4 Clustering of Disease

Clive E. Sabel and Markku Löytönen

CONTENTS

4.1 DISEASE CLUSTERS AND CLUSTERING

In these times of increased public awareness of and concern toward sources of environmental pollution and their potential links to disease, reports of a cluster or hot spot of disease in a particular area are commonplace. These concerns have no doubt been heightened, in the United Kingdom at least, by public health issues such as the so-called childhood leukemia cluster around the Sellafield nuclear reprocessing plant in northwest England (Gardner, 1993), over bovine spongiform encephalopathy (popularly known as mad cow disease), and more recently, the foot and mouth disease outbreak in 2001. Much of the controversy surrounding clustering lies in the difficulty in providing an adequate definition. To remove any potential for confusion, we draw the reader's attention to the distinction between the identification of disease clusters, the topic of this chapter, and the entirely separate statistical technique of cluster analysis, which aims to aggregate variables with similar characteristics in a dataset together to simplify subsequent analysis.

0-415-30655-8/04/$0.00+$1.50

Cluster detection methods can be classified into either global or local tests. Global tests detect the presence or absence of clustering over the whole study region without specifying spatial location. Local tests additionally specify the location and, if extended to consider temporal patterns, can specify spatiotemporal clusters. A special case of local tests is the focused test, which is used to detect raised incidence of disease around some prespecified source, such as an incinerator. In this chapter, we concentrate on local tests because we are concerned with locating disease clustering spatially and temporally.

In general lay conversation, the terms cluster and clustering are used interchangeably, usually applied loosely to mean any unusual collection of events. The Centers for Disease Control (CDC) discuss "any unusual aggregation of health events, real or perceived." Diggle, however, suggests that there are three distinct and separate problems in spatial epidemiology, namely cluster detection, clustering, and spatial variation in risk, but acknowledges that the distinctions between them are often blurred (Diggle, 2000). Cluster detection, according to Diggle, might be better named anomaly detection or surveillance, whereas clustering is a departure from complete spatial randomness (i.e., the hypothesis that cases occur independently of each other), which, he suggests, invites an interpretation in terms of genetic susceptibility or infectious transmission. Here it is useful to acknowledge that aggregation can occur by random processes, but cluster investigations seek to identify excess aggregation. For a lucid, diagrammatic presentation of the statistical model of complete spatial randomness, the reader may wish to consult Waller (2000). Spatial variation in risk is a departure from the hypothesis that all members of the population are at equal risk, which Diggle (2000) suggests invites an environmental interpretation. Wakefield et al. (2000) also discuss the underlying risk surface and suggest a wider definition with a cluster corresponding to an area and time period in which the (residual) risk surface is elevated (after adjusting for known risk factors), i.e., excess disease risk. Risk surfaces have the additional advantage of potentially highlighting areas of apparent low risk, which may be of value etiologically.

Clustering, however, according to Alexander and Cuzick (1992), might be defined as "a more heterogeneous and clumped distribution of disease cases than would be expected from the variations in the population density and chance fluctuations." It is thus important to recognize the distinction between the notion of an individual cluster and the concept of a general tendency for clustering.

Knox (1989) suggests three alternative definitions, with a cluster being a geographically or temporally bounded group of occurrences:

1. Of a disease already known to occur characteristically in clusters
2. Of sufficient size and concentration to be unlikely to have occurred by chance
3. Related to each other through some social or biological mechanism, or having a common relationship with some other event or circumstance

The second of Knox's definitions introduces the important concept of significance testing and begs the question of possible etiology; whereas the third definition relies on some notion of causation. The utility of cluster studies might thus be

summarized as the identification, statistical confirmation or rejection, and the suggestion of potential clues as to etiology.

4.2 WHY INVESTIGATE DISEASE CLUSTERING?

The detection and investigation of disease clusters has a long and controversial history in the field of spatial epidemiology. The basic interest in analyzing disease patterns is in determining whether the observed events exhibit any systematic pattern as opposed to being distributed at random over the study region. Wartenberg and Greenberg (1993) consider cluster studies to be a form of preepidemiology, placing them in an investigative niche prior to confirmatory epidemiological studies. Questions that one might wish to pose include:

- Is the observed clustering due mainly to natural background variation in the at-risk population from which events arise?
- Over what scale does any clustering occur?
- Are clusters merely a result of some obvious *a priori* heterogeneity in the study region?
- Are clusters associated with proximity to other features of interest, such as transport arteries or possible point sources of pollution?
- Are events that aggregate in space also clustered in time?

Disease clustering investigations might be used to generate ideas and hypotheses regarding disease etiology, but perhaps also to calm public fears of a local excess. Wartenberg (2001) suggests that public concern (often based on personal tragedy, perhaps with a specific point-source environmental contaminant in mind, but involving perhaps only a few cases), cannot and should not simply be dismissed as "lying within acceptable statistical limits" or explained away with demographic, statistical, or sampling error fluctuations. Reassurance is often required, which can be established via a carefully designed investigation.

Rothman (1990) has suggested that the payoff from clustering research comes from specific hypotheses that emerge to explain the observed pattern of excess occurrence. Whether infectious agents, genetic susceptibilities, or environmental pollutants, determining mechanisms is the goal, but only rarely have etiological insights resulted from cluster investigations.

Disease clustering investigations may prove most useful, however, in actively identifying outbreaks, particularly for infectious diseases. There have been attempts to establish national active-cluster surveillance programs, which might regularly scan register-based data for evidence of elevated risk. However, investigating incidence data over a relatively large population systematically is costly and so most investigations are more passive and are often the result of an initial request from a member of the public. For example, in the United Kingdom, the Rapid Inquiry Facility developed by the Small Area Health Statistics Unit (SAHSU) (Aylin et al., 1999) aims to produce a report within three days of a request by routinely collecting morbidity, mortality, and population data at a small spatial scale in anticipation of requests for an investigation.

4.3 CHOOSING BETWEEN METHODS

What is an appropriate public health and scientific response to a report of a suspected disease cluster? Great care needs to be taken in the presentation and interpretation of results to avoid unwarranted alarm among local residents, while not diminishing the importance of a real and observed public health concern. Obviously, the response varies depending on whether we are concerned with infectious diseases, typically operating over relatively short time scales, or noninfectious diseases with processes operating, say, over many decades. Naturally, disease can cluster spatially, temporally, or spatiotemporally. Particularly with infectious diseases, it is important to adopt methods that simultaneously test spatial and temporal clustering.

As with all research, one should establish what the research question is prior to considering what approach to take, what limitations the data impose, and whether there are other external factors to be taken into consideration. Then specific methods for the analysis can be selected.

There are two critical aspects — statistical power and confounding — to consider when selecting an appropriate method for the task at hand (Wartenberg and Greenberg, 1993). Statistical power is the ability to detect a real effect. The reader will become acquainted in the literature with the ability of methods to identify true clusters (true positives), but also the frequency with which the methods report clusters falsely (false positives). Comparative evaluations of statistical power, often by running competing cluster methods against a set of simulated data with known properties, can provide guidance in the choice and application of particular methods. Confounding is the distortion of the apparent effect of an exposure on risk brought about by the association of the exposure with other factors that can influence the outcome. In this situation, confounding may be the erroneous attribution of a disease cluster to the exposure under study, but the cluster has in fact been caused by a confounding factor associated with the exposure and independently influences disease outcome. It might be as simple as a change in background population density or involve known demographic factors such as age, gender, or ethnicity. Methods that can control for known confounding effects should be used in the first instance.

Wartenberg and Greenberg (1993) suggest that the scattered scientific literature on the evaluation of cluster methodology shows little consistency and thus has limited utility for the public health researcher seeking to distinguish between methods. Dozens of methods exist in the literature for the detection of spatial clustering over and above that due to the natural distribution of the population. Only a subset is regularly used, often due to the complexity of the methods and perhaps a lack of direction about the relative benefits of each and when it is appropriate to apply which method, because different methods suit different scenarios. Wartenberg and Greenberg suggest that users, when confronted with the multitude and complexity of methods available, often select methods arbitrarily basing choices on software availability or ease of implementation. Consideration of the statistical power and ability to adjust for potential confounding effects in differing tests are overlooked.

Wakefield et al. (2000) have devised a useful classification of methods, separated into four groups which we shall mention briefly:

1. Traditional methods
2. Distance/adjacency methods
3. Moving windows methods
4. Risk surface estimation methods

The first group — traditional methods — primarily detects overdispersion in areally aggregated data. They are global tests and hence do not provide an indication of location, but detect the presence or absence of clustering over the whole study region. Examples of such tests are Pearson's chi-squared statistic and the Potthoff and Whittinghill's method.

The second group — distance/adjacency methods — consists of global tests that assess the spatial dependence in a set of data. Here one considers such techniques as autocorrelation statistics (among the most common of which are Moran's I and Geary's c), Whittemore's method, Tango's method, and K-functions.

The third group — moving windows — has been developed since the mid-1980s and has been designed to assess whether the number of cases within a window exceeds that expected by chance. A window can be defined in a number of ways, such as a circle of a particular radius or a 3-×-3 grid in which observations are assessed. The window moves systematically throughout the study region. These are local tests with the ability to detect spatial locations of excess. Here Wakefield et al. consider Openshaw's method, Besag and Newell's method, scan statistics, and Cuzick and Edwards's method. Scan methods are particularly recommended where there are sparse data (Wartenberg and Greenberg, 1993).

The fourth group concerns risk surface estimation. Here such methods include kernel estimation, generalized additive models, and geostatistical methods. In these relatively recently developed techniques, the emphasis is less on hypothesis testing and more on estimation of the underlying residual risk surface. These methods can potentially offer more insight into the nature of the clusters since they produce continuous surfaces of risk across the whole study region and not just the statistically identified clusters. The downside of risk surfaces is a less well-developed statistical understanding and often the necessity for the use of specialized software. They do also tend to be computationally expensive and more difficult to implement.

For further details on the methods discussed in these four groups, we refer the reader to Bailey and Gatrell (1995) and Fotheringham et al. (2000). For further discussion of the relative merits of the alternative clustering methods, see Alexander and Boyle (1996), Alexander and Cuzick (1992), Elliott et al. (2000), Kulldorff (1998), Lawson and Kulldorff (1999), and Wakefield and Elliott (1999).

Spatial statisticians often perform analyses within specialist statistical software such as S-Plus or using homegrown code, rather than within GIS. This undoubtedly is because proprietary GIS lacks real statistical modeling sophistication, despite some advances recently such as the ArcGIS® Geostatistical Analyst. Other products such as CDC's Epi Info™ 2002 (including its allied mapping product, Epi Map) have little to no cluster detection functionality.

4.4 SOME ACCESSIBLE METHODS

A daunting number of methods exist from which investigators need to select a method suitable for their specific circumstances. In this section, we concentrate on three of the most common local cluster methods accessible to GIS users, because intuitively we feel pinpointing the location of clusters is of greater utility in this context than global tests that merely detect the presence or absence of clustering in the study region. We also look for methods that are methodologically sound and can control for covariates.

4.4.1 Openshaw's Method

Within the GIS community, probably the best known is Openshaw's Geographical Analysis Machine (GAM) (Openshaw et al., 1987). GAM is freely available on the Internet together with a comprehensive user guide. GAM is an exploratory cluster detection approach, which works by examining a large number of overlapping circles at a variety of scales and assesses the statistical probability of the number of events occurring by chance. Where the number of observations is statistically significant, a circle is plotted, which results in a visual impression of where clusters might occur. A more recent generation — GAM/K — makes use of kernel estimation to display the results of the iterative process. Openshaw et al. (1999) compared the performance of several exploratory geographical methods to identify patterns spatially and temporally. The aforementioned GAM/K worked well with spatially distributed data and the GAM/K-T (GAM/K plus time) method correctly identified temporal clustering. However, Openshaw's method has been heavily criticized in the literature, largely due not only to the multiple testing problems where there were a large number of tests, but also due to the dependency of the test. The original version was also heavily computer intensive.

4.4.2 Kulldorff's Spatial Scan Statistic

A method gaining increasing attention is Kulldorff's Spatial Scan Statistic (SaTScan) (Kulldorff, 2002). SaTScan™ software is freely available over the Internet, but is installed locally on a user's PC. The spatial scan statistic has been used to test for disease clustering in a number of recent studies, including:

- A focused test investigating potential clusters of soft-tissue sarcoma and non-Hodgkin's lymphoma around a solid waste incinerator in France (Viel et al., 2000)
- Childhood leukemia in Sweden (Hjalmars et al., 1996)
- Breast cancer in the United States (Kulldorff et al., 1997)
- Amyotrophic lateral sclerosis (ALS) birthplace clustering in Finland (Sabel et al., 2003)

The scan statistic is a spatial, temporal, or spatiotemporal, local cluster detection method for aggregated data. It can be applied to both focused and nonfocused investigations and, importantly, adjusted for confounders, including adjusting for a

heterogeneous background population density. The method imposes a circular scanning window on the map and lets the center of the circle move over the study area so that at each position the window includes different sets of neighboring administrative areas. For each circle centroid, the radius varies continuously from zero to a user-defined maximum. Although the choice of maximum cluster circle size is somewhat arbitrary, and there are no clear guidelines for its choice, it is important to make the choice of maximum cluster size *a priori* to avoid the problems of multiple hypothesis testing. The test statistic adopted is the likelihood ratio, which is maximized over all the windows to identify the most likely disease clusters. A criticism of the test is that it has good power to report clusters in approximate circular forms, but poor power to detect linear clusters that perhaps follow rivers or overhead power lines. If one does not know, *a priori,* what shape a cluster might form, the test will impose a circular one regardless. Kulldorff argues, however, that it is not the exact borders of the cluster that one is most interested in, but rather the general area and centroid. Unlike some other techniques such as Openshaw's GAM, the test statistic does take into account the problem of multiple hypothesis testing and reports the significance of each reported cluster.

Sabel et al. (2003) used the scan statistic to analyze the impact of residential migration between places of birth and death for the rare neurological disease ALS in Finland for deaths between 1985 and 1995. In Figure 4.1, we present an adapted figure from this paper using a maximum cluster size of 20 percent of the total population. Background population-at-risk data was taken from the population census in 1990 for each municipality. Both significant and nonsignificant clusters are highlighted, reflecting the full output from SaTScan. The figure shows significant and widespread clustering at time of death in two areas of the south and southeast of the country, indicating areas to investigate further.

From Table 4.1, where all nine clusters are detailed, one can see that the first two clusters have a p value of less than 0.05, which might be interpreted as being significant. Cluster 1 and Cluster 2 each comprise more than 100 cases and have relative risk (RR) estimates of 1.79 and 1.32, respectively. Cluster 3 through Cluster 9 all have higher RRs, but note the small numbers of cases involved; hence, the lack of significance of these areas.

4.4.3 Kernel Estimates

A solution of how to estimate spatial variation in relative risk was first proposed in an epidemiological setting by Bithell (1990). He proposed adopting probability density estimation techniques of which kernel estimation is the most commonly used and most well understood statistically. Bithell's ideas have been developed by Kelsall and Diggle (1995) and Sabel (1999). Sabel et al. (2000) extended them to deal with the temporal component on the third dimension by weighting each location by a value representing the length of residence at that location. Sabel et al. (2000) also investigated space–time interaction by creating separate density estimates of temporal slices of the data, which were then sequenced together in an animation or movie, to enable the authors to obtain a greater understanding of the time lag of the etiology of the disease.

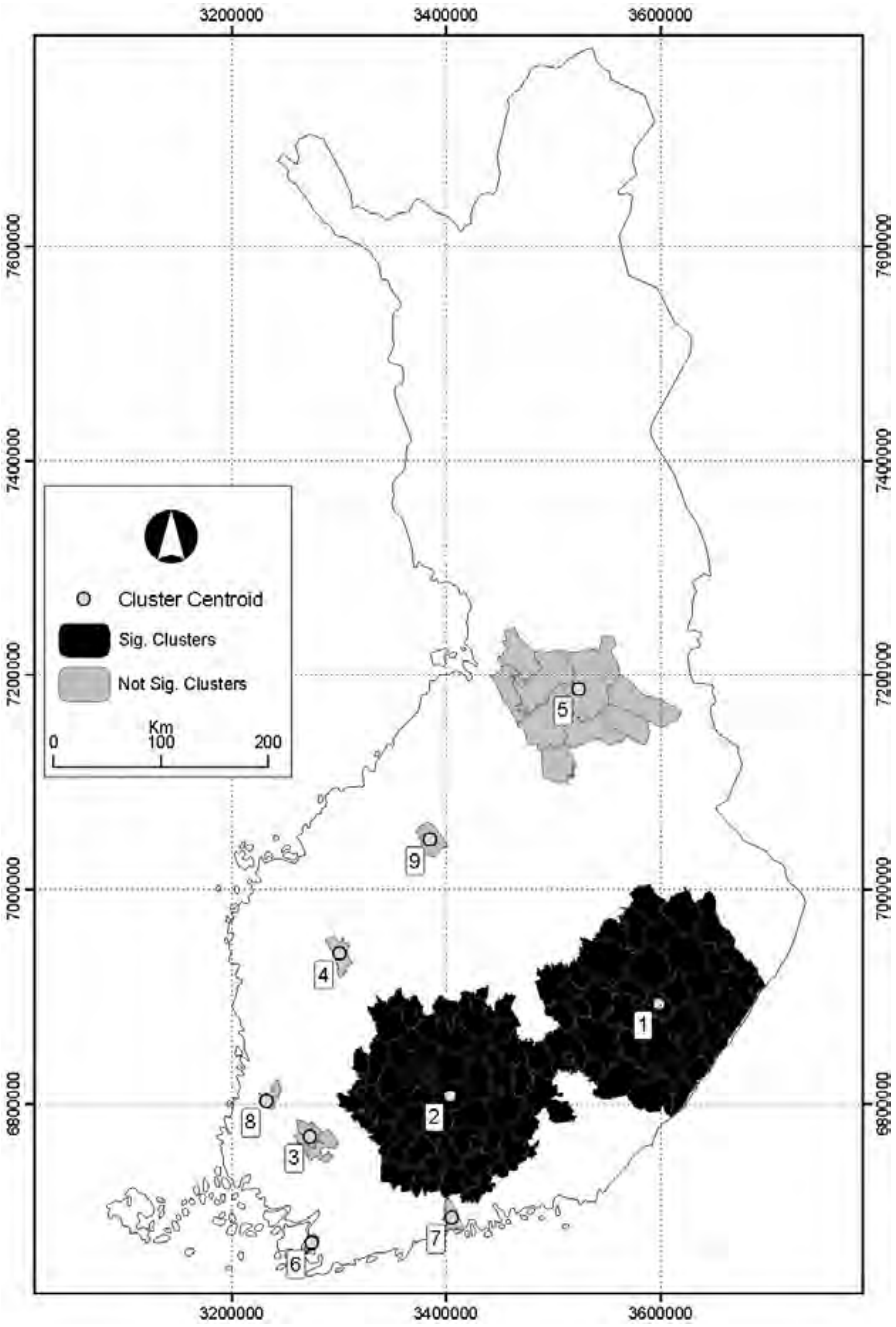

FIGURE 4.1 SaTScan identified clusters. ALS death place clusters in Finland, 1985 to 1990. (Axes labels are in meters.)

Kernel estimation is a statistical technique whereby, in epidemiological applications, a distribution of discrete points or events representing incidence of disease is transformed into a continuous surface of disease risk. Essentially, a moving three-dimensional function (the kernel) of a given radius or bandwidth visits each of the

TABLE 4.1
ALS Death Place Clusters in Finland (1985 to 1995) Using the Spatial Scan Statistic

Cluster	Approximate Location		Cases	Exp	Relative Risk Estimate (RR)	Log Likelihood Ratio (LLR)	*p* value
1	3598250	6893290	120	66.99	1.791	18.49	0.00001
2	3403860	6807980	229	174.13	1.315	9.73	0.013
3	3273130	6769600	10	2.63	3.797	6.00	0.368
4	3301020	6940090	5	0.82	6.105	4.87	0.716
5	3523310	7186900	16	7.59	2.109	3.56	0.983
6	3275130	6670440	2	0.17	11.716	3.09	0.998
7	3405190	6694410	8	2.93	2.731	2.98	0.999
8	3232900	6802670	7	2.51	2.794	2.71	1.000
9	3385530	7046240	2	0.22	8.992	2.62	1.000

points or events in turn and weights the area surrounding the point proportionately to its distance to the event. The sum of these individual kernels is then calculated for the study region and a smoothed surface is produced. Varying the bandwidth determines the degree of smoothing achieved. By taking a ratio of the kernel estimates of case intensity and underlying population, one can produce a map of relative risk, which can be tested statistically using Monte Carlo simulation techniques. Both of ESRI's ArcGIS and ArcView® GIS products include the functionality to calculate kernel estimates.

In Figure 4.2, we demonstrate output from a kernel estimate, using similar data to that shown in Figure 4.1. Here we have adapted output from Sabel et al. (2000) to again show ALS data from Finland. In this analysis, the authors took the place of residence in 1985 of ALS cases to create a relative risk estimate using kernel estimation. Population at risk was estimated by adopting a case control methodology, which has the added advantage of enabling controlling for age, gender and other covariates. A large 60-kilometer bandwidth was chosen to reveal broad trends. Significance of the surface was obtained by running Monte Carlo simulations to produce the 95 percent confidence intervals shown in the figure. Concentrating on the significant highs, a cluster in southeast Finland again manifests itself together with some other smaller areas in low-density locations. Note instances of the small number problem occurring in the north of Finland, where the applied techniques break down when spurious significant lows are shown.

Figure 4.1 and Figure 4.2 are not directly comparable. Apart from different cluster methods adopted, one uses aggregated-level data, while the other uses individual level data. Population at risk is also estimated differently between the two figures. These observations notwithstanding, the presence of the southeast cluster in both analyses suggests areas worthy of follow-up for this disease with unknown etiology.

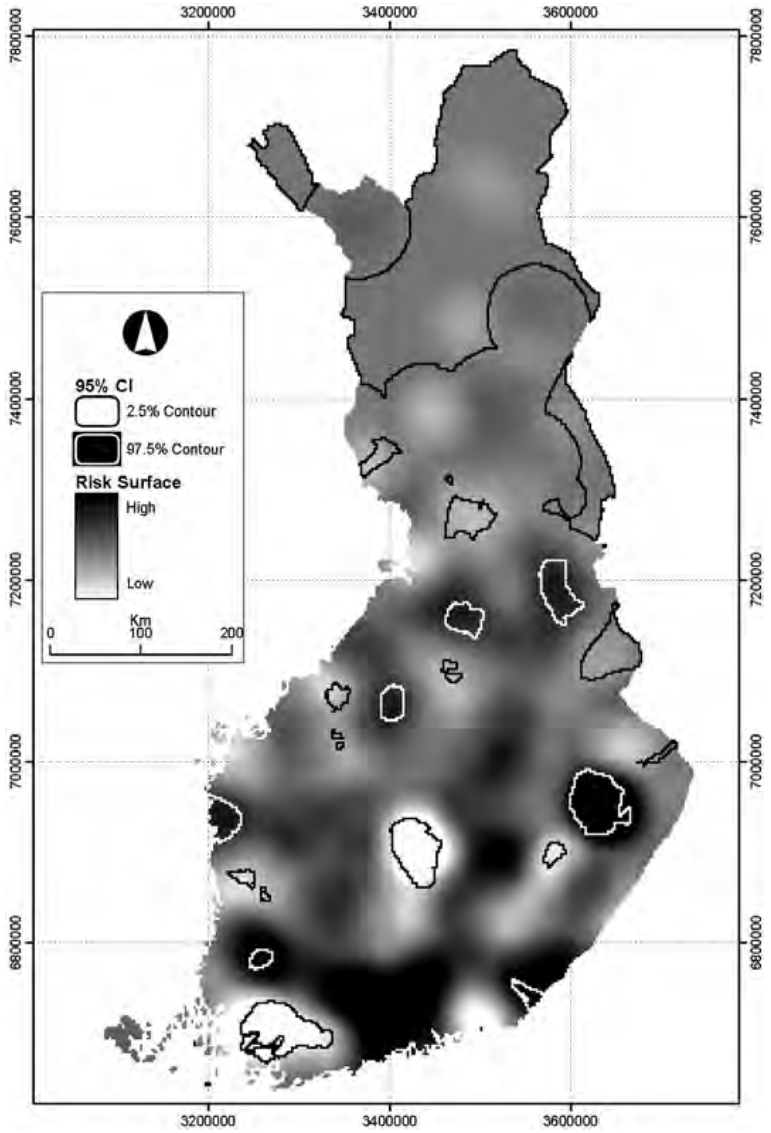

FIGURE 4.2 Relative risk surfaces derived using kernel estimation: ALS case residences, Finland, 1985. (Axes labels are in meters.)

4.5 CONSIDERATIONS AND LIMITATIONS

As important as the choice of the cluster detection test in any analysis are issues concerning data, the scale of analysis, correction for confounders and the underlying background population. It is tempting to conclude that the study of disease clustering involves so many assumptions and caveats that such studies should perhaps be avoided. Indeed, Rothman (1990) has even gone as far as to suggest that searching for individual clusters or indeed overall clustering is of little scientific value due to the (in)accuracy of the methods and the (poor) quality of data.

4.5.1 Mapping and GIS

Recall that the themes running throughout this volume are the possibilities and limitations of the use of GIS in public health. GIS has developed; it is no longer driven from the command line. Instead, it uses Microsoft® Windows®-based user interfaces, which are more intuitive. Although making the discipline less daunting for newcomers, this development risks uncritical and uninformed use of techniques that are now no more than a few clicks away. Here we might be concerned with, and warn against, the problems of naïve disease mapping, such as ignoring basic cartographic principles or, in the case of cluster investigations, adopting methods not suitable for the data being examined.

GIS is very good at exploratory visualization, but confirmatory (analytical) modeling has not developed as fast, which is often why specialized statistical packages are brought in. Ideally, data might be visualized in GIS prior to analysis within specialist cluster-detection software or routines. GIS analysis needs to develop well beyond throwing a few concentric circles or buffers around a site under investigation (Gatrell, 1999), but, dangerously for the newcomer, these are the most accessible tools available.

4.5.2 Data Quality

In studies investigating disease clustering, the quality of incidence data is paramount. Often data collection anomalies, on closer inspection, have been the cause of many clusters. Data might be inaccurate geographically or temporally, biased or missing. In instances where just one or two more cases can alter the result of the statistical test, it goes without saying that capturing all cases is highly desirable, if not a prerequisite. The use of accurate registers of disease or national death certificate registers can greatly help, but again, the quality of registers does differ. In small isolated studies, often triggered by concerns by the public, it is particularly vital to ensure that complete case ascertainment has been achieved.

Although a concern is minimizing false negatives, minimizing false positives through diagnostic inaccuracies is also an issue. This differential diagnosis is not easily handled without closely inspecting the data and perhaps even returning to the case notes of the patients. When studies involve thousands of cases, perhaps diagnosed by hundreds of doctors, this is naturally not practical and it should be noted that for some diseases, diagnosis is not clear-cut. As important as case ascertainment is accurate estimation of the underlying population at risk against which cases are compared. This is a nontrivial problem and is considered in greater depth later in this section.

Once cases have been identified, geocoding or address matching becomes an issue. Here an attempt is made to match as accurately as possible the given address to geographical space, whether latitude and longitude, census areas, or postal or zip codes. Inevitably, errors can be introduced at this stage and some cases will remain unmatched and therefore lost to the analysis.

Balancing the desire to improve healthcare through historical analysis of health events with the individual's right to privacy poses a severe limitation on the types

of spatial analysis one can achieve. This fear of violating medical confidentiality often results in individual level data not being released to researchers. In this case, out of necessity, aggregated methods must be used.

4.5.3 Geography

Studies of disease clustering are often criticized due to the way in which boundaries of space and time have been chosen. Methodologies should be adopted that are "as released as possible from preclusions in the shape of artificial spatial and temporal units and population aggregates" (Schærström, 1996). To avoid the problems of "boundary shrinkage" (Openshaw, 1984), geographic scope should be defined at the outset. Otherwise, there is the temptation to fit the study region to the desired result because the tighter the boundaries chosen around the cluster, the higher the risk will be relative to the population at risk.

When considering count or aggregated data, investigators should be aware of the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984). One component of MAUP is where a possibly false interpretation is made from analyses purely because of arbitrary aggregations of the data. This arises because data is often collected and aggregated for convenience by local administrative, postcodes or zip codes, or health areas, borders that diseases are reluctant to respect. A second issue concerns the scale of aggregation. At one level of aggregation (e.g., postcode sectors of 1000 population), no significant result may be observed, whereas if the data were reaggregated (e.g., to postcode districts of 5000 people), an effect may be observed. The data remains the same, all that changes is the scale of analysis.

Analyzing data at the individual level using point data with no spatial aggregation has obvious advantages in attempting to overcome some of these traditional concerns of MAUP. However, there remain scale issues, such as the degree of smoothing to adopt in point methods such as kernel estimation.

4.5.4 Residential Migration

Many studies examining associations between geographical patterns of disease and causal factors assume that current residence in an area can be equated with exposure to conditions that currently (and historically) pertain there (Bentham, 1998). This is important, since epidemiologists and geographers often adopt the place of residence at the time of diagnosis or death as the location for further analysis of the disease in question. Yet, people move and hence previous exposure to causative agents will not be included in the study. The problems will be greater for diseases such as many cancers that have a long lag or latency period, which allows plenty of time for mobility of the population. By adopting only the current residential address, not only will an individual's migration history be neglected, but also the daily activity spaces of the patient will be ignored, where perhaps occupational exposure might be crucial.

4.5.5 Statistical Issues

In addition to method-specific statistical issues, there remain some generic statistical considerations relevant to all geographical investigations.

A common issue arising in cluster detection is the problem of multiple hypothesis testing. Here, if any chosen method repeatedly tests multiple hypotheses with the same data in the same geographical area, there is a high probability that at least some of the tests will be spuriously significant. For example, if one performed 1000 tests, while adopting a significance level of 0.01, then a false-positive result might be expected to occur ten times. Clearly, this raises the possibility of accepting a falsely positive cluster.

In studies of rare diseases in sparsely populated areas or simply where there are few cases, the small number problem may arise (Kennedy, 1989). This is where a difference of perhaps just one or two cases can make a huge difference to rates. In these cases, cluster methods rarely have enough data with which to generate sufficient power to test specific hypotheses, although some, such as the scanning window type, are better at dealing with sparse data.

Within the medical profession, there is a heavy reliance on formal significance testing. We would like to challenge the statistical fallacy whereby a lack of formal significance is equated with a lack of effect. Ritual testing of hypotheses is often performed without adequate consideration over whether or not this is appropriate (Nester, 1996), which has led to a preoccupation with or blindness toward *p*-values. Scientific method or the biomedical tradition appears to dictate that hypothesis testing is integral to good science; however, is the positivist model of significance testing valid in all disease clustering investigations? In studies with large sample sizes, statistical significance is relatively easy to attain due to its dependence on sample size, making the acceptance of the hypothesis relatively meaningless. Whereas when we have small number problems, say, with three isolated cases of a rare disease in a rural area, significance might well be impossible to achieve. To the concerned local resident, does the "significant" cluster appear any more important than an "insignificant" one that happens to occur in the workplace? It is suggested that estimation of relative risk to examine the strength of the relationship might be a way to proceed.

It is also well worth repeating the mantra that statistically significant findings do not equate with the establishment of causal relationships in cluster detection studies. Most clustering work is essentially exploratory spatial data analysis, where specific causal hypotheses are not even tested. Causal mechanisms can only be established with follow-up studies, perhaps triggered by the initial cluster investigation.

4.5.6 Confounding

Confounding is the single most important problem affecting cluster studies (Wartenberg and Greenberg, 1993). Any spatial or temporal variation of confounders (also known as covariates), such as demographics (e.g., gender, age, ethnicity), lifestyle characteristics (e.g., diet, smoking behavior or physical activity), or simply population density (considered separately later), can mask or exacerbate real disease patterns. To illustrate this concept, consider the idea that people of similar ethnic origin have traditionally tended to live close together, although this is less the case now with increased population migration. Because we know that some diseases are

inherited, one would expect to observe spatial clusters of genetic diseases. Thus, studies investigating excess clusters in these diseases would have to examine evidence for clustering over and above that exhibited in the background (ethnically skewed) population, after adjusting for the genetic confounding effect.

Ideally, methods such as Kulldorff's SaTScan should be adopted to adjust for confounding. If a case control-study design is adopted, however, known and unknown confounders could potentially be accounted for at the risk of adjusting for a real-but-unknown etiological effect.

4.5.7 Correcting for Population Density

A special case of confounding concerns the underlying or background population at risk. Correcting for spatial variations in the population at risk is an integral part of spatial epidemiological research, simply because any observed patterning of health events needs to be tempered by the underlying population distribution due to the heterogeneously distributed pattern of population settlement where most individuals congregate together in cities.

The suitable selection of an appropriate correction method is a nontrivial research problem. One often applied method in point-pattern analysis is to define a suitable set of controls from the at-risk population. By comparing the spatial arrangement of the observed cases with that of the controls, a relative risk estimate can be produced that has adjusted for the underlying inhomogeneity. Defining the control set as an accurate sample or representation of the whole underlying population distribution then becomes an issue, because inaccuracies introduced here might reveal false results in the analysis.

A fundamentally different method makes use of cartograms (or density equalized map projections) (Koch and Denike, 2001; Schulman et al., 1988) to correct for the heterogeneity in the background population data. Here, if one were to transform the disease data by the geography of the underlying population density or some measure of the population at risk, one would produce a distorted space map with the disease events still in their relative positions.

With aggregated count data, adjustment for variations in the underlying population density is normally achieved using figures obtained from a population census. In the United Kingdom and United States, the census occurs every ten years and thus may not reflect population changes and migrations between censuses. Censuses are subject to underenumeration, which, if left uncorrected, result in elevated disease risk estimates, particularly among those most difficult to enumerate and most susceptible to disease: the homeless, the most deprived, the young, and the elderly.

4.6 CHALLENGES

This review aimed to highlight major issues regarding cluster detection. It has demonstrated that the discipline remains of great importance in public health studies, with both active surveillance and *post hoc* confirmatory modeling being useful modes of investigation. However, there remain some significant outstanding methodological and technical challenges if the fields of cluster detection and GIS are to develop

their relationship beyond anything other than a loose coupling. Here we present our wish list, extending Gatrell (1999) while concentrating on cluster detection.

One key advantage GIS has over bespoke cluster detection software is its ability to integrate complex (environmental) datasets with health data. Where GIS falls down currently is in its ability to assimilate cutting-edge cluster detection models within its framework. Is there merit in attempting to integrate these cluster models within the accepted GIS user interfaces to bring sound methodology to a wider audience?

We have already discussed the validity of the current residential address as an adequate geographical marker for disease exposure. Many, perhaps most, of us spend much time at work or at school and may be as exposed to potential environmental contamination there as at home. The occupational setting may be at least as important as the residential. Historical exposures may be more relevant in diseases with long latency periods. We also need to conduct research that looks at the practicality and the relative merits of collecting and using data on space–time activity on a daily scale. Related to this, we must do much more research that builds knowledge of migration paths into our analyses. We remain convinced that GIS has a role to play here, a view that is endorsed conceptually by Löytönen (1998) and Schærström (1996) and demonstrated technically and empirically by Sabel et al. (2000).

Further work is also needed on the coupling of physical models to GIS, whether concerned with air or water contamination. The key to successful environmental epidemiology is access to good data on the exposure of interest. Too often, such environmental information is only available at a crude geographical scale, which leads to some rather optimistic attempts to link exposure to health outcomes. Remember that disease incidence or exposures may not follow a simple circular pattern (as assumed by many cluster detection methods), but rather may follow air dispersion plumes, drainage basins, or ethnic population boundaries. At the very least, we need as far as possible to collect data on potential confounders; much of this essential information is missing from routine, geocoded databases and is only available via large-scale surveys. It is, however, essential if we are to exploit GIS in an epidemiological context.

We need to be much more sensitive to issues of error and scale and resolution problems. This also ties in with questions of exposure assessment. What error bars are attached to our locational and attribute data, whether the data are environmental or social? At what scale should we conduct our investigations? Are our data available at a suitably fine level of resolution? For example, if we are attempting to investigate the link between radon and lung cancer or electromagnetic fields and childhood leukemia, surely we can only make real progress if data are collected for individuals and properties that have locational coordinates at a resolution of less than ten meters.

Finally, we need to challenge the implicit assumption of cluster studies that single exposures lead to single outcomes. Should we not, as Wartenberg and Greenberg (1993) suggest, consider more realistic situations where single exposures can result in multiple outcomes and in which diseases are considered multifactorial, and thereby adapt our models and conceptual thinking accordingly?

References

Alexander, F.E. and Boyle, P., 1996, *Methods for Investigating Localised Clustering of Disease,* Lyon, France: IARC Scientific Publications.

Alexander, F.E. and Cuzick, J., 1992, Geographical and environmental epidemiology: Methods for small-area studies, in Elliott, P. et al., Eds., *Methods for the Assessment of Disease Clusters*, Oxford: Oxford University Press.

Aylin, P. et al., 1999, A national facility for small area disease mapping and rapid initial assessment of apparent disease clusters around a point source: The U.K. Small Area Health Statistics Unit, *Journal of Public Health Medicine*, 21, 289–298.

Bailey, T.C. and Gatrell, A.C., 1995, *Interactive Spatial Data Analysis,* Harlow, U.K.: Longman.

Bentham, G., 1998, Migration and morbidity: Implications for geographical studies of disease, *Social Science & Medicine*, 26, 49–54.

Bithell, J.F., 1990, An application of density estimation to geographical epidemiology, *Statistics in Medicine*, 9, 691–701.

Diggle, P.J., 2000, Overview of statistical methods for disease mapping and its relationship to cluster detection, in Elliott, P. et al., Eds., *Spatial Epidemiology: Methods and Applications*, Oxford: Oxford University Press.

Elliott, P. et al., Eds., 2000, *Spatial Epidemiology: Methods and Applications,* Oxford: Oxford University Press.

Fotheringham, A.S., Brunsdon, C., and Charlton, M., 2000, *Quantitative Geography: Perspectives on Spatial Data Analysis*, London: Sage Publications.

Gardner, M.J., 1993, Investigating childhood leukaemia rates around the Sellafield nuclear plant, *International Statistical Review*, 61, 231–244.

Gatrell, A.C., 1999, GIS in public and environmental health: Visualisation, exploration and modelling, available at http://geog.queensu.ca/h_and_e/healthandenvir/gatrell.html

Hjalmars, U. et al., 1996, Childhood leukaemia in Sweden: Using GIS and a spatial scan-statistic for cluster detection, *Statistics in Medicine*, 15, 707–715.

Kelsall, J.E. and Diggle, P.J., 1995, Non-parametric estimation of spatial variation in relative risk, *Statistics in Medicine*, 14, 2335–2342.

Kennedy, S., 1989, The small number problem and the accuracy of spatial databases, in Goodchild, M. and Gopal, S., Eds., *Accuracy of Spatial Databases*, London: Taylor and Francis.

Knox, G., 1989, Detection of clusters, in Elliott, P., Ed., *Methodology of Enquiries into Disease Clustering*, London: Small Area Health Statistics Unit.

Koch, T. and Denike, K., 2001, GIS approaches to the problem of disease clusters: A brief commentary, *Social Science & Medicine*, 52, 1751–1754.

Kulldorff, M., 1998, Statistical methods for spatial epidemiology: Tests for randomness, in Gatrell, A.C. and Löytönen, M., Eds., *GIS and Health: GISDATA 6*, London: Taylor and Francis.

Kulldorff, M., 2002, SaTScan v. 3.0: Software for the Spatial and Space-Time Scan Statistics, Information Management Services Inc., Bethesda, MD: National Cancer Institute.

Kulldorff, M. et al., 1997, Breast cancer in northeastern United States: A geographical analysis, *American Journal of Epidemiology*, 146, 161–170.

Lawson, A.B. and Kulldorff, M., 1999, A review of cluster detection methods, in Lawson, A.B. et al., Eds., *Disease Mapping and Risk Assessment for Public Health*, New York: Wiley.

Löytönen, M., 1998, GIS, time, geography and health, in Gatrell, A.C. and Löytönen, M., Eds., *GIS and Health: GISDATA 6*, London: Taylor and Francis.

Nester, M.R., 1996, An applied statistician's creed, *Applied Statistics*, 45, 401–410.

Openshaw, S., 1984, *The Modifiable Areal Unit Problem: CATMOG 38*, Norwich, U.K.: Geo Books.

Openshaw, S. et al., 1987, A Mark 1 geographical analysis machine for the automated analysis of point data sets, *International Journal of Geographical Information Systems*, 1, 335–358.

Openshaw, S. et al., 1999, Testing space-time and more complex hyperspace geographical analysis tools, in *GISRU.K. '99*, Southampton, U.K.: University of Southampton.

Rothman, K.J., 1990, Keynote presentation: A sobering start for the cluster buster's conference, *American Journal of Epidemiology*, 132, S6–13.

Sabel, C.E., 1999, GIS, environmental exposure and health: An exploratory spatial data analysis of motor neurone disease, Ph.D. thesis, Lancaster, U.K.: University of Lancaster.

Sabel, C.E. et al., 2000, Modelling exposure opportunities: Estimating relative risk for motor neurone disease in Finland, *Social Science & Medicine*, 50, 1121–1137.

Sabel, C.E. et al., 2003, The spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death, *American Journal of Epidemiology,* 157, 10, 898–905.

Schærström, A., 1996, Pathogenic paths? A time geographical approach in medical geography, Ph.D. thesis, Lund, Sweden: Lund University Press.

Schulman, J., Selvin, S., and Merrill, D.W., 1988, Density equalized map projections: A method for analyzing clustering around a fixed point, *Statistics in Medicine*, 7, 491–505.

Viel, J-F., 2000, Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels, *American Journal of Epidemiology*, 152, 13–19.

Wakefield, J. and Elliott, P., 1999, Issues in the statistical analysis of small area health data, *Statistics in Medicine*, 18, 2377–2399.

Wakefield, J.C., Kelsall, J.E., and Morris, S.E., 2000, Clustering cluster detection and spatial variation in risk, in Elliott P. et al., Eds., *Spatial Epidemiology: Methods and Applications,* Oxford: Oxford University Press.

Waller, L.A., 2000, A civil action and statistical assessments of the spatial pattern of disease: Do we have a cluster? *Regulatory Toxicology and Pharmacology*, 32, 174–183.

Wartenberg, D., 2001, Investigating disease clusters: Why, when, and how? *Journal of the Royal Statistical Society Series A, Statistics in Society*, 164, 13–22.

Wartenberg, D. and Greenberg, M., 1993, Solving the cluster puzzle: Clues to follow and pitfalls to avoid, *Statistics in Medicine*, 12, 1763–1770.

Section 2

GIS Applications in Communicable Disease Control and Environmental Health Protection