University of Alberta

Applied and Methodological Geographies of Disease Cluster Detection

by

Nikolaos William Yiannakoulias 

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Department of Earth and Atmospheric Sciences

Edmonton, Alberta
Fall, 2006

Your file Votre référence
ISBN: 978-0-494-23132-6
Our file Notre référence
ISBN: 978-0-494-23132-6

Abstract

In this study we consider methodological and applied issues in geographic disease cluster detection. In the early chapters of our study, we discuss rationale for using cluster detection methods. In Chapter 2, we outline features of geographic health surveillance, and suggest that some cluster detection methods offer important advantages over other methods of spatial analysis when used in surveillance applications. In Chapter 3, we present a normative rationale for using cluster detection methods in the context of chronic disease prevention. In the last two chapters, we describe and evaluate an approach to cluster detection that is able to find clusters of irregular shape. In Chapter 4, we present the method in detail and evaluate it in the context of simulated data. In Chapter 5, we observe the method's performance with respect to real disease data. We conclude the study by arguing that our new approach may be of use when applied in conjunction with more traditional techniques of disease cluster detection.

Acknowledgement

My first acknowledgement goes to my supervisor, Dr. John Hodgson. John's guidance and mentorship was not only important to my development as a student and academic, but as a human being. I will be forever thankful. Thanks also to members of my supervisory committee. Thanks to Dr. Donald Schopflocher who has always been willing to philosophize, share ideas, and give thoughtful advice. Thanks to Dr. Rhonda Rosychuk for her careful readings and comments on my thesis, and to Dr. Karen Tomic for the support and help that she provided over my graduate career. Thanks also must go to my thesis examiners, Dr. Phillipe Erdmer and Dr. Martin Kulldorff. Special appreciation must go to Mr. Larry Svenson for his support and mentorship throughout my academic career, as well as providing data required for the completion of my dissertation.

On a personal level, thanks to my family and friends, and all the people who have helped and supported me over the length of my graduate studies.

Dedication

For my mom.

Table of Contents

CHAPTER 1: Introduction - 1 -

1.1. Philosophical setting: methodology and pragmatic context - 1 -

1.2 Rationale and Outline - 3 -

1.3. References - 6 -

CHAPTER 2: A review of geographic health surveillance methodology - 7 -

2.1 Introduction - 7 -
2.1.1 Public health surveillance and explanatory research - 8 -
2.1.2 Public health surveillance and data mining - 10 -

2.2 Geographic health surveillance - 12 -
2.2.1 General applications of geographic health surveillance - 12 -
2.2.2. Identifying new potential risks - 13 -
2.2.3. Information for public health policy decisions (decision support) - 14 -
2.2.4. Information for communities (community support) - 14 -
2.2.5 Geographic alarm system for infectious disease outbreaks and bioterrorism - 15 -

2.3 Geographic health surveillance methodology - 16 -
2.3.1 Geographic surveillance methods: disease mapping - 16 -
2.3.1.1. Rate maps - 17 -
2.3.1.2. Probability Maps - 17 -
2.3.1.3 Topological smoothing - 18 -
2.3.1.4 Empirical Bayes methods of disease mapping - 20 -
2.3.1.5 Spatial models - 21 -
2.3.2 Geographic surveillance methods: cluster detection - 23 -
2.3.2.1 Distance-based approaches - 24 -
2.3.2.2 Quadrat approaches - 25 -

2.4 Discussion - 30 -
2.4.1 Modifiable Areal Unit Problem (MAUP) - 33 -
2.4.2 Diagnostic inconsistency over time and space - 33 -
2.4.3. Clinical versus statistical significance - 34 -

2.4 Conclusion - 34 -

2.6 References - 38 -

3.1 Introduction - 53 -

3.2 Disease prevention in public health - 55 -
3.2.1 Disease cluster detection - 55 -
3.2.2 Population-based disease prevention and the problem of scope - 58 -

3.3 Population-based utilitarianism - 61 -

3.4 The difference principle in disease prevention - 63 -

3.5 A geographic contribution to chronic disease prevention methodology - 70 -
3.5.1 Comparison of theory to methodology - 70 -
3.5.2 Example applications - 74 -
3.5.2.1 Cluster-communities as proxy geographies - 74 -
3.5.2.2 Cluster-communities as formal geographies - 76 -
3.5.2.3 Cluster-communities in facility location - 77 -

3.6 Conclusion - 78 -

3.7 References - 80 -

CHAPTER 4: Comparison of a structured and data-directed cluster search strategy - 92 -

4.1 Introduction - 92 -

4.2 Background - 93 -
4.2.1 Clusters and cluster detection - 93 -
4.2.2 The spatial scan approach to cluster detection - 94 -
4.2.3 Structural and data-directed searches - 99 -

4.3 Methods - 101 -
4.3.1 A greedy hot-graph search - 101 -
4.3.2 The experiment - 103 -
4.3.2.1 Power estimation - 105 -
4.3.2.2 Geographic precision: graphs - 105 -
4.3.2.3 Geographic precision: maps - 106 -

4.4 Results - 107 -
4.4.1 Detection of significant clusters - 107 -
4.4.2 Sensitivity and Positive Predictive Value - 107 -
4.4.3 Maps of true positives and false positives - 108 -

4.5 Discussion - 109 -
4.5.1 General findings - 110 -
4.5.2 Operational issues related to the hot-graph method - 111 -
4.5.3 Comparison of the methods - 113 -
4.5.4 Study Limitations - 114 -

4.6 Conclusion - 116 -

4.7 References - 140 -

Winston W L, Venkataramanan M (2003*) Introduction to Mathematical Programming.* fourth edition, Thomson Brooks/Cole: Pacific Grove. - 143 -

CHAPTER 5: Geographic discovery through cluster detection - 144 -

5.1 Introduction - 144 -

5.2 Methods - 146 -
5.2.1 Data - 146 -
5.2.2 Approach - 147 -

5.3 Results - 149 -

5.4 Discussion - 150 -
5.4.1 Differences between methods - 150 -
5.4.2 Hot-graph searches in cluster detection and geographical analysis - 152 -

5.5 Conclusion - 155 -

5.6 References - 169 -

Yiannakoulias N, Svenson L W, Schopflocher D P (2005). Diagnostic uncertainty and medical geography: what are we mapping? *The Canadian Geographer* 49:291-300.CHAPTER 6: Conclusion - 172 -

CHAPTER 6: Conclusion - 173 -

6.1 Summary of chapters - 173 -
6.1.1 A review of geographic health surveillance methodology (Chapter 2) - 173 -

6.1.2 Clusters as unfairness: geographic analysis in chronic disease prevention (Chapter 3) - 173 -
6.1.3 Structured and data-directed cluster search strategies (Chapter 4) - 173 -
6.1.4 Geographic discovery through cluster detection (Chapter 5) - 174 -

6.2 Discussion of the research project - *175* -

6.3 Future Research - *177* -
6.3.1 Detection of severity clusters - 177 -
6.3.2 Cluster detection within operational constraints - 179 -
6.3.3 Cluster detection and the modifiable areal units problem - 180 -
6.3.4 Comparing statistical and cognitive heuristic approaches to cluster detection - 181 -

6.4 References - *183* -

7. APPENDICES **- 186 -**

List of Tables

Table 2.1 A schematic table of geographic surveillance methods and application - 36 -

Table 5.1 Selected diseases - 156-

Table 5.2 Tabulation of cluster detection results - 156-

List of Figures

Figure 2.1 Crude rates of prevalent Parkinson's disease using 3 common colour schemes - 37 -

Figure 4.1 A graph representation of a corresponding polygon tessellation - 118 -

Figure 4.2 A conceptual illustration of the spatial scan - 119 -

Figure 4.3 An operational illustration of the spatial scan - 120 -

Figure 4.4 An illustration of the hot-graph method - 121 -

Figure 4.5 The simulated study area - 122 -

Figure 4.6 The four cluster patterns - 123 -

Figure 4.7a Proportion of tests correctly identifying the presence of a simulated circular cluster - 124 -

Figure 4.7b Proportion of tests correctly identifying the presence of a simulated 'ring' cluster - 125 -

Figure 4.7c Proportion of tests correctly identifying the presence of a simulated 'line' cluster - 126 -

Figure 4.7d Proportion of tests correctly identifying the presence of a simulated 'network' cluster - 127 -

Figure 4.8a Mean sensitivity for circle cluster pattern - 128 -

Figure 4.8b Sensitivity for ring cluster pattern - 129 -

Figure 4.8c Sensitivity for simulated 'network' cluster - 130 -

Figure 4.8d Sensitivity for line cluster pattern - 131 -

Figure 4.9a Positive predictive values for circle cluster pattern - 132 -

Figure 4.9b Positive predictive values for ring cluster pattern - 133 -

Figure 4.9c Positive predictive values for network cluster pattern - 134 -

Figure 4.9d Positive predictive values for line cluster pattern - 135 -

Figure 4.10a Circle cluster pattern: proportion of true positives and proportion of false positives - 136-

Figure 4.10b Ring cluster pattern: proportion of true positives and proportion of false positives - 137 -

Figure 4.10c Line cluster pattern: proportion of true positives and proportion of false positives - 138 -

Figure 4.10d Network cluster pattern: proportion of true positives and proportion of false positives - 139 -

Figure 5.1 Province of Alberta and major communities - 157 -

Figure 5.2 Bounding-box centroids and sub-rha boundaries - 158 -

Figure 5.3 Maps of identified clusters: asthma - 159 -

Figure 5.4 Maps of identified clusters: food poisoning (other, bacterial) - 160 -

Figure 5.5 Maps of identified clusters: diabetes - 161 -
Figure 5.6 Maps of identified clusters: hypertension - 162 -
Figure 5.7 Maps of identified clusters: Parkinson's - 163 -
Figure 5.8 Maps of identified clusters: influenza - 164 -
Figure 5.9 Maps of identified clusters: illness related to alcohol/drug use - 165 -
Figure 5.10 Maps of identified clusters: food poisoning (salmonella) - 166 -
Figure 5.11 Maps of identified clusters: giardiasis - 167 -
Figure 5.12 Maps of identified clusters: gonorrhoea - 168 -

CHAPTER 1: Introduction

1.1. Philosophical setting: methodology and pragmatic context

Traditional notions of scientific objectivity have been subject to considerable scrutiny over the last 50 years. Critical discussion of how comparative methodological studies are undertaken has received much of the attention. Some of the best known modern criticism comes from Kuhn (1962), who offers a critical view of how science perceives the advancement of new methodology. Small changes can occur within an accepted scientific paradigm, but methodological breakthroughs are not based on the accumulation of evidence towards better methodology, but occur in a framework of scientific crisis management. When a breakthrough does occur, the new method is chosen based on its elegance and 'neatness' as much as its ability to solve problems or answer questions.

Though these ideas continue to fuel epistemological debates in the philosophy of science, the most lasting observe that science is influenced by the participants—the scientists—in their choices of standards, subjects of study, methods of comparison and modes of communication. In the social sciences, this observation is particularly important. For example, many social scientists use language to acquire, analyze and describe information on human behaviour and human institutions. However, language is laden with uncertainties, complicating both the analysis and communication of certain types of social scientific research. This has led some to question the validity of the social sciences as sciences; to Popper, a notable portion of social theorizing is not testable (and more particularly, not falsifiable) and therefore, is not scientific or fruitful (Popper 1993). Winch (1958) comes to a similar conclusion, seeing the study of human society as a philosophical, rather than scientific exercise. Thinkers in 'critical social science' often put aside questions of truth and validity altogether, and instead, ask questions about power and motivation. Ambitiously, this often involves notions of ethics, and a suggestion that the scientific establishment must be challenged to meet particular normative goals.

As a result of these perspectives, methodological work in the social sciences is

especially sensitive to the challenges of post-modernism and critical science. There are many tools available to measure, evaluate and interpret information across research disciplines. Competing methods result in different research paradigms, and disagreements of approach; for example, interpretivists often dispute the numeric quantifications of positivists and holists often dispute the reductionism of individualists. Similar disagreements occur within particular intellectual domains. This is particularly well illustrated in the health sciences. Some define diseases based on descriptions of behaviour, while others analyze physiology, chemistry and genetics. Some research focuses on finding clinical causes of illness, other research focuses on finding causes in the social environment. Some methods conform to strict standards of experimentation (like randomized control trials) and others consist of interpretive theorizing. These differences of method have obvious effects on how states of health are identified, treated, prioritized and understood. Sometimes these differences in method translate into important differences in theory, and disputes of fact.

According to some, disagreements of approach present an intractable challenge; science has no capacity to rule on what methods should be used to approach a problem, even within a particular scientific paradigm. For Kuhn, different methods are often 'incommensurable'; there is no neutral standard to which practitioners of different methodologies can compare their respective approaches. This is because the selection of the standard of measurement is itself part of a scientist's methodological perspective or paradigm. For Quine, debates of approach are not formed accidentally; he argues that scientists are selective about the hypotheses they are willing to test, and careful about which methodologies they scrutinize. Scientists often behave in a manner that 'minimally mutilates' their theories; they preferentially reject methods and evidence that are least important to upholding their theories (Quine 1991 pp.14).

One way to manage these criticisms and still engage in systematic methodological study is to accept a pragmatic approach, which uses application, rather than 'truth', as a framework of evaluation. Though pragmatism (sometimes referred to as instrumentalism) comes in many forms, most pragmatists hold a

view that ideas and methods should be judged based on suitability to human affairs and fulfillment of human needs (subjective though they may be). Lewis, for example, argues that there are many conceivable systems of mathematics and logic, and that the justification of the current system is not based on a formal evaluation of its properties, but based on "intellectual convenience." (Lewis 1923). Similarly, James defines scientific truth as "what is good in way of belief"; that is, whatever is satisfying or in agreement with things that are considered important is true (Rorty 1982). Some natural scientists accept a pragmatic interpretation of scientific methodology. Hawking writes:

> I take the positivist viewpoint that a physical theory is just a mathematical model and that it is meaningless to ask whether it corresponds to reality. All that one can ask is that its predictions should be in agreement with observation. (Hawking and Penrose 1996 pp.4)

Hawking's position is clearly pragmatic (in spite of his probable misuse of the word 'positivist'). Theory is validated based on performance (in this case, the ability to predict observations) not a correspondence with truth. Theories that result in less error between observations and reality and that offer useful predictions are favoured, whatever their connection to the real world. This view does not represent a denial of the physical world, or even scientific truths, but acknowledges the difficulty of measuring such truths in a meaningful way independent of application and context. For the pragmatist, decisions are made within a chosen framework of 'normal science', and are evaluated with respect to applications within this paradigm.

1.2 Rationale and Outline

The broad purpose of this project is to investigate a methodology of disease cluster detection within two contexts of application. The next two chapters are a discussion of these contexts. In Chapter 2, I review various well-known methods of spatial analysis of disease in the context of geographic disease surveillance. The best geographic disease surveillance methods can simultaneously inform decision makers, the public and the academic community in a manner that is systematic, and relatively easy both to understand and explain. Recently,

researchers have acknowledged a role for geographic cluster detection in health surveillance. For example, public health systems have a duty to provide rapid response to serious infectious disease (and bioterrorism) and require tools to inform decision makers about the current and changing states of population health. In these applications, clusters help us identify where changes are occurring, and help inform decision makers in a timely and efficient manner. In Chapter 3, I discuss how geographic cluster detection can be justified for use in chronic disease prevention. I argue that from a particular standpoint of health equity (as opposed to health efficiency) chronic disease prevention resources should, at times, be distributed unequally to ensure that persons with the highest burden of illness are treated fairly. Cluster detection methods that identify worst-off spatial sets provide information that can be used in geographically specific disease prevention strategies, such as environmental modification and service facility allocation.

Following these conceptual chapters, I narrow my focus on issues related to cluster detection methodology. In Chapter 4, I present a simple approach designed to find geographic clusters of disease. This approach is based on the spatial scan statistic (Kulldorff and Nagarwalla 1995; Kulldorff 1997) but is constrained by adjacency, rather than circular geometry and is able to find clusters of disease with irregular shapes. I compare this adaptation to the spatial scan method with several simulated disease scenarios, and evaluate the power of the methods to find and precisely describe the shape of several simulated disease cluster patterns.

In Chapter 5, I evaluate these two methods on real disease data. My evaluation compares ten different diseases at a single resolution of observation, and observes how results differ between these two methods. I interpret the results in terms of statistical inference and geographical analysis. Based on these findings, I recommend the integration of these methods into a system of geographic cluster detection analysis.

In my concluding chapter, I briefly review the proceeding chapters, and comment on approaches to cluster detection methodology with respect to their

applications in geographic surveillance and geographic chronic disease prevention. Finally, I make four recommendations for future research related to the topics covered in this dissertation.

1.3. References

Hawking S, Penrose R (1996) *The Nature of Space and Time*. Princeton University Press: Princeton.

Kuhn T (1962) *The Structure of Scientific Revolutions*. University of Chicago Press: Chicago.

Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. *Statistics in Medicine* **14**:799-810.

Kulldorff M (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods* **26**:1481-1496.

Lewis C I (1923). A pragmatic conception of the a priori. *The Journal of Philosophy* **20**:169-177.

Popper K (1993) *The Poverty of Historicism.* Routledge: London

Quine W V (1991) *Pursuit of Truth*. Harvard: Cambridge.

Rorty R (1982). Pragmatism, relativism and irrationalism. In *Consequences of Pragmatism.* University of Minnesota Press: Minneapolis.

Winch P (1958). *The Idea of Social Science and its Relation to Philosophy*. Routledge and Kegan Paul: London.

CHAPTER 2: A review of geographic health surveillance methodology

2.1 Introduction

In spite of modern critiques of scientific objectivity, most health research aspires to follow a traditional process of scientific discovery—a refutation or confirmation of theory based on empirical evidence. In many of the social sciences, including fields of health research, experiments are often used to develop explanatory models based on observations of real-world phenomena. These explanatory models are the basis of both exploratory research and theory building. The merits of the experimental approach to theory building are frequently debated amongst philosophers, but even its harshest critics accept that empirical experimentation in the medical sciences has some instrumental value (whatever the 'truth' may be) in prolonging the length and quality of human lives.

In the field of public health administration, explanation and theory building are not of primary importance. Analytical work is required, but it most often deals with decision making issues—such as feasibility, cost effectiveness and priority setting. Although these analytical exercises are often informed by explanatory research based on experimentation, they require methodologies that support efficient, ethical and timely decision making. Within the health sciences, considerably more effort has been put into the development and examination of research methods that support the explanation paradigm rather than the decision making paradigm. Consequentially, decision makers are often forced to use the tools of the explanatory analyst when other, more specialised tools might be more appropriate.

Health surveillance is an important component of decision support and policy application in public health. The goals of surveillance are to collect data, monitor data systems and provide feedback that informs policy, the public, and the research community. The goals of *geographic* surveillance—to routinely monitor geographic data and disseminate geographic information—require a unique methodological perspective. In this review we discuss characteristics of

geographic health surveillance that distinguish it from spatial epidemiology, geography and most fields of quantitative explanatory study concerned with health and space. We then review a selection of methods in terms of their application to geographic health surveillance. In doing so, we hope to contribute to the understanding of methodology in geographic surveillance, and secondarily, illustrate the general importance of linking methodology to application.

2.1.1 Public health surveillance and explanatory research

The practice of public health surveillance goes back to the 17th century. Early examples involved routine collection and reporting on the health of the public, and the environment in which they lived (Thacker and Berkelman 1988). These monitoring systems differed fundamentally from John Snow's well known study of cholera in London. Snow was responding to evidence of an outbreak, whereas the contemporary surveillance activities in England, Germany and the United States were ongoing, and concerned with a large number of outcomes simultaneously rather than specific disease concerns (Declich and Carter 1994). The modern definition of health surveillance emphasizes the routine "collection, analysis and interpretation of outcome-specific data, closely integrated with the timely dissemination [...] to those responsible" (Thacker and Stroup, 1994).

Public health surveillance is motivated by different goals than other disciplines in the health sciences. Surveillance is most often dedicated to the identification and exploration of patterns rather than to the identification of causes and provisional explanations—which are of more concern in epidemiological and clinical research. As such, most health surveillance conforms to an exploratory and descriptive model of analysis. For surveillance purposes, understanding causes and processes is of secondary importance. Public health surveillance is concerned with detecting and describing changes in patterns of disease, and providing information useful for timely decision making. For example, where explanatory research may seek to understand why *E. coli* occurs in periodic local outbreaks, surveillance is concerned with where, when and if these outbreaks are occurring. Where explanatory research seeks to understand the relationship between ethnicity and health, surveillance is concerned with identifying deprived

or at-risk ethnic groups. Both approaches are important to managing public health, but surveillance is more directly connected to policy application and decision support.

There is a considerable literature describing surveillance programs (e.g. Lazarus et al., 2002; Beckett et al., 2004) and surveillance information systems (e.g. Nobre et al., 1997; Tsui et al., 2003; Wang, Ramoni, Mandl and Sebastiani 2005), however methodological literature has evolved more slowly. This may be explained by a belief that surveillance is a subdiscipline of epidemiology. Although there may be considerable overlap between these disciplines, many quantitative epidemiological tools are not well suited to the health surveillance paradigm. One of the goals of epidemiology, for example, is to study causality (or at least association) between determinants and health. One of the prerequisites of medical causality is biological plausibility (Hennekens and Buring 1987 pp. 40). To have a credible causal explanation that exposure to asbestos causes asbestosis, there must be a plausible explanation of how properties of exposure cause the symptoms of illness. In this case, small asbestos fibres build up in the lung causing scarring and a chronic restriction of lung capacity. In order to make a causal connection between poor lung functioning and asbestos exposure, it is important that disease definitions are as precise and specific as possible. This is particularly true when it is possible to identify a necessary cause—in this case, prolonged exposure to a respiratory irritant.

In surveillance, disease reporting is often more general. Rapid changes in disease patterns cannot always wait for laboratory testing or other clinical verification. More general indicators of change—such as an unusual incidence of similar symptoms presenting to a local emergency room—will provide more timely information even though causality may not be clear. For example, acute respiratory infections, bronchitis, influenza and pneumonia present similar clinical symptoms. Outside of laboratory tests, it is often difficult to separate one from the other—even though the aetiologies of these diseases are quite different. By restricting disease diagnoses to influenza (which can only be verified in the laboratory setting) the sample of outcomes may be precise, but it may also be too

small to make observations about important trends in illness, especially when data are reported over multiple dimensions—such as age, sex and geography. This is particularly true when the speed of response to an outbreak is important. By 'rolling-up' diseases into sensible symptomatically or causally similar groups, patterns are more visible, and furthermore, observations are less likely to be affected by diagnostic inconsistency (Yiannakoulias, Svenson and Schopflocher 2005). New research in syndromic surveillance even expands health monitoring beyond disease itself—using drug prescriptions, employee absenteeism and other indirect sources of information as indicators of changes in disease patterns (e.g. Tsui et al. 2003).

Another important difference between epidemiology and surveillance methodology relates to reporting. Many epidemiological studies report risks in terms of magnitude of effect, emphasizing measures such as odds ratios and regression model coefficients. These indicators can provide important evidence of the relationship between exposures and disease. Often surveillance is more concerned with incidence—over time, space, demographic and socioeconomic dimensions. Changes in rates of disease, patterns of variation that exceed expectations, and clusters of outcomes are all important in surveillance reporting. Surveillance systems must be able to offer information that is robust, and depending on the recipients of information, relatively easy to understand. The development and study of methods for distributing information that meet the needs of surveillance activities are unlikely to be inspired strictly by clinical or epidemiological research, whose audiences are often health care professionals and academics.

2.1.2 Public health surveillance and data mining

Many sources of data can be used for public health surveillance. Notifiable disease reporting systems, field investigations, routine laboratory testing and administrative health data are all useful resources (Declich and Carter 1994). Since surveillance data must be collected in an ongoing fashion, the quantity of data collected per person tends to be small (Thacker and Berkelman 1988) and the cost of obtaining them is usually inexpensive (Virnig and McBean 2001). The

type of data involved is usually related to the outcome under study; infectious disease surveillance often relies on notifiable disease reporting systems (Nelsen 2001) and chronic disease surveillance frequently relies on administrative health data (Thacker et al. 1995). Administrative data resources can offer a rich supply of health-related data. When updated in a timely manner, these large systems satisfy many of the requirements of population-wide public health surveillance. They are inexpensive, unobtrusive, reasonably comprehensive and readily accessible.

Within such large systems of data, observations of importance can be obscured by data that have little or no informative value. This problem is commonly discussed in the data mining literature (Fayyad, Piatetsky-Shapiro and Smyth 1996). In general terms, data mining is an automated process that searches through large systems of data for patterns of similarity and association. Data mining is often necessary when the data are too numerous for purely human-directed analysis. Usually, data mining does not presume to search for explanations of phenomena or offer direct support of theory, but rather, aims to automate processes that are beyond the capacity of (and have no need for) immediate human intervention (Fayyad, Piatetsky-Shapiro and Smyth 1996). The information gleaned from such processes provides the basis for a secondary analysis that informs decision makers and identifies new research questions. Data mining approaches are found in business (Bose and Mahapatra 2001), telecommunications (Cox et al. 1997), intelligence analysis (Xu and Chen, 2003), bioinformatics (Luscombe, Greenbaum and Gerstein 2001), astronomy (Voisin 2001) and even text analysis (Losiewicz, Oard and Kostoff 2000). Recently, data mining has become an important component of public health surveillance (Liao and Lee 2002; Obenshain 2004).

In several important ways, however, health surveillance is also different from traditional data mining. The operational issues that dominate methodological developments in data mining are often secondary to policy issues in health surveillance applications. Ethical concerns are of particular importance in health surveillance (Fairchild and Bayer 2004; Middaugh, Hodge and Cartter, 2004). In

addition to matters related to privacy, one must consider the social and ethical consequences of policy decisions made in the spirit of public health surveillance. What kinds of changes should motivate action? At what point are routinely identified inequalities of health considered deserving of formal concern? What diseases should be part of the routine surveillance processes? Answers to these questions have an impact on the economics of health care, and on the well being of society in general.

Another difference is the relationship between surveillance systems and the public, and in particular, the challenges of public dissemination. Traditional data mining is often centralised—outputs are in the hands of analysts and managers. Public health is often faced with accusations of paternalism (Beauchamp 1980; Hawks 1997; Bayer and Fairchild 2004), and is increasingly unable to engage in top-down decision processes without incurring political criticism. Public involvement is crucial to building credibility and trust in disease prevention and health promotion (Reed 1994). As a result, the challenge of surveillance is to extract and disseminate information that satisfies the needs of all stakeholders: managers, advocacy groups, the media and perhaps most importantly, the public.

Surveillance methods must be flexible enough to meet the operational and ethical requirements of public health. These requirements are not strictly epidemiological, nor entirely in the domain of data mining and knowledge discovery. Though recent work has significantly advanced the field of public health surveillance (e.g., Brookmeyer and Stroup 2003; Lawson and Kleinman 2005), further work is required to develop new, and refine existing, methods that meet the complex methodological and ethical demands of public health surveillance. The shortage of literature directly addressing health surveillance methods, particularly in chronic disease, continues to limit the advancement of surveillance applications in many areas. The remainder of this discussion narrows focus to the study of geographic health surveillance methodology.

2.2 Geographic health surveillance

2.2.1 General applications of geographic health surveillance

Like public health surveillance in general, geographic surveillance involves

routine collection, analysis and dissemination of health data, but specifically, health data in geographic space. A growing literature has developed methods useful for geographic surveillance (e.g. Rogerson 1997; Lawson, Clark and Vival-Rodeiro 2004; Kuldorff et al. 2005; Lawson 2005; Rogerson 2005). Unfortunately, little literature has formally evaluated these methods in terms of the particular goals of geographic surveillance.

There are at least four applications of geographic disease surveillance that are relevant to public health activities: identification of new potential risks, policy decision support, community support and alarm detection. Thacker and Stroup (1994) recommend an application not discussed here, that of building 'surveillance archives'. These archives are collections of historical surveillance data, and can be used in secondary research or in monitoring historical disease trends. For the most part, these archives are a composite of the other surveillance activities, and require no further discussion here.

2.2.2. Identifying new potential risks

Information obtained from geographic surveillance can help identify the presence of otherwise unknown risk factors in the physical or social environment. Notable geographic variation in disease and injury could indicate an abundance of hazards or shortage of protective factors. Structured research activities outside the scope of surveillance are required to more carefully investigate these observations when they occur. Clearly, feedback between surveillance systems and the research community is critical; surveillance systems can identify curious patterns or changes in disease, which experts can investigate more thoroughly. Information based on these investigations is then used to alter the focus of surveillance activities.

Considerable applied research in the social sciences explores the relationship between health and the environment. Examples include studies of air pollution and mortality (Jerrett et al. 2005), heat waves and various health-related outcomes (Smoyer-Tomic, Kuhn and Hudson 2003), insect-related illness (Mawby and Lovett 1998; Kazmi and Pandit 2001) and social determinants of disease (Mohan, Twigg, Barnard and Jones 2005). For such activities to constitute surveillance,

they must be part of an ongoing system of collection, analysis and dissemination. Some methods used in the study of health and the environment may not migrate to the surveillance setting. For example, some researchers have identified cancer clusters around sites of industry (Waller et al., 1994; Roberts, Steward and John 2003; Parodi et al. 2003) but these examples employ a cross-sectional design, and were not developed for ongoing observation. This is noteworthy since methods developed to answer isolated research questions may face multiple testing problems when used in routine reporting systems common to surveillance (Kulldorff 2001).

2.2.3. Information for public health policy decisions (decision support)

Routinely reported descriptive information on health and disease, independent of the factors that cause patterns to occur, are important in many different public health settings. This is particularly true for publicly administered health care systems in which resources are allocated based on demographic and health characteristics. Patterns of resource allocation and intervention may need to reflect patterns of disease, particularly when policy makers are concerned with equity issues. Geographic surveillance may also help inform resource allocation activities at more local levels. Mobile health resources (like needle exchanges, outpatient screening and immunization services) are likely to be more efficient when responsive to the geographic distribution of sick or at-risk populations, which may change over time.

Adjusting policy to accommodate change in public priorities also requires routine health information. The health of particular subgroups of the population—such as children and the aged—may receive more or less policy and public emphasis over time. Certain diseases may also receive more or less emphasis over time. In either case, shifting resources to a new priority area is more timely when important information on the disease is already part of a routine system of data collection and dissemination.

2.2.4. Information for communities (community support)

Modern discussion of the relationship between technology and society has considered a revitalised role for communities and citizen action in public decision

making (Guthrie and Dutton 1992; Morrow 1999). Providing the data and analytical resources to communities has been an important feature of public participation geographic information systems (PPGIS) (Aberley and Sieber 2005). A well designed surveillance system ensures that information is disseminated in a form that is accessible to the public and that the information is provided in an manner that facilitates fair comparisons between communities. For some communities, information about risk may be less influential than a general mistrust of public officials (Freudenburg and Rursch 1994). In addition to providing resources for community action, routinely collected information may help build public trust.

Protocols for responding to local health concerns must strike a balance between community and broader social interests; public health interventions must manage competing interests of different communities. False alarms can be wasteful, and a drain on communities where genuine need is greater. On the other hand, it is irresponsible to ignore community concerns altogether (Center for Disease Control 1990; Bender et al. 1990; Wartenberg and Greenberg 1992). A mature and well examined system of routine geographic surveillance could provide critical information to permit open and timely communication between government and stakeholders.

2.2.5 Geographic alarm system for infectious disease outbreaks and bioterrorism

Routine geographic surveillance is an important component of communicable disease control. Localised outbreaks of infectious illness may require immediate intervention measures to prevent spread of disease to the population as a whole. Recently, this has included the development of methods for prospective geographic disease surveillance (Kulldorff 2001; Mostashari et al. 2003) and the application of temporal surveillance tools in the spatial setting (Rogerson 2005). In these and similar applications, the emphasis has been on outbreak detection—with a particular emphasis on bioterrorism, and new disease outbreaks (such as West Nile virus and new strains of influenza).

These surveillance methods are not always as useful for detecting where future infectious disease outbreaks will occur over longer periods of time,

however. For spatially dependent processes, prediction requires understanding of spatial diffusion—which is difficult to account for in many surveillance activities—and predicting where a seminal event will occur—which is often impossible. Alternative approaches, such as those being developed in the field of cellular automata, may offer important opportunities for predicting trends in infectious disease and other dependent spatial processes.

2.3 Geographic health surveillance methodology

Much of the following discussion concentrates on methods that apply to discrete aggregate (polygon-based or centroid) rather than disaggregate atomic (point) geographic representations of data (see Cressie 1993 pp.10-13 for a discussion of spatial data types). In the study of disease, atomic point data usually represent the locations of persons with and without a condition of interest. In the case of aggregate data, the representation usually consists of a geographic boundary (in the case of polygons) or an average of locations (in the case of centroids), of more than a single individual. By concentrating on geographically aggregate data, some important methodological developments in the spatial analysis of disease (e.g., Cuzick and Edwards 1990; Diggle 1990; Diggle and Chetwynd 1991) will not be covered in this discussion. We justify our emphasis on aggregate methods on the grounds that this is the form in which a large quantity of health data are collected and stored, particularly in passive surveillance systems. Though some information administrators may have access to data at the level of individuals, privacy concerns often prohibit release of such detail outside the control of data custodians—even within a governmental agency.

2.3.1 Geographic surveillance methods: disease mapping

Disease maps convey visual information about how disease varies over space. The earliest known maps of disease come from the 18th and 19th century, and consisted of dots representing locations of infectious diseases like cholera and yellow fever (Walter 2001). Disease mapping literature has grown considerably in recent years. Marshall (1991a) provides a review of statistical and other methodological concerns. More up to date discussion can be found as well (Lawson et al. 2000; Bithell 2000; Lawson 2001a; Rushton 2003). Several

recently published books have been dedicated in part or entirely to describing and reviewing disease mapping techniques (Lawson et al. 1999; Elliott et al. 2001; Lawson 2001b; Waller and Gotway 2004).

2.3.1.1. Rate maps

Next to dot maps, one of the simplest methods of representing disease occurrence involves mapping disease rates. Some of the earliest rate maps reported mortality statistics, particularly of cancer (see Walter 2001 for a review). The simplest approach to rate estimation is to divide cases of disease by a population at risk. Age and sex standardisation is common in rate mapping—particularly in epidemiology (Hennekens and Buring 1987 pp. 54-98). Standardisation allows one to factor-out known (typically demographic) characteristics that could obscure more interesting geographic variations in the disease rate. Direct standardization is preferred when stratum specific rates are large enough (and available), otherwise the indirect method can be appropriate. Estimating standard errors of standardized rates can be complicated and traditional methods tend to produce standard error estimates that are artificially small (Carriere and Roos 1994).

Rate maps remain one of the most popular methods of representing the geographic characteristics of disease, especially for public consumption (for a current example, see http://dsol-smed.phac-aspc.gc.ca/dsol-smed/cancer/n_prov_e.phtml). Though rates only report historical information (what has happened) they are frequently used as informal tools for anticipating where disease is likely to occur in the future. Unfortunately, rates can be very unreliable indicators of underlying patterns, especially when a disease is rare. This unreliability is often a function of stochasticity observed in Poisson distributed data; when a disease is rare, small changes in the number of cases can result in large changes in rates. This is often referred to as the 'small numbers problem'.

2.3.1.2. Probability Maps

In response to the small numbers problem, Choynowski introduced the probability map (Choynowski, 1959). His method uses the Poisson distribution to

obtain probability estimates for disease counts (corrected for variations in population) for each area on a map. Maps of probability can be an improvement on maps of rates when inferential certainty is more important than difference in magnitude. A similar method is employed by Alberta Health and Wellness (Alberta Health and Wellness 2003). Alberta Health often reports data at numbers sufficiently high to assume underlying disease frequencies follow the normal distribution. By using the normal standard deviations as categorical cut-offs, the method simplifies presentation by eliminating the subtle (and generally not meaningful) variations that may be visible from location to location. This is particularly important when routine data are being reported to the public.

One problem with probability maps is that large population areas obtain greater significance than small population areas (which may have higher rates of disease) by virtue of statistical stability. In this way, probability maps tend to over-emphasize the relative importance of large population areas (Kennedy-Kalafatis 1995) and make local comparisons difficult. This has been referred to as the 'large numbers problem'.

2.3.1.3 Topological smoothing

Topological smoothing refers to methods in which geographic locations borrow statistical strength from each other based on some measure of geographic closeness—such as proximity or adjacency. These approaches are conceptually similar to smoothing approaches applied to time series data—like moving average filters. Neighbouring areas are combined to build statistical stability without losing important local variations. Methods that use uniform filters (such as circles or squares) to define topological relationships are typically referred to as 'spatial filter' approaches. These filtering approaches are based on the idea that proximal areas are the most sensible from which to borrow statistical strength. We use the phrase 'topological smoothing' to emphasize that definitions of closeness can be more flexible than a fixed geometry, and may occasionally need to depart from smoothing approaches defined by strict geographic proximity.

Rushton and Lolonis (1996) use a method that smoothes high-resolution spatial data in order to better represent estimates of incidence. The original study

data were street addresses, but the method has been used with polygon centroids (Nelson 1999). The first step is to overlay a uniform series of points over the original data. Each one of these points is the centre of a circle with a radius smaller than the distance between each point in the lattice. This ensures that circles overlap, but that each circle does not include more than one grid point. The degree of overlap between circles is chosen based on the degree of smoothing required—larger circle radii produce greater smoothing, smaller radii produce less. Case and population data are aggregated to each node in the lattice based on whether or not they fall in a circle. Mungiole, Pickle and Simonson (1999) use a distance and directional threshold to select local neighbouring areas to be used in a weighted smoothing approach. Their method smoothes localised noise while preserving other spatial characteristics of visual importance—such as edges and clusters. When area data are of sufficiently high resolution, median polish smoothing (Cressie 1993 pp.184-193) is able to extract simple patterns or 'signals'. This method is performed by iteratively subtracting row and column medians from data aggregated into grid cells overlying the original data.

Generally speaking, topological smoothing methods are more aggregation invariant (and less affected by the modifiable areal unit problem) than methods that rely on single discrete aggregate units for analysis. This is particularly true when data are available at high resolution. However, results from a topological smoothing exercise can be strongly influenced by prior analytical decisions. For instance, the median polish approach can only capture row and column effects (such as patterns in the North and South directions), and is highly sensitive to the alignment of the overlying grid; different alignment can result in different signals of trend. Topological smoothing methods can also be sensitive to large variations in population density. In study areas where population density varies considerably, fixed window smoothing approaches will either over-smooth densely populated areas or under-smooth sparsely populated areas depending on window size. This can be offset by varying the filter size so that the smoothing kernel is not of fixed radius, but of fixed population size (Talbot et al., 2000). This technique could also benefit from additional constraints that ensure that

certain areas are not 'windowed' together—such as the suburbs of large urban centres and neighbouring rural regions. Otherwise, rural and semi-rural areas near cities may be forced to share data characteristics with neighbouring urban areas, ignoring important geographic differences between urban and rural physical and social environments.

2.3.1.4 Empirical Bayes methods of disease mapping

In a response to both the large and small numbers problems, some have advocated the use of empirical Bayes estimates. Most generally, these methods combine observed rates with prior information in order to smooth out variability resulting from small numbers. Clayton and Kaldor (1987) offer a global adjustment method which uses the Gamma distribution as a prior model, and an iterative method of arriving at the distribution parameters from the entirety of data available. These distributional parameters are then used to 'shrink' local rates. Their approach has the advantage of allowing covariates and spatial autocorrelation to be added directly into the estimation procedure (Waller and Gotway 2004 pp.95). Marshall (1991b) offers global or local shrinkage—local methods restrict smoothing influence to neighbouring areas rather than an entire study region. This has the appeal of preserving more local variations, though it leaves more decision making for the analyst.

Although these methods have been criticised for over-smoothing the tails of distributions of regional rates (Cressie, Stern and Wright 2000) and for underestimating uncertainty (Mollie 2001), they are reasonably effective for exploratory purposes (Leyland and Davies 2005). These methods are practical and relatively easy to apply; most can be implemented in existing statistical software environments (such as R, SAS and Stata). Fully Bayesian approaches are also available, but are considerably more difficult to implement since they require specifications of prior distributions rather than using estimates of the priors based on the data (as is the case in empirical Bayes methods). Among other complicating issues, these methods require a way to sample from a prior distribution, usually performed with specialised software that may be insufficiently generalisable for certain tasks in spatial modelling—such as the

specification of spatial covariance (Lawson 2005).

2.3.1.5 Spatial models

Many statistical modelling methods are available for understanding the geographic patterns of disease. When observations are likely to be non-normal (as is often the case in health applications), generalized linear models—such as Poisson regression—have been used (Lovett, Bentham and Flowerdew 1986; Luyao et al. 1993; Reynolds et al. 2002). Spatial models are distinguished from other statistical models by their explicit consideration of spatial autocorrelation/spatial dependence. In general, these methods involve the direct inclusion of the relationship between proximity and similarity into the modelling process. When unaccounted for, spatial autocorrelation adversely affects a number of regression modelling diagnostics (Anselin and Griffith 1988) and is well known to result in variance estimates that are too small.

Fortunately, the increased availability and understanding of spatial modelling methods has enabled their use in a wide variety of applications. Iterative re-weighted regression—where spatial autocorrelation in residual error is iteratively re-worked into the modelling procedure—offers one of the simplest approaches. This involves solving a model, identifying a theoretical semivariogram that characterises autocorrelation in the model residuals, and then iteratively re-incorporating the parameters of the semivariogram back into the model until the parameter estimates converge (Waller and Gotway 2004 pp. 337-338). Generalized linear mixed models offer a similar method of adjusting for spatial autocorrelation in normal and non-normally distributed data. The adjustment occurs through introducing the structure of a theoretical semivariogram as a random effect (Littell et al. 1996 pp. 303-330; Waller and Gotway 2004 pp. 380-409). The SAS™ implementations of mixed models (PROC MIXED and PROC GLIMMIX) provide an estimate of the semivariogram range (the furthest distance at which spatial autocorrelation exists) which gives an indication of the magnitude of spatial autocorrelation (Littell et al. 1996). The recent inclusion of generalized linear mixed models in standard software packages has enabled a number of studies to employ these methods (e.g. Mendell et al., 1996; Witte, Greenland, Kim

and Arab 2000; Kleinschmidt et al. 2001; Kleinman, Lazarus and Platt 2004; Yiannakoulias, Svenson and Schopflocher 2006).

Simultaneous autoregressive models offer a different approach to the same problem, though are generally constrained to linearly dependent variables (or a linearization through data transformation). Rather than requiring a parametric structure *a priori* (as is the case in the spatial mixed model approach), simultaneous autoregressive methods include nearby observations as covariates in the model, and define spatial covariance empirically. Griffith et al. (1998) use simultaneous autoregressive modelling to identify risk factors associated with lead poisoning in children. Antunes and Waldman (2001) use a similar method to identify predictors of tuberculosis in Brazil. Lorant et al. (2001) compare a simultaneous autoregressive to a non-spatial regression model in the prediction of premature mortality. A thorough discussion of simultaneous autoregressive methods can be found in Griffith and Layne (1999).

Spatial modelling methods are the backbone of theory building in much of quantitative human geography. Information from such models can indicate the degree of autocorrelation present, the possibility of missing covariates, and an indication of the presence of a contagious process. The success of these methods is contingent on proper model specification. Failure to include appropriate variables is almost guaranteed to bias the resulting model (in addition to affecting the power of statistical inferences), and in turn, the theory extrapolated from it. As a result, such methods require considerable intellectual participation and content expertise. This is not an unmanageable burden in a research process; well informed researchers using spatial models to understand geographic features of disease will be able to specify and interpret these models properly. However, these models are not as useful within an automated and/or routine decision support system. Furthermore, the results of such models are not easily digested by lay consumers of information—like policy makers and the public—and may be limited in their role as components of routine disease surveillance system. Further consideration of this matter is reserved for the discussion below.

2.3.2 Geographic surveillance methods: cluster detection

Non-random clusters of spatial phenomena can occur as a result of trend (where an intermediate causal mechanism with a clustered pattern influences the distribution of events) or dependence (an event at one location causes additional identical events to occur at or nearby this location). To assign one of these causal mechanisms to an observed cluster of phenomena is difficult for two reasons. First, there is no way to distinguish dependence from trend with a cross-section of observations (Bartlett 1964). Second, many apparently non-random trend effects are actually latent dependence effects. Variables that influence disease trends are very often the product of spatially dependent processes. For example, people of similar income status tend to live together and to have offspring of similar socioeconomic status. A non-infectious disease correlating with income may exhibit a clustered pattern that is influenced by income, thereby exhibiting an intermediate dependence effect, even though there is no direct causal path between two nearby disease events. To avoid these pitfalls, this discussion treats cluster detection as a problem of classification (as discussed by Diggle 2000) regardless of spatial processes influencing the distribution of observations.

There are two conceptual subdivisions within the cluster detection literature (Besag and Newell 1991). The first considers *focus*. Focused tests examine whether or not a cluster of events have occurred around specific locations (such as power plants, pulp mills, etc.). General tests make no prior specifications about sites. Focused tests are useful when it is important to monitor disease phenomena around sites of concern, and general tests are more exploratory. The second subdivision refers to *scope*. Methods with global scope test for the presence of a clusters without specifying the location or structure of the cluster pattern. Methods with local scope identify the locations of clusters of observations that meet some (usually statistical) threshold of interest. The former method is easier to apply, but less informative; the identification of an anomalous pattern of clustering at a global scope does not give an indication of structure or location, but simply tells us that it exists in some form. The latter method is much more difficult to apply since it presents an inferential challenge—test the statistical

significance of a pattern without knowing about its presence or location ahead of time. Most methods of cluster detection can be classified using these subdivisions. A focused test indicates whether or not a disease clusters around particular locales. A global general test indicates whether or not there is clustering of disease somewhere in a study region. A local general test indicates whether or not there is clustering of disease in a particular region. A local focused tests indicates that there is clustering around one or more particular known locations.

The methods and applications of geographic cluster detection have evolved considerably in the last few decades. Most of the early methods were general tests of global scope. Some methods were based on cell occupancy tests; clusters were detected when cases occurred in pre-determined temporal and/or spatial categories more often than what was expected by chance alone. Modern methods of cluster detection are generally distance-based or quadrat-based (Ripley 1977).

2.3.2.1 Distance-based approaches

Most distance-based methods are formed from the idea that short distances between events must indicate an interesting spatial process. As such, most distance-based methods use individual event or person data, rather than spatial or temporally aggregate data. However, there are some exceptions. Though conceptually, Knox (1964) uses inter-event distances as part of his test statistic, his method requires that one aggregates data into near and far groups, and therefore, precise inter-event distance data is not really required beyond knowing whether observations are 'near' or 'far' from each other. Whittemore et al. (1987) employ a distance-based method that accommodates data aggregated to areal centroids and takes into account inhomogenous population density. Tango (1995) presents a test statistic based on the pairwise distances between observed and expected numbers of regional cases. This method incorporates a scaling factor on distance so that regions very distant from each other have a reduced role on the evaluation procedure. An adaptation of this technique (Tango 2000) makes it robust to the choice of the scale parameter. This method has been found to be particularly well suited to find global patterns of clustering (Kulldorff, Tango and

Park 2003).

Most distance based approaches are global tests for the presence of clustering rather than local tests that identify the location of clusters. One possible exception comes from Anselin (1995) in the form of local indicators of spatial association (LISAs). The method is a general framework for finding local geographic groups of systematically similar (or dissimilar) values. The method reports structure and patterns of clustering in phenomena, and not necessarily anomalies of concern. Similar methods were also explored by Getis and Ord (1996) and Munasinghe and Morris (1996). A number of global indicators of clustering can be decomposed into LISAs—such as traditional measures of spatial autocorrelation as well as other distance-based methods of detecting the presence of clustering. LISA methods suffer from some inferential challenges, in particular, the multiple testing of significance of local clusters when the observations under study are stochastic (like disease events) (Waller and Gotway 2004 pp. 238).

2.3.2.2 Quadrat approaches

Although some new methods of cluster detection still employ distance measures (e.g. Forsberg et al. 2005) many of the most recently developed approaches are based on searching study regions with a moving 'window'. In general, these methods search and evaluate the significance of clusters by observing groups of geographically compact observations and then comparing these groups to their complement or the study region as a whole. Since the focus is on local groups of observations these techniques are well suited for identifying local disease clusters of the focused or unfocussed variety. In large heterogeneous study regions (such as Canadian provinces) these kinds of tests provide important information about local patterns.

One of the first implementations was by Openshaw et al. (1987). The original method was described as an automated point pattern detection system, though practical applications (e.g., Openshaw et al. 1988) involved a form of aggregate health data. Cases are aggregated into a uniform grid tessellation with population estimates based on census data. The intersection points of the grid cells are the

centres of a finite number of circular windows ranging from a small to large radius. Case and population data are grouped into each of these circles at each of the grid locations. Monte Carlo simulations are then used to test the significance of each circle independently. This is done as follows. For each simulation, each individual in the study area is randomly assigned case/non-case status, with risk determined by the overall rate. Then the same set of circular filters are applied to each simulated data set. At each location, a circular cluster is identified as 'significant' if the real number of cases exceeds the simulated number of cases over a pre-selected threshold (e.g., 500 times). The original geographical analysis machine (or GAM) method was computationally costly for large data sets, and was frequently criticised for its tendency to falsely reject null hypotheses of constant risk (e.g., Besag and Newell 1991).

The Besag and Newell (1991) approach to detecting clusters was originally developed to detect focused clusters of very rare disease but can be applied more generally, and for unfocused tests. For each centroid in a study region, a table of ordered nearest neighbours (from closest to farthest) is calculated. A pre-determined case threshold (K) is set, and for each centroid, neighbouring centroids are accumulated (in order of nearest to farthest) up to the case threshold size. The number of centroids aggregated forms the basis of the test of significance, which is the probability of having attained the observed number of cases by accumulating fewer centroids. The success of the approach is dependent on the value of K. Le, Petkau and Rosychuk (1996) developed a modification of this method to deal with the difficulty of choosing the appropriate value of K. A testing algorithm is undertaken to determine significance of a centroid using a small set of increasing Ks, depending on the underlying population of the centroid and its neighbours. If significant at a particular K, no further tests are done at higher Ks. If insignificant, larger Ks are used in sequence until the centroid is significant or the set is exhausted. Once the testing for a centroid is complete, the algorithm continues to the next centroid.

The challenge of formulating sensible experiments to identify and test the significance of local clusters without performing a large number of statistical

inferences (which necessarily increases the risk of false rejecting true null hypotheses) encouraged some to develop cluster detection methods that find the location of a single most-likely cluster. Most of these methods share a similar null hypothesis: the rate of illness among people living in the most-likely cluster is no different from the rate of illness among those living in the complement (non-cluster areas) of the most-likely cluster. The inferential simplicity of these methods make them particularly well-suited to automated disease surveillance, and will be given a more detailed review here.

The cluster evaluation permutation procedure (Turnbull et al. 1990) determines the location and significance of a most-likely cluster. This method sequentially adds the nearest neighbours *j* to a cluster *i* (so called since it starts at centroid *i*) until a critical population threshold (R) is reached. The total number of cases in cluster *i* is recorded, and then the same procedure is conducted at a new centroid. When all clusters *i* have obtained a case count, the cluster *i* with the largest case count is selected as the most- likely cluster. A Monte Carlo simulation procedure is then used to test the likelihood of this most-likely cluster occurring by chance alone. Data are simulated similar to Openshaw et al. (1987) however, for each simulation, the case count of the most-likely cluster is retained. If the case count of the real most-likely cluster exceeds the case counts of a large number of the simulated most-likely clusters, the real cluster is considered significant. Thus unlike Openshaw et al. (1987) a single test of significance (of the most-likely cluster) is performed, rather than a large number of local tests. Generally speaking, the larger the size of R, the greater the sharing of observations between seed locations, and the more stable the underlying rates. Selecting an excessively large value of R will shrink variation unnecessarily; a value of R that is too small will leave the patterns obscured by stochasticity in the data.

The spatial scan (Kulldorff 1997) method of cluster detection is widely used, and has been implemented in a variety of disciplines (see Kulldorff (2005) for a review of applications). The method was developed for use on atomic or aggregate data; our focus is on the latter. The approach is capable of detecting

spatial, temporal and/or spatial-temporal clusters without the threshold specifications of some other methods. Similar to Turnbull et al. (1990) the spatial scan uses a circular window of varying sizes to aggregate neighbouring centroids to a seed centroid. However, rather than deriving an evaluation statistic based on windows of a particular population size, the spatial scan evaluates each window with a likelihood ratio test. Once the window increments to a size that includes all centroids the search starts over at a new centroid, and continues until all centroids have been searched. Since every centroid is the seed of a varying window search, the spatial scan statistic is an exhaustive search of clusters based on a nearest neighbour aggregation process (though typically, the algorithm is prevented for searching for new clusters once the sequence reaches 50% of the population). A most-likely cluster is chosen based on the highest likelihood ratio of all circular windows at all centroids. The likelihood ratio associated with the most-likely cluster is used to test the significance of the most-likely cluster as a true local anomaly of disease. Significance is tested using a Monte Carlo simulation where case locations are randomized (through random labelling) and the same likelihood ratio statistics are determined for each simulated disease scenario (Kulldorff and Nagarwalla 1995). The likelihood ratios associated with most-likely simulated clusters are saved, and then compared to the likelihood ratio associated with the true most-likely cluster. If the likelihood ratio of the true most-likely cluster is larger than a large proportion of the likelihood ratios associated with the simulated most-likely clusters, then the detected cluster is considered significant.

Recently, new methods have expanded the search process to allow for clusters of non-circular and unusual shapes. Tango and Takahashi (2005) developed a more complete scan technique able to find clusters of any shape constrained to relatively small size. Other approaches use heuristic strategies to solve larger problems without the burden of full enumeration. Conley, Gahegan and Macgill (2005) use a genetic approach to find clusters of circular or elliptical shape. Potential clusters that exhibit a desirable trait (high fitness to the data) are preferentially combined (or 'mated'), generating offspring clusters which (may)

mutate to further improve the search process. Duczmal and Assunçao (2004) use a circular most-likely cluster as a starting point for a more complete search of clusters without assuming a pre-defined shape. They use a simulated annealing strategy to ensure that different and (in the long-term) suitable zones will be added and/or removed from the most-likely cluster. The genetic and simulated annealing approaches balance the need for high quality solutions with the need for efficiency, while still allowing for interesting shapes to be detected.

The nature of simulated annealing, genetic searches and other heuristic strategies can make inferences (and interpretation) complicated, however. For heuristic approaches to return good solutions to a cluster detection problem, they often require proper specification—or tuning—to the problem at hand. Evolutionary approaches require decisions about the probability of mutation; tabu search methods require decisions about tabu list sizes and the duration of tabu list membership; simulated annealing strategies require decisions about the 'cooling off' rate and various other parameters. Other times methods require additional rules of thumb or 'meta-heuristics' to ensure good solutions. For example, the Duczmal and Assunçao (2004) approach includes a procedure to ensure that the search process is able to find 'interesting' clusters (and in particular, clusters that are not compact) by preferential selecting areas in the neighbourhood of recently added zones. The decision to tune and/or employ meta-heuristics must be established ahead of time—through cross-validation methods (tuning the problem on a sub-set of data first), underlying theory or guesswork. When this tuning process is done with knowledge of the data (or in conjunction with it) then the objectivity of the inferences may be called into question; that is, the method is being customized to find something that an analyst expects to see. Although this is not an uncommon or necessarily terminal problem, it does suggest that such methods must be employed with care, and perhaps a large degree of supervision.

Others have concentrated on adapting the evaluation procedures underlying the spatial scan. One of the challenges of the maximum likelihood ratio approach is that estimates of relative risk within clusters are only an informal approximation of the highest risk area (since the maximization is on the likelihood

ratio, not magnitude of risk). Gangnon and Clayton (2000) adapt a Bayesian cluster detection method able to obtain estimates of risk within detected clusters. The authors suggest that their method is fairly robust, and the difficulty of selecting an appropriate prior model can be mitigated by performing a sensitivity analysis. Their experiments include tests for compact and linear clusters, but not any more non-circular shapes. The effort of choosing a prior model does not impose an undue challenge for epidemiological purposes, but such approaches may be too complex for many routine surveillance applications. This is particularly true when the prior model includes constraints on cluster shape. Gangnon and Clayton (2001) also offer several weighted average likelihood statistics as an alternative to the maximum likelihood ratio. When the weights are properly specified, these methods can offset the tendency of the traditional scan statistic to preferentially find clusters in areas of higher resolution (Gangnon and Clayton 2001). More recent developments in outbreak surveillance have incorporated Bayesian models in anomalous pattern detection (e.g. Lawson, Clark and Rodeiro 2004; Lawson and Kleinman 2005).

2.4 Discussion

Table 2.1 offers a simple taxonomy relating the four dimensions of geographic surveillance to the methodological approaches described above. The final column includes selected references to applications of these methods. In cases where literature was originally applied or discussed in terms of public health surveillance, the reference is in bold font. Each cell contains a brief summary of some strengths and weaknesses of each methodological category that pertains to the tasks of geographic surveillance.

Much recent emphasis in geographic surveillance methodology has been on developing methods suitable for rapid response to outbreaks of disease and bioterrorism. These contributions have resulted in an increased ability to implement systems to detect sudden changes in spatial disease patterns. Many of these tools are also available for the monitoring and study of new risks. These methods are shared with more explanatory research in the social and health sciences, and not surprisingly, have been applied in numerous settings.

Unfortunately, discussion of information dissemination has been missing from many of the most recent developments of geographic health surveillance methodology. Disease mapping methods will continue to have many important applications in health research. Methodological advancements in these fields will always be important to the study of health and disease—including public health surveillance. However, the chief disadvantage of most disease mapping approaches is that they are not easy to convert from abstract research findings into information usable in the public health setting. A local estimate of disease is composed of a measure of magnitude (the rate estimate or population count) and, in some form, a measure of uncertainty (a p-value, shrinkage factor, or sampling distribution, for example). Spatial models add to this complexity by including spatial dependence and covariates into the process. Collectively, local rate estimates can never be interpreted outside a detailed discussion of methodology. Estimates are subject to a large number of caveats that are difficult for non-experts to understand (like the correct specification of the model, proper distributional assumptions, etc.).

Furthermore, disease maps require analytical processing in order to be represented on a map. Formal categorization offered by a common geographic information system can be used to classify the range of local estimates into discrete colour categories. The choice and method of classification can have a dramatic effect on the appearance of a map. Figure 2.1 shows how different categorization schemes can influence the appearance of mapped rates. For epidemiological purposes, these variations do not obscure the general observations—for example, that the pattern of Parkinson's disease is low in Northern and South-western areas and higher in the populated rural areas of the South-east. However, several areas shift one and some even two shade positions simply based on the type of categorization scheme used. For residents in these areas, such changes could represent a considerable difference in perception, and one is obliged to ask whether methods so influenced by the choice of categorization are suitable for reporting information to the public. These observations about disease maps are not new to the cartography, medical

geography or spatial statistical community, but the implications are particularly important in the public health setting when information is being communicated for public, media or political consumption. The appearance of high rates may spawn a public outcry, media attention and extra-governmental investigation.

The complexity of spatial modelling also inhibits its use in some surveillance applications. For one, direct human supervision is required in the modelling process to ensure that the procedures operate properly. For some methods, one must decide on the structure of spatial dependence—which requires prior analytical work. Decisions must be made as to model specification—not only in the choice of variables, but in their mathematical form (for example, as interaction and/or polynomial terms). Spatial models have great power to build theory and understanding from empirical data and are central to explanatory research. However, their complexity—in implementation and in explanation—may inhibit their use in many true geographic surveillance applications.

Most cluster detection methods offer considerably less information than disease maps. Cluster detection methods do not typically portray trends, and many do not even offer accurate estimates of local risk. Like data mining applications in general, cluster detection methods answer simple questions subject to various input constraints. The inferential simplicity of these methods is gained by restricting their inferential breadth. Global unfocussed cluster tests—like Knox's (1964) space-time method—have a relatively simple null hypotheses: distances between cases in space are independent of distance between cases in time. The information provided is very restricted—it does not indicate where the cluster is, and it assumes that the distance categories are properly specified. Nonetheless, given the relatively simple inputs (the choice of distances metric and categories) the presence or absence of a cluster can be determined.

Other cluster detection methods offer more information, but with a trade-off in inferential simplicity. Openshaw et al. (1987) do not test a single global hypothesis, but explore patterns of clusters for which local statistical testing was performed. Similarly, LISA methods provide information about the geographic structure of autocorrelation without a simple inferential objective. On the other

hand, several cluster detection methods do offer discrete, repeatable and relatively simple answers to questions about variation of health. Methods that follow the example of Turnbull et al. (1990)—by focussing on a most-likely cluster—combine the inferential simplicity of historical cluster detection methods with details about the geographic location of clusters.

The benefits of such inferential simplicity arise in routine alarm detection and in information dissemination. In the former case, methods that can operate without constant human supervision are easier (and cheaper) to implement in routine surveillance settings. They are able to mine the increasingly large health data warehouses at very low cost, and return information that would otherwise have never been discovered. Decision makers can evaluate the success and validity of these outputs much easier than they can engage in the searches first-hand. In the latter case, inferential simplicity provides discrete information that can be communicated to the public and to policy makers—a significant cluster was found or not found (subject to the parameters of the search method).

Some general methodological challenges that present themselves in explanatory research applications are also important to geographic surveillance and are worthy of brief mention.

2.4.1 Modifiable Areal Unit Problem (MAUP)

Although much of the literature on MAUP indicates that means and rates are not systematically biased by different aggregation arrangements (when properly weighted), means and rates are still affected by the arrangement of boundaries and scale of observation. A local estimate of disease in a high incidence neighbourhood can be obscured by combining it with neighbouring low-incidence areas. This may be somewhat rectifiable when the units of aggregation are well chosen. Some work in this area has been done in this area (Huel, Petiot and Lazar 1986; Morris and Munasinghe 1993; Haining, Wise and Blake 1994) but these approaches do not concern themselves with reporting issues relevant to surveillance. Future research on health region districting may be an important component of developing better routine reporting systems.

2.4.2 Diagnostic inconsistency over time and space

Some recent research suggests that factors unrelated to disease may explain some apparent disease patterns, particularly from administrative data. Historical changes in diagnostic methods and technology sometimes account for apparent temporal changes in disease patterns, amounting to 'diagnostic inconsistency' over time (Burnand and Feinstein, 1992). Similar observations have been made about geographic variation in diagnostic methods (Forand et al. 2002; Yiannakoulias et al. 2003; Yiannakoulias et al. 2004). The problem of diagnostic inconsistency is an important reminder of the need for careful clinical research following a surveillance exercise, particularly when an illness is likely to garner public or policy concern. Diagnoses used in surveillance activities are suited to detecting potential changes in disease, but still require considerable groundwork to validate the changes in precise terms before definitive conclusions can be made.

2.4.3. Clinical versus statistical significance

When data are numerous, and observations have marginal sampling error, traditional statistical techniques often identify patterns that are of statistical but not clinical significance. These patterns are genuinely non-random, but effect sizes are so small as to be beyond any level of explanatory, predictive or decision making importance. Under these circumstances, there is a need to ensure that the observations of certainty and magnitude are not confused, or perhaps that effect size is incorporated directly into the surveillance process. Some relatively simple modifications of traditional cluster detection techniques would accomplish this—for example, requiring that all significant clusters exceed a certain minimum magnitude threshold in order to be reported. However, such an adjustment (or more accurately, constraint) complicates the detection procedure, and traditional methods would need adaptation to ensure their effectiveness.

2.4 Conclusion

Population-wide administrative health systems provide some of the most important data required for comparative analysis of disease and injury frequency between communities. Although these systems continue to be used for explanatory research in epidemiology and the social sciences, this captures only

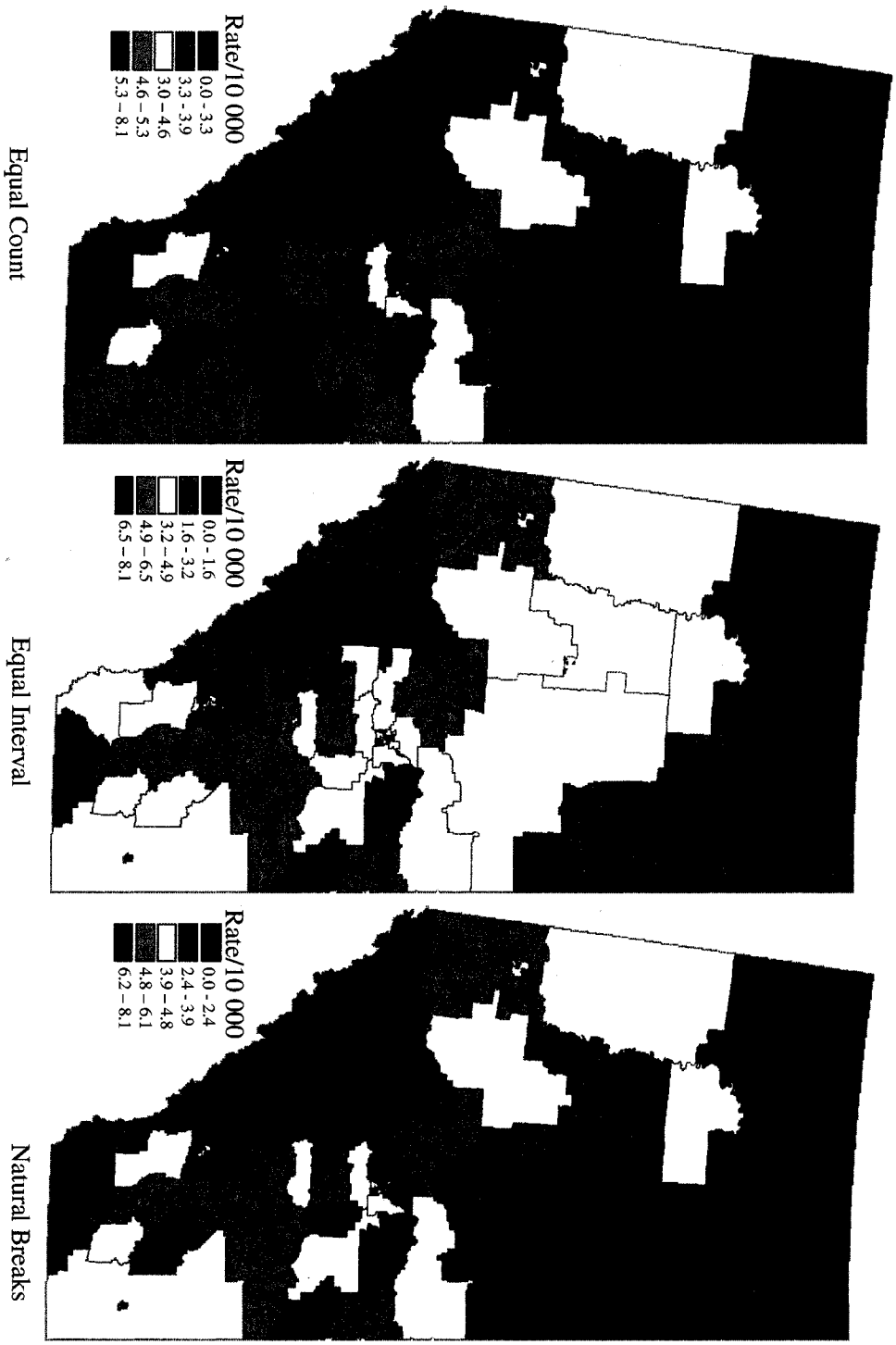
part of their potential. As quantities of data increase, policy makers are challenged to find the resources to monitor health and disease within these systems. Automated surveillance methods may be able to fill in some of the analytical gaps that more supervised analytical procedures will inevitably leave. However, these methods must be mindful of the limitations of various analytical techniques, and must be designed to suit the needs of the users.

Choice of methodology influences perceptions of spatial information. Cluster detection methods, which are generally less informative about disease patterns than disease maps, offer relatively simple information to users. They indicate where disease is anomalously high or low. This kind of information is important for policy makers, the public and the research community. Policy makers can use this information to identify areas where intervention may be important. The public can use this information to assess whether or not their community is at particularly high risk, or has a particularly high burden of disease. Researchers can use this information to build new theories, develop new experiments, and offer explanatory information back to the policy makers and the public. It is important that methods of spatial analysis that are developed in the future are also evaluated in terms of application. Methods that aspire to be used in the geographic surveillance setting must be assessed with regard to the needs and applications of geographic surveillance.

Table 2.1 A schematic table of geographic surveillance methods and application

	Identifying new risks	Information for public health decision support	Information for communities	Alarm system	References
Rate mapping	Exploratory information Can control for known risk factors	Results are simple to understand Easy to derive and present Small numbers problem can exaggerate variation		Greatest potential when effect size is large Do not typically provide discrete information indicating whether or not an alarm has been set off Multiple comparisons problem	
Probability maps	Exploratory information	Results are relatively simple to understand Large numbers problem can over-emphasize large population areas			Choynowski (1959)
Topological smoothing	Exploratory information Can control for known risk factors	General patterns with information about trend Simple output	May provide unrealistic or inaccurate local information Simple output		**Rushton and Lolonis (1996)**; Mungiole, Pickle and Simonson (1999); Ali, Emch and Donnay (2002); Trooskin et al. (2005)
Empirical Bayes smoothing	Exploratory information Can control for known risk factors	General patterns with information about trend Output is not easy to describe	May provide unrealistic or inaccurate local information Output is not easy to describe		Clayton and Kaldor (1987); Marshall (1991); Langford (1994); Kennedy-Kalafatis (1995); Leyland and Davies (2005)
Spatial models	Informative of ecological-level associations Can control for known risk factors Directly model spatial autocorrelation	Informative of ecological-level associations Can identify areas of concern while controlling for known risks or risks that cannot be specifically addressed Method and output are not easy to describe		Most require input, supervision and decision making from a knowledgeable user May be too complex for ongoing automated analysis	Griffith, Doyle, Wheeler and Johnson (1998); Kleinschmidt et al. (2001); **Kleinman, Lazarus and Platt (2004); Lawson (2005)**
Cluster detection (Distance)	Exploratory information Some methods can control for known risk factors Global or local information	Discrete results that are relatively simple to understand		Operationally efficient Discrete output available Multiple comparisons problem when used for local surveillance	Tango (1995); Anselin (1995); **Forsberg et al. (2005)**
Cluster detection (Quadrat)	Exploratory information Some methods can control for known risk factors Typically local information	Discrete results that are relatively simple to understand When a noteworthy cluster is found, its location is identifiable		Operationally efficient Discrete output available Local alarm monitoring	Openshaw et al. (1987); Kuldorff and Nagarawalla (1995); **Kuldorff (2001); Motashari et al. (2003); Neill and Moore (2005)**

Figure 2.1 Crude rates of prevalent Parkinson's disease using three common classification systems.



2.6 References

Aberley D, Sieber R (2005) Public participation GIS http://www.iapad.org/ppgis_principles.htm. Extracted October 29, 2005.

Alberta Health and Wellness (2003). A Rate Mapping Template for Alberta Regional Health Authorities. Geographic Methodology Series No. 3. Alberta Health and Wellness: Edmonton.

Anselin (1995). Local indicators of spatial association—LISA. *Geographical Analysis* **27**:93-115.

Anselin L, Griffith D (1988) Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association* **65**:11-34.

Antunes J L, Waldman E A (2001) The impact of AIDS, immigration and housing overcrowding on tuberculosis deaths in Sao Paulo, Brazil, 1994-1998. *Social Science and Medicine* **52**:1071-80.

Bartlett M S (1964) The spectral analysis of two-dimensional point processes. *Biometrika* **51**:299-311.

Bayer R, Fairchild A L (2004) The genesis of public health ethics. *Bioethics* **18**:473-492.

Beauchamp D E. (1980) Public health and individual liberty. Annual Review of Public Health **1**:121-136.

Beckett C G, Kosasih H, Ma'roef C, Listiyaningsih E, Elyazar R F, Wuryadi S, Yuwono D, McArdle J L, Corwin A L, Porter K R (2004). Influenza surveillance in Indonesia:1999-2003. *Clinical Infectious Diseases* **39**:443-339.

Bender A P, Willians A N, Johnson R A, Jagger H G (1990). Appropriate public health responses to clusters: the art of being responsibly responsive. *American Journal of Epidemiology* **132**(Suppl 1):48-52.

Besag, J, Newell J (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society Series A **154***: 143-155.

Bithell J F (2000) A classification of disease mapping methods. *Statistics in Medicine* **19**:2203-2215.

Bose I, Mahapatra R K (2001) Business data mining—a machine learning perspective. *Information & Management* **39**:211-225.

Brookmeyer R, Stroup D F (eds.) (2003) Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance. Oxford University Press: New York.

Burnand B., Feinstein A (1992) The role of diagnostic consistency in changing rates of occurrence for coronary heart disease. *Journal of Clinical Epidemiology* **45**, 929-940

Carriere K C, Roos L L (1994) Comparing standardized rates of events. *American Journal of Epidemiology* **140**:472-482.

Center for Disease Control (1990) Guidelines for investigating clusters of health events. Morbidity and Mortality Weekly Reports **39**(RR-11):1-16.

Choynowski M (1959) Maps based on probabilities. *Journal of the American Statistical Association.* **54**:385-388.

Clayton D G, Kaldor J (1987) Empirical Bayes. estimates of age-standardized relative risks for use in. disease mapping, *Biometrics* **43**:671–681.

Conley J, Gahegan M, Macgill J (2005) A Genetic Approach to Detecting Clusters in Point Data Sets. *Geographical Analysis* **37**:286-314.

Cox K C, Eick S G, Willis G I, Brachman R J (1997) Visual data mining: recognizing telephone calling fraud. *Data Mining and Knowledge* Discovery **1**:225-231.

Cressie N (1993). *Statistics for Spatial Data.* John Wiley & Sons: New York.

Cressie N, Stern H S, Wright D R (2000). Mapping rates associated with polygons. *Journal of Geographical Systems* **2**:61-69

Cuzick J, Edwards R (1990) Spatial clustering for inhomogenous populations. *Journal of the Royal Statistical Society B* **52**:73-104.

Declich S., and Carter A.O. (1994) Public health surveillance: historical origins, methods and evaluation. Bulletin of the World Health Organization **72**:285-304.

Diggle P J (1990). A point process modelling approach to raised incidence of a rare phenomena in the vicinity of a prespecified point. *Journal of the Royal Statistical Society A* **153**:349-362.

Diggle P J, Chetwynd A G (1991) Second-order analysis of spatial clustering for inhomogenous populations. *Biometrics* **47**:1155-1163.

Diggle P J (2000). Disease mapping and its relationship to cluster detection. In *Spatial Epidemiology: Methods and Applications.* Elliot P, Wakefield J, Best N, Briggs D (eds.). Oxford University Press: Oxford.

Duczmal L, Assunçao R (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* **45**:269-286.

Elliot P, Wakefield J, Best N, Briggs D (eds.) (2001) *Spatial Epidemiology: Methods and Applications*. Oxford University Press: Oxford.

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996). From data mining to knowledge discovery in databases. *AI Magazine* **17**:37-54.

Fairchild A L, Bayer R. (2004) Ethics and the conduct of public health surveillance. *Science* **30**:631-632.

Forand S P,Talbot T O, Druschel C, Cross P K (2002) Data quality and the spatial analysis of disease rates: congenital malformations in New York State. *Health & Place* **8**:191-199.

Forsberg L, Bonetti M, Jeffrey C, Ozonoff A, Pagano M (2005) Distance-based methods for spatial and spatio-temporal surveillance. *Spatial and Syndromic Surveillance for Public Health* Lawson A B, Kleinman K (eds.) Wiley: West Sussex.

Freudenburg WR, Rursch J A (1994) The risks of "Putting the numbers in context": a cautionary tale. *Risk Analysis* **14**:949-958.

Gangnon R E, Clayton M K (2000) Bayesian detection and modelling of spatial disease clustering. *Biometrics* **56**:922-935.

Gangnon R E, Clayton M K (2001) A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine* **20**:2977-2987.

Getis A, Ord J K (1996) Local spatial statistics: an overview. *Spatial Analysis: Modelling in a GIS Environment.* Longley P, Batty M (eds.) Geoinformational International: Cambridge.

Griffith D A, Doyle P G, Wheeler D C, Johnson D L (1998) A tale of two swaths: urban childhood blood-lead levels across Syracuse, New York. *Annals of the Association of American Geographers* **88**:640-665.

Griffith D A, Layne L J. (1999) *A Casebook for Spatial Statistical Data Analysis.* Oxford : New York

Guthrie K K, Dutton W H (1992) The politics of citizen access technology: the development of public information utilities in four cities. *Policy Studies Journal* **20**:574-597.

Haining R, Wise S, Blake M (1994) Constructing regions for small area analysis: material deprivation and colorectal cancer *Journal of Public Health Medicine* **16**:429-438.

Hawks D (1997) The new public health: nanny in a new hat. *Addiction* **92**:1175-1177.

Hennekens C H, Buring J E (1987). *Epidemiology in Medicine.* Little, Brown: Boston.

Huel G, Petiot J F, Lazar P (1986) Algorithm for the grouping of contiguous geographical zones" *Statistics in Medicine* **5**:171-181.

Jerrett M, Burnett R T, Ma R J, Pope C A, Krewski D, Newbold K B, Thurston G, Shi Y L, Finkelstein N, Calle E E, Thun M J (2005) Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology* **16**:727-736.

Kazmi J H, Pandit K (2001) Disease and dislocation: the impact of refugee movements on the geography of malaria in NWFP, Pakistan. *Social Science and Medicine* **52**:1043-1055.

Kelsall J E, Diggle P J (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics* **47**:569-573.

Kennedy-Kalafatis S (1995). Reliability-adjusted disease maps. *Social Science and Medicine* **41**:1273-1287.

Kleinman K, Lazarus R, Platt R (2004). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology* **159**:217-224.

Kleinschmidt I, Sharp B, Clarke G, Curtis B, Fraser C. (2001) Use of generalised linear mixed models in the spatial analysis of small-area malaria incidence rates in KwaZulu Natal, South Africa. *American Journal of Epidemiology* **153**:1213–21.

Knox E G (1964). The detection of space-time interactions. *Applied Statistics* **13**:25-29.

Kulldorff M, Nagarwalla N. (1995) Spatial Disease Clusters: Detection and Inference. *Statistics in Medicine* **14**:799-810.

Kulldorff M (1997) A spatial scan statistic. *Communications in Statistics—Theory and methods* **26**:1481-96.

Kulldorff M, Tango T, Park P J (2003) Power comparisons for disease clustering tests, Computational Statistics & Data Analysis **42**:665-684.

Kulldorff M (2005) Scan statistics for geographical disease surveillance: an overview. *Spatial and Syndromic Surveillance for Public Health.* Lawson A B, Kleinman K (eds.) Wiley: West Sussex.

Kulldorff M (2005) Prospective time periodic geographical surveillance using a scan statistic. *Journal of the Royal Statistical Society* **164**:61-72.

Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine.* **2**:216-224

Langford J H (1994) Using empirical Bayes estimates in the geographical analysis of disease risk. *Area* **26**:142-149.

Lawson A B, Biggeri A, Böhning D, Lesaffre E, Viel J F, Bertollini R (eds.) (1999) *Disease Mapping and Risk Assessment for Public Health.* Wiley: West Sussex.

Lawson A B, Biggeri A B, Boehning D, Lesaffre E, Viel J F, Clark A, Schlattmann P, Divino F (2000) Disease mapping models: an empirical evaluation. *Statistics in Medicine* **19**:2217-2241.

Lawson A B (2001a) Tutorial in biostatistics: disease map reconstruction. *Statistics in Medicine* **20**:2183-2204.

Lawson A B (2001b) *Statistical Methods in Spatial Epidemiology.* Wiley: West Sussex

Lawson A B, Clark A, Vidal-Rodeiro C L (2004). Developments in general and Syndromic surveillance for small area health data. *Journal of Applied Statistics* **31**:951-966.

Lawson A B, Kleinman K (eds.) (2005) *Spatial and Syndromic Surveillance for Public Health.* Wiley: West Sussex.

Lawson A B (2005) Spatial and spatio-temporal disease analysis. *Spatial and Syndromic Surveillance for Public Health.* Lawson A B, Kleinman K (eds.) Wiley: West Sussex.

Lazarus R, Kleinman K, Dashevsky I, Adams C, Kludt P, DeMaria A, Platt R (2002). Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerging Infectious Diseases* **8**:753-760.

Le N D, Petkau A J, Rosychuk R (1996) Surveillance of clustering near point sources. *Statistics in Medicine* **15**:727-40.

Leyland A H, Davies C A (2005) Empirical Bayes methods for disease mapping. *Statistical Methods in Medical Research* **14**:17-34

Liao S C, Lee I N (2002) Appropriate medical data categorization for data mining classification techniques. *Medical Informatics* **27**:59-67.

Littell R C, Milliken G A, Stroup W W, Wolfinger R D (1996). *SAS © System for Mixed Models.* SAS © Institute: Cary.

Lorant V, Thomas I, Deliege D, Tonglet R (2001) Deprivation and mortality: the implications of spatial autocorrelation for health resources allocation. *Social Science and Medicine* **53**:1711-1719.

Losiewicz P, Oard D W, Kostoff R N (2000) Textual data mining to support science and technology management. *Journal of Intelligent Information Systems* **15**:99-119

Lovett A A, Bentham C G, Flowerdew R (1986) Analysing geographic variations in mortality using Poisson regression: the example of ischaemic heart disease in England and Wales 1969-1973. *Social Science and Medicine* **23**:935-943.

Luscombe N M, Greenbaum D, Gerstein M (2001) What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine* **40**: 346-358.

Luyao G L, McLerran D, Wasson J, Wennberd J E (1993). An assessment of radical prostatectomy-time trends, geographic variation and outcomes. *JAMA* **269**:2633-2636.

Marshall R J (1991a) A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society* **154**:421-441.

Marshall, R J (1991b). Mapping Disease and Mortality Rates using Empirical Bayes Estimators. *Applied Statistics* **40**: 283-294.

Mawby T V, Lovett A A (1998) The public health risks of Lyme disease in Breckland, UK: An investigation of environmental and social factors. *Social Science & Medicine* **46**:719-727

Mendell M J, Fisk W J, Deddens J A, Seavey W G, Smith A H, Smith D F, Hodgson A T, Daisey J M, Goldman L R (1996) Elevated symptom prevalence associated with ventilation type in office buildings. *Epidemiology* **7**: 583-589.

Middaugh J P, Hodge J G, Cartter M L (2004) Letter: the ethics of public health surveillance. *Science* **304**:681.

Mohan J, Twigg L, Barnard S, Jones K (2005) Social capital, geography and health: a small-area analysis for England. *Social Science and Medicine* **60**:1267-1283.

Mollie A (2001). Bayesian mapping of Hodgkin's disease in France. In *Spatial Epidemiology: Methods and Applications.* Elliot P, Wakefield J, Best N, Briggs D (eds.). Oxford: Oxford.

Morris R D, Munasinghe R L (1993) Aggregation of existing geographic regions to diminish spurious variability of disease rates. *Statistics in Medicine* **12**:1915-1929.

Munasinghe R L, Morris R D (1996) Localization of disease clusters using regional measures of spatial autocorrelation. *Statistics in Medicine* **15**:893-905.

Mungiole M, Pickle L W, Simonson K H (1999) Application of a weighted head-banging algorithm to mortality data maps *Statistics in Medicine*. **18**:3201-3209.

Morrow B H (1999) Identifying and mapping community vulnerability. *Disasters* **23**:1-18.

Motashari F., Kulldorff M., Hartman J.J., Miller J.R., Kulasekera V. (2003) Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases* ***9***:641-646.

Nelson R (1999). Low birthweights: spatial and socioeconomic patterns in Kamloops. *Western Geography* **8/9**:31-59

Nelson K E (2001). Surveillance. *Infectious Disease Epidemiology*. Nelson K E, Masters-Williams C, Graham N (eds.) Aspen: Gaithersberg.

Nobre F F, Braga A L, Pinheiro R S, dos Santos Lopes J A (1997). GISEpi: a simple geographical information system to support public health surveillance and epidemiological investigations. *Computer Methods and Programs in Biomedicine* **53**:33-45.

Obenshain M K (2004) Application of data mining techniques to healthcare data. *Infection Control and Hospital Epidemiology* **25**:690-695.

Openshaw S, Charlton M, Wymer C, Craft A (1987) A mark I geographical analysis machine for the automated analysis of point datasets. *International Journal of Geographical Information Systems* **1**:335-358.

Openshaw S, Craft A, Charlton M, Birch J (1988) Investigation of leukemia clusters by use of a geographical analysis machine. *Lancet* **331**:272-273.

Parodi S, Vercelli M, Stella A, Stagnaro E, Valerio F (2003) Lymphohaematopoietic system cancer incidence in an urban area near a coke oven plant: an ecological investigation. *Occupational and Environmental* Medicine **60**:187-194

Reed J (1994) Health promotion—a community-based perspective. *American Journal of Preventative Medicine*. **10**(Suppl.):26-29.

Reynolds P, Von Behren J, Gunier R B, Goldberg D E, Hertz A, Harnly M E (2002) Childhood cancer and agricultural pesticide use: an ecological study in California. *Environmental Health Perspectives* **110**:319-325.

Ripley B D (1977) Modelling spatial patterns. *Journal of the Royal Statistical Society Series B.* **39**:172-212.

Roberts R J, Steward J, John G (2003) Cement, cancers and clusters: an investigation of a claim of a local excess cancer risk related to a cement works. *Journal of Public Health Medicine* **25**: 351-357.

Rogerson P A (1997). Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine* **16**:2081-2093.

Rogerson P A (2005) Spatial surveillance and cumulative sum methods. *Spatial and Syndromic Surveillance for Public Health.* Lawson A B, Kleinman A B (eds.) Wiley: West Sussex.

Rushton G, Lolonis P (1996) Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine* **15**:717-726.

Rushton G (2003) Public health, GIS, and spatial analytic tools. *Annual Review of Public Health* **34**:43-56.

Smoyer-Tomic K E, Kuhn R, Hudson A (2003) Heat wave hazards: An overview of heat wave impacts in Canada. *Natural Hazards* **28**:463-485

Talbot T O, Kulldorff M, Forand S P, Haley V B (2000). Evaluation of spatial filters to create smoothed maps of health data. *Statistics in Medicine* **19**:2399-2408.

Tango T (1995) A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine* **14**:2323-2334.

Tango T (2000) A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine* **19**:191-204.

Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* **4**:1-15

Thacker S.B. and Berkelman R.L. (1988) Public health surveillance in the United States. *Epidemiologic Reviews* **10**: 164-190

Thacker S.B. and Stroup D.F. (1994) Future directions for comprehensive public health surveillance and health information systems in the United States. *American Journal of Epidemiology* **140**:383-397.

Thacker S B, Stroup D F, Rothenberg R B, Brownson R C. (1995) Public health surveillance for chronic conditions: a scientific basis for decisions. *Statistics in Medicine* **14**:629-641.

Trooskin S B, Hadler J, St. Louis T, Navarro V J (2005) Geospatial analysis of hepatitis C in Connecticut: a novel application of a public health tool. *Public Health* **119**:1042-1047.

Tsui F C, Espino J U, Dato V M, Gesteland P H, Hutman J, Wagner M M (2003) Technical description of RODS: a real-time public health surveillance system. *Journal of the American Medical Informatics Association* **10**:399-408.

Turnbull B W, Iwano E J, Burnett W S, Howe H L, Clark L C. (1990) Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* **132**:S136-S143.

Virnig B A, McBean M (2001) Administrative data for public health surveillance and planning. *Annual Review of Public Health* **22**:213-230.

Voisin B (2001) Mining astronomical data. *Lecture Notes in Computer Science* **2113**: 621-631.

Waller L A, Turnbull B W, Clark L C, Nasca P (1994). Spatial Pattern Analyses to Detect Rare Disease Clusters. In *Case Studies in Biometry*. Lange N, Ryan L, Billard L, Brillinger D, Conquest L, Greenhouse J (eds.) Wiley: New York.

Waller L A, Gotway C A (2004) *Applied Spatial Statistics for Public Health Data.* Wiley: Hoboken

Walter S D (2001) Disease mapping: a historical perspective. In *Spatial Epidemiology: Methods and Applications* Elliot P,. Wakefield J, Best N, Briggs D (eds.) Oxford University Press: Oxford.

Wang L, Ramoni M F, Mandl K D, Sebastiani P (2005) Factors affecting automated Syndromic surveillance. *Artificial Intelligence in Medicine* **34**:269-278.

Wartenberg D, Greenberg M (1992). Methodological problems in investigating disease clusters. *The Science of the Total Environment* **127**:173-185.

Whittemore A S, Friend N, Brown B W, Holly E A (1987). A test to detect clusters of disease. *Biometrika* **74**:631-635.

Witte J S, Greenland S, Kim L L, Arab L (2000) Multilevel modeling in epidemiology with GLIMMIX *Epidemiology* **11**:684-688.

Xu J., Chen H.C. (2003) Untangling criminal networks: a case study. *Intelligence and Security Informatics, Proceedings Lecture Notes in Computing Science* **2665**: 232-248.

Yiannakoulias N, Svenson LW, Hill MD, Schopflocher DP, James RC, Wielgosz AT, Noseworthy TW (2003). Regional comparison of inpatient and outpatient patterns of cerebrovascular disease diagnosis in the province of Alberta. *Chronic Diseases in Canada* **24**:9-16.

Yiannakoulias N., Svenson L W , Hill M D , Schopflocher D P, Rowe B H, James R C, Wielgosz A T, Noseworthy T W (2004) Incident cerebrovascular disease in rural and urban Alberta *Cerebrovascular Diseases* **17**:72-78.

Yiannakoulias N, Svenson L W, Schopflocher D P (2005). Diagnostic uncertainty and medical geography: what are we mapping? *The Canadian Geographer* **49**:291-300.

Yiannakoulias N, Svenson L W, Schopflocher D P (2006). Modelling geographic variations in West Nile virus. *Canadian Journal of Public Health. In Press.*

CHAPTER 3: Clusters as unfairness: geographical analysis in chronic disease prevention

3.1 Introduction

A large part of modern public health policy is concerned with the purposeful planning and implementation of disease prevention strategies.[1] Although treatment of disease often involves an interaction between individuals (usually clinicians and patients) disease prevention and health promotion are frequently a matter of public health. These forms of public health intervention are meant to benefit general welfare by providing services, information and other resources that reduce the risk and mitigate the severity of undesirable health outcomes.[2] Inevitably, intervention strategies are influenced by philosophical issues—such as weighing public good against individual rights (Kass 2001) balancing differences in preference, and resolving competing claims of resource entitlement. Debates underlying these issues are historically complex, responding to trends in human epidemiology (such as the shift from infectious to chronic disease as the primary cause of human illness) but also to changes in beliefs and social norms (Karhausen 1987).

These philosophical issues are connected to practical constraints; given the scarcity of resources, where and how should they be allocated? Such questions require information about public priorities, assessments of intervention efficacy and assessments of consequences of both action and inaction. Often these questions are resolved through the analysis of empirical data. The transformation of these data into information that policy makers can use is an analytical exercise, and can involve a variety of methods and perspectives.

Analysis informs public health intervention in at least three ways. First, it provides evidence of outcomes, and their patterns in the population. Many approaches to managing public health care are at least partly dependent on understanding the current state of affairs—how healthy is the population? How much variation in health is there? How many people live below the threshold of reasonable

[1] Our use of the term 'prevention' includes secondary prevention (shortening the duration of illness), tertiary prevention (reducing complications associated with illness) in addition to primary prevention (preventing illness from occurring) (Blaney 1987).

[2] We use the term 'outcomes' as a catch-all for disease, injury and other typically negative states of interest in health research.

opportunities for good health? Second, analysis is important for identifying the feasibility and predicting the effectiveness of different intervention options. Analytical work can help anticipate the relative costs and benefits of different strategic activities. Once this information has been compiled, it can then be compared to the goals and/or needs of the public. Finally, analysis can be important for resolving disputes once a strategy is in place. Analytical information may indicate gaps between stated policy, actual policy and the expressed interests of the public. It may also be important for the ongoing revision of intervention policies; 'empirical ethics', for example, involves an iteration between information on outcomes and policy decisions (Richardson and McKie 2005).

The present discussion aims to position a particular quantitative analytical approach—disease cluster detection—into a framework of public health intervention, and in particular, chronic disease prevention. Since disease prevention resources are subject to the constraints of economic scarcity, we propose that strategic decisions fall under questions of distributive justice—what is the ethical way to distribute a finite quantity of prevention resources? Chronic diseases present a unique challenge to public health. They have permanency, cause sufferers long-term pain, negatively affect personal relationships and employment, and many are associated with co-morbid conditions that further reduce quality of life. However, clusters of chronic disease—and in particular, geographic clusters of chronic disease—often constitute small heterogeneous groups. Unlike infectious disease prevention (for which isolated interventions can be widely beneficial) small-scale chronic disease prevention has no immediately obvious impact on the health of society at large. According to some definitions of fairness (and many definitions of efficiency) allocating disproportionate resources to clusters of the chronically ill could represent bad policy, especially if the benefits are enjoyed only by a small minority of the population. In order to justify cluster detection methods as tools for assisting chronic disease prevention, these issues require specific attention.

We introduce this discussion by briefly describing disease cluster detection methods, and then outlining an historical dispute between disease prevention policy advocates that parallels an important dispute in the philosophy of the social sciences.

We then consider the relevance of cluster detection with respect to two well-known, but historically complex, positions of distributive justice: *utilitarianism* and the *difference principle*. Although there have been some attempts to review the ability of particular analytical approaches to answer questions of health ethics (e.g., Wagstaff, Paci and Doorslaer 1991; Asada 2005), the critical study of how different quantitative strategies can inform public health intervention has been largely ignored. Most modern work in disease prevention conforms to a model of resource allocation associated with utilitarianism—the maximization of an objective over the whole of the population. As an alternative framework, the difference principle, emphasizes the importance of directing prevention efforts to those who are 'worst-off'. Following this, we discuss the idea of chronic disease cluster detection within these contexts—covering points of agreement and disagreement between method and moral context. Finally, we present some ideas and examples that illustrate how systematic cluster detection methodology might inform geographic chronic disease prevention.

3.2 Disease prevention in public health

3.2.1 Disease cluster detection

The methods and applications of geographic cluster detection have evolved considerably in the last few decades. Some of the earliest methods were based on cell occupancy tests; clusters were detected when cases occurred in pre-determined temporal and/or spatial categories more often than what was expected by chance alone (Ederer, Myers and Mantel 1964). Other early applications involved detecting the interaction of events in time and space (Knox 1964). Another approach was to compare the spatial distribution of cases with the spatial distribution of controls (Diggle 1990; Cuzick and Edwards 1990). Considerable work has been concerned with searching for clusters defined by a pre-specified spatial structure, including circles and nearest neighbour distance (Openshaw, Charlton, Wymer and Craft 1987; Turnbull et al. 1990; Besag and Newell 199; Kulldorff and Nagarawalla 1995) connected graphs (Duczmal and Assuncao 2004) and ellipses (Conley, Gahegan and Macgill 2005; Kulldorff et al. 2006). What almost all approaches to geographic cluster detection seek to do is identify spatial arrangements of disease that are

anomalously high. These arrangements or 'systems' exhibit a structure that is meaningful in various contexts of health research.

The traditional explanatory research model sees cluster detection as a tool to identify (usually environmental) causes of disease that were otherwise unknown. Concerns about nuclear power generating facilities, pulp mills and other industry motivated much of the work on cancer clusters, for example. In these cases, cluster detection is part of a research process that often requires one to 'control for' known risk factors as much as possible. Many modern cluster detection techniques easily allow researchers to adjust for known factors in the detection process, thereby ruling out important (but uninteresting) covariates that could explain how a disease varies geographically. Age, sex and socioeconomic status are common examples. Conceptually, this approach is similar to devising a spatial model and mapping residuals; notable spatial variability in what is left unexplained by our model tells us that our model is missing something that could be of causal importance. This framework also has predictive value when spatial dependence is either absent, controlled for, or modelled directly.

Cluster detection also has applications in public health surveillance. Some cluster detection methods are designed as automated surveillance tools—to search for patterns of clustering in large sets of data is daunting for human directed methodology. For surveillance applications, explanation is not as important as description, especially in the case of rapidly developing disease outbreaks. In this case, knowing where an outbreak is occurring now (regardless of where it may be next month, or next year) is of immediate administrative importance. It could indicate the location of contaminated water sources, poorly ventilated housing, ill effects from airborne pollution, or even acts of bioterrorism.

An alternative application of disease cluster methods is for strategic planning and implementation of disease prevention programs. In this context, cluster detection supports decision makers who must decide where certain public health interventions should take place. Disease clusters are not informative in all areas of disease prevention; acute diseases and injury, for example, cannot be prevented once they have occurred, so a cluster indicating the presence of incident illness is not as useful

as some alternative methods (like the spatial modelling of disease and related risk factors). Hazard clusters—locations where the density of known risk factors is high—may offer this kind of information, but only if the causal relationship between hazard and illness is strong, and well understood. However, prevention of some chronic medical conditions can benefit from knowledge generated through the study of disease clusters in at least two ways. First, some chronic conditions are risk factors for a number of other serious illnesses. Clusters of hypertension inform public health officials where the burden of hypertension is high, but also, where a variety of related illnesses—such as cerebrovascular and heart disease—may be more likely to occur. Similarly, diabetes clusters provide information on prevalence of a relatively serious chronic disease and also indicate where burdens of future disease (like blindness, neuropathy and cardiovascular illness) may also be high. Second, sometimes disease burden has a causal connection with non-disease related outcomes—such as the relationship between AIDS and homelessness (Culhane et al. 2001) and the relationship between mental illness and unemployment (Dooley, Fielding and Levi 1996). Similarly, spinal cord and brain injuries are known to increase risk of mental illness (Dryden et al. 2005) but also unemployment (Doctor et al. 2005), education and personality changes (Klonoff, Clark and Klonoff 1993).

For both of these reasons, a notable geographic cluster of disease may be a practical focal point of health promotion, disease prevention and other types of health intervention. Clusters are indicators of high disease burden, but also an indicator of where future co-morbid illness and life challenges (unemployment, poverty and homelessness) may occur, particularly when causes of disease are poorly understood. When risk factors are well understood, clusters of hazard (or hazard clusters) may provide important information as well. Directing disease prevention efforts to places where clusters of disease and/or hazards are highest focuses attention on places where the burdens of these diseases are likely to be highest. In cases where the intervention is geographical, this may offer opportunities for efficient service provision to those in greatest need.

3.2.2 Population-based disease prevention and the problem of scope

Durkheim's well known observations about changes in patterns of suicide (as a product of society, and not strictly the characteristics of individuals) have been important inspiration to the study of health and the social environment (Marmot 1998). In recent decades, a large literature has emerged which is focused on identifying the 'social causes' of disease—such as those related to neighbourhood effects (Pickett and Pearl 2001) social capital (Kawachi, Kennedy and Glass 1999) and social inequality (Marmot 2005). This research often incorporates definitions of causation that support a broader perspective of influences on health (Susser 1991; Kaufman and Poole 2000). In the field of disease prevention, many of these ideas were first discussed by Rose (Rose 1985; Rose 1992). Rose's work was a critical response to the traditional biomedical approach to disease prevention which was based on isolated interactions between clinicians and patients. This approach aims to change the behaviour of individuals that fit the clinical indicators of high-risk, and has several well-discussed advantages. It is cost effective, maximizes the ratio of prevention benefits to prevention risks, and fits within the operation of existing medical systems. However, individual-level prevention strategies do not address the social causes of disease, and come with some adverse social consequences—like sickness labelling and negative social conformity (Rose 1985).

As an alternative, Rose advocates an approach that "shift[s] the whole distribution of exposure in a favourable direction." This involves making a small uniform change in the entire population (such as decreasing the average daily serving of saturated fats by 50 grams) that results in a population-wide benefit in the reduction of disease, injury and/or mortality. Rose's 'shift' is beneficial when small reductions in risk to the whole population can reduce the entire burden of disease. When the strategy works, persons with low and high risk both experience the shift, and under the assumption of a roughly linear relationship between exposure and outcome, even a small overall shift will result in a notable reduction in overall illness. Rose advocates the population-based approach partly because he believes in the notion of social causes of disease—risk is something that a society shares, and is a product of complex social, cultural and historical mechanisms. Prevention, therefore, should be

about large-scale social change, rather than attempts to change the behaviour or habits of individuals in isolation.

The most substantive examples of successful population-based interventions are seen in the reduced rates of cigarette smoking in Western countries (Pierce 1989) and the successes of HIV prevention programs (Rietmeijer et al., 1996). Smaller scale successes have also been observed (e.g. Buchbinder and Jolley 2004; Laaser, Breckenkamp, Ullrich and Hoffmann 2001; Gortmaker et al., 1999). Nevertheless, there are some practical criticisms of the population-based approach. Without adequate research directed at individuals and the 'necessary causes' of disease in those individuals, we can never have the intellectual certainty required for prevention of any form (Charlton 1995). Some empirical evidence also questions the merits of the population-based approach. In spite of some successes, many attempts at population-based prevention (and in particular, health promotion) have had unsatisfying results (Merzel and D'Afflitti 2003). Finally, the ethics of the population-based approach have come into question; evidence of "J-shaped" relationships between exposure and outcomes suggests that some people can be adversely affected by population-based approaches (Adams and White 2004). Preventative strategies that are good for the majority may sometimes adversely affect the health of a minority. For example, although health promotion programs that encourage lower fat intake may be an important message for much of North American society (given the increasing rates of obesity) a minority do not benefit from this particular message (e.g., those with naturally low levels of body fat), and some may even be harmed by it—such as those for which dieting culture is a risk factor for eating disorders (Austin 2001).

In response, advocates of population-based prevention have argued that the real benefit of a population-based approach is an indirect product of policy. By implementing population-wide change, there is a potential to permanently modify social norms, and encourage healthier behaviour in the future, independent of active, health promoting interventions (Rose, 1985; Rose 1992 pp.73-94). Proponents argue that this not only results in a healthier society, but also reduces the effort and costs associated with prevention in the long term. Second, the biomedical model does not

provide a complete picture of causation, and is not always the most logical point of intervention. Individual-level lifestyle factors have often shown weak predictive power for common health outcomes (Ericsson 1997). This may be related to the functional distance between the adverse outcome and its causes. Society has characteristics that are indirectly causal to disease, but are accessible and efficient foci of intervention (Schwartz and Diez-Roux 2001). When more immediate causes of disease are inaccessible, then the most important causes (from the standpoint of disease prevention) are those that can be points of intervention, regardless of their position or strength in the chain of causality (Rose 1992 pp. 98-100; Schwartz and Diez-Roux 2001).

Much of the dispute between population-based and biomedical philosophies of prevention is chiefly a matter of approach, since both advocate a similar general goal—that is, good health, and the prevention of illness. On the one side, there is an emphasis on the notion of *public* health—to seek improvements in the well being of society as a whole. On the other side, the emphasis is on treating and preventing disease in *individuals*, which often includes a mistrust for patronizing (and usually centralised) intervention on the part of government or society. The division runs parallel to an old, yet recurring debate in the social sciences (and more generally, philosophy) between methodological holism (e.g., Kincaid 1996; Mooney 2005), and methodological individualism (e.g. Popper 1993; Elster 1989).[1] Methodological holists argue that the structures of social relationships, should be studied independent of the individuals that comprise them. Communities, groups, classes, sexes and other organized relationships do not sum to the total of individuals within them, and require their own methodological approaches, and dedicated research. Methodological individualists, on the other hand, argue that all human relations can be understood from the disaggregate constituents, and that there are no 'group-facts' that cannot be attributed to individuals. From this perspective, the study of group-facts and social institutions is just "shorthand for talk about individuals who interact with one another and with people outside the institutions" (Elster 1989 p.158).

[1] Excellent reviews of this debate, including historical the roles of Weber and Durkheim, can be found in Hollis (1994) and Braybrooke (1987).

In addition to this debate about the proper scope of preventative intervention (whether interventions should focus on individuals, communities, genders, classes, etc.) there is a further challenge of defining a fair or just distribution of resources. The population-based strategy of prevention implies a goal of maximizing health improvement, that at least in its traditional form, seems clearly a matter of efficiency and ultimately, a measurement of utility; i.e., a preventative strategy is justified in terms of its costs and benefits. There is a history of philosophical alternatives, however, going as far back as Platonic idealism, to the moral philosophy of Kant, to modern feminist ethics of care. Rather than taking on the ambitious task of reviewing disease prevention policy with respect to all frameworks of moral philosophy, we discuss prevention policy in the context of two specific theories of distributive justice popular in modern Western democracies: utilitarianism and the difference principle.

3.3 Population-based utilitarianism

The ethical justification for the population-based strategy of disease prevention appears to be tied to some form of utilitarian philosophy. Utilitarianism takes many forms, but in its simplest terms, is a decision system that attempts to explicitly maximize some overall measure of social good. The 18th century English philosopher Jeremy Bentham was the first to develop a 'calculus' to measure the pleasures associated with a particular act. The utility of an act (individual or collective) equals the product of intensity, duration and number of people involved. Among a possible set of acts, the act that results in the greatest utility is the most just. John Stuart Mill developed a more sympathetic form of utilitarianism which observes that some measures of utility are of greater value than others, and that personal utility is often connected to (or dependent on) the utility of others (Mill 1979). Mill is often regarded as a proponent of 'rule utilitarianism', in which competing rules (as opposed to acts) are compared against overall measures of utility, with the utility maximizing rule codified as the best (and most just) choice. 'Preference utilitarianism' represents a more modern perspective in which the satisfaction of preferences is maximised rather than 'pleasure' or 'the good' which are difficult to quantify (Gandjour and Lauterbach 2003). In public health applications, a utility objective could be general welfare, length or quality of life, preferences for care, or even a combination of

various measures. Typically, utilitarians define moral policy as that which maximizes utility (and sometimes the weighted utility) summed over all individuals. When different acts, rules or policy options are compared, they are judged based on their consequences (when known) or an expectation of consequences (when unknown) to the individuals involved. Given the option of vaccinating or not vaccinating children against chicken pox, the right policy is determined by comparing the sum of outcomes under both scenarios, and then choosing the option which results in the greatest total utility.

The consequence-weighing (or 'consequentialist') practice common to utilitarianism has been the basis of considerable discussion among philosophers and social scientists. Most classical criticism has come from two directions. First, critics suggest that it is inconsistent with the kinds of decisions humans are inclined to make in the real world. If a majority of the population saw value in the enslavement of a minority, classical utilitarianism would offer a formal mechanism to justify it. More modestly, this criticism suggests that utility maximising functions are not sufficiently sensitive to issues of distribution; very unequal distributions of goods and opportunities are indistinguishable from equitable distributions of the same goods and opportunities that result in the same total utility. Critics consider this to be in conflict with the conventions that most human societies are likely to accept; all else being equal, extreme differences in distributions are generally undesirable. The second, and more abstract criticism challenges the teleological nature of all consequentialism. Morality, critics have argued, is not a function of ends, but has (or should have) intrinsic value; rules should be followed for their own sake, not some opportunistic assessment of consequences. This position was held by Kant, who argued that obligatory moral behaviour was determined by discovering universal rules rather than measuring consequences (Johnson 2004). Murder is always wrong because it fails to survive a test of universality—if everyone murdered all the time (i.e., universally), everyone would be dead. Not-murdering survives the test (since it can be applied universally) therefore it is an incontrovertible moral rule, or 'categorical imperative'.

In spite of this and other historical alternatives, utilitarian, and more generally, consequentialist reasoning is an important idea in many areas of public health policy.

Thinking in modern public health has been strongly influenced by Rose, whose arguments seem to flow from a fairly strict form of utilitarianism: the best preventative intervention is often that which maximizes health over the whole of the population. Many health economists also remain committed to consequentialist ideas—such as quality adjusted life-years (or QALYs) (McKie et al. 1998) fair-innings (Williams 1997) as well as 'softer' measures such maximizing the fulfilment of reasoned claims to care (Savulescu 1998).

Although the population-based strategy is consequentialist (since its chief concern is effectiveness of prevention) it is not always a utility maximizing perspective. If it were, then the preventative focus might be on economic definitions of efficiency—getting the greatest reduction in disease for the least cost. As noted above, a classical utilitarian may advocate that prevention resources are distributed to specific subgroups of the population—such as the middle of the bell curve, or the people who are most likely to respond to prevention advice. In these examples, the allocation of prevention resources may be efficient, but would ignore broader responsibilities to the public as a whole. The very notion of public health suggests a concern for the health of the entire population, even at some costs of efficiency. Furthermore, efficient distribution does not guarantee changes in social norms, and the population-based strategy may also find this insufficient. What the population-based approach does inherit from utilitarianism is its global perspective—that each individual is part of the decision formula. Information is gathered on the status of the population as a whole, and a strategy is implemented based on the policy that best meets the goals of disease prevention and health promotion.

3.4 The difference principle in disease prevention

Disease and injury vary between and within human societies. In some cases, human intervention plays a direct role in the differential allocation of disease and injury—such as in the distribution of violent conflict. In other instances, there are important social and cultural realities which constitute an indirect, and often unanticipated, effect on the distribution of disease and injury. For example, inadequate nutrition (Levy et al. 2005) and inadequate housing (Konradsen et al. 2003) have associations with adverse health status. Income shows relatively

consistent regional, national and international associations with many forms of disease, injury and mortality. Fair access to these and similar resources is an issue of distributive justice. Whatever the mechanisms that actively distribute disease, notable variation persists in many of the causal factors attached to social issues (Woodward and Kawachi 2000; Braveman and Gruskin 2003).

Definitions of equality can take many forms: equality of capabilities, equality of rules, equality of freedom from constraint, and equality of goods just to name a few (Marchand, Wikler and Landesman 1998). Most, if not all theories of ethics are concerned with some notion of equality—in the applicability of rules or the distribution of goods, opportunities or capabilities (Rice 2002). These same theories recognize that inequality is permissible as long as it is in some sense fair, and in some way morally justified. From this idea, ethicists and social theorists derive the term 'equity'—equality in some sort of moral context.

Issues of equality and equity, independent of absolute levels of wealth and prosperity, have received considerable attention from health researchers in the last two decades. Some have argued that social inequalities (in income, education and social class, for example) spawn a chain connecting the psychosocial effects of low social class, low social capital and low public participation to poorer health (Wilkinson 1997; Kawachi et al. 1997; Wilkinson 1999). Some have also argued that reducing the overall inequality of illness may even have a more immediate positive impact on health than a widespread growth in resources and wealth (Daniels, Kennedy and Kawachi 1999). Research has supported (Kahn et al. 2000; Koch and Denike 2001; Cooper 2001; Hou and Myles 2005) and challenged (Sturm and Gresenz 2002; Deaton 2003) this theory, but public support for the idea of equity, in some sense, is relatively persistent (Andersson and Lyttkens 1999).

The suggestion that a social phenomenon like inequality could affect the health of individuals is not inconsistent with Rose's theory of population-based prevention. Social determinants of disease are well accepted components of Rose's vision of population-based public health. However, research has suggested that such broad consequential assessments of health care resource distribution do not satisfy public senses of fairness (Wailoo and Anand 2005). Some basic human intuition seems to

include a desire to help those who are particularly sick and vulnerable, even if there is a net cost to the well-being of society as a whole (McKie and Richardson 2003). Any goal to maximise the health of the population does not take into account that some people are in the tail of the distribution, and may need more than an 'equal' share of preventative and even treatment resources. Thus, in spite of the appeal of population-based approaches to disease prevention, they may be criticised as insufficient to ensure that equitable resources are distributed to cultural, social or geographical subgroups of the population that are in particularly high need.

In 1971, John Rawls first published *A Theory of Justice*, a treatise in which he offered a framework of social interaction focused on issues of fairness (rather than consequences), and with a basis in the beliefs common to citizens of modern Western democracies (Rawls 1971). His theory is often regarded as coming from the social contract tradition, where social participation is rationalized from an asocial 'original position'. Social contract theorists use the idea of an original position as an hypothetical starting point of political reasoning; in order to decide on how political society should be structured, one must determine if political society has value in the first place. This is done by transcending existing political arrangements, and imagining a state in which competing rational agents are not protected (or regulated) by enforced political order. One of the earliest social contractarians, the 17th century thinker Thomas Hobbes, argued that persons in this position accept the structures of government because life outside government-enforced social order is 'solitary, poor, nasty, brutish and short' and above all, painful (Hobbes, 1991). Although no real negotiations take place between the agents in this state of nature, the simple thought experiment leads to certain conclusions, namely, that sacrificing some of the personal freedom afforded in the 'state of nature' is sensible given the dangers of apolitical existence. Other early social contractarians emphasized the importance of liberty (Jean-Jacques Rousseau (1997)) and personal ownership of labour (John Locke (1980)), but similarly regard social participation (and in general, government) as a means to achieve the personal goals of rational individuals. For social contract theorists, the agreement between individuals and society, though formed

conceptually, represents a critical starting point for resolving ethical and political questions.

Rawls proposes a social contract negotiated from a position of impartiality—behind what he refers to as a 'veil of ignorance'—where participants are ignorant of their own attributes (e.g. race, sex, natural abilities), personal life objectives, and who are 'mutually disinterested' in other persons. Since the veil of ignorance enforces a personal neutrality, Rawls argues that decisions about the rule of law and social structures will not be biased by historical personal objectives or status. Furthermore, the moral importance of individuals (an important theme in Rawls' framework) is preserved since it is individuals (or their representatives) that adjudicate morality. Rules are not determined by measuring consequences for society, but through the negotiations of self-interested agents behind the veil. Rawls goes on to argue that such impartial agents would choose to live in a society that:

1) maximizes equal basic civil liberties and

2) a) permits unequal distribution of social and economic goods when such inequality benefits the worst-off most, and b) ensures positions of social opportunity (like employment, political leadership, etc.) are equally accessible to all

Condition 2a is often referred to as the 'difference principle' or 'maxmin' (maximize the minimum). It stresses the importance of equality in the application and design of rules in spite of the potential for inequality in ends. In this way, his ideas are about procedural equality, not about the equality of outcomes; some inequality is acceptable so long as the procedures are fair, and in particular, the worst-off most benefit from inequality.

Rawls' original work paid little attention to the implications of the difference principle on public health and health policy (Arrow 1973). However, other thinkers have extrapolated from his original theory to include the distribution of health resources. Daniels adopts the "fair and equal opportunity" approach to justifying equitable health resource allocation in a Rawlsian framework (Daniels 1997; 2001). His argument rests on the assumption that a social contractarian behind the veil of ignorance would view health in the same way she would view fair access to

opportunities. In order to participate in the important institutions of society (such as political participation and employment), she would want to be secured of a "reasonable range to opportunity". This includes the liberties and sufficient health status required to occupy important social positions. In this way, health is viewed as more than just a material good, but as something of normative value, and as a prerequisite for participation in society. As such, it is equitable to preferentially favour those who suffer from the most debilitating health outcomes, and in particular, outcomes that inhibit their ability to participate in politics or the economy. Part of Daniels' argument rests on research which emphasizes social causes of disease; social inequality in health causes bad health in individuals. As a result, it is practical, as well as ethical, to view poor health as an obstacle to fairness, and to seek equality of outcomes and care as much as reasonable.

Green disagrees with Daniels' position, and argues that health care should be treated as a social/economic primary good subject to the procedures of the difference principle (Green 1976; 2001). Impartial actors would view access to health care as an entitlement (subject to resource constraints) within a framework ensuring that inequalities most benefit the worst-off. Access to health care resources is not a right (in the way freedom of speech may be) but it is just to ensure that persons who are disproportionately worse-off are secured of disproportionately greater resources. Where Daniels' view interprets health care resources in a similar manner as civil liberties (and subject to either tenet 1 or 2b), Green sees health care resources as a good to be allocated subject to the difference principle (2a). The remainder of our discussion assumes position 2a; that is, health care resources (and specifically, prevention resources) are considered a primary social good subject to the difference principle.

In many ways, the Rawlsian view represents a modest compromise; it is an individual-based, procedural framework that supports the concept of equal liberties, and relative equality of primary social goods. Though based on traditions of Western liberalism and individualism, in practice, Rawls' theory is equality seeking since he believes that neutral rational agents are likely to limit inequality in the distribution of resources. Although it is not clear how (or if) Rawls intended the difference principle

to apply to matters of health resource allocation, it seems reasonable that a definition of social goods should include resources designed to promote health and prevent disease whether or not health is itself considered a social good. The administration of disease prevention strategies, under the difference principle could be justified as equitably unequal; for example, persons who suffer a high burden of disease and/or risk of disease are entitled to a larger proportion of treatment and prevention resources.

Research has suggested that people, when asked to imagine themselves in an self-interested position behind Rawls' veil of ignorance, do not readily accept the maxmin strategy (Frohlich, Openheimer and Eavey 1987; Miller 1992) although this may be culturally specific (Bond and Park 1991). Others argue that the difference principle is exceedingly conservative, and ignores the human tendency to take risks; some people may be willing to sacrifice the security of relative equality if there is a chance (and in some cases, even a small chance) of great reward. There is also evidence that equal outcomes are not favoured when variation is the result of natural ability or even good fortune (Bukszar and Knetsch 1997).

Others have questioned the efficiency of the difference principle. In response to the difference principle in the context of disease treatment practices, Arrow writes:

> "[...] there can easily exist medical procedures which serve to keep people barely alive but with little satisfaction and which are yet so expensive as to reduce the rest of the population to poverty. A maxmin principle would apparently imply that such procedures be adopted." (Arrow 1973, pp.251)."

Provision of resources to worst-off individuals is inefficient since it provides scarce resources to people who are likely to consume them at little individual benefit. A terminally ill, brain-dead patient may possess the worst-off health status, but investing massive efforts to extend his life (or marginally improve the quality of his life) seems wasteful, especially since the resources could prevent the less intense suffering of other people with a much greater chance of success. Some medical technology offers treatment that is very expensive, and that provides only marginal benefits to health (in years lived and/or quality of life). In societies where rights to treatment are guaranteed (however small the chance of survival or miniscule the reduction of pain and suffering), individuals will be inclined to exploit these rights to

consume expensive 'low-yield' treatment options, and in doing so, eventually overburden public resources (Dworkin 1993). The continuous re-allocation of health care to the worst-off would not only cripple the system of health care, but also unjustly move resources from the majority to the minority.

There is, however, research that supports the idea of the difference principle in practice, though it is not explicitly used as justification. Recent work in housing policy suggests that the homeless can be classified into three groups: transitionally homeless (who make up the majority of homeless persons), episodically homeless and chronic homeless (Kuhn and Culhane 1998). Most current systems for treating the homeless are homogenous, and population-based. Resources are pooled together—into institutions like homeless shelters and detox programs—and the homeless have equal access to these resources to help move from homeless into non-homeless lifestyles. Critics have argued that although a shelter and resettlement system may be sufficient for the majority of homeless (who are not chronic), it is inadequate for chronically homeless, who would benefit from long-term care, and permanent housing assistance (Kuhn and Culhane 1998). The chronically homeless are high cost and repeat users of police, health treatment and other public services. Passive interventions—such as those found in the traditional homeless shelter system—do nothing to deal with persons in the extremes of the distribution of homelessness. However, a massive investment in harm-reduction targeted at the high-risk, chronically homeless could actually save more money and more lives in the long-term (Culhane 1992; Culhane et al. 2001). From the perspective of the difference principle, it would also better meet notions of fairness since the worst-off homeless may have a moral claim to disproportionate prevention, treatment and aid.

Similar to criticisms made earlier, this strategy of resource allocation may be seen as wasteful. Elster captures this line of reasoning clearly: "The welfare state [...] is like a circus: acrobats fall more frequently when they perform with a safety net." (Elster 1991). There are also questions about effectiveness, particularly in the case of disease prevention. Individuals in the worst conditions of health may often be least able to respond to prevention efforts that require their active participation. In some cases, the burden of adopting changes in lifestyle meant to reduce the risk or burden

of disease may be more trouble than the symptoms of disease themselves. Furthermore, groups with the highest prevalence of chronic illness are likely to be heterogeneous. Even if a high-risk group suffers high burdens of disease on average, some individuals within the group will not—because of genetics, behaviour or good luck. After all, though hypertension is a fairly strong predictor of various cardiovascular illnesses, it does not guarantee that a given individual will suffer strokes and heart attacks in the future. Like the population-based strategy, a worst-off disease prevention strategy paints with a broad brush, offering prevention to worst-off groups, which, depending on how the groups were formed, could often include individuals that may not be in particular need of intervention.

3.5 A geographic contribution to chronic disease prevention methodology

3.5.1 Comparison of theory to methodology

The two views discussed above do not exhaust the perspectives from which one may approach the ethics of disease prevention, but do offer strongly contrasting opinions on process, if not outcome. Consequentialism generally, and utilitarianism specifically, build notions of distributive justice on consequences. Utilitarianism supports justice defined in terms of overall measures of public good, but in most forms, also accommodates concerns of equity, civil liberties and protection of minority rights (e.g., Mill 1984). The difference principle is based on the idea that a fair procedure of adjudicating justice should benefit the worst-off the most, even if at times inequalities (usually thought of in terms of the distribution of resources) result. We now turn to the task of contrasting these theories with cluster detection methodology.

It is clear that cluster detection methods do not provide sufficient information for utilitarian, or any form of ethical reasoning in which an objective is calculated over the whole of the population. Since cluster detection is concerned with finding subsets that are anomalously high, most of the data on health outcomes (including the large number of people who hover around the average) are ignored. From a methodological standpoint, this points to a non-trivial disconnect between the tool of analysis and the method of decision making. Clearly, population-based strategies

require methodologies that describe the distribution of illness over the whole of the population in order to anticipate and observe the shifts in the population curve.

This is not to suggest that cluster detection methods have no value in population-based public health. In some cases, clusters may indicate a pattern of widespread importance—such as infectious disease outbreaks, or rapid changes in the spatial distribution of disease. Preventing the spread of infectious disease to the population as a whole sometimes requires prevention directed at high-risk individuals. For example, close contacts of a person diagnosed with smallpox are likely to receive considerable intervention from public health officials for the sake of greater public safety and disease containment. Such efforts are often referred to as 'herd protection'; to keep the population free from infection, interventions need to be directed at vectors, which are sometimes people, and these people sometimes cluster (geographically or otherwise). However, it is less easy to associate cluster detection methods to population-based chronic disease prevention. Cluster detection methods simply do not offer the information required to introduce or monitor a population-wide health promotion or disease prevention effort. If the policy goal is to change population-wide social norms, and make small widespread reductions in disease risk, then an overall measure of outcomes is required.

On the other hand, there are some obvious parallels between the language of the difference principle and the methodology of cluster detection. Disease cluster methodology is concerned with the extremes of the distribution, and usually, negative states—like high disease burden or high concentration of hazards. The difference principle is concerned with the worst-off groups—the tail of the distribution of disease burden (or risk of illness)—which, since they depart from the mean, can be thought of as noteworthy subsets of the population. These subsets, or clusters, could represent failures to meet the basic criteria of the difference principle; that is, ensuring that the worst-off benefit the most within a process of goods transfer (in this case, distribution of disease prevention resources). A vision of distributive justice concerned with fairness considers these clusters *unfair*, which may entitle their constituent population to disproportionately more resources than the rest of the population. Cluster detection does not provide information about the entire range of

variation but it does offer insight into the groups that are worst-off—in terms of health outcomes, disease burdens, and in some circumstances, disease risks. The difference principle requires this kind of information to formulate a response to distributive options; in other words, who are the worst-off groups, and how worse-off are they?

In the context of disease prevention, the difference principle may demand that some preventative resources be distributed unequally when the inequality most benefits the worst-off. Many geographic cluster detection methods provide a relatively simple answer to the problem of identifying who the worst-off actually are. Such methods (such as Turnbull et al., 1990 and Kulldorff 1997) determine whether or not a cluster of disease exists, but also identifies its approximate location. However, the difference principle is not specific about the dimension in which it operates; worst-off groups often (and probably most often) exist in non-geographical spaces—such as gender, ethnicity, income and age. For example, chronic diseases like Alzheimer's and Parkinson's typically 'cluster' in older age groups; persons of older ages suffer the highest incidence and highest burden of these diseases. Other illnesses cluster in gender space (like breast and prostate cancer) and others yet in ethnic, cultural and income space. We may also expect complex multi-dimensional clusters—such as motor vehicle related deaths among men 18-25 years of age. This points to a noteworthy challenge, however, since it appears that the choice of dimension somewhat trivializes the cluster detection exercise. By choosing to search for chronic disease clusters in a particular dimension, we may have ruled out other important dimensions within which other noteworthy clusters occur.

There are two ways of responding to this challenge. First, a search for worst-off clusters could involve a simultaneous investigation of all important dimensions (age, sex, geography, income etc.) and the selection of a worst-off cluster across all these dimensions as the point of preventative intervention. In this case, the most noteworthy cluster of all the dimensions requires the most attention. This is already implicit in many public health activities—some disease prevention concentrates on ethnic differences, some on gender differences and some on age differences. However, this approach does not totally resolve this challenge since a large number of

complex associations between the dimensions (for example, between age and sex or between ethnicity and income) are likely to exist. An alternative perspective is to explicitly link the mode of prevention with the dimension of interest. Geographic prevention efforts—such as environmental modification and facility location—should be informed, at least in part, by cluster detection in geographic space. Even if the process governing the spatial distribution of disease is governed by the spatial distribution of age (which correlates strongly with disease risk) a geographic cluster is still informative when the intervention is geographical. After all, treatment facilities cannot be allocated to age space, and environmental modification is not usually gender specific.

Geographic cluster detection methods can find geographic communities of high disease burden, and can reveal important information about community equity. Such ideas about equity and geographic community are not new. In *Social Justice and the City*, David Harvey describes one form of 'territorial social justice' in Rawlsian terms (Harvey 1973 chapter 3). A just distribution ensures that "[the] mechanisms (institutional, organizational, political and economic) should be such that the prospects of the least advantaged territory are as great as they possibly can be" (Harvey 1973 pp.116-117). However, finding the boundaries of such territories is not straightforward; social neighbourhoods are dynamic, complex and do not form mutually exclusive sets (Kawachi and Berkman 2003 pp.1-17). Research that attempts to understand the effects of geographic communities on individuals reports a contrast of findings that could be explained by these complexities (Ecob and Macintyre 2000). Furthermore, this invokes the debates of methodological scope mentioned above. While methodological holists try to identify which geographic groups require study, methodological individualists advocate reduction to the individuals who make up these groups.

The worst-off communities identified by geographic cluster detection techniques may be logical starting points for certain types of community-based chronic disease prevention programs, and in some sense, offer a compromise between the holists-individualist debate. Geographic clusters are locations where incidence and/or prevalence is anomalously high. These locations may understood as groups of

individuals, communities, or groups of several communities in geographic space. Within the constraints of the available data and the design of the cluster detection algorithm, geographic cluster detection offers a method for identifying groups of high disease burden or risk without first imposing a definition of scope. These groups, what we will refer to as *cluster-communities*, are synthetic geographic subsets of the population designed by a cluster-detection algorithm to meet a threshold of concern. Agents within a cluster-community (whether viewed as individuals, households, neighbourhoods or communities) may be different in all respects except that they all persist within an area of high disease burden. This commonality—though only an artifact of the algorithm creating the cluster-community—binds the agents to prevention policy in a way that is indifferent to epistemological debates of scope.

3.5.2 Example applications

The focus of this discussion has been on chronic disease prevention, largely because the case for using clusters to inform infectious disease prevention is fairly easily made. From a biomedical, population-based, or equity-based standpoint, preventing the spread of these diseases is of widespread interest. A cluster of serious infectious disease—such as smallpox—necessitates interventions that are in the interest of the population as a whole. The final task of this exercise is to discuss applications in which knowing about worst-off geographic groups—cluster communities—could be applied to chronic disease prevention.

3.5.2.1 Cluster-communities as proxy geographies

Disease cluster-communities may be sensible groups for intervention when cost or privacy issues make collecting certain types of individual-level data difficult. For example, it would be time consuming, expensive and intrusive to identify where all the long-term intravenous drug users reside in a large urban centre. When health data from disease registers or other administrative data are available, this information could be used as a proxy for identifying the location(s) of this high-risk community. If related diseases form a noteworthy cluster, then this information can be used to guide local interventions—like needle exchanges and community outreach activities. The actual distribution of intravenous drug users remains unknown, but this

information is indirectly informative. Its real value is not in the prevention of high-risk behaviour, but in helping to manage associated problems.

A cluster-community based approach may also offer a practical method for applying the difference principle that escapes the inefficiencies described earlier. Some may argue that individuals with the greatest burden of disease are typically less able (and often incapable) of responding to prevention or health promoting interventions. Directing disease prevention strategies to worst-off individuals may consume considerable resources with little personal or social benefit. However, health statistics are averaged or summarized at the cluster-community level. Although individuals may be 'inefficient' objects of intervention when guided by the difference principle, it is certainly possible to dedicate resources to worst-off cluster-communities without incurring this form of inefficiency. Environmental modification, for example, does not require active allocation of resources to specific individuals in a community, but passively allocates resources to everyone exposed to that environment. Bringing a worst-off cluster-community towards a regional average through such interventions may be costly, but it is not *individually* inefficient, even if some individuals in that community are not at risk. Furthermore, it offers a tangible and measurable reward—shifting the average disease burden of persons in worst-off cluster-communities to a healthier location on the outcome distribution curve.

Advocates of population-based disease prevention could point out that such changes positively alter social norms; there is a 'value-added' utility gained by implementing population-based strategies since they perpetuate healthier lifestyles, which will require less intervention in the future. Like population-based strategies, cluster-community based prevention strategies can also benefit from changes in social norms, particularly when the cluster-communities form some sort of natural group of interaction. As in the population-based strategy, community-specific interventions could be developed and implemented in order to alter local norms of health behaviour in the long run. Cluster-community based interventions may also help build positive social capital, and even increase the trust and credibility community members have for public health interventions in general.

3.5.2.2 Cluster-communities as formal geographies

'Area-based initiatives' emphasize the geographic community as a focal point of intervention, and have been of interest for several decades, particularly in the United Kingdom. Historically, the ethical justification of area-based initiatives comes from concerns about social inequality, but in particular, the status of the most deprived communities (Smith, Noble and Wright 2001). Communities with large numbers of unemployed adults, lower levels of education, poorer health and higher crime are provided various resources, usually with local participation, with the hope of improving the local social and environmental infrastructure. Examples include: new immigrant outreach programmes, the creation of healthy living centres (such as recreational facilities and affordable homes for the elderly) and neighbourhood renewal (Regional Coordination Unit, 2006).

One issue that poses a regular challenge to area-based initiatives is the difficulty in defining the areas of intervention. When the definition of an area suited for intervention is not formalized, it could lead to questions of legitimacy—why is one area receiving attention and not another? Formal methods, like those used in cluster detection, allow us to link these program decisions to decisions about methodology. Choices about statistical significance, magnitude of difference, search strategy, and statistical controls can all be debated and defined ahead of time. Once these terms are negotiated, methods can define cluster-communities that meet standards of rigour and repeatability, and make decisions easier to defend. This formalized approach also makes evaluation easier. For example, formally defined cluster-communities from a single city can be assigned two different intervention programs, and the effectiveness of these strategies can be evaluated by comparing the relative change of the cluster-communities with respect to each other and to other communities in the same city.

Some critics doubt the ability of area-based initiatives to efficiently target those in need—neighbourhoods are too heterogeneous to make neighbourhood-level programs efficient (Joshi 2001). However, even detractors agree that in cases of highly concentrated deprivation, area-level intervention may be sufficiently efficient (McCulloch 2001). Furthermore, the practical value of area-based initiatives may only be realized when the interventions are inherently spatial—as is the case in

housing issues, and changes to the physical environment (Tunstall and Lupton, 2003). On these occasions, area-based initiatives can influence the manner in which such spatially referenced interventions are provided without making the intervention process systematically inefficient.

3.5.2.3 Cluster-communities in facility location

Another area of possible application of cluster-community intervention may be in facility location and accessibility problems related to disease prevention. Some disease prevention/health promotion services are offered at local (and sometimes mobile) service centres—two noteworthy examples can be found in cancer screening (O'Malley et al. 2002) and needle exchange programs (Wood et al. 2004). In most instances, budgetary constraints force planners to limit the number of facilities provided while also ensuring some sort of fair service. A common approach to public-facility location is to locate facilities in such a way as to minimize the average distance that consumers travel to the closest facility. Some equity-based (rather than efficiency-based) models have been proposed in the past, particularly in the case of public facilities (e.g., Francis 1967; Toregas and ReVelle 1972). In these instances, problems either seek to ensure that either the distance between a facilities and their clients is no greater than a certain threshold, or seek to minimize the farthest distance that any consumers would have to travel.

In chronic disease prevention, cluster-communities offer information that could influence these kinds of facility location decisions. Under the difference principle, some 'consumers' should be identified as 'high-entitlement', and with a particular right (and need) to a nearby facility. This can be incorporated in several ways. Facilities could simply be located within cluster-communities. Efficiency calculations could even be made within these areas to ensure that persons within the cluster community are served efficiently. However, when multiple facilities need to be located, this approach seems somewhat unreasonable, and could lead to highly inefficient resource allocation. As an alternative, individuals within cluster-communities could receive disproportionate weighting within a traditional efficiency-based location model. In this case, travel distance minimizing functions could still be

employed, but not all individuals would receive equal weight; persons living in a cluster-community would receive a weight greater than the rest of the population.

3.6 Conclusion

The application of cluster detection methods to public health intervention policy has an ethical justification based in the difference principle. Cluster detection methods can identify cluster-communities—discrete and systematic answers to the question of where (in geographic space) the burden of a chronic disease is highest. Unlike many area-based initiatives, which use administrative areas based on the census to administer geographic interventions, cluster detection methods form cluster-communities based on instrumental goals that can be set (and modified) based on specific strategic public health goals. For some diseases, worst-off cluster-communities may not suffer noteworthy burdens of illness; community-specific programs may not be necessary, and population-based strategies meant to shift the entire distribution of illness downward are both ethically and practically justified.

Geography has a role in understanding and mitigating health inequalities (Curtis and Jones 1998). The difference principle, when applied at the level of geographic communities, describes a model of distributive justice ensuring that geographic inequalities benefit worst-off communities the most. Under the difference principle, an unequal geographic distribution of prevention resources is justifiable, and equitable, when it benefits the worst-off. Unfortunately, the definition of geographic community is dynamic, both operationally (as we know from studies of the modifiable areal unit problem) and philosophically. As a response, some advocate that we ignore the notion of geographic community altogether, and instead adopt methods that are 'aggregation invariant' and resistant to different and seemingly arbitrary definitions of community (King 1996). Methodological individualists would recommend that all analysis should be conducted at the level of individuals, ignoring the specific study of community and society altogether. This would link disease and injury to the simplest object of risk—the human person. However, both of these alternatives ignore the role of society on the individual, and assume that 'social facts'—including geographic communities—can be reduced to their constituent parts without losing important information. Social and physical environments seem to

affect people—through historical and social contexts, environmental influences, and personal interaction.

We have emphasized an application of cluster detection that may be useful for certain types of chronic disease prevention. In these applications, cluster detection methods can be used to identify cluster-communities of high disease. Subject to the constraints of the particular cluster detection algorithm, some methods are able to identify areas where disease burdens are anomalously high, and more specifically, most-anomalously high. In geographic space, these methods identify worst-off spatial sets, and offer inferential and systematic information useful for the application of the difference principle in geographic space. When applied to chronic disease, these worst-off clusters represent usually contiguous geographic groups that are suffering high disease burden. These methods may not always identify areas where an overabundance of hazards is present, but still give an indication where prevention and intervention may be needed the most, particularly for chronic conditions associated with serious co-morbid illness. These cluster-communities may have no meaning outside the scope of specific and local health interventions, but they have a clear instrumental purpose; they can provide information about where chronic outcomes are common, and a sense of where the burden of chronic disease is high.

Cluster detection methods have an established tradition in various areas of health research, and are of growing interest in the field of public health surveillance. This discussion has tried to show that cluster detection methods may have practical and ethical relevance to chronic disease prevention. Historically, prevention programs aim to reduce sickness, disability and death in human populations. Focus on high-risk individuals has been criticized for ignoring the social dimensions of health. Population-based prevention has been criticized for being inefficient and patronizing. Cluster detection methods can be viewed as a geographic application of the difference principle that can inform resource allocation when high-risk groups are spatially concentrated. The application of cluster detection to chronic disease interventions may make most sense when the interventions are of spatial form—like environmental modification, community-level education, and the location of health promoting/disease preventing service facilities.

3.7 References

Adams J, White M (2004). When the population approach to prevention puts the health of individuals at risk. *International Journal of Epidemiology* **34**:40-43.

Andersson F, Lyttkens C H (1999) Preferences for equity in health behind the veil of ignorance. *Health Economics* **378**:369-378.

Arrow K (1973) Some ordinalist-utilitarian notes on Rawls' Theory of Justice. *The Journal of Philosophy* **9**:245-263.

Asada Y. (2005) A framework for measuring health inequity. *Journal of Epidemiology and Community Health* **59**:700-705.

Austin S B (2001) Population-based prevention of eating disorders: An application of the Rose prevention model. *Preventative Medicine* **32**:268-283.

Besag, J, and Newell J (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society Series A* **154**:143-155.

Blaney R (1987). Why prevent disease? *In* Doxiadis S. (ed) *Ethical Dilemmas in Health Promotion.* Wiley: New York.

Bond D, Park J C (1991) An empirical test of Rawls's theory of justice: a second approach, in Korea and the United States. *Simulation & Gaming* **22**:443-462.

Braveman P, Gruskin S (2003) Poverty, equity, human rights and health. *Bulletin of the World Health Organization* **81**:539-545.

Braybrooke D, (1987) *Philosophy of Social Science*. Prentice-Hall: New York.

Buchbinder R, Jolley D (2004) Population based intervention to change back pain beliefs: three year follow up population survey. *BMJ* **328**:321-321.

Bukszar E, Knetsch J L (1997) Fragile redistribution choices behind a veil of ignorance. *Journal of Risk and Uncertainty* **14**:63-74.

Charlton R G (1995) A critique of Geoffrey Rose's 'population strategy' for preventative medicine. *Journal of the Royal Society of Medicine* **88**:607-610.

Cooper R S (2001) Social inequality, ethnicity and cardiovascular disease. *International Journal of Epidemiology* **30**:S48-S52.

Conley J, Gahegan M, Macgill J (2005) A Genetic Approach to Detecting Clusters in Point Data Sets. *Geographical Analysis* **37**:286-314.

Culhane D P (1992) The quandaries of shelter reform: an appraisal of efforts to "manage homelessness". *Social Science Review* **66**:428-440.

Culhane D P, Gollub E, Kuhn R, Shpaner M (2001) The co-occurrence of AIDS and homelessness: results from the integration of administrative databases for AIDS surveillance and public shelter utilisation in Philadelphia. *Journal of Epidemiology and Community Health* **55**:515-520.

Curtis S, Jones I R (1998) Is there a place for geography in the analysis of health inequality? *Sociology of Health and Illness* **20**:645-672.

Cuzick J,Edwards R (1990) Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society* **B52**:73-104.

Daniels N., Sabin J. (1997) Limits to health care: fair procedures, democratic deliberation, and the legitimacy problem for insurers. *Philosophy and Public Affairs* **26**:303-350.

Daniels N, Kennedy B P, Kawachi I (1999) Why justice is good for our health: the social determinants of health inequalities. *Daedalus* **128**:215-251

Daniels N. (2001) Justice, health and healthcare. *American Journal of Bioethics* **1**:2-16.

Deaton A (2003) Health, inequality, and economic development. *Journal of Economic Literature* **41**:113-158/

Diggle P J (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Association Series A* **154**:349-362.

Doctor J N, Castro J, Temkin N R, Fraser R T, Machamer J E, Dikmen S S (2005) Workers' risk of unemployment after traumatic brain injury: A normed comparison *Journal of International Neuropsychological Society* **11**:747-752.

Dooley D, Fielding J, Levi L (1996) Health and unemployment. *Annual Review of Public Health* **17**:449-465.

Dryden DM, Saunders LD, Rowe BH, May LA, Yiannakoulias N, Svenson LW, Schopflocher DP, Voaklander DC (2005) Depression following traumatic spinal cord injury *Neuroepidemiology* **25**:55-61.

Duczmal L, Assuncao R. (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* **45**:269-286.

Dworkin R. (1993) Justice in the distribution of health care. *McGill Law Journal* **38**:883-898.

Ecob R, MacIntyre S. (2000) Small area variations in health related behaviours; do these depend on the behaviour itself, its measurement, or on personal characteristics? *Health and Place* **6**:261-274.

Ederer F., Myers M. H., Mantel N. (1964) A statistical problem in space and time: do leukemia cases come in clusters? *Biometrics* **20**:626-638.

Elster J (1989) *Nuts and Bolts for the Social Sciences*. Cambridge University Press: Cambridge.

Elster J (1991) Local justice. *European Economic Review* **35**:273-291.

Ericsson A (1997) The importance of lifestyle to self-assessed health. *Health Policy* **42**:145-155

Francis R (1967) Some aspects of a minimax location problem. *Operations Research* **15**:1163-1169.

Frohlich N, Openheimer J A, Eavey C L (1987) Laboratory result on Rawls's distributive justice. *British Journal of Political Science* **17**:1-21.

Gandjour A., Lauterbach K.W. (2003) Utilitarian theories reconsidered: common misconceptions, more recent developments and health policy implications. *Health Care Analysis* **11**:229-244.

Gortmaker S L, Peterson K, Wiecha J, Sobol A M, Dixit S, Fox M K, Laird N (1999) Reducing obesity via a school-based interdisciplinary intervention among youth: Planet Health. *Archives of Pediatric and Adolescent Medicine* **153**:409-418.

Green R.M. (1976) Health care and justice in contract theory perspective. *Ethics and Health Policy.* R M Veatch, R Branson (ed.) Ballinger: Cambridge.

Green R.M. (2001) Access to healthcare: going beyond fair equality of opportunity. *American Journal of Bioethics* **1**:22-23.

Harvey D (1973). *Social justice and the city.* Edward Arnold: London.

Hobbes T. (1991) *Leviathan.* R Tuck (ed.) Cambridge University Press: Cambridge.

Hollis M (1994) *The Philosophy of Social Science.* Cambridge University Press: Cambridge.

Hou F, Myles J (2005) Neighbourhood inequality, neighbourhood affluence and population health. *Social Science and Medicine* **60**:1557-1569.

Johnson R (2004) Kant's Moral Philosophy. *The Stanford Encyclopedia of Philosophy.* E N Zalta (ed.), http://plato.stanford.edu/archives/spr2004/entries/kant-moral/

Joshi K (2001) Is there a place for area-based initiatives. *Environment and Planning A* **33**:1349-1352.

Kahn R S, Wise P H, Kennedy B P, Kawachi I (2000) State income inequality, household income, and maternal mental and physical health: cross sectional national survey. *BMJ* **321**:1311-1315.

Karhausen L. (1987) From ethics to medical ethics. *In* Doxiadis S. (ed) *Ethical Dilemmas in Health Promotion.* Wiley: New York.

Kass N E (2001) An ethics framework for public health. *American Journal of Public Health* **91**:1776-1782.

Kaufman J S, Poole C (2000). Looking back on "Causal Thinking in the Health Sciences". *Annual Review of Public Health* **21**:101-119.

Kawachi I., Kennedy B.P., Lochner K., Prothrow-Stith (1997) Social capital, income, inequality, and mortality. *American Journal of Public Health* **87**:1491-1498.

Kawachi I, Kennedy B P, Glass R (1999) Social capital and self-related health: a contextual analysis. *American Journal of Public Health* 89:1187-1193.

Kawachi I, Berkman L F (2003). *Neighbourhoods and Health* Oxford University Press: New York.

Kincaid H (1996) *Philosophical Foundations of the Social Sciences.* Cambridge University Press: Cambridge

King G (1996) Why context should not count. *Political Geography* **15**:159-164.

Klonoff H, Clark C, Klonoff P S (1993). Long-term outcome of head-injuries—a 23 year follow-up study of children with head-injuries. *Journal of Neurology, Neurosurgery and Psychiatry* **65**:410-415.

Koch T, Denike K (2001). Equality vs. efficiency: the geography of solid organ distribution in the USA. *Ethics, Place and Environment* **4**:45-56.

Konradsen F, Amerasinghe P, Van der Hoek W, Amerasinghe F, Perera D, Piyaratne M (2000) Strong association between house characteristics and malaria vectors in Sri Lanka. *American Journal of Tropical Medicine and* Hygiene **68**:177-181.

Knox G (1964) The detection of space-time interactions. Applied Statistics 13:25-29.

Kuhn R, Culhane DP (1998) Applying cluster analysis to test a typology of homelessness by pattern of shelter utilization: Results from the analysis of administrative data. *American Journal of Community Psychology* **26**:207-232.

Kulldorff M, Nagarwalla N. (1995) Spatial Disease Clusters: Detection and Inference. *Statistics in Medicine* **14**:799-810.

Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. *Statistics in Medicine In Press*

Laaser U, Breckenkamp J, Ullrich A, Hoffman B (2001) Can a decline in the population means of cardiovascular risk factors reduce the number of people at risk? *Journal of Epidemiology and Community Health* **55**:179-184.

Levy A., Fraser D., Rosen S.D., Dagan R., Deckelbaum R.J., Coles C., Naggan L. (2005) Anemia as a risk factor for infectious diseases in infants and toddlers: Results from a prospective study. European Journal of Epidemiology **20**:277-284.

Locke J (1980) *Second Treatise of Government.* Hackett: Indianapolis.

Marchand S, Wikler D, Landesman B (1998) Class, health and justice. *The Millbank Quarterly* **76**:449-467.

Marmot M. (1998). Improvement of social environment to improve health. *Lancet* **351**:57-60.

Marmot M. (2005) Social determinants of health inequalities. *Lancet* **365**:1099-1104.

McKie J, Richardson J, Singer , Kuhse H (1998) The Allocation of Health Care Resources: An Ethical Evaluation of the 'QALY' Approach. Ashgate: Vermont.

McKie and Richardson (2003). The rule of rescue. Social Science and Medicine **56**:2407-2419.

Merzel C, D'Afflitti J (2003) Reconsidering community-based health promotion: Promise, performance, and potential. *American Journal of Public Health* **93**:557-574.

Mill J S (1979) *Utilitarianism.* Hackett Publishing Company: Indianapolis

Mill J S (1984) *On Liberty.* Penguin: Middlesex

Miller D (1992) Distributive justice: what the people think. *Ethics* **102**:555-593.

McCulloch A (2001) Ward-level deprivation and individual social and economic outcomes in the British Household Panel Study. *Environment and Planning* A **33**:667-684.

Mooney G (2005) Communitarian claims and community capabilities: further priority setting. *Social Science and Medicine* **60**:247-255.

O'Malley A S, Lawrence W, Liang W C, Yabroff R, Lynn J, Kerner J, Mandelblatt J (2002) Feasibility of mobile cancer screening and prevention. *Journal of Health Care for the Poor and Underserved* **13**:298-319.

Openshaw S, Charlton M, Wymer C, Craft AW (1987) Mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of GIS* **1**:335-358.

Pickett K.E., Pearl M. (2001) Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology and Community Health* **55**:111-122.

Pierce J P, Fiore M C, Novotny T E, Hatziandreu E J, Davis R M (1989) Trends in cigarette smoking in the United States. Projections to the year 2000. *JAMA* **261**:61-65.

Popper K (1993) *The Poverty of Historicism.* Routledge: London

Popper K (1959) *The Logic of Scientific Discovery.* Routledge Classics: London

Rawls J (1971) *A Theory of Justice.* Harvard University Press: Cambridge.

Regional Coordination Unit (2006). http://www.gos.gov.uk/rcu/. Extracted April 21, 2006.

Richardson J, McKie J (2005) Empiricism, ethics and orthodox economic theory: what is the appropriate basis for decision-making in the health sector? *Social Science and Medicine* **60**:265-275.

Rice, T (2002) *The Economics of Health Care Reconsidered.* Health Administration Press: Chicago.

Rietmeijer C A, Kane M S, Simons P Z, Corby N H, Wolitski R J, Higgins D L, Judson F N, Cohn D L (1996) Increasing the use of bleach and condoms among

injecting drug users in Denver: Outcomes of a targeted, community-level HIV prevention program. *AIDS* **10**:291-298.

Rose G. (1981) Strategy of prevention: lessons from cardiovascular disease. *BMJ* **282**:1409-11.

Rose G. (1985) Sick individuals and sick populations. *International Journal of Epidemiology* **14**:32-38.

Rose G. (1992). *The Strategy of Preventative Medicine*. Oxford University Press: New York.

Rousseau JJ (1997) *The Social Contract and Other Later Political Writings*. V. Gourevitch (ed.).Cambridge University Press: New York.

Schwartz S, Diez-Roux R (2001) Commentary: causes of incidence and causes of cases—a Durkheimian perspective on Rose. International Journal of Epidemiology **30**:435-439

Savulescu J (1998) Consequentialism, reasons, value and justice. *Bioethics* **12**:212-235.

Smith G, Noble M, Wright G (2001). Do we care about area effects? *Environment and Planning A* **33**:1341-1344.

Sturm R, Gresenz C R (2002) Relations of income inequality and family income to chronic medical conditions and mental health disorders: national survey. *BMJ* **324**:1:5.

Susser M (1991) What is a cause and how do we know one? A grammar for pragmatic epidemiology. *American Journal of Epidemiology* **133**:635-647.

Toregas C, ReVelle C (1972). Optimal location under time or distance constraints. *Papers of the Regional Science Association* **28**:133-142.

Tunstall R, Lupton R (2003) Is targeting deprived areas an effective means to reach poor people? An assessment of one rationale for area-based funding programmes. *Centre for Analysis and Exclusion paper 70.*

Turnbull B W, Iwano E J, Burnett W S, Howe H L, Clark L C. (1990) Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* **132**:S136-S143.

Wagstaff A., Paci P., van Doorslaer E. (1991) On the measurement of inequalities in health. *Social Science and Medicine* **33**:545-557.

Wailoo A, Anand P (2005) The nature of procedural preferences for health-care rationing decisions. *Social Sciences and Medicine* **60**:223-236.

Williams A (1997). Intergenerational equity: an exploration of the 'fair innings' argument. *Health Economics* **6**:117-132.

Wilkinson R.G. (1997) Comment: Income, inequality and social cohesion. American Journal of Public Health **87**:1504-1506.

Wilkinson R.G. (1999) Health, hierarchy, and social anxiety. *Annals of the New York Academy of Science **896***:48-63.

Wood E, Kerr T, Small W, Li K, Marsh D C, Montaner J S G, Tyndall M W (2004) Changes in public order after the opening of a medically supervised safer injecting facility for illicit injection drug users. *Canadian Medical Association Journal* **171**: 731-734.

Woodward A., Kawachi I. (2000). Why reduce health inequalities? *Journal of Epidemiology and Community Health* **54**:923-929.

CHAPTER 4: Comparison of a structured and data-directed cluster search strategy

4.1 Introduction

Clusters of disease have been of interest since at least as far back as John Snow's 19th century investigation of cholera deaths in London. Over time, methods used to detect such clusters have evolved considerably. The earliest could detect the presence of clustering in a study area without providing information as to where specific clusters were located (e.g. Ederer, Myers and Mantel 1964; Whittemore et al. 1987; Cuzick and Edwards 1990). Kernel filtering methods were developed to model patterns of variation that offer visual clues to the presence and location of local clusters (e.g Diggle 1990; Rushton et al. 1996). A different approach to the problem of finding spatial disease clusters concerned finding local autocorrelation in patterns of disease, putting aside the explicit goal of identifying anomalous clusters (Anselin 1995; Getis and Ord 1996). Methods also evolved to identify the location of clusters and to then make statistical inferences about their significance (Openshaw et al., 1988; Turnbull et al. 1990; Besag and Newell 1991).

The spatial scan (Kulldorff and Nagarwalla1995; Kulldorff 1997) method of cluster detection has received considerable applied and methodological research attention over the last ten years. The spatial scan uses moving circular windows to scan a surface for high (or low) rates of disease, identifying the presence and significance of local clusters. With the release of SaTScan™ (Kulldorff and Information Management Services 2004), a freely available program enabling searches of time, space and space-time clusters, applications of this method have become increasingly common and varied. Research has expanded from various fields of human health into the areas of veterinary medicine, plant biology, criminology, and history (Kulldorff 2005).

Recent methodological research has sought to improve upon the original spatial scan by expanding the search process beyond the circular window and nearest-neighbour paradigms. We propose an approach to detecting clusters of disease based on some of these more recent ideas in disease cluster detection. By employing some simple graph theory concepts, our method can find clusters of irregular shapes and

sizes within data aggregated to tessellations of polygons. The presentation and evaluation of this approach serves two purposes. First, it offers an alternative method of cluster detection that may be suitable for applications in geography, epidemiology and public health. Second, it reveals some geographic observations about the properties of different approaches to cluster detection methodology.

4.2 Background

4.2.1 Clusters and cluster detection

Waller and Gotway (2004 pp.155) offer three questions involved in most disease cluster activities:

1) Are cases of disease near each other?

2) Does some area in a study region have a disproportionate collection of cases?

3) Where are the 'most unusual' collections of cases located?

For each of these questions, clusters are defined by some indicator of geographic closeness in the occurrence of disease. These questions also suggest that closeness is noteworthy only when it could not have occurred 'by chance'. However, there are also non-statistical definitions of clusters. Knox (1988) defines clusters as geographically 'bounded' systems that are similar in character or genesis. This conforms more closely to the definitions of cluster analysis and data classification. It is not always simple to distinguish cluster detection as an anomaly finding exercise from cluster detection as a spatial classification exercise, but these differences are usually related to application and data. When diseases are acute and harmful, the rapid detection of anomalies is important; the quicker an infectious disease outbreak is identified, the quicker an intervention can be implemented that reduces its further spread. When there is little or no stochasticity in the phenomena under study—such as when diseases are relatively common, or the phenomena of interest are nominal, classification can be more important. For example, a geographic cluster of children living in houses more than 75 years old may be informative for public health officials worried about health risks associated with lead paint.

Even within these definitions of clusters, applications can influence details in approach. Research in epidemiology often emphasizes detecting clusters of residual risk, rather than raw observations about excess disease rates (Marshall 1991;

Wakefield, Kelsall and Morris 2000). In these instances, it is often important to control for known factors that influence the spatial distribution of disease—such as geographic variation in age or sex—so that unknown risk factors can be discovered. Applications related to disease surveillance and alarm evaluation have a different purpose, applying cluster detection to decision-making tasks (Lawson and Kulldorff 1999). In these instances, controlling for known risk factors is not always important; mitigating a current cluster of events (for example, by quarantine or inoculation) does not always require knowledge of the processes or causal mechanisms behind it.

4.2.2 The spatial scan approach to cluster detection

Cost, privacy issues and computational constraints limit most modern applications of cluster detection to working with aggregate representations of disease risk. These representations usually consist of disease frequency data (the numerator in the calculation of a rate) and at-risk data (the denominator in the calculation of a rate). In discrete geographic aggregation systems, these data are commonly defined by the boundaries of a polygon tessellation formed by administrative census areas—such as census tracts. Population or geometrically weighted centroids are often used to represent the locations of polygons in these tessellations. Attributes (such as population and case counts) and topological features (such as adjacency) associated with these polygons are projected onto the centroids they enclose. For some applications, it may be useful to conceive of these systems of centroids as *graphs*. All graphs consist of two sets: a vertex set and an edge set. A graph is an object composed of both these set types, of which the vertex set is nonempty, and the edge set may be either empty, or composed of paired subsets of the vertex set (Trudeau 1993). In Figure 4.1, the vertex set has 5 elements {1,2,3,4,5} and the edge set has 8 elements {{1,2},{1,4},{1,5},{2,4},{2,3},{3,4},{3,5},{4,5}}. In cluster detection applications, the centroids constitute the vertex set, and adjacency characteristics constitute the edge set.

When working with tessellated data, the geometric or population-weighted centroids of the polygons can be treated as the vertex set, and the edges can be defined by some measure of adjacency or proximity. Typically, two centroids are connected by edges when the polygons enclosing them are adjacent to one another.

The graph representation can easily contain the information critical to cluster detection—such as the disease information and the topology—but not the often extraneous information related to the shape or area of the polygons. In general terms, a detected cluster is a subgraph within this graph that exhibits a pattern of disease that is in some way interesting.

In disease cluster detection, it is most common to search for groups of observations that exhibit an unusually high rate of disease. In its most unconstrained and unstructured form, and even without concern for how 'high rates' should be evaluated, the scope of this problem is enormous: evaluate all possible subgraphs within a graph. This is a combinatorial problem based on k selections from n centroids of a size determined by:

$$\sum_{k=1}^{n} \frac{n!}{k!(n-k)!}.$$

A search of a 40 centroid graph would require over one-thousand billion combinations to fully enumerate the search space. A simple heuristic can be used to find the subgraphs with the highest rates of disease in this unconstrained formulation. This is simply a matter of sorting the centroids from highest to lowest rate of disease, and then choosing the ordered set of centroids with the highest rate. Unfortunately, this method gives no indication of *spatial* clustering, since spatial relationships are ignored. Almost all spatial clustering methods are concerned, in some form, with proximity. Some methods use moving windows, some use nearest neighbours and some use adjacency. By constraining the problem these ways, many evaluations are eliminated from the formulation above, making the problem more tractable and more meaningful to the task at hand.

The spatial scan cluster detection method (Kulldorff and Nagarawalla 1995; Kulldorff 1997) borrows from the mechanics of several earlier approaches, and in particular Turnbull et al. (1990). The spatial scan approach has been implemented in a variety of disciplines over the last decade (see Kulldorff 2005 for a review of applications). Conceptually, the spatial scan uses a circular window of increasing size to progressively group neighbouring centroids, evaluating the anomalousness of the window at each step (Figure 4.2). Once all window sizes are exhausted for one

centroid, the scan moves to a new centroid, and continues in the manner above until all centroids have spawned a series of searches. Though often easier to describe in terms of circular windows, the problem can also be conceived as a nearest-neighbour ordered search in a graph of centroids (Figure 4.3). Centroids are sequentially grouped to a 'seed' centroid (in order of closest to farthest) until all neighbours have been added, at which point, the scanning process moves to a new centroid location. Since every centroid is a unique starting point of a search, the spatial scan is a fully enumerative search of sequentially aggregated nearest-neighbour centroids. An ordered search of neighbours is suitable for inhomogeneous population distributions and has a manageable computational burden so long as the number of centroids is not very large (Kulldorff 1999).

There are two models for evaluating the noteworthiness of clusters: a Bernoulli model and a Poisson model (Kulldorff 1997). The Bernoulli model assumes that the data consist of individuals with and without disease. Our focus is on the Poisson model, which is appropriate for the analysis of aggregate disease data. In most real-world situations, the spatial distribution of cases will have a similar spatial distribution as the population. As a result, the Poisson model assumes a null hypothesis of constant risk rather than a null hypothesis of random distribution of cases. Under the null hypothesis that risk is constant over a study area, persons inside a cluster have a probability of disease equal to persons outside a cluster. If clusters of higher risk are of most interest, the alternative hypothesis is that persons inside a cluster have a higher probability of disease than persons outside the circular window. The following formula

$$\left(\frac{c}{E[c]}\right)^{c}\left(\frac{C-c}{C-E[c]}\right)^{C-c}$$

serves as the test statistic calculated for each circular window under the Poisson model. C is the total number of cases, c is the observed number of cases and E[c] is the expected number of cases within the current window under the null hypothesis of constant risk. For simple applications, E[c] is derived by multiplying the population within a circular window by the overall rate of disease. The ratio of alternative hypothesis to null hypothesis is proportional to this test statistic when c is larger than

E[c], and otherwise equal to 1 (Kulldorff 1997). To avoid numerical overflow problems, it is safer to calculate the log transformed test statistic

$$c(\log(c) - \log(E[c])) + (C - c)\log((C - c) - (C - E[c])).$$

A thorough description of the likelihood ratio test statistic and its statistical properties can be found in Kulldorff (1997).

Once the test statistic has been derived for all circular windows (potential clusters), a most-likely cluster is chosen based on the highest calculated test statistic of all potential clusters. The value associated with this most-likely cluster (called the maximum likelihood estimator) is the test statistic for the spatial scan. Under the null hypothesis of constant risk, significance can be determined through a Monte Carlo sampling of all possible permutations of how cases could be allocated in the population (Turnbull et al. 1990; Kulldorff and Nagarawalla 1995). Cases are simulated by randomly assigning 'disease' and 'non-disease' status to all persons in the population based on the actual number of true cases. For example, if there are 200 true cases in a population, these 200 cases are randomly re-assigned to the population such that each person has an equal chance of being a case. Then the spatial scan algorithm is run on this simulated data set, a most-likely cluster is found, and the associated largest likelihood ratio is saved. This process is repeated—for example, 999 times—with each simulated largest likelihood ratio saved. The maximum likelihood estimator for the real data is compared to the distribution of simulated maximum likelihood estimators; if the real maximum likelihood estimator is larger than most of the simulated maximum likelihood estimators (99.9% of them, for example), then the found cluster is considered significant.

There have been a number of recent developments since the spatial scan's original publication. Gangnon and Clayton (2001) observed that the spatial scan method preferentially detects clusters in areas that are nearer to each other in space. In most cases, this would involve areas of higher population density. They recommend the use of a penalty that mitigates this tendency. A more recent challenge to the spatial scan comes from the observation that the method is less able to detect clusters of non-circular shape (Patil and Taille 2003; Patil and Taille 2004). Some innovations have attempted to deal with this limitation. A recent addition to the literature uses ideas

found in evolutionary programming to efficiently search elliptical and circular cluster shapes (Conley, Gahegan and Macgill 2005). This method reports relatively fast solution times, and the ability to detect a large range of cluster shapes with high resolution point data. Duczmal and Assunçao (2004) use the results of a spatial scan of circular windows as a starting point for a search of connected but potentially irregularly shaped clusters. They add new and/or remove old centroids to an existing solution in order to further increase the maximum likelihood ratio using a simulated annealing algorithm to ensure that the search is not confined to local optima. Since there is no geometric constraint, the search strategy permits clusters of any shape.

Tango and Takahashi (2005) also build on the capability of the traditional spatial scan to identify clusters of 'flexible' shape. Rather than employing a complex heuristic, they enumerate all connected subgraphs within a distance from each centroid. Thus, for each centroid, their method scans all sets of connected subgraphs within a certain distance. This expansion allows them to observe more potential clusters without the complexity of tuning a heuristic to their data. However, in order to keep the problem manageable, they must constrain the search to clusters of relatively small size. For some applications, this may not be a serious drawback, since many clusters are relatively small. It is worth mentioning that although Tango and Takahashi (2005) found that the Duczmal and Assunçao (2004) approach has lower power then their method, their observation may not be broadly generalisable since the effectiveness of this and other heuristic approaches is dependent on the nature of the data, and the effectiveness of heuristic tuning.

The Duczmal and Assunçao (2004) and Tango and Takahashi (2005) approaches illustrate two general methods for dealing with large combinatorial problems. In the first case, the search method is intelligent—a clever algorithm designed to search a large space efficiently. In the second case, the scope of the problem is scaled down, and a complete enumeration is undertaken within the smaller search space. Unfortunately, it is difficult to assess which technique is more appropriate in the general case, since this will be dependent on the study area, and subtle details involved in the set up of the algorithms—in these particular cases, the heuristic

settings (Duczmal and Assunçao 2004), and the distance threshold (Tango and Takahashi 2005).

4.2.3 Structural and data-directed searches

The traditional spatial scan and many contemporary methods use a particular structural form—a circular window of varying size—to search for clusters. It is a *structural* search method since it imposes a geometric structure (circularity or isotropic compactness) on the problem. This is sensible because many diseases probably do manifest themselves in compact, roughly circular form. Evidence of the behaviour of environmental pollution, recent work in cellular automata (e.g., Batty 2005), and the first law of geography that "everything is related to everything else, but near things are more related" (Tobler 1970) all make it reasonable to expect clusters of disease to be approximately circular in shape. When a disease is particularly rare, the imposition of the correct structure could help find the location of a true disease cluster; the structural knowledge of disease may help improve power to detect true clusters. Prior knowledge about a disease's spatial properties—whether it exhibits a circular, elliptical or triangular shape, for example—could make up for excessive variance in the data.

Another option is to constrain the search process to adjacency. In a simple case, centroids are treated as adjacent when the polygons that enclose them share a boundary. This first-order adjacency can be extended by 'powering-up' an adjacency matrix such that second, third and additional orders of connectedness can also be used[1]. In fact, a geographically unconstrained cluster detection problem uses a fully powered-up adjacency matrix where all centroids are treated as neighbours of all other centroids. When using adjacency as a constraint, all neighbours usually receive equal topological weight—all are equally near—and another variable can be used to resolve the ties that occur. This permits opportunities to use other measures to 'motivate' the procedure, an idea sometimes referred to as a *data-directed* search, discussed by Patil and Taille (2003; 2004) and first implemented in disease cluster detection (as far as the authors are aware) by Duczmal and Assunçao (2004). Such

[1] Two step adjacency can be tabulated by squaring a simple adjacency matrix. Combined with the original adjacency matrix, this expands the definition of neighbour to include centroids that are one and two steps away from each other. Further power operations can extend the definition of adjacency further.

searches could maximise some attribute within the confines of a graph—such as trying to form subgraphs that have rates of disease that are as high as possible. Alternatively, proximity can simply be described by different variables—such as similarity of rate or some secondary attribute (like area, features of the landscape, or socioeconomic status).

A data-directed search with an explicit adjacency constraint allows centroids to be added to a subgraph to maximize an evaluation criterion while adjacency retains the geographic nature of the search. In a similar problem, political districting, all polygons (such as census tracts or other bounded shapes) in a study region must be classified into a predetermined number of larger polygons that have certain desirable characteristics—such as similar populations, homogenous demographic characteristics or relatively compact shape. The objective (for example, minimizing population variance) is met through a process in which steps in the algorithm often interact with one another. When a small polygon is exchanged between two larger polygons, this may simultaneously improve one part of the problem (bringing its population towards the mean), and negatively affect another part of the problem (pulling its population away from the mean). This can make finding global optima (in this case, the smallest possible variance in population among districts) difficult to find. Bozkaya, Erkut and Laporte (2003) developed a method to find optimal and near optimal solutions to political districting problem relatively efficiently. They use a tabu search heuristic to ensure that their algorithm escapes poor locally optimal solutions. Simulated annealing approaches have been used in a similar fashion in automated zoning research (Openshaw and Rao 1995). When a single most-likely cluster is of most interest, the search process is similar to a political districting problem with only two sets (a cluster and non-cluster set). The interaction between these two sets is straightforward—any change that improves the local properties of a potential cluster is very likely (if not guaranteed) to benefit the global objective—finding the cluster with the highest evaluation criterion (the likelihood ratio test in the case of the spatial scan). The relative simplicity of finding a single most-likely cluster suggests that simple methods are likely to find very good solutions.

We propose a greedy approach to the problem of finding disease clusters of irregular shape. We refer to this search approach as the *hot-graph* method. This approach is data-directed, and actively searches for clusters in a similar fashion to Duczmal and Assunçao (2004). However, unlike the Duczmal and Assunçao (2004) method, our approach searches from all centroids, like the method proposed by Tango and Takahashi (2005). In this way, our technique is a new hybrid of these two ideas. However, unlike these methods, (but like the original spatial scan), our technique requires no sophisticated prior decision making—such as setting a maximum cluster size, or setting heuristic parameters. The method we propose does not assume a single neighbour structure, but obtains it from topological information defined by the specific problem. Using simulated data, we present and compare the performance of this method to the performance of the spatial scan. In addition to offering a new method for detecting disease clusters of any shape, our findings should offer a fair appraisal of the potential of a wide range of data-directed search techniques. We will discuss the strengths and weaknesses of this proposed approach, and suggest opportunities for its use in conjunction with the circular spatial scan.

4.3 Methods

4.3.1 A greedy hot-graph search

Greedy approaches to problem solving work by choosing the best currently available step in a decision process. The approach is both straightforward, and short-sighted; for many problems, seemingly good short-term decisions often come with worse long-term consequences. Good chess players are often willing to sacrifice a piece early in the game in order to arrive at a stronger strategic position later in the game. When presented with an opportunity to take an intentionally sacrificed piece (called a 'gambit' at early stages of a chess game), a greedy response is to take the piece. Although this may meet the short term goal of taking pieces from an opponent, this can often result in a significantly weaker position in the long term. In this case, the sacrifice of material expresses a willingness to lose something (a chess piece) in exchange for a positional advantage that will bring one closer to victory at a later time. The greedy response is to capitalize on the immediate opportunity, and to take

the piece at the possible expense of the global objective—i.e., winning the chess game.

Interestingly, some problems can be solved successfully through a greedy approach. A minimum spanning tree is the smallest distance-cost set of edges that connect all vertices in a graph. One of the methods for finding a minimum spanning tree, the Prim/Dijkstra algorithm (Prim 1957), uses a greedy approach. Defining any node in the graph as the starting subgraph g, the node with the lowest-cost edge connecting to g is added to g. This process continues until all nodes have been connected to g. Subgraph g is guaranteed to be a minimum spanning tree (Winston and Venkatarmanan 2003).

The algorithmic structure of the hot-graph approach comes from the Prim/Dijkstra algorithm to solve minimum spanning trees. As with the spatial scan, we treat each centroid in the graph as a starting point of a unique search. The hot-graph method adopts the same statistical theory as the likelihood ratio test used in the spatial scan. However, instead of adding centroids in sequence of nearest to farthest neighbour, the hot-graph method sequentially adds the centroids that results in the highest likelihood ratio at a particular step in the search process. Figure 4.4 demonstrates the hot-graph procedure in a very simple graph with only six centroids. Thin lines indicate adjacency, and thick lines indicate which centroids have been added to the current subgraph. Centroids with edges connected to any centroid within the current subgraph are feasible. At each stage, the feasible centroid that results in the highest subgraph objective (for simplicity, the rate of disease in this example) is added to the current subgraph and then stored. At each step, the algorithm adds a centroid until the new graph is fully connected (or until a population or other pre-set threshold is met). In this example, the last three centroids added cause the rate to drop. It is important to note that the algorithm is *forced* to add the best currently feasible centroid, even though newly added centroids may actually result in new subgraphs with lower objectives (in this case, the rate). The same search process is conducted from each centroid in the graph. For the hot-graph method, the centroid which maximizes the likelihood function is added (rather than the centroid that maximizes

the rate as presented in this example). The subgraph with the highest likelihood function of all the stored subgraphs is considered the most-likely cluster.

Adjacency information determines the feasibility of search moves, and can be defined any number of ways. In the typical case, neighbours are defined as immediate neighbours, or within a specific distance threshold. By squaring, cubing or performing higher-order power operations on an adjacency matrix, this can expand the definition of adjacency to include nodes that are normally two, three or more steps away from each other as neighbours as well. This would allow the algorithm to leap over nearer neighbours.

The statistical significance of a hot-graph identified cluster is similar to that of the spatial scan; simulated data are generated by randomly re-assigning case/non-case status to individuals in the study population. Unfortunately, this process is more time consuming for the hot-graph approach since the search process is influenced by the location of cases; since the location of cases changes with each simulation, the whole search process must be run on each simulated data set. This search process is more time consuming than that of the spatial scan, which has a fixed topology, regardless of the location of cases.

Pseudo-code for the hot-graph search algorithm can be found in Appendix I.

4.3.2 The experiment

The purpose of our experiment is to evaluate the effectiveness of the spatial scan and hot-graph methods by a) comparing their respective abilities to detect the presence and approximate location of simulated clusters of different shapes (circle, line, ring and network), and b) comparing their abilities to define the precise boundaries of clusters of regular (circle) and irregular (line, ring, network) shapes. The experiment is conducted on simulated data. These simulated data are a simplified representation of disease patterns in the real world, but give a suitable baseline for understanding the general behaviour of these two methods. The algorithms for solving and testing the hot-graph and spatial scan methods were programmed in the SAS^{TM} language (SAS Institute, 1999).

We use an approximately square tessellation of 81 hexagons (or 'zones') to derive a graph of our study region (Figure 4.5).[1] The use of an hexagonal tessellation makes adjacency easy to calculate since the queen's and rook's case adjacency rules are identical. Any two zones that share one or more line segment are considered adjacent. In order to ensure that the simulated clusters are reasonably symmetrical and in the centre of the tessellation, odd numbers of row/columns were required. We chose a 9 x 9 tessellation since it resulted in a number close to the number of zones used in Chapter 5. The geometric centres of the zones are used in distance calculations for the spatial scan, and the edge set is defined by immediate neighbours—hexagons that share line segments. For all simulations, each zone receives a population count. If all zones were assigned small populations, the between-zone variability in disease counts would be larger than if all the zones were assigned large populations. In order to observe how the methods behave with different underlying populations, we increment the base population of the scenarios from 1 000 people to 10 000 people in 1 000 person steps. In order to further increase the realism of the simulation scenarios, we introduce random variation into population counts rather than assigning each area the exact same population. Each zone receives a population derived from a normal distribution number generating function with a mean equal to the base population (for the first step, this is equal to 1 000), and a standard deviation equal to $1/10^{th}$ the base population. Though the population variations are somewhat arbitrary, they better reflect the heterogeneous population characteristics found in the real world than uniform populations across all zones.

We examine four cluster patterns: circle, ring, line and network (Figure 4.6). In each experiment, the cluster pattern is located in the middle of the tessellation. We do this in order to neutralize edge effects as much as possible. Though edge effects are an important concern in most spatial problems, we do not consider them here for the sake of parsimony. In all our experiments we simulate the number of cases in each zone in the same manner. This is done by assigning an equal risk of disease to all 'individuals' in a zone, and then using a random number generating function to

[1] In the context of our experiment, we use the term "zone" in place of "centroid". We prefer this term here for descriptive purposes as the maps we present later in the chapter are of polygons rather than points.

determine whether or not a person is a case based on this risk level. The risk of disease is assigned based on whether or not a person is inside or outside a cluster pattern (or more precisely, whether or not the zone to which they are assigned is inside or outside the cluster pattern). In all our experiments, individuals inside a cluster pattern are assigned a higher risk of incurring disease than persons outside a cluster pattern. Though the cluster analyses treat zones as the units of analysis, the 'individuals' within them form the sampling framework for the simulation of disease rates.

For each experiment, the hot-graph and spatial scan methods both search for a most-likely cluster of disease. We evaluate the methods in three ways.

4.3.2.1 Power estimation

In order to estimate the power of both methods to detect a cluster, our first experiment observes the number of times each method identifies at least part of a cluster pattern as significant. For these experiments, persons in zones that are part of a cluster pattern have a risk of disease equal to 0.005 and persons in zones that are not part of a cluster pattern have a risk of disease equal to 0.0025. For both methods significance is determined by the Monte Carlo simulation process described above. If a detected cluster is significant at the 0.001 level and at least one zone from the cluster pattern is part of the detected cluster, the method is considered successful. We do this for each population increment (from 1 000 to 10 000) in order to observe changes with increasing population. For each population increment, and for each of the four cluster shapes, the experiment is repeated 100 times, and successfully detected cluster patterns are tallied for each iteration. The proportion of successful runs for both methods and for all four cluster patterns are presented graphically. A polynomial regression line is fitted to these graphs to help visualize apparent patterns.

4.3.2.2 Geographic precision: graphs

We also assess the ability of both methods to find clusters that correctly identify whether zones are part of cluster patterns or part of their complement. We classify each zone within a solution found by a cluster algorithm into one of four potential categories: true positive, true negative, false positive and false negative. When a zone is correctly identified as part of a cluster pattern, we call this a true positive. A true

negative is a zone which is correctly identified as part of the complement of a cluster pattern. False positives are zones that identified as part of a cluster pattern but are actually part of the complement of a cluster pattern. False negatives are zones which are identified as part of the complement of a cluster pattern that are in fact part of a cluster pattern.

We generate scenarios for all four cluster patterns & population increments from 1 000 to 10 000. In addition, we run these scenarios for 7 different inside-cluster/outside- probabilities of disease risk: 0.003:0.0025, 0.0035:0.0025, 0.004:0.0025, 0.0045:0.0025, 0.005:0.0025, 0.0055:0.0025 and 0.006:0.0025. In total, 28 000 scenarios are generated (4 cluster patterns * 10 population increments * 7 ratios of disease risk * 100 iterations). Both methods are tested on the same scenario once. For both methods, a tally of false positives, true positives, false negatives and true negatives are kept for each scenario. From these tallies, sensitivity and positive predictive values are calculated. Sensitivity equals the number of zones correctly identified as part of a cluster (true positives) divided by the total number zones in the cluster pattern (true positives + false negatives). Positive predictive value equals the number of zones correctly identified as part of a cluster (true positives) divided by the total number of zones detected as part of a cluster (true positives + false positives). Mean sensitivity and mean positive predictive values are calculated over the 100 iterations. Results from the 0.005:0.0025 ratio of disease probability are presented graphically. A polynomial regression line is fitted to these graphs in order to help visualize apparent patterns. The results from all inside-cluster/outside-cluster ratios of disease probability are presented in Appendix II.

4.3.2.3 Geographic precision: maps

We map the proportion of true positives and false positives for the 0.005:0.0025 inside-cluster/outside-cluster pairings generated in the section above. The number of times each zone is identified as part a cluster is divided by 100 (since there are 100 iterations) and this proportion is reported on the map. For zones that are part of a cluster pattern, this is the proportion of true positives. For zones that are part of the complement of the cluster pattern, this is the proportion of false positives. These proportions are represented on the map and colour coded for easier visualization. A

perfectly performing method will have values of 1.00 for zones within a cluster pattern, and values of 0.00 for zones outside a cluster pattern.

4.4 Results

4.4.1 Detection of significant clusters

The findings reported in Figures 4.7a-4.7d report the proportion of significant true clusters found over the 10 population increments and the 4 different cluster patterns. When population is low, and disease frequencies are unstable, the spatial scan appears better at identifying a cluster as significant. This is true for all four cluster patterns. As population increases, the difference between the methods appears to shrink, though the shape of the curve varies somewhat across the different cluster patterns. For the network and ring shaped clusters, the estimates of power appear relatively similar for both methods. For the circular clusters, the difference between the methods appears fairly large. For the line-shaped cluster, the spatial scan appears to perform better, though compared to the other cluster shapes, both methods do comparatively worse.

4.4.2 Sensitivity and Positive Predictive Value

For the ring and circular clusters the mean sensitivity of the spatial scan is quite high once population levels reach 4-5 000 (Figure 4.8a and 4.8b). The mean sensitivity of the hot-graph method appears lower, but similar for both circle and ring clusters. For the network and line clusters, the spatial scan method has considerably lower mean sensitivity than the hot-graph method (Figure 4.8c and 4.8d). For network and line clusters patterns, the mean sensitivity of the spatial scan does not appear to improve with increasing population size, and is at or near 0.5 for all population levels. The hot-graph method exhibits roughly the same pattern of mean sensitivity for all four types of simulated clusters, approaching reasonably high levels of mean sensitivity at population levels around 10 000.

Positive predictive value takes into account the ability of a method to find true cluster zones without also capturing false positive zones. When clusters are circular, the spatial scan has a higher mean positive predictive value than the hot-graph method (Figure 4.9a). With the exception of the 1 000 and 2 000 population levels, the method has mean positive predictive values above 0.5, and at all population levels

this method appears superior to the hot-graph method. The hot-graph method appears to often include false positives within the identified cluster; the mean positive predictive value does not appear to approach the results of the spatial scan even at population levels of 10 000. For ring and network cluster patterns, the mean positive predictive value for the spatial scan and hot-graph methods is lower, but the hot-graph method appears to have slightly higher mean positive predictive values when population levels are high (Figure 4.9b, 4.9c). For line clusters, the spatial scan has higher mean positive predictive values than the hot-graph method, though both methods have relatively low mean positive predictive values for all but the largest population levels (Figure 4.9d).

4.4.3 Maps of true positives and false positives

Maps of the hexagon tessellation illustrate the spatial variation of true positives and false positives for both detection methods (Figure 4.10a-d). For zones included in a cluster pattern (refer to figure 6 as a reference) the values in black font represent the proportion of false positives for a particular zone. For the non-cluster zones, values in green font indicate the proportion of false positives. Zones are shaded into equal interval classes of 0.20 units in size to help illustrate the patterns indicated by the numbers.

Consistent with observations made above, the spatial scan is very effective at identifying zones which are part of a circular cluster pattern (Figure 4.10a). Even at low population levels (where statistical variability is high) the method has relatively a large proportion of detected true positives. Proportions of false positives are higher in zones closest to the cluster pattern, and fall off to low levels near the edges of the tessellation. The hot-graph method illustrates a similar pattern, although the proportions of true and false positives are higher. Not surprisingly, as population increases, both methods perform better in both respects, though the spatial scan more so. At population levels of 10 000, the proportion of true positives detected by the spatial scan's reaches 0.90 in all seven zones inside the cluster pattern. At the same time, the proportion of false positives is at or below 0.0 for almost all zones outside the cluster pattern. The trend is similar for the hot-graph method, though again, the magnitudes are less promising. Most importantly, even at population levels at and

above 10 000, the proportion of false positives should be of concern; many areas continue to be falsely identified as part of a cluster.

The topology of the ring cluster differs from the other simulated clusters since the cluster divides the non-cluster areas into two discontiguous groups—one inside the ring, and one outside the ring (Figure 4.10b). The spatial scan method has a tendency to falsely identify zones inside the ring as part of a cluster. The spatial scan has very high proportion of false positives for zones inside the ring, even when the population level is high. Indeed, the proportion of false positives (for zones inside the ring) and true positives are similar, confirming the intuitive; that the spatial scan method treats ring shaped clusters as circles. The hot-graph method offers greater precision, though as observed, still has a tendency to detect a large proportion of false positives. Still, as the population increases, the method is increasingly able to separate true from false cluster zones, and at the largest population levels, exhibits lower proportions of false positives.

Although able to detect parts of a line cluster pattern, the spatial scan cannot detect clusters encompassing the whole line (Figure 4.10c). The hot-graph method, on the other hand, is able to define the boundaries of the line cluster once the population level is large enough. By the time population level reaches 5 000, most proportions of true positives exceed 0.90, though the proportion of false positives remains over 0.10 for many zones.

The network cluster has more zones than the three other cluster patterns (Figure 4.10d). The spatial scan has particularly high proportions of false positives for this cluster type, particularly in the zones adjacent to the cluster pattern. Even for large population levels, some proportions of false positives are over 0.40. The effect is similar to that of the ring and line; the scanning window increases in size to incorporate a larger number of cluster zones, but in the process, includes several non-cluster zones as well. The hot-graph method has relatively higher proportions of true positives, and lower proportions of false positives as observed for the line and ring clusters. The proportions of true positives are highest in zones near the centre of the cluster network, and decrease with distance from this point.

4.5 Discussion

4.5.1 General findings

Based on these results, the spatial scan seems more likely to detect the approximate location of a cluster, regardless of its true shape, than the hot-graph method. However, the spatial scan approach worked particularly well for circular cluster patterns, even at low population levels. In this sense, imposing the correct circular geometric constraint on the search process seems to make up for variability in the underlying data. If all disease clusters could be classified into a finite number of discrete geometric shapes, then the spatial scan could be expanded to include these shapes in the search process. The method could then maximize the likelihood ratio test over all shapes of all sizes, and return well defined, and precise cluster information. Recent work has taken on this challenge (e.g. Kulldorff et al. 2006) by incorporating ellipses within spatial scan search. This increases the range of shapes that the spatial scan is able to detect with precision similar to the circular spatial scan.

The traditional spatial scan is less able to define the precise geometry of linear, ring and network clusters. This is not surprising given the structure of the search process. In order to encompass an entire line, the circular window would need a diameter equal to the cluster length. Unfortunately, this would include a large number of non-cluster zones. On average, the method is likely to identify zones in the middle of the line, and ignore the more peripheral zones. This is illustrated on our maps of proportions of true and false positives. This observation is somewhat dependent on the size of the cluster pattern relative to the whole study area, however. If our tessellation were larger, the spatial scan would be more sensitive to detecting an entire line or network cluster, though the number of false positives would certainly increase as well. The spatial scan is sensitive enough to identify the approximate location of ring-shaped clusters, but is guaranteed to identify a large number of false positives in the middle of the cluster area (the centre of the ring). Whether or not this represents a genuine failing of the spatial scan is dependent on whether or not such structures exist in the real world, and to what degree detecting them may be important.

The hot-graph method exhibits fairly similar performance for all simulated cluster shapes. The mean sensitivity curves express relatively similar trends with increasing

population levels. The mean positive predictive value curves vary somewhat depending on the shape—for line and circular cluster patterns, the mean positive predicted value is quite low at all population levels. But for the most part, apparent differences between the hot-graph method and the spatial scan seem to be accounted for by the varying performance of the spatial scan. The hot-graph method does not concern itself with structure, and should behave in a roughly similar manner, regardless of shape. Some variation can be accounted for by the fact that the four simulated clusters are of different relative size. For larger cluster patterns (like the ring and network) the hot-graph method has larger mean positive predictive values. This makes sense; as the cluster pattern gets larger, its complement shrinks, and as a result, there are fewer zones to identify as false positive.

4.5.2 Operational issues related to the hot-graph method

The hot-graph method has two notable limitations independent of its performance with respect to the traditional spatial scan. The first is concerned with the greedy structure of the search process. The method prefers immediately 'best' moves at the cost of ignoring immediately bad moves that may lead to globally better solutions (in particular, finding the true most-likely cluster). One might envision this happening when a zone of low disease rate separates two zones with high disease rates. It could be argued that a greedy search is a poor approach because it is unlikely to connect these two high-rate zones as part of a single cluster. We argue, however, that this may not be likely to occur very often. Since the method sequentially searches from every zone, at some point a search is certain be spawned from this low-frequency zone, and is likely to link the two neighbouring zones into part of a single cluster. Nonetheless, to allay this concern, the adjacency matrix can be redefined to ensure that the algorithm will 'hop' over low-rate zones to connect zones that have high rates. Of course, this would create clusters that are not contiguous, and one may argue that this somewhat defeats the purpose of spatial cluster detection in the first place. Under these circumstance, the low disease rate zones may represent a true separation between two clusters of different origin and different value. As such, it may be important that the areas are not compressed into a single most-likely cluster, and are detected as separate clusters.

A second limitation of the method is related to its computational complexity. Patil and Taille (2004) suggest that that the primary limitation of data-directed searches is not in their ability to identify clusters efficiently, but their need to re-build the topology for significance testing. For the spatial scan, the search is based on geographic distance between centroids, which is fixed for all data at these locations, and for the Monte Carlo simulations within a specific problem. A well designed spatial scan algorithm can use key indexing to quickly assign simulated data to the correct location in a topology file, and efficiently extract simulated maximum likelihoods. The hot-graph method uses a search method based on the arrangement of the attributes of the zones, and therefore, needs to re-calculate topology for each random iteration. This can be a major time burden.

For routine cluster detection tasks, the most important requirement is that the time to solve a problem is as fast or faster than the rate data are acquired. In jurisdictions with the capacity for large-scale routine disease surveillance, times to acquire, digitize and upload data may take days, weeks and months. A surveillance system that takes 10 hours to find a solution does not represent an insurmountable challenge to the system, especially since most of the time is in the Monte Carlo simulations, the number of which can be reduced in cases of an anticipated emergency (from 999 simulations to 99, for example). Simple alterations—like abandoning searches that have a low likelihood ratio after a certain number of moves—may help the hot-graph method operate efficiently for larger problems. Zones that are near each other are likely to spawn similar (and often identical) hot-graphs. The algorithm can also be sped up by ignoring searches spawned from certain zones—for example, those with a very low rates—with little risk of missing interesting clusters.

A final limitation worth mentioning is that the adjacency constraints could sometimes limit the ability of the hot-graph approach to find clusters in certain regions of a real world study area. Based on the algorithm as presented, zones with fewer neighbours will be less frequently included in a search. This may lead to slight under-inclusion of zones along edges, corners, or in more isolated areas of a tessellation (such as peninsulas). There may be ways to estimate this problem by summarizing information on connectedness, or classifying the structure of the

adjacency matrix. This information could then be used to weight the search process—similar to an approach discussed by Gangnon and Clayton (2001). However, since the hot-graph method is fairly exhaustive—each zone is used as a starting point of a sequence of searches—this problem may not occur except in cases of very complex topology.

4.5.3 Comparison of the methods

The elegance of the original spatial scan method—its speed, flexibility to detect clusters in time and space, and simplicity—make it a good all-around choice for cluster detection, in spite of some weaknesses. Precise identification of a cluster's shape may not be of critical importance when monitoring for infectious disease outbreaks or environmental hazards, for example. Preliminary exploratory information about clusters near sources of pollutants may only require a general indication of the location of high disease rates. Furthermore, its ability to detect more than one cluster may make up for its inability to precisely define the shape of a most-likely cluster. A linear chain of detected secondary and tertiary clusters could provide good indirect evidence of a linear-shaped cluster of disease. However, in some instances, more precise definitions of cluster areas may be important. Resource allocation based on information obtained from cluster detection sometimes requires precise definitions of cluster areas. Knowing where an outbreak of disease is occurring may be key to applying the proper interventions. Compared to the spatial scan, the hot-graph method is better able to detect the geometry of an irregular cluster pattern, though not without costs of its own. For one, at smaller population levels, this method has a tendency to include false positive zones in the detected cluster. This may direct attention to zones that do not require it, and reduce the effectiveness and efficiency of interventions informed by this method. The high proportion of false positives found in some zones tell us that the hot-graph method tends to find clusters of irregular shapes that in some sense 'over-represent' true patterns of disease, irrespective of inferential questions about clustering. Though this is not surprising, the illustration here is especially compelling since we know it is not based on *ad hoc* decisions about the search process, or heuristic tuning, and we should expect similar issues to arise for other data-directed searches.

Fortunately, the weaknesses of the spatial scan and the hot-graph approach are to some degree countervailing. When clusters are of a non-circular shape, the circular spatial scan may miss some areas, and may under-predict the size of the actual cluster. On the other hand, the hot-graph method is more likely to over-predict the number of zones in a cluster. When a disease is rare the hot-graph approach may be especially likely to identify irregularly shaped structures that may include several zones that are not part of a true disease cluster. Therefore, it may be sensible to combine these two methods into a single analytical process.

The key to interpreting such a combined analysis is to understand that the two methods are communicating different information. The traditional spatial scan imposes a single geometric structure and then searches for it; the hot-graph approach explicitly searches for clusters of high disease rates subject to adjacency constraints. The former may be of more importance for certain types of decision making since it meets traditional standards for statistical inductive inference—either identifying or failing to identify an anomalous cluster of observations. The latter informs us more generally about the disease as a spatial system. In this sense, the hot-graph approach tells us which geometry, and in the real world, geography, of disease stands out. This can be very informative when used in conjunction with the spatial scan. When the hot-graph approach finds a circular-shaped cluster, for example, it means that in spite of a large number of possible geometric shapes, a circular pattern of high disease incidence exists in the data. This adds supporting information to the spatial scan; not only is the cluster significant, the geometric/geographic circularity constraint seems to be sensible. On the other hand, when a cluster found by the hot-graph method is not circular, this may indicate that there is a non-circular geographic structure of disease. This has to be interpreted with caution, since the non-circular shape may simply be an accident of over-fitting. Nonetheless, since the methods provide different information, we argue that it may not be fruitful to consider one approach as a replacement of the other.

4.5.4 Study Limitations

There are a few limitations in our experiments that are worthy of note. First, we did not consider the impact of edge effects. Zones along the edges are guaranteed to

have fewer neighbours on average. We expect that the hot-graph method may have a slight bias against identifying zones along the boundary of a study area as part of a cluster. To some degree this problem can be mitigated by making changes to the adjacency information; zones along edges could be linked to two step neighbours to ensure they are more likely to be included. This could easily introduce a bias in favour of edge zones if not done with care. Since the spatial scan does not rely on a topology based on adjacency, it may be less affected by the edges in the search process. Future work should consider testing the performance of both methods with respect to clusters along the edges of study areas.

Second, the experiments we used were fairly restrictive, particularly in the experiments used to estimate differences in statistical power. We did not vary the size of the simulated clusters, or the difference in size between the study area and the cluster patterns. Further, we classified all cluster scenarios into two groups (inside and outside cluster), ignoring that other local variations are likely to exist in the real world. This generous assumption of internal and external homogeneity probably makes both methods appear more effective than they would be in the real world. Similarly, we ignored the role of secondary, tertiary and additional clusters on the process of detection. These assumptions limit the relevance of our findings in the real world, though this is not (in general) uncommon in simulated experiments. In the real world, patterns of disease variation usually too complex and too varied to perfectly duplicate in a simulated setting, and therefore all generalizations from the simulated to the real world must be undertaken with caution, particularly when the simulation is small in scope.

Finally, our observations about our experimental results are not formally evaluated—for example by hypothesis testing, structured classification or other procedures. Instead, we assess the performance of the hot-graph and spatial scan methods by observing tendencies on graphs and maps. From a traditional scientific perspective, this limits our ability to claim 'true' differences between methods. Where we point out differences between the two methods, we do so based on strongly apparent trends and common sense. As such, we refrain, as much as possible, from using language that implies certainty, and qualify our observations with phrases like

'appear to be different' rather than 'are different'. Though this may not satisfy all measures of scientific reasoning, we would point out that hypothetical-statistical models are not without their own philosophical shortcomings.

4.6 Conclusion

Based on the results of our experiments, we have observed differences in performance between the spatial scan and the hot-graph approaches to cluster detection. We believe that our observations are consistent with general observations about structured and data-directed searches. Structured searches work well when the structure is realised in the data; methods that search circular windows are very likely to find clusters if they exist, and accurately define the boundaries of clusters when they are circular in shape. Data-directed searches do not consider geometry explicitly, and are able to identify clusters of any shape. The hot-graph method had lower power, but was better able to describe the shape of irregular clusters.

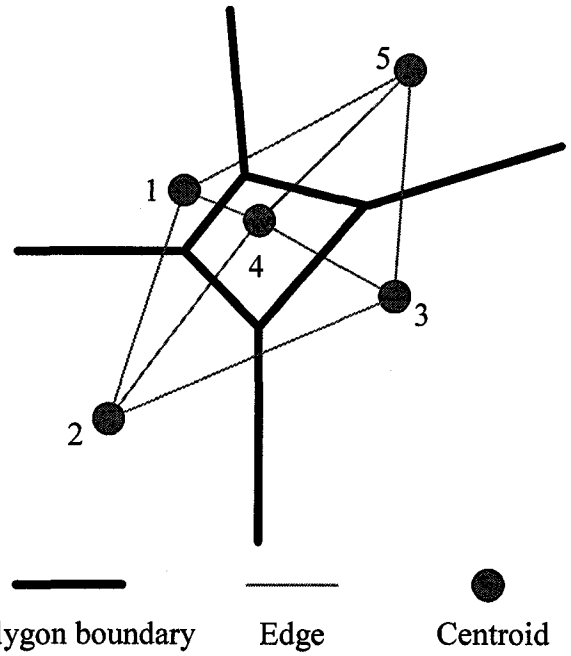
It is doubtful that data-directed methods will ever replace the elegance or functionality of the traditional spatial scan. If Tobler is right, then most clusters of disease are likely to be approximately circular in the real world. Furthermore, the spatial scan is simple to apply, efficient, and unlikely to identify clusters with a large number of false positive zones. However, there seem to be instances where a cluster detection system would benefit from (and would not be harmed by) data-directed searches. Depending on the size of the search space, heuristic modifications may be required to reduce computing time. Alternatively, the hot-graph method could be applied as an exploratory back-up; when results of the two methods are similar, it provides validating evidence of the presence and shape of clusters. When a data-directed method finds a circular cluster in the same location as the traditional spatial scan, there is especially good reason to think that the data exhibit an anomalous cluster of disease of a circular form.

This study also gives evidence of the upper limit of data-directed methods for disease cluster detection. Faster methods exist—such as that proposed by Duczmal and Assunçao 2004—but generalised comparisons of this method and the spatial scan are difficult since the performance is at least partly dependent on the data and the settings of the heuristic. The hot-graph method is slower, but still likely to find the

most-likely cluster, regardless of shape, and without heuristic tuning. Our results suggest that when (and if) clusters have irregular shapes, the spatial scan may not be as successful as the hot-graph method at defining the shapes of a true cluster. As a result, the applicability of either of these methods will depend on the expected shape of a disease cluster, as well as the consequences of finding false positives. If false positives are not of concern, the hot-graph method can be trusted to identify clusters fairly successfully. If false positives are of concern, and only general information is required about a cluster's location, the spatial scan is a preferred approach.

Although tests on simulated data allow us to offer general observations about how these two methods perform, they give little indication of how the methods would differ in the real world. Given the various distributions of risk factors in space—like plumes of pollution, the distribution of income in an urban areas, and the complex variations of geological and weather systems—future work should examine the performance of these techniques with real data. Though the conclusions of such analysis may not build our general understanding of cluster detection, it would offer important information of what we may expect to see in the real world.

Figure 4.1 A graph representation of a corresponding polygon tessellation



Vertex set= {1,2,3,4,5}
Edge set={{1,2},{1,4},{1,5},{2,4},{2,3},{3,4},{3,5},{4,5}}

Figure 4.2 A conceptual illustration of the spatial scan

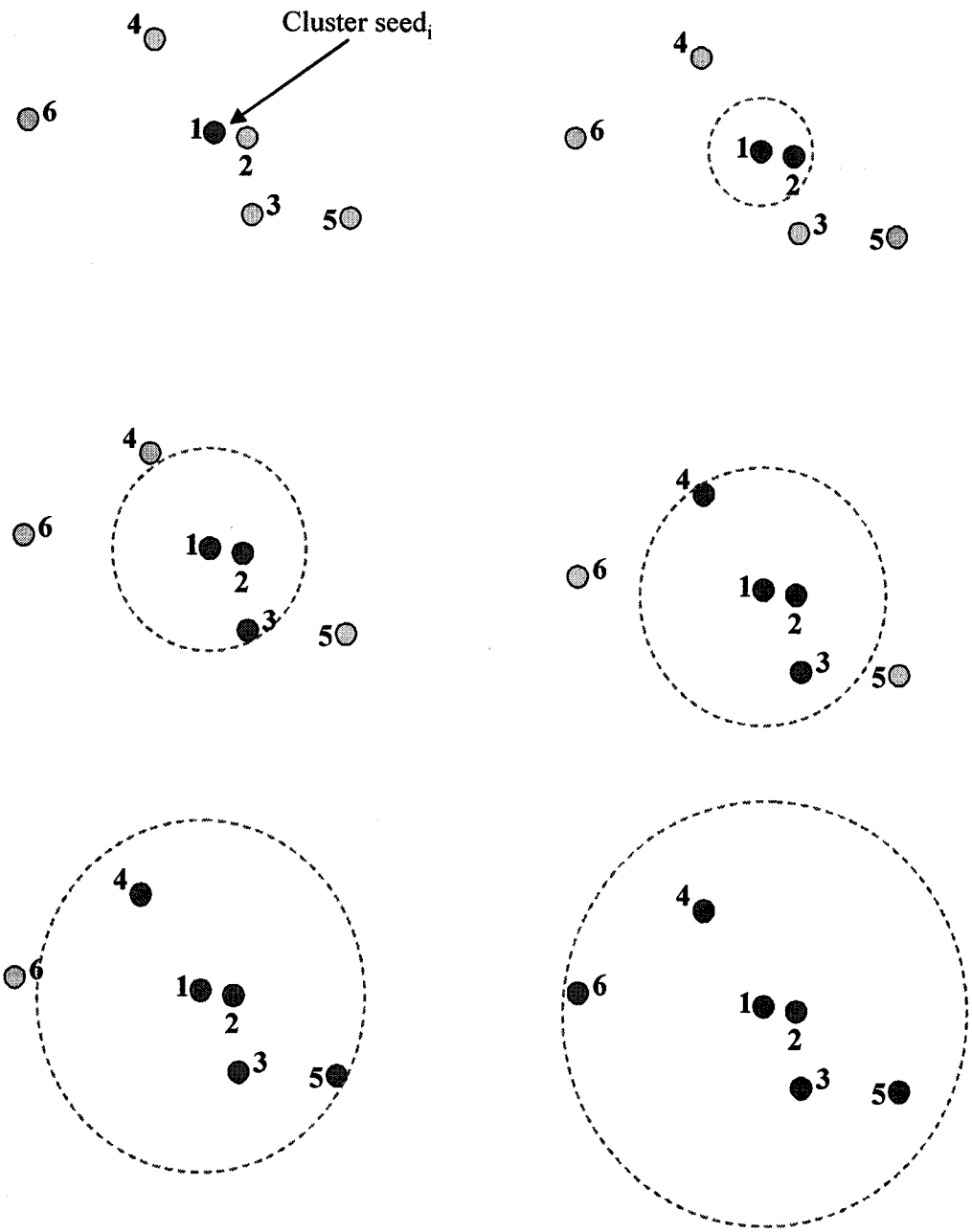


Figure 4.3 An operational illustration of the spatial scan

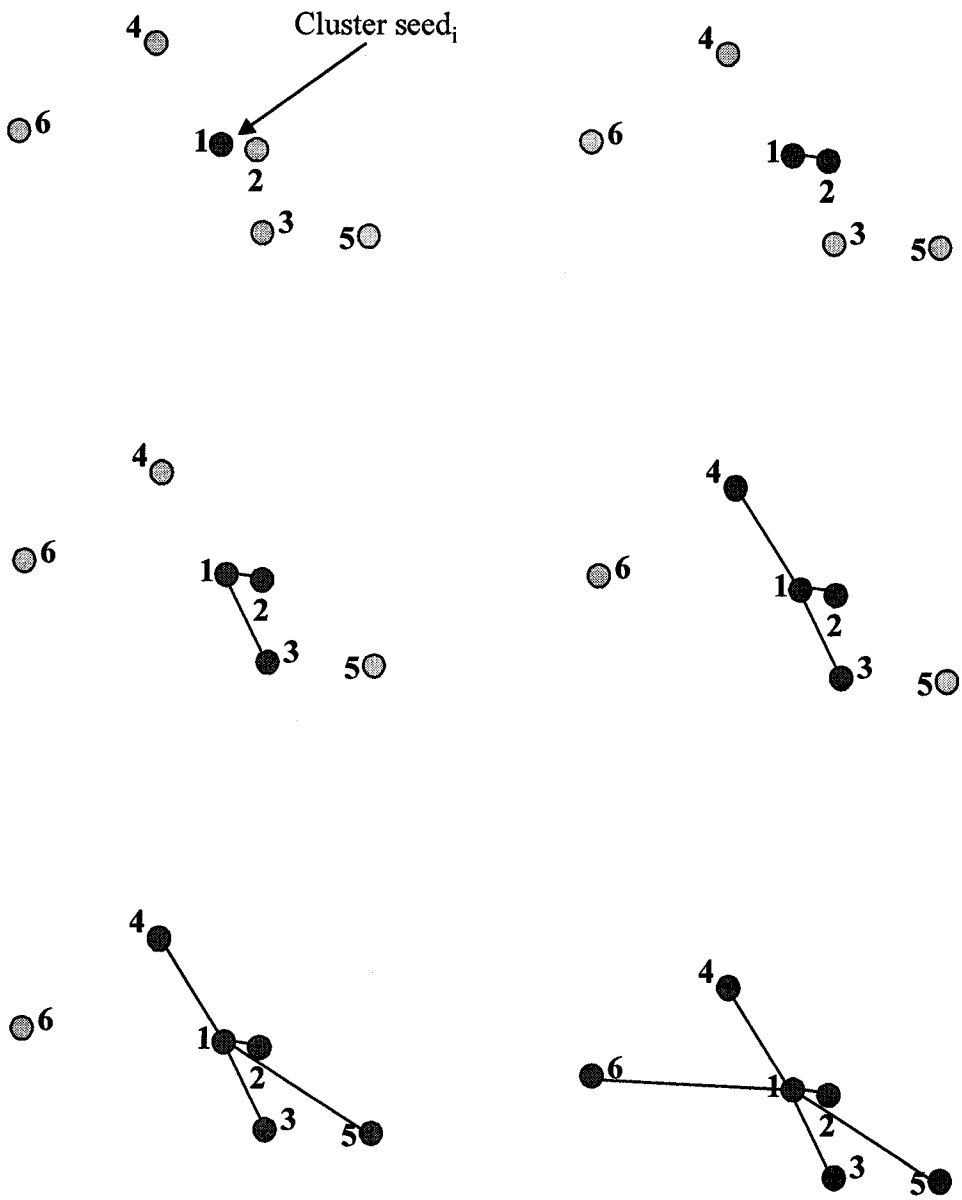


Figure 4.4 An illustration of the hot-graph method

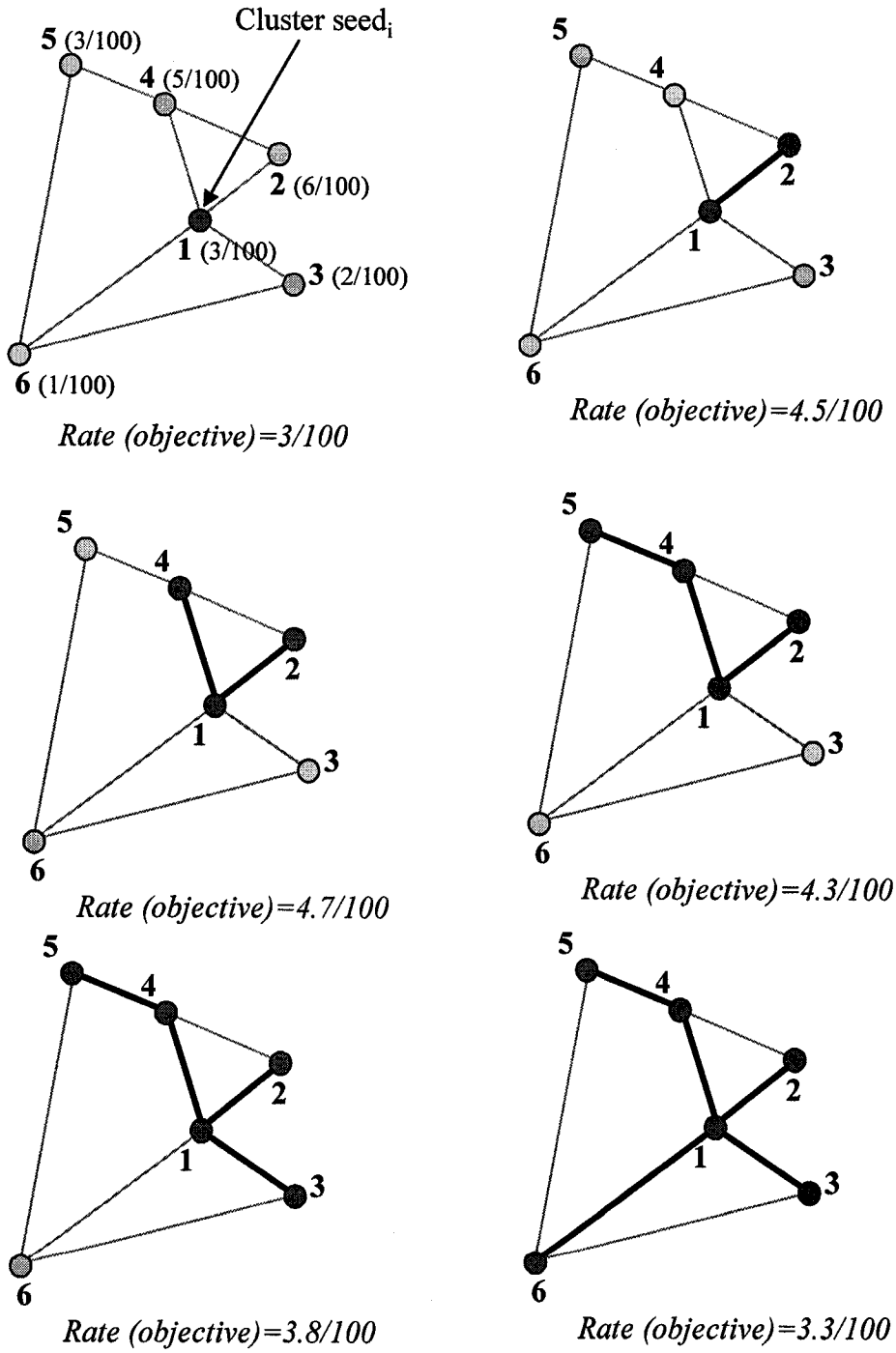


Figure 4.5 The simulated study area

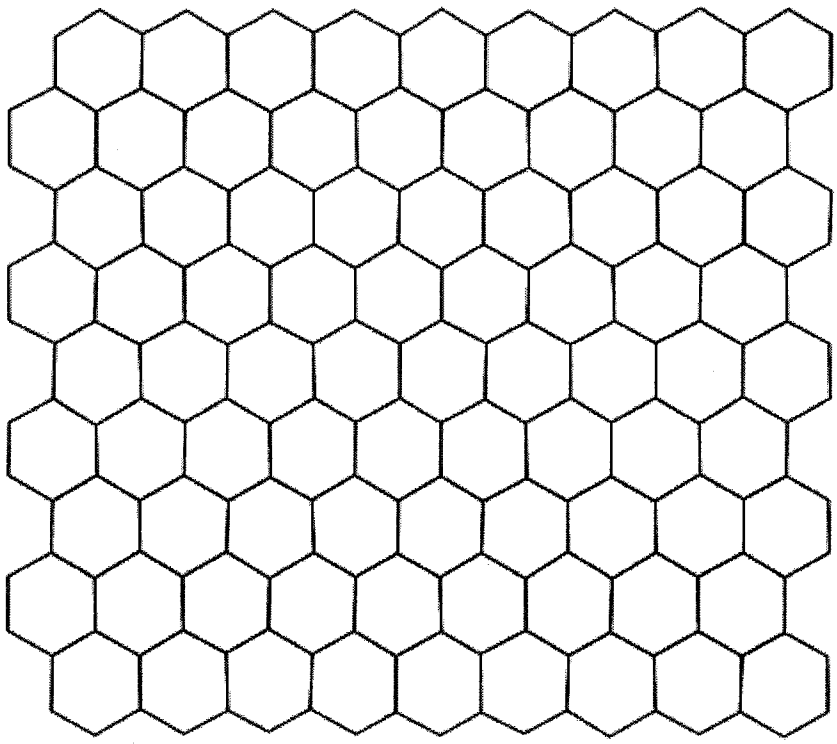


Figure 4.6 The four cluster patterns

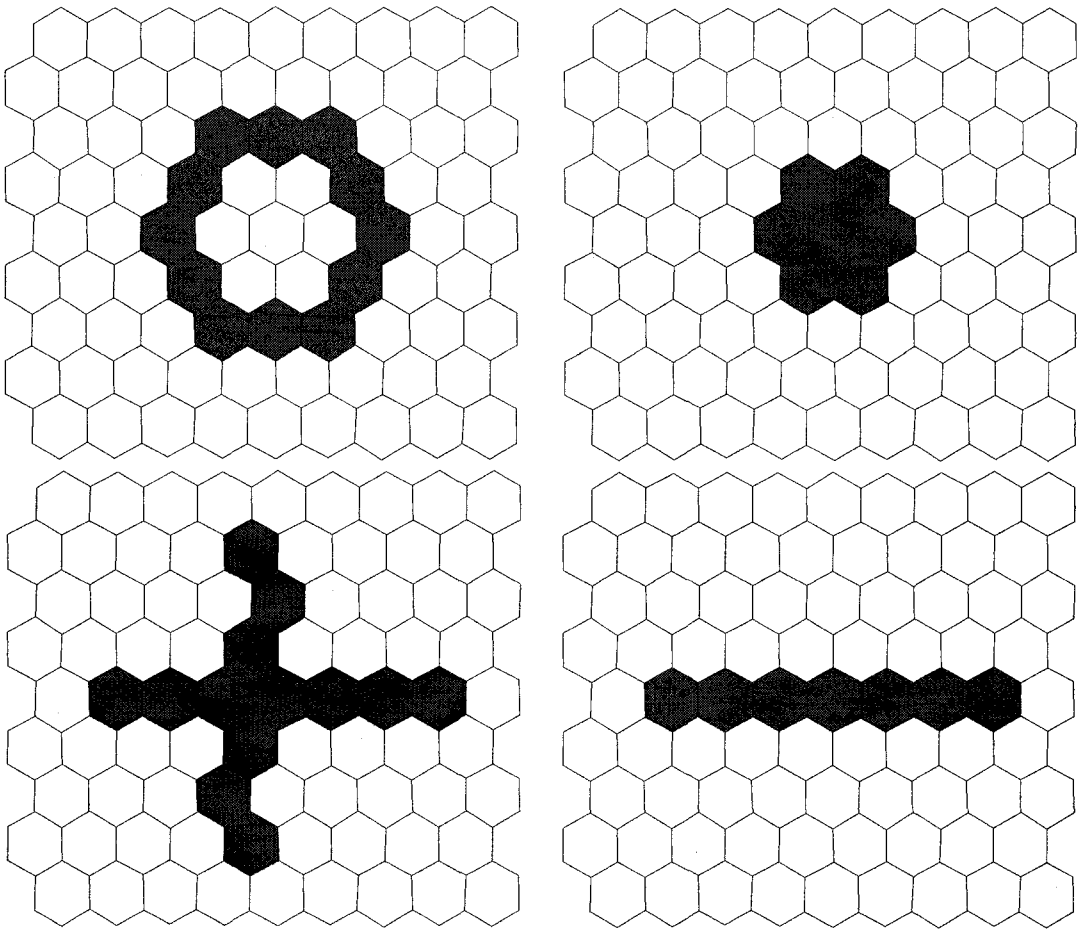


Figure 4.7a Proportion of tests correctly identifying the presence of a simulated circular cluster

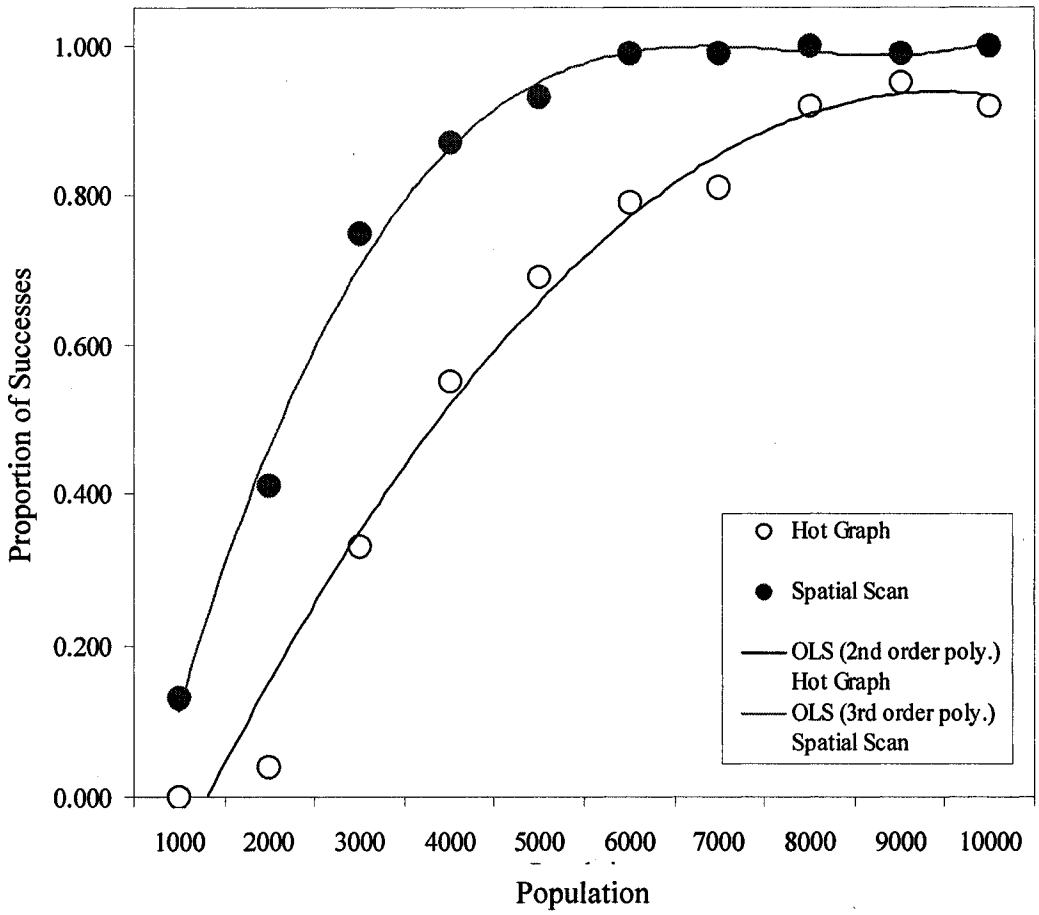


Figure 4.7b Proportion of tests correctly identifying the presence of a simulated 'ring' cluster

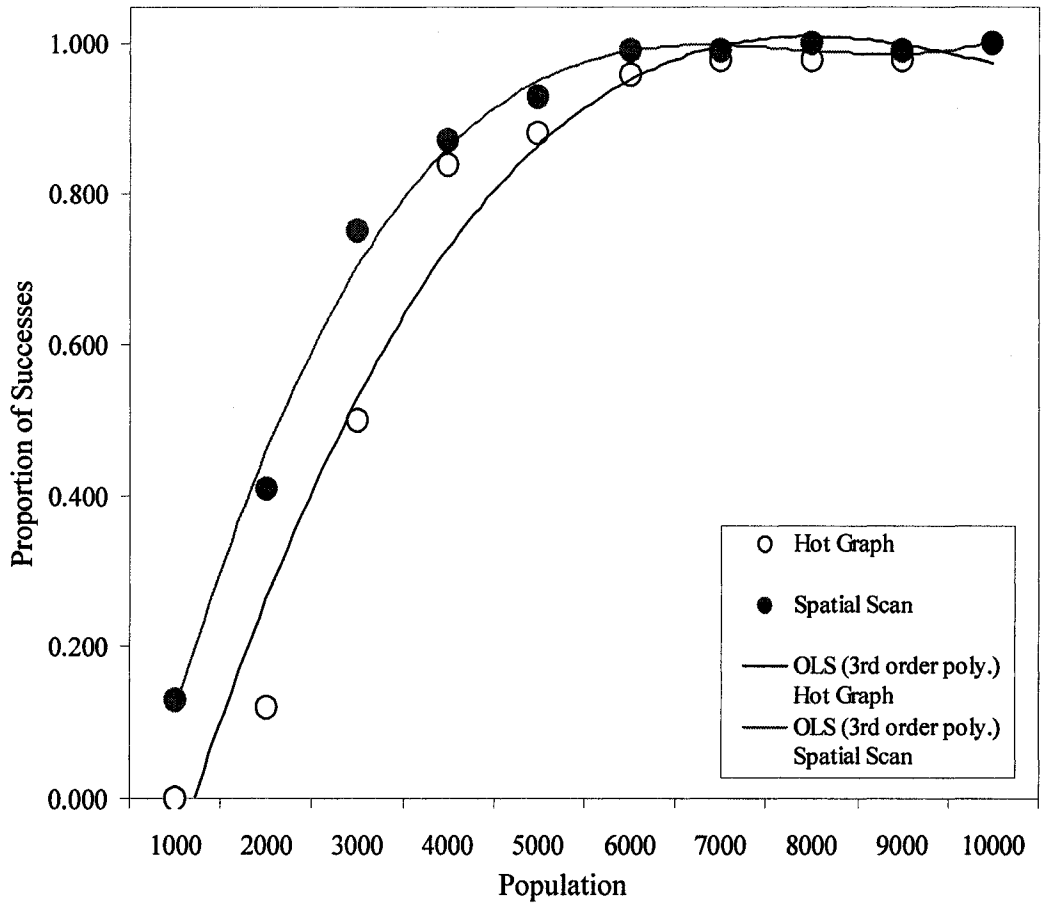


Figure 4.7c Proportion of tests correctly identifying the presence of a simulated 'line' cluster

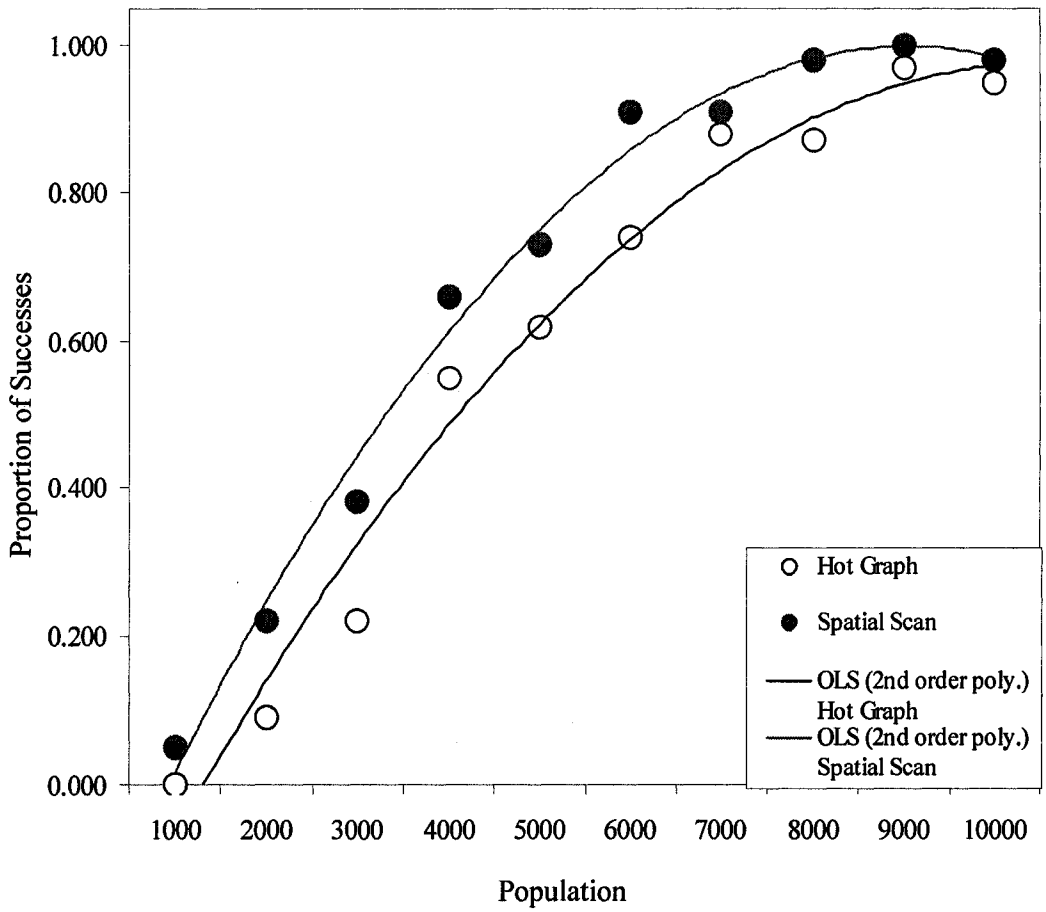


Figure 4.7d Proportion of tests correctly identifying the presence of a simulated 'network' cluster

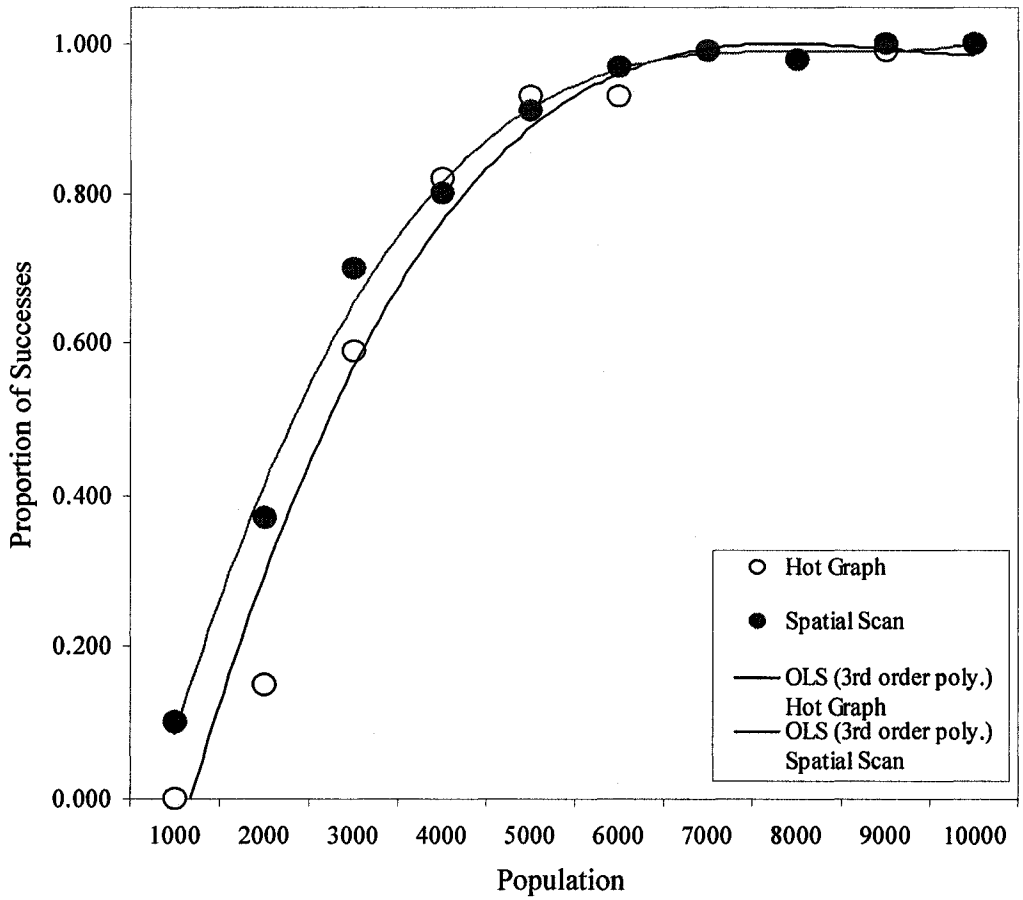


Figure 4.8a Mean sensitivity for circle cluster pattern

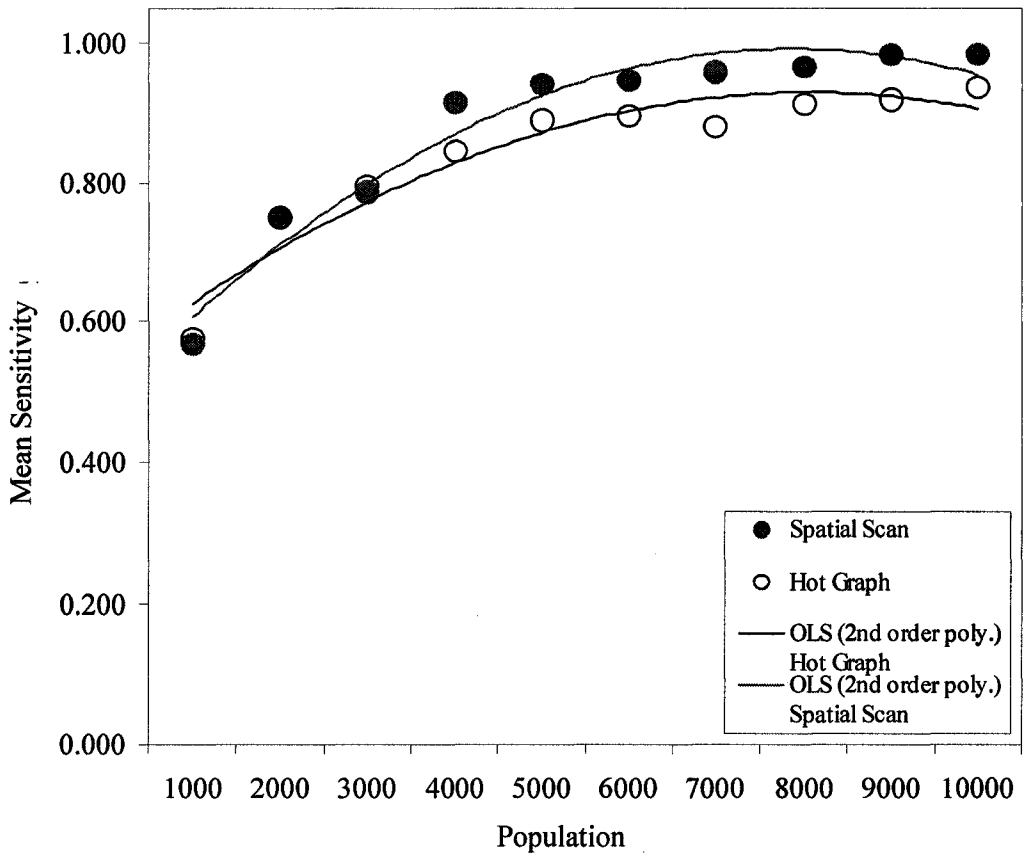


Figure 4.8b Sensitivity for ring cluster pattern

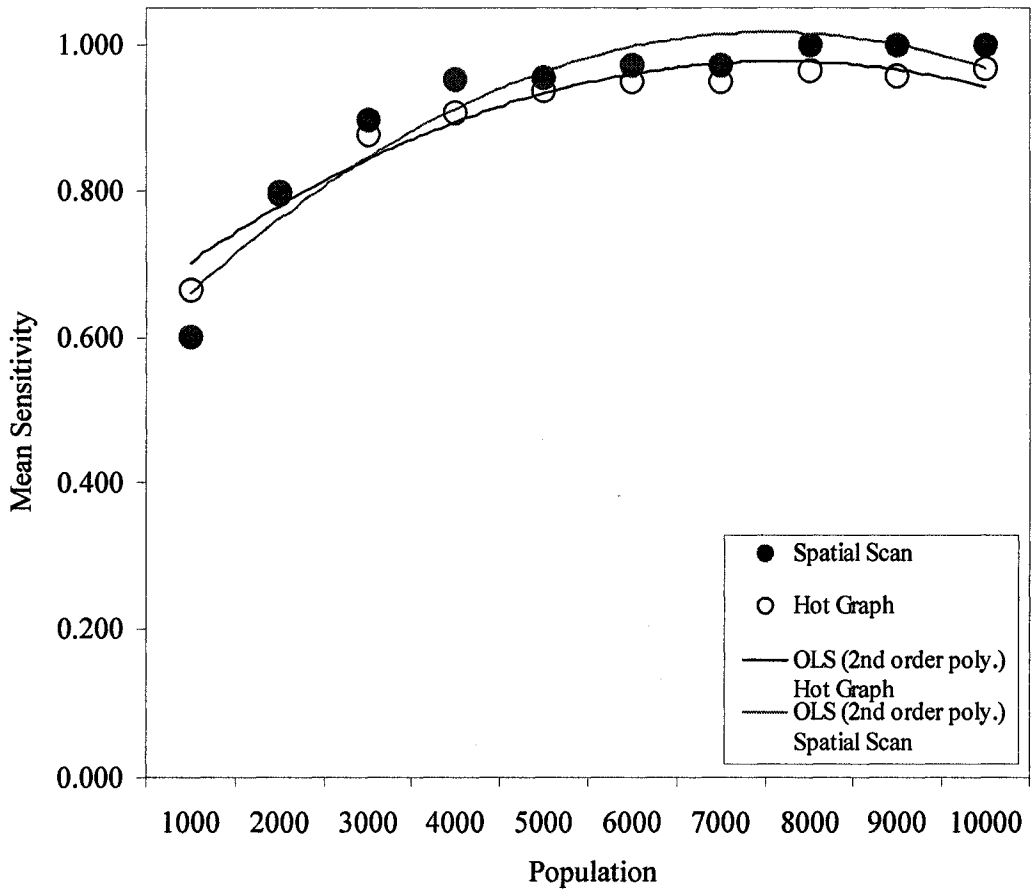


Figure 4.8c Sensitivity for simulated 'network' cluster

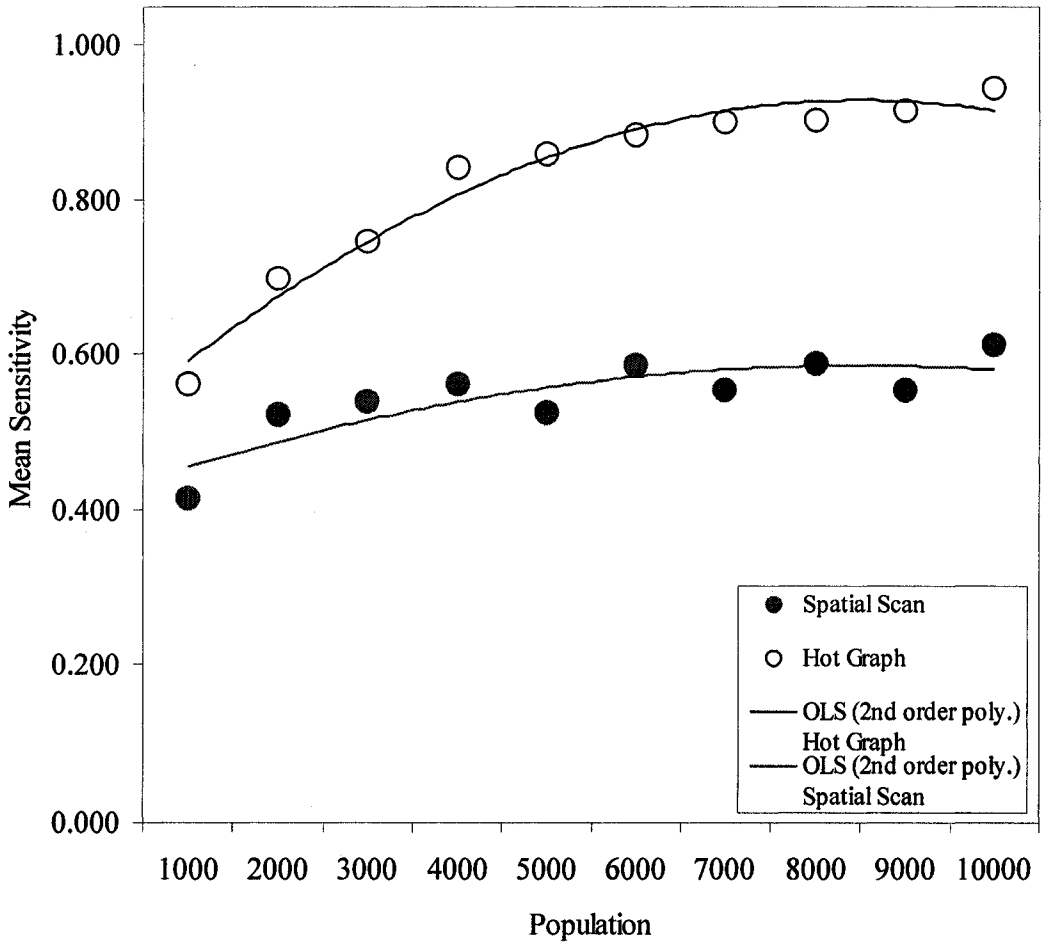


Figure 4.8d Sensitivity for line cluster pattern

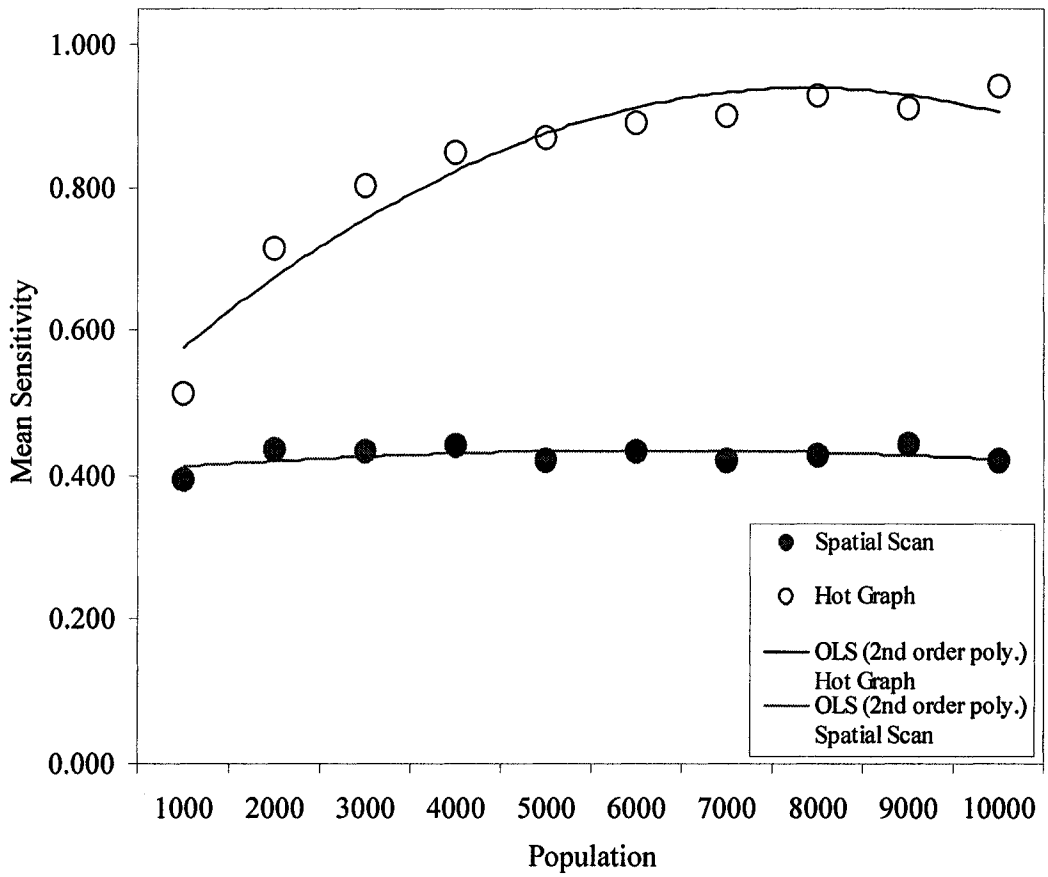


Figure 4.9a Positive predictive values for circle cluster pattern

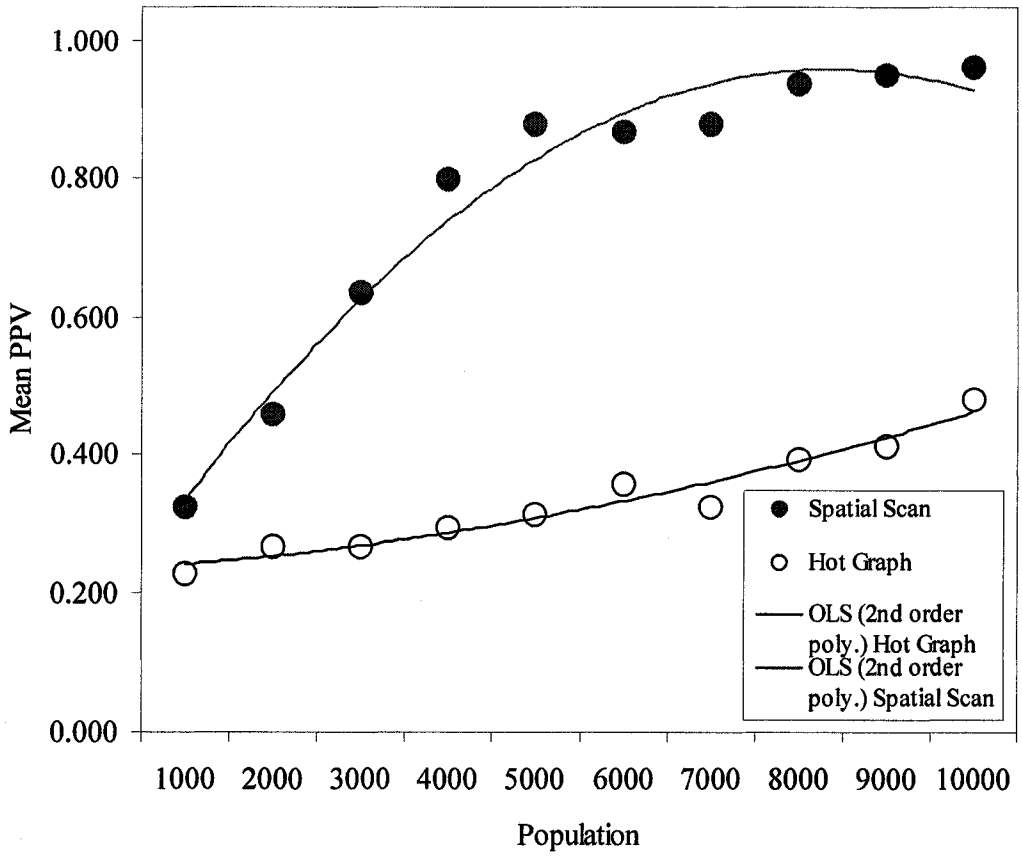


Figure 4.9b Positive predictive values for ring cluster pattern

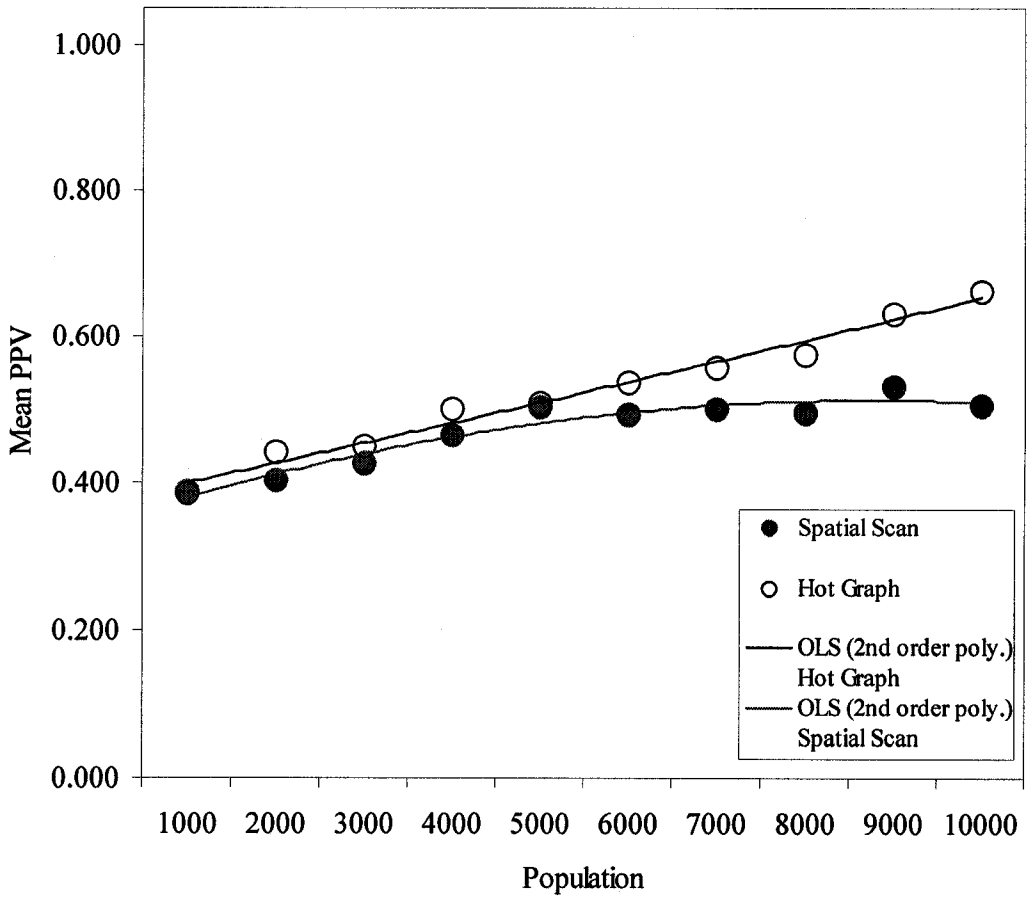


Figure 4.9c Positive predictive values for network cluster pattern

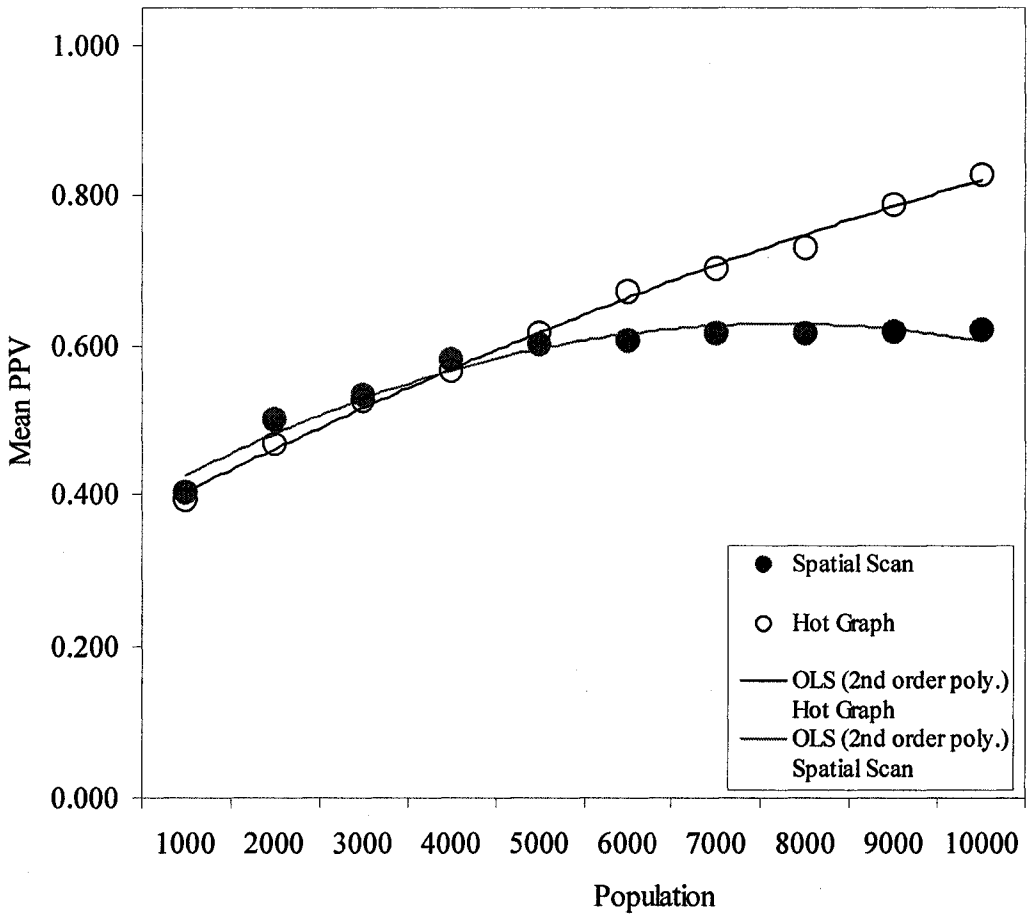


Figure 4.9d Positive predictive values for line cluster pattern

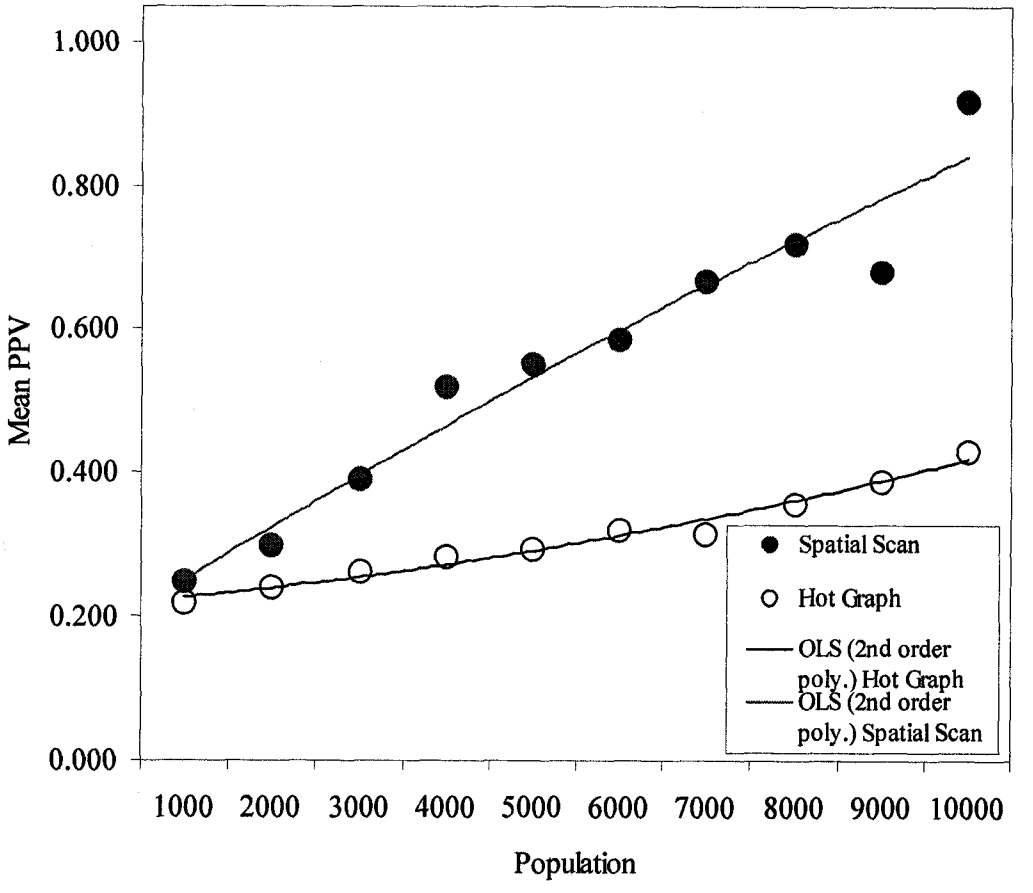
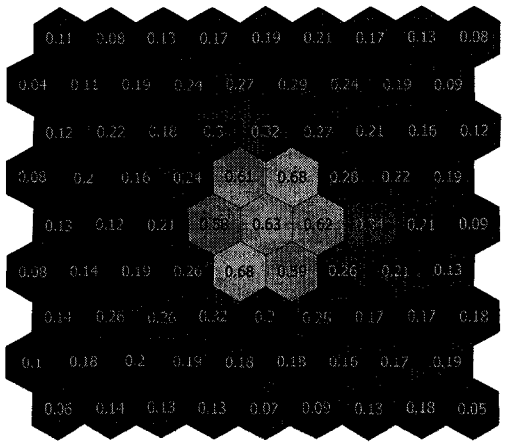
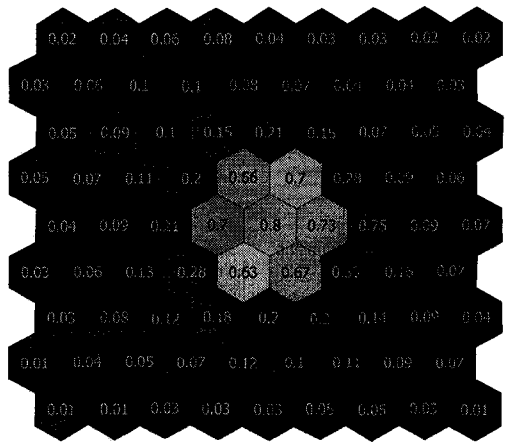


Figure 4.10a Circle cluster pattern: proportion of true positives (black font) and proportion of false positives (green font)

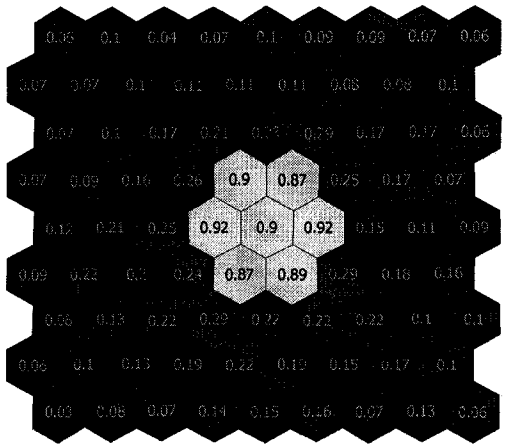
Circle, HG method Population=1 000



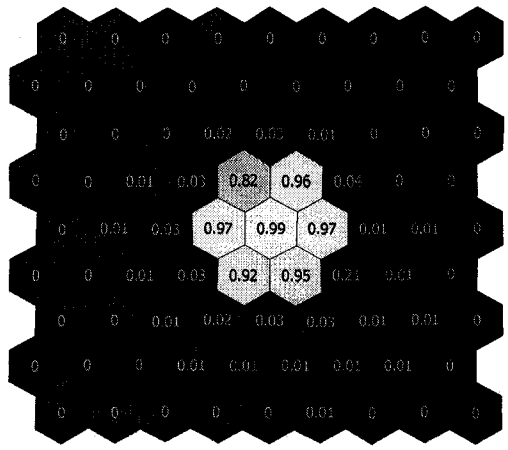
Circle, SS method Population=1 000



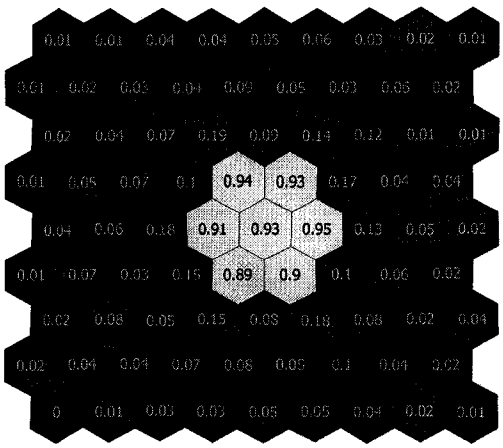
Circle, HG method Population=5 000



Circle, SS method Population=5 000



Circle, HG method Population=10 000



Circle, SS method Population=10 000

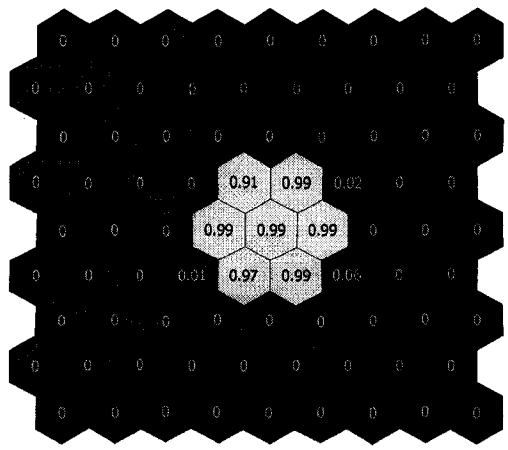
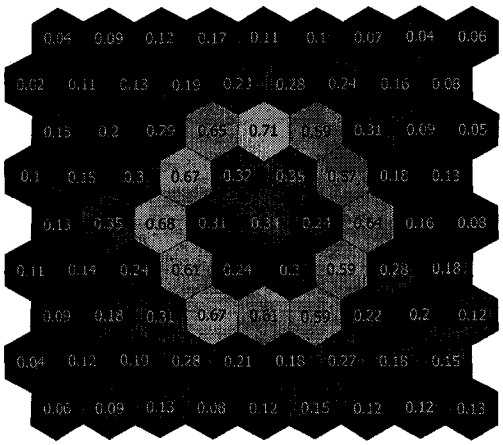
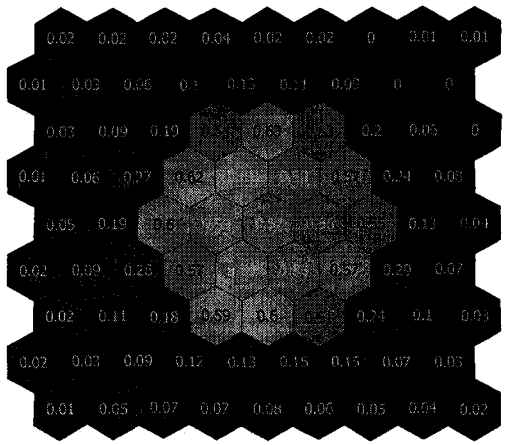


Figure 4.10b Ring cluster pattern: proportion of true positives (black font) and proportion of false positives (green font)

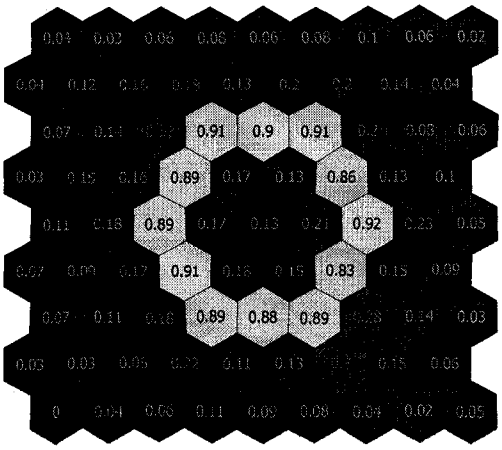
Ring, HG method Population=1 000



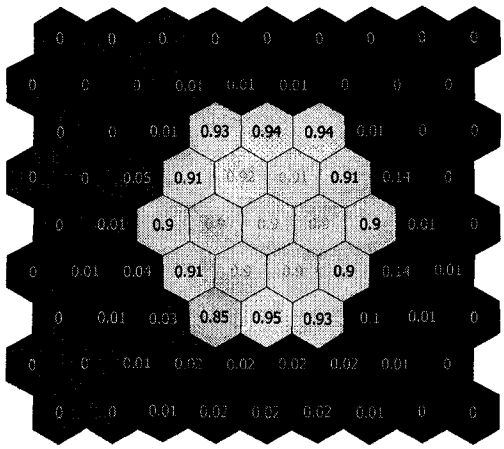
Ring, SS method Population=1 000



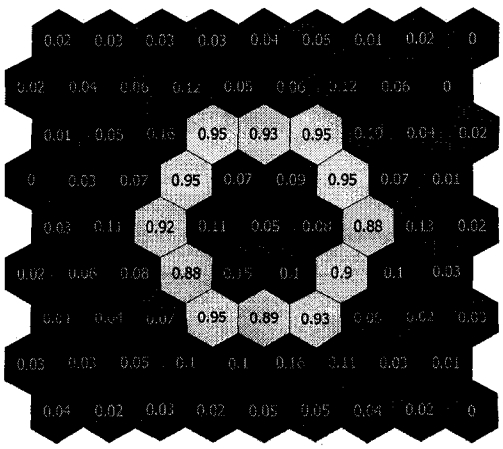
Ring, HG method Population=5 000



Ring, SS method Population=5 000



Ring, HG method Population=10 000



Ring, SS method Population=10 000

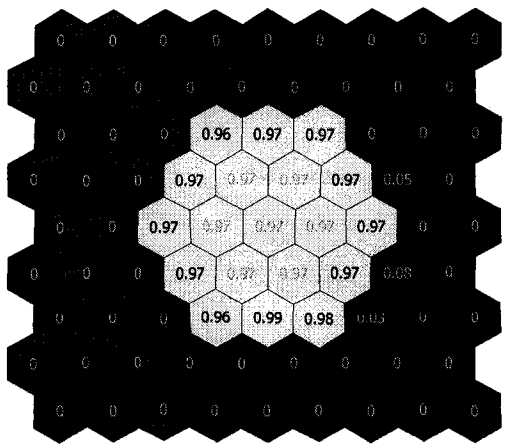
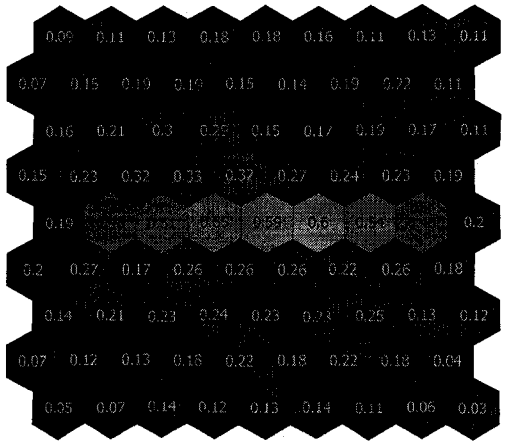
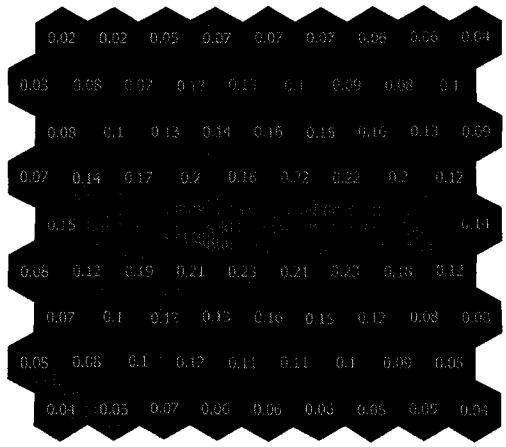


Figure 4.10c Line cluster pattern: proportion of true positives (black font) and proportion of false positives (green font)

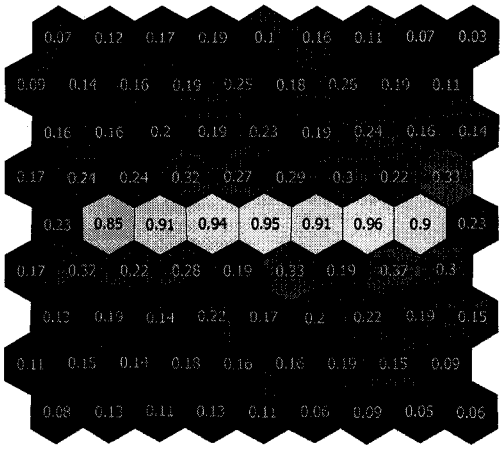
Line, HG method Population=1 000



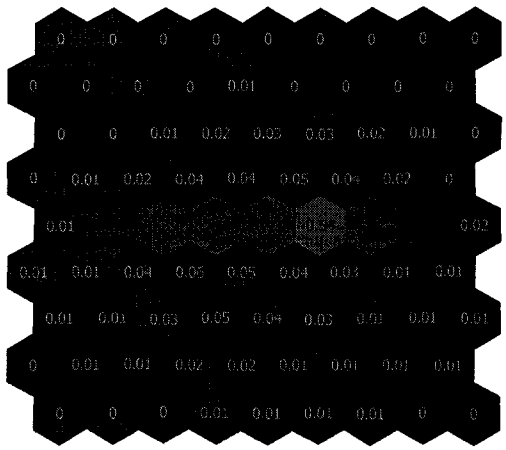
Line, SS method Population=1 000



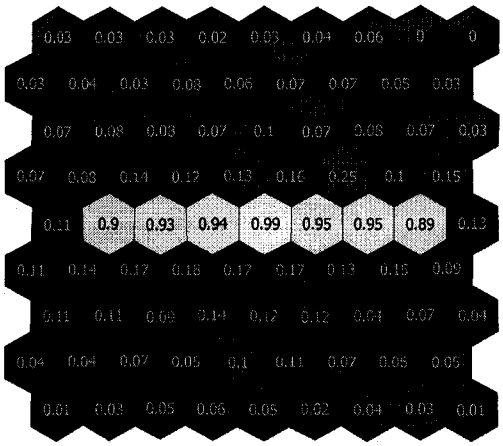
Line, HG method Population=5 000



Line, SS method Population=5 000



Line, HG method Population=10 000



Line, SS method Population=10 000

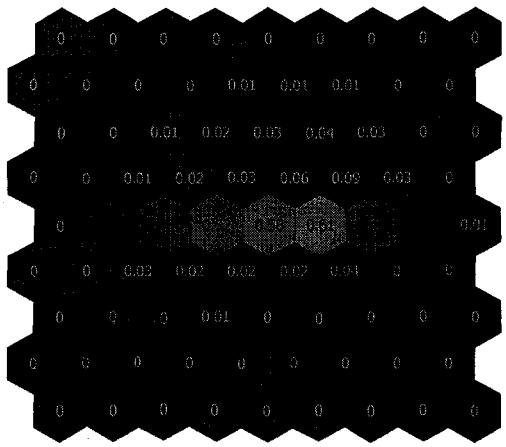
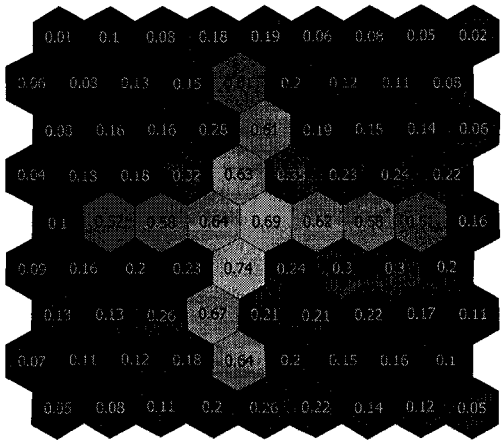
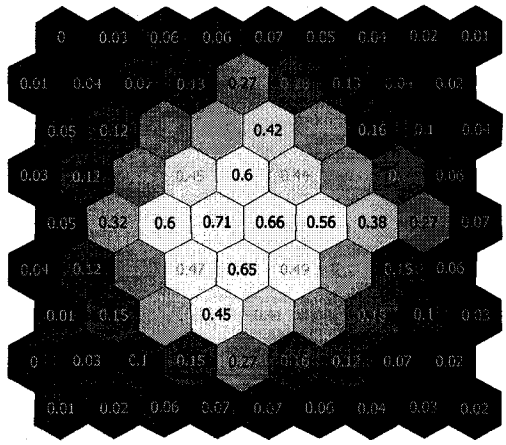


Figure 4.10d Network cluster pattern: proportion of true positives (black font) and proportion of false positives (green font)

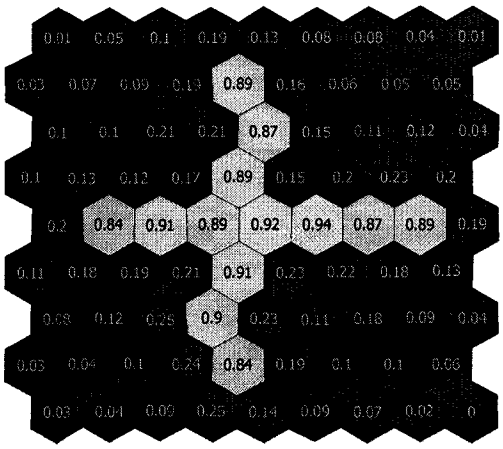
Network, HG method Population=1 000



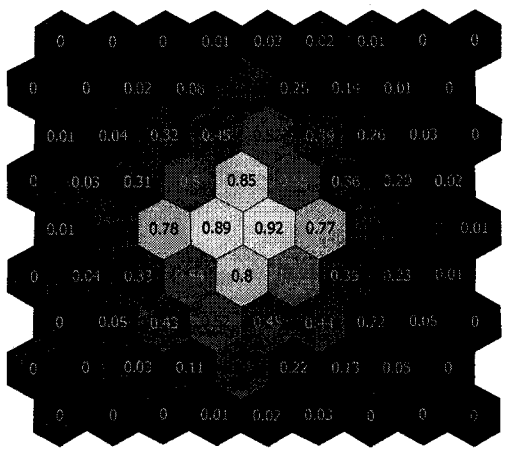
Network, SS method Population=1 000



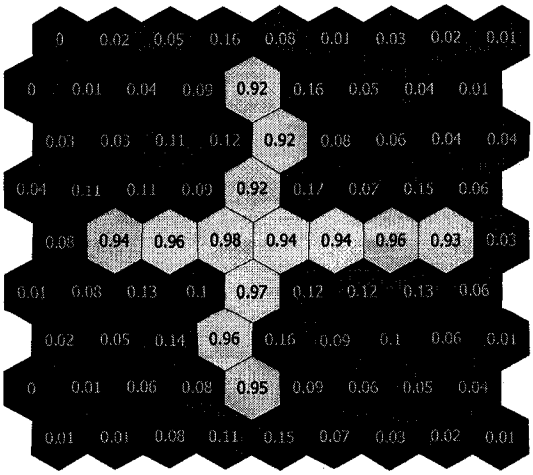
Network, HG method Population=5 000



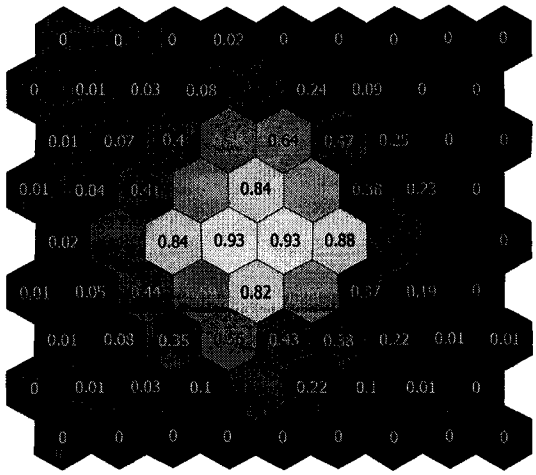
Network, SS method Population=5 000



Network, HG method Population=10 000



Network, SS method Population=10 000



4.7 References

Anselin (1995). Local indicators of spatial association—LISA. *Geographical Analysis* **27**:93-115.

Batty M. (2005). *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models and Fractals*. MIT: Cambridge.

Besag, J, Newell J (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society Series A* ***154***: 143-155.

Bozkaya B, Erkut E, Laporte G (2003) A tabu search heuristic and adaptive memory procedure for political districting. *European Journal of Operational Research* **144**:12 – 26.

Conley J, Gahegan M, Macgill J (2005) A genetic approach to detecting clusters in point data sets. *Geographical Analysis* **37**:286-314.

Cuzick J, Edwards R (1990) Spatial clustering for inhomogenous populations. *Journal of the Royal Statistical Society B* **52**:73-104.

Diggle P (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society A* **153**:349-362.

Duczmal L, Assunçao R (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* **45**:269-286.

Ederer F., Myers M. H., Mantel N. (1964) A statistical problem in space and time: do leukemia cases come in clusters? *Biometrics* **20**:626-638.

Gangnon R E, Clayton M K (2001) A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine* **20**:2977-2987.

Getis A, Ord J K (1996) Local spatial statistics: an overview. *Spatial Analysis: Modelling in a GIS Environment.* P Longley, M Batty (eds.) Geoinformational International: Cambridge.

Knox E G (1988) Detection of clusters. In *Methodology of Enquiries into Disease Clustering.* London School of Hygiene and Tropical Medicine: London.

Kulldorff M, Nagarwalla N. (1995) Spatial Disease Clusters: Detection and Inference. *Statistics in Medicine* **14**:799-810.

Kulldorff M (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods* **26**:1481-1496.

Kulldorff M (1999) Spatial scan statistics: models, calculations and applications. In *Scan Statistics and Applications.* Glaz J, Balakrishnan N (eds.) Birkhäuser: Boston.

Kulldorff M, Information Management Services, Inc. (2004) SaTScan™ v5.0: Software for the spatial and space-time scan statistics. http://www.satscan.org/.

Kulldorff M (2005) Scan statistics for geographical disease surveillance: an overview. In *Spatial and Syndromic Surveillance for Public Health.* Lawson A B, Kleinman K (eds.) Wiley: West Sussex.

Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. *Statistics in Medicine In Press*

Lawson A B, Kulldorff M (1999) A review of cluster detection methods. In *Disease Mapping and Risk Assessment for Public Health.* L Lawson A B, Biggeri A, Böhning D, Lesaffre E, Viel J F, Bertollini R (eds.) Wiley: Chichester.

Le N D, Petkau A J, Rosychuk R (1996) Surveillance of clustering near point sources. Statistics in Medicine **15**:727-40.

Marshall R J (1991) A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society* **154**:421-441.

Neill D B, Moore A W (2005) Efficient scan statistic computations. *Spatial and Syndromic Surveillance for Public Health.* Lawson A B, Kleinman A B (eds.) Wiley: West Sussex.

Openshaw S, Craft A, Charlton M, Birch J (1988) Investigation of leukemia clusters by use of a geographical analysis machine. *Lancet* **331**:272-273.

Openshaw S, Rao L (1995) Algorithms for reengineering 1991 Census geography *Environment and Planning A* **27**:425-446

Patil G P, Taille C (2003) Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science* **18**:457-465.

Patil G P, Taille C (2004) Upper level set scan statistic for detecting arbitrary shaped hotspots. *Environmental and Ecological Statistics* **11**:183-197.

Prim R C (1957) Shortest connection networks and some generalizations. *Bell Systems Technology Journal* **36**:1389-1401.

Rushton G, Krishnamurthy R, Kirshnamurti D, Lolonis P, Song H (1996). The spatial relationship between infant mortality and birth defect rates in a U.S. city. *Statistics in Medicine* **15**:1907-1919.

SAS Institute (1999) SAS 8.2. SAS Institute Incorporated: Cary

Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* **4**:1-15

Tobler W R (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46**:234-240.

Trudeau R J (1993). *Introduction to Graph Theory.* Dover: New York.

Turnbull B W, Iwano E J, Burnett W S, Howe H L, Clark L C. (1990) Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* **132**:S136-S143.

Wakefield J c, Kelsall J E, Morris S E (2000) Clustering, cluster detection, and spatial variation in risk. In *Spatial Epidemiology: Methods and Applications.* Elliot P, Wakefield J, Best N, Briggs D (eds.). Oxford University Press: Oxford.

Waller L A, Gotway C A (2004) *Applied Spatial Statistics for Public Health Data.* Wiley: Hoboken

Whittemore A S, Friend N, Brown B W, Holly E A (1987). A test to detect clusters of disease. *Biometrika* **74**:631-635.

Winston W L, Venkataramanan M (2003*) Introduction to Mathematical Programming.* fourth edition, Thomson Brooks/Cole: Pacific Grove.

CHAPTER 5: Geographic discovery through cluster detection

5.1 Introduction

The spatial scan (Kulldorff and Nagarwalla 1995; Kulldorff 1997) method of geographic cluster detection has received considerable applied and methodological research attention over the last ten years. The original spatial scan uses moving 'circular windows' to scan a surface for high rates of disease, identifying the approximate location and statistical significance of local clusters. Recent developments have sought to improve upon cluster detection methodology by expanding the search process to include shapes other than circles (e.g. Conley, Gahegan and Macgill 2005; Kulldorff et al. 2006). Other methods have further restructured how the problem is formulated, using adjacency rather than geometry to constrain the search (e.g., Duczmal and Assunçao 2004; Tango and Takahashi 2005; Chapter 4).

Many studies of cluster detection methodology offer evidence about the performance of different methods based on simulated data (e.g. Rogerson 1997; Tango 2000; Gangnon and Clayton 2001; Kulldorff, Tango and Park 2003; Song and Kulldorff 2003; Tango and Takahashi 2005; Conley, Gahegan and Macgill 2005; Puett et al. 2005; Aamodt, Samuelsen and Skrondal 2006). Simulated data studies are an important starting point for methodological innovation, but may be limited in their generalisability. For example, obtaining power estimates for cluster detection methods is complicated by topology, scale and variations in population density (Waller and Gotway 2004 pp. 260-261). As a result, the ability of a method to detect a cluster may vary depending on where the cluster occurs in a study region, even with other parameters—like shape, magnitude and size of clusters—held equal.

There are also interesting features of real world problems that are hard to recreate in a simulated computer environment. In the real world, many interacting processes can operate simultaneously, and produce complex spatial structures. Autocorrelation, trend, aggregation effects and edge effects are reasonably simple to simulate independently, but in combination, present a large number of parameters to control in a simulated setting. Indeed, it is probably impossible to

simulate and interpret how a method performs with respect to all of these processes simultaneously. Eventually, most methods of spatial analysis must be explored with real world data. This is particularly true when such methods are designed for the study of health and disease. Good methods inform applied researchers in ways that better enable them to contribute to improvements in human health and well being. Bad methods can fail to inform, provide false information, and may even cause harm.

Many studies of cluster detection methodology include tests that compare new approaches using real data sets. This has frequently involved the study of cancer (e.g. Whittemore et al. 1987; Cuzick and Edwards 1990; Le, Petkau and Rosychuk 1996; Goovaerts and Jacquez 2004) though studies have used other data on occasion, including West Nile virus (Mostashari et al. 2003), syndromic surveillance data (Kulldorff et al. 2005), birth defects (Hill, Ding and Waller 2000) and homicide (Duczmal, Assunção 2004). One data set of New York of childhood leukaemia has been used in several methodological studies (e.g. Turnbull et al. 1990; Waller and Turnbull 1993; Kulldorff and Nagarwalla 1995; Gangnon and Clayton 2001; 2003; 2004) to facilitate comparisons between methodologies. Although such applied work has obvious value in illustrating the ability of a method to identify apparent patterns, like simulated experiments, it is hard to generalize to a larger scope. In this case, the data may be real, but it is difficult to apply an instance (as opposed to a universe) of observations to a broader range of situations. Furthermore, since there is no benchmark for comparison (in the real world we have no control or knowledge about a cluster's true location or magnitude), differences that are observed have no obvious point of reference. As a result, isolated analyses of real data remain fairly limited in their ability to generalize observations about performance to different places, different diseases, or even different scales of study.

The purpose of this study is to compare results from the spatial scan (Kulldorff 1997) and hot-graph approach described in Chapter 4. By exploring these methods on ten sets of real disease data, we can gain an understanding of how similar (or different) the results of these two approaches are for some real

disease cluster detection tasks. Since we select a number of diseases—both chronic and communicable—of varying case frequency, our analysis provides some perspective on the variability we may expect to see in other applications in similar settings.

5.2 Methods

5.2.1 Data

The province of Alberta has a population of over 3.2 million people and is located in Western Canada (Figure 5.1). The majority of the population is located in the South-east quadrant of the province. Including nearby suburban municipalities, the populations of both Edmonton and Calgary are near one million people. Over 95% of Alberta's population is insured by a provincial health insurance plan. Most persons not covered are members of the military or R.C.M.P. who receive insurance coverage directly from the Federal government. At present, the province pays for most services offered by physicians, and maintains an administrative database of services and population data to facilitate the payment system. The bulk of this system is made up of a fee-for-service database (or "claims") that includes diagnostic information as well as service codes necessary for payment decisions.

We use this database as a source of information for ten different disease conditions presenting in the 2003 calendar year (Table 5.1). These conditions were chosen to include both infectious and chronic illness of varying rate of occurrence. Considerable work has gone into reviewing the accuracy of disease definitions based on claims data (Roos, Sharp and Cohen 1991; Roos et al. 1993; Muhajarine et al. 1997). We assume several caveats with the use of such data. First, we cannot make a distinction between incident and prevalent diagnoses. Although data correspond to unique individuals (rather than services), we cannot know if a diagnosis represents a follow-up visit for a persistent (prevalent) illness, or the first presentation of a new (incident) illness. Second, there is a possible rural-urban bias in the diagnosis of some diseases. Differences in physician training, the availability of equipment and practice style can complicate standardized disease definitions across different geographies, even at the intra-

provincial scale (Yiannakoulias, Svenson and Schopflocher 2005). Third, influenza, gonorrhoea, giardiasis, and food poisoning are typically confirmed through laboratory testing. Since we rely on claims diagnoses (which are infrequently validated by laboratory tests), some of our disease data may have poor diagnostic precision.

The denominator used in the analysis is based on a population registry file maintained by the provincial health ministry. For all diseases but asthma and Parkinson's, the at-risk populations were assumed to be the entire population. For asthma, cases and population data were restricted to persons less than or equal to 20 years of age. For Parkinson's, case and population data were restricted to persons 65 years of age and older. Both of these age restrictions were used to improve the precision of diagnoses—asthma is typically associated with younger ages, Parkinson's is typically associated with older ages.

For each disease, case and population data were aggregated to a polygon tessellation covering the province. The tessellation consists of the Alberta 'sub-rha' administrative health boundaries defined by Alberta Health and Wellness (Alberta Health and Wellness 2004). The 68 sub-rhas (or 'zones') were designed based on regional consultation, and a desire to create areas of roughly equal population. The area of each sub-rha is relatively small in regions where population density is high, and larger where population density is low. Geometric centroids of these zones were derived for the spatial scan method (Figure 5.2). Each centroid is calculated based on the centre of the smallest possible rectangle that encloses each zone. An adjacency matrix was built from the tessellation based on the queen's case; zones touching at one or more points (or lines) are considered connected. All programming of the hot-graph method was performed in SAS 8.2 (SAS Institute, 1999). SaTScan™ (Kulldorff and Information Management Services 2004) was used to find the spatial scan clusters.

5.2.2 Approach

We compare the results of two cluster detection approaches (the spatial scan and the hot-graph) with respect to the 10 diseases represented in Alberta sub-rhas. When working with aggregate disease data, the spatial scan uses a circular

window of varying size to progressively aggregate neighbouring centroids to a seed centroid, and calculates a likelihood ratio statistic at each window size (Kulldorff 1997). Once all window sizes are exhausted for one seed, the scan moves to a new seed until all seeds have spawned a series of searches. A most-likely cluster is chosen based on the highest likelihood ratio statistic of all windows at all centroids. Under the null hypothesis that the rate within a detected cluster is not different from the rate outside the cluster, significance is usually determined through 999 Monte Carlo simulations (Kulldorff 2004). For each simulation, case and non-case status is randomly re-assigned to all the individuals in the entire study area (through random labelling). A search for a most-likely cluster is performed for each simulated data set, and the 999 simulated most-likely clusters are compared to the real cluster to obtain a measure of statistical significance.

Inspired by the ideas of Patil and Taille (2003; 2004) and particularly Duczmal and Assunção (2004), the hot-graph cluster detection method is constrained by adjacency rather than fixed geometric structures (such as circular windows, ellipses or nearest neighbour distance). The procedure starts at a seed zone (which is itself treated as a potential cluster), and adds the adjacent zone that increases the likelihood ratio the most. These two zones are now a new potential cluster set, and neighbours of both zones are new feasible choices. Zones continue to be added to the potential cluster set until a large number of zones have been added to the current search (for example, once the potential cluster has a total population equal to 50% of the study area's population). At this point, the search moves to a new zone and starts over. The likelihood ratio test for each potential cluster is stored, and the potential cluster with the highest likelihood ratio is considered the most-likely cluster. Significance of the most-likely hot-graph is determined similar to the Monte Carlo methods described above, except a new topology must be derived for each simulated data set.

For each data set, both methods are used to identify the location and significance of a single most-likely cluster. Both methods are capable of identifying and locating secondary clusters. These additional clusters make evaluations of

performance more complicated, require consideration within the context of their use, and are therefore not examined here. Results of the two methods are tabulated and mapped so that comparative assessments can be made. We also map crude (i.e., not age-sex standardized) rates of disease (Appendix III) to compare the degree to which the different approaches correspond to simple disease maps.

5.3 Results

Table 5.2 reports the information comparing the results of both cluster detection methods. The p-value reported corresponds to the likelihood ratio test; if p is sufficiently small, we reject the null hypothesis that risk the inside and outside the cluster is not different. If we choose p<=0.001 as a cut-off of statistical significance, both methods identify significant clusters for most diseases. For the spatial scan, the null hypothesis is rejected in all cases except for giardia. The hot-graph method fails to reject the null hypothesis for three of the diseases: giardia, salmonella food poisoning and bacterial food poisoning. In every case, the two methods locate clusters that overlap by at least one sub-rha. In all but one case, the hot-graph clusters include more zones, but cluster size appears to correlate somewhat between the two methods. The relative risk (RR) estimates the ratio of observed to expected number of cases within a most-likely cluster. Since the experiment was set up to search for high clusters, the numbers are all greater than one. As numbers increase from one, risk is proportionally higher in cluster areas. When close to one, the number of cases inside a cluster is not different from the expected number of cases. The size of the clusters (N) varies between methods, with the hot-graph approach finding larger clusters of all diseases except for asthma. Except in the case of asthma and giardia, estimates of relative risk in clusters found by the spatial scan are larger than those found by the hot-graph approach.

Agreement between the methods is summarized in the four right-most columns of Table 2. The methods differ considerably with respect to the sub-rhas included as part of detected clusters. Most of these differences result from the hot-graph's tendency to define clusters with a larger number of sub-rhas than the

spatial scan. Maps help illustrate these differences (Figures 5.3 to 5.12). The colours of the polygons denote the differences noted in Table 2. Yellow identifies sub-rhas that are identified as part of a cluster by both methods. Blue identifies sub-rhas that are only part of a hot-graph identified cluster. Red identifies sub-rhas that are only part of a spatial scan identified cluster. White sub-rhas are not part of any detected cluster.

For all diseases examined here, the hot-graph and spatial scan methods identify most-likely clusters in the same general region of the province. For most of the maps, the hot-graph method identifies clusters with a large number of sub-rhas. In some cases, the hot-graph clusters also include a significant amount of the province's area. For example, in the case of alcohol and drug related illnesses, the cluster includes roughly two-thirds of the province's total area (Figure 5.9). In several other cases, the clusters include nearly half of the province's area. In most of these instances, both methods report fairly low relative risk estimates. The highest reported relative risk estimates are associated with clusters found by the spatial scan, and in these cases, the number of sub-rhas in the cluster is small. In the case of gonorrhoea, both methods find clusters with a small number of sub-rhas and with high relative risk estimates (Figure 5.12). The relative risk estimates for the asthma clusters are relatively low for the hot-graph (1.23) and spatial scan (1.19) methods, and perhaps not above the threshold of clinical importance. However, the map does illustrate differences in the geography of detected clusters (Figure 5.3). The spatial scan cluster includes two more sub-rhas than the hot-graph method. The hot-graph cluster has a more complex geometry, consisting of a half-ring shape that does not include two central sub-rhas found by the spatial scan. Rate maps (see Appendix III) support this observation; the two central sub-rhas have comparatively lower rates of disease than the other sub-rhas included in the cluster.

5.4 Discussion

5.4.1 Differences between methods

The great challenge of most methodological studies is to derive general, broadly applicable conclusions from specific experimental or theoretical work.

One way to resolve this challenge is to compare methods using simulated data. Simulated data are easy to control, and offer information about how methods perform under idealized conditions. This is not unusual, and in fact adopts a strategy common in the social sciences—how do methods perform under certain experimental conditions, all else being equal? Another option is to make comparisons with real data, and in particular, data that have been studied before. Disagreement with other research could reveal a breakthrough, or a need for further scrutiny. However, in both cases, these approaches probably fail to cover a sufficient range of scenarios that would enable comment on the breadth of real-world situations. Though our study does not exhaust all real world data, it covers a diverse range of disease outcomes, and illustrates some interesting, and possibly general, observations.

Geographic cluster detection methods are usually constrained geographically such that the cluster and non-cluster sets are contiguous, relatively compact, or limited to a pre-specified shape (like a circle, ellipse, square, etc.). Historically, the most common structures were based on geographic neighbours either explicitly (by searching for adjacent or nearby areas) or implicitly (by searching compact geometric shapes). Until recently, little work considered whether or not these approaches were sufficient in real world applications of disease cluster detection. In our study, the spatial scan method found (in most cases) clusters smaller in size and with larger relative risk estimates than the hot-graph method. The spatial scan also identified more significant clusters than the hot-graph method and in the general case, appears more able to approximate the location of anomalously high clusters of disease.

The hot-graph method identified statistically significant clusters, many of which were irregularly shaped. This result can be interpreted in at least two ways. First, it could reflect an over-fitting of data. In simulated experiments, the hot-graph method has a tendency to include false-positive areas (Chapter 4). The hot-graph search algorithm always adds the 'best' (that is, most likelihood ratio maximizing) of all the zones adjacent to the current potential cluster. As the search proceeds, the number of available neighbours increases as does the

potential for one of these zones to have a high rate by chance alone. This can result in the inclusion of zones that are not actually part of a 'true' cluster, but have a relatively high disease frequency by chance. The detected clusters may include true positives and false positives, but unfortunately, it is impossible to distinguish between the two. Under the assumption that clusters in the real world are usually circular in shape, this seems like a sensible explanation—since clusters *should* be circular, and the shapes observed here are not, the hot-graph method seems prone to sending us on 'wild goose chases', identifying some zones as part of a cluster that do not truly deviate from the average.

Alternatively, we may interpret these observations as a true reflection of geographic structure. Several of the irregular clusters were found in diseases with high baseline rates, and low local standard error. As such, it is unlikely that more than a few of the sub-rhas in any of the clusters are a result of chance over-fitting. For more common diseases, these patterns may simply illustrate larger-scale trends of spatial variation. When there are broad patterns of spatial variation in disease—for example, a North-South gradient—we should expect the hot-graph to identify a statistically significant cluster that divides the study region into North and South halves. In this case, the 'cluster' is really just a simplification of spatial trend. In other instances, the method may be characterizing the landscape of disease by identifying regions with common social and physical environments that influence disease patterns. For example, the hot-graph cluster of diseases related to alcohol and drugs (Figure 5.9) closely coincides with low-population density non-farming areas of Alberta. Hypertension (Figure 5.6) and Parkinson's (Figure 5.7) both exhibit clusters in rural farming areas of Alberta. Several of the diseases explicitly exclude Edmonton and Calgary, forming irregular shapes around one or both of these cities. In these cases, the hot-graph approach is similar to an exercise in exploratory classification—grouping high-rate areas at a large regional scale, rather than identifying anomalies of local concern.

5.4.2 Hot-graph searches in cluster detection and geographical analysis

There may be a few instances in which the hot-graph or other data-directed methods find a cluster that is not detected by a circular spatial scan, particularly

when the distribution of disease anomalies express irregular patterns. However, evidence from experiments in Chapter 4 and advances of the spatial scan to include other geometric shapes makes this seem unlikely in all but a few settings. Nonetheless, the hot-graph method offers some interesting information about detected clusters. When a spatial scan cluster includes zones that a hot-graph cluster does not include, it is possible that these zones are included because the pre-specified geometry is imprecise. This appears to occur in the case of asthma; the spatial scan method includes two sub-rhas with moderate to low disease rates. On the other hand, agreement between methods offers verification of the spatial scan's results; when the hot-graph approach finds a cluster with a shape similar to that found by the spatial scan, this represents important confirmatory evidence about the detected cluster's shape. Finally, when the spatial scan and a hot-graph approach simultaneously fail to find a statistically significant cluster, it is reasonable to assume that the circular (or other) constraint applied in the spatial scan was not prohibitive.

The hot-graph method also seems to work as a spatial classification and visualization tool. Zones with higher disease rates are grouped into contiguous geographic groups that reflect a large regional pattern of disease variation. In this application other evaluation methods may be more suitable. When working with geographically aggregate disease data, the likelihood ratio is based on a Poisson distribution model, but there may be better alternatives in situations where the disease is common. For example, one could consider maximizing relative risk subject to a minimum population threshold. The hot-graph approach may also provide some interesting information about scale and resolution of study. When one of these methods finds large irregular clusters, it may indicate that the scale of clustering is not local, but of a larger, more regional scale. In these instances there may be important spatial variation in illness that is uniquely informative at lower resolutions or at larger geographic scales.

In light of the applications and limitations of data-directed methods generally, we recommend that the hot-graph method be used as an exploration and validation tool in conjunction with the spatial scan. Though formal integration is

possible, it may be simpler and more effective to employ these approaches independently; first use a spatial scan method to identify the existence and approximate location of a diseases cluster, then apply the hot-graph method. To save time, one could forgo a statistical evaluation of the hot-graph cluster, and simply use it as an informal guide to understanding a cluster's geographical shape. The intersect (shared) zones of the two found cluster sets could be used for formal reporting. This would drop the zones not reported by both methods, and assume that they were errors due to over fitting, disagreement between geometry and disease variation, or other causes. However, this process is still reasonably systematic and can be incorporated into a routine analytical activities. Rather than changing any of the reported statistics (like the p-value or relative risk estimate), a second 'adjusted' relative risk estimate could be derived from the intersecting zones. This way analysts could ignore the contribution of the hot-graph approach if so desired.

Data-directed searches like the hot-graph method are assured of finding noteworthy patterns of diseases when local sampling error is low or non-existent. Under these circumstances, however, cluster-detection methods may not be necessary—a visual inspection of a map of rates would often be sufficient. In fact, when variance is small enough, finding the most-likely cluster is somewhat trivial—it is simply the region with the highest rate of disease. Under these circumstances, cluster detection problems could be re-formulated as more general-purpose aids to geographical analysis and spatial classification. In these situations, it may also be necessary to add additional constraints. For example, when working with large amounts of data or common diseases, it may be worthwhile to add an effect-size or population size constraint. These and similar kinds of constraints can add considerable complexity to the cluster detection process. In the future, these problems may be best viewed as having either more than one objective or a series of offsetting constraints. For example, methods could search for the most-likely clusters while simultaneously maximizing population. Alternatively, the population and effect size constraints could be added directly to the cluster search process. These kinds of formulations bring

cluster detection methodology closer to political districting/zoning problems, and may represent an important area of future research.

5.5 Conclusion

Ultimately, the decision to use a circular spatial scan or a data-directed search cannot be resolved by the isolated exploration of simulated or real data. The decision requires an understanding of methodological consequences, purpose (epidemiology, surveillance or resource allocation, for example) and clinical information. The original spatial scan constrains searches to pre-set geometric shapes. Ongoing research continues to expand the geometry that the spatial scan is able to employ. Data-directed methods in general, and the hot-graph method in particular, may, on occasion, identify distinct clusters that the traditional spatial scan misses.

One of the other benefits of applying both methods to cluster detection problems is that the comparison provides information on the robustness of the results. Ideally, a single series of simulations are undertaken to test the hypothesis that risk is elevated in the most-likely cluster. Alternatively, one could focus on the inferential information of the spatial scan, and use the hot-graph approach as an informal indicator that a structured (circular, elliptical, square, linear) process may best define the distribution of disease. When there is agreement between the methods, this is a strong indication that the disease follows the pre-determined structural pattern. When the patterns differ significantly, it may be a clue that the geography of the disease is more complex—that several adjacent (but independent) clusters may be present, or that the underlying factors causing spatial variation in disease have an interesting geographic form.

Table 5.1 Selected diseases

Disease	*ICD-9 Codes*	*Crude Rate/10 000*	*Age at Risk*
Asthma	493	629.78	0 - 20
Diabetes	350	354.77	All
Hypertension	401-405	981.96	All
Illness related to alcohol/drugs	303-305	89.90	All
Parkinson's	320	43.19	65+
Influenza	487	111.62	All
Food poisoning (salmonella)	003	0.75	All
Food poisoning (other, bacterial)	005	1.93	All
Giardiasis	007.1	0.43	All
Gonorrhea	098	1.81	All

Table 5.2 Tabulation of cluster detection results

Disease	*Hot-Graph*			*Circular Spatial Scan*			*HG & SS*	*HG not SS*	*SS not HG*	*not HG nor SS*
	N	*RR*	*P (<=)*	*N*	*RR*	*P (<=)*				
Asthma	7	1.23	0.001	9	1.19	0.001	7	0	2	59
Diabetes	24	1.21	0.001	8	1.22	0.001	6	18	2	42
Hypertension	27	1.14	0.001	19	1.10	0.001	14	13	5	36
Illness related to Alcohol/drugs	22	1.49	0.001	1	2.40	0.001	1	21	0	46
Parkinson's	24	1.23	0.001	9	1.34	0.001	8	16	1	43
Influenza	18	1.4	0.001	1	2.13	0.001	1	17	0	50
Food poisoning (salmonella)	25	1.59	0.074	1	6.06	0.001	1	24	0	43
Food poisoning (other, bacterial)	14	1.77	0.002	6	2.44	0.001	6	8	0	54
Giardiasis	9	2.36	0.640	2	2.32	0.083	2	7	0	59
Gonorrhoea	2	13.55	0.001	1	15.66	0.001	1	1	0	66

Figure 5.1 Province of Alberta and major communities

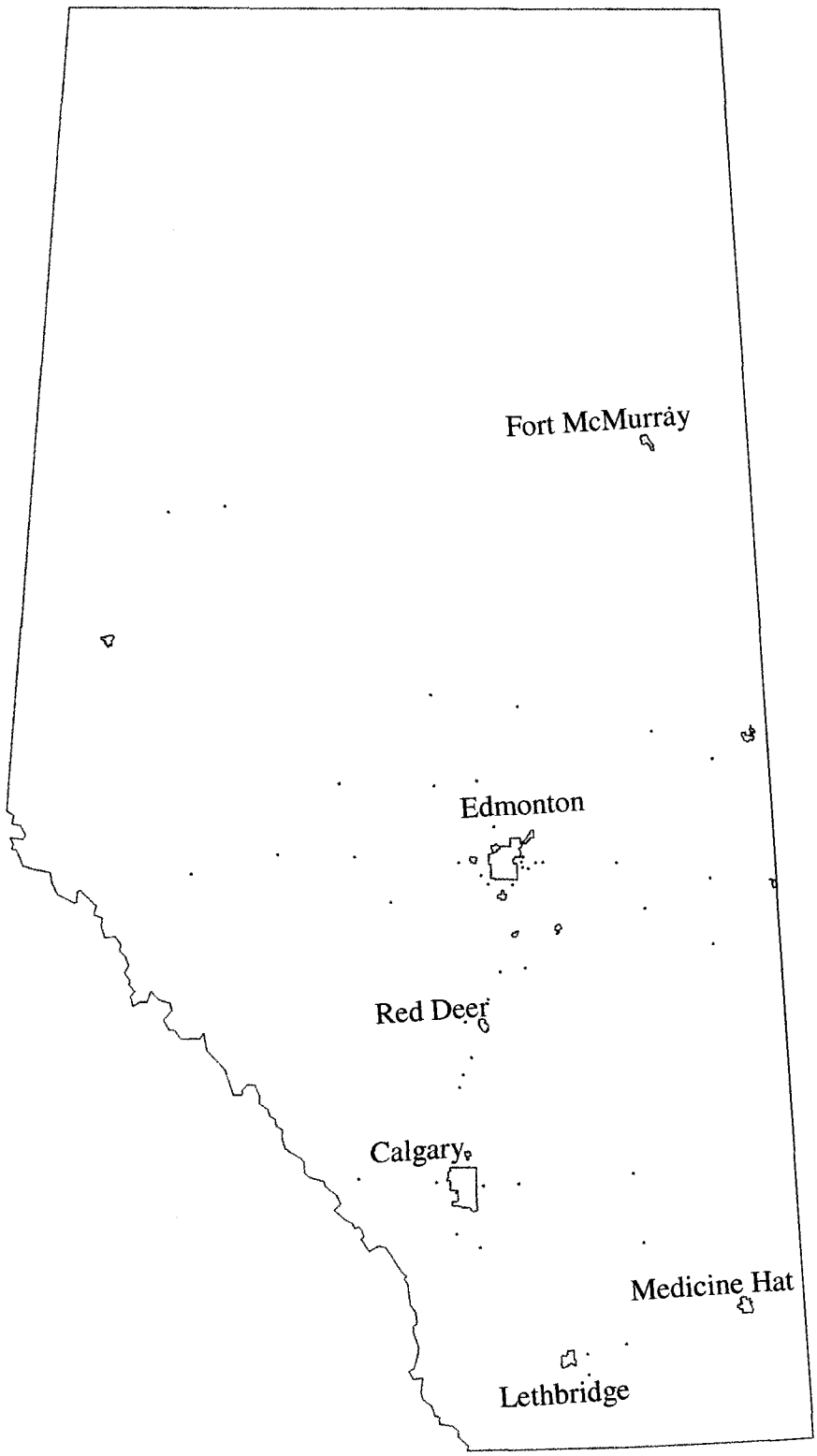


Figure 5.2 Bounding-box centroids and sub-rha boundaries

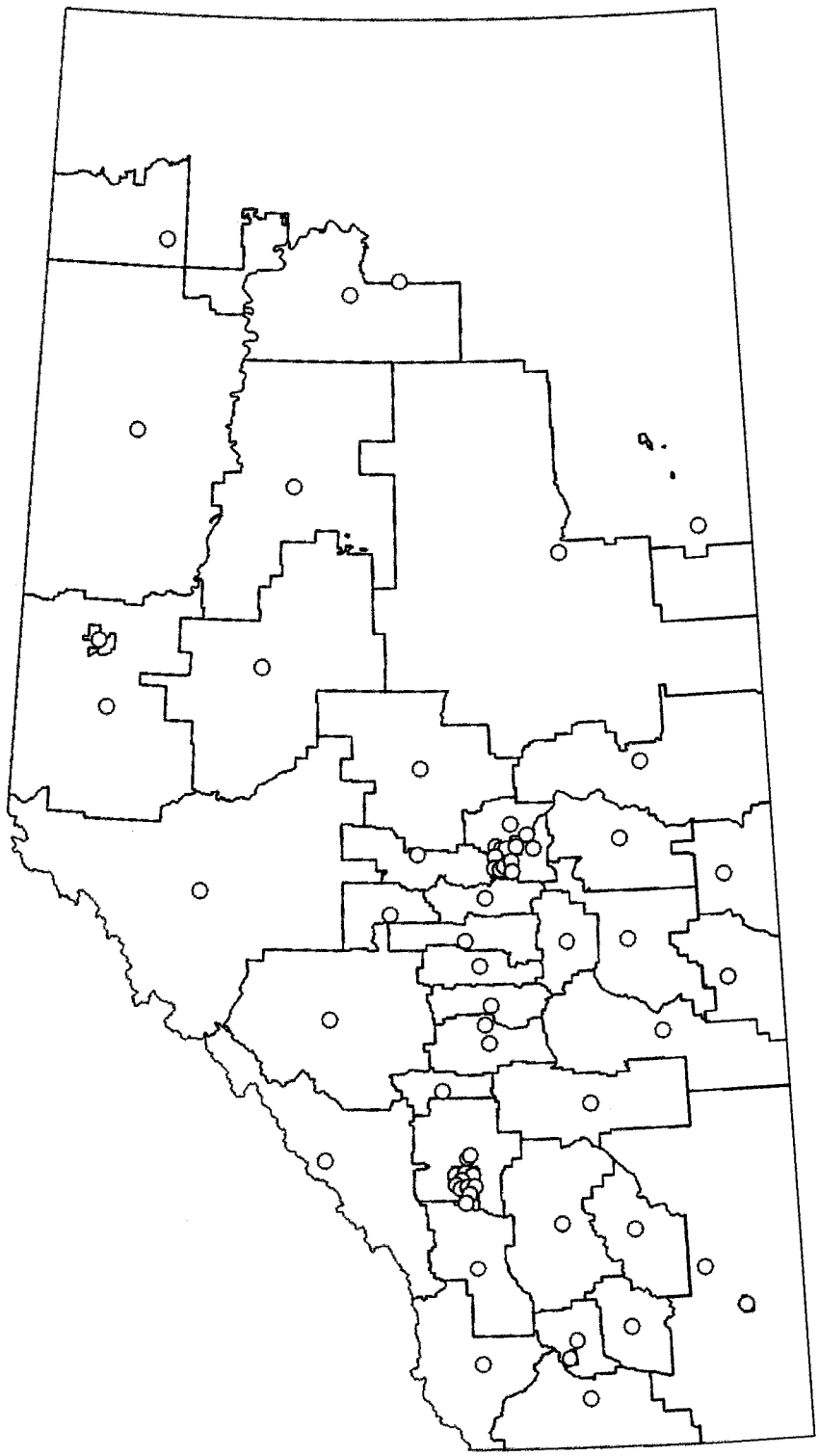


Figure 5.3 Maps of identified clusters: asthma

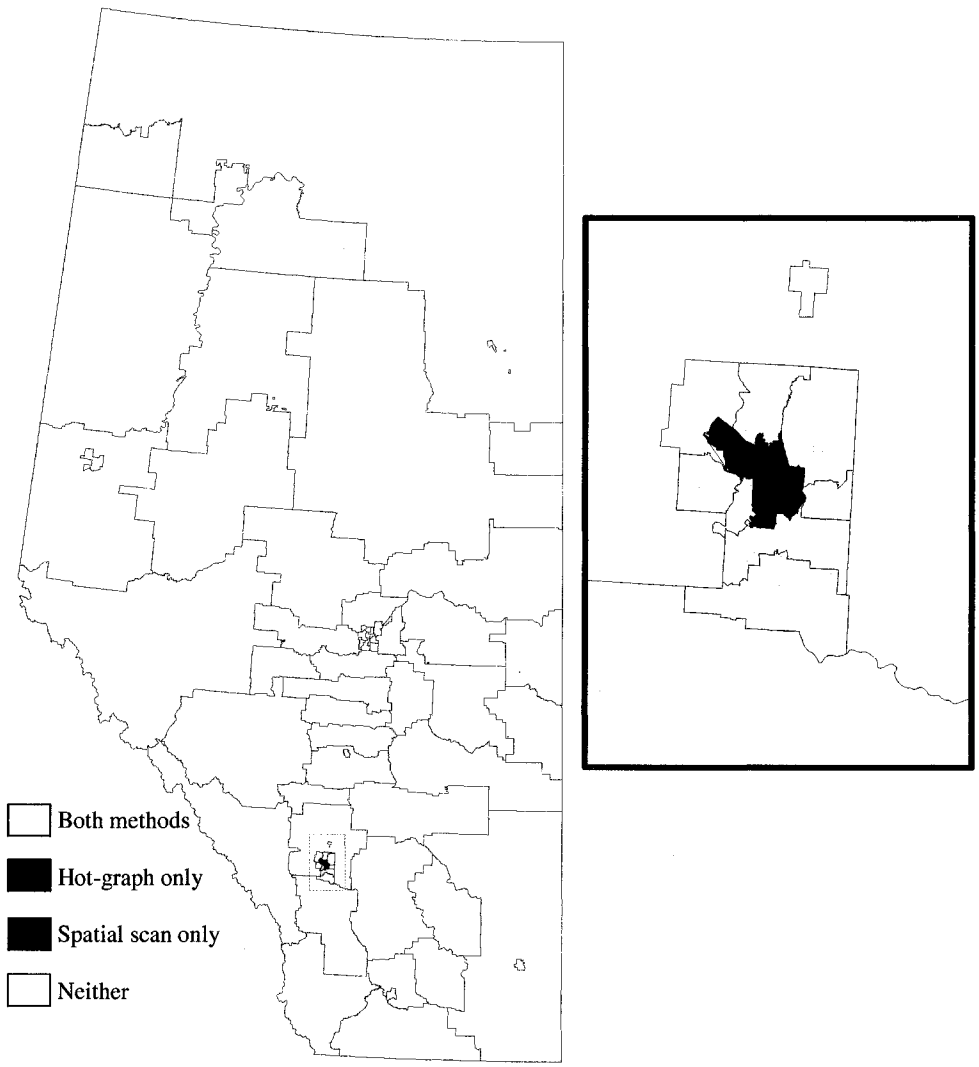


Figure 5.4 Maps of identified clusters: food poisoning (other, bacterial)

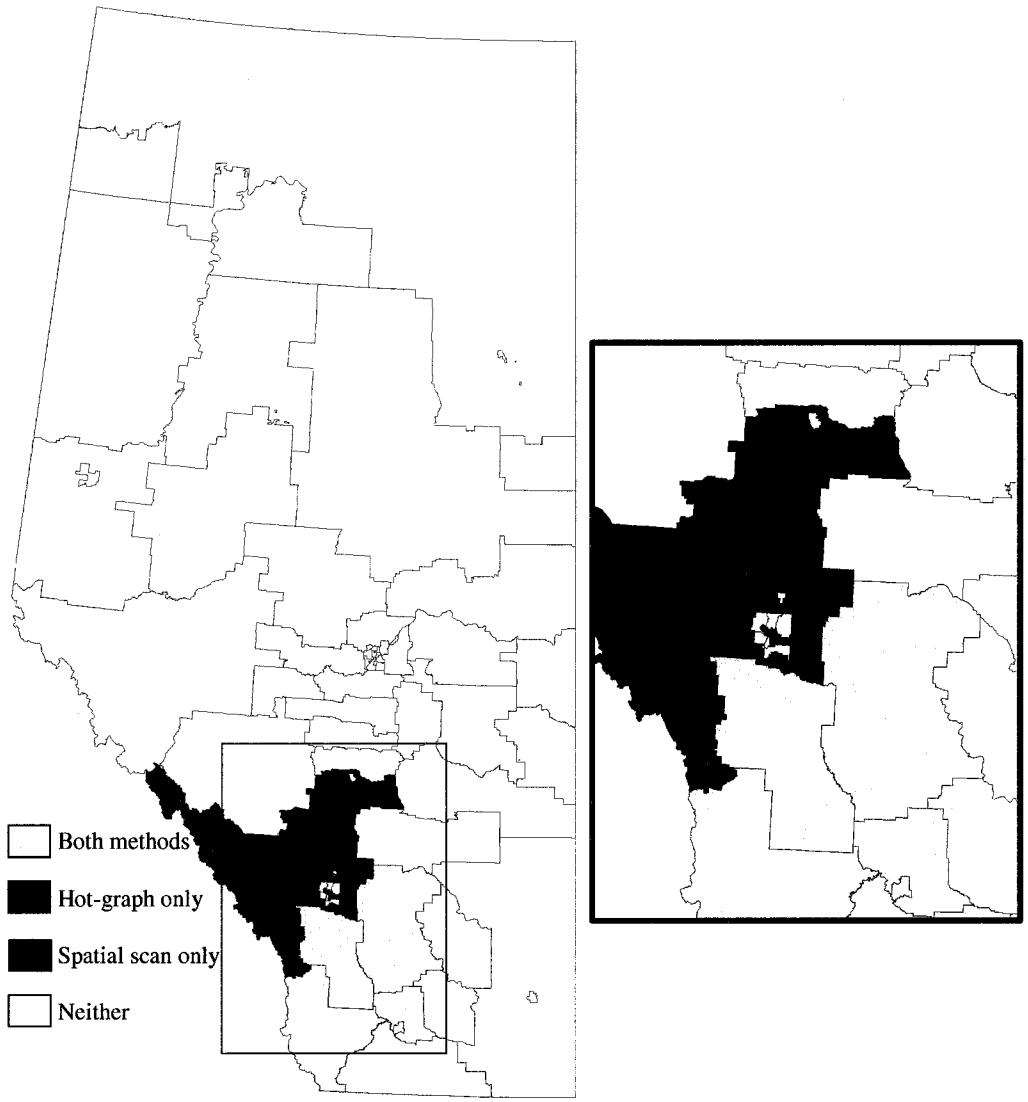


Figure 5.5 Maps of identified clusters: diabetes

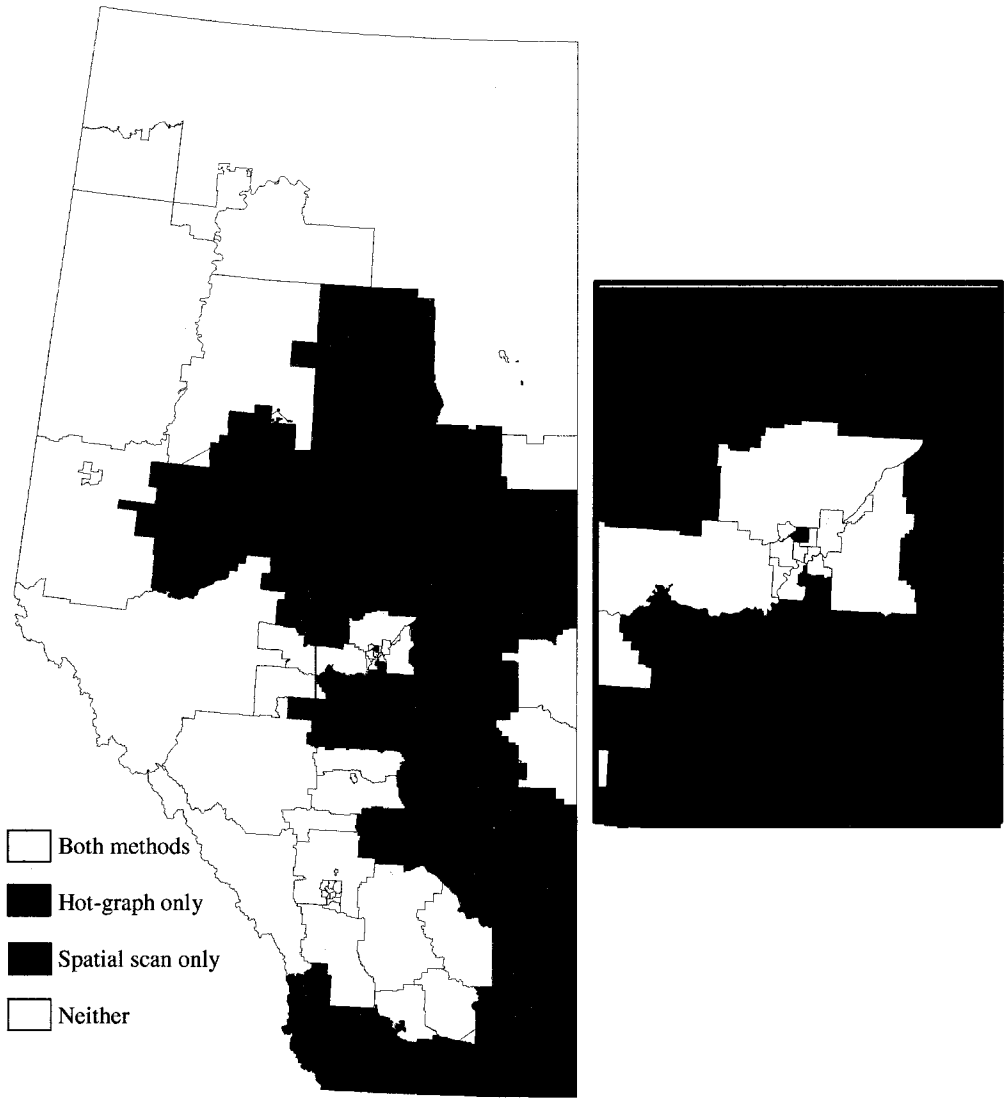


Figure 5.6 Maps of identified clusters: hypertension

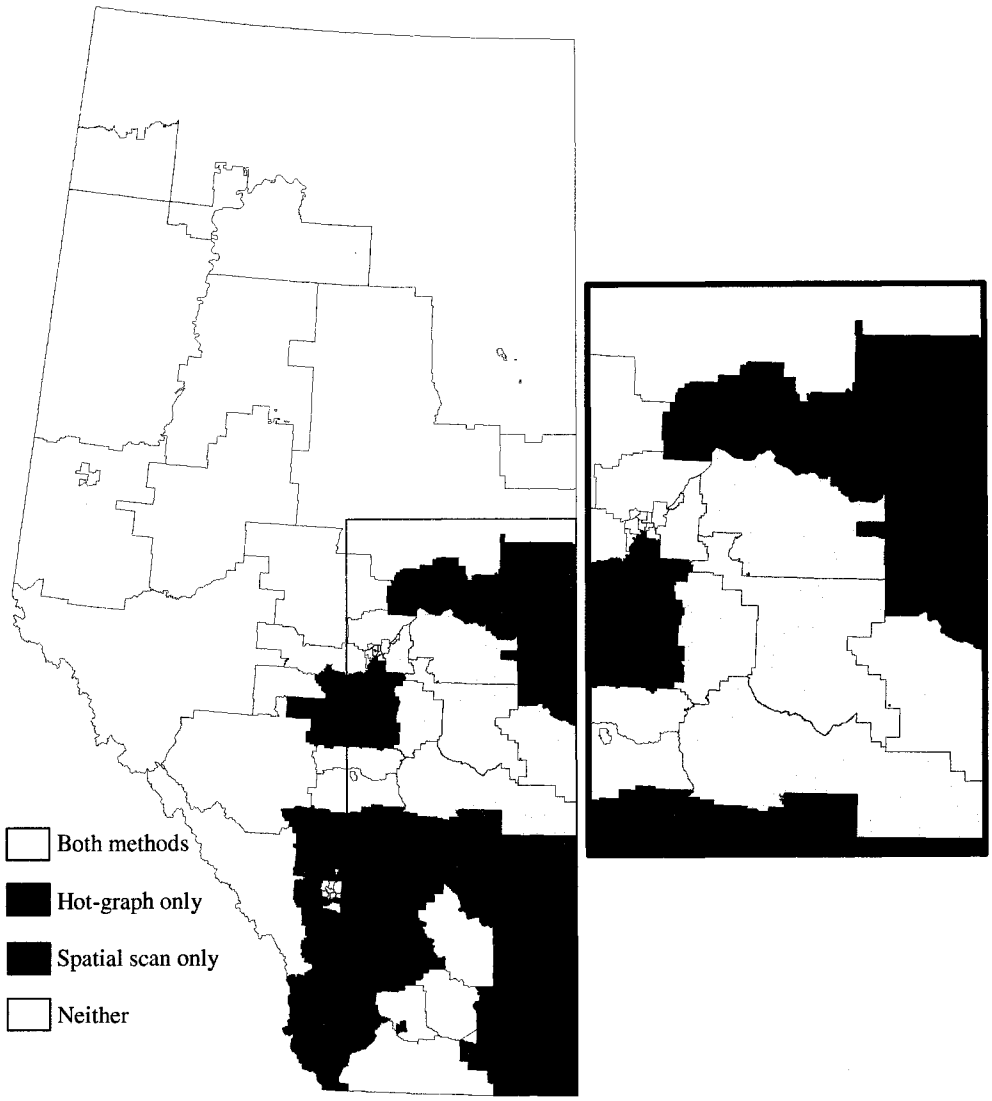


Figure 5.7 Maps of identified clusters: Parkinson's

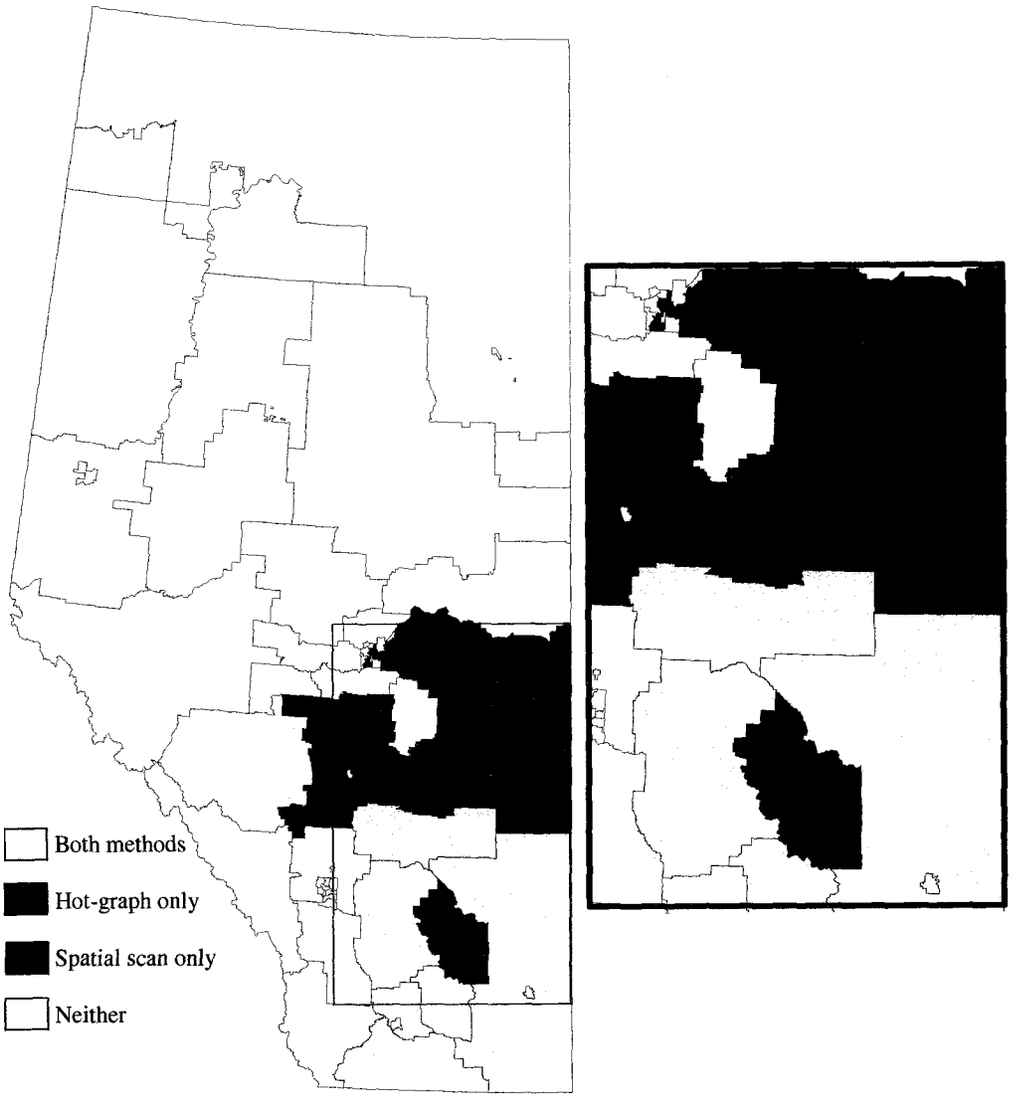


Figure 5.8 Maps of identified clusters: influenza

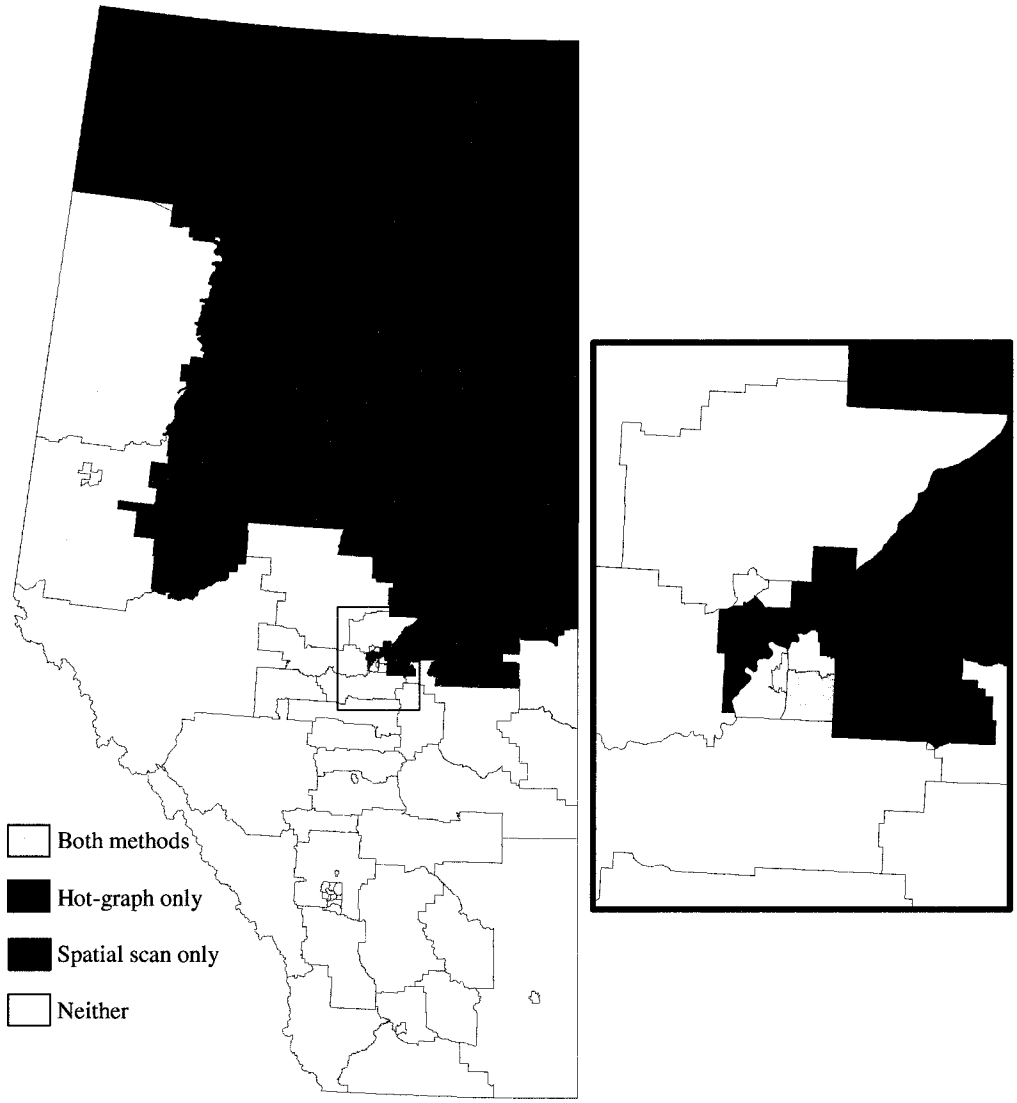


Figure 5.9 Maps of identified clusters: illness related to alcohol/drug use

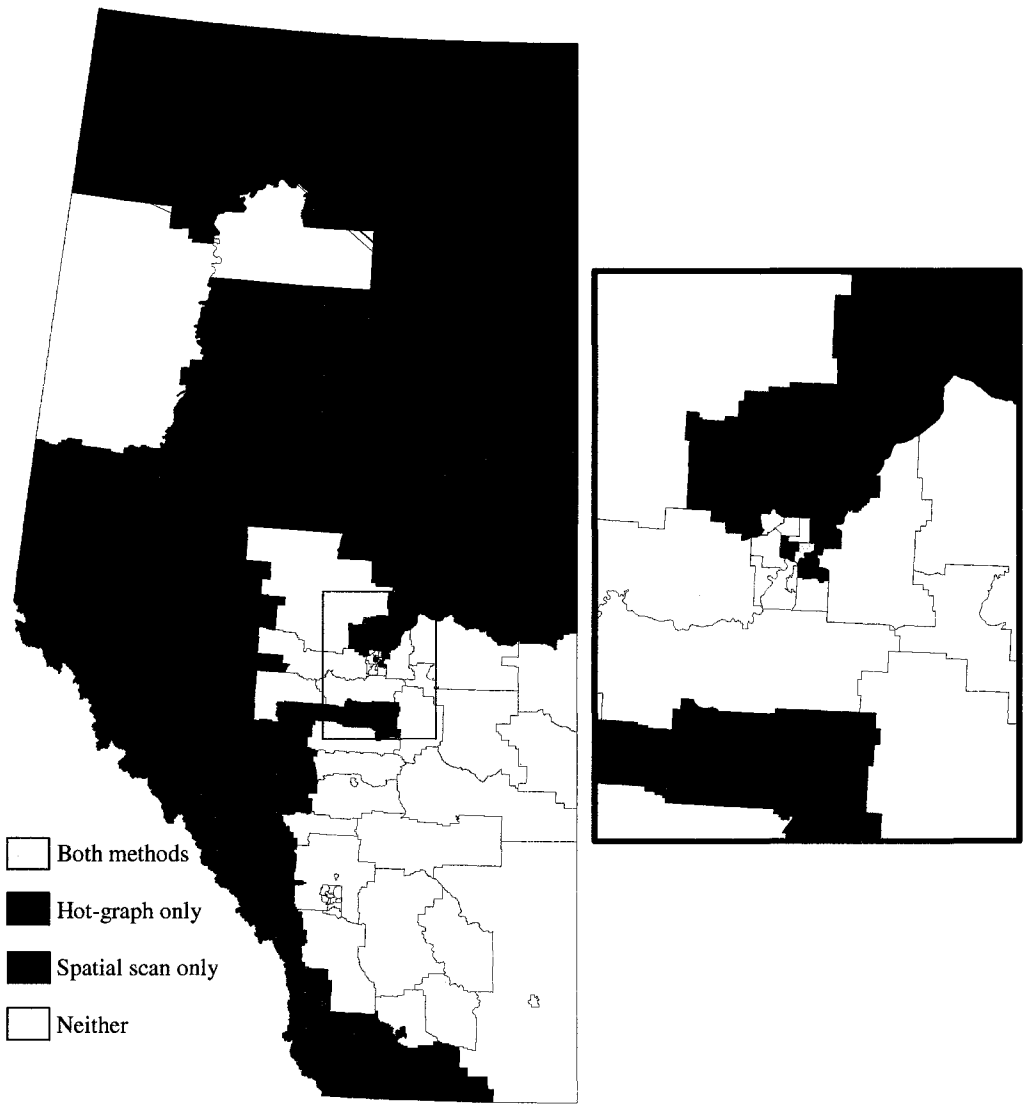


Figure 5.10 Maps of identified clusters: food poisoning (salmonella)

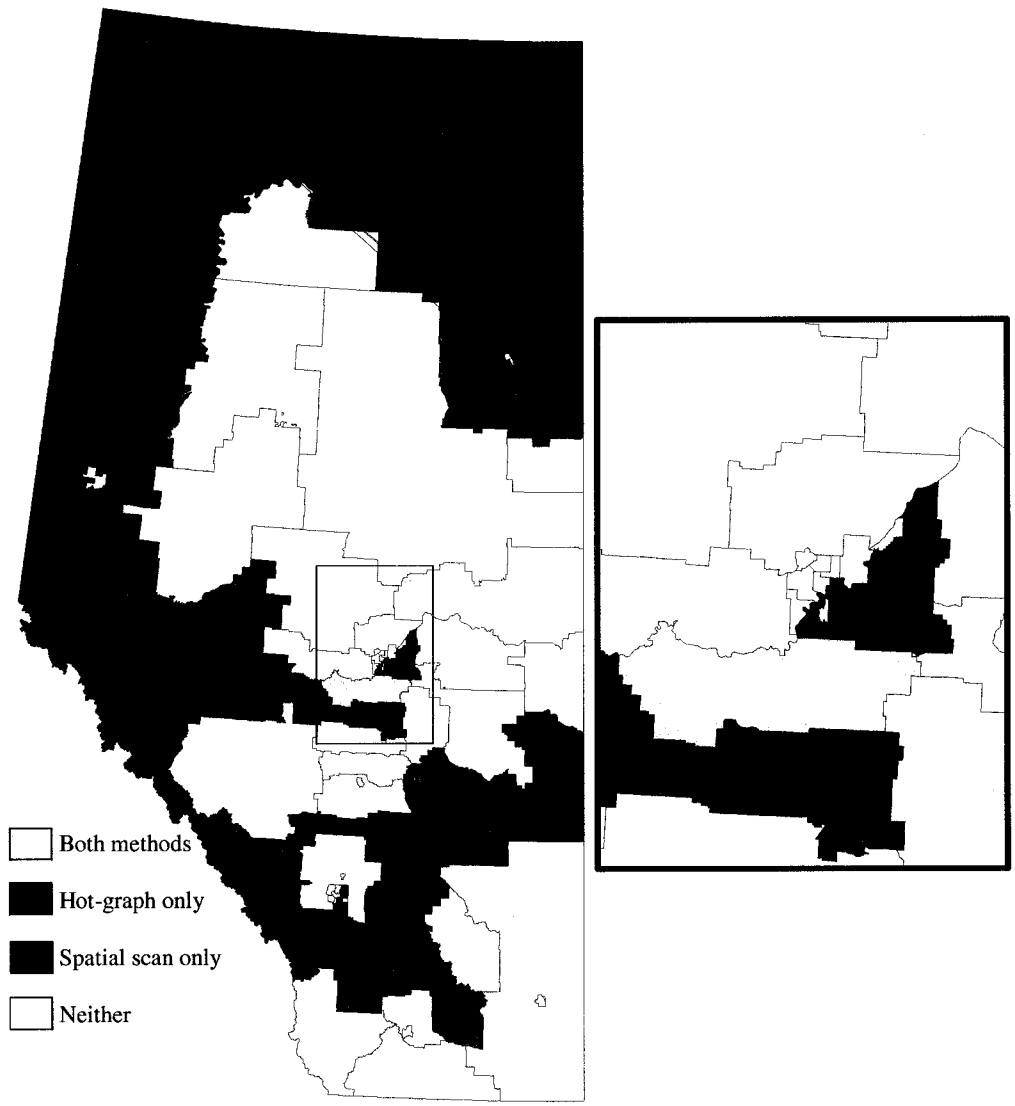


Figure 5.11 Maps of identified clusters: giardiasis

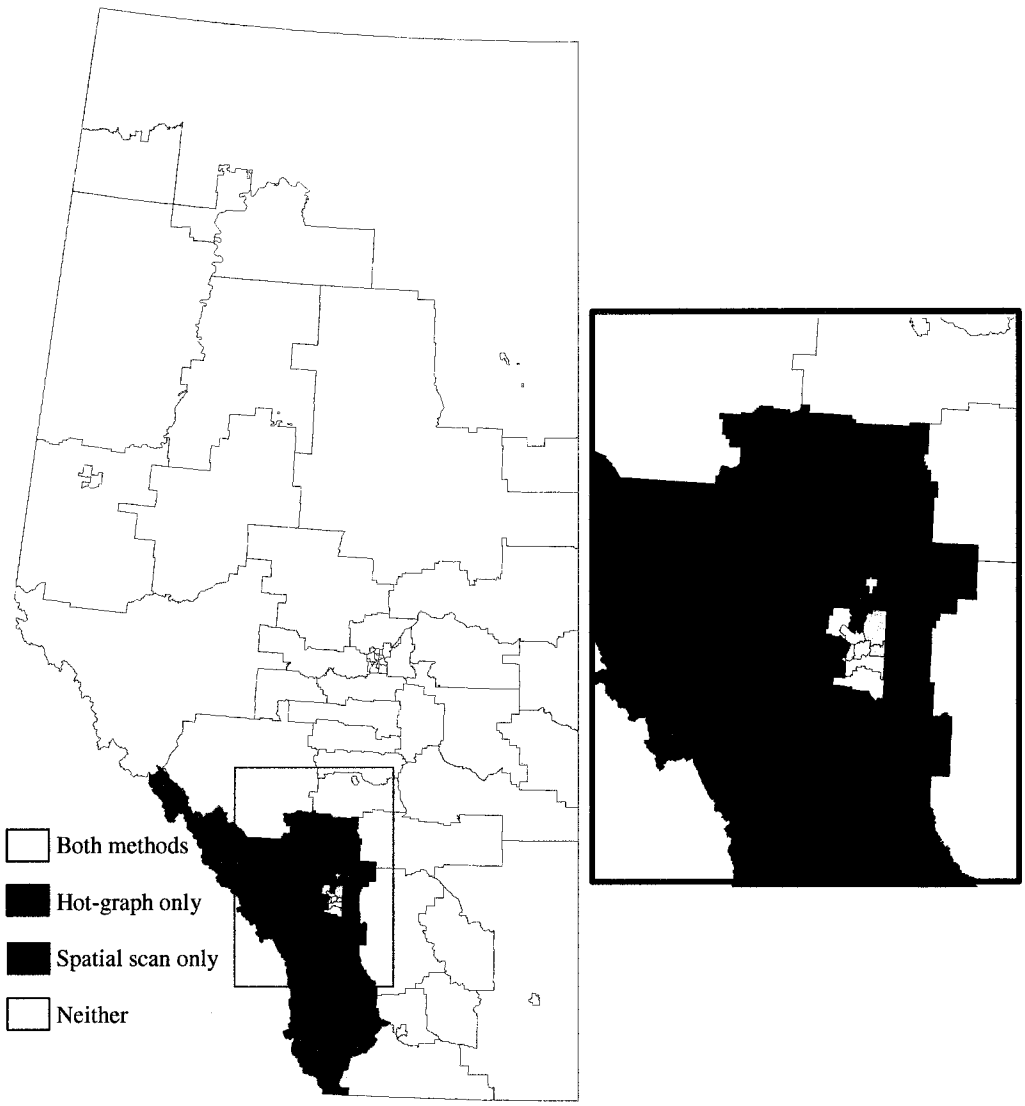
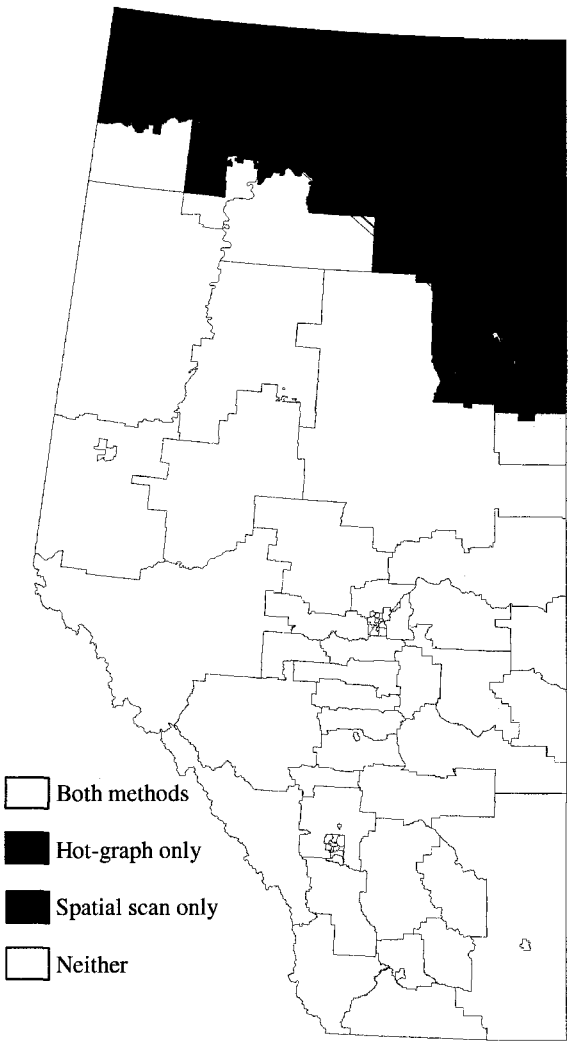


Figure 5.12 Maps of identified clusters: gonorrhoea



5.6 References

Aamodt G, Samuelsen S O, Skrondal A (2006) A simulation study of three methods for detecting disease clusters. *International Journal of Health Geographics. In Press*

Alberta Health and Wellness (2004). Calculating Small Area Analysis Definition of Sub-regional Geographic Units in Alberta Geographic Methodology Series No. 5. Alberta Health and Wellness: Edmonton.

Conley J, Gahegan M, Macgill J (2005) A genetic approach to detecting clusters in point data sets. *Geographical Analysis* **37**:286-314.

Cuzick J, Edwards R (1990) Spatial clustering for inhomogenous populations. *Journal of the Royal Statistical Society B* **52**:73-104.

Duczmal L Assunção R (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistics & Data Analysis **45**:269-286.

Gangnon R E, Clayton M K (2001) A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine* **20**:2977-2987.

Gangnon R E, Clayton M K (2003) A hierarchical model for spatially clustered disease rates. Statistics in Medicine **22**:3213-3228.

Gangnon R E, Clayton M K (2004) Likelihood-based tests for localized spatial clustering of disease. Environmetrics **15**:797-810.

Goovaerts P, Jacquez G M (2004) Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical

filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics* **3**:1-23.

Hill E G, Ding L, Waller L A (2000) A comparison of three tests to detect general clustering of a rare disease in Santa Clara County, California. *Statistics in Medicine* **19**:1363-1378.

Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. *Statistics in Medicine* **14**:799-810.

Kulldorff M (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods* **26**:1481-1496.

Kulldorff M, Tango T, Park P J (2003) Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* **42**:665-684.

Kulldorff M, Information Management Services, Inc. (2004) SaTScan™ v5.0: Software for the spatial and space-time scan statistics. http://www.satscan.org/.

Kulldorff M (2004). SaTScan™ user guide. http://www.satscan.org/

Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine.* **2**:216-224

Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. *Statistics in Medicine In Press*

Le N D, Petkau A J, Rosychuk R (1996) Surveillance of clustering near point sources. *Statistics in Medicine* **15**:727-40.

Mostashari F, Kulldorff M, Jartman J J, Miller J R, Kulasekera V (2003) Dead bird clusters as an early warnig system for West Nile virus activity. *Emerging Infectious Diseases* **9**:641-646.

Muhajarine N, Mustard C, Roos L L, Young T K, Gelskey D E (1997) Comparison of survey and physician claims data for detecting hypertension. *Journal of Clinical Epidemiology* **50**:711-718.

Neill D B, Moore A W (2005) Efficient scan statistic computations. *Spatial and Syndromic Surveillance for Public Health.* Lawson A B, Kleinman A B (eds.) Wiley: West Sussex.

Patil G P, Taille C (2003) Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science* **18**:457-465.

Patil G P, Taille C (2004) Upper level set scan statistic for detecting arbitrary shaped hotspots. *Environmental and Ecological Statistics* **11**:183-197.

Puett R C, Lawson A B, Clark A B, Aldrich T E, Porter D E, Feigley C E, Hebert J R (2005) Scale and shape issues in focused cluster power for count data. *International Journal of Health Geographics* **4**:1-16

Rogerson P A (1997). Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine* **16**:2081-2093.

Roos L L, Sharp S M, Cohen M M (1991) Comparing clinical information with claims data: some similarities and differences. *Journal of Clinical Epidemiology* **44**:881-888

Roos L L, Mustard C A, Nicol J P, McLerran D F, Malenka D J, Young K, Cohen M M (1993) Registries and administrative data: organization and accuracy. *Medical Care* **31**:201-212

SAS Institute (1999) SAS 8.2. SAS Institute Incorporated: Cary

Song C, Kulldorff M (2003) Power evaluation of disease clustering tests. *International Journal of Health Geographics* **2**:1-8

Tango T (2000) A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine* **19**:191-204.

Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* **4**:1-15

Turnbull B W, Iwano E J, Burnett W S, Howe H L, Clark L C. (1990) Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* **132**:S136-S143.

Waller L A, Turnbull B W (1993) The effects of scale on tests for disease clustering. *Statistics in Medicine* **12**:1869-1884.

Waller L A, Gotway C A (2004) *Applied Spatial Statistics for Public Health Data.* Wiley: Hoboken

Whittemore A S, Friend N, Brown B W, Holly E A (1987). A test to detect clusters of disease. *Biometrika* **74**:631-635.

Yiannakoulias N, Svenson L W, Schopflocher D P (2005). Diagnostic uncertainty and medical geography: what are we mapping? *The Canadian Geographer* **49**:291-300.

CHAPTER 6: Conclusion

6.1 Summary of chapters

6.1.1 A review of geographic health surveillance methodology (Chapter 2)

Though much modern quantitative geographic health research is concerned with theory building and explanation, geographic methodology is important in the growing field of health surveillance. I review a number of analytical methods in the context of four dimensions of geographic health surveillance: identifying new risks, providing information for public health policy decisions, providing information to communities, and acting as an alarm system to detect rapid change in disease patterns. Many methods are available to analyse geographic health data, but not all apply to these surveillance goals. In particular, there is a shortage of discussion about how different methods contribute to policy and community support in public health surveillance. I argue that cluster detection methods are particularly valuable for timely and routine information dissemination to policy makers and the public.

6.1.2 Clusters as unfairness: geographic analysis in chronic disease prevention (Chapter 3)

In this chapter, I argue that geographic cluster detection methods have value in the formulation and implementation of chronic disease prevention strategies. Geoffrey Rose's thesis—that population-wide strategies maximise the effectiveness of disease prevention programs—assumes that decision-makers operate in a utilitarian decision framework. I argue that cluster detection methods are more applicable to policies concerned with 'worst-off' geographic groups, a practice which conforms more closely to the 'difference principle' discussed by John Rawls. When applied to some chronic diseases, cluster detection methods can be used to guide prevention and intervention efforts to areas at high risk and with high burdens of illness. Geographic 'cluster-communities' of worst-off health outcomes represent a potential focus of geographic health interventions—like environmental modification, service facility location and community building.

6.1.3 Structured and data-directed cluster search strategies (Chapter 4)

This chapter considers how two different cluster search methods perform for a single method of statistical evaluation. Historically, most search methods employ a fixed structural constraint—such as a search of nearest neighbours or a search for clusters of a pre-specified shape—to find clusters of disease. Some of the most recently developed methods—termed 'data-directed' approaches—use search strategies led by attributes of the data. In these cases, search algorithms are bound by adjacency constraints, and hunt for clusters that best meet criteria related to properties of the data. I present an approach I call the 'hot-graph method' and compare it with the moving circular window strategy used by the traditional spatial scan method. I evaluate these methods two different ways. First, I test their power to detect the existence of simulated clusters. Second, I compare the methods' abilities to correctly identify the constituent areas of these simulated clusters. Results suggest a similar ability to detect the presence and approximate location of a cluster, but that the spatial scan performs better for all cluster patterns under study. However, the hot-graph method is better able to precisely define the shape of a cluster when the cluster is of an irregular shape (such as a ring, line or network). For surveillance purposes, I recommend incorporating both of these methods into a cluster detection system. I suggest that differences related to resolution, scale, relative risk, population size and frequency of disease impose a significant challenge to developing a single multi-purpose cluster detection strategy.

6.1.4 Geographic discovery through cluster detection (Chapter 5)

In this chapter I compare the performance of the traditional spatial scan and the hot-graph approach on ten different illnesses in the province of Alberta. Although results vary, they are consistent with the observations made with simulated data; the spatial scan seems to be the better tool for inferential disease cluster detection, particularly for rare diseases. However, the hot-graph method provides interesting information about the geography of disease. Analyses that combine a structured search of geometric shapes—like circles, ellipses and rectangles—still benefit from a comparative analysis with data-directed cluster detection methods such as the hot-graph approach.

6.2 Discussion of the research project

The chief goal of this project was to explore aspects of cluster detection methodology with respect to two particular contexts: geographic surveillance and chronic disease prevention. The emphasis in Chapters 4 and 5 was on a specific methodological issue, and the conclusions apply to two general classes of cluster detection methodology—structural methods (which test inferences according to pre-specified geometric constraints) and data-directed methods—which actively search for high clusters of disease subject to an adjacency constraint. I now discuss how these methods may be relevant to geographic disease surveillance and chronic disease prevention.

Methods that constrain searches for clusters within pre-specified geometries are sensible tools for inferential work, especially when disease is rare, and there is no reason to believe that the structural constraints are false. In fact, the geometric structures are an important theoretical contribution to the search process that can improve statistical power when these geometries approximately reflect the spatial characteristics of disease. In the case of very rare diseases, it appears that correct structural constraints can make a big difference in the ability of a method to correctly identify the presence and approximate location of a true cluster. For routine surveillance in infectious and chronic disease, these methods are an excellent starting point, especially in light of new work broadening the geometric structures under investigation to include ellipses (Conley, Gahegan and Macgill 2005; Kulldorff et al., 2006). Under these circumstances, failure to identify the exact geography of a disease cluster is probably not of major concern. Methods that are simple, fast, can manage large amounts of data, and report information that is easy to synthesize are effective for most purposes in health research and many applications in disease surveillance.

On the other hand, when the purpose of cluster detection is to identify cluster-communities that experience a high burden of chronic disease, it may be more important to understand geography precisely. For rare diseases, there are limits to which such geographies can be defined; based on the experimental results of Chapter 4, methods that constrain geometric structure ahead of time probably

have the best power to detect clusters of rare disease. For diseases that are more common, the hot-graph approach appears better able to identify the complete geographies of high disease burden areas. In this application, the hot-graph method is able to find cluster-communities of highest disease burden, and facilitate the allocation of intervention resources that are geographically specific. Given the tendency of the hot-graph method to include false-positive areas, the cluster-communities detected by this method may be systematically larger than they should be. However, this may be a reasonable trade-off, since the harm of failing to locate an area with a truly anomalous disease burden is, from a policy standpoint, probably greater than the harm of including a few false positive areas. When a cluster-community has a rate of disease that is statistically but not clinically significant, the cluster-community may take on irregular shapes—as seen in Chapter 5. This is itself informative, however. When the hot-graph method finds large irregular clusters, it may indicate that spatial variations in disease are small, and that perhaps no noteworthy cluster-community exists. It may also suggest that a different evaluation procedure is necessary; when statistical variability is small or non-existent measures that quantify magnitude of difference (like disease rates) or the population at risk may be more suitable criteria for the search process. In these cases, the evaluation process becomes an operational problem, and mathematical optimization techniques might be warranted.

The hot-graph and other data-directed methods can also be used to interpret the results of structural methods in a way useful to understanding the geography of disease. When there is agreement between the spatial scan and hot-graph approaches, analysts should have more faith that the structural constraints underlying the inferences are reasonable; for example, we can be more certain that a cluster found by the spatial scan is circular when a hot-graph search reports a most-likely cluster that is also circular in the same location. For research purposes, this may inform us about previously unknown hazards that persist in these shapes—such as plumes of environmental pollutants. When the methods provide very different results, and in particular, when the hot-graph cluster is

considerably different in size from a cluster detected using the spatial scan, the structural constraint should be scrutinized. In this case, alternative methods may be required to build stronger conclusions about the location and shape of the cluster.

The spatial scan continues to receive considerable methodological and applied research attention. Its relative ease of use, robustness, and recent adaptations (that enable the approach to observe new shapes) may keep it the default cluster detection tool in the foreseeable future. I believe that data-directed methods like the hot-graph approach can be important auxiliary tools to explore disease clusters in conjunction with structured approaches. Data-directed methods have three noteworthy applications. First, on occasion, they may be able to find irregularly shaped clusters that some structured searches are unable to find, or may imprecisely locate. It is hard to know how often this would occur in the real world, but some non-circular/non-elliptical structures have been observed, including core-periphery patterns (Vuorinen 1987), linear shapes (Duczmal and Assunção 2004) and the shape of the asthma cluster observed in Chapter 5.

Second, any time both a structural search and a data-directed search fail to find a significant cluster, we have some assurance that the inability of a structured search to find a cluster is not the result of the structural constraints. The search process guiding the hot-graph method is not confined to a particular geometric shape, and is fairly successful at identifying clusters of irregular shapes when they occur (as seen in Chapter 4). When this method also fails to find a cluster, we have reasonable assurance that the geometric constraints of the structural approach are not to blame. Finally, when a disease is not rare (or spatial resolution is low enough), data-directed methods may be helpful in identifying and demarcating geographic groups of high disease burden. When public health interventions are geographically specific and directed at subsets of worst-off groups, data-directed cluster detection methods offer important information about where these worst-off geographic groups are.

6.3 Future Research

6.3.1 Detection of severity clusters

Most applied work in the spatial analysis of disease classifies illness into discrete groups—persons with and without illness. For some chronic conditions, it may be only the most severe cases that are of notable concern to public health. For example, asthma has a prevalence rate of roughly 5-10% among children under 20 years of age, but fewer than 25% of these experience serious symptoms (Newacheck and Taylor 1992). Similar research of elderly sufferers of chronic lung diseases (such as bronchitis, emphysema and asthma) suggest that roughly 30% suffer serious symptoms (Selim et al. 1997). In these, examples, a large proportion of those diagnosed as sick are not seriously burdened by disease; the majority of the burden is experienced by a minority of sufferers. Health economists observe similar distributions in the expenditure of disease treatment and care; the highest per capita (and sometimes absolute) costs of treatment are often incurred by a minority of sufferers. For example, 40-95% of treatment spending on persons with bipolar disorder is account for by a mere 5% of bipolar sufferers (Simon and Unutzer 1999). Similar cost distributions have been observed in other chronic conditions, such as type 2 diabetes (Brandle et al. 2003), asthma (Smith et al. 1997) and arthritis (Kobelt et al. 1999).

Often clusters of severity occur in age space—for example, persons over 65 usually incur the greatest burden of disease and cost of treatment—but variations in income and ethnicity may be important to understanding variations in disease severity. Geographic analyses that consider disease severity (rather than disease incidence and/or prevalence) may also reveal patterns of illness that are both important and otherwise unknown. They may offer indirect exploratory evidence of ecological associations between covariates like income and severity, but may also reveal otherwise unknown geographic processes. Clusters of severity may also locate regions in high need of intervention—either in treatment or prevention. Administrative health data have been used to define disease severity in the past (Deyo, Cherkin and Ciol 1992; Clark et al. 1995) and have the advantage of supplying numbers large enough to obtain reasonably robust statistical inferences. In Alberta, data on service utilization (in terms of quantity and location), medication use (for person 65 and over) and co-morbid conditions are all

collected and maintained by the provincial ministry of health. Work in conjunction with the ministry and health care professionals may not only reveal interesting geographic patterns of disease severity, but could offer insight into more efficient and even more fair distribution of treatment and prevention resources.

6.3.2 Cluster detection within operational constraints

Much of the work in cluster detection methodology has focused on statistical power and efficiency. This is consistent with one of the most important contributions of statistical methodology in science—based on a sample, provide as much information as possible about the population from which the sample was drawn. Good methods are able to extract precise information about the location, significance and magnitude of clusters even when the data available are highly stochastic—as would often be the case in small-area studies of rare disease.

On occasion, there are also operational aspects of public health management to which cluster detection methodology can contribute. For example, health policy makers often make decisions about resource allocation that may benefit from the kind of information that cluster detection methods provide. Geography-specific resource planning (such as deciding the location of a support facility or modifying the physical environment) can always benefit from knowing where disease clusters are located. Clusters represent regions where local intervention could offer the greatest benefit. However, policy decisions are often bound by practical matters that make some cluster detection algorithms inadequate. For example, it may be important to find clusters of disease that are high, but also include a sufficiently large population. Population can be seen as a constraint in the cluster detection process—all clusters must be of a minimum specified size—or as a secondary objective in a multiple objective optimization exercise—a search for clusters that have anomalously high rates and large populations simultaneously. It may also be important to set a lower limit on any detected cluster's incidence rate; any cluster below a certain threshold of magnitude may not be of policy relevance.

These kinds of criteria can be incorporated into a detection procedure in an *ad hoc* manner. For example, simply ignore all detected clusters that fail to meet a particular population threshold. However, this may result in unnecessarily inferior solutions. For example, two neighbouring clusters that independently fail to meet the population threshold could be combined to meet that same criterion. By directly incorporating the population threshold into the search process, it is more likely that all the criteria will be met simultaneously, and more likely that the cluster identified is truly the most-noteworthy—subject to the various input constraints. Future work in this area could blend methods of cluster detection with methods of political districting (e.g., Bozkaya, Erkut, Laporte 2003); methodology could combine the anomaly finding features of cluster detection methods with the constraint management and optimization of political districting.

6.3.3 Cluster detection and the modifiable areal units problem

Decades of research on the modifiable areal unit problem (MAUP) (Openshaw 1984) have consistently revealed that most methodologies applied to spatial data are not robust to how those data are represented geographically. Analyses of a single data set aggregated to different 'levels' of resolution (for example, individual-level, family-level, neighbourhood-level, city-level) are likely to produce different results at each of these levels. Even at a single level of resolution, different zoning or grouping methods can greatly affect the behaviour of aggregate statistics.

Though the ultimate answers to the MAUP may be more philosophical than methodological (e.g. Smith and Mark 2001) it is important to explore the effects of MAUP on cluster detection algorithms. The behaviour of some statistical measures are well discussed with respect to the MAUP (e.g., Fotheringham and Wong 1991; Amrhein and Reynolds 1996) but other methodologies, like cluster detection, have been neglected. Some work has considered the role of resolution on cluster detection algorithms (Waller and Turnbull 1993), but this does not incorporate recent progress in methodology, nor does it account for issues of zoning. The MAUP is a product of spatially dependent processes like diffusion and common local influences (Holt, Steel and Tranmer and Wrigley 1996).

Cluster detection algorithms search for areas with anomalously high or low incidence of disease, which are often caused by the forms of spatial dependence underlying the MAUP. Therefore, it seems likely that the MAUP would affect cluster detection algorithms, and that not all methods would respond to different resolution and zoning schemes in the same way. Future work could reveal systematic differences in performance of methodologies with respect to the MAUP, and may also describe the degree to which MAUP should be of concern in geographic decision support.

6.3.4 Comparing statistical and cognitive heuristic approaches to cluster detection

Recent work in criminology suggests that decision makers can often use simple and easy to understand rules to make efficient and informed interpretations of spatial phenomena (Snook et al., 2005). For example, when taught basic principles of geometry and distance decay, people seem able to predict the residences of serial criminals as accurately as formal statistical methods (Snook, Canter and Bennell 2002; Snook, Taylor and Bennell 2004). Though systematic errors in judgement and judgement biases may persist (Arkes 1991; Arkes and Ayton 1999), this kind of research indicates that individuals may be able to understand the important features of space and geography without sophisticated methodology, or even the assistance of computers.

Most steps in the evolution of cluster detection methodology have involved increases in conceptual and computational complexity. Some of these methods require specialised training, specialised software, and in extreme cases, a methodological rationale that is inaccessible to all but the developers and a few of their colleagues. In cases where cluster detection methods are required for decision making purposes, small gains in statistical power and/or precision may not make up for the costs associated with these complex methodologies. Future work in cluster detection methodology should consider comparing the relative success of different techniques in conjunction with assessments of complexity. Formal experimental methods may be useful. For example, subjects are assigned tasks of manually finding clusters of spatial phenomena before and after training, and with and without the use of various formal statistical methods. These

solutions are then compared and evaluated in terms of training time, execution time, and solution quality. If application maintenance concerns are an issue (for example, in cases where the methodologies are packaged within large software systems) then 'software economics' might also be evaluated.

6.4 References

Arkes H R (1991) Costs and benefits of judgement errors: implications for debiasing. *Psychological Bulletin* **110**:486-498.

Arkes H R, Ayton P (1999) The sunk cost and concorde effects: are humans less rational than lower animals? *Psychological Bulletin* **125**:591-600.

Amrhein C G, Reynolds H (1996) Using spatial statistics to assess aggregation effects *Geographical Systems* **3**:143-158

Bozkaya B, Erkut E, Laporte G A, 2003 Tabu Search Heuristic and Adaptive Memory Procedure for Political Districting *European Journal of Operational Research* **144**:12-26.

Brandle M, Zhou H, Smith B R K, Marriott D, Burke R, Tabaei B, Brown M B, Herman W N (2003) The direct medical cost of type 2 diabetes. *Diabetes Care* **26**:2300-2304.

Clark D O, Vonkorff M, Saunders K, Baluch W M, Simon G E (1995) A chronic disease score with empirically derived weights. *Medical Care* **33**:783-795.

Conley J, Gahegan M, Macgill J (2005) A Genetic Approach to Detecting Clusters in Point Data Sets. *Geographical Analysis* **37**:286-314.

Deyo R A, Cherkin D C, Ciol M A (1992) Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *Journal of Clinical Epidemiology* **45**:613-619.

Duczmal L Assunção R (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* **45**:269-286.

Fotheringham, A S and Wong D W S (1991) The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning, A.* **23**:1025-1045.

Holt D, Steel D G, Tranmer M, Wrigley N (1996). Aggregation and ecological effects in geographically based data *Geographical Analysis* **28**:243-261.

Kobelt G, Eberhardt K, Jonsson L, Jonsson B (1999) Economic consequences of the progression of rhumatoid arthritis in Sweden. *Arthiritis and Rheumatism* **42**:347-356.

Kulldorff M, Huang L, Pickel L, Duczmal L (2006) An elliptic spatial scan statistic. *Statistics in Medicine. In Press*

Newacheck P W, Taylor W R (1992) Childhood chronic illness: prevalence, severity and impact. *American Journal of Public Health* **82**:364-371.

Openshaw S (1984) *The Modifiable Areal Unit problem* CATMOG 38, (Geo-abstracts, Norwich)

Selim A J, Ren X S, Fincke G, Rogers W, Lee A, Kazis L (1997) A symptom-based measure of severity of chronic lung disease. *Chest* **111**:1607-1614.
Simon G E, Unutzer J (1999) Health care utilization and costs among patients treated for bipolar disorder in an insured population. *Psychiatric Services* **50**:1303-1308.

Smith B, Mark D M (2001) Geographical categories: an ontological investigation. International Journal of Geographic Information Science **15**:591-612.

Smith D H, Malone D C, Lawson K A, Okamoto L J, Battista C, Saunders W B (1997) A national estimate of the economic costs of asthma. *American Journal of Respiratory Critical Care Medicine* **156**:787-793.

Snook B, Canter D V, Bennell C (2002) Predicting the home location of serial offenders: a preliminary comparison of the accuracy of human judges with a geographic profiling system. *Behavioural Sciences and the Law* **20**:1-10.

Snook B, Taylor B J, Bennell C (2004) Geographic profiling: the fast, frugal and accurate way. *Applied Cognitive Psychology* **18**:105-121.

Snook B, Zito M, Bennell C, Taylor P J (2005) On the complexity and accuracy of geographic profiling strategies. *Journal of Quantitative Criminology* **21**:1-26.

Vuorinen H S (1987) Core-periphery differences in infant mortality. *Social Science and Medicine* **24**:659-667.

Waller L A, Turnbull B W (1993) The effects of scale on tests for disease clustering *Statistics in Medicine* **12**:1869-1884.

7. APPENDICES

APPENDIX I Pseudo-code for hot-graph search algorithm

centroids = number of centroids in graph

hot_graph = the list of centroids in the current hot graph

edges = the list of centroids adjacent to centroids in **hot_graph**

objective = the objective function (likelihood ratio test)

```
do n=1 to centroids
initialize hot_graph_n and edges_n to empty
    current=n
    do k=1 to centroids (or until a population threshold is met)
            put current into hot_graph_n
            put all neighbours of hot_graph_n into edges_n
            for all j in edges_n
objective=likelihood function applied to hot_graph_n
keep j with highest resulting objective
            end
current=j
    end
    store hot_graph_n
end
```

APPENDIX II. Mean sensitivity and positive predictive value tables for different inside cluster and outside cluster rates

Base rate = 2.5/1000, Cluster rate = 3/1000

	Circle				Line				Network				Ring			
	Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV	
Population	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG
1000	0.283	0.383	0.175	0.155	0.213	0.369	0.153	0.150	0.270	0.332	0.268	0.255	0.251	0.416	0.273	0.250
2000	0.326	0.477	0.199	0.156	0.166	0.397	0.157	0.132	0.292	0.412	0.275	0.257	0.266	0.468	0.273	0.280
3000	0.389	0.534	0.216	0.170	0.249	0.460	0.167	0.151	0.275	0.482	0.294	0.284	0.332	0.467	0.277	0.276
4000	0.400	0.546	0.228	0.176	0.257	0.526	0.163	0.163	0.312	0.512	0.309	0.300	0.287	0.481	0.312	0.285
5000	0.413	0.564	0.237	0.177	0.326	0.534	0.189	0.167	0.325	0.552	0.336	0.310	0.372	0.513	0.323	0.304
6000	0.439	0.607	0.225	0.189	0.299	0.566	0.182	0.175	0.332	0.528	0.343	0.305	0.433	0.558	0.358	0.333
7000	0.407	0.639	0.247	0.190	0.300	0.559	0.201	0.168	0.345	0.570	0.319	0.322	0.395	0.546	0.334	0.326
8000	0.464	0.640	0.245	0.193	0.331	0.591	0.211	0.180	0.408	0.572	0.342	0.331	0.507	0.607	0.351	0.359
9000	0.543	0.676	0.365	0.204	0.354	0.597	0.218	0.183	0.317	0.571	0.321	0.318	0.453	0.628	0.343	0.374
10000	0.516	0.690	0.315	0.208	0.340	0.620	0.199	0.186	0.359	0.605	0.369	0.335	0.431	0.628	0.377	0.375

Base rate = 2.5/1000, Cluster rate = 3.5/1000

	Circle				Line				Network				Ring			
	Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV	
Population	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG
1000	0.401	0.454	0.214	0.178	0.303	0.370	0.180	0.154	0.255	0.396	0.277	0.282	0.356	0.532	0.312	0.316
2000	0.481	0.630	0.259	0.213	0.340	0.519	0.206	0.172	0.377	0.558	0.341	0.343	0.403	0.577	0.343	0.345
3000	0.536	0.641	0.293	0.208	0.343	0.624	0.209	0.198	0.383	0.591	0.341	0.343	0.492	0.618	0.420	0.368
4000	0.607	0.671	0.314	0.210	0.403	0.684	0.251	0.213	0.428	0.644	0.370	0.370	0.588	0.702	0.406	0.420
5000	0.669	0.754	0.403	0.233	0.394	0.726	0.269	0.227	0.435	0.665	0.395	0.381	0.653	0.725	0.462	0.428
6000	0.743	0.753	0.442	0.240	0.421	0.766	0.334	0.232	0.441	0.703	0.394	0.408	0.749	0.774	0.493	0.461
7000	0.789	0.761	0.459	0.240	0.426	0.776	0.337	0.233	0.492	0.739	0.434	0.428	0.726	0.810	0.484	0.490
8000	0.821	0.797	0.519	0.249	0.463	0.781	0.302	0.239	0.553	0.762	0.418	0.436	0.759	0.823	0.497	0.494
9000	0.851	0.817	0.533	0.266	0.417	0.773	0.389	0.226	0.495	0.748	0.453	0.434	0.839	0.842	0.528	0.507
10000	0.797	0.804	0.532	0.262	0.440	0.850	0.372	0.256	0.516	0.793	0.432	0.448	0.859	0.872	0.549	0.533

Base rate = 2.5/1000, Cluster rate = 4/1000

	Circle				Line				Network				Ring			
	Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV	
Population	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG
1000	0.501	0.496	0.242	0.203	0.299	0.466	0.241	0.197	0.392	0.481	0.353	0.352	0.462	0.627	0.383	0.375
2000	0.687	0.707	0.364	0.234	0.420	0.644	0.263	0.211	0.430	0.624	0.391	0.384	0.536	0.724	0.439	0.429
3000	0.769	0.754	0.443	0.247	0.439	0.737	0.277	0.235	0.525	0.696	0.396	0.417	0.703	0.786	0.497	0.470
4000	0.804	0.794	0.567	0.262	0.401	0.804	0.399	0.252	0.488	0.720	0.446	0.430	0.873	0.843	0.514	0.498
5000	0.853	0.834	0.571	0.267	0.489	0.830	0.306	0.263	0.510	0.798	0.453	0.465	0.828	0.864	0.553	0.541
6000	0.911	0.876	0.752	0.291	0.460	0.841	0.388	0.273	0.559	0.816	0.452	0.479	0.918	0.888	0.551	0.553
7000	0.913	0.844	0.763	0.286	0.431	0.853	0.462	0.282	0.511	0.838	0.510	0.506	0.953	0.916	0.590	0.568
8000	0.893	0.873	0.757	0.288	0.437	0.887	0.544	0.274	0.522	0.858	0.506	0.506	0.957	0.935	0.577	0.617
9000	0.903	0.894	0.818	0.300	0.430	0.893	0.602	0.296	0.568	0.899	0.464	0.525	0.964	0.944	0.589	0.615
10000	0.950	0.894	0.873	0.315	0.453	0.914	0.555	0.297	0.552	0.883	0.505	0.545	0.984	0.938	0.611	0.668

Base rate = 2.5/1000, Cluster rate = 4.5/1000

	Circle				Line				Network				Ring			
	Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV	
Population	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG
1000	0.566	0.574	0.325	0.227	0.396	0.513	0.247	0.218	0.414	0.560	0.387	0.384	0.599	0.667	0.404	0.394
2000	0.747	0.749	0.460	0.265	0.436	0.716	0.298	0.239	0.521	0.697	0.403	0.441	0.793	0.796	0.501	0.469
3000	0.786	0.793	0.636	0.266	0.434	0.801	0.389	0.260	0.538	0.745	0.424	0.449	0.896	0.877	0.535	0.527
4000	0.913	0.843	0.799	0.294	0.441	0.849	0.518	0.281	0.561	0.841	0.464	0.500	0.952	0.907	0.580	0.566
5000	0.940	0.889	0.879	0.312	0.420	0.869	0.551	0.292	0.524	0.859	0.503	0.507	0.953	0.935	0.601	0.616
6000	0.944	0.893	0.866	0.356	0.434	0.890	0.585	0.319	0.586	0.883	0.492	0.535	0.972	0.949	0.605	0.672
7000	0.957	0.877	0.879	0.322	0.421	0.901	0.666	0.313	0.553	0.900	0.499	0.557	0.972	0.949	0.617	0.702
8000	0.961	0.910	0.937	0.392	0.429	0.929	0.719	0.356	0.588	0.903	0.495	0.575	0.999	0.964	0.616	0.730
9000	0.981	0.919	0.950	0.412	0.443	0.910	0.680	0.388	0.553	0.916	0.529	0.630	0.998	0.957	0.619	0.786
10000	0.980	0.934	0.961	0.481	0.420	0.943	0.919	0.428	0.611	0.944	0.503	0.660	0.999	0.967	0.621	0.827

Base rate = 2.5/1000, Cluster rate = 5/1000

	Circle				Line				Network				Ring			
	Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV	
Population	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG
1000	0.716	0.659	0.439	0.256	0.451	0.610	0.332	0.237	0.468	0.604	0.391	0.394	0.656	0.594	0.411	0.385
2000	0.849	0.794	0.545	0.271	0.439	0.814	0.396	0.291	0.517	0.749	0.429	0.457	0.724	0.750	0.494	0.442
3000	0.909	0.844	0.872	0.315	0.433	0.851	0.536	0.295	0.558	0.822	0.477	0.493	0.869	0.849	0.564	0.474
4000	0.931	0.890	0.875	0.345	0.454	0.897	0.506	0.302	0.594	0.856	0.482	0.527	0.879	0.871	0.601	0.516
5000	0.937	0.899	0.917	0.368	0.433	0.926	0.725	0.354	0.554	0.895	0.514	0.606	0.935	0.905	0.597	0.569
6000	0.977	0.904	0.951	0.407	0.451	0.931	0.601	0.386	0.585	0.910	0.501	0.604	0.947	0.898	0.610	0.615
7000	0.971	0.904	0.965	0.485	0.441	0.939	0.724	0.412	0.546	0.935	0.540	0.669	0.929	0.921	0.612	0.650
8000	0.979	0.919	0.974	0.509	0.444	0.954	0.751	0.479	0.548	0.915	0.530	0.661	0.957	0.924	0.613	0.691
9000	0.977	0.931	0.973	0.604	0.439	0.937	0.754	0.503	0.601	0.952	0.523	0.715	0.968	0.928	0.618	0.746
10000	0.980	0.926	0.981	0.548	0.440	0.959	0.798	0.568	0.599	0.936	0.514	0.743	0.958	0.938	0.622	0.752

Base rate = 2.5/1000, Cluster rate = 5.5/1000

	Circle				Line				Network				Ring			
	Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV	
Population	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG
1000	0.771	0.673	0.497	0.265	0.433	0.660	0.322	0.263	0.498	0.645	0.411	0.441	0.827	0.796	0.489	0.474
2000	0.861	0.821	0.763	0.318	0.479	0.820	0.379	0.276	0.552	0.799	0.459	0.506	0.957	0.911	0.598	0.607
3000	0.939	0.901	0.861	0.337	0.431	0.873	0.564	0.306	0.606	0.845	0.468	0.541	0.983	0.940	0.608	0.669
4000	0.966	0.880	0.911	0.404	0.431	0.919	0.721	0.353	0.561	0.892	0.511	0.575	0.988	0.963	0.620	0.754
5000	0.976	0.900	0.963	0.446	0.439	0.943	0.764	0.414	0.623	0.928	0.484	0.649	0.998	0.958	0.621	0.844
6000	0.967	0.907	0.976	0.518	0.431	0.949	0.827	0.465	0.614	0.931	0.503	0.694	0.993	0.978	0.628	0.885
7000	0.981	0.929	0.976	0.631	0.430	0.953	0.803	0.542	0.565	0.926	0.527	0.721	1.000	0.980	0.627	0.929
8000	0.987	0.924	0.986	0.646	0.427	0.956	0.877	0.577	0.588	0.942	0.541	0.790	1.000	0.988	0.628	0.960
9000	0.987	0.936	0.987	0.693	0.433	0.947	0.899	0.649	0.616	0.962	0.522	0.820	1.000	0.983	0.629	0.963
10000	0.993	0.917	0.983	0.748	0.434	0.950	0.874	0.710	0.611	0.957	0.521	0.812	1.000	0.989	0.629	0.979

Base rate = 2.5/1000, Cluster rate = 6/1000

	Circle				Line				Network				Ring			
	Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV		Mean Sensitivity		Mean PPV	
Population	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG	SS	HG
1000	0.759	0.661	0.517	0.275	0.434	0.640	0.347	0.269	0.492	0.671	0.466	0.453	0.811	0.831	0.542	0.508
2000	0.926	0.834	0.828	0.334	0.443	0.857	0.594	0.309	0.619	0.833	0.459	0.535	0.982	0.940	0.603	0.640
3000	0.953	0.876	0.901	0.416	0.450	0.917	0.625	0.349	0.575	0.898	0.528	0.615	0.984	0.951	0.621	0.738
4000	0.980	0.904	0.984	0.509	0.426	0.946	0.821	0.424	0.609	0.916	0.494	0.646	0.999	0.974	0.622	0.822
5000	0.980	0.903	0.983	0.575	0.423	0.913	0.807	0.477	0.589	0.931	0.511	0.693	1.000	0.978	0.623	0.902
6000	0.986	0.911	0.989	0.658	0.434	0.954	0.864	0.633	0.595	0.938	0.511	0.766	1.000	0.989	0.625	0.945
7000	0.991	0.949	0.986	0.744	0.427	0.964	0.949	0.659	0.591	0.947	0.530	0.797	0.999	0.988	0.628	0.969
8000	0.981	0.940	0.987	0.814	0.431	0.960	0.880	0.716	0.576	0.950	0.541	0.814	1.000	0.993	0.628	0.984
9000	0.994	0.949	0.996	0.847	0.431	0.967	0.896	0.771	0.635	0.958	0.502	0.831	1.000	0.991	0.630	0.988
10000	0.987	0.943	0.997	0.900	0.439	0.967	0.867	0.818	0.628	0.963	0.520	0.865	1.000	0.996	0.630	0.992

APPENDIX III. Maps of crude rates into quantile (%) groups