Computational, Statistical and Graph-Theoretical Methods for Disease Mapping and Cluster Detection

by

Shannon Christine Wieland
B.S., B.A., Ohio State University, 1999

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in the field of
MATHEMATICS AND MEDICAL ENGINEERING
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2007

Author .....
Department of Mathematics
Harvard-MIT Division of Health Sciences and Technology
August 9, 2007

Certified by .....
Kenneth Mandl, M.D.
Thesis Supervisor, Assistant Professor of Pediatrics

Certified by ...
Bonnie Berger, Ph.D.
Thesis Supervisor, Professor of Applied Mathematics

Accepted by .....
Alar Toomre, Ph.D.
Chairperson, Applied Mathematics Committee

Accepted by .....
David Jerison, Ph.D.
Chairperson, Department Committee on Graduate Students

Accepted by .....
Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology

Computational, Statistical and Graph-Theoretical Methods for Disease Mapping and Cluster Detection

by

Shannon Christine Wieland

Submitted to the Department of Applied Mathematics and the Harvard-MIT Division of Health Sciences and Technology
on May 18, 2007, in partial fulfillment of the
requirements for the degree of
Doctorate in Applied Mathematics and Medical Engineering

Abstract

Epidemiology, the study of disease risk factors in populations, emerged between the 16th and 19th centuries in response to terrifying epidemics of infectious diseases such as yellow fever, cholera and bubonic plague. Traditional epidemiological studies have led to modifications in hygiene, diet, and many other practices that have profoundly altered the dynamic between humans and diseases.

In this thesis, we develop mathematical techniques to address modern challenges, including emerging diseases such as SARS and West Nile virus, the threat of bioterrorism, and stringent legislation protecting patient privacy. Within spatial epidemiology, one problem is to map the risk of disease across space (i.e., disease mapping), and another is to analyze the data for clustering. We propose a general technique, cartograms created from exact patient location data, that can address both of these problems. We also develop a graph-theoretical method to detect spatial clusters of any shape based on Euclidean minimum spanning trees. For mapping applications, we present an optimal strategy for mapping patient locations that preserves both privacy and spatial patterns within the data. For real-time disease surveillance, in which the goal is early detection of outbreaks based on time-series data, we introduce a generalized additive model that maintains constant specificity on various time scales.

Thesis Supervisor: Kenneth D. Mandl
Title: Assistant Professor

Thesis Supervisor: Bonnie A. Berger
Title: Professor

Acknowledgments

Foremost, I would like to thank my parents Sharon and David Merritt and Frank and Linda McDonald for looking after every aspect of my development throughout my life, and in particular for planning, encouraging, and sacrificing for my education. My parents-in-law Dennis and Ronnye Wieland have also been wonderfully supportive of my studies and career goals. I am indebted to my husband Aaron Wieland for his constant encouragement and for making my graduate school years happy ones, and to my daughters Bailey and Gwyneth for giving me firm deadlines, which undoubtedly sped along the process.

I am also extremely grateful for the mentorship of my Ph.D. advisors, Bonnie Berger and Kenneth Mandl. In addition to introducing me to graph theory and algorithms, Professor Berger has provided guidance during the past five years in many areas, from my coursework to planning my professional life. I am also grateful for her rare example of combining motherhood with a successful academic career. Professor Mandl introduced me to the fields of spatial epidemiology and health surveillance, and has taught me a great deal about envisioning, choosing, and collaborating on projects. I am thankful for his singular regard for my best interests, and also for his endless supply of witty and hilarious comments.

In addition to Professors Berger and Mandl, who helped develop all the ideas presented in this thesis, John Brownstein has been a helpful mentor and worked closely with me on three of the chapters of my thesis. I also collaborated with Chris Cassa on privacy protection and with Lucy Hadden, Karen Olson, and Athos Bousvaros on cartograms. I would also like to thank Daniel Kleitman for serving on my thesis committee and for his helpful comments.

I am also thankful to many other people who have enriched my intellectual life. These include my undergraduate advisor at Ohio State University, Sherwin Singer, and Edward Marcotte at the University of Texas at Austin. I have had many useful and fun conversations with several colleagues at MIT and Harvard, including Brad Friedman, Lenore Cowen, Michael Baym, Gopal Ramachandran, Clark Freifeld, Gil

Alterovitz, and Ronald Rivest, and many of their suggestions were helpful in my thesis.

I am also extremely grateful to my daughters' child care providers at the MIT Technology Children's Center, a talented and caring group of teachers who have been essential to our family: Michelle Zapatka, Alki Ikonomou, Ariel Brower, Susan Robinson, Julia Tompkins, Kettelyne Destin, Maria Bonilla, Francesca Foster and Tyhise Garay. I would also like to acknowledge the helpful and kind administrative staff at MIT and HST, especially Linda Okun, Michele Gallarelli, Andrew Kiss, Patrice Macaluso, Domingo Altarejos, Cathy Modica and Kathleen Dickey.

I am also thankful for grant support from the National Library of Medicine, the Medical Scientist Training Program, the MIT Health Sciences and Technology Bioinformatics and Integrative Genomics Program, the MIT Department of Mathematics, and the MIT Childcare Scholarship Fund from the MIT Center for Work, Life, and Family.

Contents

1 Introduction **19**

2 Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes **25**

2.1 Introduction 25

2.2 EMST Cluster Detection 27

2.2.1 Cartogram Construction 27

2.2.2 Potential Clusters 28

2.2.3 Statistical Significance 33

2.3 Results 34

2.3.1 West Nile Virus, New York City, 1999 34

2.3.2 Inhalational Anthrax, Sverdlovsk, Russia, 1979 35

2.3.3 Circular Clusters, Boston, Massachusetts 37

2.3.4 Rectangular Clusters, Boston, Massachusetts 39

2.3.5 Arbitrary Shapes 40

2.4 Discussion 41

3 Cartograms for Mapping and Analyzing Event Disease Data **45**

3.1 Introduction 45

3.2 Event Cartograms 47

3.2.1 Data 47

3.2.2 Event cartogram construction 49

3.2.3 Mapping the disease risk 51

3.3 Examples 52
3.3.1 Simulated Distributions 52
3.3.2 Pediatric Inflammatory Bowel Disease, Massachusetts, 1995-2006 54
3.4 Discussion 54

4 Optimal anomymization of patient spatial data 65
4.1 Introduction 65
4.2 Methods 68
4.3 Example 73
4.3.1 New York county census blocks 73
4.3.2 Sensitivity analysis 77
4.4 Discussion 78

5 Automated real time constant-specificity surveillance for disease outbreaks 83
5.1 Introduction 83
5.2 Methods 85
5.2.1 Data 85
5.2.2 Time series algorithms 85
5.2.3 Model predictions based on historical data 91
5.2.4 Detecting variability in the specificity 92
5.2.5 Simulated outbreaks 92
5.2.6 Estimating sensitivity, specificity, and timeliness of detection . 93
5.2.7 Comparing outbreak detection among models 93
5.3 Results 94
5.3.1 Evaluation of specificity trends over time 94
5.3.2 Comparison of sensitivity and timeliness of new and traditional methods 94
5.3.3 Temporal sensitivity trends 97
5.4 Discussion 98
5.5 Conclusions 105

List of Figures

1-1 *Die Seuche* by A. Paul Weber, depicting the bubonic plague entering a city. Image courtesy of the National Library of Medicine. 19

1-2 The English physician John Snow created a dot map showing that cholera victims lived close to one public water pump, which was the source of the outbreak. Images courtesy of the National Library of Medicine. 20

2-1 Construction of the Voronoi diagram cartogram. a) One hundred cases (green) and 50 controls (red) are distributed on a map. b) The case locations are superimposed on the Voronoi diagram constructed from the controls. c) A density-equalizing cartogram of the Voronoi diagram distorts the original map so that all Voronoi regions have the same area. New case locations are assigned on the cartogram by randomly plotting each case within its corresponding Voronoi region. 28

2-2 Procedure to locate potential clusters illustrated on a set of 15 cases. The EMST is first constructed (top left). This is a tree connecting each case (circle) that minimizes the total summed edge distance. At each step, the longest remaining edge is deleted, forming two new connected components (red). Components that were unchanged from the previous step are shown in blue. The connected components are in one-to-one correspondence with the set of potential clusters. 30

2-3 Detection of 1999 New York West Nile virus cases by SaTScan and the EMST method. a) A typical data set consisting of the 56 West Nile virus cases (red and orange) and 400 background cases (blue and gray) are shown on a map of Connecticut, New Jersey and New York. Only part of the map is shown for clarity. The West Nile virus case locations have been randomly skewed for privacy [1]. The most likely cluster identified by SaTScan is shown (red and blue). The green shading represents the density of controls in each county. b) The Voronoi diagram cartogram of part of the study area is shown along with the transformed case locations. Although the Voronoi diagram cartogram regions are not shown, the distortion of county boundaries induced by the cartogram transformation is apparent. The minimum spanning tree (black edges) connects the most likely cluster identified by the EMST method (red and blue). The control density varies by less than 2.0% over the entire map. 36

2-4 SaTScan and EMST Detection of 1979 Sverdlovsk anthrax outbreak. a) A representative data set of 63 anthrax cases (red and orange) and 400 uniformly distributed background cases (blue and gray) is shown, along with the most likely cluster determined by SaTScan (red and blue). b) The EMST method most likely cluster (red and blue) is shown for the same data set, connected by the minimum spanning tree of the cartogram-transformed cases (black edges). 38

2-5 Equally detectable potential clusters of various shapes. A most likely cluster of 35 points selected from among the Boston circular cluster data sets, along with its minimum spanning tree, is shown in the upper left. Seven other configurations of 35 points, having minimum spanning trees with exactly the same weight, are also shown. Subject to the constraint imposed by the definition of a potential cluster above, all eight clusters have equivalent detectability by the EMST method. If embedded as potential clusters in a Boston data set of 500 total cases, all would achieve the same p-value of 0.0001. 41

3-1 Applications of cartograms to spatial epidemiology. 47

3-2 Example of a Voronoi tesselation. Left: One thousand points are distributed on a map. Right: The Voronoi tesselation of the points divides the map into 1000 regions. Each region consists of the portion of the map closest to one point. The density structure is preserved in the tesselated map; regions of small Voronoi cell area correspond to high point density. 50

3-3 Dot maps and cartograms of three hypothetical disease distributions. Dot maps of 5,000 controls (blue) and 2,500 cases (red) are shown in the left column (a, c and e). The controls are distributed in proportion to the underlying population. The cases are distributed to illustrate constant relative risk (a), risk increasing linearly by a factor of four from north to south (c), and a localized cluster with a three-fold increase in relative risk in Iowa and neighboring states (e). The right column (b, d and f) shows the cartogram-transformed case locations for the three distributions. 61

3-4 Isopleth surfaces estimating the relative risk of three hypothetical disease distributions on standard maps (a,c and e) and cartograms (b,d and f). The exact locations of the cases and controls are shown in figure 3-3. The case distributions illustrate constant risk (a and b), a four-fold increase in risk from south to north (c and d) and a cluster of three-fold risk increase centered in Iowa (e and f). The patterns are obscured on the standard maps because of the presence of high relative risk artifacts, but are clear on the cartograms. 62

3-5 Pediatric inflammatory bowel disease risk in Massachusetts, 1995-2006. a) A standard map of the study area. b) A cartogram was constructed from the Voronoi diagram of the 7988 control locations. The 901 IBD cases were randomly placed within the cartogram regions corresponding to their original locations on the Voronoi diagram. An isopleth relative risk surface was calculated from the transformed case locations using kernel methods. Original case and control locations are not shown to protect patients' privacy. 63

4-1 Schematic of transition probabilities. A patient found at each location may transition to any other location. In this simple example, there are three locations (represented by houses) and nine transition probabilities (represented by arrows). The probabilities are variables solved by linear programming. 68

4-2 Total population of each census block group in New York County, NY, according to the 2000 census. 74

4-3 Transition probabilities for the optimal strategy to de-identify $s \leq 20,000$ patients from New York County, New York with a maximum re-identification probability of $\frac{s}{20000}$. Transition probabilities from three of the 988 census blocks are shown, illustrating a few of the many possible transition distributions. The shading in region j represents the value of the probability P_{ij} of transitions into the region. a) Patients in one census block (purple asterisk) may remain there, or they may transition to one of several nearby blocks. b) All patients originally in one census block (purple asterisk) are assigned to one neighboring block. c) Patients are re-assigned from one block (purple asterisk) to one of four nearby census blocks. No patients are re-assigned to the original census block (i.e. $P_{ii} = 0$). 75

4-4 Histogram of the distance between original and de-identified locations for an individual randomly chosen from the population, under the optimal strategy to de-identify a set of $s \leq 20,000$ patients in New York County, New York to a probability of $\frac{s}{20000}$. 76

4-5 Relationship between the re-identification probability, the number s of patients, and the expected transition distance for the optimal LP strategy to de-identify patients by census block group in New York county, New York. As the level of privacy protection decreases, patients are moved a smaller distance in expectation. Aggregation by zip code (green diamond) and first three zip code digits (magenta circle) are suboptimal strategies. 77

4-6 Aggregation of patients in New York County, New York by zip code and by first three zip code digits. Top) Census block groups have been aggregated by zip codes. Each census block group was assigned to the zip code containing its centroid. The expected distance moved by a randomly selected member of the population is 519 m, and the maximum probability that an individual is among a set of s de-identified patients is $\frac{s}{884}$. Bottom) Census block groups are aggregated by the first three zip code digits. The expected distance moved is 3.866 km, and the re-identification probability is $\frac{s}{8188}$. 81

5-1 Emergency department visits for respiratory presenting complaints, August 1, 1992 - July 30, 2004. Daily time series showing the number of patients presenting with respiratory complaints to the emergency department during a 12 year period. 86

5-2 Evaluating variability in specificity on three time scales. Plots of p-values for the chi-square test over various time scales for the five comparison models over a range of mean specificity values from 0.50 to 0.99, as well as p-values for the expectation-variance model. Top: calendar year of study. Middle: month of year. Bottom: day of week. The expectation-variance model has a p-value over 0.05 for the entire range of mean specificity values for all three time scales, so the null hypothesis of constant specificity is not rejected. All plots not shown are highly significant ($p < 0.001$) for non-constancy. 95

5-3 Average specificity trends over time. Average specificity for each calendar year, month, and day of week for the five comparison methods during the study period. Data shown were recorded for each model implemented at 85% mean specificity. Similar trends were observed for all methods at 97% mean specificity (data not shown). 96

5-4 Seasonal sensitivity trends. Average sensitivity for each month of the study period for the autoregressive (left), trimmed seasonal (center), and expectation-variance (right) models when applied to data containing a superimposed spike outbreak of 10 additional patients during one day. Data shown were collected at a mean specificity of 97%. The sensitivity of the trimmed seasonal and autoregression models is higher during the winter than during the summer. Sensitivity is higher during the summer than during the winter for the expectation-variance model. July receiver-operator (ROC) curves lie below February ROC curves for all three models (insets). Similar trends were observed for flat and linear outbreaks. 99

5-5 Seasonal trends in the mean and variance of ED visits. Mean number of ED visits (left axis, solid blue line) and mean variance in ED visits (right axis, dashed green line) as a function of the day of year. Data were smoothed using 5-day and 11-day moving averages, respectively. The ED utilization mean and variance are highest in the winter and lowest during the summer. 102

List of Tables

2.1 SaTScan and EMST method applied to West Nile virus. n, number of background cases added to cluster cases; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference. 35

2.2 SaTScan and EMST method applied to anthrax. n, number of background cases added to cluster cases; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference. . . . 37

2.3 SaTScan and EMST method applied to circular clusters. r, radius of cluster in kilometers; d, relative cluster density; m, mean cluster size; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference. 39

2.4 SaTScan and EMST method applied to rectangular clusters. r, ratio of cluster height to width; d, relative cluster density; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference. 40

5.1 ROC curve areas for traditional and expectation-variance detection models applied to three different types of outbreaks superimposed on respiratory visits to an urban pediatric ED, August 1998 - July 2004. 97

5.2 Mean lag in detecting outbreaks of five additional patients per day superimposed on the pediatric ED respiratory visits, August 1998 - July 2004. Detection lag calculations exclude undetected outbreaks. Hence the sensitivity of the method must be considered when interpreting the detection lag. 97

Chapter 1

Introduction

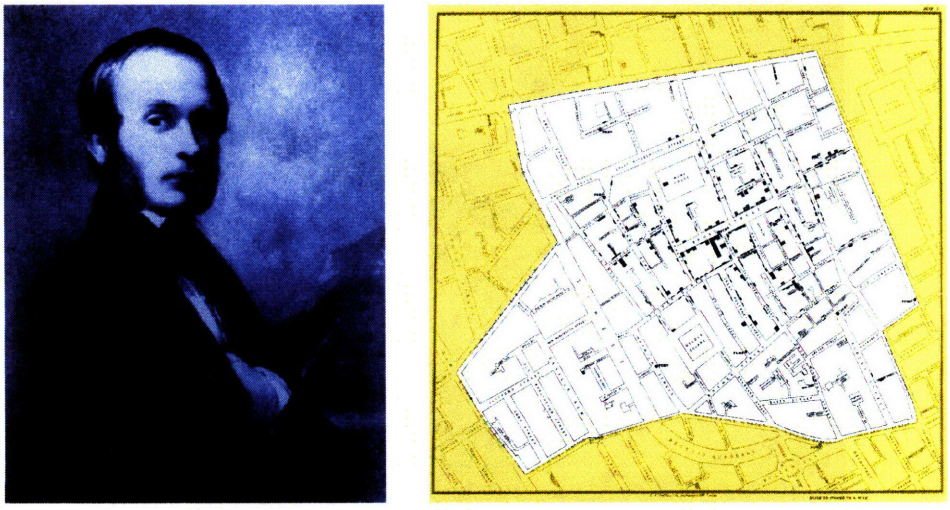
Terrifying epidemics have swept through populations throughout human history. Most famous among these is the bubonic plague, which spread in every direction from the Gobi desert in China in the 1320's, devastating parts of Asia and Africa. The plague reached Cyprus in 1347 and killed about one third of the European population in only two years [2]. In recent history, 500 million people contracted "Spanish

Figure 1-1: *Die Seuche* by A. Paul Weber, depicting the bubonic plague entering a city. Image courtesy of the National Library of Medicine.

influenza" in 1918 and 1919 [3]. The epidemic began among soldiers in March of 1918

at Camp Funston in Kansas. In late August, three nearly simultaneous outbreaks in Boston, Massachusetts, Freetown, Sierra Leone, and Brest, France signaled the start of a global pandemic which claimed at least thirty million lives [4]. The fear inspired by uncontrollable and fatal epidemics, such as the plague, yellow fever, cholera, and influenza, was a driving force for early advances in the field of spatial epidemiology.

The first disease dot maps were published by a young surgeon named Valentine Seaman in 1798, showing the locations of yellow fever victims in New York City. Seaman created the maps to support his theory that yellow fever was caused by "putrid effluvia," which was ultimately disproved [5]. An English surgeon named John Snow is often credited with founding the field of spatial epidemiology for his 1854 study of cholera in London (see figure 1-2). At the time, frequent outbreaks of cholera, resulting in severe diarrhea and death due to dehydration, were generally thought to be caused by "miasma" in the air. Snow, who happened to live close to the epicenter of a large outbreak occurring in late August, 1854, correctly theorized that cholera was spread through contaminated water. He plotted the cases, revealing that they clustered around one pump on Broad street. Seven days into the outbreak, he pre-



(a) John Snow, 1847. (b) Map of cholera cases in London, 1854.

Figure 1-2: The English physician John Snow created a dot map showing that cholera victims lived close to one public water pump, which was the source of the outbreak. Images courtesy of the National Library of Medicine.

sented his findings to the local Board of Guardians. The pump handle was removed, ending the epidemic. The map was used to support Snow's theory in a subsequent publication "On the mode of communication of cholera" [6]. Snow's success showed the potential power of spatial methods in epidemiology: his finding not only saved lives, but also gave new insight into the transmission of a poorly understood disease.

From the earliest disease dot maps, methods in spatial epidemiology have evolved to include a range of statistical and graphical techniques encompassing several distinct areas of study. Disease mapping explores spatial variations in disease risk, taking into account variations in the underlying at-risk population. Disease clustering studies investigate whether or not cases tend to cluster together more than expected, or seek to find localized subsets of patients comprising clusters. Ecological analysis investigates the relationship between the distribution of cases and environmental risk factors [7].

Despite its long and productive history, there are several challenges still facing the field of spatial epidemiology. These include the need to rapidly detect emerging diseases, such as Severe Acute Respiratory Syndrome and West Nile Virus, and bioterrorism events, such as the dissemination of anthrax through the United States postal service in 2001. There is also an increased public awareness of issues surrounding patient privacy, and more stringent legislation protecting privacy of patient-identifiable information, including geographic identifiers. Furthermore, there are recent advances in geographical information systems and cartography methods that can be leveraged for spatial epidemiology.

In this thesis, we respond to these new challenges and advances with several related projects. In chapter 2, we create a new graph-theoretical method to detect spatial clusters of any shape. Existing disease cluster detection methods cannot detect clusters of all shapes and sizes, or identify highly irregular sets that overestimate the true extent of the cluster. We introduce a graph-theoretical method for detecting arbitrarily-shaped clusters based on the Euclidean minimum spanning tree of cartogram-transformed case locations, which overcomes these shortcomings. The method is illustrated using several clusters, including historical data sets from West

Nile virus and inhalational anthrax outbreaks. Sensitivity and accuracy comparisons with the prevailing cluster detection method show that the method performs similarly on approximately circular historical clusters, and it greatly improves detection for non-circular clusters.

The use of cartograms based on exact location data, developed for this method, is explored in other contexts in chapter 3. Density-equalizing cartograms of disease case locations are used to adjust for variation in the underlying at-risk population for the purposes of visual representation and statistical analysis of disease risk. The use of cartograms has been limited to analyzing count data in a small number of settings. We show how to create and interpret cartograms from exact location data collected using various types of traditional epidemiological studies. For mapping applications, there is a simple relationship between cartogram case density and disease risk; for analysis, the cartogram simplifies the null distribution of constant disease risk, enabling the use of a variety of well-advanced statistical methods.

In chapter 4 we develop an optimal strategy for balancing the need for patient privacy with the need to share information about the spatial distributions of diseases for research and health surveillance. Ethical and legal mandates protect the privacy of patient data collected for medical care and research. Accidental disclosures sometimes occur, either because the guardians of the data do not anticipate a method of linking a released data set to individuals, or because of methodological flaws in the procedures used to ensure privacy. The prevailing solution, releasing data aggregated by large areas, usually preserves privacy but suffers from substantial information loss. We develop an alternative de-identification strategy to move individual locations based on linear programming. The method guarantees that privacy is protected. It moves patients in an optimal manner to ensure they move the minimal possible distance for the level of privacy protection. Thus the de-identified set is ideal for subsequent cluster detection or disease mapping studies. We illustrate how to de-identify patients in New York county, New York, showing that privacy is guaranteed while moving patients very short distances.

In chapter 5, we develop a temporal method to detect aberrant health events

for surveillance in real time. Detection of abnormal disease patterns is based on a difference between patterns observed, and those predicted by models of historical data. The usefulness of outbreak detection strategies depends on their specificity; the false alarm rate affects the interpretation of alarms. We evaluate the specificity of four traditional models: autoregressive, Serfling, trimmed seasonal, and wavelet-based. We apply each to 12 years of emergency department visits for respiratory infection syndromes at a pediatric hospital, finding that the specificity of the four models was almost always a non-constant function of the day of the week, month, and year of the study ($p < 0.05$). We develop an outbreak detection method, called the expectation-variance model, based on generalized additive modeling to achieve a constant specificity by accounting for not only the expected number of visits, but also the variance of the number of visits. The expectation-variance model achieves constant specificity on all three time scales, as well as earlier detection and improved sensitivity compared to traditional methods in most circumstances. Modeling the variance of visit patterns enables real-time detection with known, constant specificity at all times. With constant specificity, public health practitioners can better interpret the alarms and better evaluate the cost-effectiveness of surveillance systems.

Chapter 2

Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes

2.1 Introduction

Tests for the detection of disease clusters [8] are essential tools for identifying emergent infections and elucidating demographic and environmental factors influencing diseases. The shapes of these clusters are unpredictable [9, 10, 11, 12, 13]. However, the prevailing cluster detection method, a scan statistic that applies a likelihood ratio test to a large number of overlapping circles in a study region, reports only circular clusters [14, 15]. Straightforward extensions of the circular scan statistic, such as an elliptical scan [16] and a rectangular scan [17], are also limited to detecting specific outbreak shapes.

Originally published as: Wieland SC, Brownstein JS, Berger B, Mandl KD. Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes. Proceedings of the National Academy of Sciences. May 22, 2007.

Few methods aim to detect clusters of arbitrary shape. One class of methods based on graph theory has recently emerged to address this problem [18, 19, 20, 21]. However, these have several limitations: they are restricted to clusters that fit inside a circular region of fixed size [18], they attempt to examine a set of potential clusters too large to exhaustively search [19], they have poor specificity [20], or have yet to be implemented or evaluated [21].

In addition to the difficulties inherent in any disease cluster detection method, such as accounting for the underlying population density and controlling the level of significance given multiple potential clusters of various sizes and in various locations, arbitrary shape cluster detection presents particular challenges. As more shapes are considered, the statistical power declines, and the computational running time may become unreasonable for typical problem sizes [18]. Furthermore, if the exact case locations are available, then considering every conceivable shape is problematic; it is always possible to draw a bizarrely shaped region of infinitesimally small total area that includes every case. This problem surfaces when data are aggregated into small regions. Indeed, one study identified excessively large clusters with highly irregular shapes having greater likelihood ratios than the inserted clusters which were the detection targets [20].

In this study, we address these challenges by removing the notion of shape from consideration, and replacing it with a mathematical formalization of potential clusters based on intercase distances. We introduce a method to locate clusters of any shape based on Euclidean minimum spanning trees (EMST's), which have previously found application in heuristic methods to divide other kinds of data into a pre-determined number of subsets [22, 23]. Application of the method to synthetic, West Nile virus, and anthrax data sets show that sensitivity and accuracy are substantially improved compared to the circular scan statistic method applied to non-circular clusters, which likely include the majority of real disease clusters.

2.2 EMST Cluster Detection

Our cluster detection method consists of three sequential tasks. A density-equalizing cartogram of the study region and disease cases is first constructed from a Voronoi diagram of the controls. Second, the family of potential clusters to evaluate is defined, since it is not computationally feasible to consider all 2^n subsets of n cases. Third, the statistical significance of each potential cluster is evaluated. We address each of these tasks below.

2.2.1 Cartogram Construction

We begin with the precise spatial coordinates of a set of disease cases and controls, and a map of the study area. We first create a Voronoi diagram of the control locations, which subdivides the study area into the regions closest to each control location [24] (see figure 2-1). The density of controls within each Voronoi region is simply the number of controls in the region, which may be more than one if multiple controls can occur at the same location, divided by the region's area. We use this density function to create a density-equalizing cartogram of the Voronoi diagram. Cartograms have previously been used for aggregate data to test for clustering of several diseases [25, 26, 27, 28, 29]. To construct one, each point on the original map is essentially magnified or demagnified according to its local density. The result is a distorted map on which the density of controls is constant everywhere. Each case is placed on the cartogram at a random location within the region corresponding to its original Voronoi region, and all subsequent analyses are performed using these new case locations. Under the null hypothesis of constant relative risk, the new locations of the cases on the Voronoi diagram cartogram are uniformly and independently distributed. We use a diffusion-based cartogram construction algorithm [29], although other contiguous cartogram algorithms may also be suitable.

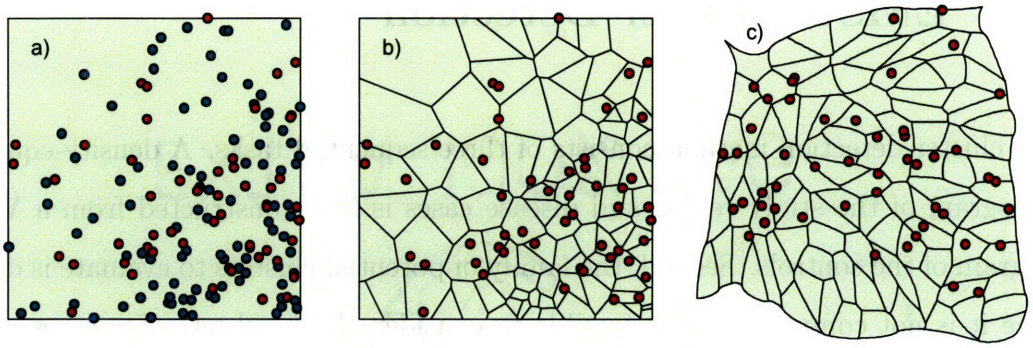


Figure 2-1: Construction of the Voronoi diagram cartogram. a) One hundred cases (green) and 50 controls (red) are distributed on a map. b) The case locations are superimposed on the Voronoi diagram constructed from the controls. c) A density-equalizing cartogram of the Voronoi diagram distorts the original map so that all Voronoi regions have the same area. New case locations are assigned on the cartogram by randomly plotting each case within its corresponding Voronoi region.

2.2.2 Potential Clusters

We call a *potential cluster* a subset of points S satisfying the property that every subset of S is "closer" to at least one other point in S than to any other point outside of S. To formalize this definition, we begin by defining the distance $\rho(X, Y)$ between two sets X and Y to be the smallest distance separating the sets:

$$\rho(X,Y) = \begin{cases} \min_{\substack{a \in X \\ b \in Y}} \rho(a,b) & \text{if } X \neq \emptyset \text{ and } Y \neq \emptyset \\ \infty & \text{otherwise} \end{cases} \tag{2.1}$$

where $\rho(x, y)$ is the Euclidean distance between two points. We also define the internal distance of a nonempty set S to be the maximum distance between any two nonempty subsets of S whose union is S:

$$\rho(S) = \max_{\substack{\emptyset \subsetneq X \subseteq S \\ \emptyset \subsetneq Y \subseteq S \\ X \cup Y = S}} \rho(X,Y) \tag{2.2}$$

We formally define a potential cluster as follows:

Definition Let V be a nonempty set of cases of a disease. A potential cluster is a nonempty set $S \subseteq V$ satisfying $\rho(S) < \rho(S, V - S)$.

Note that the entire set V is a potential cluster, as are the sets $\{v\}$ for every $v \in V$. If v is the nearest neighbor of w and w is the nearest of v, then $\{v, w\}$ is a potential cluster.

We wish to consider every potential cluster in V, but it is not straightforward from the definition how to locate potential clusters, nor how many of them are present. Progress was made toward finding potential clusters in a different application in bioinformatics [23] using the minimum spanning tree of V, a connected graph T spanning a set of points having minimal total weight

$$w(T) = \sum_{e \in E(T)} w(e) \tag{2.3}$$

where $E(T)$ denotes the set of edges of T, and the weight $w(e)$ of an edge e is in this case the Euclidean distance between the endpoints of e. (For a detailed review of graph theoretical definitions, see [30].) Given a set V of n points, every potential cluster is a connected subgraph of the EMST T of V [23]. However, even for small epidemiological data sets, the number of connected subgraphs may be extremely large; EMST's of 50 and 75 random points have approximately 10^6 and 10^8 connected subgraphs, respectively.

We prove that it is not necessary to consider all connected subgraphs of T to find the potential clusters. Remarkably, there are at most $2n - 1$ potential clusters, of which n are trivial sets consisting of only one vertex. Furthermore, the potential clusters may be quickly found from an EMST using a greedy edge deletion procedure. After constructing an EMST of the set of cartogram case locations V, we iteratively delete the longest remaining edge of T. At each iteration we consider the two newly emergent connected components, each of which is a potential cluster. In this way, we evaluate all $n - 1$ nontrivial potential clusters for statistical significance using a test described below (see Figure 2-2).

We prove that this procedure identifies the set of potential clusters by showing that that potential clusters, characterized by the definition above, are in one-to-one correspondence with a small class of subsets of an EMST T. For $w \geq 0$, we define

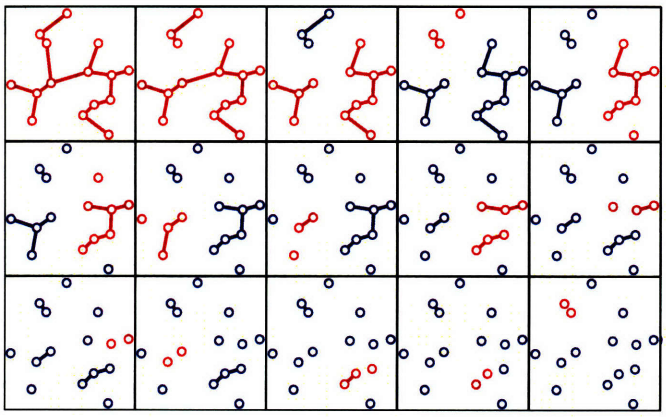


Figure 2-2: Procedure to locate potential clusters illustrated on a set of 15 cases. The EMST is first constructed (top left). This is a tree connecting each case (circle) that minimizes the total summed edge distance. At each step, the longest remaining edge is deleted, forming two new connected components (red). Components that were unchanged from the previous step are shown in blue. The connected components are in one-to-one correspondence with the set of potential clusters.

T_w to be the graph derived from T by deleting all edges of T having weight greater than w. We label the $n-1$ edges of T in order of decreasing weight, so that $w(e_1) \geq w(e_2) \geq \cdots \geq w(e_{n-1}) > 0$. If the edge weights are distinct, then there are n distinct graphs T_w; these are the graphs $T = T_{w(e_1)} \supsetneq T_{w(e_2)} \supsetneq \cdots \supsetneq T_{w(e_{n-1})} \supsetneq T_0$. $T_{w(e_{k+1})}$ is formed from $T_{w(e_k)}$ by deleting one edge, which splits one connected component of $T_{w(e_k)}$ into two components. Thus $T_{w(e_{k+1})}$ has $k+1$ connected components, $k-1$ of which are present in $T_{w(e_k)}$, and two of which are newly created. There are $2n-1$ total distinct connected components among all the graphs T_w (see Figure 2-2). If the edge weights are not distinct, then a variation of this argument shows that $2n-1$ is an upper bound on the number of distinct connected components. The following characterizes the connected components:

Lemma 2.2.1 *Let V be a nonempty set of points in a plane (representing cases of a disease). Let T be a Euclidean minimum spanning tree of V, S a nonempty subset of V, and T_S the subgraph of T induced by S. The set S is a potential cluster if and only if T_S is a connected component of T_0 or of $T_{w(e_k)}$ for some k.*

The proof is made easier by two simple lemmas.

Lemma 2.2.2 *Let T_S be a connected subgraph of T with vertex set S. Then $\rho(S)$*

(Eq. 2.2) is equal to the maximum weight of an edge in T_S if $|S| > 1$, and 0 otherwise.

Proof: If $|S| = 1$, then $S = \{x\}$ and

$$\rho(S) = \max_{\substack{\emptyset \subsetneq X \subseteq S \\ \emptyset \subsetneq Y \subseteq S \\ X \cup Y = S}} \rho(X, Y) = \rho(\{x\}, \{x\}) = \rho(x, x) = 0.$$

If $|S| > 1$, let $e = (v_1, v_2)$ be an edge of maximum weight in T_S. $T_S - e$ has two components with vertex sets V_1 and V_2. We first show that $\rho(V_1, V_2) = w(e)$, where $w(e)$ is the weight of e. We have

$$\rho(V_1, V_2) = \min_{\substack{x \in V_1 \\ y \in V_2}} \rho(x, y) \leq \rho(v_1, v_2).$$

Assume the inequality is strict, so there exist $w_1 \in V_1$ and $w_2 \in V_2$ with $\rho(w_1, w_2) < \rho(v_1, v_2)$. The graph $T - e + (w_1, w_2)$ is a spanning tree of V having lower weight than T, which is a contradiction. Hence $\rho(V_1, V_2) = w(e)$.

We now show that $\rho(S) = w(e)$. Since

$$\rho(S) = \max_{\substack{\emptyset \subsetneq X \subseteq S \\ \emptyset \subsetneq Y \subseteq S \\ X \cup Y = S}} \rho(X, Y) \geq \rho(V_1, V_2) = w(e),$$

we need only prove that $\rho(S) \leq w(e)$. This is true if $\rho(X, Y) \leq w(e)$ for every X and Y satisfying the conditions $\emptyset \subsetneq X \subseteq S$, $\emptyset \subsetneq Y \subseteq S$ and $X \cup Y = S$. Let X and Y be arbitrary sets satisfying these conditions. If X and Y share a common element, then $\rho(X, Y) = 0 \leq w(e)$. If X and Y have no common element, then since they partition the vertices of T_S into two nonempty sets, there exists some edge $f = (x, y)$ of T_S spanning X and Y. We have

$$\rho(X, Y) = \min_{\substack{a \in X \\ b \in Y}} \rho(a, b) \leq \rho(x, y) = w(f) \leq w(e).$$

Hence $\rho(S) = w(e)$.

Lemma 2.2.3 *If S is a nonempty, proper subset of V, then $\rho(S, V - S)$ is equal to*

the minimum weight of an edge in T spanning the cut $(S, V - S)$.

Proof: Let $e = (v_1, v_2)$ be an edge of T of minimum weight spanning $(S, V - S)$. We have

$$\rho(S, V - S) = \min_{\substack{a \in S \\ b \in V - S}} \rho(a, b) \leq \rho(v_1, v_2) = w(e).$$

It suffices to prove that $\rho(S, V - S) \geq w(e)$, which holds if $\rho(a, b) \geq w(e)$ for every $a \in S$ and $b \in V - S$. Suppose there exist some $a \in S$ and $b \in V - S$ for which $\rho(a, b) < w(e)$. The edge (a, b) must not be in T since e has minimum weight of all edges spanning $(S, V - S)$. The graph $T + (a, b)$ therefore contains exactly one cycle, and the cycle contains some edge $f \neq (a, b)$ spanning $(S, V - S)$. The graph $T + (a, b) - f$ is a spanning tree of V, and $w(T + (a, b) - f) = w(T) + w((a, b)) - w(f) < w(T) + w(e) - w(e) = w(T)$, contradicting the minimality of the weight of T. Hence $\rho(a, b) \geq w(e)$ for every $a \in S$ and $b \in V - S$, and so $\rho(S, V - S) \geq w(e)$.

Proof of Lemma 2.2.1: We first show that every potential cluster induces a connected component of T_0 or of $T_{w(e_k)}$ for some k. Equivalently, we show that if a subgraph H of T is not a connected component of $T_{w(e_k)}$ or of T_0, then the vertex set of H is not a potential cluster. Xu *et al.* [23] showed that every potential cluster induces a connected subgraph of T, so that if H is not connected, then its vertex set is not a potential cluster. Suppose H is a connected subgraph of T which is not a connected component of $T_{w(e_k)}$ for any k, or of T_0. H must have at least one edge; let e_j be an edge of H of maximal weight. Let C be the connected component of $T_{w(e_j)}$ containing e_j. Since H is a connected subgraph of $T_{w(e_j)}$ containing e_j, $H \subsetneq C$ We refer interchangeably to a graph and its vertex set to simplify notation. There exists some edge $e \in T$ spanning H and $C - H$, and since $e \in C$, $w(e) \leq w(e_j)$. By lemma 2.2.2, $\rho(H) = w(e_j)$, and by lemma 2.2.3, $\rho(H, V - H) \leq \rho(H, C - H) \leq w(e) \leq w(e_j)$. Hence $\rho(H, V - H) \leq \rho(H)$ and H is not a potential cluster.

To finish the proof, we must show that every connected component of $T_{w(e_k)}$ for any k or T_0 is a potential cluster. This is trivial for $T_{w(e_1)} = T$ or for T_0, whose components are the individual vertices. Let T_S be a connected component of $T_{w(e_k)} \neq T$ with vertex set S. Then $\rho(S) \leq w(e_k)$ by lemma 2.2.2. Since $V - S \neq \emptyset$, there must be

some edge $e \in T$ spanning S and $V - S$. Since the edge is not in $T_{w(e_k)}$, $w(e) > w(e_k)$. This is true for every spanning edge, so by lemma 2.2.3, $\rho(S, V - S) > w(e_k)$. Hence $\rho(S) < \rho(S, V - S)$, and so S is a potential cluster.

Note that the proof does not rely on the uniqueness of T, so degenerate EMST's do not affect the ability of the method to capture all potential clusters. If the set of cases V are continuously distributed on the cartogram, as in the present study, then in theory the EMST is unique with probability 1. However, degenerate EMST's may occur with extremely low probability due to the inability of computers to support arbitrary precision.

2.2.3 Statistical Significance

In order to assign a p-value to any potential cluster, a test statistic is required, along with its distribution under the null hypothesis H_0 of independently, uniformly distributed cases on the cartogram. Let Σ be a potential cluster generated under H_0, and let S be an observed potential cluster. We define

$$P_S = \Pr\{w(\Sigma) < w(S) \mid card(\Sigma) = card(S)\}, \tag{2.4}$$

where w is the weight of the potential cluster subgraph, and *card* denotes the number of cases. P_S is the p-value corresponding to the observed candidate cluster weight, conditioned on the number of cases in S. Because cases in a true cluster are closer together than expected, the weight $w(S)$ of a potential cluster S corresponding to a hot-spot is likely to be smaller than a random EMST potential cluster subgraph containing the same number of cases. Consequently, a hot-spot should have a low value of P_S. We define the test statistic P to be the minimum value of P_S over the set of nontrivial potential clusters containing at most half of the cases. Monte Carlo techniques are used to fit P_S as a function of $w(S)$ to a Gaussian distribution for each possible value of $card(S)$. The null distribution of P is subsequently estimated, again by Monte Carlo, and a cutoff value corresponding to the desired level of significance α is obtained.

The most significant cluster is reported, but the method could easily be modified to report all significant clusters without affecting the asymptotic running time.

2.3 Results

We applied the SaTScan circular scan statistic [15] and EMST method to several types of data sets, finding that the EMST method was substantially better able to detect non-circular clusters. The SaTScan Bernoulli model was used with a maximum geographic window size containing 50% of the cases for each data set. For each method and data set, the most significant cluster with a p-value of at most 0.05 computed using 9,999 Monte Carlo replications was reported; thus the specificity, defined as the probability of reporting no significant cluster in data generated under the null hypothesis, was 0.95 for both methods and all data sets. The sensitivity, equal to the fraction of clusters that were detected, was calculated for each data set and method. To quantify the extent of overlap between the most likely cluster and the actual cluster, we defined two other measures. We defined F_{TC} to be the fraction of true cluster cases that were correctly found in the most likely cluster, and F_{MLC} to be the fraction of cases in the most likely cluster that coincided with the true cluster.

2.3.1 West Nile Virus, New York City, 1999

The EMST method and SaTScan had similar performance detecting a 1999 outbreak of West Nile virus in New York City [31]. This was encouraging because the 56 cases appear to have an approximately circular distribution (see Figure 2-3), suggesting an advantage for the circular scan statistic. We defined a study area consisting of Connecticut, New Jersey and New York, and generated 10,000 controls within the map distributed in proportion to 2000 U.S. census county population data. In order to evaluate the methods, we required data sets with both outbreak and non-outbreak cases. In addition to the West Nile virus cases, we generated 400, 600, 800, 1000 or 1200 additional non-outbreak background cases distributed according to the underlying population distribution. As the number of background cases increased, the West

	SaTScan			EMST			Comparisons		
n	SN	F_{TC}	F_{MLC}	SN	F_{TC}	F_{MLC}	$\Delta\ SN$	$\Delta\ F_{\mathrm{TC}}$	$\Delta\ F_{\mathrm{MLC}}$
400	1.00	0.69	0.61	1.00	0.80	0.53	+0.5%	+16%	-14%
600	1.00	0.63	0.54	1.00	0.69	0.48	+0.2%	+9.1%	-11%
800	0.99	0.58	0.48	1.00	0.61	0.44	+0.7%	+5.1%	-8.5%
1000	0.99	0.55	0.44	0.99	0.55	0.41	-0.4%	-0.1%	-6.8%
1200	0.89	0.49	0.40	0.96	0.50	0.38	+8.0%	+3.4%	-4.6%

Table 2.1: SaTScan and EMST method applied to West Nile virus. n, number of background cases added to cluster cases; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference.

Nile virus cluster became harder to detect. We created 1000 data sets for each background case number. The data sets could represent, for example, emergency visits for neurological symptoms in a multi-state surveillance area, with controls drawn from all emergency visits. Figure 2-3 shows a typical data set along with its Voronoi diagram cartogram transformation and the most likely cluster obtained by both methods. The results of applying SaTScan and the EMST method to the data sets are summarized in Table 2.1.

Both methods displayed similar comparative performance for all numbers of background cases. The sensitivity of both methods declined from 1.0 for 400 background cases to 0.96 and 0.89 for 1200 background cases for the EMST method and SaTScan, respectively. The percent change in F_{TC} of the EMST method compared to SaTScan varied from -0.4% to 16%, and the percent change in F_{TC} varied from -14% to -6.8%.

2.3.2 Inhalational Anthrax, Sverdlovsk, Russia, 1979

The EMST method had greater accuracy than SaTScan when applied to a highly non-circular outbreak of 62 cases of inhalational anthrax occurring in Sverdlovsk, Russia in 1979 [9]. Because we lacked spatial references for the data necessary to geocode the case locations, we used a uniform distribution within a square study region to generate 10,000 controls. The set of cases consisted of 400, 600, 800, 1000, or 1200

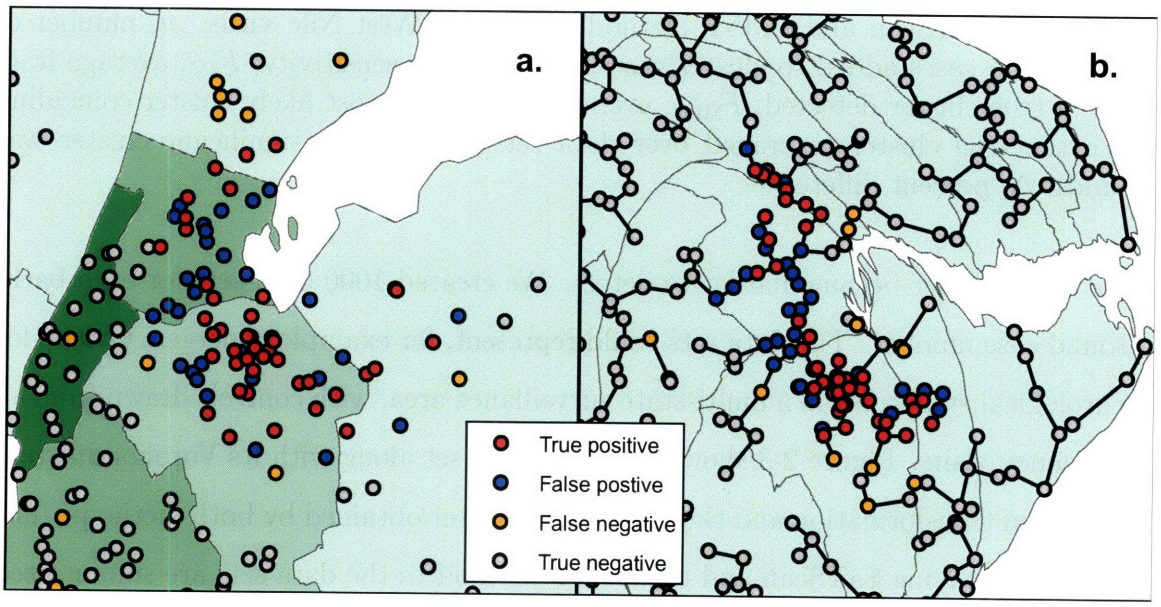


Figure 2-3: Detection of 1999 New York West Nile virus cases by SaTScan and the EMST method. a) A typical data set consisting of the 56 West Nile virus cases (red and orange) and 400 background cases (blue and gray) are shown on a map of Connecticut, New Jersey and New York. Only part of the map is shown for clarity. The West Nile virus case locations have been randomly skewed for privacy [1]. The most likely cluster identified by SaTScan is shown (red and blue). The green shading represents the density of controls in each county. b) The Voronoi diagram cartogram of part of the study area is shown along with the transformed case locations. Although the Voronoi diagram cartogram regions are not shown, the distortion of county boundaries induced by the cartogram transformation is apparent. The minimum spanning tree (black edges) connects the most likely cluster identified by the EMST method (red and blue). The control density varies by less than 2.0% over the entire map.

	SaTScan			EMST			Comparisons		
n	SN	F_{TC}	F_{MLC}	SN	F_{TC}	F_{MLC}	$\Delta\ SN$	$\Delta\ F_{\mathrm{TC}}$	$\Delta\ F_{\mathrm{MLC}}$
400	0.98	0.32	0.65	0.98	0.48	0.49	-0.4%	+48%	-24%
600	0.88	0.28	0.53	0.86	0.39	0.40	-2.3%	+38%	-25%
800	0.60	0.19	0.44	0.72	0.32	0.32	+19%	+68%	-28%
1000	0.53	0.17	0.37	0.60	0.26	0.26	+12%	+55%	-31%
1200	0.35	0.11	0.32	0.52	0.21	0.22	+46%	+100%	-31%

Table 2.2: SaTScan and EMST method applied to anthrax. n, number of background cases added to cluster cases; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference.

uniformly distributed background cases, in addition to the anthrax case locations. These could represent, for example, visits for respiratory complaints to an emergency department, with controls drawn from all visits. For each number of background cases, 1000 data sets were generated. A typical data set is shown in Figure 2-4, along with the most likely cluster detected by SaTScan and the EMST method. The mean sensitivity, F_{TC}, and F_{MLC} are summarized in Table 2.2.

The EMST method had comparable or greater sensitivity than SaTScan for all background population sizes, and it correctly identified a greater fraction of the anthrax cases (F_{TC}) for all background population sizes. Both methods' sensitivity declined as more background cases were added: from 0.98 to 0.52 for the EMST method, and from 0.98 to 0.35 for SaTScan. The EMST method had a lower value of F_{MLC} than SaTScan, indicating that it overestimated the cluster to a greater extent than SaTScan. However, the percent decline in F_{MLC} incurred by using the EMST method instead of SaTScan was about half of the gain in F_{TC}.

2.3.3 Circular Clusters, Boston, Massachusetts

We also compared the ability of the EMST method and SaTScan to detect circular clusters. Because the circular scan statistic is optimized to detect circular clusters, we were surprised to find that the EMST method was as sensitive as SaTScan. The study area consisted of the 59 zip codes within 10 km of Boston, Massachusetts. Ten

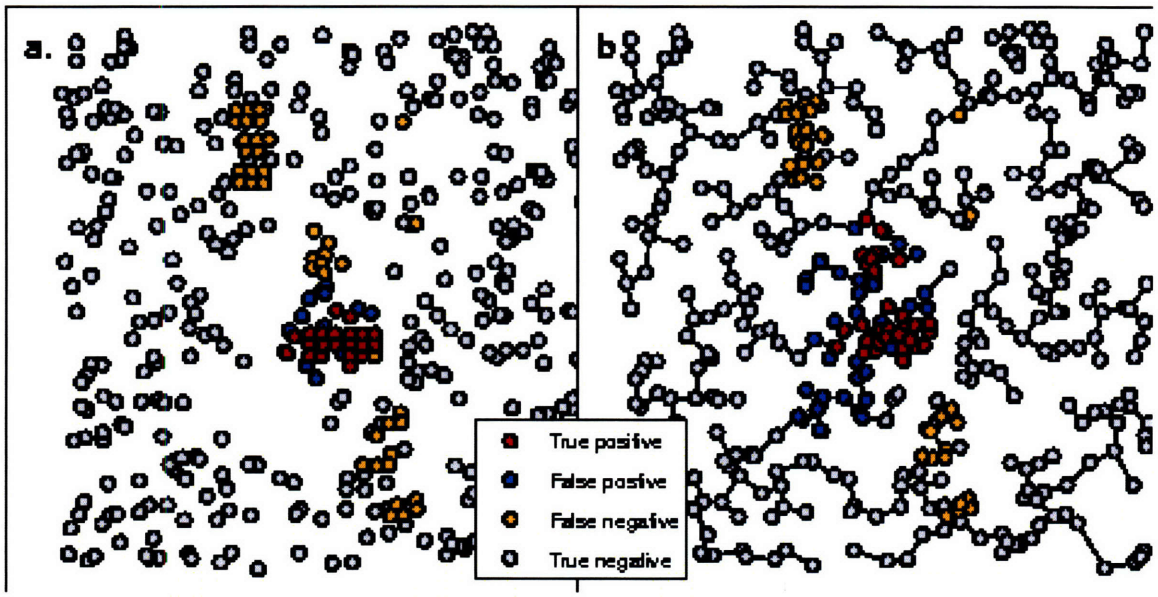


Figure 2-4: SaTScan and EMST Detection of 1979 Sverdlovsk anthrax outbreak. a) A representative data set of 63 anthrax cases (red and orange) and 400 uniformly distributed background cases (blue and gray) is shown, along with the most likely cluster determined by SaTScan (red and blue). b) The EMST method most likely cluster (red and blue) is shown for the same data set, connected by the minimum spanning tree of the cartogram-transformed cases (black edges).

thousand controls were distributed on the map in proportion to zip code population data from the 2000 U.S. census. Data sets of 500 total cases were created, each containing a synthetic circular cluster in a random location with a radius of 1, 2 or 3 km. placed within the study region. We defined the *relative cluster density* to be the case density within the cluster divided by the case density outside the cluster. This ratio varied from 2 to 5 in the data sets. For each combination of outbreak radius and relative cluster density, 1000 data sets were created.

For small clusters containing on average fewer than 35 cases, the EMST method had greater sensitivity. However, it is likely that stochastic effects caused such clusters to have non-circular shapes in general. Indeed, the smaller the cluster, the more pronounced the EMST method's relative improvement in sensitivity. For larger clusters, the EMST method had similar sensitivity to SaTScan (0.1% less to 4.1% greater) and similar values of F_{TC} (3.4% less to 0.4% greater). However, SaTScan always had a greater value of F_{MLC}, indicating that it located large circular clusters with greater

Parameters			SaTScan			EMST			Comparisons		
r	d	m	SN	F_{TC}	F_{MLC}	SN	F_{TC}	F_{MLC}	Δ SN	Δ F_{TC}	Δ F_{MLC}
1	2	8.2	0.03	0.03	0.39	0.07	0.06	0.22	+112	+128	-42
1	3	12.9	0.23	0.21	0.75	0.29	0.26	0.54	+25	+26	-28
1	4	16.3	0.45	0.41	0.84	0.49	0.45	0.66	+7.1	+8.3	-21
1	5	20.8	0.65	0.61	0.89	0.69	0.65	0.73	+5.7	+6.4	-17
2	2	33.7	0.30	0.25	0.79	0.39	0.30	0.59	+27	+20	-25
2	3	50.1	0.79	0.73	0.91	0.81	0.73	0.76	+2.3	-0.3	-17
2	4	64.4	0.94	0.89	0.94	0.95	0.90	0.82	+1.1	+0.4	-13
2	5	75.7	0.99	0.95	0.96	0.99	0.95	0.86	0.0	-0.3	-10
3	2	79.5	0.74	0.65	0.86	0.77	0.63	0.72	+4.1	-3.4	-17
3	3	108.9	0.98	0.93	0.95	0.99	0.92	0.82	+0.8	-2.0	-13
3	4	133.0	1.00	0.97	0.97	1.00	0.96	0.88	-0.1	-1.1	-9.8
3	5	153.8	1.00	0.98	0.98	1.00	0.97	0.91	0.0	-0.8	-7.3

Table 2.3: SaTScan and EMST method applied to circular clusters. r, radius of cluster in kilometers; d, relative cluster density; m, mean cluster size; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference.

overall accuracy than the EMST method. Table 2.3 summarizes the results.

2.3.4 Rectangular Clusters, Boston, Massachusetts

In a study of rectangular clusters, we found that the EMST method had greater sensitivity than SaTScan. Sets of 500 cases containing artificial rectangular clusters having a height-to-width ratio of 1, 4 or 16, and relative cluster density between 2 and 5 were generated within the same study region as above, and 10,000 controls were distributed in proportion to the background population as above. The cluster area was fixed at 20 km^2, and 1000 data sets were generated for each combination of parameters by randomly placing a rectangular cluster within the study region map. The results are summarized in Table 2.4.

In general, the EMST method had greater sensitivity than SaTScan (0.2% less to 166% greater), with the greatest percent increase in sensitivity when the cluster signal strength was weak or the height-to-width ratio was large. The EMST method captured a greater extent of the true cluster (F_{TC}) than SaTScan for all cluster types

Parameters		SaTScan			EMST			Comparisons		
r	d	SN	F_{TC}	F_{MLC}	SN	F_{TC}	F_{MLC}	$\Delta\ SN$	$\Delta\ F_{\mathrm{TC}}$	$\Delta\ F_{\mathrm{MLC}}$
1	2	0.56	0.47	0.82	0.61	0.50	0.65	+8.2%	+6.0%	-20%
1	3	0.92	0.82	0.90	0.95	0.86	0.78	+3.2%	+4.7%	-13%
1	4	0.99	0.91	0.93	0.99	0.94	0.85	-0.2%	+2.6%	-8.9%
1	5	1.00	0.93	0.95	1.00	0.97	0.88	+0.2%	+4.5%	-7.3%
4	2	0.43	0.26	0.69	0.58	0.42	0.62	+36%	+63%	-10.0%
4	3	0.95	0.64	0.77	0.97	0.86	0.74	+2.2%	+34%	-4.4%
4	4	1.00	0.73	0.79	1.00	0.95	0.80	+0.1%	+29%	+0.4%
4	5	1.00	0.78	0.81	1.00	0.97	0.84	0.0%	+25%	+3.2%
16	2	0.21	0.06	0.66	0.55	0.31	0.52	+166%	+419%	-21%
16	3	0.82	0.25	0.72	0.98	0.74	0.60	+21%	+199%	-17%
16	4	0.99	0.31	0.76	1.00	0.86	0.67	+0.9%	+177%	-11%
16	5	1.00	0.35	0.77	1.00	0.93	0.73	0.0%	+166%	-6.0%

Table 2.4: SaTScan and EMST method applied to rectangular clusters. r, ratio of cluster height to width; d, relative cluster density; SN, average sensitivity; F_{TC}, average fraction of true cluster detected; F_{MLC}, average fraction of most likely cluster coinciding with the true cluster (averaged over data sets for which a significant cluster was found); Δ, percent difference.

(2.6% to 419% greater). For most cluster types, there was a parallel decline in the fraction F_{MLC} of the most likely cluster coinciding with the true cluster (20% less to +3.2% greater).

2.3.5 Arbitrary Shapes

It is possible to gain insight into the EMST method's performance on other cluster shapes without additional intensive computer simulations. The EMST test statistic depends only on the cartogram, the total number of cases, and the cardinality and weight of a potential cluster. Hence, we can extrapolate the p-value obtained for one potential cluster to others having different shapes, but the same number of cases and weight. To illustrate this, we selected one most likely cluster of 35 cases from one of the Boston analysis data sets. The EMST method assigned a p-value of 0.0001 to this potential cluster. Figure 2-5 shows several configurations of potential clusters having the same number of cases and EMST weight, but very different shapes. If embedded as potential clusters within a Boston data set of 500 total cases, they would each

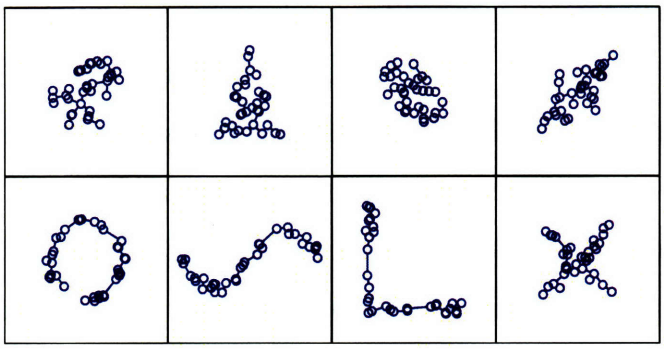


Figure 2-5: Equally detectable potential clusters of various shapes. A most likely cluster of 35 points selected from among the Boston circular cluster data sets, along with its minimum spanning tree, is shown in the upper left. Seven other configurations of 35 points, having minimum spanning trees with exactly the same weight, are also shown. Subject to the constraint imposed by the definition of a potential cluster above, all eight clusters have equivalent detectability by the EMST method. If embedded as potential clusters in a Boston data set of 500 total cases, all would achieve the same p-value of 0.0001.

achieve the same p-value of 0.0001. In fact, any potential cluster of 35 cases of any shape can be scaled in size to have the same weight, illustrating that the method can capture an infinite array of regular and irregular shapes.

2.4 Discussion

We find that the EMST method is a powerful and accurate alternative to the circular scan statistic for non-circular clusters. At a specificity of 95%, the method had comparable sensitivity to SaTScan applied to large synthetic circular clusters and to an approximately circular West Nile virus outbreak. When applied to small circular clusters, synthetic rectangular clusters, and a highly irregular anthrax cluster, the EMST method had greater sensitivity. Although SaTScan had better accuracy detecting large circular clusters, the EMST method had comparable or superior accuracy for all other cluster types. The EMST method is also able to detect a large variety of shapes, including highly irregular ones.

In addition to accurately locating clusters of any shape and size, the EMST method has two unique properties. First, its test statistic is based only on the weight

of the potential cluster subgraph. To our knowledge, all other tests that provide the location of any detected clusters while allowing the user to set the level of significance for the test utilize the likelihood ratio test statistic developed by Kulldorff and Nagarwalla [14]. This test statistic requires the area of each region considered, which in turn requires a precise definition, including the shape, of the region. Second, we formally define a cluster in mathematical terms that are independent of cluster geometry, and which depend only on intercase distances. Traditionally, clusters are often imprecisely defined; for example, Knox's frequently cited definition is "a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance" [32].

Of other cluster detection methods designed to capture clusters of any shape, the EMST method is most similar mathematically to the upper level set method of Patil and Taillie [21], which examines a well-defined family of contiguous administrative regions with high relative rates. Assunção *et al.* [20] used non-Euclidean minimum spanning trees of a graph with different vertices, edges and edge weights to consider contiguous administrative regions having similar disease rates, whether high or low. By contrast, we locate sets of individual cases corresponding to a mathematical formalization of a cluster, using specific subsets of the EMST.

General tests of clustering [8] such as Tango's maximized excess events test [33], and disease mapping methods, such as Bayesian partition models [34, 35], kriging [36], and generalized additive models [37, 38], handle arbitrary geometric configurations of cases without difficulty. However, these address separate problems within spatial epidemiology, and comparison of clustering and disease mapping methods to cluster detection methods is not straightforward [39].

The EMST method can easily be extended to analyze regional summary data, consisting of counts of observed and expected disease cases for each region on a map. A cartogram is constructed to equalize the density of expected disease cases, and each observed case is randomly placed on the cartogram within its region of occurrence. After constructing the cartogram, the procedure for case-control data is followed.

One limitation inherent in this and other methods for aggregated data is that

exact spatial locations are not used, which decreases cluster detection sensitivity and accuracy [40]. This is also a limitation for the procedure detailed above for case-control data, since a loss of spatial information is incurred by randomizing cases within their regions of occurrence on the Voronoi diagram cartogram. Because the expected area of each region on the cartogram tends toward zero as the number of control locations increases, this loss can be minimized by increasing the number of controls. For 10,000 distinct controls on a square map, as used in our study, the loss of spatial information is modest; each case is expected to move approximately 1% of the length of one side of the square.

We found that the EMST method gains in F_{TC} for non-circular clusters were partially offset by a decline in F_{MLC}, indicating that the EMST method reports fewer false negatives, but more false positives, than SaTScan. The relative cost to society of false negatives and false positives depends on many factors. The cost of false negative cases includes, for example, an increased risk of spread of a disease and the possibility that infected individuals who are unaware of the outbreak may not seek early treatment for symptoms, while the cost of false positive cases includes unnecessarily investigating and alarming the community.

In retrospective research and prospective surveillance, the shape of true clusters are not known *a priori*. Thus, in most cases, a method that is able to detect clusters of any shape is preferable. Hence the EMST method may represent a practical adjunct to methods currently used in public health practice.

Chapter 3

Cartograms for Mapping and Analyzing Event Disease Data

3.1 Introduction

From the earliest disease dot map [5] to information-rich modern maps such as the annual U.S. cancer atlases, representations of the spatial distribution of diseases have flourished. Disease maps serve several functions: describing a disease prior to more rigorous statistical study, identifying areas of increased risk or features that may be missed by mechanical mathematical analysis, and even suggesting etiology and control strategies. Dot maps showing the exact locations of disease cases compromise privacy [41] and do not account for variation in the underlying population. This has motivated the use of choropleth and isopleth maps, depicting the average risk in administrative regions, and smooth risk functions, respectively. Disease maps typically use one of several standard cartographic projections, in which areas on the map reflect the surface area of regions represented, with the degree of distortion depending on the projection. Thus a map of disease risk shows an approximation to the amount of

Joint work with John S. Brownstein, Karen Olson, Athos Bousvaros, Bonnie Berger and Kenneth D. Mandl

land area in each region of increased risk. However, as noted by Dorling [42], diseases infect people, not land. A standard map showing an increased relative risk confined to a small area does not distinguish between a major outbreak in a dense metropolis and a few cases in a rural community.

Cartograms have recently been introduced to capture this distinction. A cartogram, or density-equalizing projection, is a distorted map in which the area of each region is proportional to some quantity associated with the region, such as its population. Cartograms based on total census population have been used to simultaneously depict the relative risk and the total population affected by leukemia [27], lung cancer [29], cryptosporidiosis [28] and childhood cancers [25, 26, 43]. Cartograms have also been used to test for global clustering of disease cases [43, 27, 25].

Although the recent development of an efficient algorithm to create minimally distorted cartograms [29] has increased the practical possibilities for disease mapping and analysis, the use of cartograms has been limited in scope. First, disease cartograms have usually been based on estimates of total population taken from census data. In contrast, traditional epidemiological studies accommodate a multitude of methods for defining the underlying population at risk. Second, cartograms used for mapping and global clustering studies have only been constructed from count data, in which the number of cases and the underlying population size are aggregated by administrative regions such as counties. Aggregation results in a loss of spatial information, limiting the power for statistical analysis [40] and the ability to see trends. Exact point location data (usually termed "event" data [44]) is increasingly available due to clinical databases and fast geocoding software. The use of event data to create cartograms is an unexplored alternative to count data.

In this study, we extend the use of cartograms of event data to other areas of spatial epidemiology including disease mapping and global studies of clustering (see figure 3-1). We show how cartograms based on event data can be used to visualize and analyze several types of traditional epidemiological studies. For disease mapping, there is a simple relationship between the density of cases on the cartogram and the risk of the disease. For statistical tests related to clustering, the null hypothesis is

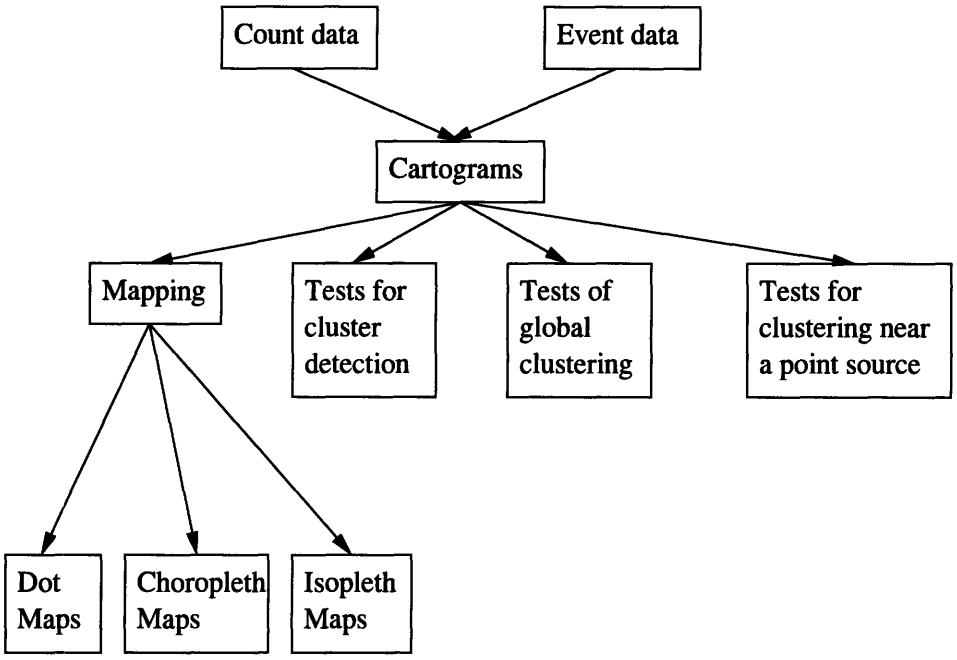


Figure 3-1: Applications of cartograms to spatial epidemiology.

simplified by the cartogram transformation. This makes it possible to apply existing statistics developed for other fields to spatial epidemiology problems. We illustrate the use of cartograms based on event data for disease mapping using several simulated distributions of cases and controls. We also apply the method to cases of pediatric inflammatory bowel disease (IBD) in Massachusetts between 1995 and 2006, finding an increased relative risk in the southern half of the state.

3.2 Event Cartograms

3.2.1 Data

We begin with a set of two-dimensional coordinates on a map, each having a binary label (case or non-case) representing disease status. Ideally, this would consist of all the members of the population at risk for the disease in a study region. Since this is not feasible in practice, the data may be collected using a standard epidemiological study design including:

1. a case-control study, in which the exposures of a group having the disease are

compared to those of a group without the disease.

2. a cohort study, in which a group is monitored (either prospectively or retrospectively) for the development of a disease over time.

3. a cross-sectional study, in which the health and exposure characteristics of a group are measured at a single point in time.

We require the collection process to preserve spatial information about disease risk. A study in which subjects are selected at random from the total population would clearly satisfy this requirement, but may not be possible. As an alternative, a collection process that is independent of spatial location would be suitable. However, even this may not be possible. For example, consider a cohort of patients enrolled at a single clinic monitored for the development of a certain disease. The collection process itself is not spatially neutral; the chance that a patient visits the hospital likely depends on his or her distance from it. However, the study may still preserve spatial information about risk if, conditional on location, the odds of entering the study for those with and without the disease is a constant. The precise requirement is:

$$\frac{p(\text{enter study}|\text{case, location } L)}{p(\text{enter study}|\text{non-case, location } L)} = c. \tag{3.1}$$

For cohort and cross-sectional studies, subjects enter the study independent of disease status, so $c = 1$. For case-control studies, c is the ratio of the total number of cases to controls.

Continuing the example above, the cohort study taking place at a single clinic may or may not satisfy this requirement. For example, if the clinic specializes in the treatment of the disease under study, its reputation may draw those with symptoms consistent with the disease from a larger area than patients without such symptoms. Patients close to the clinic may visit regardless of disease status, but distant patients may be more likely to choose the clinic if they have the disease.

A more common example of a study that does not preserve the spatial risk structure is a matched case-control study, as previously noted [45]. Choosing controls to

match the potential confounding characteristics of the cases is problematic because the spatial distribution of the matched controls may not be the same as a random sample drawn from the total population of all non-cases.

3.2.2 Event cartogram construction

To create a cartogram from event data, a Voronoi tesselation (or Voronoi diagram) is first created from the set of non-cases. (We will refer to the non-cases as controls for simplicity, without loss of generality.) First described by Voronoi in 1908 [46], Voronoi tesselations have found applications in diverse fields including forestry [47], operations research [48], and computational geometry [49], as well as epidemiology [50, 51, 35]. Given n controls $c_1, c_2, \ldots c_n$, the Voronoi tesselation partitions a map M into regions called Voronoi cells, consisting of the part of the map falling closest to each control:

$$\text{vor}(c_i) = \{p \in M \mid d(p, c_i) \leq d(p, c_j) \; \forall \; j = 1, \ldots n\} \text{ [24]}. \tag{3.2}$$

In regions with a high density of controls, the Voronoi cells tend to be smaller than in low density regions. Figure 3-2 shows an example of a Voronoi tesselation.

We assign a population to each Voronoi cell by counting the number of controls it contains. This number is usually one, but it may be greater if multiple controls may occur at the same location on the map. For example, multiple family members in a house or inhabitants of an apartment building may share the same geocoded location. Next, we use the Voronoi tesselation map to create a cartogram, on which the area of each projected cell is proportional to the number of controls in the original Voronoi cell. Hence the density of the controls is uniform throughout the cartogram. We use a fast diffusion-based cartogram construction algorithm developed by Gastner and Newman [29] that produces minimal distortion. The algorithm simulates diffusion of the population from regions of high density to low density regions, carrying map boundaries along during the process, until the density is equalized. Other contiguous cartogram construction algorithms may also be used.

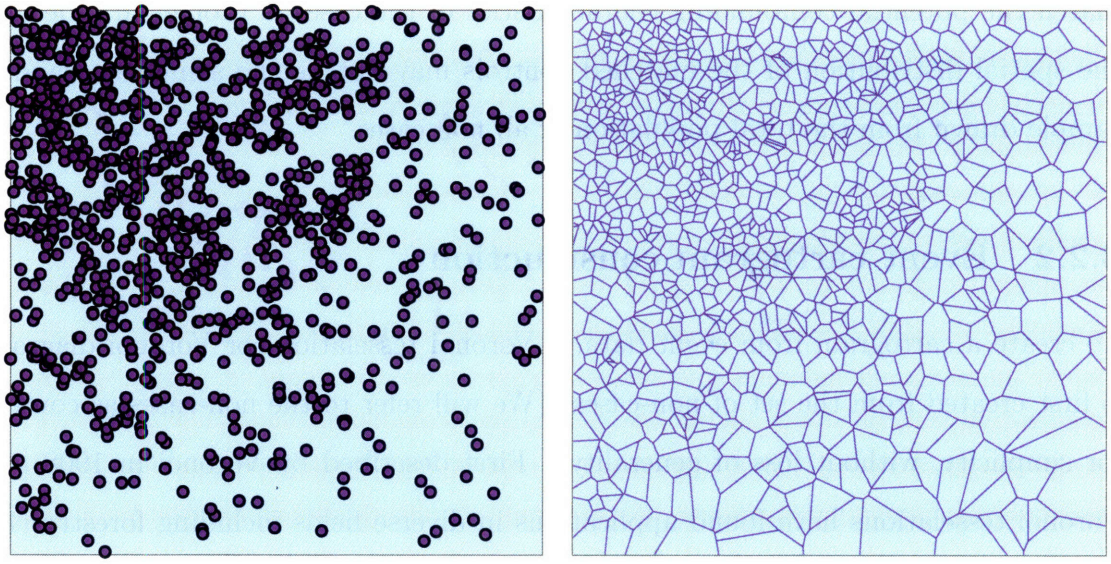


Figure 3-2: Example of a Voronoi tesselation. Left: One thousand points are distributed on a map. Right: The Voronoi tesselation of the points divides the map into 1000 regions. Each region consists of the portion of the map closest to one point. The density structure is preserved in the tesselated map; regions of small Voronoi cell area correspond to high point density.

As Bithell [52] notes, computational artifacts may arise during cartogram construction that render the cartogram unsuitable for subsequent statistical analysis. These include a failure to produce an equalized map; we apply the diffusion algorithm iteratively until the control density differs by no more than 1% over the entire map. Because an initial step of the algorithm involves replacing the continuous density function with a discrete set of values on a grid, small regions of very high or low density may be missed entirely in the digitization. If this occurs, the software may fail to contract or expand these regions. To avoid this, we position the digitization grid so that the region of highest or lowest density is always captured. Finally, Voronoi cells may be specified by very few vertices, and as the cartogram is created, edges which should bend in space remain straight and may cross one another; we eliminate this possibility by adding additional vertices on the Voronoi edges until crossing edges are not encountered.

After constructing the cartogram of the Voronoi tesselation, each case is plotted on the cartogram in a random location within the region corresponding to its Voronoi

cell on the map. Although it is possible to show cases in their exact cartogram-transformed locations, they must be randomized because density is not equalized within each Voronoi cell, as noted by Merrill [26]. Randomization represents a loss of information, but the loss is small given sufficient numbers of distinct controls. Indeed, it tends to zero as the number of distinct control locations increases.

3.2.3 Mapping the disease risk

The density of transformed cases on the cartogram is proportional to the risk of the disease. To see this, consider the set $\{R_i\}_{i=1}^{n_R}$ of $n_R \leq n$ Voronoi cells on the original map M. Let a_i and b_i denote the number of cases and controls, respectively, in region R_i for $i = 1, \ldots, n_R$. For cohort and cross-sectional studies, the quantity $\frac{a_i}{b_i}$ estimates the odds of the disease in region R_i. For case-control studies, the relative risk of the disease is approximated by $\frac{n}{m} \cdot \frac{a_i}{b_i}$, where m is the total number of cases.

Let $\tilde{R}_i$ denote the transformation on the cartogram of the region R_i for each i, and let

$$\delta(\tilde{R}_i) = \frac{\#\text{ cases in } \tilde{R}_i}{\text{area of } \tilde{R}_i} \tag{3.3}$$

denote the density of cases in the cartogram region $\tilde{R}_i$. The number of cases in $\tilde{R}_i$ is equal to the number of cases a_i on the original map falling into the Voronoi cell R_i, by the procedure outlined above. Because the cartogram scales areas to equalize the density of controls, the area of region $\tilde{R}_i$ is proportional to the number b_i of controls in R_i. That is,

$$\text{area}(\tilde{R}_i) = c \cdot b_i, \tag{3.4}$$

where c is a constant. The area of the entire cartogram $\tilde{M}$ is also proportional to the total number of controls, so

$$\text{area}(\tilde{M}) = c \cdot \sum_{j=1}^{n_R} b_j = c \cdot n. \tag{3.5}$$

The Gastner and Newman cartogram procedure preserves the total map area, so

$\text{area}(\tilde{M}) = \text{area}(M)$. Solving for the constant of proportionality c gives

$$\text{area}(\tilde{R}_i) = \frac{\text{area}(M) \cdot b_i}{n}. \tag{3.6}$$

This is simply the observation that $\frac{b_i}{n}$ of the controls lie in region R_i, so $\tilde{R}_i$ occupies $\frac{b_i}{n}$ of the total cartogram area. Thus the density of cases on the cartogram is

$$\delta(\tilde{R}_i) = \frac{n}{\text{area}(M)} \cdot \frac{a_i}{b_i}. \tag{3.7}$$

Hence for cohort and cross-sectional studies, the odds of the disease is approximately the density of cases on the cartogram divided by the average density of *controls* on the original map. For case-control studies, the relative risk of the disease is approximately the density of cases on the cartogram divided by the average density of *cases* on the original map.

3.3 Examples

3.3.1 Simulated Distributions

We first illustrate the use of cartograms of event data for visualizing hypothetical diseases having known relative risk surfaces. We generated cases and controls for three examples: a constant risk surface, risk increasing with latitude, and a circular cluster. Five thousand controls, used for all three examples, were randomly distributed on a map of the continental United States in proportion to 2000 census county population estimates [53]. To generate each control, a zip code was drawn from a multinomial distribution with probability proportional to its census population. A map location for the control was then randomly selected from a uniform distribution within the zip code boundary.

To illustrate constant risk, 2500 cases were distributed in proportion to the underlying population as above. Figure 3-3a shows the distribution of cases and controls. Figure 3-3b shows the cartogram derived from the Voronoi diagram of the controls,

as well as the transformed locations of cases. The cases appear to be uniformly distributed, consistent with a relative risk of one everywhere on the map. Figures 3-4a and 3-4b show isopleth relative risk surfaces for the same data set. The surface on the map was created by dividing a kernel case density estimate by a kernel control density estimate [54]. There are regions with extremely high relative risk estimates on the order of 1000. The cartogram surface estimate, also created using kernel methods, is much closer to the true relative risk of one everywhere.

The second example illustrates a smooth transition in relative risk. The relative risk of the 2500 cases increased linearly by a factor of four from south to north. A north-south gradient was chosen to illustrate diseases with latitude-dependent risk, such as stroke, multiple sclerosis and melanoma. Figures 3-3c and 3-3d show the map of cases and controls, and the cartogram-transformed case locations. The change in relative risk is difficult to detect on the original map, but suggested by the cartogram-transformed cases. Figures 3-4c and 3-4d show the corresponding isopleth risk surfaces. The latitudinal gradient is apparent on the cartogram, but is hard to detect on the standard map due to the presence of multiple locations having extremely high relative risk estimates.

The third example illustrates the presence of a localized cluster in the data. Within the 2500 cases, a cluster was generated to roughly approximate the geographic distribution of a large multi-state mumps outbreak occurring in 2006 [55, 56]. The circular cluster was centered in Iowa and involved neighboring states. The relative risk inside the cluster was three times the risk outside the cluster. The cases could represent, for example, national syndromic surveillance of fever, with controls drawn from all syndromes. The cluster is difficult to discern on a dot map of cases and controls (figure 3-3e). It appears on an isopleth surface created using a standard map (figure 3-4e), but is accompanied by artifactual regions of very high relative risk. The cluster is clearly apparent on a cartogram scatter plot of cases (figure 3-3f), and is captured with fidelity on a cartogram isopleth risk surface (figure 3-4f).

3.3.2 Pediatric Inflammatory Bowel Disease, Massachusetts, 1995-2006

We also used the cartogram method to visualize the spatial distribution of IBD among the pediatric population in Massachusetts, finding an increase in relative risk in the southern portion of the state. Cases were drawn from patients diagnosed with IBD at any of 10 pediatric gastrointestinal clinics affiliated with an urban tertiary care pediatric hospital in Boston, Massachusetts. All 1163 patients who were diagnosed between January 25, 1995 and March 21, 2006 were considered. Of these, the addresses of 87 patients could not be geocoded. Of the remaining 1076 patients, the cases were selected to be the 901 patients residing in Massachusetts. Controls were drawn from clinic patients (exclusive of the cases) seen during the same time period who were diagnosed with ICD9 codes 564.00 (constipation), 789.00 (abdominal pain, unspecified site) or 564.1 (irritable bowel syndrome). Of these, the 7988 patients having addresses within Massachusetts were selected to be controls. The cartogram and new case locations are shown in Figure 3-5a. The density of cases on the cartogram is increased in the southern part of the state, reflecting an increased relative risk.

3.4 Discussion

Cartograms of event data, based on Voronoi diagrams of control locations, produce maps of the variation in disease risk across space that can be used for visual displays or statistical analysis. Visual interpretation of the cartogram is straightforward: the density of cases reflects the odds or relative risk, and area reflects the size of the population at risk. The method requires no user-specified input parameters, and it minimizes the loss of spatial information. It is applicable to analyses on any scale, from local epidemiological studies to national initiatives such as BioSense and CDC influenza surveillance.

Previous applications of cartograms to disease mapping and analysis have been limited to count data, aggregated by administrative regions such as census tracts

[27, 28, 25, 26, 43] or zip codes [29]. This is a reflection of the general disease mapping literature; event data have received less attention [52] despite their greater precision and information content. It is conceivable that count data may in fact be superior because of the larger baseline of census population data. However, many epidemiological studies cannot draw baseline information from census data. For example, a cohort selected for inclusion in a study may have characteristics differing from available census categories. Furthermore, the use of event data for cluster detection has been shown to improve sensitivity compared to count data [40].

Alternative methods of creating cartograms from event data exist. In fact, any method of estimating a continuous or discrete control density function from event data may be used as the input for a cartogram. One example is binning by fixed administrative regions [28, 26]. This is essentially an aggregate method since it converts event data into count data, and hence it loses spatial information. Our proposed method "bins" data into dynamic regions that usually contain only one control. If a different estimation method is used, it must be unbiased. That is, the estimate of the mean density of controls in each region must tend to the exact distribution as the number of controls increases. Otherwise, the null hypothesis of constant, independent individual risk does not lead to complete spatial randomness of the cases on the cartogram.

Because there are existing methods for mapping and analyzing disease event data that do not involve cartograms, we wish to consider the benefits and drawbacks of using cartograms compared to these methods. In the case of mapping, cartograms simultaneously depict the risk of the disease and the size of the underlying population; there is no straightforward way for standard methods to achieve this. However, the cartogram's distortion of familiar map boundaries may detract from its visual appeal. Either cartograms or standard methods may be used to create dot, choropleth, or isopleth maps. Standard dot maps show where the preponderance of cases occurs, reflecting the burden on health care infrastructure. Although commonly used (for example, [57, 58, 59, 45, 60]), it is difficult to visually compare the spatial distributions of cases and controls to assess variation in risk. In addition, dot maps

may compromise patient privacy; maps containing minimal spatial references may be reverse-geocoded, frequently to the exact addresses of patients [61, 41]. Cartograms suffer from the same limitation if any spatial features, such as peninsulas, islands and sharp corners in administrative boundaries, can still be recognized after the cartogram transformation. Appreciating spatial variation in risk on the cartogram requires discerning deviations from complete spatial randomness. Although large deviations are obvious (for example, in figure 3-3f), the eye is not expert at this task, tending to find patterns where none exist.

Choropleth maps bin event data by pre-existing administrative regions, and use shading or color to represent the odds ratio [62] or prevalence [63] in each region. In addition to the loss of information incurred by binning, several problems with choropleth maps have been noted in the literature, which apply equally to standard and cartogram methods. Administrative boundaries are unrelated to the disease process. Their irregularity may make the map difficult to interpret, and the choice of administrative groupings used may change the appearance of the map [36]. Regions with smaller denominators are subject to greater random variation, producing more extreme rates [64]. Furthermore, sparsely populated areas are often clustered together on the map, compounding the visual effect. The gray or color scale used to produce the map may strongly affect the visual appearance [7]. In choropleth maps using standard projections, large areas may dominate the map, irrespective of the underlying population size [36].

Isopleth maps, which estimate a smooth risk surface throughout the map, are an alternative to choropleth maps. Although many methods exist for creating standard isopleth maps from count data (for example, [36, 34]), the methods for event data are limited in number. Bithell used separate kernel estimates of case and control densities at each point, dividing to estimate the odds function [54]. Kelsall and Diggle used kernel estimates of the conditional probability of being a disease case, given location [38]. These methods do not allow adjustment for covariates. To address spatial confounding, which occurs when risk factors for a disease are not uniformly distributed in space, Kelsall and Diggle also introduced the use of generalized additive models

that adjust for covariates. They modeled the log odds as the sum of a nonparametric function of the location and a parametric function of the covariates. One limitation of these methods is that under the null hypothesis of constant individual risk, the variance of the estimate changes throughout the map. In regions of high density, the variance is lower than in sparse areas.

On the cartogram, the density of cases may be estimated at each point to derive the odds (for cross-sectional and cohort studies) or relative risk (for case-control studies), giving rise to an isopleth surface. Kernel density estimation is a natural choice, which would result in a variation of Bithell's method in which the bandwidth is the same for the numerator and denominator, but adjusted (relative to the original map) at each point depending on the local density of controls. This is preferable because the variance of the estimate at each point is constant. Generalized additive models can also be used in this setting to estimate the density while adjusting for covariates. This improves the procedure proposed by Kelsall and Diggle by equalizing the variance of the estimate.

The use of cartogram-transformed data for statistical analysis may be complicated by computational artifacts arising during the cartogram construction [52]. We verify that the cartogram does indeed equalize the density of controls to within 1% throughout the map. The null hypothesis of constant, independent individual risk implies that cases are uniformly and independently distributed on the cartogram. For studies of clustering, this is a simplified version of the usual problem in which the underlying at-risk population has uniform density. In many cases, a statistical test used for the general problem can be simplified for use on cartogram-transformed data. However, the main benefit of using cartograms over standard methods for statistical tests of deviation from constant disease risk is that it enables the use of a host of methods that have been developed to analyze a single point process (cases) instead of two point processes (cases and non-cases). Because the problem is easier, the body of such methods is farther advanced than methods for two point processes.

For example, tests for the detection of clusters find localized "hotspots" of disease. The most commonly used test is the scan statistic developed by Kulldorff and

Nagarwalla [14, 15], which evaluates the statistical significance of a large number of circular regions using a likelihood ratio test. The cartogram transformation does not preclude subsequent use of the scan statistic; it may still be applied to test the null hypothesis that cases arise from a homogeneous Poisson process. However, in addition to standard cluster detection tests, the cartogram transformation allows the application of tests of deviation from a uniform null distribution. An example is the graph-theoretical cluster detection method developed in Chapter 2 to detect clusters of any shape. There are no corresponding methods available for two point processes to detect arbitrarily shaped clusters in event data, so the cartogram is necessary.

Cartograms are also useful for tests of global clustering, which detect the general tendency of cases to cluster in space. Of the methods developed to compare two point processes on standard maps, most suffer from the problem of multiple hypotheses due to a variable parameter related to the scale of clustering. This makes it difficult to calculate the probability of type I error. One exception is a statistic proposed by Tango [33], which is the minimum p-value of a parameterized statistic that sums an index of goodness-of-fit and one of autocorrelation [65, 66]. The test is for count data, so event data must first be aggregated to use it. Simpler statistics such as the mean nearest-neighbor distance and Ripley's K-function are available to test for global clustering of event data on the cartogram. Selvin and Merrill [27] developed a statistic for cartogram-transformed data based on concentric polygons, defined to be polygons within a boundary polygon having the same centroid and parallel boundaries. They first ordered the cases on the cartogram by their distance from the cartogram boundary, from furthest to nearest. Then for each i, they calculated the proportion P_i of the cartogram area contained in the concentric polygon passing through the ith case. The value of P_i approximates the value of the uniform cumulative distribution function of a uniform random variable on the interval $[0, 1]$. The expected value is $\frac{i}{n}$. The expected and observed distributions were compared using the Kolmogorov test.

We illustrated the use of cartograms for making dot and isopleth maps with a preliminary analysis of an IBD case-control study. IBD is characterized by chronic gastrointestinal inflammation believed to result from immune dysregulation. The

term encompasses both ulcerative colitis (UC), which affects the large intestine, and Crohn's disease (CD), which may affect the entire gastrointestinal tract. Both cause a range of symptoms, such as diarrhea, weight loss, fever, abdominal pain, bleeding, and bowel obstruction. They are also associated with a variety of extra-intestinal manifestations, such as arthritis, skin nodules, and thromboembolic disease [67].

IBD is an interesting disease to study for clustering because its epidemiology is not completely understood. Linkage analysis studies have implicated regions on several chromosomes in susceptibility to UC and CD [68], with the most strongly associated genetic mutation occurring on a gene involved in the innate immune response [69]. The rates of concordance among identical twins with UC and CD are 20 [70] and 67 [67] percent, respectively, indicating that environmental factors play a significant role. Several factors are known to modify the genetic risk of IBD, most notably cigarette smoking, appendectomy [71] and the use of oral contraceptives [67]. Other potential risk factors, such as childhood hygiene, Mycobacterium paratuberculosis infection, childhood viral infections, diet, and blood transfusions have shown weak associations with IBD or have not been well studied [72]. In general, the environmental IBD risk modifiers are poorly understood, and further epidemiological study is needed.

Previous IBD clustering studies have yielded some interesting preliminary results. There is evidence for concordance among spouses and friends [73], and one study of clustering at a global level among Swedish IBD cases found that patients were likely to have been born closer together in time and space than expected by chance [74]. No evidence of global clustering was detected in a study of CD cases surrounding Nottingham, England [75]. A seasonal pattern of IBD hospital admissions and a greater prevalence of IBD in the northern U.S. than in the south were found by Sonnenberg *et al.* [76]. A study of focused clustering concluded that a suspected cluster of four cases in an eight-year period in one district of France was statistically significant [77]; another found increased prevalence in a region of Ireland suspected to have a high burden of IBD [78]. While such focused studies are potentially illuminating, the investigation of previously suspected clusters raises statistical concerns [64].

Our study found an increased density of cartogram-transformed IBD cases in the southern portion of Massachusetts. While this may reflect an actual increase in relative risk in the south, it may also result from a violation of the assumption that spatial risk information is preserved by the study collection process, defined in equation 3.1. This could result, for example, from a bias in referral patterns to the gastrointestinal clinics. Our finding, although preliminary, gives cause for further investigation of IBD spatial clustering.

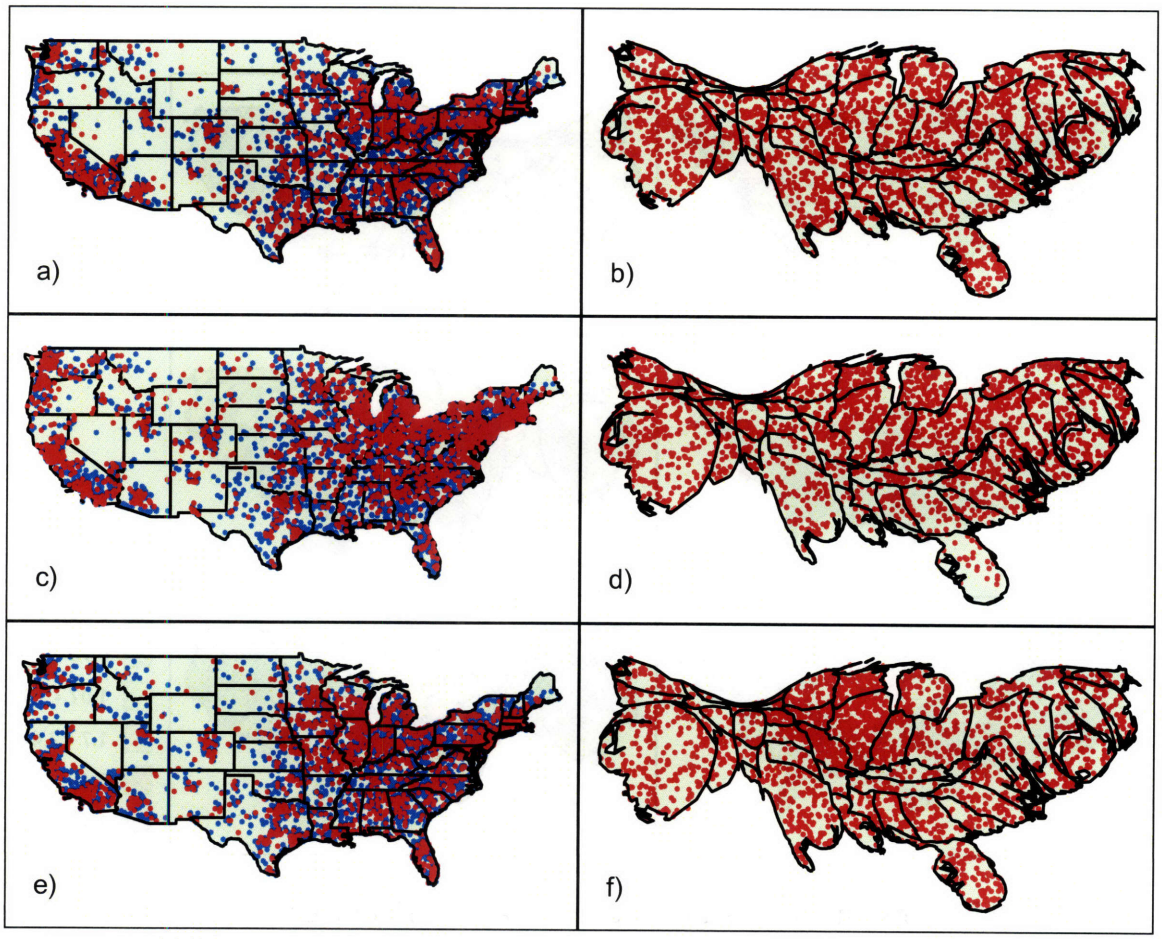


Figure 3-3: Dot maps and cartograms of three hypothetical disease distributions. Dot maps of 5,000 controls (blue) and 2,500 cases (red) are shown in the left column (a, c and e). The controls are distributed in proportion to the underlying population. The cases are distributed to illustrate constant relative risk (a), risk increasing linearly by a factor of four from north to south (c), and a localized cluster with a three-fold increase in relative risk in Iowa and neighboring states (e). The right column (b, d and f) shows the cartogram-transformed case locations for the three distributions.

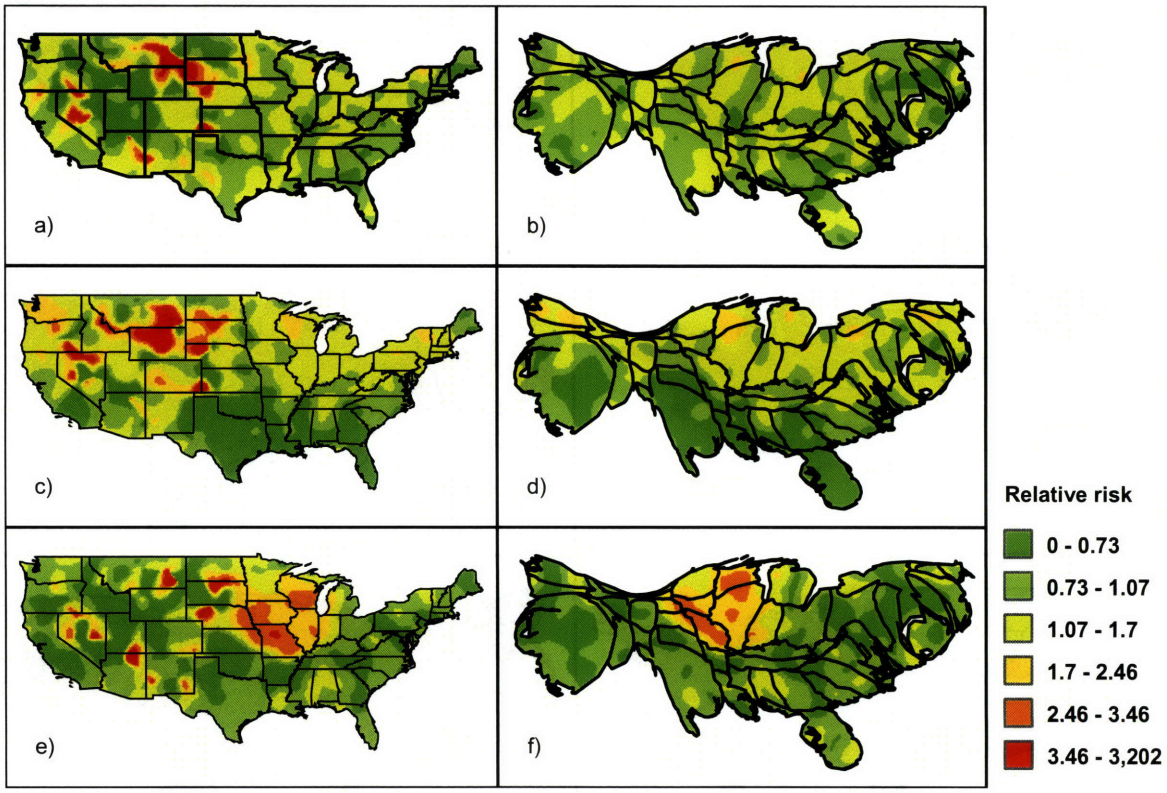


Figure 3-4: Isopleth surfaces estimating the relative risk of three hypothetical disease distributions on standard maps (a,c and e) and cartograms (b,d and f). The exact locations of the cases and controls are shown in figure 3-3. The case distributions illustrate constant risk (a and b), a four-fold increase in risk from south to north (c and d) and a cluster of three-fold risk increase centered in Iowa (e and f). The patterns are obscured on the standard maps because of the presence of high relative risk artifacts, but are clear on the cartograms.

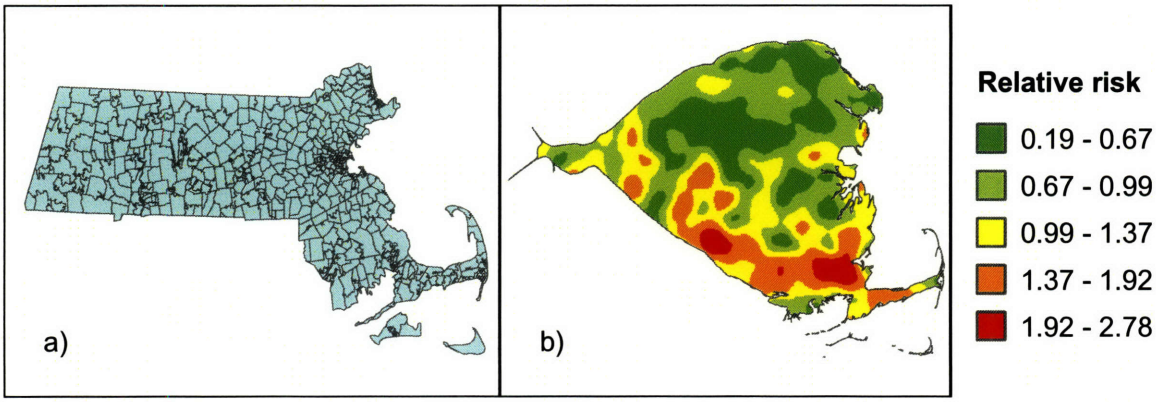


Figure 3-5: Pediatric inflammatory bowel disease risk in Massachusetts, 1995-2006. a) A standard map of the study area. b) A cartogram was constructed from the Voronoi diagram of the 7988 control locations. The 901 IBD cases were randomly placed within the cartogram regions corresponding to their original locations on the Voronoi diagram. An isopleth relative risk surface was calculated from the transformed case locations using kernel methods. Original case and control locations are not shown to protect patients' privacy.

Chapter 4

Optimal anomymization of patient spatial data

4.1 Introduction

Since the first disease dot map was published more than 200 years ago, it has become a staple of the epidemiologist's toolkit. Dot maps are frequently used as an initial step in studying the spatial epidemiology of a disease, revealing associations among the cases and with the landscape. Often published in medical journals, dot maps represent one common and useful instance of sharing information about the geographical distribution of diseases with the scientific community and the public. Spatial data in other formats are also shared between individuals and institutions for many purposes: to identify focal clusters, to study etiological factors influencing disease risk, and ultimately to improve medical care and public health.

Despite their potential for public good, geographical identifiers such as zip codes, street addresses, and locations on maps are highly identifying protected health information that pose a threat to patient privacy. Even coarse identifiers can be linked to individuals: one study found that 87% of subjects could be uniquely identified by

Joint work with Christopher A. Cassa, Kenneth D. Mandl and Bonnie Berger

their gender, zip code and date of birth [79]; another found that low-resolution dot maps of diseases published in several medical journals could be used to trace most cases to single addresses [41].

For this reason, data are frequently aggregated to preserve privacy. Aggregation by zip codes (in the United States) or census enumeration districts (in the United Kingdom) is common for published maps, spatial epidemiological studies, and prospective health surveillance. However, aggregation may erase spatial information useful for research and health surveillance; for example, cluster detection methods to detect spatial clusters are significantly less sensitive and specific when data are aggregated by zip code [40]. Furthermore, in many instances, aggregation by zip code may not be sufficiently privacy-preserving. In the 2000 U.S. census, 1210 inhabited zip codes contained fewer than 100 people, and the least populated of these contained only one person [53].

For research, disclosures of zip codes and other geographical identifiers is limited by the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Under this legislation, zip codes may be released as part of a limited data set, accompanied by a data use agreement. The rule also defines a category of "non-identifiable data sets," whose dissemination is not restricted. Either of two criteria must be met for a data set to qualify. The first specifies that only the first three digits of a zip code are included, provided that at least 20,000 people share the same first three digits. Unfortunately, this is even coarser and less useful for research and public health than zip codes. Furthermore, the level of privacy protection depends on the number of patient records. For example, if it is revealed that 20 patients having a certain disease reside in a region containing 20,000 people, then there is a $\frac{1}{1000}$ chance that a randomly selected individual from the region is one of the patients. However, if 200 patients with the disease live in the region, then the probability that a random individual from the region is among the set of patients increases to $\frac{1}{100}$.

The second criterion specifies that "the risk is very small that the information could be used, alone or in combination with other reasonably available data, to identify an individual" [80]. In line with this, several strategies have been developed to

create a de-identified data set by applying a spatial transformation to each patient in the original set. These include the family of "geographical masks" [81], deterministic or stochastic functions of geographical identifiers designed to de-identify patient locations, while preserving the approximate spatial distribution of cases. These encompass previous approaches such as aggregation and translation by fixed distances, as well as affine transformations (consisting of scaling and rotation followed by translation) and random perturbations. Cassa *et al.* [1] evaluated a probabilistic randomization scheme using a bivariate Gaussian distribution with standard deviation inversely proportional to the population density to standardize the level of privacy protection throughout the map. Although these techniques represent a significant advance over aggregation, they apply the same transformation independent of the local geography, the number of patient records, and, in several cases, the underlying population counts. Consequently, the probability that any of the de-identified records originated from a single individual depends upon all of these variables. Although it may be possible to quantify this probability, the dependence on the variables may be complicated and the method of quantification would not be straightforward. However, a quantitative measure that captures privacy protection is essential for ethical and legal reasons.

We present a principled approach to de-identifying patient locations based on linear programming that specifies the maximum probability of associating any of the transformed locations with any individual in the population. This re-identification probability is a user-specified parameter of the method. The solution is optimal in that it guarantees that patients are moved the minimum distance for the level of privacy protection offered. The method has the advantage that it does not move patients to unrealistic locations, such as lakes and rivers. It may be used to create de-identified data sets in accordance with HIPAA for publishing maps of diseases, for cluster detection, or for other epidemiological investigations. Application of the method to de-identifying patients in New York county, New York shows that a high level of privacy can be achieved while moving patients very short distances.

4.2 Methods

Given the locations of a set of patients, the aim is to randomly assign new, de-identified locations that can be associated with the original patients with very low risk. The distance between the original and new locations should be minimized. The original locations may be any discrete geographical identifiers. For example, they may be zip codes, census block groups, street addresses, or even pixels in an image. The set of available locations must be known in advance; for example, these could be all the census block groups in a state, or all the residential addresses within a city.

This problem can be captured by a linear programming (LP) model, a simple type of mathematical model that consists of a set of decision variables, constraint equations, and an objective function. Given m available locations, the decision variables are the transition probabilities P_{ij} of assigning a patient in location i to a new location j (see figure 4-1). Once values have been assigned to the decision variables, each

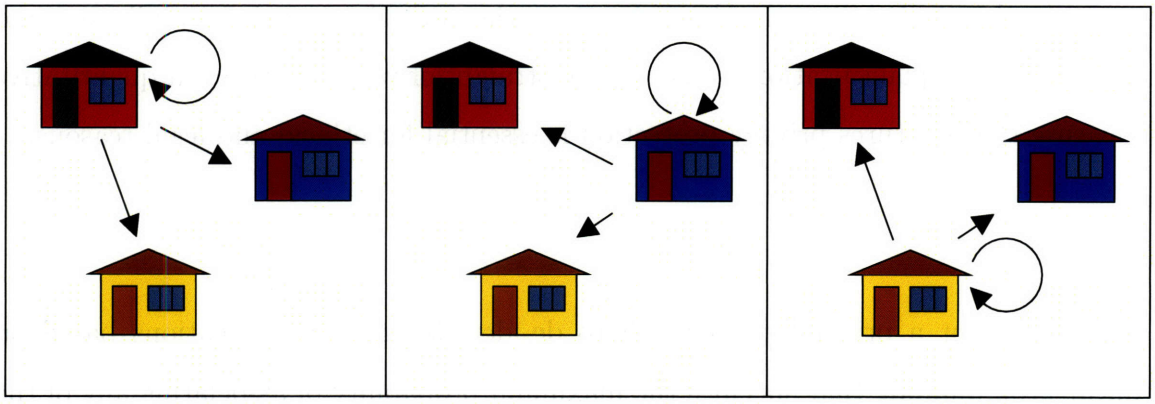


Figure 4-1: Schematic of transition probabilities. A patient found at each location may transition to any other location. In this simple example, there are three locations (represented by houses) and nine transition probabilities (represented by arrows). The probabilities are variables solved by linear programming.

of s patients in a list of original locations is moved to a new location independently of the other patients. If a patient is originally in location i, a new location is drawn from the set $j \in \{1, 2, ..., m\}$ using a multinomial distribution with probabilities P_{ij}. The goal is thus to assign a value to each decision variable P_{ij} so that this procedure ensures privacy and minimizes patient movement.

Constraint equations specify conditions that must be satisfied by the decision variables P_{ij}. Since the decision variables are probabilities, each must be nonnegative:

$$0 \leq P_{ij} \text{ for all } 1 \leq i \leq m \text{ and } 1 \leq j \leq m. \tag{4.1}$$

In addition, every case must be moved somewhere ("moves" to the original location are allowed), so

$$\sum_j P_{ij} = 1 \text{ for all } 1 \leq i \leq m. \tag{4.2}$$

A final constraint guarantees that the risk of linking any randomized location with any original patient is small. In formal terms, we specify that the probability that any location from the randomized data set originated from any specific individual in the underlying population is at most ξ:

$$P_{ij} - \frac{\xi}{s} \cdot \sum_{k=1}^{m} n_k \cdot P_{kj} \leq 0 \ \forall \ i, j. \tag{4.3}$$

In this equation, ξ is a user-specified privacy bound between zero and one, $s \geq 1$ is the number of patients in the data set to be de-identified, and n_i is the number of people in region i. For example, if the regions are census block groups, then the constants $\{n_i\}_{i=1}^m$ may be corresponding populations drawn from the same census. If the regions are exact addresses, then n_i is assumed to be 1 for each i.

To derive this constraint, consider the probability of re-identifying a set of s cases that have been randomized to new locations. Given m available locations, let P_{ij} denote the probability of transition from location i to location j for $1 \leq i, j \leq m$. Given the set of s locations comprising the de-identified data set, we require the probability that any one of these derived from one specific individual to be at most ξ. Equivalently, the probability that a location from the randomized data set originated from an arbitrary specific individual is required to be at most $\frac{\xi}{s}$. Let X and Y denote the original and transformed locations, respectively. This condition is formally expressed as:

$$p(\text{patient } q | Y = j) \leq \frac{\xi}{s} \tag{4.4}$$

for every individual q in the population and every location $j \in \{1, 2, \ldots, m\}$. The left hand side of this inequality is equivalent to

$$p(\text{patient } q \cap X = L(q)|Y = j), \tag{4.5}$$

where $L(q)$ is the location of individual q, or

$$p(\text{patient } q|X = L(q)) \cdot p(X = L(q)|Y = j). \tag{4.6}$$

Assuming that all individuals in location $L(q)$ have an equal chance of having the disease, we have

$$p(\text{patient } q|X = L(q)) = \frac{1}{n_{L(q)}}, \tag{4.7}$$

where $n_{L(q)}$ is the number of people in location $L(q)$. Hence the condition expressed by equation 4.4 is

$$p(X = L(q)|Y = j) \leq n_{L(q)} \cdot \frac{\xi}{s} \tag{4.8}$$

for every individual q and location j. Since the location of q, $L(q)$, may only take on the values $1, 2, \ldots, m$, this is equivalent to

$$p(X = i|Y = j) \leq n_i \cdot \frac{\xi}{s} \tag{4.9}$$

for every i and j in the set $1, 2, \ldots, m$. After multiplying both sides of equation 4.9 by $p(Y = j)$, the left hand side becomes $p(X = i \cap Y = j)$, or $p(Y = j|X = i) \cdot p(X = i)$. Furthermore, $p(Y = j|X = i)$ is simply the transition probability from location i to location j, so it is equivalent to the decision variable P_{ij}. Hence equation 4.9 is equivalent to

$$P_{ij} \cdot p(X = i) \leq n_i \cdot \frac{\xi}{s} \cdot \sum_{k=1}^{m} P_{kj} \cdot p(X = k) \tag{4.10}$$

for all i and j. Assuming that all individuals in the population have an equal proba-

bility of having the disease, we have

$$p(X = i) = \frac{n_i}{\sum_{r=1}^{m} n_r}. \tag{4.11}$$

Hence, we rearrange equation 4.10 to obtain

$$P_{ij} - \frac{\xi}{s} \cdot \sum_{k=1}^{n} n_k \cdot P_{kj} \leq 0 \tag{4.12}$$

for all j, and for all i having $n_i > 0$. However, if $n_i = 0$, then no patients can be found in location i, so any conditions on P_{ij} do not affect the strategy. Hence we require that the inequality holds for all i and j.

The objective function to be minimized is the expected distance that a patient is moved:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \frac{d_{ij} \cdot n_i}{\sum_r n_r} P_{ij}, \tag{4.13}$$

where d_{ij} is the distance between region i and region j. For example, this could be the distance between census block group centroids, or between exact addresses.

Several standard linear programming techniques to solve LP models, such as that specified by equations 4.1-4.13, have been developed. When applied to an LP model, they either locate an optimal solution that minimizes the objective function, or they prove that no solution exists. The latter happens if *no* probabilistic de-identification strategy has a risk of re-identification of at most ξ. For example, if there are m available individual addresses, then no strategy to de-identify $s \leq m$ patients can achieve a risk of re-identification below $\frac{s}{m}$. If no strategy exists, then a larger re-identification risk can be specified (if acceptable for privacy protection), or the set of available locations can be expanded.

Especially for address data and image pixels, there may be many available locations, and consequently a large number of decision variables and constraint equations. This affects the running time and storage requirements of linear programming methods. The problem size can be decreased by allowing only transitions from each region to its k nearest neighbors, for some fixed k. The solution to this modified problem

may be slightly sub-optimal in terms of the distance patients are moved, but the restriction does not affect the accuracy of the re-identification probability.

Simple variations of the linear program make it possible to capture other objective functions, constraint equations, or decision variable constraints. Instead of minimizing the expected distance, the expected squared distance may be used:

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{d_{ij}^2 \cdot n_i}{\sum_r n_r}P_{ij}. \tag{4.14}$$

The squared distance penalizes long distance moves more heavily than short moves, which may be less likely to affect subsequent clustering analyses of the de-identified data set. In fact, any objective function that is a linear combination of the decision variables P_{ij} may be used without complicating the analysis.

If a deterministic strategy which always gives the same answer is preferred to a randomized strategy, this may be found by converting the problem into a binary integer program. This specifies that only the values 0 or 1 may be assigned to the decision variables. An optimal solution of the binary integer problem has the property that for any fixed i, P_{ij} is equal to 1 for exactly one value of j, and is equal to 0 for all other values of j. The result is a mapping of the set of locations onto itself. For a fixed j, the set $I_j = \{i : P_{ij} = 1\}$, if nonempty, has the property that $\sum_{i\in I_j} n_i \geq \frac{s}{\xi}$. In other words, the patients are binned into a subset of the locations, the number and positions of the bins minimize the expected transition distance, and the total population assigned to each bin is at least $\frac{s}{\xi}$. In general, the optimal deterministic strategy moves patients farther than the optimal randomized strategy.

It is also simple to add additional linear constraints to the problem. For example, it is possible to guarantee that no case is assigned to its original location by specifying that $P_{ii} = 0$ for every i in the LP model. Although this would not increase the level of privacy, it may assuage fears that original locations may be released. In general, additional constraints increase the optimal value of the objective function.

4.3 Example

To illustrate the method, we determine an optimal strategy to randomize patients in New York County census block groups with a maximum re-identification probability of 0.00005. We find that this bound is achieved while moving patients an average distance of only 265 meters (m), and we show that the privacy bound is robust to inaccurate census counts. The strategy preserves privacy to a greater extent than both aggregation by zip code and aggregation by the first three digits of zip code.

4.3.1 New York county census blocks

We consider de-identifying case locations in New York County, NY grouped by census blocks. A census block is a small geographical unit typically containing approximately 1500 people [82]. According to the 2000 census, the 988 census blocks in New York County contain between 0 and 15112 people (see Figure 4-2). We devise the optimal strategy to de-identify one patient with a maximum re-identification probability of 0.00005. This is consistent with the spirit of the HIPAA legislation since the first three digits of a zip code may be released if shared by at least 20,000 people. This is also the optimal strategy to de-identify 10 patients with a maximum re-identification probability of $\frac{1}{2000}$, 100 patients with a maximum probability of $\frac{1}{200}$, or, more generally, $1 \leq s \leq 20000$ patients with a maximum probability of $\frac{s}{20000}$. Transitions from any census block were restricted to its nearest 100 neighbors. The LP model was solved using CPLEX LP software, resulting in a 988×988 matrix of transition probabilities. Each matrix row contained the transitions from a fixed census block to every other census block; by constraint, at most 100 of these were nonzero.

Under the optimal strategy, the expected distance between a patient's original and de-identified location is only 265 m. Three of the 988 matrix rows are illustrated in figure 4-3. These show three possible configurations: patients are re-assigned to the same census block group or one of a few neighboring census block groups; patients are re-assigned to a single nearby census block group; and patients are moved to one of several possible census block groups which do not include the original location.

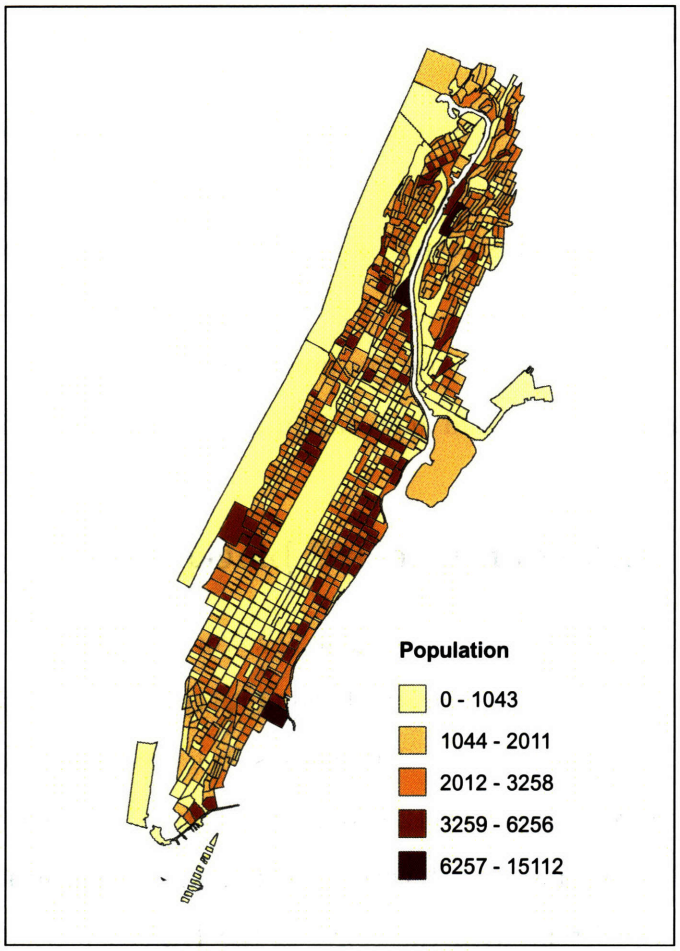


Figure 4-2: Total population of each census block group in New York County, NY, according to the 2000 census.

Even from this limited subset, it is clear that the optimal strategy would be difficult to devise by hand. In particular, the optimal transition probabilities are not a monotonic or regular function of the distance between census block groups, such as a Gaussian function.

In general, transitions are more likely to occur between nearby locations, and the likelihood of transition declines to zero as the distance between the regions increases (figure 4-4). The vast majority of the 98,800 transition probabilities are zero, indicating that transitions between most regions never occur; only 3155, or about 3.2%, of the transition probabilities are non-zero. For a fixed i, at most eight outgoing transition probabilities P_{ij} were non-zero. Thus it is unlikely that the restriction to 100 such transitions affected the optimality of the result.

Figure 4-3: Transition probabilities for the optimal strategy to de-identify $s \leq 20,000$ patients from New York County, New York with a maximum re-identification probability of $\frac{s}{20000}$. Transition probabilities from three of the 988 census blocks are shown, illustrating a few of the many possible transition distributions. The shading in region j represents the value of the probability P_{ij} of transitions into the region. a) Patients in one census block (purple asterisk) may remain there, or they may transition to one of several nearby blocks. b) All patients originally in one census block (purple asterisk) are assigned to one neighboring block. c) Patients are re-assigned from one block (purple asterisk) to one of four nearby census blocks. No patients are re-assigned to the original census block (i.e. $P_{ii} = 0$).

To examine the relationship between re-identification probability and the expected distance moved by a patient, we calculated the optimal de-identification strategies for a range of re-identification bounds. Because the total population summed over all census block groups is 1,696,038, the minimum achievable re-identification probability is $\frac{s}{1696038}$, or $s \cdot 0.00000059$. This corresponds to the complete randomization strategy of moving each person in a list of $s \leq 1,696,038$ people to census block i with probability $\frac{n_i}{\sum n_j}$. The corresponding expected transition distance is 6.4 km. As the re-identification probability is increased, the optimal strategy moves the patients less in expectation. The least populated non-empty census block group contains only one individual, so the strategy of re-assigning patients to their original locations has a re-identification probability of 1 (which would be realized if one patient in a "de-identified" set came from that census block group) and an expected transition distance of 0 km. The optimal strategies for de-identifying patients were calculated for a range of re-identification probabilities between these two extremes, and the expected distance moved by each patient is shown in figure 4-5. The results are

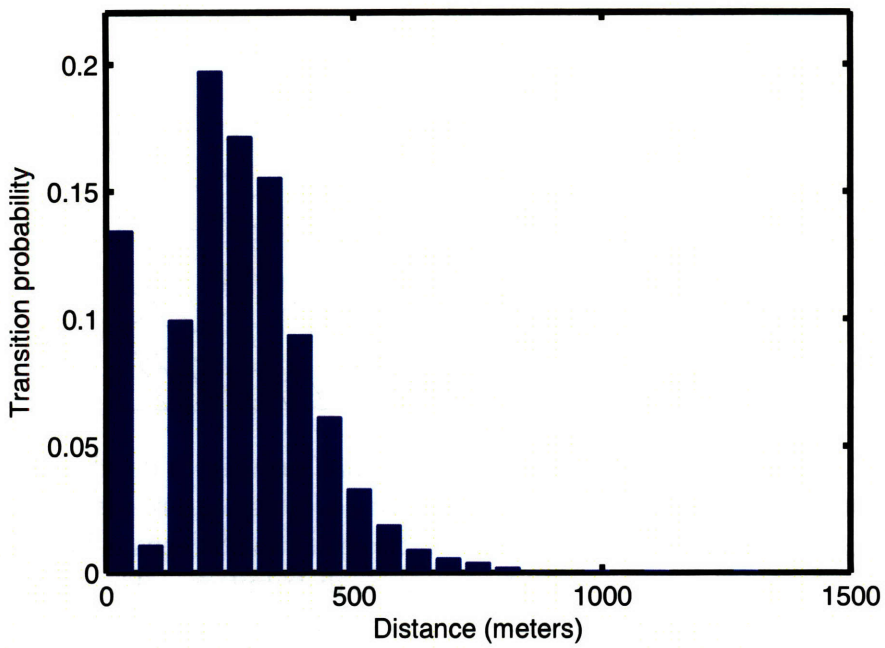


Figure 4-4: Histogram of the distance between original and de-identified locations for an individual randomly chosen from the population, under the optimal strategy to de-identify a set of $s \leq 20,000$ patients in New York County, New York to a probability of $\frac{s}{20000}$.

shown in figure 4-5.

These optimal LP strategies move patients less than other HIPAA-compliant strategies (figure 4-5). Creating a HIPAA limited data set by aggregating patients by zip code moves patients an expected 519 m. The least populated zip code contains 884 people (excluding empty zip codes and one zip code containing only one person), so there is a maximum re-identification probability of $\frac{s}{884}$ for a set of $s \leq 884$ patients under this strategy. Aggregating by the first three digits of zip code to create a HIPAA non-identifiable data set moves patients an expected 3.9 km, and has a maximum re-identification probability of $\frac{s}{8188}$. (In this case, the aggregated data set would not qualify as non-identifiable under HIPAA since some digits are shared by fewer than 20,000 people.) Figure 4-6 shows the spatial bins into which patients are aggregated by zip code and by first three zip code digits.

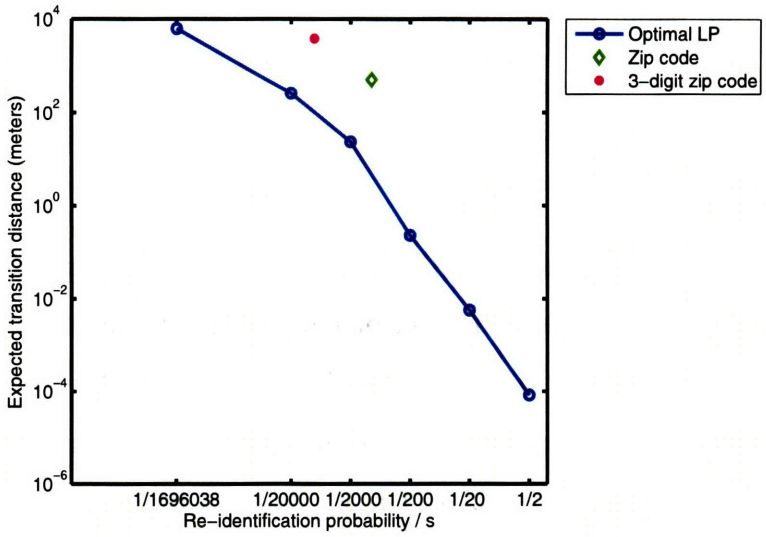


Figure 4-5: Relationship between the re-identification probability, the number s of patients, and the expected transition distance for the optimal LP strategy to de-identify patients by census block group in New York county, New York. As the level of privacy protection decreases, patients are moved a smaller distance in expectation. Aggregation by zip code (green diamond) and first three zip code digits (magenta circle) are suboptimal strategies.

4.3.2 Sensitivity analysis

Inaccuracies in the census estimates for each location used as input to the LP model may affect the re-identification probability. This happens when the number of people in a location is overestimated. It is elementary to show that overestimating *all* the census numbers by a factor of f leads to a re-identification probability of $f \cdot \xi$ instead of ξ in the worst case. For example, if every region in the New York analysis had only half the people reflected by the census, the re-identification probability in that analysis would be 0.0001 instead of 0.00005. However, in practice, overestimating the census numbers may have very little effect on the re-identification bound. To illustrate this, we randomly chose 5% of the census blocks to have actual populations 10% below the census estimates. For each i and j, we re-calculated the re-identification probability that a patient reported to be in location j of the de-identified set corresponded to any specific individual in region i using the strategy calculated above to de-identify individuals with probability $\frac{s}{20000}$. We found that only 1839 of the 98800, or fewer than 2%, of the re-identification probabilities violated the pre-specified bound of 0.00005.

The maximum probability was 0.000052.

4.4 Discussion

In the current climate of public concern for patient privacy and legislation imposing strict controls on the dissemination of patient-identifiable data, new strategies for de-identifying data sets while preserving information for disease surveillance and epidemiology are needed. It is imperative that strategies quantify the level of disclosure risk. The LP technique presented here for de-identifying spatial data has several benefits over existing methods. First, the user-specified level of privacy protection afforded by the method is mathematically well-defined. This re-identification probability is simply the maximum probability that any patient in the de-identified data set corresponds to any single individual in the population. This re-identification probability holds even if the exact randomized strategy is known to the data recipients. In other words, even knowledge of the complete set of transition probabilities $\{P_{ij}\}$ would not help re-identify patients beyond the pre-specified probability.

Second, the strategy moves patients a smaller distance than the common practice of aggregating by zip code, and it is far superior to the strategy suggested in the HIPAA legislation of aggregating by the first three digits of zip code. In fact, it moves patients a smaller distance, on average, than *every* other possible strategy, either deterministic or random, obeying the same re-identification bound that can be expressed as a matrix of transition probabilities.

Third, the technique is flexible, and can be extended based on the requirements of the user to minimize other objective functions or capture other constraint equations. It may also be used to calculate an optimal deterministic de-identification strategy, which always assigns patients to the same locations.

In addition, the LP strategy does not assign cases to unrealistic places, such as bodies of water or other uninhabited regions. While this does not help in scientific exploration of the distribution of the disease using the de-identified data set, it makes for more attractive maps of the de-identified locations.

The accuracy of the re-identification bound depends on a few assumptions. The underlying population size at each location must be known in advance, although the method appears to be robust to small inaccuracies. We also make the assumption that no other information is available to influence the *a priori* probability that any individual in the population has the disease. If any other information is available, it must be incorporated into the problem. Otherwise, the re-identification probability bound will not be correct, and privacy will not be guaranteed. For example, if the final version of the data set is to contain both the location and the race of each patient, then a de-identification strategy must be developed for each race represented. The population sizes n_i used in equations 4.3 and 4.13 must represent the number of people of that race. Similarly, if age, sex, or any other identifier is released, a new LP model reflecting the sub-population sizes must be solved for each value of the identifier. This is not always possible since stratified population data may not be available. If the population sizes are unknown, a lower bound on each population size will suffice to ensure privacy, but the solution may not be optimal in expected distance.

For individual addresses, we recommend using a population size of 1 for each address in the LP model. This limits the probability of associating any household with a case to the re-identification probability. However, the public may not feel comfortable with any addresses being released, even if the probability that an individual at that address has the disease is very small. An alternative is to use small aggregations such as city blocks or census block groups, as in the example presented in the previous section.

The measure of privacy protection proposed here, equal to the probability that any individual in the underlying population is among the de-identified patients, captures what is essentially important to a patient: "Will I be identified as having a disease as a result of the disclosure?" This measure is difficult to compute for previous strategies, since it may depend on variables such as the number of patient records and the study region itself. Several other measures of confidentiality have been proposed. These include Spruill's measure [81], equal to the proportion of records in the de-identified set that lie closer to their original location than to all other locations in the original

set. The exact value of the measure for the LP strategy depends not only on the privacy bound ξ, but also on the number and locations of original records and on the particular values for destination locations drawn from the multinomial distribution. For low values of ξ or a small number s of records, Spruill's measure is close to one. As ξ tends toward 1, Spruill's measure approaches $\frac{1}{s}$. Increasing s also decreases Spruill's measure. The interpretation of this is unclear because Spruill's measure does not always capture the intuition about privacy. For example, creating a de-identified set by shuffling the exact locations of all patients in the original set measures well by Spruill, but is clearly unacceptable for privacy protection. Conversely, assigning completely random locations to de-identify a data set of two patients measures poorly by Spruill, but would certainly preserve privacy.

Armstrong *et al.* also proposed four other measures of confidentiality. The first of these is a qualitative measure of vulnerability to geographical knowledge [81]. The LP strategy has no disclosure risk by this measure, since knowledge of all the possible locations does not decrease the re-identification probability. The second measures the ability to infer from the de-identified set regions within the map having a high disease risk. Like the de-identification strategies of aggregation and random perturbation, the LP method may reveal regions of high disease risk. However, this is a strength of the method, since the de-identified set may be used for disease mapping studies to depict variation in the spatial risk. The third measures the ability to re-identify all the patients, given the identity of some of the patients, and the final confidentiality measure is the minimum number of unlabeled locations from the original data set that can be used to compromise the entire de-identified set. There is no risk under the LP strategy by these measures; since patients are randomly moved independently of each other, relinking some of the patients cannot be used to compromise the identities of others.

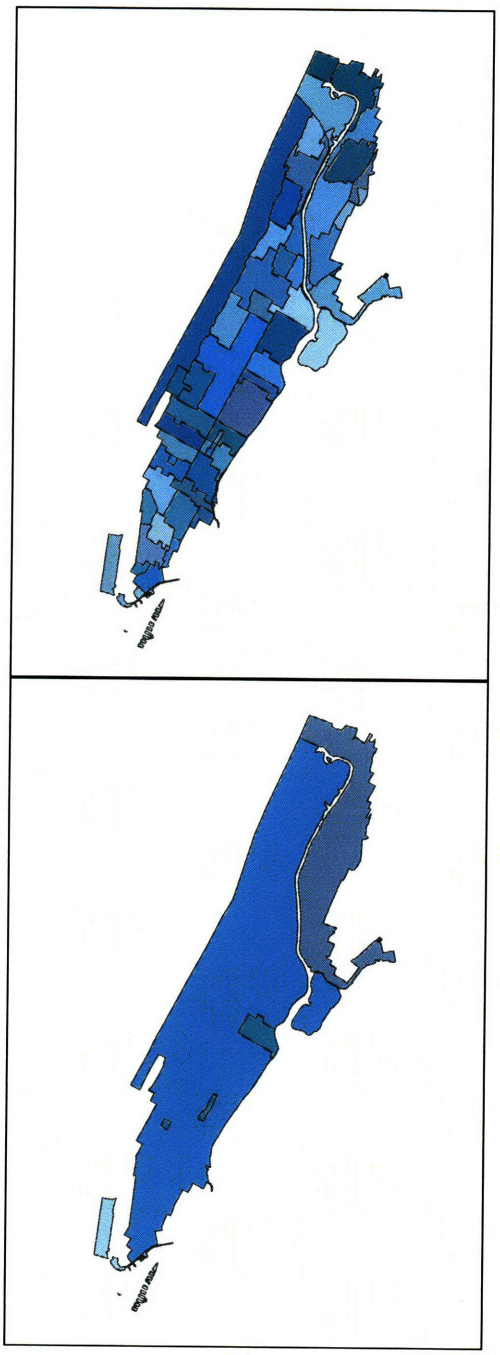


Figure 4-6: Aggregation of patients in New York County, New York by zip code and by first three zip code digits. Top) Census block groups have been aggregated by zip codes. Each census block group was assigned to the zip code containing its centroid. The expected distance moved by a randomly selected member of the population is 519 m, and the maximum probability that an individual is among a set of s de-identified patients is $\frac{s}{884}$. Bottom) Census block groups are aggregated by the first three zip code digits. The expected distance moved is 3.866 km, and the re-identification probability is $\frac{s}{8188}$.

Chapter 5

Automated real time constant-specificity surveillance for disease outbreaks

5.1 Introduction

The release of anthrax in 2001, the Severe Acute Respiratory Syndrome (SARS) outbreaks in China, Hong Kong and Toronto in 2002, and the emergence of new diseases such as West Nile virus have underscored the need for automated, real-time detection of outbreaks. Several such detection systems have been deployed in recent years at the hospital [83, 84], city [85, 86, 87], regional [88, 89, 90] and national [91, 92, 93] levels. Many systems use time series algorithms to detect aberrant conditions, such as CuSUM [94, 95, 96], variants of the Serfling method [85], multiresolution wavelet-based models [97], and trimmed seasonal models [98].

An outcome of any of these statistical methods – whether or not there is an alarm on any given day – is uninformative without an estimate of the likelihood that an

Originally published as: Wieland SC, Brownstein JS, Berger B, Mandl KD. Automated real time constant-specificity surveillance for disease outbreaks. BMC Medical Informatics and Decision Making. 2007;7(1):15.

alarm signals a true outbreak. This likelihood depends in part on the specificity of the detection method, equal to the proportion of non-outbreak days for which no alarm is raised. The specificity is related to the false alarm rate by the simple equation

$$\text{false alarm rate} = 1 - \text{specificity}.$$

Even small changes in the specificity of the detection method may have a large impact on the likelihood of a true outbreak. Despite the importance of knowing the specificity, analysis of the specificity of outbreak detection algorithms has been rudimentary, and it is common practice to report one average value of specificity that is assumed to reflect the true specificity on any day of the year or week. Implicit in this is the assumption that the specificity is constant as a function of time. If this assumption is incorrect – if instead the specificity of an outbreak detection system is a function of time that deviates significantly from its average value – then on any given day, a public health practitioner cannot know the specificity of the system or the related probability that there is a disease outbreak, and therefore cannot respond appropriately to alarms.

The sensitivity of a method, or proportion of outbreaks detected, is negatively associated with its specificity. Unlike the specificity, however, it cannot be evaluated from non-outbreak data. This is because in addition to its dependence on the specificity, it also depends on the characteristics of an outbreak, including its duration and magnitude. Hence the trade-off between sensitivity and specificity must be carefully considered in the context of the outbreak type of interest to ensure that both fall in a useful range.

We sought to characterize changes in the specificity of alarms produced by standard time series outbreak detection methods as a function of time. We further explored how these changes affect the sensitivity of detection methods to several outbreak types. We introduced a statistical technique that allows us to model properties of time series not captured by traditional models, developing an outbreak detection strategy with constant specificity that may be used by public health practitioners for biosurveillance.

5.2 Methods

5.2.1 Data

Data were collected retrospectively in the emergency department (ED) of an urban pediatric tertiary care teaching hospital. All patients with respiratory presenting complaints seen in the ED between August 1, 1992 and July 30, 2004 were included in the study. The data were divided into a six-year training period, and a test period consisting of the final six years. ED chief complaints were selected at triage from among a constrained list, and classified as respiratory or non-respiratory using a previously validated method [99]. The study was approved by the institutional review board.

During the study period, approximately 137 patients were seen each day in the ED. The number of daily visits for respiratory complaints varied from 2 to 78. The mean number of respiratory visits was 21.05, and the standard deviation was 9.03 (see 5-1). These data and other hospital visit data time series have previously been shown to depend significantly on the day of the week and the season of the year [98, 100, 101, 102].

5.2.2 Time series algorithms

We implemented five traditional time series models used for outbreak detection: a simple autoregressive model, a Serfling model, the trimmed seasonal model, a wavelet-based model, and a generalized linear model. In addition, we introduced a model of both the expectation and the variance based on generalized additive modeling techniques. The input to each algorithm was a time series of historical daily ED respiratory visit counts, and each returned a threshold number of visits for the day immediately following the historical period. An alarm occurred when the actual number of visits exceeded the threshold.

Autoregressive model. The autoregressive model predicted the number of ED respiratory visits using linear regression on the number of visits during the previous

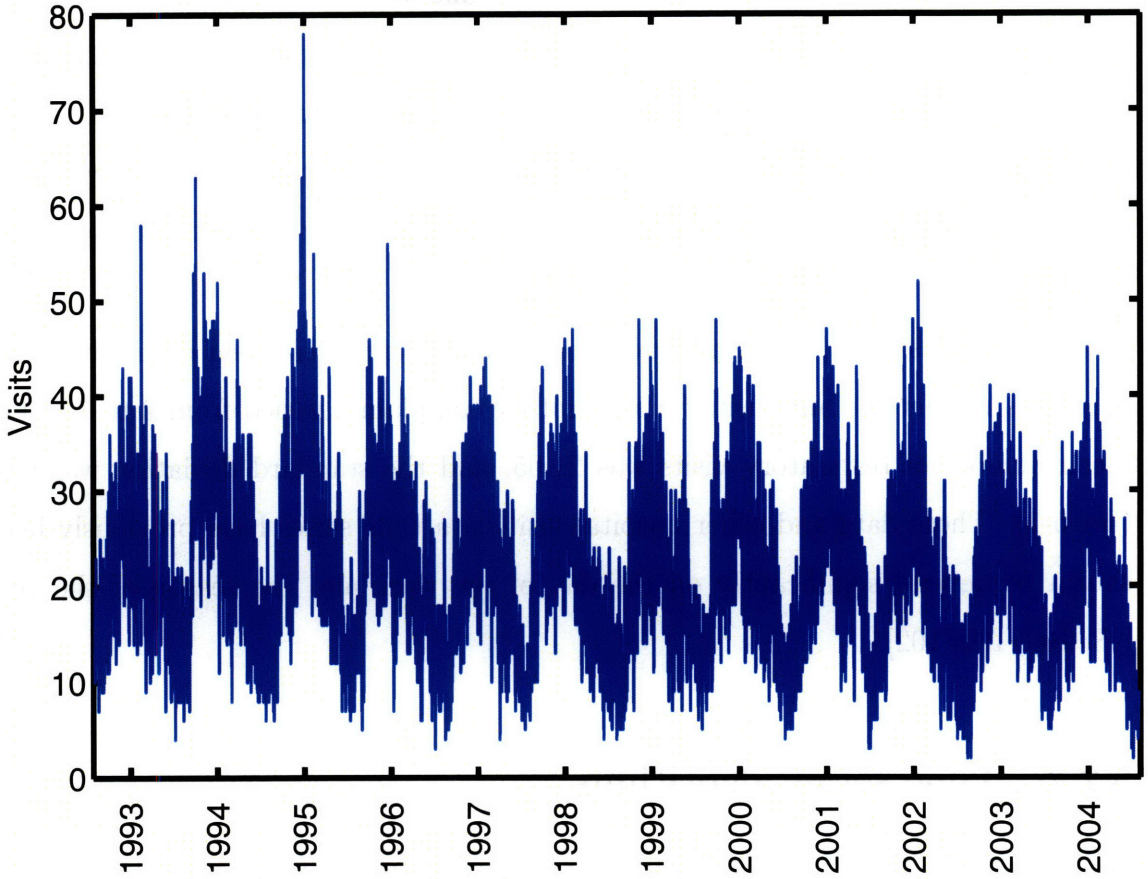


Figure 5-1: Emergency department visits for respiratory presenting complaints, August 1, 1992 - July 30, 2004. Daily time series showing the number of patients presenting with respiratory complaints to the emergency department during a 12 year period.

seven days:

$$E_t = a_0 + \sum_{k=1}^{7} a_k \cdot V_{t-k}, \tag{5.1}$$

where E_t is the predicted number of visits on day t, V_{t-k} is the actual number of visits on day $t-k$, and the coefficients a_k were fitted by least squares regression using training data.

Serfling method. The Serfling method and its variants have been extensively used for surveillance of influenza and other diseases [85, 103, 104]. Our implementation modeled the number of daily visits using linear regression on sine and cosine terms having yearly periodicities to capture seasonal effects, categorical variables for the day of week, and linear and quadratic terms. Under this model, the predicted number of visits on day t was

$$E_t = \sum_{k=0}^{6} a_k \cdot \delta_{k,\mathrm{dow}(t)} + a_7 + a_8 \cdot t + a_9 \cdot t^2 + a_{10} \cdot \sin\left(\frac{2\pi \cdot \mathrm{doy}(t)}{365}\right) + a_{11} \cdot \cos\left(\frac{2\pi \cdot \mathrm{doy}(t)}{365}\right), \tag{5.2}$$

where $\mathrm{dow}(t)$ is the day of the week from 0 to 6, $\mathrm{doy}(t)$ is the day of the year from 1 to 365, and the Kronecker delta function $\delta_{x,y}$ is equal to 1 when $x = y$ and 0 otherwise. To calculate the day of the year during leap years, each day after February 28 was treated as though it occurred on the previous day.

Trimmed seasonal model. The trimmed seasonal model is used in the AEGIS system [105] for statewide real-time population health monitoring, and was implemented as previously described [98]. Beginning with training set data, the average number of visits was calculated and subtracted from the data. From this, the average for each day of the week was calculated and again subtracted. To remove seasonal effects, the average for the day of the year was calculated after excluding the highest and lowest 25% of values for each day of the year, and again subtracted from the data. A first-order autoregressive, first-order moving average (ARMA) model was then fitted to the errors. The predicted number of visits E_t was calculated by summing the overall average, the average for the day of the week, the average for the day of the year, and the ARMA prediction for day t.

Wavelet model. The wavelet-based model was patterned after the wavelet anomaly detector developed by Zhang *et al.* [97]. The method used the number of daily visits in a training set, $V_1, V_2, \ldots, V_{t-1}$, to produce a prediction for day t. It consisted of the following steps:

1. A low-frequency wavelet component of the visit signal having periodicity of more than 32 days was calculated. This period was selected by Zhang *et al.* because it removes seasonal effects while preserving higher-frequency information, and because it is a power of 2, which is mathematically convenient for wavelet analysis. We used the Haar wavelet in our implementation of the model [106].

2. This low-frequency baseline was subtracted from the original signal, producing a residual for each day in the training set.

3. The predicted number of visits on day t was the value of the low-frequency component on the previous day.

Daily alarm thresholds for the autoregressive, Serfling, trimmed seasonal, and wavelet-based models were calculated as the sum of the expected number of visits and a multiple λ of the standard deviation of the model residuals on the historical training data. The value of λ was an adjustable parameter that affected the specificity of each model.

Generalized linear model. The generalized linear model consisted of a Poisson distribution function, an identity link function, and a linear predictor that included day of the week, month of the year, holiday and linear trend terms:

$$E_t = \beta_0 + \sum_{k=0}^{5} \beta_{k+1} \cdot \delta_{k,\mathrm{dow}(t)} + \sum_{k=0}^{11} \beta_{k+6} \cdot \delta_{k,\mathrm{moy}(t)} + \beta_{18} I_{\mathrm{holiday}}(t) + \beta_{19} t, \quad (5.3)$$

where $\mathrm{dow}(t)$ and $\delta_{x,y}$ are described in equation 5.2, $\mathrm{moy}(t)$ is the month from 1 (January) to 12 (December), and $I_{\mathrm{holiday}}(t)$ is an indicator function equal to 1 if day t is a holiday, and 0 otherwise. An alarm sounded if the value of the cumulative distribution function of a Poisson random variable with mean E_t exceeded the desired

specificity. This model was found by Jackson *et al.* [100] to have superior sensitivity to a variety of outbreak types compared to several control-chart and exponential weighted moving average models.

Expectation-variance model. In addition, we developed and implemented a novel method for outbreak detection that captures changes in the ED visit standard deviation, as well as in the expected number of visits. In contrast to previous surveillance models, which assumed that the variance is constant or proportional to the mean, it did not assume a functional form for the variance. Instead, the dependence of both the mean number of visits and the variance was modeled explicitly. In other applications, several statisticians have modeled the variance as a function of the same or additional covariates used to model the mean using iterative successive relaxation procedures (see, for example, [107] and [108]). We employed a simplified procedure involving two distinct models: an expectation model of the daily expected number E_t of respiratory ED visits, and a variance model of the daily variance σ_t^2 of respiratory ED visits. The number of daily visits is then modeled as a Gaussian with mean E_t and variance σ_t^2. Both components are generalized additive models (GAM's): nonparametric extensions of linear regression models having several variants depending on the choice of smoothing technique, the procedure used to find estimates of the nonparametric functions for multivariate models, and the number of degrees of freedom for each covariate [109, 110].

The GAM of the expectation accepted historical daily visit counts as input, and modeled them as a function of linear time to capture a long-term trend, the day of the year to account for seasonal trends, and the day of the week:

$$E_t = f_{\text{trend}}(t) + f_{\text{doy}}(\text{doy}(t)) + f_{\text{dow}}(\text{dow}(t)). \tag{5.4}$$

No smoothing was performed for the day-of-week term, since many replicates were available for each day of the week. A Gaussian kernel smoother was used for the trend term, and a Gaussian kernel smoother with circular boundaries was used for the day-of-year term since the day is a periodic covariate. Although a Gaussian was selected

for its ease of interpretation, in general the choice of kernel function has little effect on the model compared to the choice of bandwidth [109]. Optimal bandwidths of the two Gaussian smoothers were estimated by a two-step procedure. First, to optimize the bandwidth of the day-of-year Gaussian, the mean predictive squared error (PSE) on a training set consisting of the first six years of ED visit data was calculated for a range of bandwidths using 10-fold cross-validation for a model containing only the day-of-week and day-of-year covariates. The bandwidth minimizing the mean PSE was chosen, corresponding to a Gaussian distribution with a standard deviation of five days. Next, the bandwidth of the kernel used for the trend term was chosen by using 10-fold cross-validation to estimate the mean PSE on the training set of a model containing all three covariates for a range of trend bandwidths, using the previously determined optimal bandwidth of the day-of-year kernel. The minimizing bandwidth was again chosen, corresponding to a standard deviation of eight days. Because the model contained multiple nonparametric functions, an iterative backfitting procedure was used to estimate each until the model converged [109].

The residuals of the expectation GAM on the historical data were squared and used as the input to the variance GAM. This GAM was also a function of linear time, day-of-year, and day-of-week variables:

$$\sigma_t^2 = g_{\text{trend}}(t) + g_{\text{doy}}(\text{doy}(t)) + g_{\text{dow}}(\text{dow}(t)). \tag{5.5}$$

The Gaussian smoothers were chosen to minimize the PSE on the training data set using the same procedure as above. The optimal smoothers corresponded to Gaussian distributions with standard deviations of 6 and 253 days for the day-of-year and trend terms, respectively.

To set the alarm threshold for a given day, a composite expectation-variance model consisting of the two GAM's was trained on the previous six years of data. The alarm threshold for the next day was calculated as the sum of the expected number of ED visits, as predicted by the expectation GAM, and a multiple λ of the

expected standard deviation of ED visits, as predicted by the variance GAM:

$$\begin{aligned} \mathrm{A}_t &= E_t + \lambda \cdot \sigma_t && (5.6) \\ &= f_{\mathrm{trend}}(t) + f_{\mathrm{doy}}(\mathrm{doy}(t)) + f_{\mathrm{dow}}(\mathrm{dow}(t)) && (5.7) \\ &\quad + \lambda \cdot \sqrt{g_{\mathrm{trend}}(t) + g_{\mathrm{doy}}(\mathrm{doy}(t)) + g_{\mathrm{dow}}(\mathrm{dow}(t))}. && (5.8) \end{aligned}$$

The value of λ was an adjustable model parameter.

All models were implemented using the Matlab software package, Version 7.0.1 [111]. The Matlab system identification, statistics and wavelet toolboxes were used for the wavelet, generalized linear, and expectation-variance models.

5.2.3 Model predictions based on historical data

We used the expectation-variance model to generate alarm thresholds for each day during the test period from August 1, 1998 to July 30, 2004, which comprised the last six years of historical data. All of the available data could not be used for testing because a training period was required. To predict each threshold, the model was trained on the previous six years of data, ending the day before the day to be predicted, and was blind to the actual number of ED visits on the prediction day. The backfitting procedures to estimate the model successfully converged for each day of the study period. The model predictions for both the expected number of patients and the variance were always positive numbers throughout the study period. The average absolute predictive error was approximately four patients during the study period.

For each day, an alarm threshold was produced for each desired outbreak detection specificity between 0.01 and 0.99 in 0.01 increments. This was achieved by varying the threshold parameter λ appropriately. For example, to generate an alarm threshold with specificity s on day T, the model was trained on the historical visit data, $V_{T-2191}, \ldots, V_{T-1}$. This generated model estimates for the expected number of visits for each day, $E_{T-2191}, \ldots, E_{T-1}, E_T$, as well as estimates for the expected standard deviation of visits, $\sigma_{T-2191}, \ldots, \sigma_{T-1}, \sigma_T$. The parameter λ was chosen so that the

fraction of historical days for which the Z-score was at most λ was as close as possible to the desired specificity s. That is, λ was chosen to have the property that

$$\#\{t : T - 2191 \leq t \leq T - 1 \text{ and } V_t - E_t \leq \lambda \cdot \sigma_t\} \approx 2191 \cdot s. \tag{5.9}$$

The predicted threshold for day T was $E_T + \lambda \cdot \sigma_T$.

Alarm thresholds for each day of the test period and each desired specificity were similarly calculated for the autoregressive, Serfling, trimmed seasonal, and wavelet models. The alarm threshold for the generalized linear model was the largest integer A_t for which the cumulative distribution function of a Poisson random variable with mean E_t was at most s. With the exception of wavelet model thresholds, all alarm thresholds were calculated using the six years of visit data immediately preceding the prediction day. The wavelet model requires a training period having length equal to a power of two, so 2048 days of training data were used.

5.2.4 Detecting variability in the specificity

To determine whether a given model at a particular mean specificity had constant specificity as a function of the day of the week, we tabulated the proportion of alarm and non-alarm days at that mean specificity by day of the week. A chi-square analysis was performed under the null hypothesis that all days of the week had an equal fraction of alarm days. A p-value less than 0.05 indicated that the specificity was dependent on the day of the week. To determine whether the specificity was constant as a function of month and year, we performed similar chi-square analyses after tallying alarm days by month of the year and by calendar year of the study, respectively.

5.2.5 Simulated outbreaks

In order to ascertain the sensitivity of the models to outbreaks, we superimposed three synthetic outbreaks on the test data set: a flat outbreak of five additional patients per day for seven days, a linear outbreak which increased from one to five patients over five days, and a spike outbreak of 10 additional patients in one day. For each model,

each outbreak type, and each day of the test period, we created a new semisynthetic data set by adding an outbreak beginning on that day to the original data set. We then made an alarm threshold prediction for each of the outbreak days, and for each desired specificity between 0.01 and 0.99, based on training using the semisynthetic data set.

5.2.6 Estimating sensitivity, specificity, and timeliness of detection

The actual mean specificity for one model at each desired input specificity was determined by running the model on the historical data set. Specificity was estimated by calculating the fraction of days without alarms for each day of the week, month of the year, or calendar year. Sensitivity calculations used the results of applying each of the models to the semisynthetic data sets. The sensitivity was calculated as the fraction of outbreaks for which there was at least one alarm day. Exact 95 percent binomial confidence intervals were calculated for each estimate of sensitivity and specificity. Timeliness of detection was evaluated for each method by calculating the mean lag in days between the start of a flat outbreak and the first alarm sounded. Missed outbreaks, for which no alarms were sounded on any day of the outbreak, were excluded from timeliness calculations. An alarm sounding on the first outbreak day corresponded to a lag of zero. Timeliness calculations were calculated at the benchmark specificity values of 0.85 and 0.97.

5.2.7 Comparing outbreak detection among models

To compare the outbreak detection performance of the expectation-variance model with the traditional models, receiver-operator (ROC) curves were constructed for all models. ROC curves show the dependence of the mean sensitivity on the mean specificity, and the area under the ROC curve is an indicator of overall performance. The area was estimated by the trapezoidal method.

5.3 Results

5.3.1 Evaluation of specificity trends over time

As suspected, the specificity of the five standard models was not constant over time. Hypothesis testing indicated that the specificity of the Serfling, trimmed seasonal and generalized linear models varied with the study calendar year and study month ($p < 0.05$) over a range of mean specificities between 0.50 and 0.99. The autoregressive model demonstrated a variable specificity with the study month and day of the week ($p < 0.05$) for the same range of mean specificities, and the wavelet model had variable specificity ($p < 0.05$) on all three time scales (5-2). Several trends in the specificity were apparent when the analysis was limited to particular values of mean specificity. For example, at a mean specificity of 85 percent, corresponding to approximately one false alarm each week, the autoregressive, Serfling, trimmed seasonal and wavelet models had highest specificity in June and July and low specificity during the winter months. The specificity of the autoregressive and wavelet models was highest in the middle of the week and lowest on Sunday, and the Serfling, trimmed seasonal and generalized linear models had higher specificity during certain study years (5-3). Similar trends were observed at other mean specificity values, including 0.90, 0.95, and 0.97 (data not shown).

By contrast, the expectation-variance model specificity was constant as a function of the study year, study month, and the day of the week. Hypothesis testing resulted in a p-value above 0.05 for the entire range of input specificities on all three time scales, indicating that there was no evidence to suggest that the specificity was non-constant on any time scale (5-2).

5.3.2 Comparison of sensitivity and timeliness of new and traditional methods

The expectation-variance model usually outperformed traditional approaches in terms of sensitivity. The area under the expectation-variance model ROC curve was equal

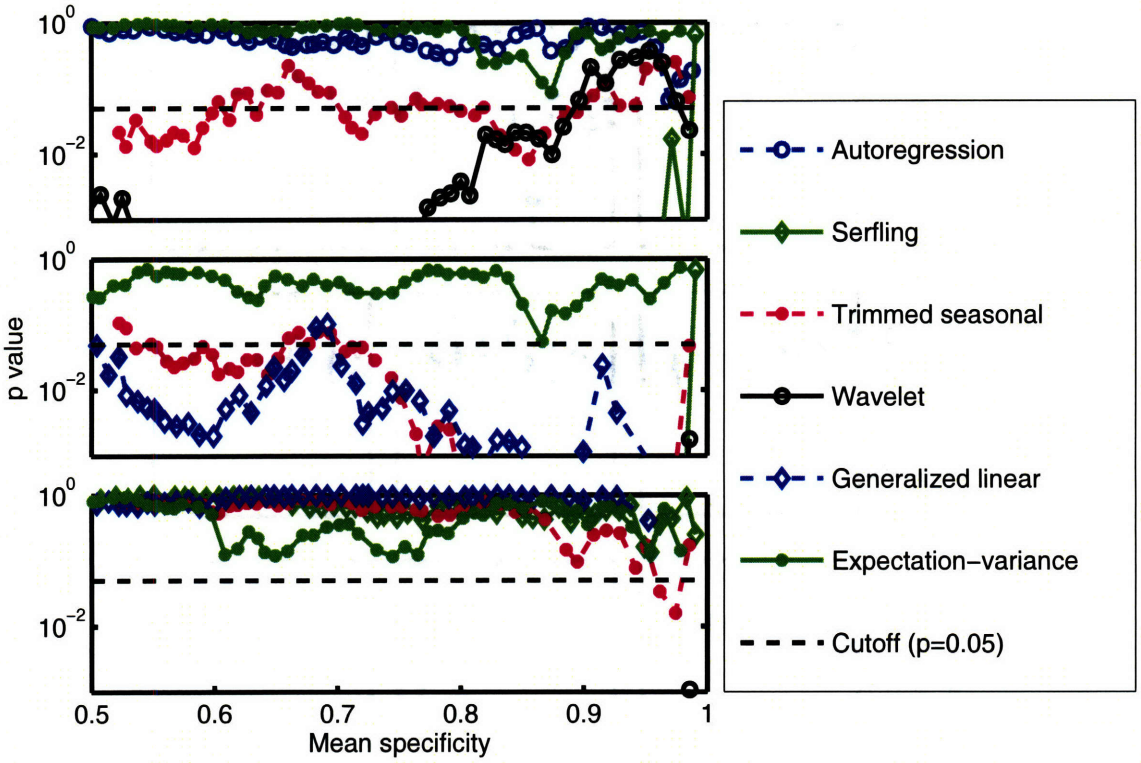


Figure 5-2: Evaluating variability in specificity on three time scales. Plots of p-values for the chi-square test over various time scales for the five comparison models over a range of mean specificity values from 0.50 to 0.99, as well as p-values for the expectation-variance model. Top: calendar year of study. Middle: month of year. Bottom: day of week. The expectation-variance model has a p-value over 0.05 for the entire range of mean specificity values for all three time scales, so the null hypothesis of constant specificity is not rejected. All plots not shown are highly significant ($p < 0.001$) for non-constancy.

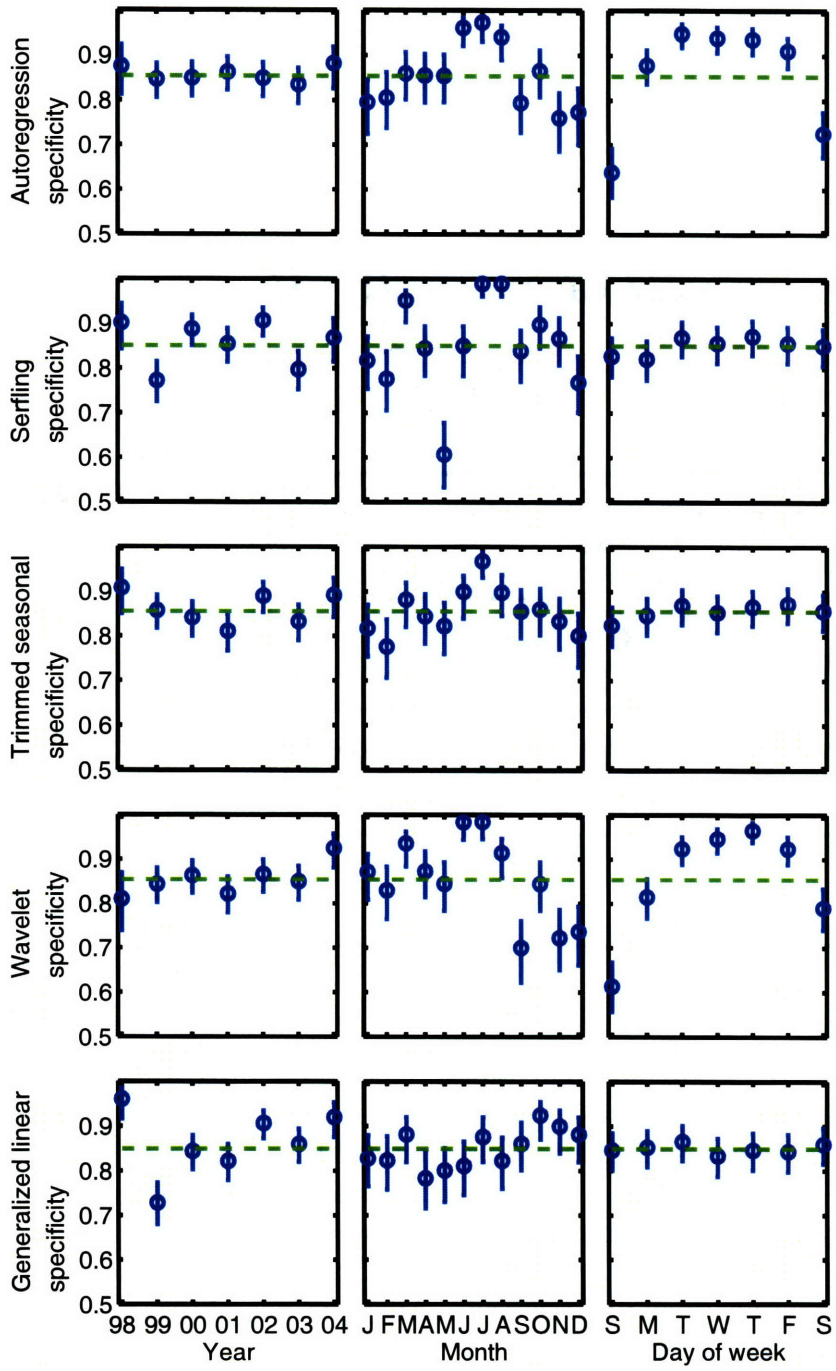


Figure 5-3: Average specificity trends over time. Average specificity for each calendar year, month, and day of week for the five comparison methods during the study period. Data shown were recorded for each model implemented at 85% mean specificity. Similar trends were observed for all methods at 97% mean specificity (data not shown).

Table 5.1: ROC curve areas for traditional and expectation-variance detection models applied to three different types of outbreaks superimposed on respiratory visits to an urban pediatric ED, August 1998 - July 2004.

Detection method	Flat outbreak	Linear outbreak	Spike outbreak
Autoregression	0.94	0.90	0.88
Serfling	0.93	0.88	0.89
Trimmed seasonal	0.95	0.91	0.89
Wavelet	0.93	0.87	0.86
Generalized linear	0.95	0.91	0.91
Expectation-variance	0.95	0.91	0.91

Table 5.2: Mean lag in detecting outbreaks of five additional patients per day superimposed on the pediatric ED respiratory visits, August 1998 - July 2004. Detection lag calculations exclude undetected outbreaks. Hence the sensitivity of the method must be considered when interpreting the detection lag.

Detection method	Mean specificity	Mean sensitivity	Mean detection lag (days)
Autoregression	0.97	0.40	2.26
Serfling	0.97	0.36	2.37
Trimmed seasonal	0.97	0.42	2.26
Wavelet	0.98	0.38	2.43
Generalized linear	0.95	0.68	1.93
Expectation-variance	0.97	0.58	1.96

to or greater than that of the five comparison models for all three outbreak types (table 5.1).

The expectation-variance method also performed well in terms of earliness of detection. At a benchmark mean specificity of approximately 97 percent, it detected a seven-day outbreak consisting of five additional patients each day with a shorter lag than the autoregressive, Serfling, trimmed seasonal, and wavelet models (table 5.2). The expectation-variance model also had earlier detection than these models at 85 percent specificity (data not shown).

5.3.3 Temporal sensitivity trends

The sensitivity of outbreak detection depends on the size and shape of an outbreak, as well as on the amount of noise in the ED utilization signal. Thus even when the specificity is held constant, it is natural for the sensitivity to vary with the season,

day of the week, and trend. The ED visit signal had the least noise in the summer and the most noise in the winter (5-4). Hence the signal-to-noise ratio was highest in the summer for any fixed type of outbreak, and the sensitivity of any reasonable detection strategy should theoretically be greater during the summer than in the winter. Summer and winter ROC curves for the expectation-variance and five comparison methods confirmed that summer sensitivity was greater than winter sensitivity when the specificity was held fixed (5-4 insets). However, at mean specificity values of 85 and 97 percent, plots of sensitivity over time for the autoregressive, Serfling, trimmed seasonal and wavelet models showed a paradoxical increase in sensitivity to synthetic outbreaks during winter months compared to summer months (5-4). These seemingly contradictory results occurred because the mean specificity of these four comparison models was not the actual specificity during either the summer or winter. The specificity was significantly higher during the summer, corresponding to a shift to the left along the summer ROC curve and a concomitant decline in summer sensitivity. The opposite occurred in winter. This anomaly was corrected by the expectation-variance model (5-4), since it operated at the same specificity during all seasons. The generalized linear model exhibited variable specificity by month, but its specificity was not highest during the summer months (5-3), and hence it also had greater summer sensitivity than winter sensitivity.

5.4 Discussion

We found that the specificity of outbreak detection was not constant for five traditional algorithms. This is important because having a standardized interpretation of the statistical characteristics of an outbreak detection test, including the specificity, aids public health practitioners in making rational decisions regarding resource allocation in the event of an alarm. The positive predictive value (PPV) of an alarm, the probability that an alarm signals a real outbreak, bears directly on the priority and

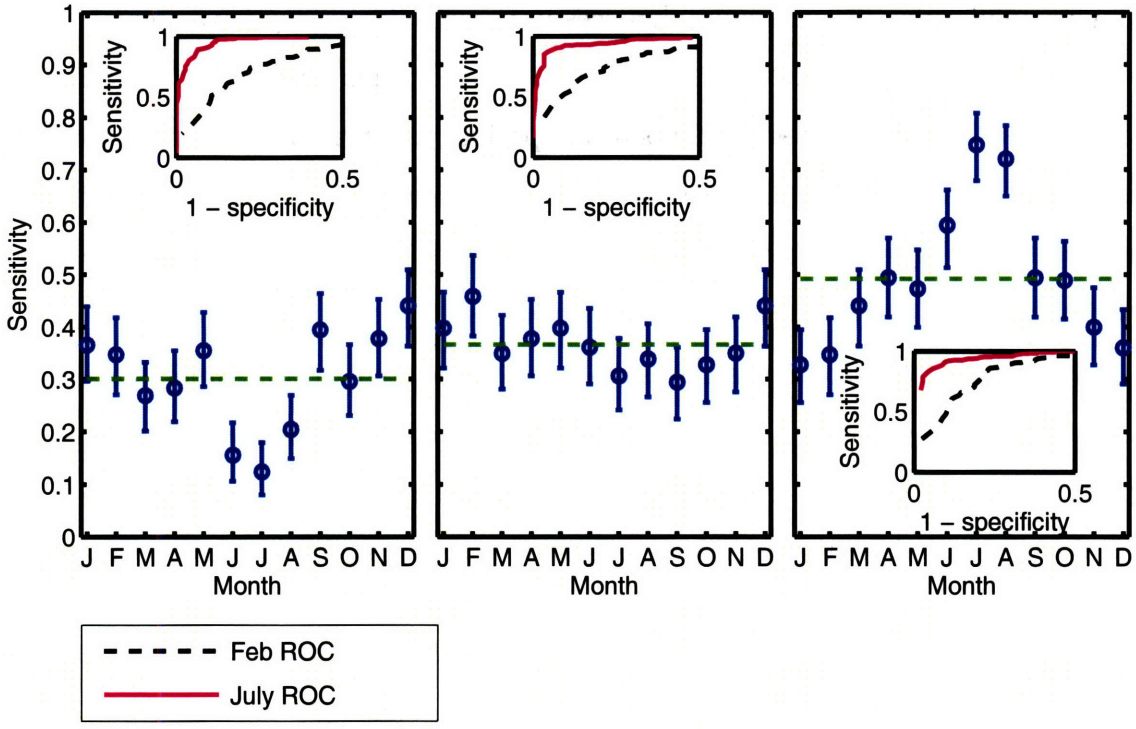


Figure 5-4: Seasonal sensitivity trends. Average sensitivity for each month of the study period for the autoregressive (left), trimmed seasonal (center), and expectation-variance (right) models when applied to data containing a superimposed spike outbreak of 10 additional patients during one day. Data shown were collected at a mean specificity of 97%. The sensitivity of the trimmed seasonal and autoregression models is higher during the winter than during the summer. Sensitivity is higher during the summer than during the winter for the expectation-variance model. July receiver-operator (ROC) curves lie below February ROC curves for all three models (insets). Similar trends were observed for flat and linear outbreaks.

extent of response required. The PPV is related to the specificity by the equation

$$\text{PPV} = \frac{\text{sensitivity} \cdot p}{\text{sensitivity} \cdot p + (1 - \text{specificity}) \cdot (1 - p)}, \tag{5.10}$$

where p is the prior probability of an outbreak. Because the specificity of an alarm strategy affects its PPV, it is crucial to have an accurate estimate of the specificity on any particular day. Even small differences in the specificity may have a great impact on the PPV; an alarm strategy at 95 percent specificity may have a PPV nearly twice as high as the same strategy at 90 percent specificity, depending on the nature of the outbreak considered and the sensitivity of the system. A public health practitioner responding to an alarm in the first case may wish to devote twice as many resources to investigating the alarm than in the second case.

The specificity also affects the overall cost associated with a surveillance model. Let c_{TP}, c_{FP}, c_{TN} and c_{FN} denote the costs associated with true positive alarms, false positive alarms, true negatives, and false negatives, respectively. Then the expected total cost of an alarm strategy on a given day is a weighted sum of these costs:

$$E[\text{cost}] = c_{TP} \cdot \text{sens} \cdot p + c_{FN} \cdot (1 - \text{sens}) \cdot p + c_{FP} \cdot (1 - \text{spec}) \cdot (1 - p) + c_{TN} \cdot \text{spec} \cdot (1 - p). \tag{5.11}$$

Lowering the specificity contributes to the cost due to fruitlessly investigating more false positive alarms, reflected in the third summand of the equation. At a specificity of, for example, 99%, one can expect to experience a false alarm every 100 outbreak-free days. Lowering the specificity to 97% increases the false alarms to approximately once per month. The cost equation can also be used to compare two alarm methods, A and B. Strategy A is more cost-effective than strategy B if and only if the expected cost of A is less than that of B:

$$(\text{sens}_A - \text{sens}_B)(c_{TP} \cdot p - c_{FN} \cdot p) < (\text{spec}_A - \text{spec}_B)\left(c_{FP} \cdot (1 - p) - c_{TN} \cdot (1 - p)\right). \tag{5.12}$$

Thus the greater the accuracy in the estimates of the specificity and sensitivity of each method, the prior probability of an outbreak p, and the costs of each scenario,

the more accurately a public health department can compare the cost-effectiveness of the various available surveillance methods.

It may be desirable under certain conditions to have non-constant specificity. For example, one may wish to adjust the specificity so that the PPV is constant as a function of the day of the week, season, and trend. Alternatively, a high profile event may merit special attention, requiring lower specificity surveillance to increase the sensitivity to outbreaks. The expectation-variance model is preferable to traditional models in these situations because its specificity is known more reliably than that of traditional models. Therefore the specificity can easily be adjusted with time according to public health needs. By contrast, current models operate with unknown specificity, and adjusting an unknown quantity presents a difficulty.

To understand the inability of traditional models to maintain constant specificity over time, it is useful to recast the outbreak detection problem in terms of percentiles instead of means. A perfect outbreak detection model operating at a specificity of 0.95 would output an alarm threshold equal to the 95th percentile for each day, above which an alarm would sound. More generally, a perfect model at specificity $\frac{k}{100}$ would model the kth percentile. The autoregressive, Serfling, trimmed seasonal and wavelet models assume that the data have normally distributed errors with constant variance. They thus make a first approximation to this percentile by modeling the mean, to which a constant (which depends on k) is added. One problem with this approach is that the ED utilization signal is heteroscedastic – that is, its variance is not constant as a function of time (5-5). In practical terms, this means that the kth percentile is sometimes farther from the signal mean than at other times. Hence it cannot be captured by adding a constant value to the mean. The result is that during periods of greatest ED utilization variance, such as the winter months (5-5), the alarm thresholds of these traditional models underestimate the kth percentile, leading to a decreased winter specificity (5-3). Conversely, all four models overestimate the alarm threshold during the summer months, when the ED utilization variance is lowest. In fact, neglecting the dependence of the ED visit variance on the day of week, day of year, or long-term trend when determining the alarm threshold introduces some

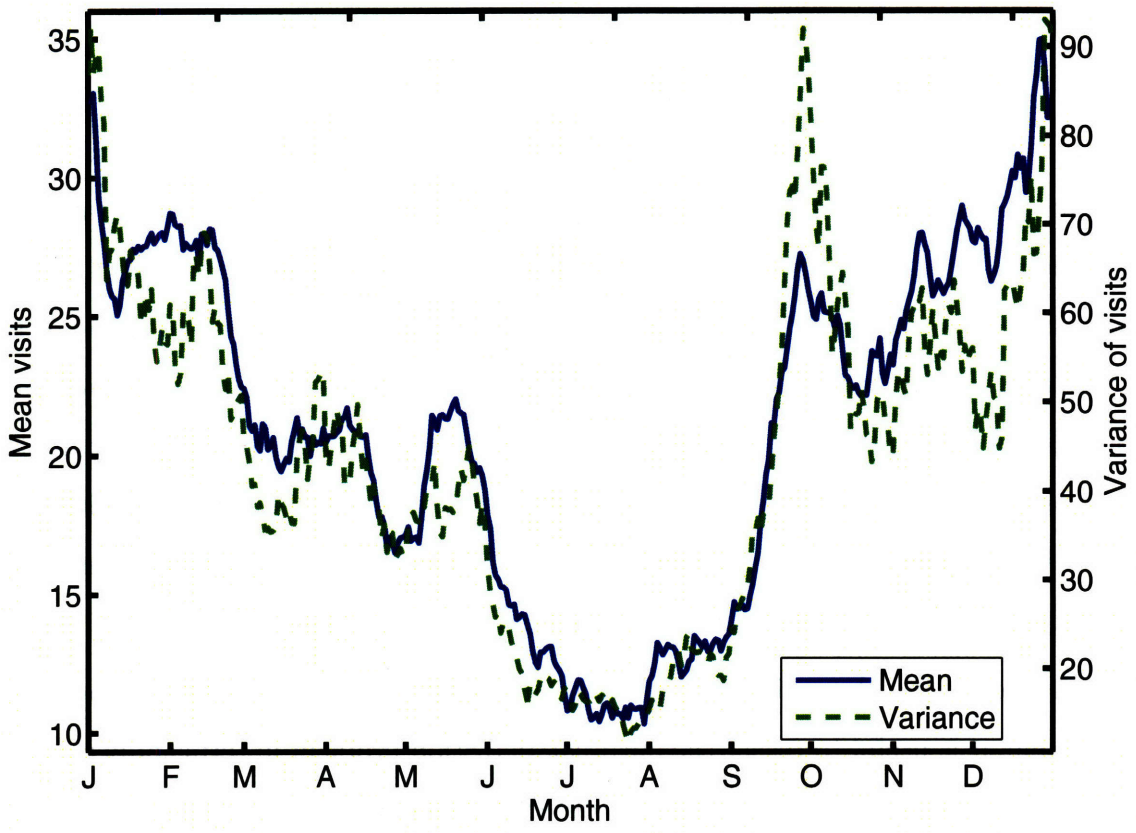


Figure 5-5: Seasonal trends in the mean and variance of ED visits. Mean number of ED visits (left axis, solid blue line) and mean variance in ED visits (right axis, dashed green line) as a function of the day of year. Data were smoothed using 5-day and 11-day moving averages, respectively. The ED utilization mean and variance are highest in the winter and lowest during the summer.

degree of systematic error in the alarm threshold, although it may not be of sufficient magnitude to cause statistically detectable variations in the specificity.

Although the generalized linear model does not assume that the variance is constant, it does assume that the data are Poisson distributed, and consequently that the signal variance is equal to the signal mean. However, the actual signal variance is greater than the mean; the ratio ranges from approximately one to more than three during the calendar year (5-5). The result is that during periods of high relative signal variance, the specificity of the method is also relatively high. For example, in October, both the ratio of signal variance to signal mean (5-5) and the specificity (5-3) are high.

Changes in specificity may also result from systematic errors in the expected

number of ED visits predicted by the algorithms. For example, our implementations of the wavelet and autoregression models do not take into account day-of-week effects on the number of ED visits. Hence during high-volume days, such as Sundays, these models underestimate the expected number of visits. This in turn lowers the alarm cutoff value and the specificity compared to low-volume days such as Wednesdays. The Serfling model constrains the seasonal effects of ED utilization to a sine wave. However, the normal seasonal pattern of respiratory visits includes a spring increase that coincides with the allergy season (5-5), which cannot be captured by a sine curve. This causes a May dip in the specificity of the Serfling model (5-3).

In addition to the approach considered here, it may be possible to apply a generalized additive or other model to the squared residuals of a traditional algorithm. A model for the alarm threshold would then be constructed in a similar manner to the expectation-variance model. Because the specificity is affected by systematic errors in both the mean and the variance, it would be necessary to apply a statistical test to ensure that the specificity was constant.

The expectation-variance model is a general time series method which could be applied to surveillance of other syndromes and populations. Implemented here in Matlab, it could easily be imported to other platforms, and it requires minimal additional computational resources for public health departments collecting surveillance visit data. It does, however, have several limitations. While useful for modeling syndromes that are predictable functions of the trend, season, and day-of-week covariates, such as respiratory or gastrointestinal illnesses, it would have limited utility compared to simpler models for rare or sporadically occurring syndromes. The present study has evaluated the specificity, sensitivity, and timeliness of detection using a training set containing six years of data. However, this much historical data is not always available for model training. Although the algorithm is easily adapted to shorter training sets, future work is needed to assess its performance with such sets. Like other detection methods, the training data must be free of an outbreak of interest in order for the specificity estimates to be accurate. Thus the training set used in the present study would be useful for detecting anthrax, other bioterrorism events,

or large influenza outbreaks due to changing viral strains, but not for reliably detecting yearly average influenza outbreaks present in the data. Like other time series methods, the model also does not take advantage of geospatial information or data streams containing different types of data.

A more subtle limitation of the expectation-variance model is that its output is a binary variable – the absence or presence of an alarm. Kleinman *et al.* [112] proposed an approach to temporal and spatial surveillance which instead provides the probability that an observed event would be expected in the absence of an outbreak. This approach represents a shift from statistical testing to more detailed statistical modeling techniques [113]. Although the current implementation of our method is binary, it can easily be converted to a "modeling" approach. For example, a graph of the specificity as a function of the alarm threshold corresponds to a predicted cumulative distribution function of the number of visits on any given day.

In addition to the limitations of the model, our study is limited in its analysis of sensitivity to various outbreak types. The sensitivity depends on the time series of additional outbreak patient visits, of which an infinite array of possibilities exist. In the absence of outbreak data capturing the essential features of the many diseases and syndromes that may be monitored, we have used synthetic outbreaks having simple functional forms or "canonical shapes" [114]. This makes comparisons between types of outbreaks easy to interpret. Alternatively, the response to one or more known outbreaks may be evaluated [100, 115]. This approach has the advantage that the outbreaks are inherently realistic, since they are instances of true outbreaks. However, they may be highly irregular and dominated by stochastic effects. Indeed, there is no guarantee that they bear resemblance to future outbreaks of the same or other diseases. The present study offers the promising conclusion that the expectation-variance model has good comparative sensitivity for a limited number of artificial outbreaks, but more detailed study in the context of outbreaks of interest would be necessary to conclude that the model is preferable to previous models for real-world surveillance.

5.5 Conclusions

The interpretation of alarms using current outbreak detection strategies is difficult because the specificity is extremely variable. The fluctuations in specificity are due to changes on the same time scales in the variance of the ED utilization signal. Unlike previous models, the model developed here accounts for changes with time of not only the expected number of ED visits, but also of the variance of the number of visits. It is our hope that this provides a useful method for achieving a signaling strategy with known, constant specificity, enhancing the ability of public health practitioners to interpret the meaning of an alarm.

Bibliography

[1] C.A. Cassa, S.J. Grannis, M. Overhage, and K.D. Mandl. A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *J. Am. Med. Inform. Assoc.*, 13:160–165, 2006.

[2] J. Kelly. *The Great Mortality: An Intimate History of the Black Death, the Most Devastating Plague of All Time.* Harper Collins, 2005.

[3] JS Oxford. Influenza A pandemics of the 20th century with special reference to 1918: Virology, pathology and epidemiology. *Reviews in Medical Virology*, 10:119–133, 2000.

[4] A.W. Crosby. *America's Forgotten Pandemic: The Influenza of 1918.* Cambridge University Press, 2003.

[5] G.W. Shannon. Disease mapping and early theories of yellow fever. *The Professional Geographer*, 33(2):221–227, 1981.

[6] H. Brody, M. Rip, P. Vinten-Johansen, N. Paneth, and S. Rachman. Map-making and myth-making in Broad Street: The London cholera epidemic, 1854. *The Lancet*, 356:64–68, 2000.

[7] Andrew B. Lawson. *Statistical Methods in Spatial Epidemiology.* Wiley, 2001.

[8] J. Besag and J. Newell. The detection of clusters in rare diseases. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 154:143–155, 1991.

[9] M. Meselson, J. Guillemin, M. Hugh-Jones, A. Langmuir, I. Popova, A. Shelokov, and O. Yampolskaya. The Sverdlovsk anthrax outbreak of 1979. *Science*, 266:1202–1208, 1994.

[10] M.O. Ruiz, C. Tedesco, T.J. McTighe, C. Austin, and U. Kitron. Environmental and social determinants of human risk during a West Nile virus outbreak in the greater Chicago area, 2002. *International Journal of Health Geographics*, 3:8, 2004.

[11] P. Diggle. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 153:349–362, 1990.

[12] M.J. Keeling, M.E.J. Woolhouse, D.J. Shaw, L. Matthews, M. Chase-Topping, D.T. Haydon, S.J. Cornell, J. Kappey, J. Wilesmith, and B.T. Grenfell. Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, 294:813–817, 2001.

[13] P. Elliott, J. Wakefield, N. Best, and D. Briggs. *Spatial Epidemiology: Methods and Applications.* Oxford University Press, 2000.

[14] M. Kulldorff and N. Nagarwalla. Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14:799–810, 1995.

[15] M. Kulldorff. A spatial scan statistic. *Commun. Stat. Theor. M.*, 26:1481–1496, 1997.

[16] M. Kulldorff, L. Huang, L. Pickle, and L. Duczmal. An elliptical spatial scan statistic. *Statistics in Medicine*, 25(22):3929 – 3943, 2006.

[17] D.B. Neill. *Detection of spatial and spatio-temporal clusters.* PhD thesis, Carnegie Mellon University, Pittsburgh, 2006.

[18] T. Tango and K. Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4:11, 2005.

[19] L. Duczmal and R. Assunção. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Comput. Stat. Data Anal.*, 45:269–286, 2004.

[20] R. Assunção, M. Costa, A. Tavares, and S. Ferreira. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, 25:723–742, 2006.

[21] G.P. Patil and C. Taillie. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environ. Ecol. Stat.*, 11:183–197, 2004.

[22] C.T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.*, C20:68–86, 1971.

[23] Y. Xu, V. Olman, and D. Xu. Clustering gene expression data using a graph-theoretic approach: An application of minimum spanning trees. *Bioinformatics*, 18:536–545, 2002.

[24] M. de Berg, M. van Kreveld, M. Overmars, and Schwarzkopf O. *Computational Geometry: Algorithms and Applications.* Springer-Verlag, 2000.

[25] D.W. Merrill, S. Selvin, E.R. Close, and H.H. Holmes. Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four California counties. *Statistics in Medicine*, 15:1837–1848, 1996.

[26] D. Merrill. Use of a density equalizing map projection in analysing childhood cancer in four California counties. *Statistics in Medicine*, 20:1499–1513, 2001.

[27] S. Selvin and D. Merrill. Adult leukemia: A spatial analysis. *Epidemiology*, 13:151–156, 2002.

[28] A. Khalakdina, S. Selvin, and D.W. Merrill. Analysis of the spatial distribution of cryptosporidiosis in AIDS patients in San Francisco using density equalizing map projections (DEMP). *Int. J. Hyg. Environ. Health*, 206:553–561, 2003.

[29] M. Gastner and M. Newman. Diffusion-based method for producing density-equalizing maps. *Proc. Natl. Acad. Sci. U.S.A.*, 101:7499–7504, 2004.

[30] B. Bollobas. *Modern Graph Theory.* Springer-Verlag, 1998.

[31] J.S. Brownstein, H. Rosen, D. Purdy, J.R. Miller, M. Merlino, F. Mostashari, and D. Fish. Spatial analysis of West Nile virus: Rapid risk assessment of an introduced vector-borne zoonosis. *Vector Borne Zoonotic Dis.*, 2:157–164, 2002.

[32] E.G. Knox. Detection of clusters. In P. Elliott, editor, *Methodology of Enquiries into Disease Clustering*, pages 17–20. Small Area Health Statistics Unit, London, 1989.

[33] T. Tango. A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine*, 19:191–204, 2000.

[34] D.G.T. Denison and C.C. Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57:143–149, 2001.

[35] J.T.A.S. Ferreira, D.G.T. Denison, and C.C. Holmes. Partition modelling. In A.B. Lawson and D.G.T. Denison, editors, *Spatial Cluster Modeling*, pages 125–146. Chapman & Hall, London, 2002.

[36] O. Berke. Exploratory disease mapping: Kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3:18, 2004.

[37] T. Webster, V. Vieira, J. Weinberg, and A. Aschengrau. Method for mapping population-based case-control studies: An application using generalized additive models. *International Journal of Health Geographics*, 5:26, 2006.

[38] J.E. Kelsall and P.J. Diggle. Spatial variation in risk of disease: A nonparametric binary regression approach. *J. R. Stat. Soc. Ser. C Appl. Statist.*, 47(4):559–573, 1998.

[39] P.J. Diggle. *Spatial Epidemiology: Methods and Applications*, pages 87–103. Oxford University Press, Oxford, 2000.

[40] K.L. Olson, S.J. Grannis, and K.D. Mandl. Privacy protection versus cluster detection in spatial epidemiology. *Am. J. Public Health*, 96(11):2002–2008, 2006.

[41] J.S. Brownstein, C.A. Cassa, and K.D. Mandl. No place to hide – reverse identification of patients from published maps. *New England Journal of Medicine*, 355:1741–1742, 2006.

[42] D Dorling. Worldmapper: The human anatomy of a small planet. *PLoS Medicine*, 4(1):e1, 2007.

[43] S. Selvin, J. Schulman, and D.W. Merrill. Distance and risk measures for the analysis of spatial data: A study of childhood cancers. *Social Science and Medicine*, 34(7):769–777, 1992.

[44] A.C. Gatrell, T.C. Bailey, P.J. Diggle, and B.S. Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21(1):256–274, 1996.

[45] A.G. Chetwynd, P.J. Diggle, and A. Marshall. Investigation of spatial clustering from individually matched case-control studies. *Biostatistics*, 2(3):277–293, 2001.

[46] M.G. Voronoi. Nouvelles applications des parametres continus a la theorie des formes quadratiques. *J. Reine Angew. Math.*, pages 198–287, 1908.

[47] CM Gold, J. Nantel, and W. Yang. Outside-in: An alternative approach to forest map digitizing. *International Journal of Geographical Information Science*, 10(3):291–310, 1996.

[48] AB Mendes and IH Themido. Multi-outlet retail site location assessment. *International Transactions in Operational Research*, 11(1):1–18, 2004.

[49] R. Klein, K. Mehlhorn, and S. Meiser. Randomized incremental construction of abstract Voronoi diagrams. *Computational Geometry: Theory and Applications*, 3(3):157–184, 1993.

[50] F. Rezende, R.M.V. Almeida, and F.F. Nobre. Diagramas de Voronoi para a definição de áreas de abrangência de hospitais públicos no Município do Rio de Janeiro. *Cad Saúde Pública*, 16:467–75, 2000.

[51] I. Hanigan, G. Hall, and K.B.G. Dear. A comparison of methods for calculating population exposure estimates of daily weather for health research. *International Journal of Health Geographics*, 5(1):38, 2006.

[52] J F Bithell. A classification of disease mapping methods. *Statistics in Medicine*, 19:2203–2215, 2000.

[53] The Massachusetts Institute of Technology Geodata Repository.

[54] JF Bithell. An application of density estimation to geographical epidemiology. *Statistics in Medicine*, 9:697–701, 1990.

[55] CDC. Mumps epidemic – Iowa, 2006. *Morbiditiy and Mortality Weekly Report*, 55(13):366–368, 2006.

[56] CDC. Update: Multistate outbreak of mumps — United States, January 1–May 2, 2006. *Morbiditiy and Mortality Weekly Report*, 55:1–5, 2006.

[57] S.B. Eng, D.H. Werker, A.S. King, S.A. Marion, A. Bell, J.L. Issac-Renton, G.S. Irwin, and W.R. Bowie. Computer-generated dot maps as an epidemiologic tool: Investigating an outbreak of toxoplasmosis. *Emerging Infectious Diseases*, 5(6):815–9, 1999.

[58] T.J. Oyana, P. Rogerson, and J.S. Lwebuga-Mukasa. Geographic clustering of adult asthma hospitalization and residential exposure to pollution at a United States-Canada border crossing. *American Journal of Public Health*, 94(7):1250–1257, 2004.

[59] G.M. Jacquez, A. Kaufmann, J. Meliker, P. Goovaerts, G. AvRuskin, and J. Nriagu. Global, local and focused geographic clustering for case-control data with residential histories. *Environmental Health: A Global Access Science Source*, 4:4, 2005.

[60] D. Han, P.A. Rogerson, J. Nie, M.R. Bonner, J.E. Vena, D. Vito, P. Muti, M. Trevisan, S.B. Edge, and J.L. Freudenheim. Geographic clustering of resi-

dence in early life and subsequent risk of breast cancer (United States). *Cancer Causes and Control*, 15:921–929, 2004.

[61] A.J. Curtis, J.W. Mills, and M. Leitner. Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics*, 5:44, 2006.

[62] M.T. Wallin, W.F. Page, and J.F. Kurtzke. Multiple sclerosis in US veterans of the Vietnam era and later military service: Race, sex, and geography. *Annals of Neurology*, 55(1):65–71, 2004.

[63] K. Torugsa, S. Anderson, N. Thongsen, N. Sirisopana, A. Jugsudee, P. Junlananto, S. Nitayaphan, S. Sangkharomya, and A.E. Brown. HIV epidemic among young Thai men, 1991-2000. *Emerging Infectious Diseases*, 9(7):881–883, 2003.

[64] S.F. Olsen. Cluster analysis and disease mapping – why, when and how? a step by step guide. *British Medical Journal*, 313:863–866, 1996.

[65] T. Tango. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, 14(21-22):2323–34, 1995.

[66] P.A. Rogerson. The detection of clusters using a spatial version of the chi-square goodness-of-fit statistics. *Geographical Analysis*, 31(1):128–47, 1999.

[67] Sonia Friedman and Richard S. Blumberg. *Harrison's Internal Medicine*, chapter 276. McGraw-Hill, 16th edition, 2006.

[68] Bonen DK and Cho JH. The genetics of inflammatory bowel disease. *Gastroenterology*, 124(2):521–536, 2003.

[69] Warren Strober, Peter J. Murray, Atsushi Kitani, and Tomohiro Watanabe. Signalling pathways and molecular interactions of NOD1 and NOD2. *Nature*, 6, 2006.

[70] Jonas Halfvarson, Lennart Bodin, Curt Tysk, Eva Lindberg, and Gunnar Jrnerot. Inflammatory bowel disease in a Swedish twin cohort: A long-term follow-up of concordance and clinical characteristics. *Gastroenterology*, 124(7):1767–1773, 2003.

[71] Edward V. Loftus. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. *Gastroenterology*, 126, 2004.

[72] T Andus and V Gross. Etiology and pathophysiology of inflammatory bowel disease–environmental factors. *Hepatogastroenterology*, 47(31):29–43, 2000.

[73] J Aisenberg and HD Janowitz. Cluster of inflammatory bowel disease in three close college friends? *J Clin Gastroenterol*, 17(4), 1993.

[74] Anders Ekbom, Matthew Zack, Hans-Olov Adami, and Charles Helmick. Is there clustering of inflammatory bowel disease at birth? *American Journal of Epidemiology*, 134(8):876–886, 1991.

[75] D.S. Miller, Andrea Keighley, P.G. Smith, A. O. Hughes, and M.J.S. Langman. A case-control method for seeking evidence of contagion in Crohn's disease. *Gastroenterology*, 71(3):385–387, 1976.

[76] Amnon Sonnenberg and Irene Wasserman. Epidemiology of inflammatory bowel disease among U.S. military veterans. *Gastroenterology*, 101:122–130, 1991.

[77] M Valenciano, B Gagniere, C Maurage, H de Valk, and JC Desenclos. Étude d'un agrégat apparent de maladies de crohn en indre-et-loire, 1990-1999 [analysis of an apparent cluster of crohn's disease cases in indre-et-loire, france (1990-1999)]. *Rev Epidemiol Sante Publique*, 50(6):509–517, 2002.

[78] D O'Donovan, D Keegan, G McEvoy, H Mulcahy, and D O'Donoghue. A cluster of Crohn's disease in Ballybrack: Fact or fiction? *Endoscopy*, 36, 2004.

[79] L. Sweeney. k-Anonymity: A model for protecting privacy. *Int J Uncertainty Fuzziness Knowledge-Based Syst*, 10:557–570, 2002.

[80] NIH publication number 04-5489: Research repositories, databases, and the HIPAA privacy rule. http://privacyruleandresearch.nih.gov/research_repositories.asp, 2004.

[81] M.P. Armstrong, G. Rushton, and D.L. Zimmerman. Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18(5):497–525, 1999.

[82] United States 2000 census. http://www.census.gov/geo/www/cob/bg_metadata.html.

[83] C M Yuan, S Love, and M Wilson. Syndromic surveillance at hospital emergency departments – Southeastern Virginia. *MMWR Morb. Mortal. Wkly. Rep.*, 53 Suppl:56–58, 2004.

[84] L. Hammond, S. Papadopoulos, C.F. Johnson, S. MaWhinney, B. Nelson, and J.K. Todd. Use of an Internet-Based Community Surveillance Network to Predict Seasonal Communicable Disease Morbidity. *Pediatrics*, 109(3):414–418, 2002.

[85] F. Mostashari, A. Fine, D. Das, J. Adams, and M. Layton. Use of ambulance dispatch data as an early warning system for communitywide influenzalike illness, New York City. *J. Urban Health*, 80 Supplement 1:i43–i49, 2003.

[86] R. Heffernan, F. Mostashari, D. Das, M. Besculides, C. Rodriguez, J. Greenko, L. Steiner-Sichel, S. Balter, A. Karpati, P. Thomas, M. Phillips, J. Ackelsberg, E. Lee, J. Leng, J. Hartman, K. Metzger, R. Rosselli, and D. Weiss. New York City syndromic surveillance systems. *MMWR Morb. Mortal. Wkly. Rep.*, 53 Supplement:25–27, 2004.

[87] M.D. Lewis, J.A. Pavlin, J.L. Mansfield, S. O'Brien, L.G. Boomsma, Y. Elbert, and P.W. Kelley. Disease outbreak detection system using syndromic data in the greater Washington DC area. *Am. J. Prev. Med.*, 23(3):180–186, 2002.

[88] F.C. Tsui, J.U. Espino, V.M. Dato, P.H. Gesteland, J. Hutman, and M.M. Wagner. Technical description of RODS: A real-time public health surveillance system. *J. Am. Med. Inform. Assoc.*, 10:399–408, 2003.

[89] M. Paladini. Daily emergency department surveillance system—Bergen County, New Jersey. *MMWR Morb. Mortal. Wkly. Rep.*, 53 Supplement:47–49, 2004.

[90] ZF Dembek, K. Carley, A. Siniscalchi, and J. Hadler. Hospital admissions syndromic surveillance–Connecticut, September 200-November 2003. *MMWR Morb. Mortal. Wkly. Rep.*, 53 Supplement:50–52, 2004.

[91] M.M. Wagner, J.M. Robinson, F.C. Tsui, J.U. Espino, and W.R. Hogan. Design of a national retail data monitor for public health surveillance. *J. Am. Med. Inform. Assoc.*, 10:409–418, 2003.

[92] R. Platt, C. Bocchino, B. Caldwell, R. Harmon, K. Kleinman, R. Lazarus, A.F. Nelson, J.D. Nordin, and D.P. Ritzwoller. Syndromic surveillance using minimum transfer of identifiable data: The example of the National Bioterrorism Syndromic Surveillance Demonstration Program. *J. Urban Health*, 80 Supplement 1(2):i25–i31, 2003.

[93] D.L. Cooper, G. Smith, M. Baker, F. Chinemana, N. Verlander, E. Gerard, V. Hollyoak, and R. Griffiths. National symptom surveillance using calls to a telephone health advice service-United Kingdom, December 2001–February 2003. *MMWR Morb. Mortal. Wkly. Rep.*, 53 Supplement:179–83, 2004.

[94] L. Hutwagner, W. Thompson, G.M. Seeman, and T. Treadwell. The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS). *J. Urban Health*, 80 Supplement 1:i89–i96, 2003.

[95] LC Hutwagner, EK Maloney, NH Bean, L. Slutsker, and SM Martin. Using laboratory-based surveillance data for prevention: An algorithm for detecting Salmonella outbreaks. *Emerg. Infect. Dis.*, 3(3):395–400, 1997.

[96] L. Hutwagner, T. Browne, G.M. Seeman, and A.T. Fleischauer. Comparing aberration detection methods with simulated data. *Emerg. Infect. Dis.*, 11(2):314–6, 2005.

[97] J. Zhang, FC Tsui, MM Wagner, and WR Hogan. Detection of outbreaks from time series data using wavelet transform. In *Proc. AMIA Symp.*, pages 748–752, 2003.

[98] B.Y. Reis and K.D. Mandl. Time series modeling for syndromic surveillance. *BMC Med. Inform. Decis. Mak.*, 3, 2003.

[99] A.J. Beitel, K.L. Olson, B.Y. Reis, and K.D. Mandl. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. *Pediatr. Emerg. Care*, 20(6):355–360, 2004.

[100] M.L. Jackson, A. Baer, I. Painter, and J. Duchin. A simulation study comparing aberration detection algorithms for syndromic surveillance. *BMC Med. Inform. Decis. Mak.*, 7:6, 2007.

[101] J.C. Brillman, T. Burr, D. Forslund, E. Joyce, R. Picard, and E. Umland. Modeling emergency department visit patterns for infectious disease complaints: Results and application to disease surveillance. *BMC Med. Inform. Decis. Mak.*, 5, 2005.

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

[103] RE Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Rep.*, 78(6):494–506, 1963.

[104] W.W. Thompson, D.K. Shay, E. Weintraub, L. Brammer, C.B. Bridges, N.J. Cox, and K. Fukuda. Influenza-associated hospitalizations in the United States. *JAMA*, 292:1333–1340, 2004.

[105] K.D. Mandl, J.M. Overhage, M.M. Wagner, W.B. Lober, P. Sebastiani, F. Mostashari, J.A. Pavlin, P.H. Gesteland, T. Treadwell, E. Koski, L. Hutwagner, D.L. Buckeridge, R.D. Aller, and S. Grannis. Implementing syndromic surveillance: A practical guide informed by the early experience. *J. Am. Med. Inform. Assoc.*, 11:141–150, 2004.

[106] A Boggess and FJ Narcowich. *A First Course in Wavelets with Fourier Analysis.* Prentice Hall Press, Upper Saddle River, NJ, 2001.

[107] M. Aitkin. Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, 36, 1987.

[108] R.A. Rigby and D.M. Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6, 1996.

[109] TJ Hastie and RJ Tibshirani. *Generalized Additive Models.* Chapman & Hall, New York, NY, 1990.

[110] F. Dominici, A. McDermott, S.L. Zeger, and J.M. Samet. On the use of generalized additive models in time-series studies of air pollution and health. *Am. J. Epidemiol.*, 156:193–203, 2002.

[111] *Matlab User's Guide.* Mathworks, Inc., Natick, MA.

[112] K. Kleinman, R. Lazarus, and R. Platt. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *Am. J. Epidemiol.*, 159:217–224, 2004.

[113] L.A. Waller. Invited commentary: Surveilling surveillance – some statistical comments. *Am. J. Epidemiol.*, 159:225–227, 2004.

[114] K D Mandl, B Y Reis, and C Cassa. Measuring outbreak-detection performance by using controlled feature set simulations. *MMWR Morb. Mortal. Wkly. Rep.*, 53 Suppl:130–136, 2004.

[115] A. Goldenberg, G. Shmueli, R.A. Caruana, and S.E. Fienberg. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Science*, 99:5237–5240, 2002.