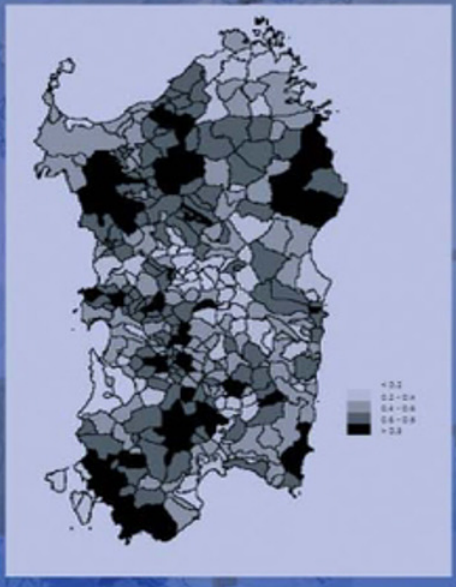
WILEY

Statistical Methods in Spatial Epidemiology

Second Edition



Andrew B. Lawson

WILEY SERIES IN PROBABILITY AND STATISTICS

Statistical Methods in Spatial Epidemiology

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Statistical Methods in Spatial Epidemiology

Second Edition

Andrew B. Lawson

Department of Epidemiology and Biostatistics,
University of South Carolina,
Columbia, USA

John Wiley & Sons, Ltd

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-470-01484-4
ISBN-10: 0-470-01484-9

"... a story is a letter the author writes to themself, to tell themself things that they would be unable to discover otherwise."

after **Carlos Ruiz Zafón**

‘to Keir, Fraser, and Hugh and all my family’

Contents

Preface and Acknowledgements to Second Edition

Since the appearance of the first edition of this book there has been a considerable development of interest in statistical methodology in the area of spatial epidemiology. This development has seen the increased output of research papers and books marking the maturity of certain areas of concern. For example, close to that date when the edited volume by Elliott *et al*. (2000) appeared, and since special issues of the *Journal of the Royal Statistical Society, Series A* (2001), *Environmental and Ecological Statistics* (2005), *Statistical Methods in Medical Research* (2005, 2006) and *Statistics in Medicine* (2006) have all contributed to the appearance of novel methodology. The development of software has also facilitated the wider use of the more advanced methods. In particular, the availability of free packages such as `R`, `WinBUGS` and `SaTScan` has led to wide dissemination of the available methods.

In particular, the area of disease map modelling has seen much development with Bayesian modelling as a particular feature. The use of mixture models and variants of likelihoods has seen development, while the routine application of sophisticated random-effect models is now relatively straightforward. The areas of disease clustering, ecological analysis and infectious disease modelling have all seen advances. In addition, the area of surveillance has re-emerged due to interest in early detection of potential bioterrorism attacks and in particular syndromic surveillance has become a major focus.

I would like to take this opportunity to acknowledge the influence and support of the following: Linda Pickle (NIH), Ram Tiwari (NIH), Martin Kulldorff, Dan Wartenburg, Peter Rogerson, Andrew Moore, Sudipto Banerjee, Ken Kleinman, William Browne, Carmen Vidal Rodeiro, Monir Hossain, Allan Clark, Yang Wang, Yuan Liu, Bo Ma, Huafeng Zhou. Finally I should also like to acknowledge the helpful interactions with staff at Wiley Europe over the years: Kathryn Sharples, Sian Jones, Helen Ramsey, Sharon Clutton and Lucy Bryan.

Andrew Lawson, Columbia, South Carolina
December 2005

Preface and Acknowledgements

The development of statistical methods in spatial epidemiology has had a chequered career. One of the earliest examples of the analysis of geographical locations of disease in relation to a putative health hazard was John Snow's analysis of cholera cases in relation to the location of the Broad Street water pump in London (Snow, 1854). However, until recently, developments in statistical methods in this area have been sporadic. While medical geography developed in the 1960s (Howe, 1963), only a number of papers on space-time clustering (Mantel, 1967; Knox, 1964) appeared in the statistical literature. More recently, developments of methods in spatial statistics, image processing, and in particular Bayesian methods and computation, have seen parallel developments in methods for spatial epidemiology (see Marshall (1991b) for a review). It is notable that methods for the analysis of case locations around a source of hazard (such as Snow's cholera map) have only recently been developed (Diggle, 1989; Lawson, 1989). The current increased level of interest in statistical methods in spatial epidemiology is a reflection, in part, of the increased concern in society for environmental issues and their relation to the health of individuals. Hence, the 'detection' of pollution sources or sources of health hazard can be seen as the backdrop to many studies in environmental epidemiology (Diggle, 1993). The correct allocation of resources for health care in different areas by health services is also greatly enhanced by the development of statistical methods which allow more accurate depiction of 'true' disease incidence and its relation to explanatory variables. Previous work in this area has been reviewed by Lawson and Cressie (2000), and Marshall (1991b) and Elliott *et al*. (1992a) discuss the general epidemiological issues surrounding spatial epidemiological problems.

It is the purpose of this book to provide an overview of the main statistical methods currently available in the field of spatial epidemiology. Inevitably, some selectivity in choice of methods reviewed will be apparent, but it is hoped that our coverage will encompass the most important areas of development. One area which we do not examine in detail is that of space-time analysis of epidemiological data, although the modelling of infectious disease data is considered in Chapter 11.

As this book is mainly a review of recent research work, its target audience is largely confined to those with some statistical knowledge and is appropriate for

third level degree and postgraduate students in statistics, or epidemiology with a strong statistical background.

A considerable number of people have directly or indirectly contributed to the production of this book. First, I acknowledge the support of Sharon Clutton and Helen Ramsey at Wiley and Tony Johnson of Statistics in Medicine for their support from Budapest onwards. Fundamental influences in the development of my ideas in spatial epidemiology have been Richard Cormack and Peter Diggle. I also acknowledge the encouragement of Noel Cressie, who has supported my work through visits to Iowa State and Ohio State Universities, and important collaborations with Martin Kulldorff, Annibale Biggeri, Dankmar Boehning, Peter Schlattmann, Emmanuel Lesaffre, Jean-Francois Viel, Adrian Baddeley, Niels Becker and Andrew Cliff.

Andrew B. Lawson, Aberdeen,
March 2000

Part I

The Nature of Spatial Epidemiology

1

Definitions, Terminology and Data Sets

Spatial epidemiology concerns the analysis of the spatial/geographical distribution of the incidence of disease. In its simplest form the subject concerns the use and interpretation of *maps* of the locations of disease cases, and the associated issues relating to map production and the statistical analysis of mapped data must apply within this subject. In addition, the nature of *disease* maps ensures that many epidemiological concepts also play an important role in the analysis. In essence, these two different aspects of the subject have their own impact on the methodology which has developed to deal with the many issues which arise in this area.

First, since mapped data are spatial in nature, the application of *spatial* statistical methods forms a core part of the subject area. The reason for this lies in the fact that the study of any data which are georeferenced (i.e. have a spatial/geographical location associated with them) may have properties which relate to the location of individual data items and also the surrounding data. For example, Figure 1.1 shows the total number of deaths from respiratory cancer found in 26 small areas (census tracts) in central Scotland over the period 1976–1983. This map displays a number of features which commonly arise when the geographical distribution of disease is examined. On this map the numbers (counts) of cases within each area are displayed. In some areas of the map the counts are similar to those found in the immediately surrounding areas (e.g. in the south and southeast of the map counts of 4 and 6 are recorded, while in the northwest of the map, lower counts are found in many areas). This similarity in the count data in groups of tracts is unlikely to have arisen from the allocation of a random sample of counts from a common statistical distribution. The counts may display some form of correlation in their levels based on their location, i.e. counts close to each other in space are similar. This form of correlation does not arise from the usual statistical models assumed to

Statistical Methods in Spatial Epidemiology/Second Edition Andrew B. Lawson

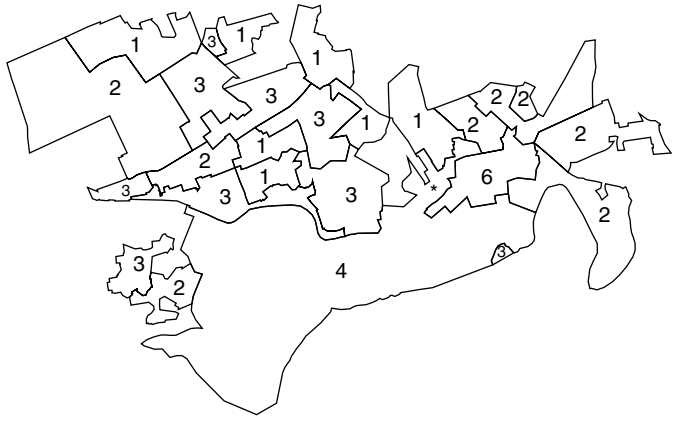


Figure 1.1 Falkirk: central Scotland respiratory cancer counts in 26 census enumeration districts over a fixed time period. * Putative health hazard.

apply to independent observations found in, for example, clinical medical studies or other conventional statistical application areas. Hence, methods which apply to the analysis of these data must be able to address the possibility of such correlation existing in the mapped data under study. Another feature of this example, which commonly arises in the study of spatial epidemiology, is the irregular nature of the regions within which the counts are observed, i.e. the census tracts have irregular geographical boundaries. This may arise as a feature of the whole study region (*study window*) or may be found associated with tracts themselves. In some countries, notably in North America, small areas are often regular in shape and size and this feature simplifies the resulting analysis. However, in many other areas irregular region geometries are common. Finally, in some studies, the spatial distribution of cases or counts of disease are to be related to other locations on the map. For example, in Figure 1.1 the location of a potential (*putative*) environmental health hazard is also mapped (a metal-processing plant), and the focus of the study may be to assess the relationship of the disease incidence on the map to that location, perhaps to make inferences about the environmental risk in its vicinity.

The second feature which uniquely defines the study of spatial epidemiology is that the mapped data are often *discrete*. Unlike other areas of spatial statistical analysis, which are often focused on continuous data, e.g. geostatistical methods, the data found in spatial epidemiology often take the form of point locations (the address locations of cases of disease) or counts of disease within regions such as census tracts or, at larger scale, counties or municipalities. Hence, the mapped data often consist of cartesian coordinates in the form of a grid reference or longitude/latitude of an address of a case, or a count of cases within a region with the associated location of that region (either as a point location of a centroid or as a set of boundary line segments defining the region). Given this form of data format,

it is not surprising that models which have been developed for applications within this area are derived from stochastic point process theory (for case locations) and associated discrete probability distributions (for counts within arbitrary regions).

Finally, the epidemiological nature of these discrete spatial data leads to the derivation of models and methods which are related to conventional epidemiological studies. For example, the case–control study, where individual cases are matched to control individuals based on specific criteria, has parallels in spatial epidemiology where spatial control *distributions* are used to provide a locational control for cases. This is akin to the estimation of background hazard in survival studies. One fundamental epidemiological issue which arises in these studies is the incorporation of the local population which is at risk of contracting the disease in question. As we must control for the spatial variation in the underlying population, then we must be able to obtain good estimates of the population from which the cases or counts arise. This estimation often leads to the derivation of *expected* rates in the region count case and further to the estimation of the ratio of count to expected count/rate or the *relative risk*, in each area. Relative risk is a fundamental epidemiological concept (Clayton and Hills, 1993) in non-spatial epidemiological studies.

1.1 Map Hypotheses and Modelling Approaches

In any spatial epidemiological analysis, there will usually be a study focus which specifies the nature and style of the methods to be used. This focus will usually consist of a hypothesis or hypotheses about the nature of the spatial distribution of the disease which is to be examined, and it is convenient to categorise these hypotheses into three broad classes: *disease mapping*, *ecological analysis* and *disease clustering*. Usually, the distribution of cases of disease, whether in the form of counts or case address locations, can be thought to follow an underlying model, and the observed data may contain extra noise in the form of random variation around the model of interest. Often, the model will include aspects of the *null* (hypothesis) spatial distribution of the cases, which captures the 'normal' variation which is expected, and also aspects of the *alternative* spatial distribution. In much of spatial epidemiology, the focus of attention is on identifying features of the spatial distribution which are not captured by the null hypothesis distribution. This is mainly related to *excess spatial aggregation* of cases in areas of the map. That is, once the normal variation is allowed for, the residual spatial incidence *above* the normal incidence is the focus. Seldom is there any need to examine areas of lower aggregation than would be normally expected. Note that 'normal' variation is usually assumed to be defined by the underlying population distribution of the study region/window and cases are thought to arise in relation to the local variation in that distribution.

The first class, that of *disease mapping*, concerns the use of models to describe the overall disease distribution on the map. In disease mapping, often the object is simply to 'clean' the map of disease of the extra noise to uncover the underlying

structure. In that situation, the null hypothesis could be that the case distribution arises from an *unspecified* or partly specified null spatial distribution (which includes the population spatial distribution) and the object is to remove the extra noise/variation. In this sense disease mapping is close in spirit to image processing where *segmentation* usually describes the process of allocating pixels or groups of pixels to classes.

The second class, that of *ecological analysis*, concerns the analysis of the relation between the spatial distribution of disease incidence and measured explanatory factors. This is usually carried out at an aggregated spatial level, and usually concerns regional incidence compared to explanatory factors measured at regional or other levels of aggregation (Greenberg *et al*., 1996). This contrasts with studies which use measurements made on individual subjects. However, many of the issues concerning interpretation of ecological studies are concerned with *change* in aggregation level and not aggregated data per se. For example, the *ecological fallacy* concerns making inference about individuals from analyses carried out at a higher scale, e.g. regional or country-wide level. Equally, the *atomistic fallacy* concerns making inferences about average characteristics from individual measurements. In what follows we assume a relatively wide definition of ecological, more in the sense of ecology itself, as any study which seeks to describe/explain the spatial distribution of disease based on the inclusion of explanatory variables. Two classic studies of this kind are presented by Cook and Pocock (1983), who examined the relation of cardiovascular incidence in the UK to a variety of variables (including water hardness, climate, location, socioeconomic and genetic factors and air pollution), and Donnelly (1995), who examined the respiratory health of school children and volatile organic compounds in the outdoor atmosphere. Note that this general definition can include the situation where case address locations are related to a pollution hazard via explanatory variables such as distance and direction from the hazard. In that case individual data are related to explanatory variables.

The final class, that of *disease clustering*, concerns the analysis of 'unusual' aggregations of disease, i.e. assessing whether there are any areas of elevated incidence of disease within a map. This type of analysis could take a variety of forms. First, the analysis could include the assessment of a complete map to ascertain whether the map is *clustered.* This is often termed *general clustering*. In this case, the null hypothesis would be that the disease map represents normal variation in incidence given the population distribution. The alternative hypothesis would include some specified clustering mechanism for the disease cases. This mechanism could be descriptive or include some notion of how the clusters form (e.g. clusters can form if infectious diseases are examined, and the contact rate of individuals can be modelled). General clustering is often treated as a form of *autocorrelation* and models for such effects are often employed. This form of clustering can be termed *non-specific* as it does not seek to determine where clusters are found but instead simply seeks to determine whether the pattern is clustered.

Second, *specific* cluster studies attempt to ascertain the locations of any clusters if they exist on the map. These clusters could have known (fixed) locations and the incidence of disease around these locations may be assessed for its relation to the location(s). Studies of putative pollution hazards fall within this category. This is often termed *focused* clustering. If the locations of clusters are unknown a priori, then the locations must also be estimated from the data; this is termed *non-focused* clustering. Often, ecological regression methods can be used in focused clustering studies, whereas, for non-focused studies, special methods must be constructed which allow the estimation of cluster locations and their form.

In all the above areas of study, fundamental to the methods employed is the inclusion of spatial location in the analysis and so spatial statistical methods are often employed to model the observed data; that epidemiological considerations should be employed in any study of the distribution of disease incidence, in that the concept of normal variation of disease (i.e. that generated from the population at risk from the disease) must be catered for in any model of incidence; and that methods used should be appropriate to the analysis of georeferenced discrete data.

1.2 Definitions and Data Examples

In this section, some basic definitions and concepts are introduced which are used throughout this book. In addition, a number of data examples make their first appearance and these will be referred to at various stages throughout the work.

In what follows we will mainly be concerned with data which are available within a single period of time. Hence, we do not provide notation for space-time problems here. Where such notation is appropriate, we provide it locally.

We define 'epidemiology' as the study of the occurrence of disease in relation to explanatory factors. A strict dictionary definition of the term implies the study of 'epidemic diseases'. However, in this work we mainly restrict attention to *fixed* time period studies and do not directly examine the dynamic behaviour of disease incidence. This area has recently been reviewed in Mollison (1995), Daley and Gani (1999) and Andersson and Britton (2000). Some discussion of epidemic models appears in Chapter 10. Here the term 'spatial epidemiology' is defined to mean the study of the occurrence of disease in spatial locations and its explanatory factors. Usually, the disease to be examined occurs within a *map* and the data are expressed as a point location (case event) or are aggregated as a count of disease within a subregion of the map. Two examples of such data are provided in Figures 1.2 and 1.3. These two data types lead to different modelling approaches, and we make specific the following definitions as a basis for further discussion.

1.2.1 Case event data

We define the study window (W), within which m disease case events occur at locations $\boldsymbol{x}_i$, $i = 1, \ldots, m$. The area of W is denoted by $|W|$, Lebesque measure on $\mathbb{R}^2$. Figure 1.4 displays these definitions.

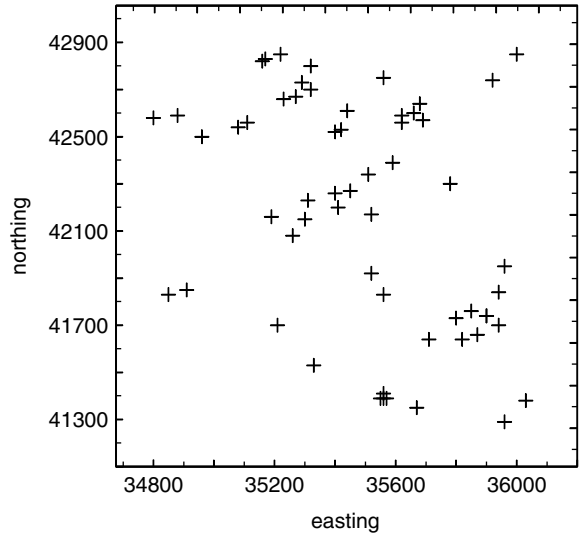


Figure 1.2 The locations of larynx cancer cases in an area of central Lancashire, UK, for the period 1974–1983.

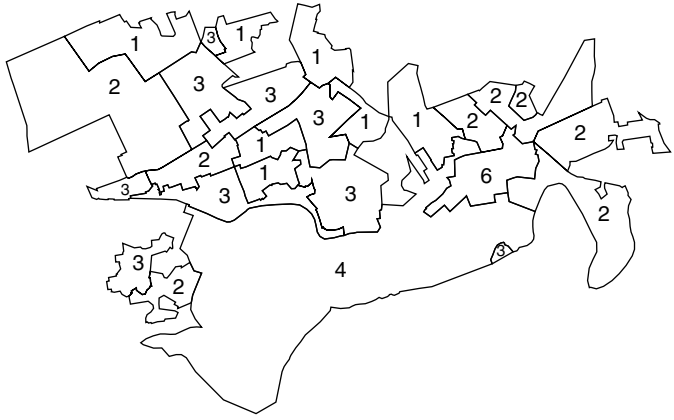


Figure 1.3 Respiratory cancer counts within census tracts (enumeration districts) of Falkirk, central Scotland, for the period 1978–1983.

1.2.2 Count data

We define the study window (W) as above, within which m *arbitrarily bounded* subregions, wholly or in part, lie. The count in m subregion tracts is denoted n_i, $i = 1, \ldots, m$. In Figure 1.5, only regions 4, 5 and 6 are wholly within the

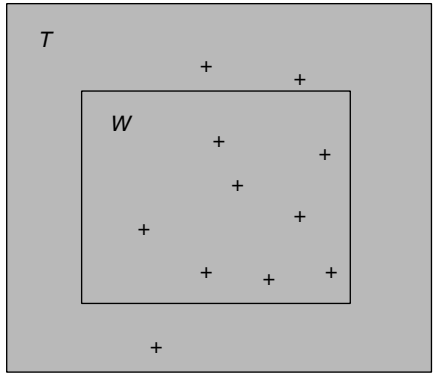


Figure 1.4 A notional study area (W) and a guard area (T).

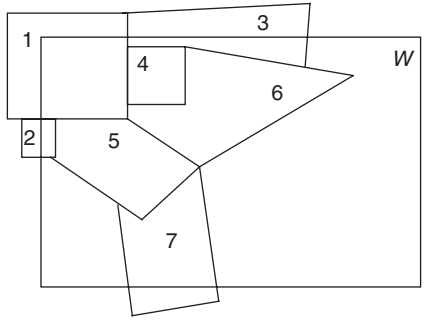


Figure 1.5 A study region within which counts are observed in subregions (tracts).

window. Regions 1, 2, 3 and 7 are cut by the window boundary. The effect of this region truncation will be discussed in detail later. However, it should be noted that, *usually*, the count available (n_i) is from the complete region and *not* from the truncated region which appears in the study window.

Usually, the m subregions are politically defined administrative regions and are often tracts defined for the purposes of population censuses. We adopt the term 'census tract' to denote an arbitrarily defined region. In addition, the counts in census tracts are just an aggregation of case event data counted within the bounding tract boundaries. Hence, the data in Figure 1.5 could be derived from the data in Figure 1.4 by counting case events in census tract subregions of the window.

The object of analysis of case event or count data can define the type of summary measures used to describe the data. Usually, as a basic summary measure it is common to compute a local measure of relative risk, or to use a local measure of relative risk as the dependent variable in a more substantial analysis. Here,

relative risk is taken to mean the measure of excess risk found in relation to that supported purely by the local population, which is 'at risk'. This population is sometimes called the 'at-risk' population or background. Relative risk is derived or computed from the relation of observed incidence to that which would be expected based on the 'at-risk' background. It is common practice within epidemiology to derive such risk estimates. In the case of spatial epidemiology it is common, when tract count data are available, to compute a standardised mortality (or morbidity) ratio (SMR), which is simply the ratio of the observed count within a tract to the expected count based on the 'at-risk' background. A ratio greater than 1.0 would suggest an excess of risk within the tract. These SMRs are often the basis for atlases of disease risk (see, for example, Pickle *et al.*, 1999).

1.3 Further Definitions

Some further definitions are required in relation to data which arise in such studies.

1.3.1 Control events and processes

Often, an additional process or realisation of disease events is used to provide an estimate of the 'background' incidence of disease in an area. Define $\boldsymbol{x}_{c_j}$, $j = 1, \ldots, m_c$, to be these m_c control event locations. The use of such data will be detailed in a later section.

1.3.2 Census tract information

The census tract count of a control disease is defined to be n_c.

Instead of using a control disease to represent 'background', the 'expected' incidence of disease can be used. This is usually based on known rates of disease in the population (Inskip *et al.*, 1983). Denote this expected incidence as e_i, $i = 1, \ldots, m$. The total population of a tract is p_i, while the extent of the tract is defined as a_i. The tract centroid, however defined, is denoted by $\boldsymbol{x}_{n_i}$.

For models involving explanatory variables measured at tract level, we define F as an $m \times p$ matrix whose columns represent p explanatory variables, and α as a $p \times 1$ vector of parameters. (For case event models the row dimension of F will usually be m also.)

1.3.3 Clustering definitions

In cases where clustering is studied, a number of additional definitions are required. First, cluster centre locations are defined as $\boldsymbol{y}_j$, $j = 1, \ldots, k$, where k is the number of centres in a suitably defined window. The term 'parent' is used here synonymously with cluster centre. This does not imply any genetic linkage with the observed data. The observed data belonging to a cluster are sometimes referred to as offspring. Again, there is no genetic linkage implied by this term. In addition,