to be sampled from the bivariate uniform distribution on $(0, 25] \times (0, 25]$, which represents the null hypothesis if the correction is used. Table 28 displays empirical null distributions, based on 1000 replications, for the Scan Statistic for a sample of size 58, given scanning squares with sides of length 2, 4, 6 and 8 km. These results also serve to indicate a feasible range of values of the statistic for the laryngeal cancer data: observed test statistics either much smaller or much larger than the simulated values would have to be treated with caution.

With $A = B = 25$, the constant d in (4.10) was chosen, initially, to be in the range $d = 1\ (1)\ 10$. The analyses based on first and second order Taylor expansions experienced the same type of numerical difficulties that were reported in Chapter 4. The observed values of the Scan Statistic for each choice of d in these two cases are listed in Table 29, along with the dimensions of the square inside which the statistic was observed. Especially for the smaller d, the values obtained were much greater than those found in Table 28. In addition, the corresponding square dimensions seemed to be too large, and the consistency across the range of d was curious.

Values of the Scan Statistic	Frequency			
	$d = 2$	$d = 4$	$d = 6$	$d = 8$
2	3			
3	419			
4	485	1		
5	88	185		
6	5	459	1	
7		268	35	
8		67	245	
9		14	347	1
10		5	226	36
11		1	100	168
12			34	265
13			10	242
14			2	149
15				85
16				41
17				8
18				5

Table 28: Empirical null distributions of the Scan Statistic for the South Lancashire data, with $d \times d$ scanning squares. From 1000 simulations, events being generated from a bivariate uniform distribution.

d	First Order, (4.12)		Second Order, (4.13)	
	Scan Statistic	Square Size	Scan Statistic	Square Size
1	11	11.5909	20	6.3097
2	14	18.3940	20	6.3106
3	14	18.4051	20	6.3121
4	22	8.1390	20	6.3143
5	22	8.1735	20	6.3171
6	22	8.2143	20	6.3204
7	22	8.2609	20	6.3244
8	22	8.3127	20	6.3289
9	22	8.3689	20	6.3340
10	22	8.4291	21	6.4137

Table 29: Observed Scan Statistics for the South Lancashire data, given population corrections (4.12) and (4.13). The size of the scanning window inside which each statistic was located is also listed.

These two factors suggested that the algorithm was consistently overestimating the size of window required and, thus, inflating the value of the Scan Statistic. The following example supports this observation.

Example

For the first order correction, (4.12), with $d = 1$, the Scan Statistic was found within a square with sides of length 11.5909 km in the region

$$[2.4091, 14.0000] \times [3.6091, 15.2000].$$

A square of this size was not typical of those calculated during the scanning procedure, as can be seen from the following stem and leaf diagram, which plots all the values of δ that were estimated from (4.12) by the FORTRAN program used to search for the Scan Statistic:

Stem unit = tenths; leaf unit = hundredths

0	0000000000
1	7888
2	000111233344445667777777888999999999
3	001111111122222333333445556666777778
4	00012222455567899
5	06778899999
6	000011111235688899
7	12222333333455567789999999
8	001233444677889
9	068
10	18
11	
12	9

High : 2.1097, 2.7292, 6.9901, 11.5909, 11.5909

Number of observations = 183
Median = 0.4227
First Quartile = 0.2931
Third Quartile = 0.7262

If the square is of the correct size, it should bound a volume under the controls k.d.e. of $d^2/AB = 0.0016$. However, by using the type of Monte Carlo integration procedure described in Section 6.2.1, with 10^6 replications, the actual volume was estimated to be 0.166895, with an approximate 95% confidence interval of (0.164844, 0.168946). Hence, the given value of the Scan Statistic is clearly erroneous.

At a number of locations, the first and second order analyses were unable to calculate square dimensions, *i.e.* there were no real, positive zeros of the polynomials (4.12) or (4.13). This type of problem was also encountered in Chapter 4.

For the third order method, analyses with $d \geq 4$ suffered from the same numerical difficulties described above, although with a lower frequency. The results in Table 30, for smaller d, seem to be much more reasonable. The only problem encountered was a small set of coordinates at which no estimate of δ could be calculated, for $d = 3$. The

Scan Statistics observed for the different values of d in Table 30 seem typical of those expected under the null hypothesis and, therefore, provide no evidence of clustering of laryngeal cancer within the Chorley and South Ribble Health Authority district between 1974 and 1983.

d	Scan Statistic	Square Size	p-value
0.25	2	0.01539	0.474
0.5	2	0.03078	0.931
0.75	2	0.15154	0.999
1	2	0.20606	1.000
2	4	0.40413	0.578
3	4	0.67287	0.998

Table 30: Observed Scan Statistics for the South Lancashire data, with population correction (4.14), listed with corresponding square dimensions and p-values.

6.4 Discussion

The conclusion reached by Diggle (1990) and Diggle *et al* (1990) was that there appeared to be a local increase in the risk of laryngeal cancer in the area around the incinerator, but that the magnitude and extent of the effect was very difficult to quantify, because of the relatively small number of cases involved. Most of the evidence for clustering was provided by the aggregation of four cases near the coordinates (35600,41400), and the results were sensitive to the random deletion of either one or two points from this group. With smoothing parameters greater than those obtained by least squares cross-validation (or, equivalently, the method of Diggle (1985)), the analysis using the ISD statistic provided some evidence of a difference between the case and control densities, although this seemed to be attributable to a region with a greater relative density of controls, rather than to the clustering of laryngeal cancer. Plots of the surface $P(x, y)$ from (6.13), due to Bithell (1990), supported this observation. The analysis based on the Scan Statistic, with the null density correction of Chapter 4 employing third order Taylor expansions, found no

evidence of clustering for any of the choices of d. However, only one choice of smoothing parameter was used with the kernel density estimate of the lung cancer data, and the results may be unreliable, because of the inaccuracies in the correction that were noted in Chapter 4 and in Section 6.3. Since it has been demonstrated that the method may overestimate the size of a scanning window, δ, or that it may find no value for δ at all, it is also possible that some square dimensions could be underestimated, an error that is more difficult to detect, and which might cause the statistic to be reduced and the power of the test to be lowered. In general, the Scan Statistic would seem to be a useful tool for the investigation of spatial clustering, but the correction method used here is unsatisfactory for most purposes, and should be replaced by a method such as the one described in Section 4.4.

The use of a series of smoothing parameters, each being a multiple of the value found by the method of Diggle (1985), with the ISD statistic was prompted by the observation that the individual case and control kernel estimates were quite noisy. In strict terms, therefore, the inference for the group of results so obtained is highly complex, since the approach was formulated *a posteriori* and led to a set of multiple comparisons. However, the results are useful for investigating the methodology, and seem to indicate that the smoothing parameters employed in the calculation of $T_{h_1 h_2}$ should be larger than those that would be used simply to estimate the density from which data points were sampled. In other contexts, this type of phenomenon has been discussed in the literature for quantities similar to $T_{h_1 h_2}$. For example, Jones and Sheather (1991) consider the nonparametric estimation of the functionals

$$\int \left\{ f^{(m)}(x) \right\}^2 dx, \quad m = 0,1,2,\ldots, \tag{6.15}$$

where a superscript "(m)" represents an mth order derivative, by replacing $f(\cdot)$ with its kernel density estimate. The authors note that the smoothing parameter suitable for estimating (6.15) is different to that appropriate for estimating $f(\cdot)$. It is possible that the same may be true for the ISD statistic, since $T_{h_1 h_2}$ is qualitatively similar to (6.15) when $m = 0$.

A standard property of a k.d.e. is that the value of the smoothing parameter controls the relationship between the bias and variance of the estimate. With small values of h, bias is low, but variability is high and, thus, the variance of the bootstrapped ISD statistics, which depend upon the control estimate, might also be increased. If a sufficient number of large, simulated values of $T_{h_1 h_2}$ were caused by this noise, the power of the test could be reduced. As the smoothing parameters increase, the biases of both $\hat{f}_1(\cdot)$

and $\hat{f}_2(\cdot)$ increase, but the variances of the estimates are reduced, which could improve the sensitivity of the test. If the two estimates were biased to approximately the same order, one might expect some form of cancellation effect (since we are working with the differences between the kernel estimates and not the estimates themselves), and that there might be some advantage (in terms of power) in choosing smoothing parameters that were artificially high. In the current example, however, a k.d.e. of the laryngeal cancer data will have considerably greater bias than that of the lung cancer data, because of the discrepancy in sample sizes, and, thus, the difference in magnitude of the two bandwidths. This may explain why the (graphical) evidence for an excess of cases is reduced as the smoothing parameters are increased, so a genuine cluster of laryngeal cancer should not be ruled out on the basis of these new results.

The calculation of $T_{h_1 h_2}$ and the associated test of significance, as described in Section 6.2.1, is computationally demanding. The use of an Epanechnikov kernel to minimise edge effects means that the test statistic must be calculated by Monte Carlo integration, which requires a large number of function evaluations and the generation of many pseudo-random deviates. Each simulation step requires the procedure to be carried out again, and requires two bootstrap samples to be drawn from the controls. The Epanechnikov kernel is normally chosen in preference to, for example, the Gaussian density if computational efficiency is an important consideration, because the former p.d.f. may be evaluated more rapidly. Paradoxically, it may prove to be more efficient, for this context, to employ a Gaussian kernel, so that $T_{h_1 h_2}$ may be calculated analytically, removing the need for the time-consuming Monte Carlo procedure. The penalty to be paid for this substitution would be the loss of probability mass from the region within which data were collected, but this might be thought to be of less importance than the reduction in the computational workload that could be achieved.

The semiparametric method of Diggle (1990) seems to be more effective than either the Scan Statistic or $T_{h_1 h_2}$ for the South Lancashire data set. This may be because the clustering effect is limited to the immediate vicinity of the incinerator site, which forms the focus of the model of intensity, (6.1). Global measures may be more appropriate in other situations, such as the detection of spatial clustering on a larger scale or in the exploratory stages of an investigation. If the study is required to exclude the location of an hypothesised point source, so that the pattern of events is assessed as if no prior knowledge is available, or if the application is the regular monitoring of a large area for anomalous event patterns, then a statistic such as $T_{h_1 h_2}$ or the Scan Statistic would, in fact, be required by design.

CHAPTER 7

CONCLUSIONS

7.1 General Considerations

The work reported in this thesis investigates ways of examining the geographical distribution of locations of events for evidence of a tendency to aggregate into clusters. This type of inferential task is commonly undertaken in studies of leukaemias, lymphomas and other types of malignant disease, but other fields, *e.g.* geography and ecology, may give rise to spatial point patterns for which the detection of spatial clustering is important. Some types of analysis used for this purpose estimate the null distribution of events from data that are not particularly suited to the task, or are based on counts within arbitrary sub-regions. The effect of the latter feature is, effectively, to discretise a problem that is actually continuous, and to make the results of the analysis dependent on the particular spatial distribution of the sub-regions.

To avoid these difficulties, the methods discussed here assumed that the exact geographical coordinates of the events were available, and based the analyses on estimates of the underlying p.d.f.'s obtained by smoothing techniques, in particular nonparametric kernel density estimation. Samples of controls were used to represent the distribution of events in the absence of any clustering effect. Suitable sampling frames and schemes with which to generate these subsidiary observations will vary from application to application and, therefore, have not been discussed in great detail. However, in practice, considerable time and effort will be required to ensure that a representative sample is drawn, and this will form an important part of the organisation of a study.

A number of other problems that could be encountered when studying spatial clustering were discussed in Chapter 1. These included the importance of the relative size of clusters and study region, the formulation of hypotheses and boundaries *a posteriori* and the low power of some statistical methods that results from the small sample sizes found in typical applications. It is likely that these difficulties will be common to all investigations, regardless of the methodology employed, and that it will not, in general, be possible to prevent them from having some influence on the analysis.

7.2 The Scan and Integrated Squared Difference Statistics

The Scan Statistic, discussed in Chapters 2 and 3, is usually employed to detect clustering in a single dimension, such as time. However, it may also be generalised to two dimensions, to provide a simple method of investigating the spatial distribution of event locations. Existing bounds for the upper tail probability of the statistic (Naus, 1965b) are unsuitable for significance testing, because they lack sharpness for ordinates of magnitudes appropriate for that purpose. Therefore, the most tractable approach for inference is to use simulation, by employing Monte Carlo significance tests.

Empirical investigations of the power of the Scan Statistic suggested that the probability of detecting a genuine cluster would be maximised by setting the size of the scanning window to approximately that of the cluster itself, which confirmed a result of Wallenstein and Neff (1987). In practice, of course, the geographical extent of the cluster will be unknown, so there will be little information to guide the choice of the scanning square's magnitude. A possible alternative would be to carry out significance tests for a range of values of d, where d is the length of one side of the window, and to apply a correction for multiple testing at each stage. This procedure could, perhaps, be interpreted as a type of diagnostic tool, for which the choices of d leading to significant results would indicate the scale at which clustering might be occurring. Although this approach would be appropriate for exploratory data analysis, for the purposes of inference it would seem to be preferable to find some objective method of choosing an optimal size of scanning window before any analysis was undertaken; however, it is not clear how this could be achieved.

If the distribution of events in the region of interest is non-uniform under the null hypothesis, the number of observations within the scanning window will be large in regions of high density and the test will tend to be significant, regardless of the presence or absence of any genuine clustering. The correction discussed in Chapter 4 alters the size of the window so that the Scan Statistic is calculated properly, and hypothesis testing may proceed as if the observations were drawn from a uniform density. Thus, simulation or approximation to the distribution of the statistic is simplified considerably. The precision of the method by which the correction procedure is implemented will rely, to a great extent, on the accuracy of the kernel estimate that is calculated from the control sample, and, therefore, the choice of the corresponding smoothing parameter and sample size will be important. The success of the method used in Chapter 4, and again in Chapter 6, was limited by the accuracy of the Taylor expansion applied to the estimator, which required either the target size of the scanning

window to be very small or for terms of higher order to be retained in the approximation, if the error was to be reduced to acceptable levels. This would place too great a constraint on the use of the correction procedure with the Scan Statistic in practice, and so other implementations, such as the one described in Section 4.4, should be investigated, with a view to accuracy, simplicity and reliability.

The integrated squared difference (ISD) statistic of Chapter 5 is a measure of the agreement between two probability density functions and, thus, should detect areas in which the density of events is either higher or lower than would be expected from the behaviour of the population, represented by the sample of controls. Therefore, it may have lower power against clustering (in the usual sense of an increase in "risk") than a statistic that is defined, explicitly, to search for this feature. However, the ISD statistic should be useful for more general applications that require the comparison of two spatial patterns of events, and so it does merit consideration. Section 5.2 demonstrates that the asymptotic distribution of $T_{h_1 h_2}$ is normal, by considering central limit theorems for certain random variables and U-statistics, but the results of a simulation study suggest that, even for quite large sample sizes, the exact distribution is clearly non-normal. Therefore, a Monte Carlo implementation of the bootstrap seems to be the most feasible approach to hypothesis testing.

We have remarked previously that the quality of the individual estimates of the two probability density functions is not of primary interest. However, Chapters 5 and 6 suggest that this issue will, nevertheless, be of some concern, because there appear to be two possible routes by which the accuracy of the kernel estimates may affect a test of significance undertaken with the ISD statistic. First, if the two bandwidths employed in calculating $T_{h_1 h_2}$ undersmooth one or both of the kernel estimates, the resulting noise may obscure a genuine difference between the two distributions. Therefore, it may be necessary to increase the magnitudes of the smoothing parameters, so that the variability of the k.d.e.'s is reduced. Secondly, the analysis of the South Lancashire data suggests that the statistic could fail to detect a small scale cluster if the associated increase in density of the cases is masked by the bias of the corresponding kernel estimate, or if the difference in magnitude of the biases of the two estimates reduces the prominence of the peak in the squared difference surface in the region of the aggregation. Further investigation is required to determine more precisely the influence of the variability and bias of the k.d.e.'s on the power of the statistic, and how such effects may be minimised. In the case of the latter problem, for example, it may be appropriate to ensure that both kernel estimates have the same bias under the null hypothesis, which would require the two smoothing parameters to be of equal value.

The above considerations indicate that it would be desirable to find some objective method of selecting smoothing parameter values for use with the ISD statistic. Existing methods for choosing bandwidths for the purpose of estimating a single density, such as least squares cross-validation (Bowman, 1984), will probably be unsatisfactory in our context, because it will be necessary for the procedure to consider the behaviour of both estimates simultaneously, so that variability is minimised, and the appropriate relationship between the two biases is maintained. The results of Jones and Sheather (1991) offer a possible line of enquiry, since the authors are interested, explicitly, in estimating the integral of a squared function, rather than the function itself.

Chapter 5 also proposes a modified integrated squared difference statistic, T, for which the two smoothing parameters are taken to be equal and fixed, *i.e.* independent of sample size. An asymptotic argument and empirical results show that T has greater power than $T_{h_1 h_2}$ for a number of clustering alternatives. This approach seems attractive, because a procedure for calculating suitable bandwidths is no longer required and, since the two parameters are equal, the kernel estimates will have the same bias under H_0, so T may be less susceptible to the second type of problem described above. However, tests of significance employing T may still be affected by kernel estimates with large variability, so that a statistic with random smoothing parameters may prove to be a better choice in certain cases.

A number of individual problems with the Scan and ISD statistics have been identified above. A further point that may have to be considered for both tests is the influence of the boundary of the region in which data were collected on the required kernel estimates. It would be hoped that the ISD statistic would not be affected to any great extent by the absence of information outside this area, because both estimates should be in error by approximately the same amount, and the test is concerned with the detection of differences between the two densities. However, it would be desirable for this to be confirmed in some way, *e.g.* by a simulation study. The Scan Statistic will normally require the correction of Chapter 4 when it is applied in practice, and so the analysis will depend on a single kernel estimate, that of the control sample. Hence, inaccuracies in the k.d.e. near the boundary would reduce the precision of the calculation of scanning window magnitudes nearby, which could alter, incorrectly, the value of the Scan Statistic, and affect the results of a test of significance. To avoid this, calculation of the kernel density estimate could, perhaps, be based on a sample of controls from a region much larger than, but containing, the one from which cases were obtained. The analysis would, however, use only that part of the estimate falling within the original area. A more complex alternative would be to incorporate directly into the estimation

procedure some form of edge-correction that would allow for a region with an irregular boundary.

The power of the two techniques described in this thesis to detect spatial clustering in the vicinity of a prespecified point, thought to have some influence on the incidence of the events under consideration, will be lower than that of methods designed for that specific application, such as the techniques proposed by Diggle (1990) or Stone (1988). The Scan and ISD statistics will be better suited to exploratory investigations or to studies in which, for example, the location of the causative agent is not clear and one wishes to avoid the bias that would result from choosing the wrong coordinate for the focus of modelling or testing. A third situation for which these methods would be particularly appropriate is the regular surveillance of one or more areas in a proactive cluster detection programme, in which there would be, normally, no specific fixed location that was hypothesised to be aetiologically relevant.

Chapter 3 indicates that the power of the Scan Statistic is low for clusters consisting of only a few events. In addition, because the statistic is discrete, it would be possible for a number of small clusters to produce a sequence of (possibly identical) window counts during the scanning process, none of which individually reached the critical value for the given sample size and significance level. Therefore, the statistic should have greater power to detect a single, large cluster, rather than a number of smaller aggregations, for a given level of "clustering activity". The ISD statistic should be more flexible, since its definition is equivalent, in some sense, to the sum over the whole region of squared deviations between the two densities. Hence, a large (significant) value of $T_{h_1 h_2}$ or T could be composed, plausibly, of a single, large deviation or several smaller deviations at different locations, the former representing one cluster and the latter, many.

7.3 Generalisations

Since significance testing for the Scan Statistic will usually be accomplished by simulation, it would be possible to extend the method to use scanning windows of different shapes. For example, it might be thought to be more relevant to some clustering alternatives to use a scanning circle, which would give a test procedure with similarities to the Geographical Analysis Machine of Openshaw *et al* (1987, 1988). This resemblance would become stronger if, as suggested above, the scanning procedure were repeated for a range of radii, with the intention of exploring the scale at which clustering might be taking place. Other shapes could be introduced to increase

power for specific alternative hypotheses, *e.g.* a rectangle or ellipse aligned with the direction of the prevailing wind, to test hypotheses regarding air dispersion of pollutants. However, dispensing with the square window would have the disadvantage that the design of an efficient computational algorithm that would ensure the necessary coverage of the region of interest by the scanning procedure would become more difficult.

The ISD statistic, $T_{h_1 h_2}$, is based on the difference between two kernel density estimators, whilst the exploratory method of Bithell (1990) consists of plotting a function of the ratio of these quantities. It would be of interest to compare other measures of the deviation between two k.d.e.'s, the most obvious choice, perhaps, being the maximum L_1 distance between the two estimates,

$$\max_{x \in A} \left| \hat{f}_1(x) - \hat{f}_2(x) \right|, \tag{7.1}$$

where A is the region of interest. This statistic may have greater power than $T_{h_1 h_2}$ to detect small clusters, since the former test could be sensitive to a single spike in the surface representing absolute difference, whilst the corresponding integrated squared difference between the two kernel estimates might not be significantly large. Obtaining distributional results for (7.1), whether exact or asymptotic, could be very difficult, and, therefore, it might be necessary to use a bootstrap significance test. If so, a smoothed bootstrap procedure might be appropriate once more, since repeated values in the simulated data, caused by the with-replacement sampling scheme, could produce artificially high spikes in the corresponding simulated absolute difference surface, reducing the power of the test.

It would also be of interest to extend the type of technique considered here, *i.e.* based on nonparametric smoothing methods in a continuous spatial domain, to the problem of detecting clustering around a fixed location, the particular concern of the methods discussed in Section 1.2.1. In fact, one-dimensional results for both the Scan Statistic, as presented earlier, and the ISD statistic could be used for this purpose with few modifications. Providing directional heterogeneity could be neglected, either technique could be applied to assess the distribution of 'event to fixed point' distances, where the null hypothesis of no clustering could be represented by the empirical distribution of 'control to fixed point' distances. This approach might be generalised further, to problems where the fixed location is replaced by a line, which might represent, say, a high voltage electrical power cable. The two samples would then consist of the

perpendicular distances between the observations and the line source. It would be reasonable to expect these methods to have relatively low power, but there may be other possibilities with greater power that retain the advantages of this type of method for detecting spatial clustering.

REFERENCES

Afflerbach, L. and Grothe, H. (1988) The lattice structure of pseudo-random vectors generated by matrix generators. *J. Comput. Appl. Math.*, **23**, 127 - 131.

Ahmad, I.A. (1980) Nonparametric estimation of an affinity measure between two absolutely continuous distributions with hypotheses testing applications. *Ann. Inst. Statist. Math.*, **32**, 223 - 240.

Alexander, F.E. (1991) Investigations of localised spatial clustering and extra-Poisson variation. In: *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966-83* (ed. Draper, G.), OPCS Studies on Medical and Population Subjects, No. 53. London: HMSO.

Anderson, N.H., Hall, P. and Titterington, D.M. (1991) *Edgeworth expansions in very high-dimensional problems*. Research Report CMA - SR28 - 91, Centre for Mathematics and its Applications, Australian National University, Canberra.

Anderson, N.H., Hall, P. and Titterington, D.M. (1992a) *Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates*. Research Report CMA - SR18 - 92, Centre for Mathematics and its Applications, Australian National University, Canberra.

Anderson, N.H., Hall, P. and Titterington, D.M. (1992b) Interpolation rules suggested by asymptotic expansions. *Biometrika*, **79**, 651 - 653.

Barnard, G. (1963) Contribution to the Discussion of "The spectral analysis of point processes" by M.S. Bartlett. *J. R. Statist. Soc. B*, **25**, 294.

Barnes, N., Cartwright, R.A., O'Brien, C., Roberts, B., Richards, I.D.G. and Bird, C.C. (1987) Spatial patterns in electoral wards and high lymphoma incidence in the Yorkshire health region. *Br. J. Cancer*, **56**, 169 - 172.

Berman, M. and Diggle, P. (1989) Estimating weighted integrals of the second-order intensity of a spatial point process. *J. R. Statist. Soc. B*, **51**, 81 - 92.

Besag, J. and Clifford, P. (1991) Sequential Monte Carlo p-values. *Biometrika*, **78**, 301 - 304.

Besag, J. and Newell, J. (1991) The detection of clusters in rare diseases. *J. R. Statist. Soc. A*, **154**, 143 - 155.

Bithell, J.F. (1990) An application of density estimation to geographical epidemiology. *Statist. Med.*, **9**, 691 - 701.

Bithell, J.F. and Stone, R.A. (1989) On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. *J. Epidem. Comm. Health*, **43**, 79 - 85.

Black, D. (1984) *Investigation of the possible increased incidence of cancer in West Cumbria.* London: HMSO.

Black, R.J., Sharp, L. and Urquhart, J.D. (1991) An analysis of the geographical distribution of childhood leukaemia and non-Hodgkin lymphomas in Great Britain using areas of approximately equal population size. In: *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966-83* (ed. Draper, G.), OPCS Studies on Medical and Population Subjects, No. 53. London: HMSO.

Bowman, A.W. (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353 - 360.

Conover, W.J., Bement, T.R. and Iman, R.L. (1979) On a method for detecting clusters of possible uranium deposits. *Technometrics*, **21**, 277 - 282.

Cuzick, J. and Edwards, R. (1990) Spatial clustering for inhomogeneous populations. *J. R. Statist. Soc. B*, **52**, 73 - 104.

David, H.A. (1981) *Order Statistics* (Second Edition). New York: Wiley.

Diggle, P.J. (1983) *Statistical Analysis of Spatial Point Patterns*. London: Academic Press.

Diggle, P.J. (1985) A kernel method for smoothing point process data. *Appl. Statist.*, **34**, 138 - 147.

Diggle, P.J. (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *J. R. Statist. Soc. A*, **153**, 349 - 362.

Diggle, P.J. and Chetwynd, A.G. (1991) Second order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, **47**, 1155 - 1163.

Diggle, P.J., Gatrell, A.C. and Lovett, A.A. (1990) Modelling the prevalence of cancer of the larynx in part of Lancashire: a new methodology for spatial epidemiology. In: *Spatial Epidemiology* (ed. Thomas, R.W.), London Papers in Regional Science, **21**. London: Pion.

Diggle, P.J. and Marron, J.S. (1988) Equivalence of smoothing parameter selectors in density and intensity estimation. *J. Am. Statist. Assoc.*, **83**, 793 - 800.

Doll, R. (1989) The epidemiology of childhood leukaemia. *J. R. Statist. Soc. A*, **152**, 341 - 351.

Draper, G. (ed.) (1991) *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966-83.* OPCS Studies on Medical and Population Subjects, No. 53. London: HMSO.

Durst, M.J. (1989) Using linear congruential generators for parallel random number generation. *Proc. 1989 Winter Simulation Conf.*, 462 - 466. San Diego: IEEE Press.

Ederer, F., Myers, M. and Mantel, N. (1964) A statistical problem in space and time: do leukaemia cases come in clusters? *Biometrics*, **20**, 626 - 638.

Eichenauer-Herrmann, J., Grothe, H. and Lehn, J. (1989) On the period length of pseudo-random vector sequences generated by matrix generators. *Math. Comput.*, **52**, 145 - 148.

Fishman, G.S. and Moore, L.R. (1982) A statistical evaluation of multiplicative congruential random number generators with modulus 2^{31} - 1. *J. Am. Statist. Assoc.*, **77**, 129 - 136.

Gardner, M.J. (1989) Review of reported increases of childhood cancer rates in the vicinity of nuclear installations in the U.K. *J. R. Statist. Soc. A*, **152**, 307 - 325.

Gibbons, J.D. (1986) Randomized Tests. In: *Encyclopedia of Statistical Sciences, Volume 7* (ed. Kotz, S., Johnson, N.L. and Read, C.B.), 548 - 549. New York: Wiley.

Glaz, J. (1989) Approximations and bounds for the distribution of the Scan Statistic. *J. Am. Statist. Assoc.*, **84**, 560 - 566.

Grothe, H. (1987) Matrix generators for pseudo-random vector generation. *Stat. Hefte*, **28**, 233 - 238.

Hall, P. (1984) Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Mult. Anal.*, **14**, 1 - 16.

Hall, P. and Hart, J.D. (1990) Bootstrap test for difference between means in nonparametric regression. *J. Am. Statist. Assoc.*, **85**, 1039 - 1049.

Hall, P. and Wilson, S.R. (1991) Two guidelines for bootstrap hypothesis testing. *Biometrics*, **47**, 757 - 762.

Heyde, C.C. (1983) Limit Theorem, Central. In: *Encyclopedia of Statistical Sciences, Volume 4* (ed. Kotz, S., Johnson, N.L. and Read, C.B.), 651 - 655. New York: Wiley.

Hills, M. and Alexander, F. (1989) Statistical methods used in assessing the risk of disease near a source of possible environmental pollution: a review. *J. R. Statist. Soc. A*, **152**, 353 - 363.

Hinkley, D.V. (1988) Bootstrap methods. *J. R. Statist. Soc. B*, **50**, 321 - 337.

Hope, A.C.A. (1968) A simplified Monte Carlo significance test procedure. *J. R. Statist. Soc. B*, **30**, 582 - 598.

Huntington, R.J. and Naus, J.I. (1975) A simpler expression for k th nearest neighbour coincidence probabilities. *Ann. Probab.*, **3**, 894 - 896.

Hwang, F.K. (1977) A generalisation of the Karlin-McGregor Theorem on coincidence probabilities and an application to clustering. *Ann. Probab.*, **5**, 814 - 817.

James, F. (1990) A review of pseudorandom number generators. *Comput. Phys. Commun.*, **60**, 329 - 344.

Jones, M.C. and Lotwick, H.W. (1984) A remark on Algorithm AS176: kernel density estimation using the Fast Fourier Transform. *Appl. Statist.*, **33**, 120 - 122.

Jones, M.C. and Sheather, S.J. (1991) Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.*, **11**, 511 - 514.

Khashimov, Sh. A. (1988) On the limit distribution of a two-sample Von Mises functional with variable kernel. *Theor. Probab. Appl.*, **33**, 179 - 182.

Kinlen, L.J. (1988) Evidence for an infective cause for childhood leukaemia: a Scottish new town compared to nuclear reprocessing sites. *Lancet*, **ii**, 1323 - 1326.

Knox, E.G. (1964a) The detection of space-time interactions. *Appl. Statist.*, **13**, 25 - 29.

Knox, E.G. (1964b) Epidemiology of childhood leukaemia in Northumberland and Durham. *Br. J. Prev. Soc. Med.*, **18**, 17-24.

Knuth, D.E. (1981) *The Art of Computer Programming, Volume 2: Seminumerical Algorithms.* Second Edition. London: Addison-Wesley.

L'Ecuyer, P. (1990) Random numbers for simulation. *Commun. A.C.M.*, **33**, 85 - 97.

Lyon, J.L., Klauber, M.R., Graff, W. and Chiu, G. (1981) Cancer clustering around point sources of pollution: assessment by a case-control methodology. *Environ. Res.*, **25**, 29 - 34.

Mantel, N. (1967) The detection of disease clustering and a generalised regression approach. *Cancer Res.*, **27**, 209 - 220.

Marshall, R.J. (1991) A review of methods for the statistical analysis of spatial patterns of disease. *J. R. Statist. Soc. A*, **154**, 421 - 441.

McAuliffe, T.L. and Afifi, A.A. (1984) Comparison of a nearest neighbour and other approaches to the detection of space-time clustering. *Comp. Stat. & Data Anal.*, **2**, 125 - 142.

Miller, K.S. (1964) *Multidimensional Gaussian Distributions*. New York: Wiley.

NAg. (1990) *FORTRAN Subroutine Library, Mark 14*. Oxford: Numerical Algorithms Group.

Naus, J.I. (1965a) The distribution of the size of the maximum cluster of points on a line. *J. Am. Statist. Assoc.*, **60**, 532 - 538.

Naus, J.I. (1965b) Clustering of random points in two dimensions. *Biometrika*, **52**, 263 - 267.

Naus, J.I. (1966a) Some probabilities, expectations and variances for the size of largest clusters and smallest intervals. *J. Am. Statist. Assoc.*, **61**, 1191 - 1199.

Naus, J.I. (1966b) A power comparison of two tests of non-random clustering. *Technometrics*, **8**, 493 - 517.

Naus, J.I. (1982) Approximations for distributions of Scan Statistics. *J. Am. Statist. Assoc.*, **77**, 177 - 183.

Naus, J.I. (1988) Scan Statistics. In: *Encyclopedia of Statistical Sciences, Volume 8* (ed. Kotz, S., Johnson, N.L. and Read, C.B.), 281 - 285. New York: Wiley.

Openshaw, S. (1990) Automating the search for cancer clusters: a review of problems, progress and opportunities. In: *Spatial Epidemiology* (ed. Thomas, R.W.), London Papers in Regional Science, **21**. London: Pion.

Openshaw, S., Charlton, M. and Craft, A. (1988) Searching for leukaemia clusters using a Geographical Analysis Machine. *Papers Reg. Sci. Assoc.*, **64**, 95 - 106.

Openshaw, S., Charlton, M., Wymer, C. and Craft, A. (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *Int. J. Geog. Inf. Sys.*, **1**, 335 - 358.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1989) *Numerical Recipes (FORTRAN Version)*. Cambridge: Cambridge University Press.

Raubertas, R.F. (1988) Spatial and temporal analysis of disease occurrence for detection of clustering. *Biometrics*, **44**, 1121 - 1129.

Ripley, B.D. (1987) *Stochastic Simulation*. New York: Wiley.

Rosenblatt, M. (1971) Curve estimates. *Ann. Math. Statist.*, **42**, 1815 - 1842.

Ross, A. and Davis, S. (1990) Point pattern analysis of the spatial proximity of residences prior to diagnosis of persons with Hodgkin's disease. *Am. J. Epidem.*, **132**, Suppl. 1, S53 - S62.

Rubinstein, R.Y. (1981) *Simulation and the Monte Carlo Method.* New York: Wiley.

Schulman, J., Selvin, S. and Merrill, D.W. (1988) Density Equalised Map Projections: a method for analysing clustering around a fixed point. *Statist. Med.*, **7**, 491 - 505.

Silverman, B.W. (1982) Algorithm AS176: kernel density estimation using the Fast Fourier Transform. *Appl. Statist.*, **31**, 93 - 99.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Silverman, B.W. and Young, G.A. (1987) The bootstrap: to smooth or not to smooth? *Biometrika*, **74**, 469 - 479.

Smith, B.T. (1967) *ZERPOL: a zero-finding algorithm for polynomials using Laguerre's Method.* Technical Report, Department of Computer Science, University of Toronto, Canada.

Stone, R.A. (1988) Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statist. Med.*, **7**, 649 - 660.

Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L. and Clark, L.C. (1990) Monitoring for clusters of disease: application to leukaemia incidence in upstate New York. *Am. J. Epidem.*, **132**, Suppl. 1, S136 - S143.

Wakeford, R. (1990) Some problems in the interpretation of childhood leukaemia clusters. In: *Spatial Epidemiology* (ed. Thomas, R.W.), London Papers in Regional Science, **21**. London: Pion.

Wakeford, R., Binks, K. and Wilkie, D. (1989) Childhood leukaemia and nuclear installations. *J. R. Statist. Soc. A*, **152**, 61 - 86.

Wallenstein, S.R. and Naus, J.I. (1974) Probabilities for the size of largest clusters and smallest intervals. *J. Am. Statist. Assoc.*, **69**, 690 - 697.

Wallenstein, S. and Neff, N. (1987) An approximation for the distribution of the Scan Statistic. *Statist. Med.*, **6**, 197 - 207.

Wartenberg, D. and Greenberg, M. (1990) Detecting disease clusters: the importance of statistical power. *Am. J. Epidem.*, **132**, Suppl. 1, S156 - S166.

Weinstock, M.J. (1981) A generalised Scan Statistic test for the detection of clusters. *Int. J. Epidem.*, **10**, 289 - 293.

Whittemore, A.S., Friend, N., Brown, B.W. and Holly, E.A. (1987) A test to detect clusters of disease. *Biometrika*, **74**, 631 - 635.