Spatial and Syndromic Surveillance for Public Health

Edited by

Andrew B. Lawson
Department of Epidemiology and Biostatistics
University of South Carolina, USA

Ken Kleinman
Department of Ambulatory Care and Prevention
Harvard Medical School, USA

John Wiley & Sons, Ltd

Spatial and Syndromic Surveillance for Public Health

Spatial and Syndromic Surveillance for Public Health

Edited by

Andrew B. Lawson
Department of Epidemiology and Biostatistics
University of South Carolina, USA

Ken Kleinman
Department of Ambulatory Care and Prevention
Harvard Medical School, USA

John Wiley & Sons, Ltd

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

Telephone (+44) 1243 779777

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

Other Wiley Editorial Offices

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

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

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

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

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

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

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

British Library Cataloguing in Publication Data

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

ISBN 0-470-09248-3 (HB)

Typeset in 10/12pt Photina by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by TJ International Ltd, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Preface

This volume hopes to fill a growing need for the description of current methodology in public health surveillance. Recent advances in syndromic surveillance and, more generally, in spatial and multivariate surveillance have never been collected in a single volume. The field of syndromic surveillance now attracts a wide audience due to the perceived need to implement wide-ranging monitoring systems to detect possible health-related bioterrorism activity. In addition, many computer systems have been, and are being, developed that have the capability to store and display large volumes of health data and to link dynamically between different data streams. This capability has not been matched with extensive statistical research into the properties of the methods used in these systems. The ability of these systems to correctly sound health alarms when needed is of paramount importance. It is the task of statisticians to develop and evaluate the methodologies and to ensure that the correct interpretation is made of evaluated data. This volume seeks to provide a synopsis of current practice as well as a starting point for the development and evaluation of methods.

In the production of this volume we have been helped by a great range of people. First we would like to thank our families who, throughout this venture, have been a great source of support. In addition we would like to thank the contributors for their timely submission of interesting articles, as well as our colleagues in respective departments for helping with evaluation and criticism. Finally we would like to thank the staff of Wiley Europe for their continual help during the sometimes fraught stages of production. In particular, we thank Kathryn Sharples the Statistics sub-editor and Lucy Bryan in production, as well as Richard Leigh, the copy-editor.

Andrew Lawson (Columbia)
Ken Kleinman (Boston)

List of Contributors

Marco Bonetti Instituto di Metodi Quantitativi Univerisita' Bocconi Viale Isonzo, 25 Milano Italy bonetti@jimmy.harvard.edu

Laura Forsberg Department of Biostatistics Harvard School of Public Health 655 Huntington Ave Boston MA 02115 USA lforsber@hsph.harvard.edu

Marianne Frisén Statistical Research Unit Göteborg University Box 660 SE 40530 Göteborg Sweden Marianne.Frisen@statistics.gu.se

Caroline Jeffery Department of Biostatistics Harvard School of Public Health 655 Huntington Ave Boston MA 02115 USA cjeffery@hsph.harvard.edu

Ken Kleinman Department of Ambulatory Care and Prevention Harvard Medical School & Harvard Pilgrim Health Care 133 Brookline Avenue Boston MA 02215 USA ken_kleinman@harvardpilgrim.org

Martin Kulldorff Department of Ambulatory Care and Prevention Harvard Medical School and Harvard Pilgrim Health Care 133 Brookline Avenue 6th Floor Boston MA 02215 USA martin_kulldorff@hms.harvard.edu

Andrew B. Lawson Department of Epidemiology and Biostatistics Arnold School of Public Health University of South Carolina Columbia SC 29208 USA alawson@gwm.sc.edu

David Madigan Faculty of Arts and Sciences Rutgers University 77 Hamilton Street New Brunswick NJ 08901 USA dmadigan@rutgers.edu

Andrew W. Moore Carnegie Mellon University School of Computer Science 5000 Forbes Avenue Pittsburgh PA 15213 USA awm@cs.cmu.edu

Daniel B. Neill Carnegie Mellon University School of Computer Science 5000 Forbes Avenue Pittsburgh PA 15213 USA neill@cs.cmu.edu

Al Ozonoff Department of Biostatistics Boston University School of Public Health 715 Albany Street Talbot East 512 Boston MA 02118 USA aozonoff@bu.edu

Marcello Pagano Department of Biostatistics Harvard School of Public Health 655 Huntington Ave Boston MA 02115 USA pagano@hsph.harvard.edu

Peter A. Rogerson Departments of Geography and Biostatistics and National Center for Geographic Information and Analysis University at Buffalo Buffalo NY 14261 USA rogerson@buffalo.edu

Christian Sonesson Statistical Research Unit Göteborg University Box 660 SE 40530 Göteborg Sweden Christian.Sonesson@statistics.gu.se

Yann Le Strat Département des Maladies Infectieuses Institut de Veille Sanitaire 12 rue du Val d'Osne 94415 Saint-Maurice Cedex France

Weng-Keen Wong Center for Biomedical Informatics 200 Lothrop Street 8084 Forbes Tower Pittsburgh PA 15213 USA wwong@cbmi.pitt.edu

1

Introduction: Spatial and Syndromic Surveillance for Public Health

Andrew B. Lawson and Ken Kleinman

1.1 WHAT IS PUBLIC HEALTH SURVEILLANCE?

The Centers for Disease Control and Prevention (CDC) define public health surveillance as:

> the ongoing, systematic collection, analysis, and interpretation of health data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know. The final link of the surveillance chain is the application of these data to prevention and control. A surveillance system includes a functional capacity for data collection, analysis, and dissemination linked to public health programs.
>
> (Thacker, 1994)

It is clear from this that a broad definition of surveillance is implied and that it relates to a wide range of monitoring methods related to health. From a statistical point of view it is relevant to consider how statistical methods can be developed or employed to best aid the task of surveillance of populations. This will require using all of the relevant data available for analysis. It will certainly include information about *where* the data was recorded as well as *when* it was observed.

1.1.1 Spatial Surveillance

There is thus a need to combine the thinking in two previously mostly distinct fields of statistical research, namely surveillance, which generally constitutes

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

monitoring statistics for evidence of a change, and spatial techniques, which are often used to find or describe the extent of 'clustering' across a map. While both endeavors pre-date the formal study of statistics as a discipline, they have rarely been combined. More often, as in the famous case of John Snow and cholera in London in 1854, note of an increase in a global statistic has been followed by a spatial analysis to determine whether the increase is localized or general. Interest in doing spatial monitoring of data as it accrues has been greatly enhanced by two developments: the perceived need to quickly detect bioterrorism after the terrorist dissemination of *Bacillus anthracis* in October 2001, and the increasing availability of data that contains spatial (geographical location) information.

1.1.2 Syndromic Surveillance

Another result of the burgeoning availability of data has been the recognition of a need and an opportunity. The need is for the ability to group symptoms together in broad groups that combine similar types of complaints – this being necessary to ensure that two cases attributable to the same cause are not considered separately due to variable coding practice on the part of health care providers. Misappropriating from medical nomenclature, these groups of symptoms are loosely designated as 'syndromes'. The opportunity presented by the increasing availability of data is to use, for public health purposes, data that has not often been recognized as useful in this way. Examples include information about school absenteeism and over-the-counter sales of remedies such as anti-diarrheals.

Together, grouping of large numbers of symptoms and data regarding nontraditional sources of information are labeled as 'syndromic surveillance'. The putative advantage of syndromic surveillance is that detection of adverse effects can be made at the earliest possible time, possibly even before disease diagnoses can be confirmed through unmistakable signs or laboratory confirmation.

1.2 THE INCREASED IMPORTANCE OF PUBLIC HEALTH SURVEILLANCE

In addition to the CDC definition, we might consider a dictionary definition of surveillance: 'the close observation of a person or group, especially one under suspicion'. In this light we would define public health surveillance as the monitoring of the health of the public for the onset or outbreak of illness. The illness surveilled may be rare (plague) or recurrent (influenza), natural or intentional (bioterrorism).

Since the intentional release of anthrax in the USA in October 2001, there has been a great deal of interest in establishing systems to detect another such

attack as early as possible, should it occur. The need for early detection is motivated by two facts. First, many agents that might be used for such an attack have a prodromal phase that is relatively nonspecific, with symptoms that often resemble those of the common cold. This describes anthrax, botulism, plague, smallpox, and tularemia – all of the CDC class A bioterrorism agents except for viral hemorrhagic fevers (CDC, 2004). If the attack can be detected while most victims are in this phase, they may be helped by specialized care, and future onsets may be prevented by prophylaxis. Second, for contagious diseases, earlier interdiction can slow down or stop the epidemic curve; the latter might be impossible if detection were delayed.

A great deal of resources are being expended on mechanical detection of airborne organisms, such as anthrax spores. We do not discuss such efforts here and merely observe that from the statistical perspective, detection is finished once a spore of anthrax has been positively identified. (Determining the locations in need of treatment or prophylaxis is a separate question). In this book, we focus on individual human beings, in contrast to disease organisms. While it is true that a single definitive diagnosis of anthrax or any of the organisms cited above also ends the statistical interest in detection, the fact of the nonspecific prodrome opens a window for detection of an attack before a definitive diagnosis has been made. To wit, since the prodromal symptoms are so common, one might search for unusual increases in symptoms consistent with the prodrome. Such an increase could be due either to natural variation in the symptom incidence or to an attack with some agent that causes those symptoms in the prodrome.

In this context, we are most interested in detecting attacks while they are ongoing rather than retrospectively. In statistical terms, we might refer to this as 'cluster detection' or 'incident cluster detection', where by 'cluster' we mean the occurrence of extra cases in a short time span. In the literature on surveillance, this is sometimes referred to as 'on-line' surveillance (Chapter 3). Many techniques exist for ongoing monitoring or surveillance of a count; these come from industrial applications – for example, Shewhart control tables and cumulative sum (CUSUM) methods (Chapter 2) – as well as from public health surveillance (Huttwagner, 2003; Sonesson and Bock, 2004).

1.3 GEOGRAPHIC INFORMATION, CLUSTER DETECTION AND SPATIAL SURVEILLANCE

The increased need for cluster detection has coincided with an increasing availability of data, especially data on the location of events. This is often obtained by geocoding the addresses of individual cases. This can be done 'on the fly' as cases are encountered (Beitel *et al.*, 2004) or with static databases that retain the location of all patients eligible for surveillance (Lazarus *et al.*, 2002). In its simplest form, geocoding could imply merely obtaining the zip or postal code, but it may also include finding the exact latitude and longitude of an address

using geographical information systems (GIS). In statistical jargon, such data about location is often referred to as 'spatial' data.

The value of spatial data for cluster detection is twofold. First, all attacks are localized at some spatial scale. That is, an attack could conceivably target a neighborhood, but on a city-wide scale this would be a small area. Alternatively, an attack could include a whole metropolitan area, but on a national scale this would be a small region. When surveillance is limited to a single daily count from a neighborhood or city, even sharp increases in relatively small regional counts may be hidden within the natural variation found in the count across a larger area. Spatial surveillance thus promises to increase the power to detect events that occur in small regions, relative to surveillance of the total count only. Secondly, if an incident cluster is identified, public health officials will need to respond. If the data are nonspatial, surveillance can only give vague messages of the sort 'there is an excess of cases in the Boston metropolitan region'; this is unlikely to be of much practical use. In contrast, spatial surveillance would allow more-specific messages, such as 'there are excess cases in zip code 02474'. The job of identifying small regions with extra cases is also referred to as 'cluster detection', where the clustering in this case refers to extra cases in an area on the map.

The coincidence of suddenly increased need and increasingly available spatial data has generated new interest in statistical methods for spatial surveillance, which might be described as the detection of incident clusters in space. The goal of this book is to provide a snapshot of the state of the nascent art of incident spatial cluster detection, provided by statisticians involved in traditional surveillance (of a single statistic), in spatial clustering, and in spatial surveillance.

1.4 SURVEILLANCE AND SCREENING

An idea related to surveillance is that of screening. The use of screening to allow the early detection of disease onset is well established, though possibly controversial, in such areas as cervical or mammarian cancer. These examples of screening involve testing individuals at regular time points to attempt to assess if onset of a condition has occurred or is likely or imminent. Screening could be applied to populations as well as individuals, in that changes in public health might trigger interventions. Such interventions could be designed to redirect health resources towards attempts to improve the health status of the population. However, screening is usually associated with individual assessment or monitoring, while surveillance is usually carried out at an aggregate population level.

Surveillance and screening share an implicit temporal dimension: populations or individuals are assessed (often repeatedly over time) to assess whether changes have occurred which may warrant action. In general, a change is

defined as exceeding limits describing the acceptable results of current observation and actions taken if these limits are passed. In screening individuals, the limits may be based on a previously observed known or stable abnormal baseline or on 'normal' standards thought to obtain in healthy persons.

In surveillance, the 'normal' case is rarely known, and most attention is directed to detect passing limits based on observed or expected patterns. These limits may be fixed or may depend on the status of ancillary variables. For example, incidence of influenza-like illness would be expected to vary seasonally, so similar numbers of cases would be more or less alarming at different times of year.

To carry the screening analogy further, the location of the public health incident is as important as the fact that it occurred. A public health report indicating only a disease outbreak is comparable to a garbled mammography report that only indicates a cancer but no suggestion of which breast is affected, let alone a location in which a biopsy would be appropriate. In population-level analysis, statisticians use 'spatial' statistics to discuss location. Further still, mammography uses the spatial information to help identify the existence of the node in the first place.

1.5 OVERVIEW OF PROCESS CONTROL AND MAPPING

Process monitoring is necessary for quality control in a manufacturing context. The subject of statistical process control (SPC) has received the most methodological attention of all surveillance questions. SPC has formed the basis for many disease surveillance systems. In this section we describe some basic SPC methods that could be applied in this context.

1.5.1 Process Control Methodology

A number of methods have been developed for the detection of changes in populations over time. These methods are characterized by the estimation of changepoints in a sequence of disease events or a time series of population rates, or the determination of or application of control limits to the behavior of a system. In this area there are some simple methods available to assist in the assessment of change or 'in control' behavior. Some of these methods are derived from SPC, which was developed for the monitoring of industrial processes over time, and could be applied within a disease surveillance program, with due care. For example, it is well known that the temporal variation in count data can be monitored by using a Poisson control chart (C or U chart), upon which specific limits can be plotted beyond which corrective action should be taken. These charts are based on normal pivotal approximations.

An exact interval could be constructed for independent Poisson counts in an attempt to utilize SPC methods. However, if the counts were correlated even under the null hypothesis, then some allowance must be made for this correlation in the chart. A further issue, when such methods are to be used within disease monitoring, is the issue of how to incorporate any changes in the background 'at-risk' population which may arise. One possibility, in the temporal domain, is to employ relative risk estimates. For large aggregation scales, time-series methods have been employed which allow temporal dependence (see Chapter 2 in this volume).

In addition, special types of chart (CUSUM charts) have been developed specifically to detect changes in pattern over time (changepoints). These are constructed by cumulative recording of events over time, the accumulation being found to be sensitive to changepoints in the process under consideration. Some recent work in the application of these ideas in medical surveillance and monitoring has been done by Frisén and co-workers (Chapters 3 and 9 in this volume). These methods require special adaptations to be developed to deal with the spatial and spatio-temporal nature of geographical surveillance.

The main issues within temporal surveillance which impact on spatial surveillance and spatio-temporal surveillance can be categorized into three classes: detection of changepoints (mean level, variance), detection of clusters, and the detection of overall process change. Conventional SPC would use control limits to detect shifts in single or multiple parameters where the target parameters are usually constant. However, disease incidence varies naturally in time and so allowance must be made for this variation in any monitoring system, particularly with variation in population at risk. In addition, particular departures from the 'normal' variation are often of greater interest than simple shifts of parameters. Changepoints, where jumps in the incidence occur, could be a focus of interest. Alternatively, clusters of disease may be important. Finally, there may be an overall process change, where various parameters change. Any disease surveillance system is likely to be focused on one or all of these changes. Indeed, it is the multiple focus of such systems that is one of the greatest challenges for the development of statistical methodology.

1.5.2 The Analysis of Maps and Surveillance

In the spatial case, there is a wide range of methods that can be applied to a single map of case events within a fixed time frame/period. Many of the methods applied in disease mapping, clustering or ecological analysis could be applied as a surveillance tool. For example, general clustering tests could be applied or residuals from disease maps fitted in each time period could be examined. Questions which might be appropriate to answer with these methods are such as: Is there evidence of unusual variation in incidence in the map? Is there evidence of 'unusual' clustering on the map? Is there a spatial trend on the map related to, for example, a putative source?

However, when the question relates to a spatio-temporal pattern or change in pattern, then there are few methods currently available which are designed for this purpose. There is a correspondence between the temporal surveillance foci, and features which are important to detect in the spatial domain. First, localized discontinuities in mean level or variance of risk could be of concern (changepoints). Second, spatial clusters of disease could be a focus. Finally, overall process change could also be envisaged spatially.

1.6 THE PURPOSE OF THIS BOOK

We hope that this book may serve a dual purpose. First, we hope that the potential users of spatial surveillance – the public health authorities – will use it as an introduction to the value of spatial data and as guide to analytic methods competing for scarce resources. Second, we hope that the statistical community will use it as a spur to further development of techniques and to resolution of questions unanswered by the chapters which follow.

1.6.1 Statistical Surveillance and Methodological Development in a Public Health Context

The ongoing aim of public health surveillance, since the time of John Snow, has been to identify public health problems as they occur and respond appropriately when they do. Breaking this down into discrete steps, this involves determining from whom to collect data, collecting the data, summarizing the data, evaluating the summarized data, and taking action if the evaluation warrants it. 'Action' can be defined as any additional steps that are not performed on a routine basis, which might include everything from asking for additional data to the vigilante removal of a pump handle or launching some other direct interdiction to prevent further illness.

1.6.2 The Statistician's Role in Surveillance

As statistical analysts, it is important that we remember that our role in public health surveillance is in evaluating the summarized data; this may be the main factor in the decision whether to take action *now*. This is different from academic work and most scientific work in several important ways.

First of all, we do not have the luxury of academic distance from the subject. This means that the case for making the best-supported decision needs to be made in a way that is powerful and easily absorbed by decision-makers. These authorities may lack the time, inclination, centralization, or training to follow complicated arguments or abstract presentations of our evaluations.

One example of simplifying the evaluation is included in Chapter 5, on modeling, where we discuss inverting the *p*-value as a means of making the unusualness of the result under the null more approachable. A second difference is that the time available to develop analyses is very brief. In contrast, the motivation to complete statistical work quickly mainly derives from fear of competition or career goals. The impact of this difference is that simpler, more easily generalizable methods have a practical advantage – they can be deployed in practical situations. Methods tailored to one situation may fit that data better but be essentially useless when the data changes or when applying the method to a different situation. Similarly, health departments may never use methods so complicated that they require the active participation of a Ph.D. statistician. Finally, in our experience, decision-makers often urgently request results immediately after data becomes available. This is another impetus towards easily generalizable, easily used methods.

On the other hand, while statistical evaluation may be a key datum informing a decision, we should also remember that it is not actually the decision itself. Other data important to decision-makers include financial, political, and organizational feasibility considerations.

1.7 THE CONTENTS OF THIS BOOK

In the initial section (Part I), we provide an introduction and grounding in traditional temporal surveillance. This includes the current chapter, plus an overview and an evaluation of methods used in purely temporal surveillance. The goal of these chapters is to bring a reader unfamiliar with surveillance to a level that subsequent chapters can be more easily digested. This is necessary because those chapters may take as read concepts native to traditional temporal surveillance.

We begin with a discussion of purely temporal surveillance by Yann Le Strat (Chapter 2). Purely temporal surveillance is commonly used in most public health departments, and is an area studied little outside the areas of statistical process control and surveillance. Thus statistical readers may find a review helpful. The chapter includes an introduction to surveillance as well as a survey of typical methods. Methods considered include historical (nonstatistical) limits, process control charts (Shewhart charts, moving average charts, exponentially weighted moving average charts, CUSUM charts), time-series analysis, combinations of process control and time-series methods, integer-valued autoregressive processes, Serfling's method, and log-linear and other parametric models.

We also provide a discussion, by Marianne Frisén and Christian Sonesson (Chapter 3), of optimality in surveillance, and how detection methods might be designed with optimality in mind. This includes a discussion of evaluation metrics for surveillance (including false alarms, delay before alarm, and predictive value of alarms) and optimality criteria (including minimal expected

delay, minimax optimality, and average run length). The chapter goes on to discuss the optimality and performance features of several methods described in Chapter 2. It concludes by discussing several features of the public health environment that differentiate it from other applications of surveillance.

In Part II of the book, we provide a summary and some development of statistical approaches currently applied for spatial surveillance. First, Chapter 4 provides an overview of spatial and spatio-temporal health analysis outside of surveillance. This includes a discussion of disease mapping in the cases where individual locations of each case are known and alternatively when cases are aggregated into regions, as well as assessment of maps through residuals and goodness of fit. Finally, spatio-temporal and surveillance issues are introduced in the spatial context.

In Chapter 5, a summary of generalized linear models and generalized linear mixed models, including the use of binomial and Poisson models, is offered. Another purpose of the chapter is to note advantages that are realized through Poisson models (including variable-duration cluster signals) and to compare the surveillance resulting from the various models in an example data set.

In Chapter 6, Peter Rogerson addresses how CUSUM methods can be adapted to spatial surveillance. This includes a discussion of statistical process control that can be tailored for use in spatial applications, followed by a demonstration. Uses include surveillance of multiple local regions as well as of global statistics.

Martin Kulldorff (Chapter 7) discusses how scan statistics can be used in this context, and recent developments in this approach. The chapter mentions tests of spatial randomness, then introduces scan statistics. This is followed by a thorough introduction to the practical application of scan statistics for spatial health surveillance. This includes a discussion of the null and alternative hypotheses for the test, as well as the power and methods for displaying the suggested clusters. Finally, some applications in cancer clustering, infectious disease, other human diseases, veterinary medicine, and plant disease are surveyed.

In Chapter 8, Laura Fosberg and co-workers discuss distance methods for cluster detection and identification. This includes a motivation and summary of distance-based methods, the introduction of a new statistic based on distances, and a simulation-based evaluation of the new statistic. An example of syndromic spatial surveillance using the statistic is provided.

Next Christian Sonesson and Marianne Frisén (Chapter 9) consider multivariate surveillance, what is often described as multiple streams of surveillance data. This topic addresses the common case where either different data sources supply information regarding a single syndrome, or where a single data provider reports on multiple syndromes. The approaches mentioned include a reduction of dimensionality (to one or a few statistics) for each time point, parallel surveillance, vector accumulation methods, and simultaneous solution. They also discuss evaluation in this context.

In Part III, advanced approaches to syndromic and spatial surveillance are considered, including Bayesian models and data mining techniques.

In Chapter 10, Neil and co-workers discuss the use of Bayesian networks and the development of computational algorithms; in Chapter 11 they consider speeding up spatial processing of large data sets. In Chapter 12, David Madigan provides an example of Bayesian modeling of temporal surveillance using hidden Markov models. Finally, in Chapter 13, general issues in the Bayesian analysis of syndromic data and the model-based detection of spatial and spatio-temporal clusters as they evolve in time are discussed.

PART I

Introduction to Temporal Surveillance

2

Overview of Temporal Surveillance

Yann Le Strat

2.1 INTRODUCTION

The threat of emerging infections and the increased potential for bioterrorist attacks have introduced an additional importance for surveillance systems. The main objective of surveillance is to monitor the incidence or prevalence of specific health problems over time, within a well-defined population. But a wide range of objectives can be considered. Among these, detecting or monitoring outbreaks and monitoring trends represent statistical challenges. Temporal health surveillance is a vast domain. After a brief and incomplete description of surveillance systems and attributes, this chapter will focus on a review of statistical methods for the detection of unusual health events.

2.1.1 Surveillance Systems

Most disease surveillance systems are passive. A passive approach means that the organization conducting surveillance leaves the initiative for reporting to potential reporters. Contrary to an active system, the organization does not regularly contact physicians or hospitals to obtain reports. Surveillance can be conducted in many ways. To simplify, one can briefly identify two main approaches: exhaustive reporting and voluntary reporting.

In the former approach, notifiable diseases, essentially infectious diseases, are designated by public health agencies and by law, and their occurrence must be reported. In the USA, each state can designate which diseases are reportable by law. A physician, a laboratory where the diagnosis is made or a hospital where

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

the patient is treated may be included in the system. While each case should be declared, the surveillance system rarely detects every case in practice.

In the voluntary reporting approach, laboratory-based surveillance relies on clinicians, laboratory staff, microbiologists or infection control personnel to voluntarily report test results on a standard form to the public health system. Conventional reporting methods include mail, fax, and telephone. One advantage of this system is to give detailed information about the results of diagnostic tests. However, patients having laboratory tests may not be representative of all persons with the disease. From a statistical point of view, this can become problematic when inference is made from the sample to the general population of interest. It is reinforced by the fact that the laboratories do not represent a random sample and are simply those laboratories that volunteer to participate. Other specific surveillance networks are developed when more detailed information is required. Participants of these networks are also volunteers. Sampling of sites, hospitals or individuals is more statistically suitable, but only people interested in the surveillance participate, and the construction of a sample is generally more time-consuming than identifying volunteer participants. As a representative sample cannot be obtained, volunteers should be as heterogeneous as their patients.

Other approaches are possible. Registries which contain listings of a disease within a defined area can be used for the surveillance of diseases. Data from registries such as the national cancer registries include demographic characteristics, exposures, and treatments. Periodic surveys allow the monitoring of behavior associated with disease. A more detailed description of surveillance methods can be found in Thacker *et al.* (1983) or Buehler (1998).

2.1.2 Surveillance Attributes

The success of a surveillance system depends on a number of attributes, including simplicity, flexibility, acceptability, sensitivity, predictive value positive, representativeness, and timeliness. Surveillance systems are judged using these attributes (Centers for Disease Control and Prevention (CDC), 2001). We will discuss briefly four of these attributes.

- Sensitivity can be assessed by estimating the proportion of cases of a disease or health condition detected by the surveillance system. Sensitivity can also be considered as the ability of the system to detect unusual events. If the main objective of the system is to monitor trends, a reasonably low but constant sensitivity over time may be acceptable. However, if the objective is to detect epidemics, high sensitivity is required.
- The predictive value positive (PVP) is, firstly, the proportion of persons identified as cases who really are cases. Secondly, if the aim is detection, PVP is the proportion of epidemics identified by surveillance that are true epidemics. A low value of PVP will indicate that unnecessary investigations are being made.

- Representativeness is based on the comparison of the characteristics of reported events with those, partially unknown, in the target population. Representativeness of a surveillance system can be judged using knowledge of characteristics of the population (age, socioeconomic status, geographic location, etc.) and of the disease (latency period, mode of transmission, etc.). In most countries, evaluation of notifiable disease surveillance systems has found that communicable illnesses are underreported.
- Timeliness reflects the delay between steps in the surveillance system, from information collection to dissemination. One of the most crucial time intervals is between the onset of the health event and the report of this event to the public health agency. The control and prevention measures greatly depend on timeliness. There is a need to make disease surveillance more sensitive, specific, and timely. The development of automated reporting systems seems to be a valuable alternative. Electronic laboratory reporting will deliver more timely notifications than paper-based methods.

2.1.3 Early Detection of Unusual Health Events

Of critical importance to public health practitioners is an ability to rapidly detect any substantial changes in disease, thus facilitating timely public health interventions. Over the last 20 years, a number of statistical methods to detect changes in public health surveillance time-series data have been proposed. All of these methods look at the occurrence of a health event and test for a departure from an expected number based on the historical incidence of the event. Stroup *et al.* (1993) used the term 'aberration' when a change in the occurrence of a health event was statistically different from historical data.

2.1.3.1 Detection of an outbreak or detection of an aberration?

An outbreak is classically defined by the CDC as (i) a single case of a communicable disease long absent from a population, or (ii) the first invasion by a disease not previously recognized in that area requiring immediate reporting and epidemiologic investigation, or (iii) two or more cases of a disease associated in time and place (American Public Health Association, 2000). This definition is not adapted to the prospective detection of outbreaks because a statistical alarm must be triggered before any epidemiologic investigation and thus before the determination of a potential epidemiologic link between cases. The statistical alarm signals an aberration which can be a potential outbreak but can also be sporadic cases occurring at the same time or artifacts of the surveillance system. An aberration is suspected when the number of reported cases exceeds expected levels derived from historical data. Then a statistical test determines if, for each time period, the number of reported cases is significantly higher than the expected values. If the observed number is significantly higher than the expected number, an aberration is declared and a statistical alarm

triggered. In a further step, epidemiologic investigations allow the classification of the aberration as an outbreak or not. Depending on the gravity of the disease, interventions can be initiated to control and prevent the disease.

2.1.3.2 What information is collected and how is it used?

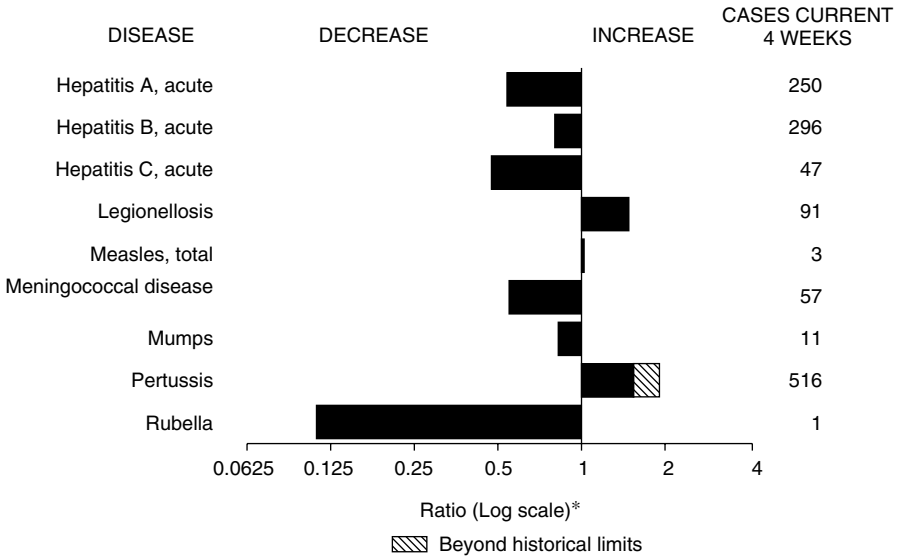
Surveillance is said to be prospective (or on-line) when the number of cases of a disease is recorded and analyzed sequentially over time. In contrast to retrospective surveillance, the time period between the onset of the health event and the report of this event to the public health agency is crucial when the aim of surveillance is the outbreak detection. Low reporting delays are essential if the aim is to detect an increased incidence as quickly as possible. One important question is how reporting delays can be taken into account in a statistical detection method. One solution is to consider a correction factor based on the distribution of reporting delays to impute the dates of onset. This approach has been successfully used in the epidemiology of AIDS (Brookmeyer and Gail, 1994; Lui and Rudy, 1989), but it requires that all dates are known. A second solution is to work with the dates of receipt of the case reports. The use of dates of receipt may be viewed as more reliable than imputation, but the main drawback of this approach is the loss of sensitivity and specificity. While the mean of the reporting delay is a major component of the timeliness of a surveillance system and consequently of outbreak detection, the variance of reporting delays (between participants involved in the system) affects its sensitivity (Farrington and Andrews, 2004).

Recent bioterrorist threats have increased the need for very early detection of outbreaks and hence the need to reduce the mean and variance of reporting delays. To this end, funds have been allocated by the US government to various public health agencies to build innovative surveillance systems. New syndromic surveillance systems have been designed for the early detection of the first symptomatic cases. These systems use different sources of information (primary care physician visits, emergency department admissions, infectious disease specialists, etc.) and make use of electronic reporting and the internet. The timeliness of syndromic surveillance is potentially better, but these systems do not use the same reporting sources and focus on first cases rather than on an increase in the number of reported cases as in traditional systems.

2.2 STATISTICAL METHODS

2.2.1 Historical Limits Method

State health departments in the USA report weekly the numbers of cases of a set of notifiable diseases to the CDC National Notifiable Diseases Surveillance System (NNDSS). Since 1990, these data have been published in graphical format in the *Morbidity and Mortality Weekly Report* as shown in Figure 2.1. A bar graph



* Ratio of current 4-week total to mean of 15 4-week totals (from previous, comparable, and subsequent 4-week periods for the past 5 years). The point where the hatched area begins is based on the mean and two standard deviations of these 4-week totals.

Figure 2.1 Selected notifiable disease reports, United States. Comparison of provisional 4-week totals June 19, 2004 with historical data.
Source: Reprinted from Centers for Disease Control and Prevention, *Morbidity and Mortality Weekly Report*, **53**, 537, June 25, 2004.

Table 2.1 Summary of provisional cases of selected notifiable diseases, United States, cumulative, week ending June 19, 2004 (24th Week)*.

	Cum. 2004	Cum. 2003
Anthrax	—	—
Botulism:	—	—
foodborne	7	7
infant	26	31
other (wound & unspecified)	4	10
Brucellosis†	47	41
Chancroid	14	28
Cholera	2	1
Cyclosporiasis†	59	24
Diphtheria	—	—
Ehrlichiosis:	—	—
human granulocytic (HGE)†	41	50
human monocytic (HME)†	29	41
human, other and unspecified	1	8
Encephalitis/Meningitis:	—	—
California serogroup viral†	—	—
eastern equine†	—	1

Table 2.1 (*continued*)

	Cum. 2004	Cum. 2003
Powassan[†]	—	—
St. Louis[†]	—	3
western equine[†]	—	—
Hansen disease (leprosy)[†]	36	35
Hantavirus pulmonary syndrome[†]	7	12
Hemolytic uremic syndrome, postdiarrheal[†]	35	49
HIV infection, pediatric[†§]	78	102
Measles, total	16[¶]	27[**]
Mumps	90	111
Plague	—	1
Poliomyelitis, paralytic	—	—
Psittacosis[†]	3	5
Q fever[†]	22	35
Rabies, human	—	—
Rubella	13	4
Rubella, congenital syndrome	—	1
SARS-associated coronavirus disease[†††]	—	7
Smallpox[†§§]	—	NA
Staphylococcus aureus:	—	—
Vancomycin-intermediate (VISA)[†§§]	4	NA
Vancomycin-resistant (VRSA)[†§§]	1	1
Streptococcal toxic-shock syndrome[†]	53	109
Tetanus	7	3
Toxic-shock syndrome	49	68
Trichinosis	3	—
Tularemia[†]	19	11
Yellow fever	—	—

-: No reported cases.
[*] Incidence data for reporting years 2003 and 2004 are provisional and cumulative (year-to-date).
[†] Not notifiable in all states.
[§] Updated monthly from reports to the Division of HIV/AIDS Prevention — Surveillance and Epidemiology, National Center for HIV, STD, and TB Prevention. Last update May 23, 2004.
[¶] Of 16 cases reported, nine were indigenous, and seven were imported from another country.
[**] Of 27 cases reported, 19 were indigenous, and eight were imported from another country.
[††] Updated weekly from reports to the Division of Viral and Rickettsial Diseases, National Center for Infectious Diseases (notifiable as of July 2003).
[§§] Not previously notifiable.

shows, for a set of infectious diseases under surveillance, a comparison between the number of reported cases in the current 4-week period and a baseline value (Centers for Disease Control & Prevention, 1988). This baseline is the average of the reported number of cases for the preceding 4-week period, the corresponding 4-week period and the following 4-week period, for the previous 5 years. Fifteen

values are obtained and a ratio is calculated by dividing the current 4-week total by the mean of the 15 values (Stroup *et al.*, 1989). For each disease, the ratio is displayed on a logarithmic scale. Historical limits of the ratio are calculated as

$$1 \pm \frac{2\sigma}{\mu},$$

where the mean μ and the standard deviation σ are calculated from the 15 historical incidence values. Kafadar and Stroup (1992) discuss the estimation of the variance of the ratio when surveillance data exhibit correlation. Three major drawbacks of this method can be noted: (i) it does not incorporate a trend; (ii) it ignores correlation between counts; and (iii) the underlying normality assumption is not always verified, in particular for rare health events. However, this technique provides a weekly synthetic summary of unusually large numbers of reported cases to epidemiologists, clinicians and other public health professionals and the method can be applied easily – see, for example, Birnbaum (1984) for the analysis of hospital infection surveillance data.

2.2.2 Process Control Charts

Assume that the observations $x = (x_1, \ldots, x_t, \ldots)$ are a realization of the stochastic process $X = (X_1, \ldots, X_t, \ldots)$. Usually, process control charts require random variables which are independent and normally distributed when the process is in statistical control. The basic idea of process control charts is to construct a statistic, denoted by y_t. When this control statistic exceeds predetermined control limits, the process under study is said to be statistically out of control. An alarm is then triggered, meaning a statistical aberration, that is, the existence of an unusual event. As mentioned by Williamson and Hudson (1999), the choice of the appropriate control limits is sometimes difficult. However, upper and lower control limits are usually expressed as a multiple of the process standard deviation (e.g. ± 3 standard deviations).

2.2.2.1 Shewhart chart

The Shewhart chart (Shewhart, 1931) is the simplest form of control chart. An alarm is triggered for the first time t when the value of $|x_t|$ exceeds predetermined control limits. This indicates that the process level has shifted from its previous level, that is, that the process is statistically out of control. The Shewhart chart is known to be slow in detecting small changes, but it rapidly detects large shifts in the process (see Chapter 3). An illustration of the Shewhart chart is given in Figure 2.2. Applied to the monthly poliomyelitis cases in the USA between January 1970 and December 1983, the Shewhart chart clearly identified four values above the upper control limit (UCL). The first value

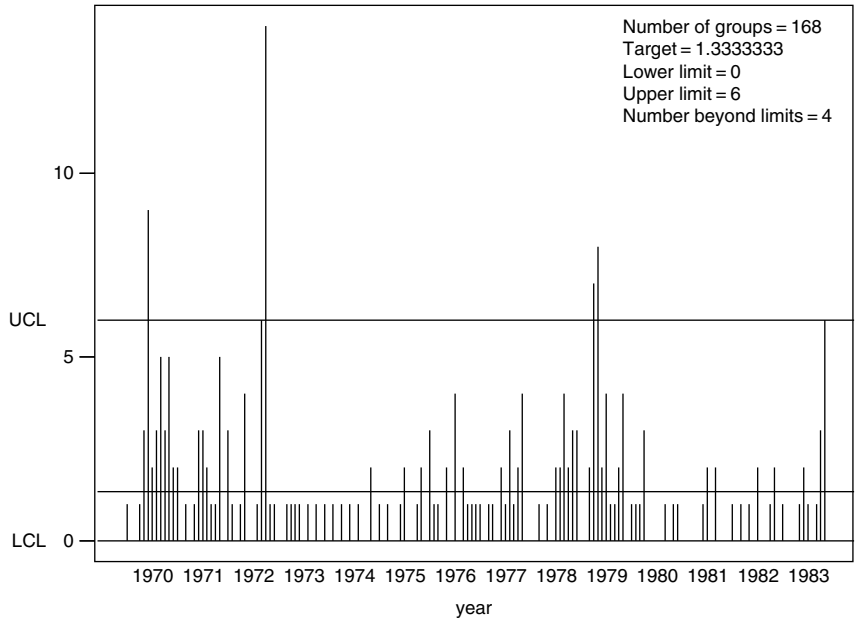


Figure 2.2 Monthly poliomyelitis reported cases in the USA between January 1970 and December 1983. Application of the Shewhart chart.

corresponds to an outbreak in Texas (1970). The second value corresponds to an outbreak in Connecticut (1972). The third and fourth values coincide with a third poliomyelitis outbreak reported in Pennsylvania, Wisconsin, Iowa, and Missouri (1979). For more details of these outbreaks, see Moore *et al.* (1982).

2.2.2.2 Moving average charts

An automated warning system was proposed by Stern and Lightfoot (1999), based on moving averages and applied to surveillance data of enteric pathogens. The statistic of the moving average (MA) control chart is given by:

$$y_t = \frac{1}{m} \sum_{k=0}^{m-1} x_{t-k},$$

where m is the number of past observations used in the moving average. A statistical aberration is identified when $|y_t|$ exceeds control limits. Similar to the Shewhart chart, control limits are a multiple of the standard deviation of y_t. This chart is more effective than the Shewhart chart in detecting small shifts in process level. As in the Shewhart chart, only the last observation is taken into account. In an MA chart, a sensible choice of m determines the suitability of

the chart. A suitable chart allows a good balance between the false positive rate (chart fails to indicate a shift in process level) and the false negative rate (chart indicates a non-real shift). For example, VanBrackle and Williamson (1999) used $m = 2$ in their study.

2.2.2.3 Exponentially weighted moving average control chart

The exponentially weighted moving average (EWMA) control chart (Hunter, 1986) gives less weight to data as they get older (less weight to more historical data, more weight to more recent data). The control statistic is defined by the following recursive equation:

$$y_t = (1 - \lambda)y_{t-1} + \lambda x_t,$$

where $0 < \lambda \leq 1$ is the EWMA weighting parameter and $y_0 = 0$. When the value of λ increases, the influence of the data in the more distant past decreases. A value for λ is usually chosen between 0.1 and 0.5, but this value can be chosen more or less subjectively. The EWMA control chart is less sensitive to the assumption of an underlying normal distribution, and Williamson and Hudson (1999) suggest that this method provides a more flexible tool than the Shewhart control chart for the monitoring of surveillance data.

2.2.2.4 The cumulative sum control chart

Cumulative sum (CUSUM) charts were introduced by Page (1954) and originally used in manufacturing processes to monitor production defect rates. They have been used by epidemiologists for the surveillance of congenital malformations (Gallus *et al.*, 1986), mortality due to respiratory diseases (Rossi *et al.*, 1999) or nosocomial clusters (Brown *et al.*, 2002). At the CDC, the CUSUM method is routinely applied to laboratory-based salmonella serotype data to detect salmonella outbreaks (Hutwagner *et al.*, 1997). Assuming that $X_t \sim N(\mu_t, \sigma_t^2)$, the control statistic is defined iteratively by:

$$y_t = \max\left(0,\ y_{t-1} + \left(\frac{x_t - \mu_t}{\sigma_t} - k\right)\right),$$

with $y_0 = 0$ and $k > 0$.

Extensive statistical literature, including review papers, exists on this topic. For more details about CUSUM and more generally about process control charts see, for example, Frisén (2003) and Chapter 3 (this volume). The average run length (ARL) is a standard measure used in quality control (Wetherill and Brown, 1990). It is the expected number of the surveillance time units

(e.g. weeks) before the chart indicates a shift in the process level. It can be used for the comparison of process control charts.

2.2.3 Time-series Analysis

Public health surveillance data are collected at regular intervals over time. Thus the surveillance data often exhibit correlation and seasonality. Adapted methods are required to take into account these specific features and to provide forecasts of future incidence values. A natural orientation is to consider the extensive literature concerning the Box–Jenkins (seasonal) autoregressive integrated moving average (ARIMA) models (Box and Jenkins, 1970). Box–Jenkins models have been used in many applications, including the analysis of surveillance data (Choi and Thacker, 1981; Helfenstein, 1986; Nobre *et al.*, 2001; Reis and Mandl, 2003; Schnell *et al.*, 1989; Stroup *et al.*, 1989; Watier *et al.*, 1991; Zaidi *et al.*, 1989). Forecasts estimate the expected incidence values, and these are compared with the most recently observed disease incidence value. Several steps are necessary:

(1) *Stationarity*. The time series must be stationary in terms of both mean and variance. A stochastic process $\{X_t\}$ is stationary if, for all t, the mean of the process is constant and the covariance between $\{X_t\}$ and $\{X_{t-1}\}$ depends only on the time lag k. If the time series has a nonconstant mean, traditional transformations are required to generate a stationary series from the nonstationary series. Time lag differencing is used when nonstationary means are encountered. Square root transformations are applied when variances depend on time.

(2) *Identification and estimation*. Identification of an adequate stochastic process to describe the observed time series is needed. The tools used for identification are the autocorrelation function (ACF), the partial autocorrelation function (PACF) and the inverse autocorrelation function (IACF). The PACF and IACF indicate the order (q) of the autoregressive part, while the ACF indicates the order (p) of the moving average part. When the orders of the process are determined, estimation of the parameters is performed by the maximization of a likelihood function.

(3) *Diagnostic checking*. Residuals, defined as the difference between the observed values and the model estimations, have to fulfill three conditions: (i) the mean of the residuals should not be significantly different from zero; (ii) the distribution of residuals should be normal; (iii) there should be no residual autocorrelation. The Kolmogorov–Smirnov test (see Daniel, 1995) and the Box–Ljung statistic (Ljung and Box, 1978) can be used respectively to verify the last two conditions. Once the residuals have been analyzed, the model can be used to forecast one-step-ahead values and their corresponding confidence limits. The forecasts are assumed to be normal in order

to calculate the 95% forecast interval defined by the forecast plus or minus the square root of the forcast variance. Box–Jenkins modeling can be carried out using classical statistical software. In addition, the Statistical Software for Public Health Surveillance (SSS1) developed by the CDC (Stroup *et al.*, 1994) provides several methods for analyzing surveillance data, including the Box–Jenkins method.

2.2.3.1 Combination of process control methods and Box–Jenkins models

Williamson and Hudson (1999) describe a combination of the Box–Jenkins models and statistical process control methods. In the first stage, an ARIMA model is developed as described in the previous section. In the second stage, the forecast errors assumed to be approximately independent and identically distributed are tracked in a statistical process control. Their two-stage monitoring system was performed on data from the NNDSS. The Shewhart, EWMA and MA charts were used. Several types of control charts are implemented in the CDC's statistical software to monitor the forecasting performance of ARIMA models.

2.2.3.2 Combination of wavelets and Box–Jenkins models

In a recent paper on the early detection of infectious disease outbreaks associated with bioterrorism (Goldenberg *et al.*, 2002), the authors present a set of tools for the analysis of time series. Their approach is original in the sense that they avoid public health data in favor of, for example, grocery and pharmacy data, school attendance records, and web sources. The idea is to detect infected people through their purchase of medication rather than from medical or public health sources. The timeliness of the surveillance may be superior if people pursue self-treatment before seeking medical assistance. From a statistical point of view, after several layers (denoising filter, decompositions by a discrete wavelet transform, simple autoregressive model applied to each decomposition), an upper threshold for the next day's forecast is computed, based on the addition of the forecast and an error. As in the traditional methods for detection, the system flags an alarm when the threshold is exceeded.

2.2.3.3 Integer-valued autoregressive processes

Integer-valued autoregressive (INAR) models represent a class of models for the analysis of time series. They have been studied theoretically by many authors (Al-Osh and Alzaid, 1987; Du and Li, 1991; Latour, 1997, 1998) and applied on time series of infectious disease incidence (Cardinal *et al.*, 1999). This class of models is an interesting alternative to the real-valued time-series models which

do not respect the nonnegative integer-valued characteristics of surveillance values. Real-valued models applied to nonnegative integer-valued observations may be an inappropriate strategy, especially for the analysis of rare events. An INAR process of order p is defined by:

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \varepsilon_t,$$

where $\{X_t\}$ is a nonnegative integer-valued stochastic process associated with the observed disease incidence time series. The Steutel and van Harn's convolution operator (Steutel and van Harn, 1979), denoted '$\circ$', is defined by:

$$\alpha \circ X_t = \sum_{k=1}^{X} Y_k,$$

where $\{Y_k; k \in \mathbb{N}\}$ is a sequence of identically and independently distributed random variables which follow a Bernoulli distribution with parameter α. If we consider an integer-valued autoregressive process of order 1, the first formula can be rewritten as

$$X_t = Y_1 + Y_2 + \cdots + Y_{X_{t-1}} + \varepsilon_t.$$

An epidemiologic interpretation of this formula is to consider that X_t is the prevalence of the disease at time t. The prevalence at time t is the sum of individuals remaining infected with a probability α in the time interval $(t-1, t)$ and individuals contracting the disease in the same interval (represented by ε_t). INAR models are identified using the same tools as for ARIMA models, that is, the ACF and the PACF. Autoregressive parameters are estimated using either the Yule–Walker estimation technique or the conditional least-squares method. Cardinal *et al.* (1999) concluded that an INAR model provides a smaller relative forecast error than ARIMA models for meningococcal disease.

2.2.3.4 Serfling's method

Serfling (1963) proposed a statistical analysis of weekly pneumonia and influenza deaths in 108 US cities. Based on this work, several authors have proposed a regression model which fits the nonepidemic data and predicts a nonepidemic level curve. Costagliola *et al.* (1991) applied Serfling's method to the French influenza-like syndrome data collected from a sentinel network from 1984 to 1988. They deleted the cases for the past epidemic periods, defined as periods above three cases per sentinel general practitioner (SGP). Then they fitted the following regression equation to forecast the expected nonepidemic level for the following winter:

$$y_t = \alpha + \beta t + \gamma_1 \cos\frac{2\pi t}{52} + \gamma_2 \sin\frac{2\pi t}{52} + \gamma_3 \cos\frac{4\pi t}{52} + \gamma_4 \sin\frac{4\pi t}{52} + \varepsilon_t,$$

where y_t is the number of cases per SGP in week t and ε_t follows a centered normal distribution. The parameters were estimated by the least-squares method.

The first main drawback of this approach is that one must define what the epidemic periods are, that is, at what number of cases per SGP we can consider that past observed data should be deleted when fitting the model. The second limitation is that the model imposes both a seasonal period and very specific terms in the regression equation. This means that the process under study must be relatively regular over time. Finally, this method cannot be easily applied to a wide range of time series exhibiting different features in terms of seasonality, number of cases, etc. However, despite the strong underlying hypotheses of this method, this approach represents a simple tool to analyze surveillance data for relatively well-known diseases. This is the case for the detection of epidemics of influenza-like syndromes or gastroenteritis (Flahault *et al.*, 1995).

Figure 2.3 illustrates Serfling's method applied to weekly number of *Salmonella paratyphi* B infections in France from 1992 to 1996. An aberration is defined when the number of cases exceeds, for two consecutive weeks, the expected number of cases represented by the upper 95% confidence limit of the expected value.

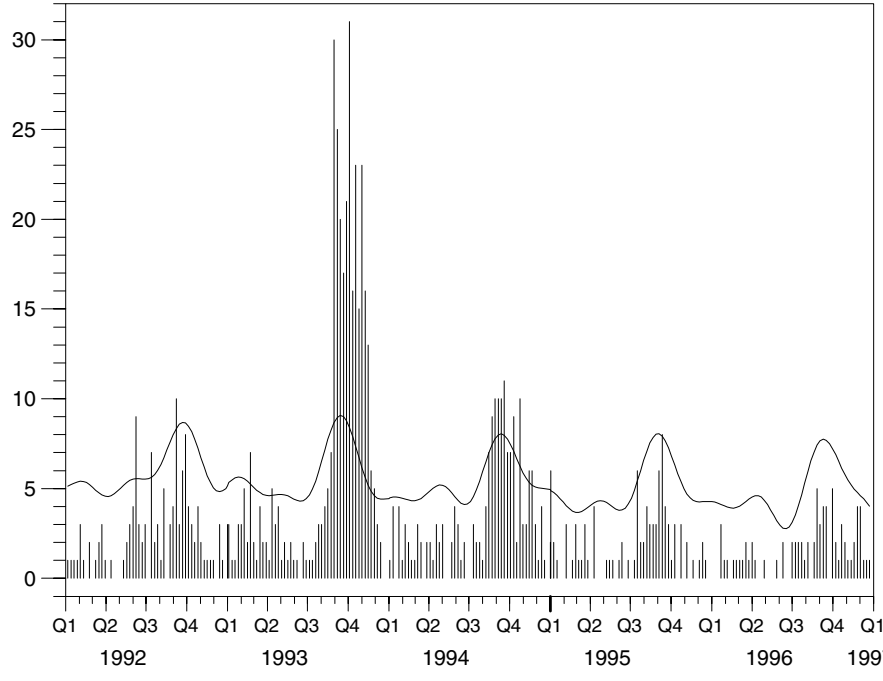


Figure 2.3 Weekly number of *Salmonella paratyphi* B infections in France from 1992 to 1996 with the Serfling upper 95% confidence limit.

2.2.3.5 Log-linear regression model

A regression model was developed by Farrington *et al.* (1996) and dedicated to the early detection of outbreaks from reports received at the Communicable Diseases Surveillance Centre (CDSC). The general formulation is:

$$\begin{aligned} \log(\mu_i) &= \alpha + \beta t_i, \\ E(X_i) &= \mu_i, \\ V(X_i) &= \phi\mu_i. \end{aligned}$$

The baseline count x_i, corresponding to baseline week t_i, is assumed to be distributed with mean μ_i and variance $\phi\mu_i$, where ϕ is the dispersion parameter. Estimates are obtained by a quasi-likelihood method. This model represents one of the most interesting tools for detection as it includes for the majority of data characteristics a statistical solution. Trends are incorporated into the regression by fitting a linear time variable (the first line in the above equation). Seasonality is handled, as in the historical limits method, by using only observations from comparable periods in the threshold calculation. Serial correlations between baseline counts are estimated and included in the threshold expression. The influence of baseline counts in time periods coinciding with past outbreaks is reduced by constructing weights based on adequate residuals (Davison and Snell, 1991). The idea is to associate low weights with large residuals, that is, high baseline counts. Finally, this log-linear regression, adjusted for overdispersion, is highly sensitive and detects small increases in rare disease reporting, as well as large excesses in common disease reporting. Since 1996, this method has been applied to the detection of aberrations for a set of 200–350 different types of organisms reported from laboratories. Each week, an exceedance score is given for each organism. If the exceedance score is higher than one, an alarm is triggered. A part of the algorithm output representing the monthly number of serologic tests for leptospirosis reported to the French Reference National Centre is illustrated in Figure 2.4.

2.2.3.6 Other parameter-driven models

Another Poisson log-linear regression model was developed but not applied to the detection of aberration (Zeger, 1988). In this parameter-driven model an underlying hidden (unobserved) stochastic process generates the dependence between random variables of the process of interest. This class of models represents an alternative to observation-driven models described in the previous sections. In observation-driven models X_t is a function of past observations $X_{t-1}, X_{t-2}, \ldots$. Among parameter-driven models, dynamic linear models, formalized by West and Harrison (1989) and methods based on the Kalman filter (Kalman, 1960) seem to be useful for forecasting time-series values.

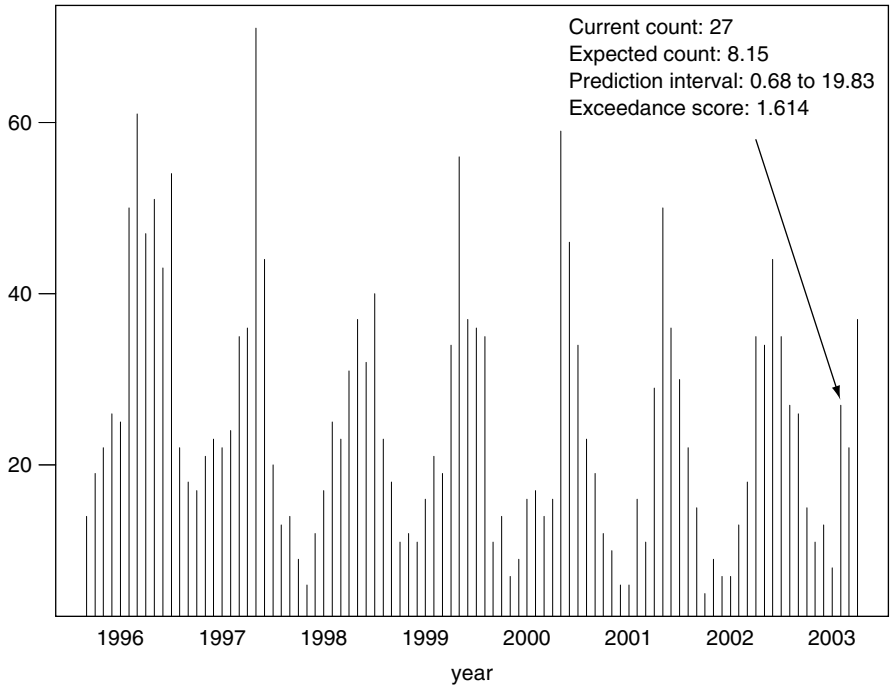


Figure 2.4 Monthly number of serologic tests for leptospirosis reported to the French Reference National Centre. Application of the CDSC method.

For example, the Kalman filter was applied to the monitoring of AIDS surveillance data (Stroup & Thacker, 1995). Other parameter-driven models called hidden Markov models (HMMs), have been applied to the monitoring of surveillance data (Le Strat and Carrat, 1999; Rath *et al.*, 2003) and the analysis of hospital infection data (Cooper & Lipsitch, 2004). The basic idea is to associate with each X_t an unobserved random variable S_t that determines the conditional distribution of X_t. Parameter estimations are obtained by the maximization of a likelihood function. The most likely sequence of states is reconstructed using a specific statistical method. Figure 2.5 gives an illustration of the reconstruction sequence of states by a two-state HMM applied to weekly influenza-like illness incidence rates in France, between 1984 and 2004. Figure 2.5(a) shows the weeks classified in one of the Markov states. This state is considered as the nonepidemic state. Figure 2.5(b) shows the weeks classified in the second Markov state and interpreted as epidemic weeks. Finally, Figure 2.5(c) represents the incidence for the totality of the weeks. HMMs provide a very flexible tool for the analysis of time series of discrete values. Trend, seasonality, and covariates can be easily introduced into the model and different distributions can be considered (normal, Poisson, etc.).

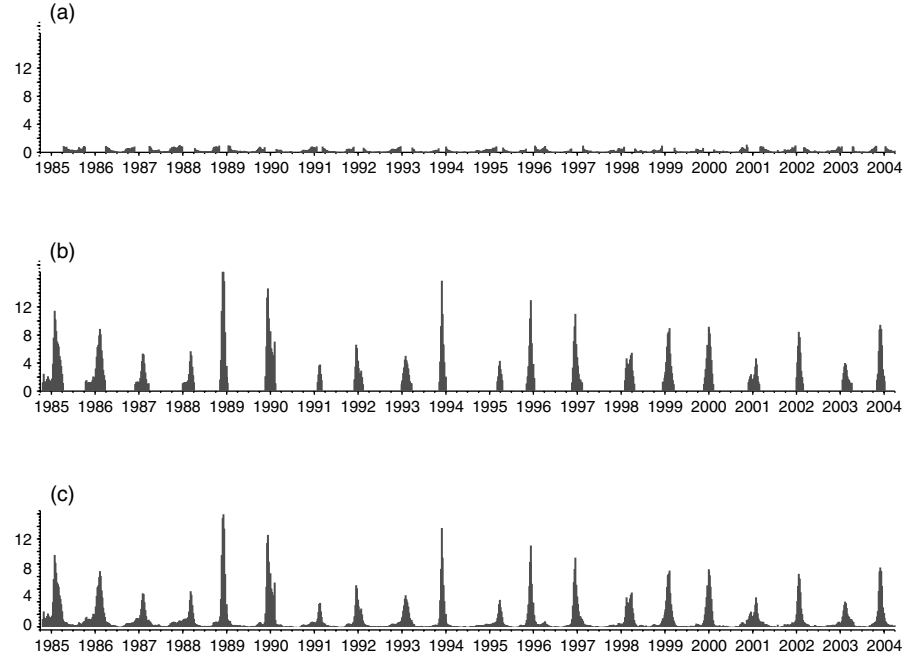


Figure 2.5 Weekly influenza-like illness incidence rates between 1984 and 2004. Application of a two-state hidden Markov model: (a) nonepidemic state; (b) epidemic state; (c) both states combined.

2.3 CONCLUSION

The goals of surveillance data analysis are to answer the following classical questions: Is there a trend and/or a seasonality and/or abrupt change in the observed incidence/prevalence time series of a disease? It is much more challenging to carry out prospective surveillance and try to forecast future values. The main objective, in this kind of surveillance, is the early detection of aberrations, which has become very important following recent bioterrorism threats. The detection of unusual events involves a combination of a forecasting method and a decision mechanism which permits a decision to be taken as to whether the observed value is significantly different from the forecast value. The decision mechanisms routinely used in surveillance centers are relatively similar, but forecast methods can have well-marked differences. Their capacity to detect unusual events with good sensitivity and specificity depends on their ability to take into account trend, seasonality, correlations between random variables of the stochastic process, and the disease amplitude in terms of number of cases or deaths. In addition, the most refined methods allow the weighting of each observation used in the analyses in order to lessen the influence of past observations

that correspond to past outbreaks. In this sense, the log-linear regression model (Farrington *et al.*, 1996) seems to be a more elaborate approach.

A real comparison between all of these models based on a calculation of the sensitivity and specificity and using different time surveillance series should be performed. However, the conclusion is likely to be that there is no unique method which can be applied to all surveillance series of a disease. Each disease and each surveillance system has its own unique characteristics. Detection is strongly reliant not so much on the statistical methods as on the characteristics of the system such as the data collected, the professionalism of the participants in the surveillance system, the reporting mechanism and its stability over time, and its reactivity. If the surveillance system itself is not of good quality, it is easy to produce spurious statistical results. It is, therefore, crucial to be familiar with the structure of the surveillance system before analyzing the systems data. If the system is viable and the statistical detection method rigorous, it is possible for an epidemiologist to use the model results as a statistical aid. The decision to investigate should be easier when the experience of the epidemiologist is coupled with a statistical result.

Even if the statistical detection of temporal unusual events is useful, the spatial component of disease distribution is generally not taken into account. For a given time period, when the observed number of cases is significantly different from the expected number, it is natural to look at the spatial distribution of the cases to make observations concerning the existence or absence of clusters. As in the detection of temporal aberrations, a spatial statistical method can sharpen the epidemiologist's judgment and corroborate (or not) the occurrence of a temporal unusual event that indicates the need for an epidemiologic investigation.

In conclusion, a vast literature on time detection methods exists and several applications have been described. However, improvements are still needed in order to suitably incorporate surveillance system features such as reporting delays, the time evolution of 'denominators' (i.e. the population size and the number of serological tests). Similarly, the inclusion of spatial information, available in almost all surveillance databases, seems to be essential for the improvement of the early detection of unusual events.

3

Optimal Surveillance

Marianne Frisén and Christian Sonesson

3.1 INTRODUCTION

In public health the timely detection of various types of adverse health events is an important issue. Kaufmann *et al.* (1997) stated that a delay of one day in the detection of and response to an epidemic due to a bioterrorist attack could result in a loss of thousands of lives and millions of dollars. Thus optimality is important. Public health surveillance is described by Thacker and Berkelman (1988) as the ongoing systematic collection, analysis, and interpretation of outcome-specific data essential to the planning, implementation, and evaluation of public health programs, closely integrated with the timely dissemination of these data to those responsible for prevention and control. Källén and Winberg (1969) and Hill *et al.* (1968) discussed the importance of detection of an increased birth rate of babies with congenital malformations. This was especially apparent during the thalidomide tragedy in the early 1960s. Monitoring of mortality rates in primary care is treated by Aylin *et al.* (2003). Other examples include the detection of bioterrorism, outbreaks of infectious diseases and the spatial clustering of various forms of cancer. In all of these examples quick detection is beneficial both at an individual level and to society. Examples of different public health surveillance data sources are given by Stroup *et al.* (2004).

For spatial surveillance of public health there are many important issues, such as data collection and data quality, to consider. In Chapter 2 surveillance systems are classified in different ways, for example according to how the information is collected. Here, the focus will be on issues of statistical inference. There is a need for continual observation of time series, with the goal of detecting an important change in the underlying process as soon as possible after it has occurred. Statistical methods are necessary to separate important changes in the process from stochastic variation. Broad surveys and bibliographies on statistical

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

surveillance are given by Lai (1995), who concentrates on minimax properties of stopping rules, by Woodall and Montgomery (1999), who concentrate on control charts, and by Frisén (2003), who concentrates on optimality properties of methods. A review of methods for surveillance in public health is given by Sonesson and Bock (2003). The statistical methods suitable for this differ from the standard hypothesis testing methods. Surveillance, statistical process control, monitoring, and change-point detection are different names for methods with this goal. Also the criteria for optimality differ. In Section 3.2 the important difference between optimality for a fixed sample and optimality for on-line surveillance will be discussed.

In Section 3.3 the notation and specification used in the chapter are described. Most of the theory for surveillance is derived for normal distributions, but Poisson processes are of special interest in public health surveillance. A bibliography of surveillance for attribute data is given by Woodall (1997).

In spatial surveillance of public health, evaluations and optimality are very important in order to choose which surveillance method to use (and parameters in the method) for the specific aim. The requirements are different for short-term, high-risk and long-term, low-risk situations. In applied work a single optimality criterion is not always enough, but evaluations of different properties might be necessary (Frisén, 1992). These properties are also the base for the formal optimality criteria. In Section 3.4 we will describe some measures for evaluations of surveillance methods.

To choose the optimal method you have to specify what 'optimal' means in a surveillance context. Optimality plays an important role both in applied work and in theoretical research. There are many papers which claim to give the optimal surveillance method. However, the suggested optimality criteria differ in important aspects described by Frisén (2003). In Section 3.5 some general criteria of optimality are described, which are based on the expected delay, the minimax principle, and the average run length.

Most of the commonly used methods are optimal in some respect. Some commonly used methods are described in Section 3.6. The correspondences between the criteria of optimality and methods are examined. The situations and parameter values for which some commonly used methods have optimality properties are thus determined. Thus, the commonly used methods are characterized by their optimality properties. One of the methods described is the full likelihood ratio (LR) method. The LR method corresponds to the use of the posterior distribution and fulfils important optimality criteria. This method, which relies on generally accepted principles of inference, can then be used as a benchmark for the other methods discussed.

Cardinal *et al.* (1999) describe the problems of using methods for continuous variables when studying the incidence of a disease which is based on count data. Methods and evaluation for distributions of special interest for public health studies are treated throughout the chapter. However, we treat some subjects separately. In Section 3.7 we describe methods for some more complicated situations of special interest for public health. A discussion of methods

and optimality is given for gradual changes from an unknown baseline as well as spatial and other multivariate surveillance situations. Section 3.8 contains some concluding remarks.

3.2 OPTIMALITY FOR A FIXED SAMPLE AND FOR ON-LINE SURVEILLANCE

In the comparison of disease patterns in different regions many questions can be answered by hypothesis tests based on a fixed sample of data. For reviews, see Lawson *et al.* (1999) and Lawson and Cressie (2000). In the prospective surveillance situation repeated analysis of data accumulating over time is used. Then, there is no fixed data set and not even a fixed hypothesis to be tested. A decision concerning whether, for example, an incidence has increased or not has to be made sequentially, based on the data collected so far. The statistics derived for a fixed sample might be of great value also in the surveillance case, but there are great differences concerning the system for decisions. In complicated surveillance problems a stepwise reduction of the problem might be useful. Then, the statistics derived to be optimal for the fixed sample problem can be a component in the construction of the prospective surveillance system. How this can be done is described in the Chapter 9.

Error rates suitable for a fixed decision time can be used as components in evaluation measures for on-line surveillance. Different error rates and their implications for a system of decisions were discussed by Frisén and de Maré (1991). The maximal detection probability for a fixed false alarm probability for each decision time is a simple criterion. The LR method of Section 3.6.1 satisfies this criterion. Using a constant probability of exceeding the alarm limit for each decision time means that we have a system of repeated significance tests. This might work well also as a system of surveillance and is often used. The Shewhart method described in Section 3.6.2 has this property. This is also the motivation for using the limits with the exact variance in the exponentially weighted moving average (EWMA) method described in Section 3.6.5. However, the probability of exceeding the alarm limit conditional on no earlier alarm is not constant for this type of EWMA method.

Evaluation by the significance level, power, specificity, and sensitivity which is useful for a fixed sample is not appropriate in a surveillance situation without a modification since they have no unique value unless the time period is fixed. Also, a formulation of an optimality criterion for surveillance must naturally take into account the delay time in detection, since the aim of a surveillance method is quick detection.

There are close relations between the methods for a fixed sample and for on-line surveillance. However, both the methods and the optimality criteria suitable for on-line surveillance differ from the standard hypothesis testing situation. The choice of an optimality criterion in on-line surveillance is an interesting and important issue which will be treated throughout the chapter.

3.3 SPECIFICATION OF THE STATISTICAL SURVEILLANCE PROBLEM

We will specify the situation with a change in distribution at a certain change-point time τ. The variable under surveillance could be an age-adjusted incidence or some other derived statistic depending on the specific situation. We denote the process by $Y = \{Y(t) : t = 1, 2, \dots\}$, where $Y(t)$ is the observation made at time t. The random process that determines the state of the system is denoted by $\mu(t)$. At each decision time, s, we wish to discriminate between two states of the monitored system, the in-control and the out-of-control state, here denoted by $D(s)$ and $C(s)$, respectively. To do this we use the accumulated observations $Y_s = \{Y(t);\ t \le s\}$ to form an alarm criterion such that if this is fulfilled it is an indication that the process is in state $C(s)$ and an alarm is triggered. Usually this is done by using an alarm statistic, $p(Y_s)$, and a control limit, $G(s)$, where the time of an alarm, t_A, is

$$t_A = \min\{s;\ p(Y_s) > G(s)\}.$$

Different types of in-control and out-of-control states are of interest depending on the application. The most frequently studied case is when $D(s) = \{\tau > s\}$ and $C(s) = \{\tau < s\}$; see Figure 3.1. The time τ of the change is regarded as a random variable with probabilities $\pi(t) = P(\tau = t)$. These probabilities can also be regarded as priors. The intensity, $\nu(t)$, of a change is defined as $\nu(t) = P(\tau = t|\tau \ge t)$, which is usually assumed to be constant over time.

The change to be detected also differs depending on the application. Most studies in literature concerns a step change, where a parameter changes from

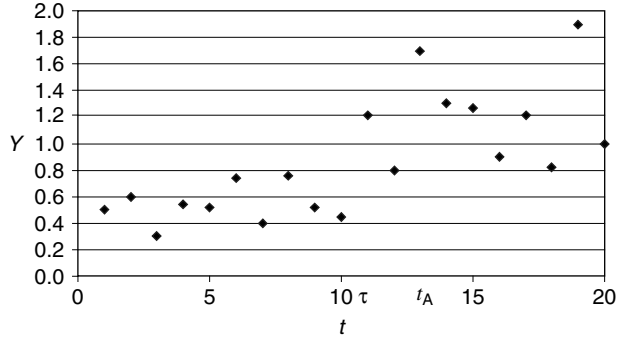


Figure 3.1 Illustration of concepts in evaluation. The first $\tau - 1 = 10$ observations $Y_{\tau-1} = \{Y(t); t \le \tau - 1\}$ are in state D. The subsequent observations (from $t = 11$ onwards) are in state C with a higher mean. The alarm time is t_A, which might happen to be 13, in which case the delay would be $t_A - \tau = 2$.

one constant level to another constant level. Often, the case of a shift in the mean of a normally distributed variable from an acceptable value μ^0 (say, zero) to an unacceptable value μ^1 is considered, but with the same known standard deviation σ. For clarity, when suitable, standardization to $\mu^0 = 0$ and $\sigma = 1$ is used and the size of the shift after standardization is denoted by μ. The case $\mu > 0$ is described here. The case $\mu < 0$ is treated in the same way. We have $\mu(t) = \mu^0$ for $t = 1, \ldots, \tau - 1$ and $\mu(t) = \mu^1$ for $t = \tau, \tau + 1, \ldots$. Even though autocorrelated time series are studied by, for example, Schmid and Schöne (1997), Petzold *et al.* (2004), and in Chapter 2, processes which are independent given τ are the cases most often studied. This is a simple situation which we will use to describe general concepts of evaluations, optimality and standard methods. A sudden sharp increase might be realistic in the case of bioterrorist attack. Methods optimal for detection of a step change might be good also for gradual change if the change rate is high. However, other types of changes are also of interest. Some cases of special interest in the surveillance of public health are discussed in Section 3.7.

For the detection of an increased incidence rate different assumptions concerning the underlying process can be made depending on the setting and the data collected. Often a Poisson process for the cases of disease is assumed. In some cases the intervals between the adverse events have been of interest. These intervals can be measured by the continuous time intervals between the events, which are exponentially distributed, or by using a discrete time scale measuring the number of acceptable events between adverse events. Neither of these ways implies any loss of information about the process. The increased intensity would then be recognized as shorter intervals between the adverse events and fewer acceptable events between adverse events, respectively. Sometimes only the numbers of events in certain fixed time windows are available. These numbers are usually assumed to be Poisson distributed. A normal approximation is frequently used. There are also other situations where the normal distribution is a natural assumption.

3.4 EVALUATIONS OF SYSTEMS FOR SURVEILLANCE

We need some measures for evaluation as a basis for the formal optimality criteria of the next section. Good properties are quick detection and few false alarms. When monitoring is used in practice, knowledge about the properties of the method is important. If an alarm is triggered it is otherwise hard to know how strong an indication this is of a change. In applied work a single optimality criterion is not always enough, and evaluation by several measures might be necessary. In this chapter we will discuss the measures given in Table 3.1. Computer illustrations of the interpretation of some of the measures mentioned below are given in Frisén and Gottlow (2003). Formulae for the numerical approximations of some of the measures are available in the literature.

Table 3.1 Measures given in Sections 3.4.1–3.4.4 classified by whether they are adopted for ongoing surveillance or not.

	Conventional measures	Special measures for ongoing surveillance
False alarms	Size α, Specificity	ARL^0, MRL^0, PFA
Detection ability	Power, Sensitivity	ARL^1, MRL^1, CED, ED, maxCED
Predictive value	PVP, PVN	$PV(t) = P(t_A \leq \tau | t_A = t)$

3.4.1 Measures for a Fixed Sample Situation Adopted for Surveillance

In the draft guidelines given by the Centers for Disease Control and Prevention (CDC) for evaluating surveillance systems (Sosin, 2003), timeliness is mentioned as one very important aspect when evaluating a surveillance system. Some measures of evaluation are stated, such as the sensitivity and the predicted value. German (2000) gives a review of the use of such measures in public health surveillance systems. Guidelines for evaluations of syndromic surveillance are given by Mandl *et al.* (2004).

One problem with evaluation measures originally suggested for the study of a fixed sample of, say, n observations is that the measures depend on n. The specificity will for most methods tend to zero and the size of the test tend to one as n increases, as shown in Figure 3.2.

Chu (1995) and others have suggested methods with a size less than one:

$$\lim_{n\to\infty} P(t_A \leq n | D) < 1.$$

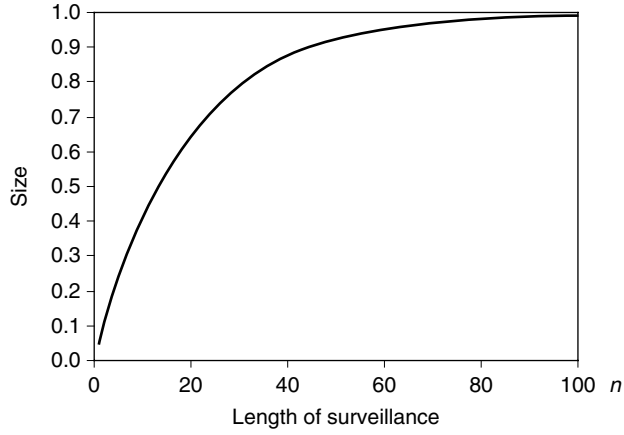


Figure 3.2 The size, α, of a surveillance system which is pursued for n time units, when the probability of a false alarm is 5% at each time point.

However, Frisén (2003) proved that the detection ability for methods with this property declines rapidly with increasing values of time τ of the change.

The performance of a method for surveillance depends on the time τ of the change. The sensitivity will in general not be the same for early changes as for late ones. It also depends on the length of time for which the evaluation is made. Thus there is not one unique sensitivity value at surveillance, but other measures might be more useful. Thus, conventional measures for fixed samples should be supplemented by other measures designed for statistical surveillance, as will be discussed in the following.

3.4.2 False Alarms

The erroneous false alarm is more complicated to control in surveillance than in hypothesis testing. There are special measures of the false alarm properties which are suitable for surveillance. The most commonly used measure is the average run length when there is no change in the system under surveillance, $\mathrm{ARL}^0 = \mathrm{E}(t_\mathrm{A}|D)$. This measure is closely related to the 'recurrence interval' and 'number of alarms per month' discussed in Chapter 5. In Chapter 5 the average time to a false alarm, if the alarm limit was the observed value, is calculated as a measure of how extreme a result is. The relation between ARL^0 and the recurrence interval is thus the same as that between the significance level and the p-value in an ordinary hypothesis test. Numerical methods to calculate ARL^0 are discussed in Chapter 6. A variant of the ARL is the median run length (MRL).

Another kind of measure commonly used is the false alarm probability, $\mathrm{PFA} = P(t_\mathrm{A} < \tau)$. This is the probability that the alarm occurs before the change. In theoretical work, the standard procedure is to assume that τ is geometrically distributed, implying a constant intensity of a change.

3.4.3 Delay of the Alarm

The delay time in detection should be as small as possible. Shiryaev (1963) suggested measures of the expected value of the delay. Let the expected delay from the time of change, $\tau = t$, to the time of alarm, t_A, be denoted by

$$\mathrm{ED}(t) = \mathrm{E}[\max(0, t_\mathrm{A} - t)|\tau = t].$$

$\mathrm{ED}(t)$ will typically tend to zero as t increases. The conditional expected delay,

$$\mathrm{CED}(t) = \mathrm{E}[t_\mathrm{A} - \tau | t_\mathrm{A} \geq \tau = t] = \mathrm{ED}(t)/P(t_\mathrm{A} \geq t),$$

is the expected delay for a specific change point τ. The expected delay is generally not the same for early as for late changes. CED will for most methods converge

to a constant value. This value is sometimes named the 'steady state ARL' (Srivastava and Wu, 1993). The summarized expected delay is

$$\mathrm{ED} = \mathrm{E}[\mathrm{ED}(\tau)],$$

where the expectation is with respect to the distribution of τ. To minimize the expected delay of detection is important in most practical situations. In the literature on signal detection (see Gustavsson, 2000) the mean time to detection is used. This is defined as

$$\mathrm{MTD}(t) = \mathrm{E}[t_\mathrm{A} - t | \tau = t].$$

This measure is related to $\mathrm{CED}(t)$ and $\mathrm{ED}(t)$ but the expected value, with respect to the distribution of τ, differs from ED and there is no simple relation to the ED optimality described in Section 3.5.1.

The most commonly used measure of the delay is the average run length until detection of a true change (that occurred at the same time as the surveillance started) which is denoted ARL^1 (see Page, 1954; Ryan, 2000). The part of the definition in the parentheses is seldom spelled out, but is generally used in the literature. Note that

$$\mathrm{ARL}^1 = \mathrm{ED}(1) + 1.$$

For some situations and methods the properties are roughly the same, regardless of when the change occurs, but this is not always true, as illustrated by Frisén and Wessman (1999). The run length distributions are often very skewed, and the skewness depends on important parameters. Instead of the average, Gan (1993) advocates that the median run length should be used on the grounds that it might be more easily interpreted. However, the main problem is that only the case $\tau = 1$ is considered.

Sometimes there is a limited time available for rescuing action. The probability of successful detection, suggested by Frisén (1992), measures the probability of detection with a delay time no longer than d:

$$\mathrm{PSD}(d, t) = \mathrm{P}(t_\mathrm{A} - \tau < d | t_\mathrm{A} \geq \tau = t).$$

This measure is a function both of the time of the change and the length of the interval in which the detection is defined as successful. It has been used by, for example, Petzold *et al.* (2004) in connection with a monitoring system for pregnancies. Also when there is no absolute limit for the detection time it is often useful to describe the ability to detect the change within a certain time. In those cases it might be useful to calculate PSD for different time limits d. This has been done by Marshall *et al.* (2004) in connection with monitoring of health care quality. The ability for very quick detection (small d) is important for surveillance of sudden major changes, while the long-term detection ability (large d) is more important for ongoing surveillance where smaller changes are expected.

3.4.4 Predictive Value

In order to know which action is appropriate at an alarm we need to know if we should act as if we were sure about a change or just have a vague suspicion. For a diagnostic test based on a fixed data set, it is common to use the predictive values of a positive diagnosis (PVP) and a of a negative diagnosis (PVN). For ongoing surveillance we have corresponding measures. The probability that a change has occurred when the surveillance method signals was suggested by Frisén (1992) as

$$\mathrm{PV}(t) = P(t_A \leq \tau | t_A = t).$$

When you get an alarm ($t_A = t$), PV tells you whether there is a large probability that the change has occurred ($t_A \leq \tau$). Some methods have a constant PV. Others have a low PV at early times but better later. In those cases, the early alarms will not motivate the same serious action as later alarms. Also the predictive value of the lack of an alarm at a certain time point might sometimes be of interest.

3.5 OPTIMALITY CRITERIA

We will now use the measures of the previous section to formulate and discuss some criteria of optimality for surveillance.

3.5.1 Minimal Expected Delay

Shiryaev (1963) suggested a very general utility function where the expected delay of a desired alarm plays an important role. He treated the case of constant intensity of a change where the gain of an alarm is a linear function of the value of the delay, $t_A - \tau$. The loss associated with a false alarm is a function of the same difference. This utility can be expressed as $U = \mathrm{E}\{u(\tau, t_A)\}$, where

$$u(\tau, t_A) = \begin{cases} h(t_A - \tau), & \text{if } t_A < \tau, \\ a_1(t_A - \tau) + a_2, & \text{otherwise.} \end{cases}$$

The function $h(t_A - \tau)$ is usually a constant (say, b), since the false alarm causes the same cost of alerts and investigations, irrespective of how early the false alarm was given. In this case we have

$$U = bP(t_A < \tau) + a_1 \,\mathrm{ED} + a_2.$$

Thus, we would have a maximal utility if we have a minimal (a_1 is typically negative) expected delay from the change-point for a fixed probability of a false

alarm (see Section 3.4.3). This is termed the ED criterion. The full likelihood ratio method satisfies this criterion. The ED criterion seems to be a suitable optimality criterion in a public health setting because of its generality of including changes occurring at different time points.

Variants of the utility function leading to different optimal weighting of the observations are suggested by, for example, Poor (1998) and Beibel (2000).

3.5.2 Minimax Optimality

The next criterion concerns the minimax of the expected delay after a change. It is related to the ED criterion as several possible change times are considered. However, instead of an expected value, which requires a distribution of the time of change, the worst value of CED(t) is used.

Moustakides (1986) uses a still more pessimistic criterion , the 'worst possible case', by using not only the worst value of the change time, but also the worst possible outcome of $Y_{\tau-1}$ before the change occurs. This criterion is pessimistic since it is based on the worst possible circumstances. The cumulative sum (CUSUM) method, described in Section 3.6.3, provides a solution to the criterion proposed by Moustakides. The merits of studies of this criterion have been thoroughly discussed by Yashchin (1993) and Lai (1995). Much theoretical research is based on this criterion.

3.5.3 Average Run Length

In the literature on statistical process control, optimality is often stated as minimal ARL^1 for a fixed ARL^0. ARL^1 and ARL^0 are expectations under the assumption that there are equal distributions for all observations under each of the two alternatives. Statistical inference with the aim of discriminating between the alternatives that all observations come from one of two specified distributions should, by the ancillarity principle, not be based on the time of the observation. However, the ARL criterion does not necessarily have to agree with generally accepted principles of inference. From the point of view of optimal decision theory it is hard to motivate a cost function with no cost for a delay of the alarm when $\tau > 1$.

As pointed out by Pollak and Siegmund (1985), the maximal value of CED(t) is equal to CED(1) for many methods, and with a minimax perspective this can be a motivation for the use of ARL^1 since $\text{CED}(1) = \text{ARL}^1 - 1$. However, this argument is not relevant for all methods. In particular it is demonstrated by Sonesson (2003) that the maximal value is not CED(1) for the EWMA method of Section 3.6.5. For this method, there is no similarity between the solution to the ARL criterion and the minimax criterion, while the solutions to the criterion of expected delay and the minimax criterion demonstate good agreement.

The widespread use of the ARL criterion has been questioned. Consequences of this criterion which make it unsuitable for many applications were demonstrated by Frisén (2003). Methods useless in practice are ARL optimal. Thus this optimality should only be used with care. The ARL can be used as a descriptive measure and gives a rough impression but is questionable as a formal optimality criterion.

3.6 OPTIMALITY OF SOME STANDARD METHODS

The optimality of some important methods will now be described. Some methods are very flexible with several parameters. The parameters can be chosen to make the method optimal for the specific conditions (e.g. the size of the change or the intensity of changes) of the application. Many methods for surveillance are based in one way or another on likelihood ratios. Thus, we will start by describing the likelihood ratio method as it is a benchmark for other methods. Commonly used methods are compared with it in order to clarify their optimality properties.

3.6.1 The Likelihood Ratio Method

The likelihood ratio (LR) method, is optimal with respect to the criterion of minimal expected delay and also a wider class of utility functions (Frisén and de Maré, 1991). Several methods can be described by approximations or combinations of likelihood ratios (Frisén, 2003). However, the LR method, with its relation to the posterior probability, has a special motivation. The full likelihood is a weighted sum of the partial likelihoods

$$L(s,t) = f_{Y_s}(y_s|\tau = t)/f_{Y_s}(y_s|D(s)).$$

The alarm set consists of those Y_s for which the full likelihood ratio exceeds a limit. When the event to be detected at decision time s is $C(s) = \{\tau < s\}$, with the alternative $D(s) = \{\tau > s\}$, the time of an alarm for the LR method is

$$\begin{aligned} t_A &= \min\left\{ s; \frac{f_{Y_s}(y_s|C(s))}{f_{Y_s}(y_s|D(s))} > \frac{P(\tau > s)}{P(\tau \le s)} \cdot \frac{K}{1-K} \right\} \\ &= \min\left\{ s; \sum_{t=1}^{s} w(s,t) \cdot L(s,t) > G(s) \right\}, \end{aligned}$$

where K is a constant and $G(s)$ is an alarm limit. The time of an alarm can equivalently be written as the first time the posterior probability of a change into state C exceeds a fixed level

$$t_A = \min\{s; P(C(s)|Y_s = y_s) > K\}.$$

The posterior probability of a change has been suggested as an alarm criterion by, for example, Smith and West (1983). When there are only two states, C and D, this criterion leads to the LR method (Frisén and de Maré, 1991). In cases where several changes may follow after each other, the process might be characterized as a hidden Markov chain and the posterior probability for a certain state determined (Harrison and Stevens, 1976; Hamilton, 1989). Le Strat and Carrat (1999) use hidden Markov models with two states (outbreak or not) to retrospectively estimate the times for outbreaks of influenza-like illness and poliomyelitis and found that the times agreed well with other judgments. They also suggested that the method should be used in public health monitoring. Sometimes the use of the posterior distribution, or equivalently the likelihood ratio, is termed 'the Bayes method'. However, it depends on the situation whether the distribution of τ is considered as a 'prior', or as an observed frequency distribution (e.g. from earlier cases in intensive medical care) or just reflects the situation for which optimality is desired. When the intensity, ν, of a change (see Section 3.3) tends to zero, the weights $w(s,t)$ of the partial likelihoods do not depend on t and the limit $G(s)$ of the LR method does not depend on s. Shiryaev (1963) and Roberts (1966) suggested the method, which is now called the Shiryaev–Roberts method, for which an alarm is triggered at the first time s, such that

$$\sum_{t=1}^{s} L(s,t) > G,$$

where G is a constant, which is thus the limit of the LR method as ν tends to zero. The Shiryaev–Roberts method can also be derived as the LR method with a noninformative prior for the distribution of τ. Both the LR and the Shiryaev–Roberts method can be expressed recursively. A valuable property of the methods is an approximately constant predictive value (Frisén and Wessman, 1999), which allows the same interpretation of early and late alarms.

For the case of a normal distribution specified in Section 3.3, the LR method is optimized for the values of the change size μ and the change intensity, ν used in the alarm statistic, and gives an alarm at

$$t_A = \min\left\{ s; \sum_{t=1}^{s} P(\tau = t) \exp\{t\mu^2/2\} \exp\{\mu \sum_{u=t}^{s} Y(u)\} > \exp\{(s+1)\mu^2/2\} P(\tau > s) \frac{K}{1-K} \right\},$$

where the constant K determines the false alarm probability.

Stroup and Thacker (1993) discuss the use of likelihood ratios and posterior probabilities to detect aberrations of public health data. They apply LR statistics to public health surveillance data collected in the USA to detect sudden,

sustained changes in reported disease occurrence, changes in the rate of change of health event occurrence, as well as unusual reports or outliers. For the detection of a changed intensity in a Poisson process. Sonesson and Bock (2003) derived the LR method based on the exponentially distributed time intervals between events. The stopping time is

$$t_A = \min\left\{s; \sum_{t=1}^{s} P(\tau = t) \cdot \exp\left\{(-\lambda_1 + \lambda_0)\sum_{i=t}^{s} Y(i)\right\} \cdot \left(\frac{\lambda_1}{\lambda_0}\right)^{s-t+1} > L \cdot P(\tau > s)\right\},$$

for some constant L. Here λ_0 denotes the baseline intensity and λ_1 the intensity after the change. For the same situation, the Shiryaev–Roberts method can be found in Kenett and Pollak (1996) and gives an alarm at

$$t_A = \min\left\{s; \sum_{t=1}^{s} \exp\left\{(-\lambda_1 + \lambda_0)\sum_{i=t}^{s} Y(i)\right\} \cdot \left(\frac{\lambda_1}{\lambda_0}\right)^{s-t+1} > G\right\}.$$

In the case where the observed data consist of the number of events recorded in fixed intervals of length k, the alarm time for the LR method is derived by Sonesson and Bock (2003) to be

$$t_A = \min\left\{s; \sum_{t=1}^{s} P(\tau = t) \exp\{(-\lambda_1 + \lambda_0)k(s-t+1)\} \cdot \left(\frac{\lambda_1}{\lambda_0}\right)^{\sum_{i=t}^{s} Y(i)} > L \cdot P(\tau > s)\right\},$$

and the alarm time for the Shiryaev–Roberts method is

$$t_A = \min\left\{s; \exp\{(-\lambda_1 + \lambda_0)k(s-t+1)\} \cdot \sum_{t=1}^{s} \left(\frac{\lambda_1}{\lambda_0}\right)^{\sum_{i=t}^{s} Y(i)} > G\right\}.$$

If the counts are recorded for intervals of different lengths, a slight modification has to be made.

Linear approximations of the LR method are of interest for two reasons. One is to obtain a method which is easier to use and analyze, but has good properties similar to the LR method. Another is to get a tool for the analysis of approximate optimality of other methods, as was done by Frisén (2003) and will be seen in subsequent sections.

3.6.2 The Shewhart Method

The Shewhart method, which was suggested by Shewhart (1931), is simple and is certainly the most commonly used method for surveillance. It can be regarded as performing repeated significance tests. An alarm is triggered as soon as an

observation deviates too much from the target. Thus, only the last observation is considered in the Shewhart method. An alarm is triggered at

$$t_A = \min\{s;\, Y(s) > L\},$$

where L is a constant. More detailed descriptions can be found in textbooks such as Ryan (2000). The alarm criterion for independent observations can be expressed by the condition $L(s, s) > G$, where G is a constant. The alarm statistic of the LR method reduces to that of the Shewhart method when the event to be detected at decision time s is $C(s) = \{\tau = s\}$ and the alternative is $D(s) = \{\tau > s\}$. Thus, the Shewhart method has optimal error probabilities for these alternatives for each decision time s. For large shifts, Frisén and Wessman (1999) showed that the LR method and the CUSUM method converge to the Shewhart method. By several criteria, the Shewhart method performs poorly for small and moderate shifts. However, by the minimax criterion it is nearly as good as the LR method for some situations.

Woodall (1997) gave a review of the Shewhart method as applied to attribute data. The observation is often a proportion or number of objects with a certain property (e.g. having a certain disease). The p-chart, np-chart, c-chart, and u-chart are all based on a normal approximation of the binomial distribution but differ depending on whether numbers or proportions (possibly with varying sizes of the populations at risk) are used. Shewhart methods have also been suggested for geometrical, negative binomial, and Poisson data.

3.6.3 The CUSUM Method

The CUSUM method, first suggested by Page (1954), is closely related to the minimax criterion. Yashchin (1993), Siegmund and Venkatraman (1995), Hawkins and Olwell (1998), and Chapter 6 of this book give reviews of the CUSUM method. The alarm condition of the method can be expressed by the partial likelihood ratios as

$$t_A = \min\{s;\, \max(L(s, t);\, t = 1, 2, \ldots, s) > G\},$$

where G is a constant. The method is sometimes called *the* likelihood ratio method, but this combination of likelihood ratios should not be confused with the full LR method.

The most commonly described application of the CUSUM method is to the case of normally distributed variables as specified in Section 3.3. The CUSUM statistic in this case reduces to a function of the cumulative sums

$$C_r = \sum_{t=1}^{r} (Y(t) - \mu(t)).$$

There is an alarm for the first time s for which

$$C_s - C_{s-i} > h + ki, \qquad \text{for some } i = 1, 2, \ldots, s,$$

where $C_0 = 0$ and h and k are chosen constants. In the case of a step change, the value of the parameter k is usually $k = (\mu^0 + \mu^1)/2$. Sometimes the CUSUM alarm statistic is presented recursively by the formula

$$S_s = \max(0, S_{s-1} + Y(s) - k),$$

where $S_0 = 0$. This alarm statistic is used with a constant alarm limit.

In the study of events of diseases the case of a Poisson process is of special interest. Sometimes the exact time of the event is not known or convenience motivates that only the number of events in fixed intervals are recorded. This number is usually assumed to be Poisson distributed. To account for possible over-dispersion, a negative binomial distribution can be used instead (Gallus *et al.*, 1991). A direct analogue to the cumulative sums for the normal case is the Poisson CUSUM which compares the recorded number of events in each time period with the expected number and uses the cumulated sum of deviations to form an alarm statistic. The value of k can be derived from the likelihood. Ewan and Kemp (1960) described this and tried different values of k in the formula. Hill *et al.* (1968) used this method for monitoring congenital malformations. In Barbujani (1987) comparisons with earlier methods for this problem are made. Hutwagner *et al.* (1997) use the CUSUM method for the case of salmonella outbreaks. Lucas (1985) described the Poisson CUSUM method and suggested the value of k derived by analogy to sequential probability ratio tests. This makes the method agree with the CUSUM as expressed by the likelihood expressions in the beginning of this section. Different approximations to the Poisson CUSUM were suggested by Rossi *et al.* (1999) in order to overcome problems with an unknown or varying baseline.

When using the continuous, exponentially distributed time between adverse events the exponential CUSUM can be constructed in the same way as for the Poisson distribution (see Lucas, 1985; Vardeman and Ray, 1985; Gan, 1992b; Mathers *et al.* 1994). In contrast to the Poisson CUSUM, the exponential CUSUM is not based on an initial reduction of the information. Such reduction of the information from the interval data should be avoided if possible. Other types of information reduction include, for example, only recording the time when a certain number of events has happened or only recording whether the time between events is larger than a threshold value or not. These types of reductions are used in the sets method (Chen 1978; Gallus *et al.*, 1986) and the cuscore method (Radaelli, 1992). Sometimes the time to event is measured as the number of positive cases (e.g. the number of healthy newborn babies) between negative ones (e.g. a baby born with congenital malformation). In those cases the negative binomial distribution might be an appropriate distribution (Radaelli, 1994) to base a CUSUM method on.

Different variants of CUSUM methods have been proposed for spatial surveillance by, for example, Raubertas (1989), Rogerson (1997, 2001), and Rogerson and Yamada (2004).

Closely related to the CUSUM method are the generalized likelihood ratio (GLR) and mixture likelihood ratio (MLR) methods. For the MLR method, suggested by Pollak and Siegmund (1975), a prior for the shift size is used in the CUSUM method. For the GLR method the alarm statistic is formed by maximizing over possible values of the shift (besides the maximum over possible times of the shift). Lai (1998) describes both GLR and MLR and prove a minimax result for a variant of GLR suitable for autocorrelated data.

The CUSUM method satisfies the minimax criterion of optimality described in Section 3.5.2. Also other good qualities of the method have confirmed by, for example, Srivastava and Wu (1993) and Frisén and Wessman (1999). With respect to the expected delay, the CUSUM method is almost as good as the LR and the Shiryaev–Roberts method.

3.6.4 Moving Average and Window-Based Methods

The alarm condition for the moving average method can be expressed by the likelihood ratios $L(s,t)$ as

$$L(s, s-d) > G,$$

where G is a constant and d is a fixed window width. For the standard case of normally distributed variables described in Section 3.3 this will be a moving average. It will have the optimal error probabilities of the LR method with $C = \{\tau = s-d\}$, and will thus have optimal detection abilities for changes which occurred d time points earlier.

Previously, the Food and Drug Administration (FDA) recommended a window-based method to detect increased frequencies of adverse events related to drugs. In this case the number of reported adverse events in a 'report interval' was compared to that in a 'comparison interval' and reported to the FDA. This recommendation was revoked in 1997 on the ground that this type of report had not contributed to the timely identification of safety problems.

The number of events in a moving window of fixed length is compared to an expected number based on the previous years in a retrospective setting by Stroup *et al.* (1989, 1993). For prospective use, Wharton *et al.* (1993) used data from the National Notifiable Diseases Surveillance System (NNDSS) for a 4-month period, and Rigau-Perez *et al.* (1999) applied it to dengue outbreaks in Puerto Rico. However, window-based methods do not utilize all available information. Using data recorded in moving windows will reduce the information about the observed process. If the window is wide it will smooth over possible shifts in the process. If, on the other hand, the window is narrow, the

information lost will be larger since only a small amount of the observations are used at each time point. One motivation for the use of moving windows is to overcome the problem of an unknown baseline.

Sometimes, as in Lai (1998), advanced methods such as the GLR method are combined with a window technique in order to ease the computational burden.

3.6.5 Exponentially Weighted Moving Average Methods

A variant of the moving average method which does utilize all information is the EWMA method. The alarm statistic is based on exponentially weighted moving averages,

$$Z_s = (1-\lambda)Z_{s-1} + \lambda Y(s), \qquad s = 1, 2, \ldots,$$

where $0 < \lambda < 1$ and Z_0 is the target value, which is normalized to zero. The EWMA statistic gives the most recent observation the greatest weight, and gives all previous observations geometrically decreasing weights. If λ is near zero, all observations have approximately the same weight. Note that if $\lambda = 1$ is used, the EWMA method reduces to the Shewhart method. The asymptotic variant, EWMAa, will give an alarm at

$$t_A = \min\{s : Z_s > L\sigma_Z\},$$

where L is a constant. For the EWMAe version of the method, the exact standard deviation (which is increasing in s) is used instead of the asymptotic in the alarm limit. A comparison between the EWMAa and the EWMAe method is given in Sonesson (2003), where it is found that the EWMAa version is preferable for most cases.

The EWMA method was described by Roberts (1959). Positive reports of the quality of the method are given by Crowder (1989), Lucas and Saccucci (1990), Domangue and Patch (1991), and Knoth and Schmid (2002). The choice of λ is important, and the search for the optimal value of λ has been of great interest in the literature. Small values of λ result in good ability to detect early changes, while larger values are necessary for changes that occur later.

Most reports on optimal values of the parameter λ concern the ARL criterion. Frisén (2003) demonstrated that there exist methods with equal weights for all observations which are ARL optimal. This is a reason for choosing equal weights for the EWMA method. To get equal weights for all observations by the EWMA method, λ should approach zero. Methods which allocate the power to the first time points will have good ARL properties but worse ability to detect a change that happens later. In fact, wisely enough, no one seems to have suggested that λ should be chosen to be zero, even though that should fulfill the ARL criterion.

The EWMA method can be seen as a linear approximation of the LR method for a certain value of the parameter of the method,

$$\lambda^* = 1 - \exp(-\mu^2/2)/(1-\nu),$$

which has a specific value if the change μ and the intensity of change ν are specified. This was shown by Frisén (2003), and an additional adaptation of the EWMA method by changing the alarm limits was suggested. This modification of the EWMA method leads to a method which is approximately optimal with respect to the minimal expected delay.

Adaptations of the EWMA method for binomial and Poisson data are made, for example, by Gan (1991) and Borror *et al.* (1998). Public health applications are considered by Williamson and Hudson (1999) and VanBrackle and Williamson (1999).

3.7 SPECIAL ASPECTS OF OPTIMALITY FOR SURVEILLANCE OF PUBLIC HEALTH

Some aspects of public health, such as the kind of processes of special interest, have been treated throughout this chapter. In this section some special aspects of optimality which are of concern in spatial and other surveillance of public health are treated separately. When the states (between which the change occurs) are completely specified, the LR method, with its good optimality properties, can be used. Pollak and Siegmund (1985) point out that the martingale property (for continuous time) of the Shiryaev–Roberts method makes it more suitable than the CUSUM method (which does not have this property) for adaptation to complicated problems. On the other hand, Lai (1995, 1998) and Lai and Shan (1999) argue that the good minimax properties of generalizations of the CUSUM method make the CUSUM suitable for complicated problems. In complicated problems it is not always easy to achieve, or even define, exact optimality.

3.7.1 Gradual Changes during Outbreaks of Diseases

Most of the literature on surveillance treats the case of an abrupt change, which might be caused, for example, by a sudden bioterrorist attack. However, in many cases of public health surveillance the change is gradual, for example at the outbreak of a contagious disease. The change is thus more complicated than the standard situation with a sudden shift from one level to another. A change in slope of a linear regression is the case most studied. Aerne *et al.* (1991) and Gan (1992a) suggest CUSUM methods for the residuals from a known regression. For the case of an unknown pre-change regression, Krieger *et al.* (2003) suggest CUSUM and Shiryaev–Roberts methods based on a statistic which does not

depend on the unknown parameters. Arteaga and Ledolter (1997) compare several procedures with respect to ARL properties for several different monotonic changes. One of the methods suggested in that paper is a window method based on the likelihood ratio and isotonic regression techniques. In general, window methods (Section 3.6.4 are inefficient for detection of gradual changes (Järpe, 2000). Yashchin (1993) discusses generalizations of the CUSUM and EWMA methods to detect both sudden and gradual changes.

At an outbreak the incidence typically increases gradually and then possibly declines (Buehler *et al.*, 2003). It might be hard to model exactly the shape of the rise and the decline, or even to estimate the baseline accurately. The timely detection of a change in monotonicity is then of interest. The start of an increase is of course of special interest, but the decline might also be of interest since influenza-like symptoms after the influenza season might indicate, for example, a bioterrorist attack.

When the knowledge of the shape of the curve is uncertain, a nonparametric method is of interest. Frisén (2000) suggested surveillance that is not based on any parametric model but only on monotonicity restrictions. The surveillance method was described and evaluated by Andersson (2002). The method is developed for cyclical processes with the aim of detecting a turn (peak or trough) as soon as possible. It is a Shiryaev–Roberts variant of the maximum likelihood ratio method based on the statistic

$$\frac{\max f_{Y_s}(y_s|C(s))}{\max f_{Y_s}(y_s|D(s))}$$

using the maximum likelihood over the class of all monotonic (D) or unimodal (C) functions. The maximum likelihood estimator of μ under the monotonicity restriction is described by, for example, Robertson *et al.* (1988). The maximum likelihood estimator of μ under the unimodality restriction was given by Frisén (1986).

3.7.2 Change between Unknown Incidences

The value of the incidence after a change is seldom known. However, false alarm properties will remain even if the size of the change is not known. For the design of methods, it is required only to specify a specific type of change which the methods should be as optimal as possible in detecting. The Shewhart method does not involve the size of the shift as a parameter. This is a disadvantage.

Knowledge of the baseline is important. Often the baseline rate is estimated and used as a plug-in value in the method. The estimated baseline value will affect the performance of the method. If the baseline rate is underestimated we will get more false alarms than if the true value is used. The opposite is true if the baseline rate is overestimated. A 'self-starting' technique is described in Chapter 6. To detect emerging clusters of a disease, Kleinman *et al.* (2004)

estimated the baseline by generalized linear mixed models using a history of naturally occurring disease. The approach was illustrated using data on health care visits in the context of syndromic surveillance for anthrax. For spatial surveillance, not only is the baseline level important, but also the baseline spatial correlation between regions during the in-control period. Rogerson and Yamada (2004) showed that if the spatial correlation was ignored in the construction of a multivariate CUSUM method, when there in fact was true spatial correlation between regions, then ARL^0 was worse than anticipated.

One way to avoid the problem of unknown parameters is to transform the data to invariant statistics. Frisén (1992) and Sullivan and Jones (2002) use the deviation of each observation from the average of all previous ones. Gordon and Pollak (1997) use invariant statistics combined by the Shiryaev–Roberts method to handle the case of an unknown pre-change mean of a normal distribution. Krieger *et al.* (2003) use invariant statistics combined by the CUSUM and by the Shiryaev–Roberts method for surveillance of change in regression from an unknown pre-change one. In Chapter 5 a model (and corresponding statistic) which does not require the number of people at risk is demonstrated to be useful.

When both the baseline incidence and the increase at a change are unknown we aim for the detection of a change to a stochastically larger distribution. Bell *et al.* (1994) suggested a nonparametric method geared to the exponential distribution. They apply their method to the detection of a change of the parameter in a Bernoulli process to an unknown but larger value. Asymptotic efficiency for their method is reported. The nonparametric method of Section 3.7.1 designed for detection of a change from monotonicity also avoids the problem of unknown values of the baseline and the change by only using the monotonicity properties.

The problem of unknown parameter values can be handled by a statistic which involves the maximum difference (measured by the likelihood ratio, for example) between the baseline and the changed level. The GLR method (Lai, 1995, 1998) uses the maximum likelihood estimator of the value after the change. Kulldorff (2001) used the same technique for detection of clustering in spatial patterns.

The MLR method suggested by Pollak and Siegmund (1975) uses priors for the unknown parameters in the CUSUM method. Priors are also used by Radaelli (1996) for the sets method. Lawson (2004) used priors for the unknown parameters to calculate the posterior means in a Bayesian space-time interaction model.

To control the false alarms is usually more important than to optimize the detection ability. The unknown parameters can be handled within different frameworks corresponding to different restrictions on possible optimality.

3.7.3 Spatial and Other Multivariate Surveillance

In many public health surveillance programs measurements are obtained not only in time but also at various locations. For example, the cases of disease

reported to the CDC through the NNDSS are collected at various places all over the USA. Several interesting new approaches for public health surveillance have recently been developed, such as EARS (described by Hutwagner *et al.*, 2003), ESSENCE (Lombardo *et al.*, 2003), RODS (Tsui *et al.*, 2003), and WSARE (Wong *et al.*, 2003) (also described in Chapter 10).

The inferential problems involved in spatial surveillance are multivariate and such techniques are discussed in this book both in Chapters 6 and 7 on spatial surveillance and in Chapter 9 on multivariate surveillance. Also, when many symptoms are considered we have a situation of multivariate surveillance.

When confronted with a problem involving both spatial and temporal components, which is the case in surveillance of spatial structures, different approaches can be used. A stepwise reduction of the surveillance problem is common. One is to perform parallel surveillance for each spatial component (e.g. location) and sound a general alarm when there is an alarm for any of the components. This approach was used for different cluster sizes and cluster locations by Kulldorff (2001).

Another way is to first reduce the information to one statistic which expresses the spatial pattern, and then monitor this statistic in time. Then usually the correlations between the variables are used in the transformation. Rogerson (1997, 2001) uses this approach together with a CUSUM method. One can alternatively use a vector accumulation approach, where the information is cumulated in vectors, and for each time point transform these into alarm statistics; see, for example, Rogerson and Yamada (2004). It is also possible to construct the multivariate method while aiming to satisfy some global optimality criterion. Järpe (1999) suggested an ED optimal surveillance method of clustering in a spatial log-linear model.

The multivariate methods can be evaluated by the measures and criteria described above. For example, Frisén and Wessman (1999) suggested a generalization of the ARL measure to allow for the possibility of different change times for different variables. Control of the false discovery rate is of interest when conclusions are made about several variables and is used by Wong *et al.* (2003). However, optimality is always complicated in multidimensional cases. No approach will be uniformly optimal for all kinds of changes. The methods which are optimal for detection of changes at a few prespecified locations are different from those which are optimal for detection of a change in all or many locations. This is exemplified by ARL^1 in Rogerson and Yamada (2004) for the case of spatial surveillance.

3.8 CONCLUDING REMARKS

The need for proper statistical evaluation is evident and the importance of timeliness of on-line surveillance is more and more accepted. It is important to recognize the sequential type of the decision situation.

Optimality and evaluation of methods for spatial surveillance of public health in practice are very important. It is necessary to know the basic properties of a system before it is implemented. This involves many important aspects. One of them is the use of proper statistical measures for evaluation which take care of the special features of a surveillance system. There is a great difference between surveillance and hypothesis testing in this respect.

The most commonly used formal optimality criterion specially designed for surveillance is the ARL criterion. The logical drawbacks of this criterion and the advantages of other ones, such as the minimal expected delay for a fixed value of the probability of a false alarm, are discussed by Frisén (2003).

For the ED criterion knowledge of the distribution of the time of the change is used. In practice, some knowledge should be available and should influence the choice of method. Frisén and Wessman (1999) demonstrated that the LR method is very robust in this respect.

The lack of need of a distribution for the change time might be seen as an advantage for the minimax optimality criterion. However, this pessimistic view might not reflect the situation in public health. Besides, it is not self-evident that the possibility to optimize a method for a parameter should be seen as a disadvantage even though it is hard to specify which value is of most interest.

In many applications, including public health surveillance, one measure of performance alone is not enough. Therefore, one should aim at a complete and thorough evaluation of proposed systems. We suggest using measures such as the expected delay, the probability of successful detection and the predictive value. Another important aspect of a system for surveillance is the robustness against misspecification.

Knowledge of the properties of a system for surveillance is very important both for the choice of the appropriate method and for the interpretation of an alarm.

ACKNOWLEDGMENT

Supported by grant F0473/2000 from the Swedish Council for Research in the Humanities and Social Sciences.

PART II

Basic Methods for Spatial and Syndromic Surveillance

4

Spatial and Spatio-Temporal Disease Analysis

Andrew B. Lawson

4.1 INTRODUCTION

The representation and analysis of maps of disease incidence data is now established as a basic tool in the analysis of regional public health. One of the earliest examples of disease mapping is the map of the addresses of cholera victims related to the locations of water supplies, by Snow (1854). In that case, the street addresses of victims were recorded and their proximity to putative pollution sources (water supply pumps) was assessed.

The subject area of disease mapping has developed considerably in recent years. This growth in interest has led to a greater use of geographical or spatial statistical tools in the analysis of data both routinely collected for public health purposes and found within ecological studies of disease relating to explanatory variables. The study of the geographical distribution of disease can have a variety of uses. The main areas of application can be conveniently broken down into the following classes: disease mapping; disease clustering; and ecological analysis. In the first class, usually the object of the analysis is to provide (estimate) the true *relative risk* of a disease of interest across a geographical study area (map): a focus similar to the processing of pixel images to remove noise. Applications for such methods lie in health services resource allocation and in disease atlas construction (see, for example, Pickle and Hermann, 1995).

The second class, that of disease clustering, has particular importance in public health surveillance, where it may be important to be able to assess whether a disease map is clustered and where the clusters are located. This may lead to examination of potential environmental hazards. A particular special

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

case arises when a known location is thought to be a potential pollution hazard. The analysis of disease incidence around a putative source of hazard is a special case of cluster detection.

The third class, that of ecological analysis, is of great relevance within epidemiological research, as its focus is the analysis of the geographical distribution of disease in relation to explanatory covariates, usually at an aggregated spatial level. Many issues relating to disease mapping are also found in this area, in addition to issues relating specifically to the incorporation of covariates.

In the following, the issues surrounding the first class of problems, namely disease mapping, are the focus of attention. While the focus here is on *statistical* methods and issues in disease mapping, it should be noted that the results of such statistical procedures are often represented visually in mapped form. Hence, some consideration must be given to the purely cartographic issues that affect the representation of geographical information. The method chosen to represent disease intensity on the map, be it color scheme or symbolic representation, can dramatically affect the resulting interpretation of disease distribution. It is not the purpose of this review to detail such cognitive aspects of disease mapping, but the reader is directed to some recent discussions of these issues: MacEachren (1995), Monmonier (1996), Pickle and Hermann (1995), and Walter (1993).

4.2 DISEASE MAPPING AND MAP RECONSTRUCTION

To begin, we consider two different mapping situations which clearly demarcate approaches to this area. These situations are defined by the form of the mapped data which arises in such studies. First, the lowest level of aggregation of data observable in disease incidence studies is the case itself. Its geographical reference (georeference), usually the residential address of the case, is the basic mapping unit. This type of data is often referred to as *case event* data. We usually define a fixed study area, denoted as W, the study window, within which occur m case events. We term this a *realization* of events within W. The locations of the residences of the cases are denoted by $\{\mathbf{x}_i\}, i = 1, \ldots, m$.

The second type of data commonly found in such studies is a count of disease cases within arbitrarily defined administrative regions (*tracts*), such as census tracts, electoral districts or health authority areas. Essentially the count is an aggregation of all the cases within the tract. Therefore the georeference of the count is related to the tract location, where the individual case spatial references (locations) are lost. Denote the counts of disease within p tracts by $\{y_i\}, i = 1, \ldots, p$. Often the latter form of data is more commonly available from routine data sources such as government agencies than the first form. Confidentiality can limit access to the case event realization.

4.3 DISEASE MAP RESTORATION

4.3.1 Simple Statistical Representations

The representation of disease incidence data can vary from simple point object maps for cases and pictorial representation of counts within tracts, to the mapping of estimates from complex models purporting to describe the structure of the disease events. In this section, we describe the range of mapping methods from simple representations to model-based forms. The geographical incidence of disease has as its fundamental unit of observation the address location of cases of disease. The residential address (or possibly the employment address) of cases of disease contains important information relating to the type of exposure to environmental risks. Often, however, the exact address locations of cases are not directly available, and one must use instead counts of disease in arbitrary administrative regions, such as census tracts or postal districts. This lack of precise spatial information may be due to confidentiality constraints relating to the identification of case addresses or may be due to the scale of information gathering.

4.3.1.1 Crude representation of disease distribution

The simplest possible mapping form is the depiction of disease rates at specific sets of locations. For case events, this is a map of case event locations. For counts within tracts, it is a pictorial representation of the number of events in the tracts plotted at a suitable set of locations (e.g. tract centroids). The locations of case events within a spatially heterogeneous population can display a small amount of information concerning the overall pattern of disease events within a window. Ross and Davis (1990) provide an example of such an analysis of leukemia cluster data. However, any interpretation of the structure of these events is severely limited by the lack of information concerning the spatial distribution of the background population which might be 'at risk' from the disease of concern and which gave rise to the cases of disease. This population also has a spatial distribution, and failure to take account of this spatial variation severely limits the ability to interpret the resulting case event map. In essence, areas of high density of 'at risk' population would tend to yield high incidence of case events and so, without taking account of this distribution, areas of high disease intensity could be spuriously attributed to excess disease risk.

In the case of counts of cases of disease within tracts, similar considerations apply when crude count maps are constructed. Here, variation in population density also affects the spatial incidence of disease. It is also important to consider how a count of cases could be depicted in a mapped representation. Counts within tracts are totals of events from the whole tract region. If tracts are irregular, then a decision must be made either to 'locate' the count at some tract

location (e.g. tract centroid, however defined) with suitable symbolization, or to represent the count as a fill color or shade over the whole tract (choropleth thematic map). In the former case, the choice of location will affect interpretation. In the latter case, symbolization choice (shade and/or color) could also distort interpretation, although an attempt to represent the whole tract may be attractive.

In general, methods that attempt to incorporate the effect of background 'at risk' population are to be preferred. These are discussed in the next section.

4.3.1.2 Standardized mortality/morbidity ratios and standardization

To assess the status of an area with respect to disease incidence, it is convenient first to attempt to assess what disease incidence should be locally 'expected' in the tract area and then to compare the observed incidence with the 'expected' incidence. This approach has been traditionally used for the analysis of counts within tracts and can also be applied to case event maps.

Case events Case events can be depicted as a map of point event locations. For the purposes of assessment of differences in local disease risk it is appropriate to convert these locations into a continuous surface describing the spatial variation in *intensity* of the cases. Once this surface is computed, then a measure of local variation is available at any spatial location within the observation window. Denote the intensity surface as $\lambda(\mathbf{s})$, where $\mathbf{s}$ is a spatial location. This surface can be formally defined as the first-order intensity of a point process (Lawson and Waller, 1996), and can be estimated by a variety of methods, including density estimation (Härdle, 1991). To provide an estimate of the 'at risk' population at spatial locations, it is necessary first to choose a measure that will represent the intensity of cases 'expected' at such locations. Define this measure as $\lambda_0(\mathbf{s})$. Two possibilities can be explored.

First, it is possible to obtain rates for the case disease from either the whole study window or a larger enclosing region. Often these rates are available only at an aggregated level (e.g. census tracts). The rates are obtained for a range of subpopulation categories which are thought to affect the case disease incidence. For example, the age and sex structure of the population or the deprivation status of the area (see, for example, Carstairs, 1981) could affect the amount of population 'at risk' from the case disease. The use of such external rates is often called external standardization (Inskip *et al.*, 1983). It should be noted that rates computed from aggregated data will be less variable than those based on density estimation of case events.

An alternative method of assessing the 'at risk' population structure is to use a case event map of another disease, which represents the background population but is not affected by the etiological processes of interest in the case disease. For example, the spatial distribution of coronary heart disease (CHD: ICD code, list A 410–414), could provide a *control* representation for

respiratory cancer (ICD code, list A 162) when the latter is the case disease in a study of air pollution effects, as CHD is less closely related to air pollution insult. Other examples of the cited use of a control disease would be: larynx cancer (case) and lung cancer (control) (Diggle, 1990), although this control is complicated by the fact that lung cancer is also related to air pollution risk; lower body cancers (control) and gastric cancer (case), where lower body organs may only be affected by specific pollutants such as nickel (Lawson and Williams, 2000); birth defects (case) and live births (control).

While exact matching of diseases in this way will always be difficult, there is an advantage in the use of control diseases in case event examples. If a realization of the control disease is available in the form of a point event map, then it is possible also to compute an estimate of the first-order intensity of the control disease. This estimate can then be used directly to compare case intensity with background intensity. Note that $\lambda_0(\mathbf{s})$ can be estimated, equally, from census tract standardized rates (see, for example, Lawson and Williams, 1994).

The estimates of $\lambda(\mathbf{s})$ and $\lambda_0(\mathbf{s})$ can be compared in a variety of ways. First, it is possible to map the ratio form,

$$\widehat{R}(\mathbf{s}) = \frac{\widehat{\lambda}(\mathbf{s})}{\widehat{\lambda}_0(\mathbf{s})}, \tag{4.1}$$

as suggested by Bithell (1990). Modifications to this procedure have been proposed by Lawson and Williams (1993) and Kelsall and Diggle (1995). Care must be taken to consider the effects of study/observation window edges on the interpretation of the ratio. Some edge-effect compensation should be considered when there is a considerable influence of window edges in the final interpretation of the map. A detailed discussion of edge effects can be found in Lawson *et al.* (1999) and Vidal-Rodeiro and Lawson (2005).

Apart from ratio forms, it is also possible to map transformations of ratios (e.g. $\log \widehat{R}(\mathbf{s})$) or to map

$$\widehat{D}(\mathbf{s}) = \widehat{\lambda}(\mathbf{s}) - \widehat{\lambda}_0(\mathbf{s}). \tag{4.2}$$

The choice of (4.1) or (4.2) will depend on the underlying model assumed for the excess risk.

In all the approaches above to the mapping of case event data, some smoothing or interpolation of the event or control data has to be made. The statistical properties of this operation depend on the method used for estimation of each component of the map. Optimal choices of the smoothing constant (i.e. bandwidth) are known for density estimation and kernel smoothing (Härdle, 1991).

Tract counts As in the analysis of case events, it is usual to assess maps of count data by comparison of the observed counts to those counts 'expected' to arise given the 'at risk' population structure of the tracts. Traditionally, the ratio of observed to expected counts within tracts is called a standardized mortality/morbidity ratio (SMR) and this ratio is an estimate of *relative risk* within each tract (i.e. the ratio describes the odds of being in the disease group rather than the background group). The justification for the use of SMRs can be supported by the analysis of likelihood models with multiplicative expected risk (see, for example, Breslow and Day, 1987).

Define y_i as the observed count of the case disease in the ith tract, and e_i as the expected count within the same tract. Then the SMR is defined as

$$\widehat{R}_i = \frac{y_i}{e_i}. \tag{4.3}$$

The alternative measure of the relation between observed and expected counts, which is related to an additive risk model, is the difference,

$$\widehat{D}_i = y_i - e_i. \tag{4.4}$$

In this case it must be decided whether to express the $\widehat{R}_i$ or $\widehat{D}_i$ as fill patterns in each region, or to locate the result at some specified tract location, such as the centroid. If it is decided that these measures should be regarded as continuous across regions then some further interpolation of $\widehat{R}_i$ or $\widehat{D}_i$ must be made (see Breslow and Day, 1987, pp. 198–199). Figure 4.1 displays the SMR map for congenital abnormality deaths for 1990 in South Carolina, USA.

SMRs are commonly used in disease map presentation, but have many drawbacks. First, they are based on ratio estimators and hence can yield large changes in estimate with relatively small changes in expected value. In the extreme, when a (near-)zero expectation is found the SMR will be very large for any positive count. Also the zero SMRs do not distinguish variation in expected counts, and the SMR variance is proportional to $1/e_i$. The SMR is essentially a saturated estimate of relative risk and hence is not parsimonious.

4.3.1.3 Interpolation

In many of the mapping approaches mentioned above, use must be made of interpolation methods to provide estimates of a surface measure at locations where there are no observations. For example, we may wish to map contours of a set of tract counts if we believe the counts to represent a continuously varying risk surface. For the purposes of contouring, a grid of surface interpolant values must be provided. Smoothing of SMRs has been advocated by Breslow and Day (1987). Those authors employ kernel smoothing to interpolate the surface (in a temporal application). The advantage of such smoothing is that the

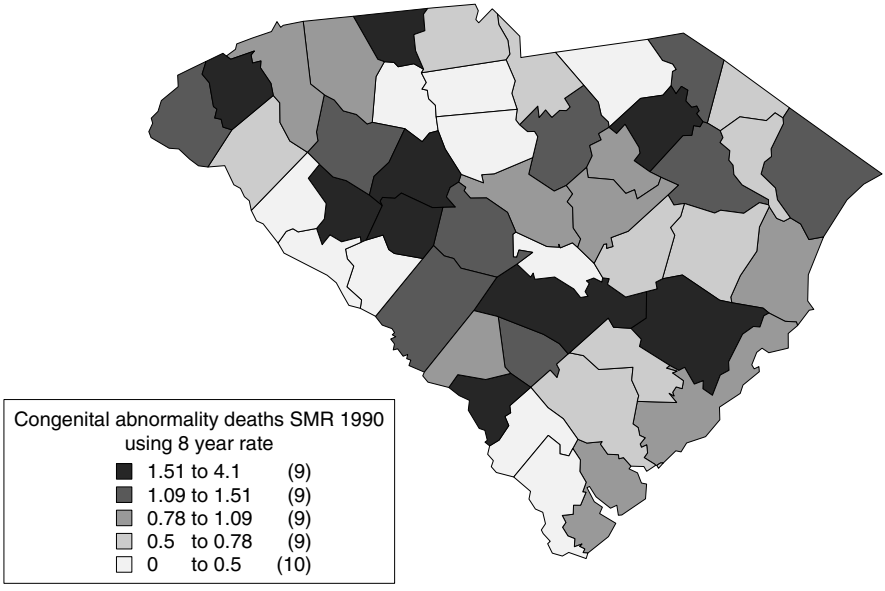


Figure 4.1 SMR for congenetial abnormality deaths, South Carolina, 1990.

method preserves the positivity condition of SMRs: that is, the method does not produce negative interpolants (which are invalid), unlike kriging methods (for discussion of this issue, see Lawson and Cressie, 2000). Many mapping packages utilize interpolation methods to provide gridded data for further contour and perspective view plotting (e.g. ArcView, R, S-Plus). However, often the methods used are not clearly defined or they are based on mathematical rather than statistical interpolants (e.g. the Akima or Delauney interpolator).

Note that the comments above also apply directly to case event density estimation. The use of kernel density estimation is recommended, with edge correction as appropriate. For ratio estimation, Kelsall and Diggle (1995) recommend the joint estimation of a common smoothing parameter for the numerator and denominator of $R(\mathbf{s})$ when a control disease realization is available.

4.3.1.4 Exploratory methods

The discussion above, concerning the construction of disease maps, could be considered as exploratory analysis of spatial disease patterns. For example, the construction and mapping of ratios or differences of case and background measures is useful for highlighting areas of incidence requiring further consideration. Contour plots or surface views of such mapped data can be derived. Comments concerning the psychological interpretation of mapped patterns also

apply here (Walter, 1993; Ripley, 1981). However, inspection of maps of simple ratios or differences cannot provide accurate assessment of the statistical significance of, for example, areas of elevated disease risk. Proper inference requires statistical models, and that is the subject of the next section.

4.3.2 Basic Models

In the previous section we discussed the use of primarily *descriptive* methods in the construction of disease maps. These methods do not introduce any particular model structure or constraint into the mapping process. This can be advantageous at an early or exploratory stage in the analysis of disease data but, when more substantive hypotheses and/or greater amounts of prior information are available concerning the problem, it may be advantageous to consider a model-based approach to disease map construction. Model-based approaches can also be used in an exploratory setting, and if sufficiently general models are employed then this can lead to better focusing of subsequent hypothesis generation. In what follows, we first consider likelihood models for case event data and then discuss the inclusion of extra information in the form of random effects.

4.3.2.1 Likelihood models

Denote a realization of n case events within a window W, as $\{\mathbf{x}_i\}, i = 1, \ldots m$. In addition, define the count of cases of disease within the ith tract $\{W_i\}, i = 1, \ldots, p$, of an arbitrarily regionalized tract map as y_i.

Case event data Usually the basic model for case event data is derived from the following assumptions:

(1) Individuals within the study population behave independently with respect to disease propensity, after allowance is made for observed or unobserved confounding variables.

(2) The underlying 'at risk' population intensity has a continuous spatial distribution, within specified boundary vertices.

(3) The case events are unique, in that they occur as single spatially separate events.

Assumption 1 allows the events to be modeled via a likelihood approach, which is valid conditional on the outcomes of confounder variables. Further, assumption 2, if valid, allows the likelihood to be constructed with a background continuous modulating intensity function $\{\lambda_0(\mathbf{s})\}$ representing the 'at risk' population. The uniqueness of case event locations is a requirement of point process theory (the property called orderliness: see Daley and Vere-Jones,

1988), which allows the application of Poisson process models in this analysis. Assumption 1 is generally valid for noninfectious diseases. It may also be valid for infectious diseases if the information about current infectives were known at given time points. Assumption 2 will be valid at appropriate scales of analysis. It may not hold when large areas of a study window include zones of zero population (e.g. harbors or industrial zones). Often models can be restricted to exclude these areas, however. Assumption 3 will usually hold for relatively rare diseases but may be violated when households have multiple cases and these occur at coincident locations. This may not be important at more aggregate scales, but could be important at a fine spatial scale. Remedies for such nonorderliness are the use of declustering algorithms (which perturb the locations by small amounts), or analysis at a higher aggregation level. Note that it is also possible to use a conventional case–control approach to this problem (Diggle *et al.*, 2000).

Given the assumptions above, it is possible to specify that the case events arise as a realization of a Poisson point process, modulated by $\lambda_0(\mathbf{s})$, with first-order multiplicative intensity:

$$\lambda(\mathbf{x}) = \rho\lambda_0(\mathbf{s})\lambda_1(\mathbf{s};\boldsymbol{\theta}). \tag{4.5}$$

In this definition, $\lambda_1(\mathbf{s};\boldsymbol{\theta})$ represents a function of confounder variables as well as location, $\boldsymbol{\theta}$ is a parameter (vector) and ρ is the overall constant rate of the process. The confounder variables can be widely defined, however. For example, a number of random effects could be included to represent unobserved effects, as well as observed covariates, as could functions of other locations. The inclusion of random effects could be chosen if it is felt that unobserved heterogeneity is present in the disease process. This could represent the effect of known or unknown covariates which are unobserved. The likelihood associated with this is given by:

$$L = \left[\prod_{i=1}^{m} \lambda(\mathbf{s}_i)\right] \exp\left\{-\int_W \lambda(\mathbf{u})d\mathbf{u}\right\}. \tag{4.6}$$

For suitably specified $f(.)$, a variety of models can be derived. In the case of disease mapping, where only the background intensity is to be accounted for, a reasonable approach to intensity parameterization is $\lambda(\mathbf{x}) = \rho\lambda_0(\mathbf{s})\lambda_1(\mathbf{s})$. The preceding definition can be used as an informal justification for the use of intensity ratios $(\widehat{\lambda}(\mathbf{s})/\widehat{\lambda}_1(\mathbf{s}))$, in the mapping of case event data; such ratios represent the local 'extraction' of 'at risk' background, under a multiplicative hazard model. On the other hand, under a pure additive model, $\lambda(\mathbf{s}) = \rho[\lambda_0(\mathbf{s}) + \lambda_1(\mathbf{s};\boldsymbol{\theta})]$ say, differencing the two estimated rates would be supported.

Tract count data In the case of observed counts of disease within tracts, the Poisson process assumptions given above mean that the counts are Poisson

distributed with, for each tract, a different expectation : $\int_{W_i} \lambda(\mathbf{u})d\mathbf{u}$, where W_i denotes the extent of the ith tract. Then the log-likelihood based on a Poisson distribution is, bar an additive constant only depending on the data, given by

$$l = \sum_{i=1}^{p} \left\{ y_i \log \int_{W_i} \lambda(\mathbf{u})d\mathbf{u} - \int_{W_i} \lambda(\mathbf{u})d\mathbf{u} \right\}, \tag{4.7}$$

where p is the number of tracts.

Often a parameterization in (4.7) is assumed where, as in the case event example, the intensity is defined as a simple multiplicative function of the background $\lambda_0(\mathbf{s})$. An assumption is often made at this point that the integration over the ith tract area can be regarded as a parameter within a model hierarchy, without considering the spatial continuity of the intensity. That is, $y_i \sim \text{Pois}(\lambda_i)$, where λ_i is the rate in the ith region.

The mapping of 'extracted' intensities for case events or modified SMRs for tract counts is based on the view that once the 'at risk' background is extracted from the observed data, then the resulting distribution of risk represents a 'clean' map of the ground truth. Of course, as the background function $\lambda_0(\mathbf{s})$ must usually be estimated, then some variability in the resulting map will occur by inclusion of different estimators of $\lambda_0(\mathbf{s})$. For example, for tract count data, the use of external standardization alone to estimate the expected counts within tracts may provide a different map from that provided by a combination of external standardization and measures of tract-specific deprivation (e.g. deprivation indices: see Carstairs, 1981). If any confounding variables are available and can be included within the estimate of the 'at risk' background, then these should be considered for inclusion within the $\lambda_0(\mathbf{s})$ function. Examples of confounding variables could be found from national census data, particularly relating to socioeconomic measures. These measures are often defined as 'deprivation' indicators, or could relate to lifestyle choices. For example, the local rate of car ownership or the percentage unemployed within a census tract or other small area could provide a surrogate measure for increased risk, due to correlations between these variables and poor housing, smoking lifestyles, and ill health. Hence, if it is possible to include such variables in the estimation of $\lambda_0(\mathbf{s})$, then any resulting map will display a close representation of the 'true' underlying risk surface. When it is not possible to include such variables within $\lambda_0(\mathbf{s})$, it is sometimes possible to adapt a mapping method to include covariables of this type by inclusion within $\lambda_1(\mathbf{s})$ itself.

4.3.2.2 Random effects and Bayesian models

In the sections above some simple approaches to mapping intensities and counts within tracts have been described. These methods assume that once all known and observable confounding variables are included within the $g(\mathbf{x})$ estimation then the resulting map will be clean of all artifacts and hence will depict the

true excess risk surface. However, it is often the case that unobserved effects could be thought to exist within the observed data and that these effects should also be included within the analysis. These effects are often termed *random* effects, and their analysis has provided a large literature both in statistical methodology and in epidemiological applications; for recent views, see Elliott *et al.* (2000) and Lawson (2001). Within the literature on disease mapping, there has been a considerable growth in recent years in modeling random effects of various kinds. In the mapping context, a random effect could take a variety of forms. In its simplest form, a random effect is an extra quantity of variation (or variance component) which is estimable within the map and which can be ascribed a defined probabilistic structure. This component can affect individuals or can be associated with tracts or covariables. For example, individuals vary in susceptibility to disease, and hence individuals who become cases could have a random component relating to different susceptibility. This is sometimes known as frailty. Another example is the interpolation of a spatial covariable to the locations of case events or tract centroids. In that case, some error will be included in the interpolation process, and could be included within the resulting analysis of case or count events. Also, the locations of case events might not be precisely known or subject to some random shift, which may be related to uncertain residential exposure. (However, this type of uncertainty may be better modeled by a more complex integrated intensity model, which no longer provides an independent observation model.) Finally, within any predefined spatial unit, such as tracts or regions, it may be expected that there could be components of variation attributable to these different spatial units. These components could have different forms, depending on the degree of prior knowledge concerning the nature of this extra variation. For example, when observed counts, thought to be governed by a Poisson distribution, display greater variation than expected (i.e. the variance is greater than the mean), it is sometimes described as overdispersion. This overdispersion can occur for various reasons. Often it arises when clustering occurs in the counts at a particular scale. It can also occur when considerable numbers of cells have zero counts (sparseness), which can arise when rare diseases are mapped. Furthermore, in spatial applications it is important to distinguish two basic forms of extra variation. First, as in the aspatial case, a form of independent and spatially uncorrelated extra variation can be assumed. This is often called *uncorrelated heterogeneity* (Besag *et al.*, 1991). Another form of random effect is that which arises from a model where it is thought that the spatial unit (case event, tract or region) is correlated with neighbouring spatial units. This is often termed *correlated heterogeneity*. Essentially, this form of extra variation implies that there exists spatial autocorrelation between spatial units: see Cliff and Ord (1981) for an accessible introduction to spatial autocorrelation. This autocorrelation could arise for a variety of reasons. First, the disease of concern could be naturally clustered in its spatial distribution at the scale of observation. Many infectious diseases display such spatial clustering, and a number of apparently

noninfectious diseases also cluster (Cuzick and Hills, 1991; Glick, 1979). Second, autocorrelation can be induced in spatial disease patterns by the existence of unobserved environmental or frailty effects. Hence, the extra variation observed in any application could arise from confounding variables that have not been included in the analysis. In disease mapping examples this could easily arise when simple mapping methods are used on SMRs with just basic age–sex standardization.

In the discussion above on heterogeneity, it is assumed that a global measure of heterogeneity applies to a mapped pattern. That is, any extra variation in the pattern can be captured by including a general heterogeneity term in the mapping model.

4.3.3 A Simple Overdispersion Model

A common assumption made when examining tract counts is that $y_i \sim \text{Pois}(e_i\theta_i)$ independently, and that $\theta_i \sim G(\alpha, \beta)$. The latter gamma distribution is often assumed for the Poisson rate parameter and provides for a measure of overdispersion relative to the Poisson distribution itself, depending on the α, β values used. The joint distribution is now given by the product of a Poisson likelihood and a gamma distribution. At this stage a choice must be made concerning how the random intensities are to be estimated or otherwise handled. One approach to this problem is to average over the values of θ_i to yield what is often called the *marginal* likelihood. Having averaged over this density, it is then possible to apply standard methods such as maximum likelihood. This is usually known as marginal maximum likelihood (Bock and Aitkin, 1981; Aitkin, 1996b). In this approach, the parameters of the gamma distribution are estimated from the integrated likelihood. A further development of this approach is to replace the gamma density by a finite mixture. This approach is essentially nonparametric and does not require the complete specification of the parameter distribution (Aitkin, 1996a). Although the example specified here concerns tract counts, the method described above can equally be applied to case event data, by inclusion of a random component in the intensity specification.

It is natural to consider modeling random effects within a Bayesian framework. First, random effects naturally have prior distributions and the joint density discussed above is proportional to the posterior distribution for the parameters of interest. Hence, full Bayes and empirical Bayes (posterior approximation) methods have developed naturally in the field of disease mapping. The prior distribution(s) for the ($\boldsymbol{\theta}$, say) parameters in the intensity specification $\rho\lambda_0(\mathbf{s})\lambda_1(\mathbf{s};\boldsymbol{\theta})$, have hyperparameters (in the Poisson–gamma example above, these were α, β). These hyperparameters can also have hyperprior distributions. The distributions chosen for these parameters depend on the application. In the full Bayesian approach, inference is based on the posterior distribution of $\boldsymbol{\theta}$ given the data. However, as in the frequentist

approach above, it is possible to adopt an intermediate approach where the posterior distribution is approximated in some way, and subsequent inference may be made via frequentist-style estimation of parameters or by computing the approximated posterior distribution. In the tract count example, approximation via intermediate prior parameter estimation would involve the estimation of α and β, followed by inference on the estimated posterior distribution (see Carlin and Louis, 1996, pp. 67–68).

Few examples exist of simple Bayesian approaches to the analysis of case event data in the disease mapping context. One approach which has been described (Lawson *et al.*, 1996) can be used with simple prior distributions for parameters and the authors provide approximate empirical Bayes estimators based on Dirichlet tile area integral approximations. For count data, a number of examples exist where independent Poisson distributed counts (with constant within-tract rate, λ_i) are associated with prior distributions of a variety of complexity. The earliest examples of such a Bayesian mapping approach can be found in Manton *et al.* (1981) and Tsutakawa (1988). Also, Clayton and Kaldor (1987) developed a Bayesian analysis of a Poisson likelihood model where y_i has expectation $\theta_i e_i$, and found that with a prior distribution given by $\theta_i \sim G(\alpha, \beta)$, the Bayes estimate of θ_i is the posterior expectation

$$\frac{y_i + \alpha}{e_i + \beta}. \tag{4.8}$$

Hence, one could map these Bayes estimates directly. Now the distribution of θ_i conditional on y_i is $G(y_i + \alpha,\ e_i + \beta)$ and a Bayesian approach would require summarization of θ_i from this posterior distribution. In practice, this is often obtained by generation of realizations from this posterior and then the summarizations are empirical (e.g. Markov Chain Monte Carlo (MCMC) methods). Other approaches and variants in the analysis of simple mapping models have been proposed by Tsutakawa (1988), Marshall (1991) and Devine and Louis (1994). In the next section, more sophisticated models for the prior structure of the parameters of the map are discussed.

4.3.4 Advanced Bayesian Models

Many of the models discussed above can be extended to include the specification of prior distributions for parameters and hence can be examined via Bayesian methods. In general, we distinguish here between empirical Bayes methods and full Bayes methods, on the basis that any method which seeks to approximate the posterior distribution is regarded as empirical Bayes (Bernardo and Smith, 1994). All other methods are regarded as full Bayes. This latter category includes maximum a posteriori estimation, estimation of posterior functionals, as well as posterior sampling.

Full posterior inference for Bayesian models has recently become feasible, largely because of the increased use of MCMC methods of posterior sampling. The first full sampler reported for a disease mapping example was a Gibbs sampler applied to a general model for intrinsic autoregression and uncorrelated heterogeneity by Besag *et al.* (1991). Subsequently, Clayton and Bernardinelli (1992), Breslow and Clayton (1993), and Bernardinelli *et al.* (1995) have adapted this approach to mapping, ecological analysis, and space-time problems.

This has been facilitated by the availability of general Gibbs sampling packages such as BUGS (GeoBUGS and WinBUGS) and MLwiN. Such Gibbs sampling methods can be applied to focused clustering problems as well as mapping/ecological studies. However, specific variations in model components (e.g. variation in the spatial correlation structure) cannot be easily accommodated in this general Bayesian package. Alternative, and more general, posterior sampling methods, such as the Metropolis–Hastings algorithm, are currently not available in a packaged form, although these methods can accommodate considerable variation in model specification.

Generalized linear mixed models have as their focus the inclusion of random effects within the generalized linear modeling framework. This is a general class of models allowing a range of data likelihoods (including Poisson and binomial) and the inclusion of uncorrelated and correlated heterogeneity. Often these models are fitted, after suitable approximations, using general software packages such as SAS or R. For example, for small area counts within m tracts a Poisson likelihood can be assumed with a log-linear model for the additive covariate and (random) effect of heterogeneity:

$$y_i \sim \text{Pois}(e_i\theta_i),$$
$$\log(\theta_i) = \boldsymbol{x}_i^t\boldsymbol{\beta} + u_i + v_i,$$

where $\boldsymbol{x}_i^t\boldsymbol{\beta}$ is a linear predictor of fixed effects, $\boldsymbol{x}_i^t$ is the ith row of the covariate design matrix, and $\boldsymbol{\beta}$ a parameter vector, v_i is an uncorrelated heterogeneity term, and u_i is a correlated heterogeneity term. In a full Bayesian analysis $\boldsymbol{\beta}, \{u_i\}\{v_i\}$ all have prior distributions. Approximations to the likelihood or posterior distributions allow the use of SAS or R to fit such models. WinBUGS can be used to carry out full Bayesian analysis. Chapter 5 examines in more detail aspects of these models in the surveillance context.

4.4 RESIDUALS AND GOODNESS OF FIT

The analysis of residuals and summary functions of residuals forms a fundamental part of the assessment of model goodness of fit in any area of statistical application. In the case of spatial or spatio-temporal analysis there is no exception, although full residual analysis is seldom presented in published work in the area. Often goodness-of-fit measures are aggregate functions of piecewise

residuals, while measures relating to individual residuals are also available. A variety of methods are available when full residual analysis is to be undertaken. Define a piecewise residual as the standardized difference between the observed value and the fitted model value. Usually the standardization will be based on a measure of the variability of the difference between the two values.

Within a frequentist paradigm, it is common practice to specify a residual as

$$r_{1i} = y_i - \widehat{y}_i \tag{4.9}$$

or

$$r_{2i} = r_{1i}/\sqrt{\text{var}(r_{1i})}$$

where $\widehat{y}_i$ is a fitted value under a given model. When complex spatial models are considered, it is often easier to examine residuals such as $\{r_{1i}\}$ using Monte Carlo methods. In fact it is straightforward to implement a parametric bootstrap approach to residual diagnostics for likelihood models. The simplest case is that of tract count data, where for each tract an observed count can be compared to a fitted count. In general, when Poisson likelihood models are assumed with $y_i \sim \text{Pois}\{e_i\theta_i\}$ then it is straightforward to employ a parametric bootstrap by generating a set of simulated counts $\{y_{ij}\}, j = 1, \ldots, J$, from a Poisson distribution with mean $e_i\widehat{\theta}_i$. In this way, a tractwise ranking, and hence p-value, can be computed by assessing the rank of the residual within the pooled set

$$\{y_i - e_i\widehat{\theta}_i; \{y_{ij} - e_i\widehat{\theta}_i\},\ j = 1, \ldots, J\}.$$

Denote the observed standardized residual as r_{2i} and the simulated residuals as $\{r^s_{2ij}\}$. Note that it is now possible to compare functions of the residuals as well as making direct comparisons.

The spatial distribution of residuals is also important. For example, in a spatial context, it may be appropriate to examine the spatial autocorrelation of the observed residuals. Hence, a Monte Carlo assessment of degree of residual autocorrelation could be made by comparing Moran's I statistic for the observed residuals, say, $M(\{r_{2i}\})$, to that found for the simulated count residuals $M(\{r^s_{2ij}\})$.

In the situation where case events are available, it is not straightforward to define a residual. As the data are in the form of locations, it is not possible to directly compare observed and fitted values. However, by a suitable transformation, it is possible to compare *measures* which describe the spatial distribution of the cases. A model which fits the data well should provide a good fit to the spatial distribution of the cases. It is possible to examine the difference between a local estimate of the case density, $\widehat{\lambda}(\mathbf{x}_i)$, and that predicted from a fitted model, $\widehat{\lambda}^*(\mathbf{x}_i)$, that is, at the ith location:

$$r_i = \widehat{\lambda}_i - \widehat{\lambda}^*_i, \tag{4.10}$$

where $\lambda_i \equiv \lambda(\mathbf{x}_i)$.

This approach has been proposed in the derivation of a deviance residual for modulated heterogeneous Poisson process models (Lawson, 1993). This residual can incorporate estimated expected rates. It is possible to simulate J realizations of events from the fitted model, and the local density of these realizations could be compared pointwise with $\widehat{\lambda}_i^*$. Of course, these proposals rely on a series of smoothing operations. More complex alternative procedures could be pursued.

In a Bayesian setting it is natural to consider the appropriate version of (4.9). Carlin and Louis (1996) describe a Bayesian residual as

$$r_i = y_i - \frac{1}{G}\sum_{g=1}^{G} E(y_i|\theta_i^{(g)}) \tag{4.11}$$

where $E(y_i|\theta_i)$ is the expected value from the posterior predictive distribution, and (in the context of MCMC sampling) $\{\theta_i^{(g)}\}$ is a set of parameter values sampled from the posterior distribution.

In the tract count modeling case, this residual can therefore be approximated, when a constant tract rate is assumed, by:

$$r_i = y_i - \frac{1}{G}\sum_{g=1}^{G} e_i\theta_i^{(g)}. \tag{4.12}$$

This residual averages over the posterior sample. An alternative possibility is to average the $\{\theta_i^{(g)}\}$ sample, $\widehat{\theta}_i$ say, to yield a posterior expected value of y_i, say $\widehat{y}_{\widehat{\theta}i} = e_i\widehat{\theta}_i$, and to form $r_i = y_i - \widehat{y}_{\widehat{\theta}i}$. A further possibility is simply to form r_{2i} at each iteration of a posterior sampler and to average these over the converged sample (Spiegelhalter *et al.*, 1996). These residuals can provide pointwise goodness-of-fit measures as well as global goodness-of-fit measures, and can be assessed using Monte Carlo methods. Surveillance residuals are based on these constructs (Lawson *et al.*, 2004).

Deletion residuals and residuals based on conditional predictive ordinates can also be defined for tract counts (Stern and Cressie, 2000). To further assess the distribution of residuals, it would be advantageous to be able to apply the equivalent of the parametric bootstrap in the Bayesian setting. With convergence of a MCMC sampler, it is possible to make subsamples of the converged output. If these samples are separated by a distance (h) which will guarantee approximate independence (Robert and Casella, 1999), then a set of J such samples could be used to generate $\{y_i\}j = 1, \ldots, J$, with $y_i \leftarrow Pois(e_i\widehat{\theta}_{ij})$, and the residual computed from the data r_i can be compared to the set of J residuals computed from $y_i - E(y_i)$, where $E(y_i)$ is the predictive expected value of y_i. In turn, these residuals can be used to assess functions of the residuals and goodness-of-fit measures. The choice of J will usually be 99 or 999, depending on the level of accuracy required.

In the situation where case events are examined it is also possible to derive a Bayesian residual as we can evaluate $E\{\lambda(\mathbf{x}|\theta^{(g)})\}$ based on the $\{\theta_i^{(g)}\}$ posterior samples. Hence it is possible to examine:

$$r_i = \widehat{\lambda}_i - \frac{1}{G}\sum_{g=1}^{G}\widehat{\lambda}_i^{*(g)}$$

where $\widehat{\lambda}_i^{*(g)}$ is the fitted model estimate of intensity corresponding to the gth posterior sample. Further, it is also possible with subsampling for approximate independence to use a parametric bootstrap approach to residual significance testing.

4.5 SPATIO-TEMPORAL ANALYSIS

As in other application areas, it is possible to consider the analysis of disease maps which have an associated temporal dimension (a map evolution). The sequential analysis of georeferenced data will be discussed in the following section and elsewhere in this volume. The two most common formats for observations are: georeferenced case events which have associated a time of diagnosis or registration or onset, that is, we observe, within a fixed time period T, m cases at locations $\{\boldsymbol{x}_i, t_i\}, i = 1, \ldots, m$; and counts of cases of disease within tracts are available for a sequence of T time periods that is, we observe a binning of case events within $p \times T$ space-time units $y_{it}, i = 1, \ldots, p,\ t = 1, \ldots, T$.

Figure 4.2 displays an example of space-time count data: a sequence of five biweekly standardized incidence ratio maps for parishes of Cumbria, UK, for the foot-and-mouth disease epidemic of 2001. While this is an animal epidemic example, this does provide a glimpse of the data available and the nature of space-time variation of infectious disease. Surveillance of animal populations is also important in the bioterrorism context of course.

In the case event situation, few examples exist of mapping analysis. However, it is possible to specify a model to describe the first-order intensity of the space-time process (as in the spatial case). The intensity at time t can be specified as:

$$\lambda(\mathbf{x}, t) = \rho g(\mathbf{x}, t) f_1(\mathbf{x}; \theta_x) f_2(t; \theta_t) f_3(\mathbf{x}, t; \theta_{xt}), \tag{4.13}$$

where ρ is a constant background rate (in space $\times$ time units), $g(\mathbf{x}, t)$ is a modulation function describing the spatio-temporal 'at-risk' population background in the study region, f_k are appropriately defined functions of space, time, and space-time, and $\theta_x, \theta_t, \theta_{xt}$ are parameters relating to the spatial, temporal, and spatio-temporal components of the model.

Here each component of the f_k can represent a *full* model for the component, that is, f_1 can include spatial trend, covariate, and covariance terms, and f_2 can

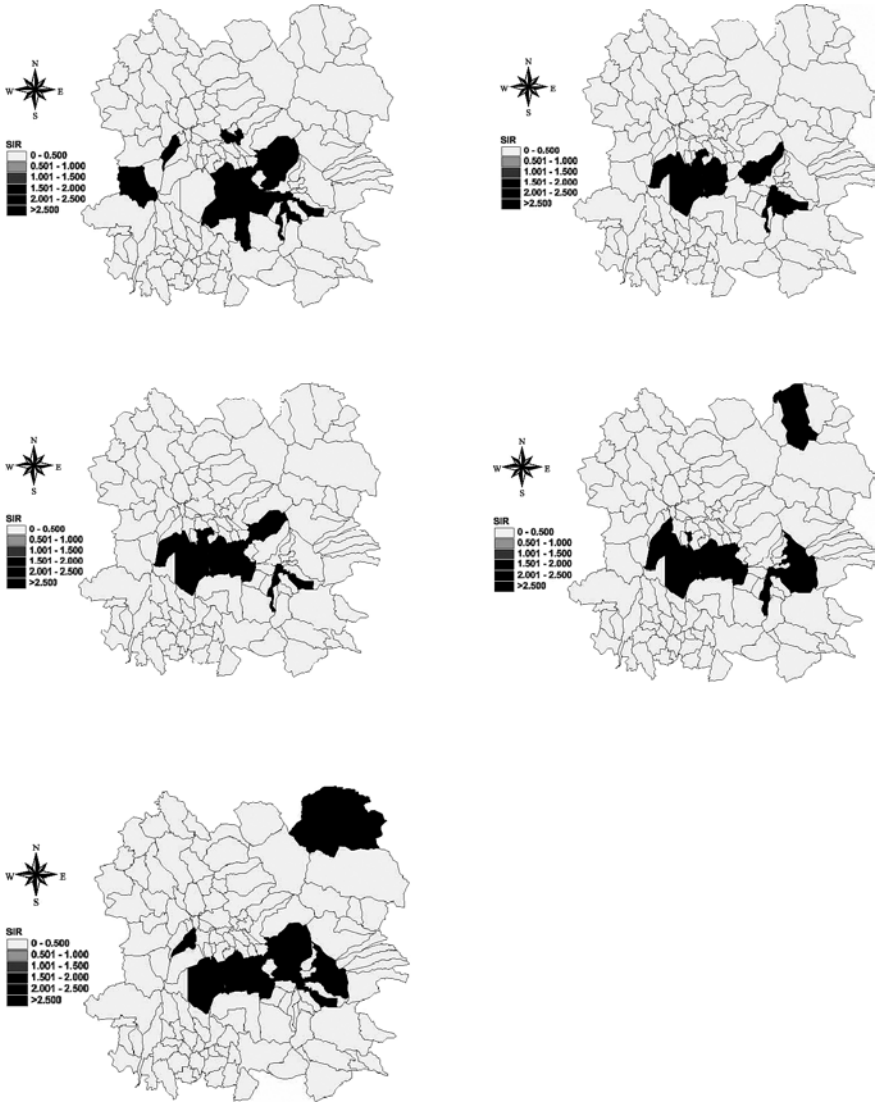


Figure 4.2 Standardized incidence ratio map sequence (five periods) for the UK foot-and-mouth disease epidemic of 2001. The sequence is rowwise from top (biweekly).

contain similar terms for the temporal effects, while f_3 can contain *interaction* terms between the components in space and time. Note that this final term can include *separate* spatial structures relating to interactions which are not included in f_1 or f_2. The exact specification of each of these components will depend on the application, but the separation of these three components is helpful in the formulation of components.

The above intensity specification can be used as a basis for the development of likelihood and Bayesian models for case events; if it can be assumed that the events form a modulated Poisson process in space-time then a likelihood can be specified as in the spatial case.

Note that the above case event intensity specification can be applied in the space-time case where small area counts are observed within fixed time periods $\{l_j\}, j = 1, \ldots, T$. In addition, the counts are independent conditional on the intensity given, and this expectation can be used within a likelihood modeling framework or within Bayesian model extensions. If a constant rate approximation is valid then it is straightforward to derive the minimal and maximal relative risk estimates under the Poisson likelihood model assuming $E\{y_{ij}\} = \lambda_{ij} = e_{ij}\theta_{ij}$, where e_{ij} is the expected rate in the required region/period. The maximal model estimate is $\widehat{\theta}_{ij} = \frac{y_{ij}}{e_{ij}}$, the space-time equivalent of the SMR, while the minimal model estimate is $\widehat{\theta} = \frac{\sum_i \sum_j y_{ij}}{\sum_i \sum_j e_{ij}}$. Smooth space-time maps of relative risk estimates will usually lie between these two extremes.

Development of count data modeling based on tract/period data has recently seen considerable development. The first example of such modeling was by Bernardinelli *et al.* (1995). In their approach, they assumed a Poisson model with $E\{y_{ij}\} = \lambda_{ij} = e_{ij}\theta_{ij}$ and log relative risk of the form

$$\log(\theta_{ij}) = \mu + \phi_i + \beta t_j + \delta_i t_j \tag{4.14}$$

where μ is an intercept (overall rate), t_j is the (suitably defined) time of the jth interval, ϕ_i is an area (tract) random effect, βt_j is a linear trend term in time t_j, and δ_i an interaction random effect between area and time. Suitable prior distributions were assumed for the parameters in this model and posterior sampling of the relevant parameters was performed via Gibbs sampling. Note that in this formulation there is no spatial trend, only a simple linear time trend and no temporal random effect. The components in (4.13) above allow a range of effects in each of the spatial and temporal components, however, and this model could be extended in a variety of directions. A variant of this model is discussed in Chapter 5 of this volume.

Waller *et al.* (1997) and Xia and Carlin (1998) (see also Carlin and Louis, 1996) subsequently proposed a different model where the log relative risk is parameterized as

$$\log(\theta_{ijkl}) = \phi_i^{(j)} + \delta_i^{(j)} + \text{fixed covariate terms } (kl)$$

where $\phi_i^{(j)}$ and $\delta_i^{(j)}$ are uncorrelated and correlated heterogeneity terms which can vary in time. This model was further developed by Xia and Carlin (1998) who also examined a smoking covariate which has associated sampling error and spatial correlation. Their model was defined as:

$$\log(\theta_{ijkl}) = \mu + \zeta t_j + \phi_{ij} + \rho p_i + \text{fixed covariate terms } (kl)$$

where an intercept term μ is included with a spatial random effect nested within time $\{\phi_{ij}\}$, a linear time trend ζt_j, and p_i is a smoking variable measured within the tract unit. In these model formulations no spatial trend is admitted and all time-based random effects are assumed to be subsumed within the ϕ_{ij} terms.

To allow for the possibility of time-dependent effects in the covariates included (race and age), Knorr-Held and Besag (1998) formulated a different model for the same data set (88 county Ohio lung cancer mortality, 1968–1988). Employing a binomial likelihood for the number at risk $\{y_{ijkl}\}$ with probability π_{ijkl}, for the counts, and using a logit link to the linear predictor, they proposed

$$\eta_{ijkl} = \ln\{\pi_{ijkl}/(1-\pi_{ijkl})\},$$

where

$$\eta_{ijkl} = \alpha_j + \beta_{kj} + \gamma_{lj} + \delta z_i + \theta_i + \phi_i. \tag{4.15}$$

Here α_j is a time-based random intercept, β_{kj} a kth age group effect at time j, γ_{lj} a gender $\times$ race effect for combination l at the jth time, δz_i a fixed covariate effect term where the z_i is an urbanization index, and θ_i, ϕ_i are correlated and uncorrelated heterogeneity terms which are not time-dependent. No time trend or spatial trend terms are used, and these effects will (partially) be subsumed within the heterogeneity terms and the $\alpha_j + \beta_{kj} + \gamma_{lj}$ terms.

More recent examples of spatio-temporal modeling include extensions of mixture models (Böhning *et al.*, 2000), which examines time periods separately without interaction, and the use of a variant of a full multivariate normal spatial prior distribution for the spatial random effects (Sun *et al.*, 2000). Other developments include the extension the Knorr-Held and Besag model to include different forms of random interaction terms (Knorr-Held 2000), and the use of covariates at different levels of aggregation (Zhu and Carlin, 2000). A recent brief review of this area in relation to fitting models is provided in Lawson *et al.* (2003).

Overall, there are a variety of forms which can be adopted for spatio-temporal parameterization of the log relative risk, and it is not yet clear which of the models so far proposed will be most generally useful. Many of the above examples exclude spatial and/or temporal trend modeling, although some examples absorb these effects within more general random effects. Allowing for temporal trend via random walk intercept prior distributions provides a relatively nonparametric approach to temporal shifting, while it is clear that covariate interactions with time should also be incorporated. Interactions between purely spatial and temporal components of the models have not been examined to any extent, and this may provide a fruitful avenue for further developments. If the goal of the analysis of spatiotemporal disease variation is to provide a parsimonious description of the relative risk variation then it would seem to be reasonable to include spatial and temporal trend components in any analysis (besides those defined via random effects).

Finally, it is relevant to note that there are many possible variants of the two basic data formats which may arise, partly due to mixtures of spatial aggregation levels, but also to changes in the temporal measurement units. For example, it may be possible that the spatial distribution of case event data is only available within fixed time periods, and so a hybrid form of analysis may be required where the evolution of case event maps is to be modeled. Equally, it may be the case that repeated measurements are made on case events over time so that attached to each case location is a covariate (possibly time-dependent) which is available over different time periods. In that case a form of spatio-longnitudinal analysis might be considered. A special case of this might be the analysis of time to endpoint for georeferenced cases of disease (e.g. death/recovery/remission). This could be regarded as a spatial survival analysis (Banerjee *et al.*, 2003).

4.6 SURVEILLANCE ISSUES

The above comments concerning spatio-temporal (ST) modeling carry over to surveillance. Most ST models have been developed for retrospective analysis of complete data sets. However, a fundamental characteristic of surveillance is that it is carried out on-line within real time or near-real time and an emphasis is placed on detection of changes. Hence, although a good ST model may be useful, there are many new issues relating to model fitting that should be considered. A brief list of these is as follows:

(1) At a new monitoring point in time, an assessment must be made as to whether the process has changed (see Chapter 10 of this volume).

(2) Changes are to be detected beyond the 'normal' ST behavior of the disease or diseases.

(3) Multiple diseases may need to be monitored.

(4) As time progresses the data set enlarges, and the parameter space can also enlarge.

(5) As time progresses the model assumed for the 'normal' ST behavior may deteriorate.

The first and second points require the monitoring of change beyond background 'normal' variation. Hence, a well-designed ST model should include the 'normal' variation but must also be capable of allowing detection of changes. This suggests that the model should be flexible enough to capture normal historical variation but also should not 'model out' changes. A model that is too sophisticated may absorb the changes in the model fit, and so a balance must be struck. The second issue relates to the need for multivariate and syndromic surveillance as often there will be no indication about which disease is to be targeted with any insults. The previous discussion relates only to single diseases, although there is recent work on sophisticated models for multiple diseases

(Carlin and Banerjee, 2003). Of course, single diseases can be monitored in parallel, but this ignores correlation between the disease incidences that may contain important clues for the detection of early changes.

The fourth and fifth points relate to difficulties in refitting models with enlarging parameter and data spaces and also the lack of fit which could develop over time. Often complex spatial models are fitted using computationally expensive simulation methods. Over time these models will have to be refitted to new larger data sets with enlarged parameter sets. This could lead to computational problems. Sliding windows have been proposed to allow for data reduction, but these also lose information about distant historical data. Filtration can also be used. Progressive lack of fit of a model is a major problem as model readjustment could reduce the chance of detecting new events. There is no simple satisfactory answer for this problem (see Lawson, 2004).

5

Generalized Linear Models and Generalized Linear Mixed Models for Small-Area Surveillance

Ken Kleinman

5.1 INTRODUCTION

As outlined in Chapter 1, we are interested in performing surveillance, as a practical matter, when spatial data are available. Without such data, one of the methods outlined in Chapter 2 and evaluated in Chapter 3 would be employed. In addition, it may be valuable to use such methods on a summary statistic (such as the total count) even when spatial information is available, since they may have power against alternatives for which spatial methods are not especially sensitive. Several methods have been proposed for this type of surveillance. For example, scan statistic methods are summarized in Chapter 7 and spatial versions of the CUSUM approach are discussed in Chapter 6.

In Kleinman *et al.* (2004a) we introduced the concept of using generalized linear mixed models for surveillance when the location of each case and of potential cases was available to within a small area. We have since dubbed this the 'SMART scores' (Small Area Regression and Testing scores) approach. That paper also reviews some of the literature of spatio-temporal modeling, and its usefulness for large-scale surveillance. In essence, the extremely large size of spatio-temporal surveillance data sets precludes some complex models; the additional complication of reliable repeated fitting, rather than exploratory one-off models, also suggests simpler approaches may be particularly valuable in this context.

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

In brief, Kleinman *et al.* (2004a) treat each small area as if it were an individual, and fit a random effect to account for the repeated counts for each area. This allows variability in the baseline risk of a case in each small area. We presented this approach using a logistic regression model. In this chapter, we extend the example provided in Kleinman *et al.* to use Poisson regression; this allows models that can be fitted without knowing the number of people at risk, as demonstrated below. We also evaluate the assertions in Kleinman *et al.* (2004a, 2004b) regarding the utility of the random effects, relative to the fixed effects versions of the models. Finally, we also examine the impact of modeling the data monthly, as opposed to daily. We do this using an example data set.

We use y_{st} to denote the number of cases observed in a small area s at time t. We assume here that time is discrete. To concretize, we will refer to areas s as census tracts or tracts, and to times t as days. Data available in addition to y_{st} include n_{st}, the maximum number of cases possible in tract s on day t, and $c_{kt}, k = 1, \ldots, K$, covariates describing day t, such as day of the week, days since surveillance started, or trigonometric functions of the day of the year. We refer to n_{st} as the number of people eligible for surveillance. It is possible that in practice covariates describing the cases and individuals eligible for surveillance, such as age and gender, may be available. However, as discussed in Kleinman *et al.* (2004a, 2004b), these may make model estimation impractical and actually obtaining surveillance data of this sort may introduce privacy issues. Therefore, for this discussion, we assume that no covariate data of this sort will be used.

5.2 SURVEILLANCE USING SMALL-AREA MODELING

We will model the counts $y_{st}, s = 1, \ldots, S, t = 1, \ldots, T$, in some historical period ending on day T. We will use parameter estimates from these models to estimate the distributional parameters of $y_{s,T+r}, s = 1, \ldots, S, r > 0$, for some future day $T + r$. Using these estimated parameters, we will find the probability of seeing as many cases as were seen, or more, assuming the distribution and parameter are accurate. For data of this sort, Kleinman *et al.* (2004a) propose a generalized linear mixed model (Breslow and Clayton, 1993). Another approach to small areas involves conditional autoregressive models (Lawson *et al.*, 2003).

5.2.1 Example

For example, consider the method proposed in Kleinman *et al.* (2004a). There, we suggested a logit link on the probability of a case in each tract:

$$\text{logit}(pr_{st}) = \frac{pr_{st}}{1 - pr_{st}} = \beta_0 + \sum_{i=1}^{K} c_{it}\beta_i + b_s, \tag{5.1}$$

where pr_{st} is the probability of a case and b_s, the random effect, is assumed normally distributed with mean 0. We note here that if it seems desirable to incorporate spatial correlation into the model, it may be conveniently included via the variance–covariance matrix of the bs. Here, we assume that this is equal to $\sigma_b^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Thus, in some sense, the model and approach in this case is essentially nonspatial – the relative locations of the small areas do not enter into the model. For comments on this, see Chapters 2 and 9. For the remainder of the chapter, we refer to model (5.1) as the random effects logistic regression model.

After obtaining estimates $\hat{\beta}$ and $\tilde{b}_s$, we insert covariate values $c_{k,T+r}$ into the regression equation, and invert the logit to get

$$\hat{pr}_{s,T+r} = \frac{e^{\hat{\beta}_0 + \sum c_{k,T+r}\hat{\beta}_i + \tilde{b}_s}}{1 + e^{\hat{\beta}_0 + \sum c_{k,T+r}\hat{\beta}_i + \tilde{b}_s}}$$

for each tract s. After collecting surveillance data for day $T+r$, we then calculate the probability that a binomial random variable with parameters $(n_{s,T+r}, \hat{pr}_{s,T+r})$ has a value greater than or equal to $y_{s,T+r}$;

$$\begin{aligned}&\text{Prob}(Y \geq y_{s,T+r} | n_{s,T+r}, \hat{pr}_{s,T+r}) \\ &= \tilde{p}_{s,T+r} = 1 - \sum_{i=0}^{y_{s,T+r}-1} \binom{n_{s,T+r}}{i} (\hat{pr}_{s,T+r})^i (1 - \hat{pr}_{s,T+r})^{n_{s,T+r}-i}.\end{aligned}$$

5.2.2 Using the Model Results

The ordinary use of these probabilities is as p-values assessing the null hypothesis that the data were drawn from binomial distributions with the parameters above. A rejection of the null should be taken to mean that the observed cases may not have a natural origin, and certainly that so many cases appearing by chance is small, assuming the null is true. However, since S may be large, there is a reasonable concern about multiple testing. To address this problem, we suggest reporting the 'recurrence interval' of each p-value, calculated as the expected number of days of surveillance required so that exactly one p-value as small as the one observed would be expected. This is simply

$$(\tilde{p}_{s,T+r} \times S)^{-1}, \tag{5.2}$$

though note that this estimate may be conservative (Waller *et al.*, 1994).

One nice feature of the recurrence interval is with respect to classical inference. Ordinarily, one must perform the mental acrobatics of imagining repeating the same experiment a large number of times and thinking of the p-value roughly as the proportion of times the results would be as extreme as the one you observed in the one experiment you actually did. In contrast, in the

surveillance setting, you are in the position of actually repeating approximately the same experiment many times, meaning each day. The recurrence interval then is the number of such experiments one would have to perform in order to expect to see a result as extreme as or more extreme than the one you happened to see today. The mental movement required is simply to think of some period of days, and imagine only seeing results like these on one occasion.

It also seems that interpretation of small *p*-values is much simpler through the recurrence interval. For instance, *p*-values of 0.001 and 0.0001 both seem rather small, and similar in that regard. In contrast, the corresponding recurrence intervals (assuming S=1) of 100 days (0.27 years) and 1000 days (2.74 years) seem very different; one might react very differently to the two recurrence intervals. The advantage is compounded for the modeling approach discussed here; $\tilde{p}$s of 0.000 02 and 0.000 002 correspond to recurrence intervals of 0.27 and 2.74 years if S=500.

In practice, consumers of public health surveillance tend to think in terms of the number of alarms per month or per year, and this standpoint is easy to meet with the recurrence interval approach as well. A final boon of this way of thinking is that any *p*-value can be presented this way. Thus, a complex approach to prospective surveillance, such as that presented in Kulldorff (2001) is not necessary; any spatial clustering test can be performed repeatedly; and the recurrence interval presentation protects readers from forgetting that a *p*-value of 0.05 or smaller should be expected in any 20 (0.05^{-1}, since S=1) day period, under the null hypothesis.

5.3 ALTERNATE MODEL FORMULATIONS

5.3.1 Fixed Effects Logistic Regression

The fixed effects logistic regression approach is the simpler expression of model (5.1), with fixed rather than random effects for the small-area base rates. In theory, estimators incorporating the estimated fixed effects are unbiased but may be expected to have larger standard errors than estimators incorporating the estimated random effects estimates, which are biased or 'shrunken'. The model may be written as

$$\text{logit}(pr_{st}) = \frac{pr_{st}}{1-pr_{st}} = \sum c_{it}\beta_i + \theta_s, \tag{5.3}$$

where θ_s denotes an additional fixed effect for each area *s*.

Analysis proceeds as in model (5.1). Note that differences between analysis via model (5.1) and model (5.3) should mainly be due to differential amounts of information contributed by each area *s*.

5.3.2 Poisson Regression Models

Assuming a Poisson regression for the counts has a number of extremely attractive features. First, it will be possible to fit the model without knowing the number of people eligible for surveillance, opening additional data sources to the possibility of analysis. This is impossible with the binomial formulation. Second, the basic probability result that the sum of Poisson variates is Poisson distributed with parameter equal to the sum of the constituent parameters can be leveraged to perform multi-day and/or multi-area surveillance. This is also not possible in the binomial formulation, assuming that there are meaningful day-to-day changes in pr_{st}. These advantages are laid out more explicitly below in Section 5.3.2.3. Third, the bias involved with fitting logistic regressions with very small proportions of events can be avoided. However, these advantages are purchased at cost of depending on the accuracy of the Poisson approximation to the binomial in a case where typically $NP \ll 5$.

5.3.2.1 Fixed effects Poisson regression

This model is represented as

$$\log(\mathrm{E}(y_{st})) = \sum c_{it}\beta_i + \theta_s, \tag{5.4}$$

where all terms on the right-hand side are as defined in model (5.3).

Analysis of the surveillance data proceeds as in model (5.1) except that the p-value is calculated from the probability under the Poisson distribution with parameter $\hat{\lambda}_{s,T+r} = e^{\sum c_{i,T+r}\hat{\beta}_i + \hat{\theta}_s}$ that $Y \geq y_{s,T+r}$. This is

$$\tilde{p}_{s,T+r} = 1 - \sum_{i=0}^{y_{s,T+r}-1} \frac{(e^{-\hat{\lambda}_{s,T+r}})(\hat{\lambda}_{s,T+r})^i}{i!}.$$

It is common in Poisson regression applications to include an 'offset' when, for example, differential follow-up times obtain for the individuals. The offset is simply a covariate, often the log of time, with coefficient constrained to be 1. In data that are naturally binomial in nature, the offset should be based on the number of trials. In the present application, however, the number of cases is typically so small relative to n_{st} that including $\log(n_{st})$ as the offset results in fitting problems. Instead, we would have to consider models with $\log(\log(n_{st}))$ as the offset. In addition, note that the change in n_{st} over time is usually small in this type of application, relative to the average daily eligible subjects $\bar{n}_s$. This may lead to nonidentifiability problems with respect to the θ_s in (5.4). This occurs because, since an intercept is omitted and a fixed effect is estimated for every area s, the θ, as a group, are almost collinear with a constant; if $n_{st} = n_s$ for each s, it is such a constant. Finally, incorporating n_{st} into the model removes

one key advantage of the Poisson approach relative to the binomial approach by requiring the number eligible for surveillance to be known. We thus consider only model (5.4) as the fixed effects Poisson regression approach, and will not consider Poisson regression with an offset.

5.3.2.2 *Random effects Poisson regression*

Here we allow random effects for each tract, while maintaining the Poisson assumption for the counts:

$$\log(\mathrm{E}(y_{st})) = \beta_0 + \sum c_{it}\beta_i + b_s \tag{5.5}$$

where $b_s \sim N(0, \sigma_b^2)$ and with $\hat{\lambda}_{s,T+r} = e^{\hat{\beta}_0 + \sum c_{i,T+r}\hat{\beta}_i + \tilde{b}_s}$. As with the fixed effects Poisson approach, we do not consider model (5.5) with the inclusion of an offset.

5.3.2.3 *Multi-day surveillance*

To concretize the notion of multi-day surveillance, suppose the event we were surveilling for had a variable onset. Then the most powerful surveillance would not be to check each day separately, but to check a series of days simultaneously for an excess of cases. To evaluate three future days for a cumulative deviation from the expectation, we would simply compare $y_{s,T+r} + y_{s,T+r+1} + y_{s,T+r+2}$ to a Poisson distribution with parameter $\hat{\lambda}_{s,T+r} + \hat{\lambda}_{s,T+r+1} + \hat{\lambda}_{s,T+r+2}$, where $\hat{\lambda}_{st}$ is as defined in Section 5.3.2.1 or the equivalent based on model (5.5).

If a fixed number of days is identified as the ideal and only surveillance summary period, the recurrence interval can be calculated as in Section 5.2.2. If the number of days is unknown, then surveillance of different lengths can be incorporated accurately by multiplying the denominator in the recurrence interval calculation (equation (5.2)) by the number of different lengths that will be considered. So if we wanted to surveil for increases over one, two, or three days, the proper recurrence interval would be $(S \times 3 \times \tilde{p}_{st'})^{-1}$, where t' implies any of the p-values based on one, two, or three days.

5.4 PRACTICAL VARIATIONS

In applying these models, some consideration of real applications is important. One example is that while there is no theoretical problem in fitting the model each day and using the results for evaluation of tomorrow's observations, this is not so simple in practice. Data sets in this context are typically large and cumbersome, making repeated analysis time-consuming. In addition, as discussed in Chapter 4, application of too complex a model too often may result

in incorporating additional cases from an outbreak into the model and thus masking the appearance of future outbreak-related cases. Automating the model fitting is a possibility, but is not recommended, since some models are complex enough that fitting may fail somewhat frequently. Even getting the data from a collection point to the desk of an analyst may take several hours.

For all of these reasons, it is worthwhile to consider how much the fit is improved with each day's data. If the model could be refitted, say, monthly, then the time involved in moving the data around daily and in daily analysis could be avoided. In addition, this would allow model fitting at some central location; lookup tables with the p-value associated with a range of counts could be easily produced and could be used with or without computers to evaluate the surveillance data at the site of data collection. We will compare the results achieved using daily model fitting with those seen using monthly fitting. Symbolically, this is a question of the difference between surveillance when $r = 1$ for all days or $r = 1, \ldots, 31$ depending on the number of days since the model was fitted.

5.5 DATA

We will evaluate the performance of models (5.1) and (5.3)–(5.5) in a typical surveillance case, described in some detail here. In Lazarus *et al.* (2002) we described the surveillance system that uses the automated electronic medical record system at Harvard Vanguard Medical Associates (HVMA), currently a multi-specialty practice group with 14 clinics in the greater Boston, Massachusetts, area. During the initial period of data collection described below, HVMA was the staff model division of Harvard Pilgrim Health Care (HPHC). Briefly, the data set represents the ambulatory medical encounters of a dynamic population of approximately 250 000 individuals, representing about 10 % of the population in a region of eastern Massachusetts.

At each office visit, the clinician entered diagnoses, to which International Classification of Disease, 9th Revision, Clinical Modification (ICD9) codes were attached. The computerized record was available for data analysis within a day. The data set used for this example incorporates all such visits between January 1, 1997 and December 31, 1999.

In addition, a linked database includes information on all eligible individuals; this database includes the patients' billing addresses, ages, genders, and their dates of eligibility for care. Billing addresses are geocoded, providing the exact latitude and longitude, as well as the census block group and census tract.

Census block groups typically are constructed to have 1000 residents; census tracts are more populous, with approximately 4000 residents. Census regions are generally to be preferred in surveillance, since the census draws them for homogeneity as well as roughly consistent population size. In contrast, zip codes

are drawn solely for the convenience of the Postal Service and are not regular in size, population, or any other known characteristics. In addition, they change often and unpredictably. Their only desirable property for surveillance is their ubiquity.

5.5.1 Developing and Defining Syndromes

Encounters are categorized into syndrome groups by examining all of the ICD9 codes assigned at the time of consultation. The surveillance software considers each encounter record in turn and merges related ICD9 diagnosis codes into syndrome groups using a modification of a provisional classification scheme developed as part of the Department of Defense ESSENCE project (provided to us by J. Pavlin, Department of Defense Global Emerging Infections System). This scheme reduces the complexity of the ICD9 into eight syndrome categories – coma/shock, neurological, upper gastrointestinal, lower gastrointestinal, upper respiratory, lower respiratory, dermatological, and sepsis/fever.

As an example, we consider lower respiratory infection (LRI). For more complete information on our definition of this syndrome, see Lazarus *et al.* (2001). Briefly, the syndrome incorporates 119 ICD9 codes. These codes include influenza, pneumonia, bronchitis, and cough; incidence rates are much higher in the winter than in the summer. Spatial clusters as well as temporal clusters are expected to occur naturally, because of the contagious nature of illnesses that contribute a large proportion of the visits associated with this syndrome.

LRI is of particular interest, because one bioterrorism agent it is designed to detect is anthrax. Typically, inhalational anthrax begins with a nonspecific prodromal phase in which the sufferer may experience fever, dyspnea, cough, and chest discomfort. (Intestinal and dermal anthrax are both less lethal and less acute and are therefore of little interest in this application.) During this phase neither physical examination nor any widely used diagnostic test will suggest an unusual illness. Diagnosis usually occurs after two to four days, when respiratory failure and hemodynamic collapse may ensue. By that time chest X-rays show an unusual pattern of mediastinal widening (MMWR, 2001). It is hoped that a victim of an anthrax release would receive a diagnosis in the LRI category if they visited their health care provider during the prodromal phase.

An individual patient may have multiple encounters associated with a single episode of illness (for example: initial consultation, consultation one or two days later for laboratory results, follow-up consultation a few weeks later, and so on). In order to avoid double counting from this common pattern of ambulatory care, the first encounter for each patient within any single syndrome group is reported, but subsequent encounters with the same syndrome are not reported as new episodes until six weeks or more have elapsed since the most recent encounter in the same syndrome. (On the other hand, it may be the case that there is information relevant to bioterrorism in repeat visits; this approach

ignores that information). We have reported previously that grouping of respiratory illness visits into episodes reduces the total number of events by 38% in this clinical setting.

Between January 1, 1997 and December 31, 1999, there were 133 853 lower respiratory infection episodes for which it was possible to determine that patient's residence was in one of the 565 census tracts with centroids in the greater Boston area between longitude 70.85° and 71.4° W and latitude 42.15° and 42.66° N.

5.6 EVALUATION

We emulated one year of surveillance beginning on January 1, 1999. In other words, we treated each day of 1999, sequentially, as if it were the day of surveillance, with no later data being known during the surveillance of that day. In all, we will analyze the data in five ways. These include the four models described in equations (5.1) and (5.3)–(5.5), fitted on a monthly basis, plus one model fitted each day. In the monthly models, we fit the models as if during the first day of each month, including all data from January 1, 1997, through the end of the previous month.

5.6.1 Fixed and Random Effects Monthly Models

We fitted 12 models each of the type described in models (5.1) and (5.3)–(5.5), each time increasing the value of T to include the last day of the previous month. Covariates describing day t included six indicators for the day of week, indicators of holiday or day after holiday, sine and cosine functions of the day of the year, and a linear secular time trend. Each of these covariates is necessary in this context: ambulatory visits are most common on Mondays, decrease during the week through Thursday, rise again on Friday, and are scarce on weekends. Visits are especially rare on holidays, but the day following a holiday typically has more visits than the same day of the week not following a holiday. (From this we might infer that even visits for current illnesses tend to be scheduled by the patient so as not to interfere with social obligations, but that is beyond our scope here.) Similarly, data exhibit pronounced seasonal trends, as shown in Lazarus *et al.* (2001), necessitating the trigonometric functions, and secular time trends as well.

After fitting the models, we calculated the recurrence intervals associated with each tract on each day of surveillance. We tabulate the 10 most unusual days as identified by each model in Table 5.1.

Table 5.1 shows one look at the ultimate effect of using the different models. The table shows the census tract identifier, date, and count for the 10 tract days with the largest recurrence intervals. For each of the four methods, the rank of the recurrence interval and the the recurrence interval itself are reported.

Table 5.1 Rank and recurrence interval (RI) in millions of days for the 10 largest RI among each of four models.

			Model							
			(5.5)		(5.4)		(5.1)		(5.3)	
Tract	Date	Count	Rank	RI	Rank	RI	Rank	RI	Rank	RI
25025101001	2/8	8	1	465.0	1	446.0	1	420.0	1	423.0
25025091400	5/24	5	2	76.4	2	77.2	2	91.1	2	85.8
25025110401	5/3	6	3	53.2	3	51.8	3	61.9	3	51.9
25017317100	2/17	8	5	27.4	5	26.5	4	38.3	4	39.3
25025140102	10/15	6	4	28.7	4	28.0	5	29.8	5	28.8
25025092000	3/1	6	6	13.1	6	12.8	6	15.6	6	14.2
25017340100	12/30	6	8	0.998	8	0.975	7	1.61	7	1.44
25021419700	2/24	5	7	1.30	7	1.29	8	1.04	8	1.02
25009254301	3/11	4	9	0.638	9	0.698	9	0.605	9	0.904
25017350200	5/24	5	10	0.517	10	0.505	10	0.597	10	0.568

While the differences between the recurrence intervals assigned by each of the models may appear large, note that all 10 tract-days that appear in the table have recurrence intervals greater than 365 000 days, or 10 000 years, under each of the models.

In practice, we find that public health authorities do not interpret the recurrence intervals as continuous values, but trigger alarms of increasing intensity if the recurrence interval exceeds increasing thresholds. One set of such alarm thresholds is 14, 30, 60, 180, 365, 730, and 1825 days. Under any reasonable set of alarm thresholds all of the events listed in Table 5.1 will generate the highest possible degree of alarm.

An alternative way to look at this would be to consider not just the most extreme events, but to consider all tract-day results. We can do this by examining scatterplots of the recurrence intervals, as shown in Figures 5.1–5.4; we use the log of the recurrence interval for visual clarity and provide a reference line for equal values. The images show a remarkable consistency of results from

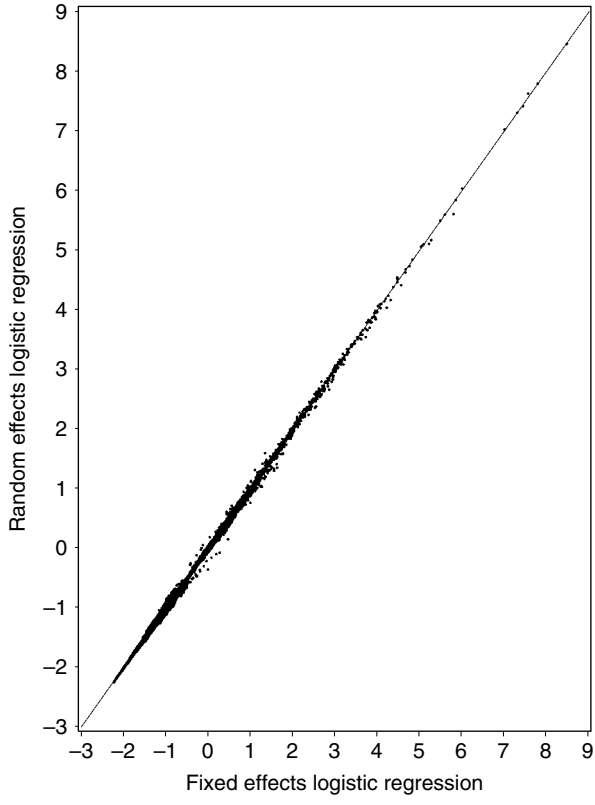


Figure 5.1 Pairwise scatterplots comparing recurrence intervals (on the $\log_{10}$ scale) from models (5.1) and (5.3).

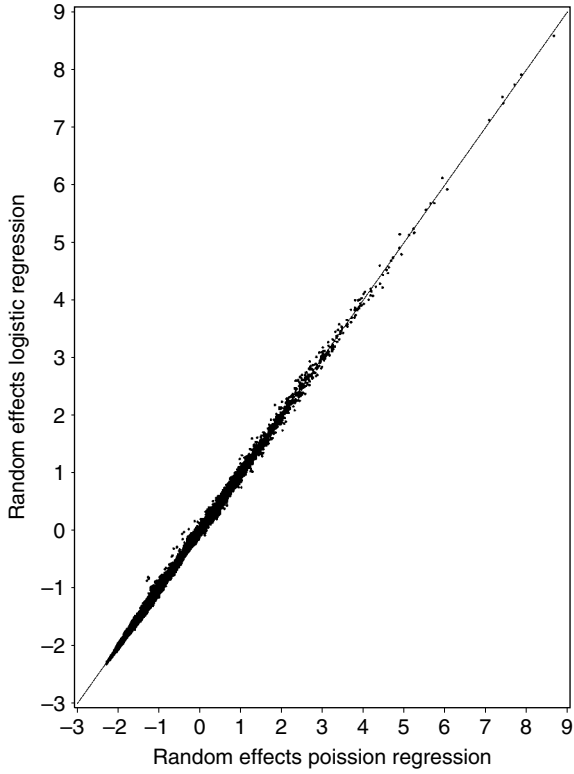


Figure 5.2 Pairwise scatterplots comparing recurrence intervals (on the $\log_{10}$ scale) from models (5.1) and (5.5).

the four models. The fixed effect and random effect methods agree well for both the Poisson (Figure 5.4, Pearson correlation = 0.999 96 for untransformed values) and logistic (Figure 5.1, Pearson correlation = 0.999 63) regression models. Similarly, the Poisson and logistic fixed effects (Figure 5.3, Pearson correlation = 0.999 04) and random effects (Figure 5.2, Pearson correlation = 0.997 75) models also agree extremely well. Even the cross-distribution, cross-method correlations are quite high; models (5.1) and (5.4) correlate at 0.998 20 and models (5.5) and (5.3) correlate at 0.998 75. The associations are as linear as those shown in Figures 5.1–5.4. Tract-days with counts of 0 are omitted, as they are constrained to be equal.

Instead, one might consider the alarm threshold approach to evaluation. Tables 5.2–5.5 show the agreement between the pairs of models shown in Figures 5.1–5.4. Each table shows the cross-classification of each tract-day based on the category of recurrence interval assigned under the various models. For example, in Table 5.2, we see that there were 14 tract-days that had recurrence intervals between 14 and 30 days under model (5.3) but 0 to 14 days under

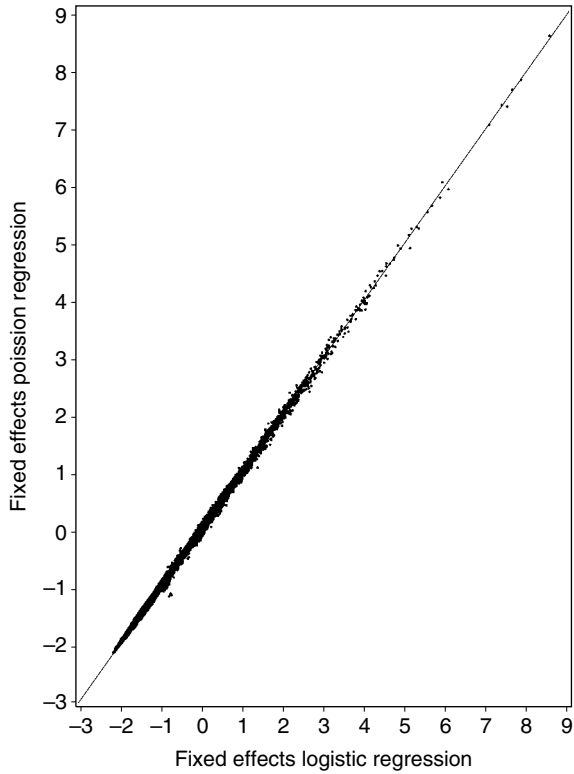


Figure 5.3 Pairwise scatterplots comparing recurrence intervals (on the $\log_{10}$ scale) from models (5.3) and (5.4).

model (5.1), while 246 tract-days had recurrence intervals between 14 and 30 days under both models. Again, the degree of consistency is remarkable. In each table, the percent of exact agreement is at least 99.9 %, and only one tract-day recurrence interval is more than one alarm category different across all these pairings of methods. This occurs when comparing the two random effects methods, which would appear to be the least similar theoretically as well as in this exploration.

The large number of tract-days with small recurrence intervals reflects mostly the 167 885 tract-days on which no cases were observed. We might consider these to be irrelevant to the amount of agreement between methods. In that case, a better sense of the agreement might be found by omitting the cell where both methods assign a recurrence interval of 14 days or less. With that restriction, there is 90.1 % agreement between the two logistic models (Table 5.2), 79.7 % agreement between the two random effects models (Table 5.3), 83.2 % agreement between the two fixed effects models (Table 5.4), and 97.4 % agreement between the two Poisson models (Table 5.5).

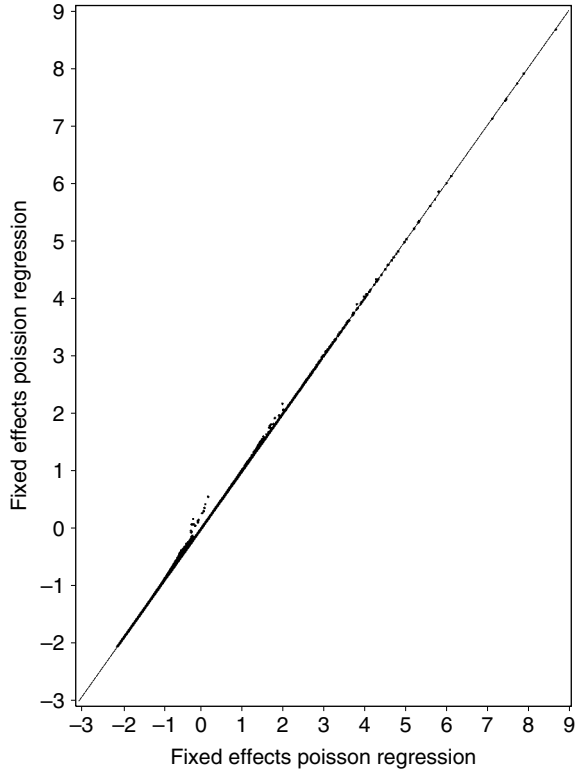


Figure 5.4 Pairwise scatterplots comparing recurrence intervals (on the $\log_{10}$ scale) from models (5.4) and (5.5).

Table 5.2 Recurrence interval (RI) in days, categorized into alarm levels; random effects logistic regression (model (5.1)) vs. fixed effects logistic regression (model (5.3)).

Random effects logistic regression	Fixed effects logistic regression							
	0–14	14–30	30–60	60–180	180–365	365–730	730–1825	1825+
0–14	204 151	14	0	0	0	0	0	0
14–30	12	246	14	0	0	0	0	0
30–60	0	16	159	9	0	0	0	0
60–180	0	0	5	180	3	0	0	0
180–365	0	0	0	5	90	8	0	0
365–730	0	0	0	0	2	55	4	0
730–1825	0	0	0	0	0	2	59	0
1825+	0	0	0	0	0	0	3	93

Table 5.3 Recurrence interval (RI) in days, categorized into alarm levels; random effects logistic regression (model (5.1)) vs. random effects poisson regression (model (5.5)).

Random effects logistic regression	Random effects poisson regression							
	0–14	14–30	30–60	60–180	180–365	365–730	730–1825	1825+
0–14	204 150	15	0	0	0	0	0	0
14–30	40	220	12	0	0	0	0	0
30–60	0	31	140	13	0	0	0	0
60–180	0	1	20	164	3	0	0	0
180–365	0	0	0	18	77	8	0	0
365–730	0	0	0	0	11	46	4	0
730–1825	0	0	0	0	0	13	43	5
1825+	0	0	0	0	0	0	5	91

Table 5.4 Recurrence interval (RI) in days, categorized into alarm levels; fixed effects Poisson regression (model (5.4)) vs. fixed effects logistic regression (model (5.3)).

Fixed effects poisson regression	Fixed effects logistic regression							
	0–14	14–30	30–60	60–180	180–365	365–730	730–1825	1825+
0–14	204 155	33	0	0	0	0	0	0
14–30	8	232	25	0	0	0	0	0
30–60	0	11	142	23	0	0	0	0
60–180	0	0	11	170	15	0	0	0
180–365	0	0	0	1	78	12	0	0
365–730	0	0	0	0	2	52	11	0
730–1825	0	0	0	0	0	1	48	4
1825+	0	0	0	0	0	0	7	89

Table 5.5 Recurrence interval (RI) in days, categorized into alarm levels; fixed effects Poisson regression (model (5.4)) vs. random effects Poisson regression (model (5.5)).

Fixed effects poisson regression	Random effects poisson regression							
	0–14	14–30	30–60	60–180	180–365	365–730	730–1825	1825+
0–14	204 186	2	0	0	0	0	0	0
14–30	4	259	2	0	0	0	0	0
30–60	0	6	167	3	0	0	0	0
60–180	0	0	3	192	1	0	0	0
180–365	0	0	0	0	90	1	0	0
365–730	0	0	0	0	0	65	0	0
730–1825	0	0	0	0	0	1	51	1
1825+	0	0	0	0	0	0	1	95

5.6.2 Daily Versus Monthly Modeling

Here we compare the daily model ($r = 1$ for all days) to the monthly model ($r = 1, \ldots, 31$). Having concluded that Poisson and logistic models as well as fixed effects and random effects models perform approximately the same way, we will use fixed effects Poisson models in this section as they are the most attractive models, as described in Section 5.3.2, as well as the least time-consuming to fit. Note that the experiment is designed so that the values must be exactly equal on approximately one-thirtieth of the days – the first of each month, when both approaches will use data from the beginning of the data set through the previous day.

As in Section 5.6.1, we show a scatterplot of the recurrence intervals obtained with each method for each tract-day (Figure 5.5) as well as tabulating the effect on the various alarm thresholds (Table 5.6). The scatterplot reflects a correlation of 0.999 91. The proportion of complete agreement, as before, is greater than 99.9 %. Omitting the cases where both methods agree that the

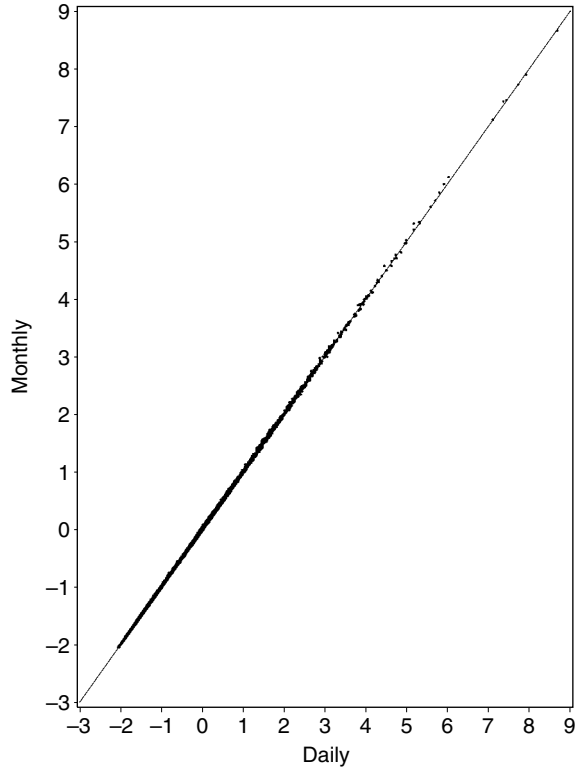


Figure 5.5 Pairwise scatterplots comparing recurrence intervals (on the $\log_{10}$ scale) from model (5.4) fitted once each month vs. daily.

Table 5.6 Recurrence interval (RI) in days, categorized into alarm levels; fixed effects Poisson regression (model (5.4)) fitted once each month vs. daily.

Monthly models	Daily models							
	0–14	14–30	30–60	60–180	180–365	365–730	730–1825	1825+
0–14	204 184	4	0	0	0	0	0	0
14–30	6	257	2	0	0	0	0	0
30–60	0	7	164	5	0	0	0	0
60–180	0	0	5	189	2	0	0	0
180–365	0	0	0	6	85	0	0	0
365–730	0	0	0	0	1	64	0	0
730–1825	0	0	0	0	0	2	49	2
1825+	0	0	0	0	0	0	3	93

recurrence interval is at or below 14 days, the agreement is at 95.2 %. There is no tract-day for which the two methods disagree by more than one alarm category.

5.7 CONCLUSION

In this chapter, I have reviewed the use of generalized linear models and generalized linear mixed models for use in the surveillance of counts from small areas. In particular, I have reviewed the utility of the Poisson model with respect to multi-day and multi-area modeling. I have compared results from fixed and random effect models based on binomial and Poisson distributional assumptions and also examined the importance of fitting the models each day, as compared to less frequent modeling.

The main results were that the Poisson models result in very similar recurrence intervals to the logistic models in data sets like ours – meaning that in our practice it matters little which method is used. Thus the Poisson model, which does not require knowledge of the number of people at risk and which allows multi-day surveillance, can be used for surveillance with little different effect in the applications where either model is suitable. Similarly, the fixed effects models generated results quite similar to those from random effects models in these data. Thus, in this sort of application, the 'strength-borrowing' features of the random effects seem to make little difference. This means that the more stable, more computationally efficient, and fully maximum likelihood fittable fixed effects models may be applied, instead of the generalized linear mixed model, which with this quantity of data is cumbersome and requires approximate and biased (McCulloch and Searle, 2001) fitting methods. Finally, the practical question of whether there is much lost by fitting the model monthly appears also to have the happy answer that less frequent fitting will generate values that closely approximate those that could be obtained with more frequent fitting.

Generalized linear models hold out some prospect of using spatial information, be it as coarse as zip codes or as fine as census block groups, to public health officials who may have previously used only gross time-series or CUSUM methods. This spatial information may well allow detection of events that would go undetected by methods that sum across small areas for want of statistical techniques. The models discussed here are simple and easily applied using commercial statistical software. Until other methods discussed in this book can be developed further (and made simpler to apply in an automated way), these models may serve as a first step toward using spatial data in practical surveillance.

6

Spatial Surveillance and Cumulative Sum Methods

Peter A. Rogerson

6.1 INTRODUCTION

The methods of statistical process control have a long history of application to problems in public health surveillance. Hill *et al.* (1968) and Weatherall and Haskey (1976) were among the first to propose and implement such systems, with their applications to the surveillance of malformations. Barbujani (1987) provides a review of these methods, with particular emphasis on applications to monitoring birth defects. Farrington and Beale (1998) and Sonneson and Bock (2003) provide more recent and more general reviews of statistical surveillance in public health.

Methods of statistical process control (see, for example, Montgomery, 1996; Hawkins and Olwell, 1998, Chapters 2 and 3) include, but are not limited to, Shewhart charts, cumulative sum (CUSUM) methods, and exponentially weighted moving average methods. Shewhart charts are designed to detect large deviations from the mean of a process; single, outlying observations can trigger an alarm or signal that the process mean may have changed. CUSUM methods maintain a running total of the deviations between observed and expected values; if this total exceeds a predetermined threshold, an alarm is sounded, again indicating a potential change in the underlying mean of the process. In this chapter, we give some attention to Shewhart charts, but focus primarily on the use of CUSUM methods. We will pay particular attention to issues that arise when there is a desire to carry out surveillance in a multiregional setting.

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

In the next section, I first review and illustrate the fundamentals of CUSUM methods. The discussion is general, and is initially oriented toward variables that come from normal distributions. Public health data are often not normally distributed – especially small counts that are collected at frequent intervals, and data associated with uncommon diseases. Later in the section and chapter, both transformations to normality and methods designed to handle directly data that are other than normally distributed are covered. I also point out in the next section several developments associated with CUSUM methods that have not been widely used in a public health context, and may ultimately prove to be of value. More specific treatment of temporal surveillance is covered elsewhere in this book (see Chapter 2). In Section 6.3, I focus on the use of CUSUM methods when data are available for multiple regions. At least four separate perspectives on spatial surveillance with CUSUM methods have been taken, and these four approaches are summarized. Section 6.4 provides a summary.

6.2 STATISTICAL PROCESS CONTROL

6.2.1 Shewhart Charts

Shewhart charts plot individual observations, or the means of groups of observations, as they are observed over time. Limits or thresholds are placed on the charts, and an 'out-of-control' signal is sent if an observation is found to be outside of these limits. When observations come from a standard normal distribution, the establishment of a threshold of ± 3 would imply that while in control, any observation would cause a signal with probability of 0.0027 (where this is the area corresponding to the tails of the standard normal distribution). This in turn implies a false alarm (where significant change is declared when it in fact has not occurred), on average, every $1/0.0027 = 370$ observations. If false alarms were more (or less) tolerable, the threshold could be adjusted accordingly.

Although Shewhart charts are good at detecting large changes in the mean of a variable, they fare less well (e.g., in comparison with CUSUM methods) in the quick detection of more subtle changes in the mean. Chapter 3 fully discussed these optimality issues.

6.2.2 Cumulative Sum Charts

CUSUM methods are designed to detect sudden changes in the mean value of a quantity of interest; they are widely used in industrial process control to monitor production quality. The methods rely upon the assumption that that the variable exhibits no serial autocorrelation. In the most common case, it is also assumed that the quantity being monitored is normally distributed, although it

is also possible to monitor observations that come from other distributions (and some of these will be discussed subsequently).

We first review the case of normally distributed observations. This could apply, for example, to the number of people with a particular disease in a large population. In particular, if all individuals have the same probability of disease, the underlying distribution of disease counts is binomial, but this can often be approximated by the normal distribution. Without loss of generality, let the variable of interest be converted to a z-score with mean zero and variance one. One way to achieve this for Poisson counts is to use $z = (O - E)/\sqrt{E}$, where O and E represent observed and expected counts, respectively. The CUSUM, following observation t, is defined as

$$S_t = \max(0, S_{t-1} + z_t - k), \tag{6.1}$$

where k is a parameter, and the CUSUM is started at zero (i.e., $S_0 = 0$). A change in mean is signaled if $S_t > h$, where h is a threshold parameter. Signals will sometimes occur when no actual change has taken place; the expected time until a false alarm is called the 'in-control' average run length, and it is designated by the notation ARL_0.

Note that values of z in excess of k are cumulated. The parameter k in this instance, where a standardized variable is being monitored, is often chosen to be equal to $1/2$; in the more general case where the variable of interest may not have been standardized, k is often chosen to be equal to one-half of the standard deviation associated with the variable being monitored. The choice of $k = \frac{1}{2}$ minimizes the average out-of-control run length (i.e., the average number of observations between the time of change and the time of detection) for a given value of ARL_0, when a true increase of one standard deviation has occurred. More generally, k is chosen to minimize the time needed to detect a change of $2k$ standard deviations in the mean.

The threshold parameter h is chosen in conjunction with a predetermined, acceptable rate of 'false alarms'; high values of h lead to a low probability of a false alarm, but also a lower probability of detecting a real change. Most texts on statistical process control have tables and charts that may be used to find the value of h that is associated with chosen values of ARL_0 and k. When $k = 1/2$, an approximation for ARL_0 may be derived from:

$$\mathrm{ARL}_0 = 2(e^a - a - 1), \tag{6.2}$$

where $a = h + 1.166$ (Siegmund 1985). One can make practical use of this approximation to choose the parameter h by first deciding upon a value of ARL_0, and then solving the approximation for the corresponding value of h. In the more general situation where a nonstandardized variable is being monitored, the critical value of the CUSUM is determined by multiplying the value of h by the standard deviation of the variable being monitored.

Siegmund's approximation requires numerical methods to solve for the threshold parameter h, for a desired and specified value of ARL_0. Rogerson (2004) has shown that Siegmund's equation may be solved, approximately, for the threshold parameter as a function of the in-control average run length:

$$h \approx \left(\frac{\mathrm{ARL}_0 + 4}{\mathrm{ARL}_0 + 2}\right) \ln\left(\frac{\mathrm{ARL}_0}{2} + 1\right) - 1.166. \tag{6.3}$$

When k is not necessarily equal to ½, the more general form of the equation for h is:

$$h \approx \left(\frac{2k^2\mathrm{ARL}_0 + 2}{2k^2\mathrm{ARL}_0 + 1}\right) \frac{\ln(1 + 2k^2\mathrm{ARL}_0)}{2k} - 1.166. \tag{6.4}$$

6.2.2.1 Illustration

To illustrate how the CUSUM methodology is implemented, and some of the issues that arise, data were simulated for a nine-region spatial system, constructed by assuming a three-by-three structure of square regions in a square study area. Simulated data are in the form of standardized z-scores. The regions were numbered from 1 to 9, beginning in the upper-left hand corner, and proceeding row by row, with the lower right region designated as region 9 (a map of this hypothetical spatial system is not shown). In Table 6.1, the simulated z-scores are depicted for each region, for each time period. Each column represents a region, and each row represents a time period. The data in Table 6.1 were developed by first choosing random variates from a standard normal distribution for the first 15 time periods, for each of the nine regions. Beginning in period 16, each region's mean value increased; the mean increased by 0.2 in regions 1, 3, 7, and 9; by 0.3 in regions 2, 4, 6, and 8; and by 0.75 in region 5. This corresponds to an increase that is centered on region 5, and dampens as one goes outward from there.

Now suppose that each region maintains its own CUSUM. If we use $k = 0.5$, and assume a desired ARL_0 of 100, this leads to a threshold of 2.84 that will be used in each region:

$$h \approx \left(\frac{\mathrm{ARL}_0 + 4}{\mathrm{ARL}_0 + 2}\right) \ln\left(\frac{\mathrm{ARL}_0}{2} + 1\right) - 1.166 = 2.84. \tag{6.5}$$

Maintaining CUSUMs for each region using equation (6.1) reveals the following signals: region 1, periods 17–19; region 2, periods 22–23; region 4, periods 17–29; region 5, periods 24–30; and region 9, periods 7–11. Note that one of the regions (region 9) signals even before the change occurs at period 16. This is clearly a false alarm. Three other regions (1, 2 and 4) exceed the threshold, but only for a temporary period. By time period 30, an increased mean is indicated only in region 5.

Table 6.1 Hypothetical *z*-scores for nine hypothetical regions, for 30 time periods.

Time Period	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Region 7	Region 8	Region 9
1	−0.80	−2.04	−0.69	−1.47	1.05	−0.05	1.34	−0.52	1.58
2	0.30	0.42	−0.44	0.96	0.07	−0.20	−1.20	−0.44	0.15
3	0.31	0.11	−0.29	−0.01	−1.28	1.47	0.40	0.23	0.84
4	1.76	−1.65	−0.99	0.78	−0.85	0.54	0.82	−0.39	0.95
5	−0.75	1.48	−0.34	0.41	−1.33	−1.12	0.64	−1.41	0.65
6	0.38	0.05	−0.41	−0.07	−0.05	1.02	−1.60	0.23	0.91
7	0.57	−0.78	−1.40	0.25	−1.97	1.02	0.70	0.19	1.90
8	−0.79	−1.04	−1.34	−0.32	−0.13	0.48	−0.64	1.18	1.78
9	−1.04	1.14	−3.37	0.47	−0.88	−0.18	−1.83	1.11	0.55
10	1.14	−1.17	−1.00	0.61	1.62	0.48	−0.10	0.70	−0.03
11	1.40	0.03	0.94	−1.80	0.33	−0.91	−1.13	1.25	−0.66
12	0.35	−0.28	0.55	−0.18	0.37	−0.95	−0.22	−2.13	−2.56
13	1.38	−1.12	0.00	−1.11	0.37	−0.63	0.12	0.44	1.18
14	−0.01	0.83	−0.98	−0.55	0.17	−0.97	0.62	−0.24	0.74
15	1.56	1.67	1.73	2.89	−0.03	0.90	−0.39	−0.64	0.51
16	−0.57	−0.85	0.82	0.52	1.25	−1.03	0.97	−1.54	0.12
17	1.98	0.40	−0.01	1.45	0.13	0.02	0.36	0.93	0.92
18	1.77	−0.66	−0.12	0.33	0.27	1.10	−1.58	1.08	0.84
19	−0.89	0.07	0.09	0.55	0.00	−1.13	−0.02	1.51	−2.30
20	−0.81	1.40	0.61	1.69	−0.35	0.38	−1.02	1.18	0.76
21	−0.52	2.05	−0.11	1.59	1.21	−1.12	−0.57	−0.78	0.35
22	1.18	1.75	−0.37	−0.09	−0.74	−0.08	0.87	0.76	0.00
23	0.55	−0.01	0.41	1.65	1.87	1.15	−0.13	−0.59	−0.05
24	0.63	−1.29	0.87	1.65	2.26	−0.05	−0.68	−2.12	0.10
25	−0.20	1.12	1.94	−1.74	1.65	0.44	1.15	−0.30	−0.38
26	−0.44	−0.52	1.22	−0.56	0.30	0.73	2.40	−0.21	−0.25
27	0.65	−0.80	0.10	0.50	1.65	−0.36	−0.70	−1.47	−0.77
28	−1.99	1.92	−1.90	0.54	0.94	0.68	1.17	0.26	−1.71
29	1.28	0.77	0.44	−0.35	1.70	−0.29	−1.88	−1.07	−0.40
30	0.06	0.38	−0.93	−0.33	0.47	−0.37	−0.14	0.73	−1.74

These nine separate surveillance systems might be suitable for each of nine individual, regional health departments. However, there are a number of important aspects of surveillance pertaining to the spatial and hierarchical structure of the study area that merit further discussion. For example:

(1) A state health official desiring an ARL_0 of 100 (i.e., an average time of 100 time periods before witnessing an alarm in *any* of the nine regions) would have to set the threshold higher than the value of 2.84 found above; otherwise, alarms would occur too frequently. This is discussed further in Section 6.3.1.

(2) Regional officials could conceivably miss a change that is spread across several regions. Note in the example above that small changes have occurred in each region, but the regional alarms are not necessarily persistent or timely. In general, the magnitude of the change might be relatively small in any particular region, but if many such small changes across clusters of counties are viewed in their totality, the change may become more apparent and detectable. This is discussed further in Sections 6.3.2 and 6.3.3.

6.2.2.2 Cumulative sum charts for Poisson data

Surveillance of public health data often requires methods that are able to handle the monitoring of rare events effectively. In this case, the approach described above is not adequate, since frequencies do not have a normal distribution when the mean count is low. One approach is to use the Poisson CUSUM (Lucas, 1985). When the variable being monitored has a Poisson distribution, the CUSUM is

$$S_t = \max(0, S_{t-1} + y_t - k), \tag{6.6}$$

where y_t is the count observed at time t. We now discuss determination of the parameters k and threshold h. Let $\lambda^{(a)}$ be the mean value of the in-control Poisson parameter. Following Lucas, the corresponding k-value that minimizes the time to detect a change from $\lambda^{(a)}$ to the specified out-of-control parameter ($\lambda^{(d)}$) is

$$k = \frac{\lambda^{(d)} - \lambda^{(a)}}{\ln \lambda^{(d)} - \ln \lambda^{(a)}}. \tag{6.7}$$

Then the threshold parameter h may be found from the values of the parameter k and the desired ARL_0 by using either a table (see Lucas, 1985), Monte Carlo simulation, or an algorithm such as the one provided by White and Keats (1996), which makes use of a Markov chain approximation. To illustrate, if $\lambda^{(a)} = 4$, one might desire to detect quickly a one standard deviation increase

to $\lambda^{(d)} = 6$; in this case we would first find $k = 4.93$ from equation (6.7). Then, if we desired an ARL_0 of approximately 420, we could use Table 2 of Lucas to find $h = 10$.

If the in-control value of λ is larger than about 2, then it is feasible to transform the counts to a standard normal random variable, z, using the transformation suggested by Rossi *et al.* (1999),

$$z = \frac{y - 3\lambda + 2\sqrt{\lambda y}}{2\sqrt{\lambda}}, \tag{6.8}$$

where y is the observed count and λ is the expected count.

As Rogerson and Yamada (2004a) indicate, this transformation can give misleading results for small values of λ. For example, when desired values of $ARL_0 = 500$ and $ARL_1 = 3$ (where ARL_1 is the average time taken to detect an actual increase) are used in situations where $\lambda < 2$, simulations show that using this transformation will almost always yield actual values of ARL_0 that are significantly lower than the desired value of 500. In some cases (e.g., $\lambda \approx 0.15$), the actual ARL will be lower than 100, indicating a much higher rate of false alarms than desired. The performance is better when $ARL_0 = 500$ and $ARL_1 = 7$, but use of the transformation will again lead to substantially more false alarms than desired when λ is less than about 0.25. They also note the instability with respect to similar values of λ: $\lambda = 0.56$ will lead to an ARL_0 of around 400, while $\lambda = 0.62$ is associated with an ARL_0 of over 700. This is also true when $ARL_1 = 3$: with $\lambda = 0.96$, the transformation has an ARL_0 of about 212, while with $\lambda = 0.98$, the transformed data has a very different ARL_0 of 635.

There is another reason to be cautious when applying the normalizing transformation. As Hawkins and Olwell (1998) point out, the times to detection for Poisson data will be shortest when the proper Poisson CUSUM procedure is employed. Using the CUSUM procedure for normal variables on the transformed data will generally lead to longer (though not usually substantially longer) detection times.

There have been several applications of Poisson CUSUMs in a public health context; examples include the surveillance of congenital malformations (Hill *et al.*, 1968; Weatherall and Haskey, 1976), salmonella outbreaks (Hutwagner *et al.*, 1997), and lower respiratory infection (Rogerson and Yamada, 2004a). In the latter, data on the number of visits to clinics made from each of 287 census tracts are monitored for the first 303 days of 1999, based upon expectations formed using daily data for the period 1996–1998. Equations (6.6) and (6.7) were used to form the Poisson CUSUM, with the modification that the k and λ parameters were allowed to vary over time. The temporally varying parameters reflected the fact that the expectations of daily counts, which were estimated as a function of month, a dummy weekend/weekday variable, and a time trend, were not constant.

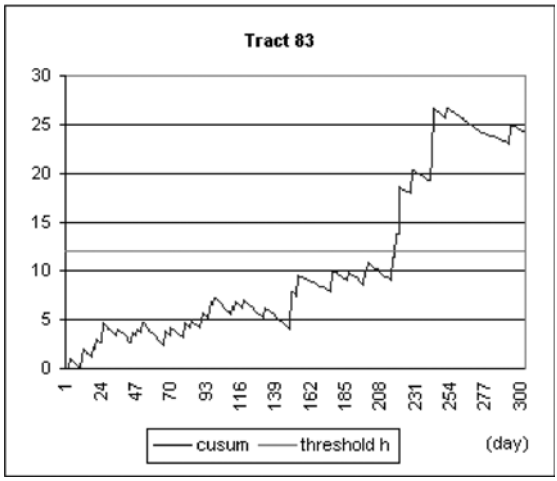


Figure 6.1 CUSUM chart for tract 83.

Figure 6.1 shows the Poisson CUSUM for one of the census tracts. During the base period, this tract had an average of 0.12 cases per day; this rose to 0.135 cases/day during 1999. The CUSUM crosses the threshold of $h = 12$ in early August 1999, around the 220th day of monitoring. Cases leading to the alarm occurred on August 4, 6, and 9 (there were two cases observed on August 9). These four cases in six days (0.67 cases/day) caused the Poisson CUSUM to rise above the critical threshold.

6.2.2.3 Cumulative sum charts for exponential data

An alternative approach for monitoring rare events is to use the fact that the times between Poisson-distributed events are exponentially distributed. Chen and her colleagues have written a series of papers on the sets method, which assumes exponential waiting times for events in a homogeneous Poisson process (Chen, 1978; Chen *et al.*, 1993, 1997). Sonesson and Bock (2003) discuss extensions to heterogeneous Poisson processes.

Lucas (1985) discusses the exponential (or time-between-events) CUSUM, and Gan (1994) compares its performance with the Poisson CUSUM. Gan finds that when there is interest in detecting an increase in the frequency of events, the exponential CUSUM outperforms the Poisson CUSUM, especially when there are large changes in the event frequency. This is due to the fact that the Poisson CUSUM does not signal until the end of the time period; the exponential CUSUM is able to capitalize on the data it uses by signaling *during* a period in which the frequency has increased. Similarly, Wolter (1987) notes that monitoring the gaps between events is more efficient than monitoring the number of events per time period when the number of events per period is small. Borror *et al.* (2003) have recently

shown that the exponential CUSUM is relatively robust with respect to departures from the assumption that the underlying distribution is exponential.

For an exponential distribution with mean $1/\theta$, given by

$$f(x) = \theta \exp(-\theta x),$$

a potential change from an in-control value of θ_0 to θ_1 can be monitored by first defining

$$k = \frac{\theta_1 - \theta_0}{\ln(\theta_0 \theta_1)} \tag{6.9}$$

The CUSUM, designed to detect an increase from θ_0 to θ_1 (corresponding to a decrease in the mean time between events, and an increase in the frequency of events, and where $\theta_1 > \theta_0$), is

$$S_t = \max(0, S_{t-1} - kx_t + 1), \tag{6.10}$$

where x_t is the time between events $t-1$ and t. To determine the threshold associated with the calculated value of k and a desired ARL_0, Gan provides charts (or nomographs). An alternative approach, suggested by Alwan (2000), is to transform the data to normality by raising the observed x values to the power 0.2777 (i.e., $y = x^{0.2777}$). Alwan also gives the expectation and variance associated with these transformed values:

$$\begin{aligned} E[y] &= 0.90\,11\theta_0^{-0.2777}, \\ V[y] &= 0.27\,80\theta_0^{-0.2777}. \end{aligned} \tag{6.11}$$

This allows one to implement the more common CUSUM based on the assumption of normality (along with approximations such as (6.4) for determining the appropriate threshold).

Hawkins and Olwell emphasize that transformations will adversely affect the performance of the CUSUM. In this case, the normality transformation will increase the time to detection when a change has occurred (although Alwan shows that the effect is not large). It is therefore of interest to determine the threshold for the exponential CUSUM directly. Without loss of generality, the problem is transformed into one having an in-control parameter of 1, and an out-of-control parameter equal to $\tilde{\theta}_1 = \theta_1/\theta_0$. This is achieved by normalizing the observed values by dividing each by θ_0. Based on the work of Siegmund (1985), it is possible to derive

$$\text{ARL}_0 \approx \frac{e^{\ln(\tilde{\theta}_1)(h+1.33)} - \ln(\tilde{\theta}_1)(h+1.33) - 1}{\ln(\tilde{\theta}_1)|1-k|}. \tag{6.12}$$

For a desired value of ARL_0, equation (6.12) may be solved for the threshold h.

Using arguments similar to those in Rogerson (2004), this equation may be solved approximately for the threshold, h, in terms of ARL_0:

$$h \approx \frac{q+2}{q+1}\frac{\ln(q+1)}{\ln(\tilde{\theta}_1)} - 1.33, \tag{6.13}$$

where

$$q = \mathrm{ARL}_0 \ln(\tilde{\theta}_1)|1-k|. \tag{6.14}$$

6.2.2.4 Other useful modifications for cumulative sum charts

One common modification is to start the CUSUM at a value other than zero. Lucas and Crosier (1982) recommend starting the CUSUM at a value of $h/2$, instead of zero. This has the benefit of signaling changes much more quickly if the series of observations begin out of control. This benefit is achieved at the cost of a slightly higher rate of false alarms if the usual value of h is used; if the value of h is raised slightly to maintain the desired ARL_0, then the time to detection is slightly longer than it would have been without using this fast initial response (FIR) feature. The FIR CUSUM is employed widely because the benefits of quicker detection that result in out-of-control startup situations generally outweigh the small costs described above.

Perhaps even more important in the context of public health surveillance is the fact that the 'in-control' parameters are often not known, and are instead often estimated from recent or historical data. For example, the 'true' rate of malformations in a health authority's geographical area that is to be used as a baseline expectation for the future rate of malformations is most often based upon either recent data for that area or some larger geographic area.

Hawkins and Olwell demonstrate that if a historical sample is used to estimate an unknown mean, subsequent surveillance can have false alarm rates that are much higher or much lower than the desired, nominal value of ARL_0. This is because the 'true' rate is not known. To account for this, they suggest a self-starting approach that may be summarized as follows (for a normally distributed variable; other self-starting approaches also may be devised for variables with other distributions). As each observation is made, the quantities

$$T_n = \frac{X_n - \bar{X}_{n-1}}{s_{n-1}} \tag{6.15}$$

are found, where $\bar{X}_{n-1}$ and s_{n-1} are the sample mean and standard deviation based upon the first $n-1$ observations. The quantity

$$V_n = \sqrt{\frac{n-1}{n}}T_n \tag{6.16}$$

has a t-distribution with $n-2$ degrees of freedom. This can be transformed into a quantity that has a standard normal distribution as follows:

$$U_n = \Phi^{-1}[F_{n-2}(V_n)] \tag{6.17}$$

where F_{n-2} is the cumulative distribution function for the t-distribution with $n-2$ degrees of freedom, and Φ^{-1} is the inverse of the normal distribution. Thus the Us represent the value from a standard normal distribution that would have an area equal to the area observed for the quantity V_n under the t-distribution with $n-2$ degrees of freedom.

6.3 CUMULATIVE SUM METHODS FOR SPATIAL SURVEILLANCE

6.3.1 Maintaining a Cumulative Sum Chart for Each Region

To maintain a desired ARL over a set of m regional charts that are monitored simultaneously, the threshold for each chart should be adjusted. An approximate adjustment is found by using the product of m and ARL in place of ARL when the threshold is determined. For the illustrative data in Table 6.1, $m = 9$ regions; if ARL $= 100$ for a state official maintaining all nine regional charts, then h is found by using ARL $= 9 \times 100 = 900$ (using $k = 0.5$):

$$h \approx \frac{904}{902}\ln(451) - 1.166 = 4.96. \tag{6.18}$$

This ensures that the average time until the first false alarm over the set of m charts is equal to ARL.

This method for determining the threshold is an approximation; a more precise threshold may be found by using the fact that the distribution of run lengths is approximately exponential (Page, 1954). Then, following Raubertas (1989), the average run length between false alarms observed over the set of m charts is

$$ARL_0^* = \frac{1}{1-(1-1/ARL_0)^m}. \tag{6.19}$$

This can be rearranged to find the value of ARL_0 to be used on each chart, in terms of the desired value of ARL_0^*:

$$ARL_0 = \left[1 - \left(1 - \frac{1}{ARL_0^*}\right)^{1/m}\right]^{-1}. \tag{6.20}$$

For the example, with $\text{ARL}_0^* = 100$ and $m = 9$, $\text{ARL}_0 = 895.99$. This leads to $h = 4.95$, which is very close to the value found in (6.18).

For the data in Table 6.1, only the CUSUMs for regions 4 and 5 attain this threshold:

Region	Time periods where $S > h$
4	21–25
5	27–30

6.3.2 Maintaining Cumulative Sum Charts for Local Neighborhoods around Each Region

The implementation of the CUSUM approach in a regional setting has, to this point, been rather uninteresting from a spatial context – each region is simply monitored separately, and there is no explicitly spatial connection between the temporal evolutions of regions that are near to one another. Raubertas (1989) suggested that the CUSUM methodology be generalized by maintaining CUSUMs not for each individual region, but for each individual region and its surrounding neighborhood.

An extension of the approaches outlined above is to construct 'local statistics' in association with each geographic unit. These are defined as a weighted sum of the region's observation and surrounding observations, where the weights could potentially decline with increasing distance from the region. CUSUMs associated with these local statistics may be monitored. Because the local statistics are spatially autocorrelated, a Bonferroni adjustment would result in too high a value of h, making it difficult to detect change when it actually occurs.

We now implement this idea; as a reminder, we desire rapid detection of the shift from the null hypothesis (where there is no spatial pattern, and all regions have zero means) to the situation where a set of adjacent regions witnesses a change from regional means of zero to alternative, higher regional means.

At each location, we construct a local statistic, y_{it}, by using a Gaussian kernel, represented by a weighted sum of the regional values:

$$\begin{aligned} y_{it} &= \sum_j w_{ij} x_{jt}, \\ w_{ij} &= (\sqrt{\pi}\sigma)^{-1} \exp(-d_{ij}^2/2\sigma^2), \end{aligned} \tag{6.21}$$

where σ is the width of the Gaussian kernel (chosen to coincide with the likely size of any emergent spatial cluster), and d_{ij} is the distance from the centroid in region i to the centroid in region j. The local statistics constructed at or near edges will not have as many regional neighbors as other regions. Consequently, the sum of the squared weights ($\sum_j w_{ij}^2$), and the variance of the local statistic (which is based on the sum of the squared weights), will be smaller for regions

near edges than for other regions. To help address these edge effects, it is useful to use modified, scaled weights in place of the original weights to ensure equal variances for all local statistics. The modified weights are defined in terms of the original weights as follows:

$$w_{ij}^{*} = \frac{w_{ij}}{\sqrt{\sum_j w_{ij}^2}} \tag{6.22}$$

The local statistics, y_i, all have normal distributions with mean zero and variance one (see, for example Siegmund and Worsley, 1995; Rogerson, 2001a).

If there are m regions, one possibility for surveillance would be to monitor each local statistic individually, as in Section 6.3.1. However, the Bonferroni adjustment used to determine the threshold is conservative since the local statistics are correlated (i.e., local statistics that are near to one another in space are correlated, since there is some commonality in the information they make use of).

One alternative for the determination of appropriate thresholds would be Monte Carlo simulation; a value of $s < m$ could be used in $s\text{ARL}_0$, and the value of s that leads to a systemwide average run length of ARL_0 could be found via simulation.

An alternative approach that accounts for the spatial correlation of local statistics is to use the number of effectively independent regions. Rogerson (2001a) shows that, for a single test, the effective number of independent tests, e, when using a Gaussian kernel is approximately

$$e \approx m/(1 + 0.81\sigma^2). \tag{6.23}$$

The following simulation experiments are designed to evaluate whether this is of use in a monitoring context. The null hypothesis was simulated as follows. A 16×16 grid of cells was filled with normal standard deviates for each successive time period, and then smoothed with a Gaussian kernel using various values of σ. The central 8×8 portion of the grid was then taken as the study area, to avoid possible edge effects. Cusums were kept for each of the 64 local statistics (i.e., for each of the $y_{it}, i = 1, 2, \ldots, 64$).

Results are shown in Table 6.2. The first two columns give the parameters chosen for particular simulation runs. For each pair of σ and h values shown in the table, 200 trial runs were carried out, and the time until the first signal was recorded. To speed the simulations, censoring points were chosen; if the number of time periods needed for a false alarm exceeded a value of C (where the value of C was particular to each pair of σ and h values), this was noted, and the ARL_0 was estimated using the fact that average run lengths tend to have an exponential distribution (see Appendix to this chapter).

The third column gives the estimated ARL under the null hypothesis and is based upon the simulations. The fourth column gives the ARL that would result

Table 6.2 Effective number of independent tests when monitoring smoothed, neighborhood z-scores.

(1) σ	(2) h	(3) Estimated ARL_0 from simulations	(4) ARL_0 based on h and one region	(5) Effective no. of indep. tests	(6) Estimated no. of indep. tests (eq. (6.1))	(7) h determined using cols. (3) and (6)
1	5.5	55.8	1 555	1 555/55.8 = 27.9	35.4	5.74
1	6.0	81.7	2 573	2 573/81.7 = 31.5	35.4	6.12
1	6.7	137.0	5 196	5 196/137 = 37.9	35.4	6.63
1	7.4	277.0	10 481	10 481/277 = 37.8	35.4	7.33
1	8.0	433.5	19 112	19 112/433.5 = 44.1	35.4	7.78
1	8.7	960.5	38 506	38 506/960.5 = 39.3	35.4	8.58
1	9.2	1334.0	63 499	63 499/1334 = 47.6	35.4	8.90
2	4.5	52.0	564	564/52 = 10.8	15.1	4.83
2	5.5	111.5	1 555	1 555/111.5 = 13.9	15.1	5.58

if the corresponding value of h was used in a CUSUM with only one region. The number of effectively independent tests, as determined from the simulation, is shown in the fifth column; it is found by dividing column (4) by column (3). This is to be compared with column (6), which is the estimated number of independent tests based upon equation (6.23).

In general, columns (5) and (6) are quite similar. For low values of ARL_0, there is a tendency for the estimated number of independent tests to be too high (i.e., conservative). For large values of ARL_0, the opposite is true – the estimated number of independent tests is too low, and this would result in a somewhat higher number of false alarms than desired.

This comparison can also be viewed by comparing columns (2) and (7); the latter is the value of the threshold h that would be used if one used equation (6.23) to estimate the number of effectively independent tests, and if one desired an ARL_0 equal to that given in column (3). Again low values of ARL_0 would result in the use of thresholds (h) that were conservative; that is, the thresholds would be too high (in comparison with the h values of column 2, which are the ones that *should* be used to achieve the ARLs in column 3). In contrast, for high values of ARL_0, the estimated values of h are lower than the values in column 2 that in principle should be used to achieve the ARLs in column 3.

One more illustration of monitoring local statistics is now made by reconsidering the data in Table 6.1. From (6.21) and (6.22) (with $\sigma = 1$), the local statistic constructed for the center square would have weights of 0.2119 for corner squares, 0.3421 for squares adjacent by rook's adjacency, and a weight of 0.576 for the central region (region 5). The results of using a CUSUM based upon equation (6.21) for this weighted local statistic are shown in Table 6.3.

The CUSUM exceeds the threshold of 2.84 used for a single variable in time periods 18 and 20–30. It exceeds the conservative threshold of 4.96 associated with monitoring nine independent local statistics in time periods 23–30.

6.3.2.1 Poisson variables

For Poisson variables, one can monitor the quantities $y_{it} = \sum_j w_{ij} x_{jt}$, where x_{jt} is the observed count in region j at time t, and w_{ij} is a weight associated with, for example, the distance from region i to region j. These observed quantities are then compared with their corresponding expectations, $\sum_j w_{ij} \lambda_{0,jt}$ (where the subscript jt refers to region j at time t), and used in a CUSUM for region i.

To determine appropriate critical thresholds for each region, Monte Carlo simulation of the null hypothesis may be used (where observed counts are realizations from Poisson or normal distributions with parameters set equal to the corresponding expectations). In particular, with a desired average run length of ARL_0, the critical thresholds should be determined using $s ARL_0$; the value of s is less than the number of regions (r), and is determined via simulation to lead to the desired average run length. The greater the correlation between the local regional statistics, the lower s will be relative to r (Rogerson and Yamada, 2004a).

Table 6.3 Cumulative sum for the local neighborhood statistic of the center square (region 5) in a hypothetical nine-region system.

Time	Cumulative sum for region 5
1	0.0
2	0.0
3	0.0
4	0.0
5	0.0
6	0.0
7	0.0
8	0.0
9	0.0
10	0.65
11	0.0
12	0.0
13	0.0
14	0.0
15	1.85
16	1.36
17	2.59
18	3.07
19	2.26
20	3.05
21	3.67
22	3.90
23	5.39
24	5.77
25	6.59
26	6.69
27	6.26
28	6.53
29	6.56
30	5.89

6.3.3 Cumulative Sum Charts for Global Spatial Statistics

A strategy for monitoring spatial patterns to detect quickly any deviation from the expected pattern is to place a global spatial statistic within a CUSUM monitoring system. For example, Rogerson (1997) begins by adopting Tango's (1995) statistic as a measure of spatial pattern. In principle, other spatial statistics could also be used. Suppose that Tango's statistic is found for a particular set of observations. Then, using the null hypothesis, one is able to compute the probability that a newly computed Tango statistic, based on one more observation, will take on a particular value. This leads to calculation of the expected

value and variance of the Tango statistic after the next observation, conditional upon the current value of the statistic. In turn, the expected value and variance may be used to convert the Tango statistic that is observed after the next observation into a z-score. Finally, these z-scores are used in a CUSUM framework.

Rogerson and Sun (2000) show how a similar approach may be used to monitor changes in the nearest neighbor statistic, while Rogerson (2001b) monitors changes in the space-time Knox statistic as new observations are collected.

6.3.4 Multivariate Cumulative Sum Methods

Monitoring all regions at once, as in Section 6.3.1, is one approach to multivariate or multiregional surveillance; Woodall and Ncube (1985) were early advocates of this approach. However, it does not allow for the potential correlation between variables (in this case, each variable is a region).

Pignatiello and Runger (1990) describe multivariate monitoring in the presence of a known variance–covariance matrix. Multivariate monitoring begins by cumulating the differences between the observed and expected number of cases in each region,

$$\mathbf{S}_t = \sum_{j=t-n_t+1} (\mathbf{O}_j - \mathbf{E}_j), \tag{6.24}$$

where $\mathbf{O}_j$ and $\mathbf{E}_j$ are vectors of observed and expected counts at time j; there are m elements in each vector, corresponding to entries for each of the m regions. It is assumed that the vector of observed counts is approximately multivariate normal. This will necessitate either a sufficiently large number of counts, or some transformation to normality. The quantity n_t is defined as the number of time periods since the CUSUM was last reset to zero.

The norm of $\mathbf{S}$ is a scalar representing the multivariate distance of the cumulated differences from the target:

$$||\mathbf{S}_t|| = \sqrt{\mathbf{S}_t' \boldsymbol{\Sigma}^{-1} \mathbf{S}_t} \tag{6.25}$$

where $\boldsymbol{\Sigma}$ is the variance–covariance matrix associated with the m regions. The quantity monitored is

$$MC1_t = \max\{0, ||\mathbf{S}_t|| - kn_t\}, \tag{6.26}$$

where

$$n_t = \begin{cases} n_{t-1} + 1, & MC1_{t-1} > 0, \\ 1; & \text{otherwise.} \end{cases} \tag{6.27}$$

The value of the parameter k is chosen to be equal to one-half of the multivariate distance from the target vector to the hypothesized alternative vector. The parameter k is approximately equal to one-half of the off-target multivariate distance one would like to quickly detect; this choice is thought to minimize the time taken to detect such an off-target process. The threshold parameter h is found by simulating the hypothesis of no change; choosing different values for h naturally leads to different average times until the threshold is first crossed. From this simulation, the threshold leading to the desired ARL_0 may be chosen. Crosier (1988) outlines a procedure that is very similar to this.

In a multivariate setting, the actual ARL_0s can be less than or greater than desired because the covariance matrix that captures spatial dependence is typically assumed or estimated, and is not known exactly. When the covariance matrix is incorrectly taken to be the identity matrix (i.e., when it is incorrectly assumed that there is no spatial dependence), the presence of spatial autocorrelation leads to much lower ARL_0s in comparison with the nominal ARL_0. Thus premature signals might be caused by spatial autocorrelation, and may not represent true increases in incidence. Furthermore, the effect on ARL_0 is more serious when k is high than when it is low. Rogerson and Yamada (2004b) provide additional examples showing that the effects of overestimating spatial autocorrelation appear to be less serious that the effects of underestimating it.

For the data in Table 6.1, we assume no spatial autocorrelation in the maps of z-scores that are observed each period, $\boldsymbol{\Sigma} = I$. Setting $k = 0.5$, simulation of the null hypothesis of no change in the mean reveals $h = 7.95$. For the data in Table 6.1, the multivariate CUSUM exceeds this threshold only for periods 8–11 and 18; the former is a false alarm.

Rogerson and Yamada (2004b) compare Pignatiello and Runger's multivariate approach with the multiple univariate approach described in Section 6.3.1. They find that the multiple univariate approaches is limited by its lack of ability to account for the spatial autocorrelation of regional data; the multivariate methods are limited by the difficulty in accurately specifying the multiregional covariance structure. When the degree of spatial autocorrelation is low, the univariate method is generally better at detecting changes in rates that occur in a small number of regions; the multivariate approach is better when change occurs in a large number of regions.

More extended discussion of multivariate surveillance can be found in Chapter 9.

6.4 SUMMARY AND DISCUSSION

In this chapter, we have reviewed several aspects of CUSUM methods and their application to public health surveillance. Although the CUSUM approach has been widely employed in this context, the majority of applications have been limited to the most common form of CUSUM, where the normal distribution is

assumed, and where a single region is monitored. The primary purposes of this chapter have been, first, to indicate how other developments in statistical process control allow for improved CUSUM surveillance (e.g., through implementation of the FIR feature and through the use of CUSUMs designed for statistical distributions other than the normal), and second, to indicate how these methods may be extended to a more explicitly spatial context.

Given that the CUSUM accumulates deviations between observed and expected values, one of the biggest challenges is to model expectations well. If the model for expectations is poor, errors in the model will eventually cumulate sufficiently to send a signal, and this signal would have nothing to do with underlying change in the rate or frequency of health events. It is important therefore to build any known variability, such as seasonality, into the model for expectations. One possibility may be to attempt to account for lack of model fit directly in the construction of the CUSUM; this would be a difficult, but useful and promising direction for further study.

Signals may be also be a consequence of poor data quality, in addition to being caused by poor models for expectations. It should not be surprising if initial applications of CUSUM methods provide more clues on the issues of data quality and model fit and are less reliable indicators of true change. Only when these difficult questions are addressed satisfactorily can the methods achieve their full potential.

Finally, it is important to recognize that CUSUM methods represent just one approach to sequential decision-making. In Chapter 3, Frisén and Sonesson describe the desirable optimality of CUSUM and other methods.

ACKNOWLEDGMENTS

The support of Grant 1R01 ES09816-01 from the National Institutes of Health and National Cancer Institute Grant R01 CA92693-01 is gratefully acknowledged.

APPENDIX

Suppose one has n uncensored and m censored observations (the latter known to be greater than some value C) from an exponential distribution with unknown parameter θ. An estimate of the unknown parameter may be found by first taking the likelihood:

$$L = \prod_{i=1}^{n} \theta e^{-\theta x_i} \prod_{i=1}^{m} e^{-C\theta}$$

After finding the log-likelihood,

$$\ln L = n \ln \theta - \theta \sum_{i=1}^{n} x_i - C\theta m,$$

one takes the derivative with respect to θ. Equating to zero and solving for θ yields

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i + Cm}$$

The average of this exponential variable is the average run length of the CUSUM under the null hypothesis; it is equal to the reciprocal of this estimate:

$$ARL = \frac{1}{\hat{\theta}} = \frac{\sum_{i=1}^{n} x_i + Cm}{n}.$$

7

Scan Statistics for Geographical Disease Surveillance: An Overview

Martin Kulldorff

7.1 INTRODUCTION

7.1.1 Geographical Disease Surveillance

Public health surveillance has been defined as 'the systematic ongoing assessment of the health of a community, based on the collection, interpretation, and use of health data and information' (Teutsch and Churchill, 2000). A key concept is the vigilance for unsuspected relationships, which is in contrast to most epidemiological studies where the main goal is to evaluate clear and predefined hypotheses. In geographical disease surveillance, the interest is in the spatial and/or spatio-temporal distribution of disease. Here are some examples.

Sheehan *et al.* (2000), Roche *et al.* (2002), Gregorio *et al.* (2002), and Thomas and Carlin (2003) looked at the geographical distribution of the proportion of late versus early stage breast cancer incidence in Massachusetts, New Jersey, Connecticut, and Minnesota respectively. Areas with a high proportion of late stage breast cancer indicate places where public health officials may want to make special efforts to increase mammography screening and other early detection efforts.

The geographical distribution of Creutzfeldt–Jakob disease was studied in Great Britain, and investigators found a small statistically significant cluster with five cases in Charnwood, Leicestershire, England ($p = 0.004$). A detailed local epidemiological investigation identified specific and unusual butcher shop

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

practices as the likely cause for the outbreak (Cousens *et al.*, 2001; Bryant and Monk, 2001).

Viel *et al.* (2000) did geographical surveillance of soft-tissue sarcoma and non-Hodgkin's lymphoma in France, finding statistically significant high incidence rates around a municipal solid waste incinerator with high dioxin emission levels, a known carcinogen. As a result of the report, additional emission controls were installed at the incinerator.

In the 1990s, the Washington State Health Department evaluated a glioblastoma cancer cluster alarm around Seattle-Tacoma International Airport that the community was very concerned about. The earliest analyses were inconclusive as results depended on the geographical boundaries chosen to define the cancer cluster. There were also problems with pre-selection bias due to testing for an increased incidence in an area that was chosen because it had an increased incidence (the 'Texas sharpshooter effect', named after the Texan who first fired his gun at the barn door and then drew the bull's-eye around the bullet hole). A geographical surveillance approach incorporating the county as a whole revealed a nonsignificant cluster around the airport, adding weight to other evidence that it was probably a chance occurrence (VanEenwyk *et al.*, 1999).

In New York City, the health department has studied the spatio-temporal distribution of dead birds reported by the public, with an increase in the number of reported dead birds from an area used as a signal of increased West Nile virus activity (Mostashari *et al.*, 2003). This provides information on where preventive measures such as the use of mosquito larvicides is warranted.

In the fall of 2001, the New York City Department of Health set up a real-time syndromic surveillance system for the early detection of disease outbreaks. On a daily basis, they receive information from hospitals about the number of emergency room visits, broken down by syndrome/symptom (respiratory, vomiting, fever, diarrhea, etc.) and residential zip code (Heffernan *et al.*, 2004). Every morning the data is analyzed for unusual patterns to see whether there are any indications of either citywide or localized disease outbreaks. The rationale for using syndromes rather than diagnosed diseases is the hope of picking up an outbreak as early as possible before lab results and other tests can confirm a particular disease diagnosis.

The geographical distribution of disease varies naturally and many apparent disease clusters are simply due to random fluctuations. What all these examples have in common is that they used the spatial and/or space-time scan statistics to differentiate clusters that are likely chance occurrences from those that are unlikely to be due to chance. Of course, there are many other types of geographical disease surveillance of equal importance, for which other statistical methods should be used. The aim of this paper is to provide a brief overview of one simple and user-friendly tool that epidemiologists and public health officials can use in their disease surveillance work: the spatial and space-time scan statistics.

7.1.2 Tests for Spatial Randomness

There are different types of statistical methods used to test whether a spatial or space-time pattern of counts is random or not. Besag and Newell (1991) distinguished between focused and general tests. Among general tests, Kulldorff (1998) distinguished between global clustering and cluster detection tests.

Tests for global clustering (Alt and Vach, 1991; Besag and Newell, 1991; Cuzick and Edwards, 1990; Grimson, 1991; Tango, 1995, 2000; Walter, 1994) are used when we want to investigate whether there is clustering throughout the study region, without being interested in the specific location of clusters. For example, we might want to know if a particular disease is infectious or not, in which case we would expect cases to be found close to each other no matter where they occur. The most extreme example of this type of clustering arises if we have one parent case with all other cases being its closest neighbors, but more typical is a situation where groups of cases are spread throughout the region.

Cluster detection tests are concerned with local clusters. They are used when there is simultaneous interest in detecting the location of clusters and testing their statistical significance. For disease outbreak detection and the prevention and control of disease this is critical, since public health officials need to know the location of an outbreak in order to know where to intervene. The spatial and space-time scan statistics, described in detail below, are cluster detection tests.

Focused cluster tests are also concerned with local clusters, but used when there is a prespecified hypothesis, not generated by the data, about the location of the cluster (Bithell, 1995; Lawson, 1993; Lawson and Waller, 1996; Stone, 1988; Waller *et al.*, 1992). For example, the hypothesis could be that disease incidence is high around a specific toxic waste site that is known to leak. If the location hypothesis was not generated by the data, then a focused test will have higher statistical power than a cluster detection test. On the other hand, if the hypothesis is generated from the data, then we will have the 'Texas sharpshooter' type preselection bias if a focused test is used.

7.1.3 Scan Statistics

The scan statistic is a statistical method with many applications, designed to detect a local excess of events and to test if such an excess may reasonably have occurred by chance. Scan statistics were first studied in detail by Naus (1965a, 1965b), who looked at the problem in both one and two dimensions. The field has recently been summarized in two excellent books by Glaz and Balakrishnan (1999) and by Glaz *et al.* (2001). In the simplest form of a scan statistic there is a time period of length T and a number C of events at times $t_i, i = 1, \ldots, C$. A scanning window of fixed length $w < T$ is continuously moved across time,

and the definition of the scan statistic S is the maximum number of events in the window as it is scanning the time period. In mathematical notation,

$$S = \max_{0<s<T-w} \sum_{i=1}^{C} I(t_i \in [s, s+w]), \tag{7.1}$$

where $I(\cdot)$ is the 0/1 indicator function. The value of S is compared to what would be expected by chance based on the event times being generated from a homogeneous Poisson process, either with or without conditioning on the total number of observed events C. One-dimensional temporal scan statistics have been used for disease surveillance since the early 1980s (Wallenstein, 1980; Weinstock, 1981).

Three basic properties of the spatial and other multidimensional scan statistics are the geometry of the area being scanned, the probability distribution generating events under the null hypothesis, and the shapes and sizes of the scanning window. In terms of the region being scanned, Naus (1965b), Loader (1991), Alm (1997, 1998), and Anderson and Titterington (1997) all considered a two-dimensional rectangle. Alm (1998) also looked at a three-dimensional rectangular volume. Chen and Glaz (1996) studied a regular grid of discrete points within a rectangular area. Turnbull *et al.* (1990) used an irregular grid, where points may be anywhere within an arbitrarily shaped area. Under the null hypothesis, Naus (1965b), Loader (1991), and Alm (1997, 1998) looked at a homogeneous Poisson process, Turnbull *et al.* (1990) considered an inhomogeneous Poisson process, while Anderson and Titterington (1997) considered both types. Chen and Glaz (1996) considered a Bernoulli model. As for the scanning window, Naus (1965b), Loader (1991), Chen and Glaz (1996), Alm (1997, 1998), and Anderson and Titterington (1997) all considered rectangles. Alm (1997,1998) also looked at circles, triangles, and other convex shapes. Turnbull *et al.* (1990) considered a circular window centered at any of the grid points making up the data. The window is, in all cases, of fixed shape as well as of fixed size in terms of the expected number of events, with the exception of Loader (1991), who also considered a variable size window.

Except for Turnbull (1990), all of the above-mentioned authors have had the aim of mathematically finding the exact or approximate distribution of various scan statistics. These are very hard probability theory problems even for simple one-dimensional scan statistics, and a lot of the mathematical work has been focused on finding lower and upper bounds on the distribution probabilities. In disease surveillance, the scan statistics needed are much more complex due to (i) the fact that the population at risk is unevenly distributed geographically, with higher population density in cities than in the country side, (ii) the need to use a variable size scanning window, since we do not know the size of potential disease clusters a priori, and (iii) the need to make adjustments for natural spatial and temporal variation because of known risk factors.

7.2 SCAN STATISTICS FOR GEOGRAPHICAL DISEASE SURVEILLANCE

Based on the likelihood ratio test, Kulldorff and Nagarwalla (1995) and Kulldorff (1997) presented a general mathematical model for spatial scan statistics that adjusts for the uneven geographical population density and allows for a variable size scanning window. These were later generalized to prospective space-time scan statistics for the early detection of disease outbreaks (Kulldorff, 2001). By using Monte Carlo hypothesis testing (Dwass, 1957), there is no longer a need to worry about the very difficult mathematics entailed in finding approximate or asymptotic solutions. With this approach, random data sets are generated under the known null hypothesis, and the value of the scan statistic is calculated for the real data set and compared to its value for the random data sets. Rather than worrying about complicated probability theory, though, we must now have efficient algorithms (Kulldorff, 1999). While computer-intensive, the Monte Carlo approach need not to be overly so, and users routinely use spatial scan statistics to analyze data sets with more than 10 000 geographical locations.

7.2.1 Probability Models

Spatial and space-time scan statistics can be used to analyze incidence, mortality, prevalence, treatment, survival, survey, and many other types of data. Depending on the nature of the data, different underlying probability models should be used. Here we describe when to use the Bernoulli, Poisson (Kulldorff, 1997) and exponential (Huang *et al.*, 2004) probability models.

When there are 'cases' and 'noncases' represented by a 0/1 variable, a Bernoulli model should be applied. The 0/1 variable may represent people with or without a disease or people with different types of disease such as early and late stage breast cancer. They may reflect cases and controls from a larger population, or they may together constitute the population as a whole. Separate locations may be specified for each case and each noncase, or the data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc., with multiple cases and noncases at each location. To do a space-time analysis, it is necessary to have a time for each case and each noncase as well. The analyses are conditioned on both the total number of cases and the total number of noncases observed.

In many applications, there are not individuals who are either a case or not a case, but rather a population at risk mass that reflects births, deaths, and migration of people into and out of an area as they occur over time. It is then appropriate to use a Poisson probability model where the number of cases in each location is Poisson distributed. Under the null hypothesis, when there are no covariates, the expected number of cases in each area is proportional to the person-time in that area. In most situations, the person-time for each

location and time interval will be an approximation based on the census, and the population is therefore usually aggregated into some political entities such as counties, parishes, census tracts, or zip code areas. To do a space-time analysis, it is also necessary to have a time for each case. The population in each location is either constant or changing according to some known temporal trends that may be different for different locations. The analyses are conditioned on the total number of cases observed, as well as on the estimated population numbers.

It is also possible to use spatial scan statistics to analyze survival data, looking for geographical areas with exceptionally short or long survival times, determining if such areas are statistically significant. In such as setting, we are not interested in the geographical distribution of individuals diagnosed with the disease compared to those without, but rather in the distribution of diagnosed individuals with short survival times compared to diagnosed individuals with long survival times. Huang *et al.* (2004) have developed a spatial scan statistic for survival data that uses the exponential probability distribution to model the survival times. It is possible to analyze both censored and noncensored data.

Unlike the Bernoulli and Poisson models, where the random data is generated under the known null hypothesis, that is not possible for survival times. The null hypothesis is that the survival time distribution is the same across space, but we almost never know what the actual distribution is, and even if we are willing to assume the shape of the distribution we do not know the mean or variance of it. To solve this, the randomization is done by conditioning on the collection of survival times observed, and then permuting the geographical coordinates and survival times. When there is censored data, the 0/1 censoring indicator will be permuted together with the corresponding censoring time as a pair. This randomization procedure ensures that the statistical inference (p-value) is unbiased even when the true survival distribution is not exponential.

7.2.2 Likelihood Ratio Test

Traditionally, scan statistics are simply defined as the maximum number of cases in the window. In most applications though, the cluster size is not known a priori, and one should then use a variable window size. It does then not work to simply use the maximum number of cases, since larger windows will have more cases just because they are larger. The scan statistics are then based on the likelihood, and defined as a likelihood ratio test (Loader, 1991; Kulldorff, 1997). For each location and size of the scanning window, the alternative hypothesis is that there is an elevated rate within the window as compared to outside.

Under the Poisson model, the likelihood function for a specific window is proportional to

$$\left(\frac{c}{n}\right)^{c}\left(\frac{C-c}{C-n}\right)^{C-c} I(c>n), \tag{7.2}$$

where C is the total number of cases, c is the number of cases within the window, and n is the covariate-adjusted expected number of cases within the window under the null hypothesis. $I(\cdot)$ is the indicator function. When SaTScan (see Section 7.8) is set to scan only for clusters with high rates, we use $I(c > n)$, which is equal to 1 when the window has more cases than expected under the null hypothesis and 0 otherwise. When only scanning for clusters with low rates we use $I(c < n)$ instead. When the program scans for clusters with either high or low rates the indicator function is removed from the likelihood formula.

For the Bernoulli model the likelihood function when scanning for high rates is

$$\left(\frac{c}{n}\right)^{c}\left(1-\frac{c}{n}\right)^{n-c}\left(\frac{C-c}{N-n}\right)^{C-c}\left(1-\frac{C-c}{N-n}\right)^{(N-n)-(C-c)} I\left(\frac{c}{n}>\frac{C-c}{N-n}\right), \quad (7.3)$$

where c and C are defined as above, n is the total number of cases and noncases in the cluster, N is the total number of cases and noncases in the data set. When scanning for low rates, '>' is changed to '<' in the indicator function, and when scanning for either high or low rates, the indicator function is removed.

The likelihood function is maximized over all window locations and sizes, and the one with the maximum likelihood constitutes the most likely cluster. This is the cluster that is least likely to have occurred by chance. The likelihood ratio for this window constitutes the maximum likelihood ratio test statistic. Its distribution under the null hypothesis is obtained by repeating the same analytic exercise on a large number W of random of replications of the data set generated under the null hypothesis, conditioning on the total number of cases and noncases observed. The p-value is obtained through Monte Carlo hypothesis testing (Dwass, 1957) by comparing the rank of the maximum likelihood from the real data set with the maximum likelihoods from the random data sets. If this rank is R, then $p = R/(1 + W)$.

7.2.3 Scanning Window

Whether the data is purely temporal, purely spatial, or space-time, the scanning window of the scan statistic can be defined as any collection of geographical 'zones' (Kulldorff, 1997). Most commonly, the spatial scan statistic imposes a circular window on the map. The circle is in turn centered on each of several possible grid points positioned throughout the study region. For each grid point, the radius of the circle varies continuously in size from zero to some upper limit specified by the user. In this way, the circular window is flexible both in location and size. In total, the method creates an infinite number of distinct geographical circles with different sets of neighboring data locations within them, where each circle is a possible candidate cluster. Other window shapes have also been used, such as ellipses (Kulldorff *et al.*, 2004b), rectangles (Chapter 11) and irregular shapes defined in a nonparametric fashion (Duczmal and Assunção, 2004;

Patil and Taillie, 2003, 2004). By increasing the number of windows considered in a given analysis, either through more grid-points, sizes or shapes, the power increases for detecting clusters conforming to many (but not necessarily all) of the newly included windows, while the power decreases for detecting clusters conforming to the originally included windows due to the increased amount of multiple testing that needs to be adjusted for.

The space-time scan statistic is most often defined by a cylindrical window with a circular geographic base and with height corresponding to time (Kulldorff *et al.*, 1998). The base is defined exactly as for the purely spatial scan statistic, while the height reflects the time period of potential clusters. The cylindrical window is then moved in space and time, so that for each possible geographical location and size, it also visits each possible time period. In effect, we obtain an infinite number of overlapping cylinders of different size and shape, jointly covering the entire study region, where each cylinder reflects a possible cluster.

The space-time scan statistics may be used for either retrospective or prospective analyses. In a retrospective analysis, a data set is analyzed once, scanning for current as well as past space-time clusters (Kulldorff *et al.*, 1998). This means that both the window start and end dates are flexible within some limitations specified by the user, such as a maximum temporal cluster size. In the prospective setting (Kulldorff, 2001), analyses are repeated every day, week, month, or year, and the interest is only in current clusters. In this case, the start date of the window is flexible as before, but the end date of the scanning window is always identical to the last date for which data is available. The prospective space-time scan statistic is useful for the early detection of disease outbreaks. It is possible to adjust the inference for the repeated analyses conducted over time, either by adjusting the p-values (Kulldorff, 2001) or, preferably, by using reccurrence intervals as proposed by Kleinman *et al.* (2004) and in Chapter 5, where a cluster is described as having a strength so that it would occur by chance only once every X number of days, months or years.

7.2.4 Adjustments

As described above, the spatial and space-time scan statistics adjust for the unevenness in the underlying population density, taking into account that there will be more cases per square mile in New York City than in Wyoming, in proportion to the population. It is often important to adjust for other factors as well. For example, based on raw population numbers there is higher cancer mortality in Florida than in other parts of the United States, simply because older people are at higher risk for cancer and there is a higher proportion of older people in Florida. For the Poisson model, such adjustments are easily done by indirect standardization, replacing the raw population number with the new expected counts (Kulldorff, 1997). For example, let c_{is} and pop_{is} be the number of cases and the population respectively in age group s in location i.

The mortality rate in age group s is then $r_s = \sum_i c_{is} / \sum_i pop_{is}$, and the expected count for location i is $\sum_s pop_{is} r_s$, which replaces n in equation (7.2). The same method can be used to adjust for other categorical covariates such as gender, ethnicity or education, as well as for area level covariates such as urbanicity or socioeconomic neighborhood variables (Kulldorff *et al.*, 1997; Klassen *et al.*, 2004; Sheehan *et al.*, 2004).

Other types of adjustments may also be of interest. For example, when strata counts are only available for the denominator, while missing from the numerator, one can use risk estimates from an epidemiological study in place of the r_ss above (Kulldorff *et al.*, 1997). The expected counts may also be calculated through Poisson or other types of regression analysis, as a preprocessing step, which also allows for the adjustment of continuous variables (Sheehan *et al.*, 2004), or in a space-time analysis, for purely temporal and purely spatial variation (Kleinman *et al.*, 2004).

7.3 SECONDARY CLUSTERS

With scan statistics it is also possible to identify secondary clusters in the data set in addition to the most likely cluster, and then order them according to the value of their likelihood. There will often be secondary clusters that are almost identical to the most likely cluster and that have almost as high likelihood values, since marginally expanding or reducing the size of a medium or large cluster will not change the likelihood that much. Most clusters of this type provide little additional information, but their existence means that while it is possible to pinpoint the general location of a cluster, its exact boundaries must remain uncertain. There may also be secondary clusters that do not overlap with the most likely cluster, and they may be of great interest. In Monte Carlo hypothesis testing, the likelihood of secondary clusters in the real data set should be compared with the likelihood of the most likely clusters in the simulated data. A consequence of this is that p-values for secondary clusters are conservative.

While it should be clear how the scan statistic adjusts for the multiple testing in terms of the multiple window location and sizes evaluated, one could reasonably ask whether it also adjusts for the multiple p-values obtained for the secondary clusters detected. The answer to this is that even though there are many p-values there is still only one test, but we need to dissect what we are doing into two parts. The first is whether or not to reject the null hypothesis. To do this, we only need to know the likelihood from the most likely cluster in the real data set and compare it with the most likely cluster in each of the random data sets. Secondary clusters are irrelevant for this. The second part is pinpointing the specific cluster causing the rejection. The most likely cluster is clearly causing the rejection, since that was the likelihood that was high when compared to the ones from the random data sets. But, it is possible that there is more than one cluster in the real data set that is strong enough to cause

a rejection of the null hypothesis, so rather than doing multiple tests we are simply tallying the clusters capable of rejecting the null hypothesis.

One useful way of thinking about it may be if there are two different clusters in different parts of the map, both with 20 cases when exactly two were expected. With identical likelihoods, either could be assigned as the most likely cluster, causing the rejection, and the scan statistic may arbitrarily select one as the primary and the other as the secondary cluster. In reality though, we should clearly treat them as equals. For clusters with excess risk, another way to look at it is that if we take the cases in the most likely cluster and move them to other locations on the map, either randomly or according to some deterministic rule, then no matter how they are distributed, we would still reject the null hypothesis due to the secondary cluster.

7.4 NULL AND ALTERNATIVE HYPOTHESES

7.4.1 The Null Hypothesis

After adjusting for population density and covariates such as age or gender, scan statistics are based on the null hypothesis of complete spatial randomness. For most disease data that is not true. Does this mean that the null hypothesis is wrong?

When accepting the notion of statistical hypothesis testing one must also accept the fact that the null hypothesis is never true. For example, when comparing the efficacy of two different surgical procedures in a clinical trial we know for sure that the efficacy cannot be equal, but we still use equality as the null hypothesis since we are interested in finding out whether one is better than the other. Likewise, with geographical data we know that disease risk is not the same everywhere but we still use it as the null hypothesis since we are interested in finding locations with excess risk. Hence, the null hypothesis is wrong in the sense that we know it is not true but it is not wrong in the sense that we should not use it.

7.4.2 Spatial Autocorrelation

Spatial autocorrelation means that the location of disease cases is dependent on the location of other disease cases in such a way that there is a tendency for them to occur close together. It is natural to ask whether spatial scan statistics assume that there is no spatial autocorrelation in the data. The answer is no. Rather, it is a test of whether there is spatial autocorrelation or other divergences from the null hypothesis. In this sense it is equivalent to a test for normality, which does not assume that the data is normally distributed but tests whether it is.

If one is interested in whether there is spatial autocorrelation in the data, one should not necessarily use a scan statistic though. If one does not care about

cluster locations, there are tests for global clustering that have higher power than the spatial scan statistic and should be used instead (Song and Kulldorff, 2003). As mentioned before, the spatial scan statistic should be used when one is interested in the detection and statistical significance of local clusters.

In much of spatial statistics, it is of critical importance to adjust for spatial autocorrelation, but that is typically not done for spatial and space-time scan statistics, and for good reasons. Whether to adjust for spatial autocorrelation depends on the question being asked from the data. As an example, let us assume that we have geographical data on people who get sick due to food poisoning. In such data there is clearly spatial autocorrelation, since bad food sold at restaurants or grocery stores is often sold to multiple customers, many of who will live in the same neighborhood, city, or county. If we are doing spatial regression trying to determine what neighborhood characteristics – such as the grocery store chains that are present, the mean income, educational levels, and ethnic origin – contribute to a higher risk for food poisoning, it is critical to adjust for the spatial autocorrelation in the data. If not, the risk relationships will be overestimated with biased p-values that are too small, providing 'statistically significant' results when none exist. Here, the null hypothesis should be that there is spatial autocorrelation and the alternative hypothesis that there are geographical differences in the risk of food poisoning. On the other hand, if we are interested in quickly detecting food poisoning outbreaks, we should not adjust for the spatial autocorrelation since we are interested in detecting clusters due to such correlation, and if they are adjusted away, important clusters may go undetected. Here, the null hypothesis is that the food poisoning cases are geographically randomly distributed (adjusted for population density, etc.) and the alternative hypothesis is that there is some clustering either due to differences in underlying risk factors or spatial autocorrelation. Once the location of a cluster has been detected, it is for the local health officials to determine the source of the cluster to prevent further illness.

7.4.3 The Alternative Hypothesis

The spatial scan statistic uses a particular alternative hypothesis with an excess risk in, for example, a circular cluster. Does this mean that it can only be used to detect such alternative hypotheses?

The answer is no. Many widely used test statistics do not specify an alternative hypothesis at all. This means neither that they cannot be used for any alternative hypotheses nor that they are good for all alternatives. Likewise, if an explicit alternative is defined, as with the spatial scan statistic, that does not mean that it cannot be used for other alternative hypotheses as well. It is simply a question of the test statistic having good power for some alternative hypotheses and low power for others. The advantage of having a well-specified alternative hypothesis is that it gives some information about the alternatives for which the test can be expected to have good power.

7.5 POWER

The statistical power of the spatial and space-time scan statistics has been evaluated in a number of different settings (Kulldorff and Nagarwalla, 1995; Kulldorff *et al.* 2003, 2004a; Huang *et al.*, 2004). For the Poisson and Bernoulli probability models, the power depends primarily on the total number of cases in the study, the expected number of cases in the cluster area under the null hypothesis, and the relative risk (RR). The higher each of these is, the higher is the power to detect the cluster.

The power depends on other factors as well. By reducing the upper limit on the window size, the power slightly increases for clusters that are smaller than the new upper limit, while the power decreases for clusters that are larger. Moreover, while the circular spatial scan statistic has good power for many noncircular cluster shapes, the power decreases for less compact clusters.

7.6 VISUALIZING THE DETECTED CLUSTERS

Once a cluster has been detected using the spatial or space-time scan statistic, it can be depicted on a map in various ways. The easiest is to simply draw the circle corresponding to the most likely cluster, using the coordinates of its centroid and the radius. For aggregated data, a more common approach is to color the census areas within the cluster. Since a census area is either completely inside or complete outside the cluster, depending on the location of the census area centroid, such clusters are not perfect circles but have rough edges according to the boundaries of the census areas included. A third very nice approach first used by the New York State Cancer Surveillance Improvement Initiative (New York State Department of Health, 2001), is to use hatching to show the location of the spatial scan clusters, and overlaying it on top of a regular map with incidence rates. With hatching, the exact borders are diffuse, accurately reflecting the imprecise nature of the cluster borders generated by the spatial scan statistic.

When there are both clusters of high and low rates, it is common to show them in different colors or shadings.

Boscoe *et al.* (2003) wrote a paper on the visualization of spatial scan statistic results using nested circles, by which various combinations of overlapping clusters are used to create a smoothed map of RRs. Walsh and DeChello (2001) combined the spatial scan statistic with an empirical Bayes smoothing technique (Clayton and Kaldor, 1987), to construct a map of standardized mortality rates (SMRs) that shows a mix of scan-based clusters and spatial smoothing.

There is no one right way to map the detected scan statistic clusters. Rather, different approaches may be used at different times based on the particular study, and the above variations are great reflections of that.

7.7 A SAMPLE OF APPLICATIONS

The spatial and space-time scan statistics have been used for a wide variety of spatial disease surveillance problems, most often for cancer, infectious diseases and in veterinary medicine. In some studies it was the only statistical method used, but more commonly it was one of several methods used as part of a larger research or surveillance project. Here we provide a sample of various applications.

7.7.1 Cancer Surveillance

In light of many reports of childhood cancer clusters around the world, Hjalmars *et al.* (1996, 1999) used the spatial scan statistic to see if there were any childhood leukemia or childhood brain cancer incidence clusters in Sweden, finding none that was statistically significant. As part of their cancer surveillance initiative, the New York State Department of Health (2001) used the spatial scan statistic to look at the geographical variation of breast, lung, prostate, and colorectal cancer incidence in New York State, finding various statistically significant clusters but no local hotspots with greatly elevated risk. Hsu *et al.* (2004) looked at the geographical distribution of breast cancer incidence in Texas among different ethnic groups, finding significant clusters for some ethnic groups but not for others. Viel *et al.* (2000) investigated the geography of soft-tissue sarcoma and non-Hodgkin's lymphoma in the *département* of Doubs, France. Michelozzi *et al.* (2002) used the spatial scan statistic to evaluate childhood leukemia incidence in Rome, Italy, finding no statistically significant clusters. Buntinx *et al.* (2003) used it to confirm a bladder cancer incidence cluster in Limburg, Belgium ($p = 0.0001$), that was first found using more descriptive geographical surveillance methods. VanEenwyk *et al.* (1999) used it to evaluate a potential glioblastoma cluster around Seattle-Tacoma International Airport that was of great community concern, finding that it was not statistically significant, and hence a likely chance occurrence.

Kulldorff *et al.* (1997) looked at the geographical distribution of breast cancer mortality in the northeastern USA, finding a statistically significant cluster in the New York City – Philadelphia metropolitan area ($\text{RR} = 1.07, p = 0.001$). Jemal *et al.* (2002) explored the geography of prostate cancer mortality in the USA, while Fang *et al.* (2004) investigated brain cancer mortality in the USA (Figure 7.1). Both of the latter two studies found statistically significant clusters with very modest RRs, but no localized hotspots with high RRs.

In addition to cancer incidence and mortality, scan statistics can also be used to evaluate the geographical distribution of cancer stage, treatment or survival. The interest is then not in whether people are at higher risk to get cancer in certain areas. Rather, the question is whether there are some geographical areas where the cancer patients have shorter survival, higher risk of late state

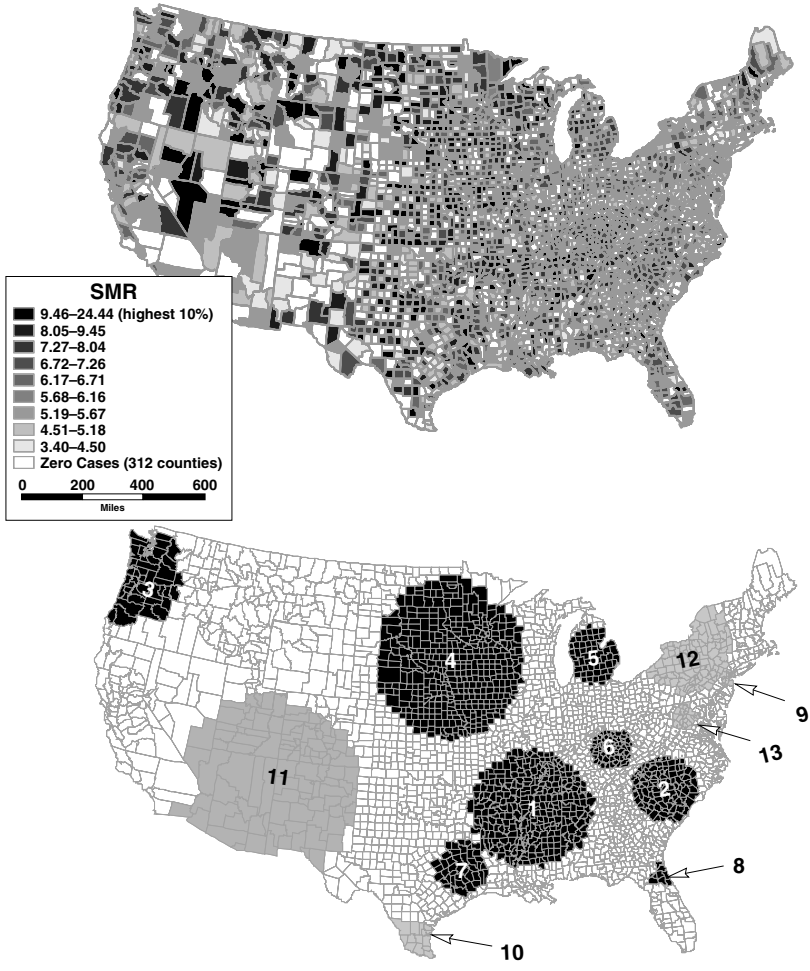


Figure 7.1 Brain cancer mortality rates by country for adults aged 20 and above in the USA, 1986–1995, adjusted for age (5-year age groups), gender, and ethnicity (white, black, other), using indirect standardization. Top: Standardized mortality rates. Bottom: The result of the spatial scan statistic. Dark clusters are areas of excess mortality, while the light clusters are areas with lower mortality. Within each category, clusters are numbered in order of the likelihood ratio, with the lowest number for the highest likelihood. All clusters except no. 8 are statistically significant at the 0.05 level. The most likely cluster was around Arkansas and Mississippi (no. 1), with 6251 cases when 5322 were expected (RR = 1.18, $p = 0.0001$). Fang *et al.* (2004) provide detailed information about the remaining clusters.

disease or where they get inferior treatment, compared to cancer patients elsewhere. In this manner, Sheehan *et al.* (2000), Roche *et al.* (2002), Gregorio *et al.* (2002) and Thomas and Carlin (2003) looked at the geographical distribution of late stage cancer in Massachusetts, New Jersey, Connecticut, and Minnesota,

respectively. In some cases the results led the state health department to increase mammography screening promotion in areas with a high proportion of late stage disease. Thomas and Carlin (2003) also looked at the geographical distribution of late stage colorectal cancer. Gregorio *et al.* (2001) investigated the geographical variation in breast cancer treatment in Connecticut, finding some areas lagging behind with the implementation of state-of-the-art treatment protocols. Huang *et al.* (2004) evaluated the geographical distribution of prostate cancer survival in Connecticut.

7.7.2 Infectious Diseases

During the last few years, the spatial and space-time scan statistics have been used for a number of different infectious diseases. As mentioned in the introduction, Cousens *et al.* (2001) investigated the geographical distribution of Creutzfeldt–Jakob disease in Great Britain, and Huillard d'Aignaux *et al.* (2002) did the same in France. Saunders *et al.* (2003) used the space-time scan statistic to search for human listeriosis clusters in New York State, by different ribotypes and pulsed-field gel electrophoresis types.

Scan statistics has also been used for vector-borne diseases. Chaput *et al.* (2002) used it to evaluate the spatial patterns of human granulocytic ehrlichiosis, a newly recognized tick-borne disease, finding a statistically significant cluster in the towns of Lyme and Old Lyme (RR = 2.6, $p = 0.001$) in southern Connecticut. Fevre *et al.* (2001) studied the geographical distribution of early cases of a sleeping sickness outbreak in eastern Uganda, finding a statistically significant cluster around a cattle market, which is logical in light of cattle being an important reservoir for the disease. Mostashari *et al.* (2003) used it for a surveillance system based on dead bird reports for the early detection of West Nile virus outbreaks.

7.7.3 Other Human Diseases

In addition to cancer and infectious diseases, the spatial scan statistic has also been used for other human disease and health events. In pediatrics, Kharrazi *et al.* (1998), Forand *et al.* (2002), the Colorado Department of Public Health and Environment (2002), and Boyle *et al.* (2004) used it to evaluate the spatial distribution of birth defects. George *et al.* (2001) looked at sudden infant death syndrome in Sweden, while Sankoh *et al.* (2001) looked at all types of childhood mortality in rural Burkina Faso.

In one of the earliest applications of the spatial scan statistic for disease surveillance, Walsh and Fenster (1997) studied systemic sclerosis mortality in the southeastern USA, reporting a statistically significant cluster for white men in parts of Tennessee, Kentucky, and Alabama (SMR=1.2, $p = 0.0004$) and another for black men around Northampton in North Carolina (SMR = 3.9, $p = 0.02$).

Sabel *et al.* (2003) studied the geographical distribution of amyotrophic lateral sclerosis in Finland. A very interesting aspect of this study is that they analyzed the data not only by place of death but also by place of birth, to better reflect geographical variation in genetic and early childhood determinants of the disease. The results from the two analyses were similar, with both detecting a statistically significant cluster in southeastern Finland for both the place of death (RR = 1.8, $p = 0.000\,01$) and place of birth (RR = 1.5, $p = 0.000\,01$). Walsh and DeChello (2001) evaluated the geography of systemic lupus errythematosus in the USA, finding four clusters with low rates (SMR = 0.56–0.68, $p < 0.001$) and three strong clusters with high rates (SMR = 1.41–1.65, $p < 0.0001$) and one weak cluster with a high rate (SMR = 1.51, $p = 0.02$). Enemark *et al.* (2002) evaluated the geographical variation in the proportion of different genotypes of *Cryptosoridium parvum* parasites in Denmark.

While less common, it also possible to use the spatial scan statistic for prevalence data. For example, Green *et al.* (2003) used it to investigate the geographical variation of diabetes mellitus prevalence in Winnipeg, Canada, finding several high- and low-risk areas ($p < 0.001$). López-Abente *et al.* (2003) studied the geographical distribution of Paget's disease prevalence in Spain.

The spatial and space-time scan statistics have also been used for other types of health data not directly related to a particular disease. Hanson and Wieczorek (2002) used it to evaluate the geographical distribution of alcohol-related mortality in New York State, finding a number of statistically significant clusters ($p < 0.01$) of either high or low mortality. Yiannakoulias *et al.* (2003) studied the geography of fall injuries in the elderly in Edmonton, Canada. Margai and Henry (2003) evaluated the geographical variation in learning disabilities in Binghamton, New York. Sudakin *et al.* (2002) used the space-time scan statistic to evaluate regional variation in pesticide exposure in Oregon. In two different criminology research projects, Kaminski and Jefferis (2000) and Beato Filho *et al.* (2001) used scan statistics for the spatial analysis of homicides in the USA and in Belo Horizonte, Brazil, respectively.

7.7.4 Veterinary Medicine

The spatial and space-time scan statistics have been used in a wide variety of veterinary disease surveillance projects. Studies of domestic animals include acute respiratory disease outbreaks in Norwegian cattle herds (Norström *et al.*, 2000), blowfly strike in Australian sheep (Ward, 2001), West Nile virus in US equids (United States Department of Agriculture, 2001), bovine spongiform enephalopathy in Swiss cattle (Doherr *et al.*, 2002), bovine tuberculosis in Argentinian cattle (Perez *et al.*, 2002), psoroptic sheep scab in Swiss sheep (Falconi *et al.*, 2002), leptospirosis in North American dogs (Ward, 2002), Aujeszky's disease in German pigs (Berke and Grosse Beilage, 2003), and viral diseases in farmed and wild Swiss salmonids (Knuesel *et al.*, 2003). The scan statistics have also been used for wildlife. For example, Smith *et al.* (2000) studied

anthrax outbreaks among animals in Kruger National Park, South Africa; Berke *et al.* (2002) looked at *Echinococcus multilocularis* parasites in red foxes in Lower Saxony, Germany; Miller *et al.* (2002, 2004) investigated *Toxoplasma gondii* parasites in sea otters in California; Hoar *et al.* (2003) studied sylvatic plague and *Bartonella vinsonii* bacterial infections in coyotes in California; Olea-Popelka *et al.* (2002) evaluated the geographical distribution of bovine tuberculosis in badgers in Ireland; and Joly *et al.* (2003) investigated chronic wasting disease in white-tailed deer in Wisconsin.

7.7.5 Plant Diseases

Health is not only important for humans and animals, but also for plants, and there is no reason why the spatial scan statistic cannot be used for plant disease surveillance as well. In fact, Coulston and Riitters (2003) used the method to look at the geographical variation in insect and pathogen indicators on trees in the Pacific Northwest and forest fragmentation indicators in the southeastern part of the USA.

7.8 SOFTWARE

Two software products are available for disease surveillance using the spatial and space-time scan statistics. SaTScan is a free software product that can be downloaded from www.satscan.org. It includes temporal scan statistics in addition to the spatial and space-time version, but no other statistical methods. The EpiAnalyst extension for ArcView GIS, a commercial product from Public Health Research Laboratories (www.phrl.org), contains a link between ArcView and SaTScan. ClusterSeer is a commercial software product produced by TerraSeer (www.terraseer.com) that contains both the spatial and space-time scan statistics together with many other statistical clustering methods.

ACKNOWLEDGMENT

The writing of this review was funded by National Cancer Institute grant R01-CA95979. The author thanks Alison Hemhauser for preparing the two maps.

8

Distance-Based Methods for Spatial and Spatio-temporal Surveillance

Laura Forsberg, Marco Bonetti, Caroline Jeffery, Al Ozonoff and Marcello Pagano

8.1 INTRODUCTION

The emergence of new infectious diseases and the threat of biological attacks have lead to a growing interest in methods of surveillance, including the accompanying statistical methods, for the early detection of an outbreak. Statistically, what we would like to do is detect the time point at which there is an increase in the number of infected individuals, an increase that may also be accompanied by a change in the spatial distribution of these patients, either, or both, of which might indicate an outbreak of some sort – a disturbance of normalcy. The time element is critical in that a less than timely detection would make the methods essentially useless.

The timeliness is an extra consideration that possibly distinguishes the newer surveillance methods from those in the older literature. The older ones are often related to such issues as the detection of cancer clusters (see, for example, Alexander and Boyle, 1996), and sometimes use data that was collected over a period of years prior to analysis which, parenthetically, makes the existence of a cluster questionable. This is not meant as a criticism of the classical methods as the time element is inherent in those methods, too.

When considering spatial methods for cluster detection, no method seems to be uniformly better than all others, so it is beneficial to review a number of these methods. Several reviews of statistical methods for the detection of spatial anomalies have been written (see, for example, Kulldorff, 1998; Elliott *et al.*, 2000;

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

Lawson, 2001; Brookmeier and Stroup, 2004, Chapter 7). Most of the statistical methods that have been described for the detection of spatial anomalies can be grouped into two general categories: quadrat methods and distance methods. Quadrat methods divide the geographical region into smaller areas termed quadrats and compare the incidence of events within the quadrat to the incidence in the remaining study region. The spatial scan statistic is perhaps the most widely known and used of these methods (Kulldorff, 1997). Distance-based methods, on the other hand, consider some measure of distance between events. Usually Euclidean distance is used as the measure of distance between individuals, but typically any measure of dissimilarity or similarity between events can be utilized. We focus on these distance-based methods in this chapter, and discuss two methods of more modern interest: the maximized excess events test (MEET) and the M statistic, with particular emphasis given to the latter method. We present the motivation for using distance-based methods in Section 8.2. In Section 8.3, we give a review of the MEET statistic and the M statistic, and their utility in public health surveillance. Section 8.4 introduces a data example to illustrate the implementation of the MEET and M statistic to detect spatial clusters of disease. Our focus of attention is a bivariate statistic, which simultaneously monitors case volume and the spatial distribution of the cases. This bivariate statistic is introduced to improve the power to detect suspicious patterns in the data stream. In Section 8.5, we describe and illustrate a method for determining the location of a cluster, or other spatial aberrations, once the M statistic has indicated that such an anomaly exists.

8.2 MOTIVATION

Distance-based methods consider the distribution of the pairwise interpoint distances between all the individuals in the study region. Under the null hypothesis this distribution remains stable. As time progresses, we need to be on the alert for a disturbance in the distribution. This alternative distribution should be sensitive to the detection of disturbances in how individuals are located, especially if individuals are clustered. These disturbances are those we would expect during an outbreak of a contagious disease or an outbreak resulting from one or more point source emissions of some bioterrorist agent. This places the problem in the classical hypothesis testing paradigm, and to pursue this thinking further, we seek methods that will have power against alternatives that reflect clustering of individuals. One obvious characterization of clustering is to consider pockets of individuals who consequently will have smaller average distances between themselves than they would in the null case. But this is not the only alternative one can envision; others may impact the second moment of the distribution of distances, for example.

One method considers a test of the mean of the interpoint distance distribution (Whittemore *et al.*, 1987). The statistic, usually called the δ statistic, is equal to a

weighted average of the observed distances, and thus tests for shifts in the mean of the interpoint distance distribution. Subsequent work has shown that this method is not very powerful at detecting clusters (Bonetti and Pagano, 2004a). The reason for the lack of power is that the mean is not an efficient summary of the null distribution, typically because the null distribution of distances is not normal. Furthermore, dependencies in the distances can often lead to complex deviations from the null distribution that may not necessarily lead to a shift in the overall mean. Figure 8.1 illustrates such a scenario arising from real data. Here the densities clearly differ from one another, but the mean does not lead to a powerful statistical test for detecting such a deviation.

Dealing with distances between individuals requires some thought since the usual statistical methods do not apply seamlessly. First, the distances themselves are not independently distributed. This would seem clear considering that for every n individuals there are $\binom{n}{2}$ distances. Thus considering the statistical properties of their joint distribution is not straightforward. Additionally, location data is often not reported precisely, but rather it is reported in a discretized manner. For instance, instead of individuals' home or work addresses we may only be told the census tract, postal code, or county in which they reside. Thus the distances can only assume values in a finite grid.

Additionally, the location of spatial aberrations in the study region will impact the shape that the alternative distribution will assume. For instance, a cluster

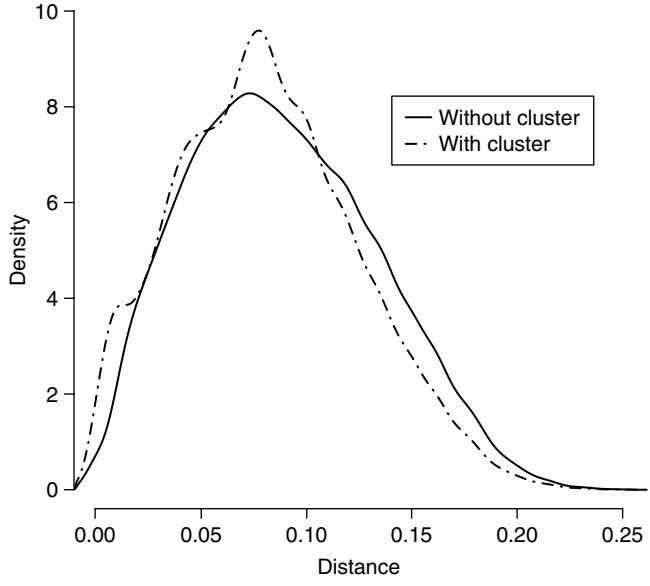


Figure 8.1 Distribution of the distances for a data set with no clusters ($\hat{\mu} = 0.090, \hat{\sigma} = 0.045$) versus a the same data set with clusters superimposed ($\hat{\mu} = 0.083, \hat{\sigma} = 0.044$).

placed in the study region will create a larger than expected number of small distances. However, the cluster will also create other abnormalities in the distribution, but these will depend upon where the cluster is placed, due to the addition of the distances between the cluster and other points in the region. This patterning increases the more clusters we have.

Several methods for analyzing distances have been proposed, although no one statistic seems to completely handle the complexities that distance data presents uniformly better than any others. K functions are one method that has been proposed (Ripley, 1976; Diggle and Chetwynd, 1991) for detecting spatial abnormalities, especially in the ecological literature (Dobbertin *et al.*, 2001; Couteron and Kokou, 1997). These functions enjoy nice mathematical properties, but can be cumbersome to implement for purposes of biosurveillance. Therefore we will direct our attention to two other methods, the MEET statistic and the M statistic, with particular emphasis on the latter.

8.3 DISTANCE-BASED STATISTICS FOR SURVEILLANCE

8.3.1 MEET Statistic

Tango (1995) describes a method of cluster detection that assumes that the data is aggregated into m regions according to some spatial boundaries, for instance by zip code or county. The statistic considers the difference between the observed rate of cases in each region and the expected rate, and then weights these differences by a measure of the distance between the regions. More explicitly, within the ith region, let y_i be the observed number of cases and e_i be the expected number of cases. Define the parameter λ such that any pair of cases that are farther than λ apart cannot be considered a cluster. Basically, λ can be thought of as some measure of the spatial extent of a cluster. Consider the vectors $\mathbf{r} = \{r_i\}$, where $r_i = y_i / \sum_{i=1}^{m} y_i$, and $\mathbf{p} = \{p_i\}$, where $p_i = e_i / \sum_{i=1}^{m} e_i$. Then the estimated events test statistic is given by

$$C_\lambda = (\mathbf{r} - \mathbf{p})^{\mathrm{T}} \mathbf{A}(\lambda)(\mathbf{r} - \mathbf{p}),$$

where $\mathbf{A}(\lambda) = \{a_{ij}(\lambda)\}$. One can consider several forms for the $a_{ij}(\lambda)$. Clearly, the choice of the form that $\mathbf{A}(\lambda)$ assumes will have an impact on the efficacy of this statistic. However, the magnitude of this effect and the sensitivity of the statistic to $\mathbf{A}(\lambda)$ have not been studied systematically. In practice the exponential threshold model has been used (Tango, 2000), such that $a_{ij}(\lambda)$ is defined as

$$a_{ij}(\lambda) = \exp\left\{-4\left(\frac{d_{ij}}{\lambda}\right)^2\right\},$$

where d_{ij} is the Euclidean distance between regions i and j. The problem with this method is that it requires specification of the parameter λ. Generally this is not known a priori, and several values of λ are tested, leading to multiple testing problems. In order to circumvent this problem, Tango (2000) developed the maximized excess events test (MEET). This statistic searches for the value of λ which gives the smallest p-value of the observed value of C_λ, denoted c_λ, as follows,

$$P = \min_{\lambda} \Pr\{C_\lambda > c_\lambda | H_0, \lambda\}.$$

This is implemented by allowing λ to assume discrete values near zero up to about half of the size of the study area and performing a line search over these values of λ. Monte Carlo simulation methods are used to obtain the null distribution of P.

8.3.2 The Interpoint Distribution Function and the *M* Statistic

The M statistic uses the interpoint distance distribution and its empirical cumulative distribution function (ecdf) to perform inference. Consider a spatial distribution $P(\mathbf{x})$ defined over a bounded region of the plane. Let the point distribution over the region be absolutely continuous, so that for two independent and identically distributed points $\mathbf{x}_1$ and $\mathbf{x}_2$ in the region, $P(\mathbf{x}_1 = \mathbf{x}_2) = 0$. For any such point distribution P, if one defines a nonnegative distance (or dissimilarity) function d, then the random variable $D = d(\mathbf{x}_1, \mathbf{x}_2)$ has some distribution $P_D(d)$. We call D the interpoint distance between two independent points. The cdf $F(\cdot)$ of D is $F(d) = EI(d(\mathbf{x}_1, \mathbf{x}_2) \leq d)$, where $I(\cdot)$ is the indicator function and E denotes expectation with respect to the $P \times P$ distribution.

Extending the usual definition of an ecdf for random samples, one can define the ecdf of the interpoint distances associated with a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as

$$F_n(d) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} I(d(\mathbf{x}_i, \mathbf{x}_j) \leq d).$$

The quantity $\sqrt{n}(F_n(d) - F(d))$, considered as a stochastic process indexed by d, converges weakly to a Gaussian process (Silverman, 1976; Bonetti and Pagano, 2004a). Because of the very definition of a Gaussian process, this general result implies that for a fixed value d the cdf $F_n(d)$ has $\sqrt{n}$-convergence to $F(d)$.

More generally, consider the empirical cdf $F_n(\mathbf{q}) = (F_n(q_1), \ldots, F_n(q_k))$ computed at a finite number k of fixed values $\mathbf{q} = (q_1, \ldots, q_k)$. The cutoff

points q_j are typically chosen to be the $(j/k)100\,\%$ percentiles of the distribution of D. If the range of D is unbounded, we set $q_k = \infty$. Then, the weak convergence implies that the joint asymptotic distribution of the centered ecdf $\sqrt{n}(F_n(\mathbf{q}) - F(\mathbf{q})) = \sqrt{n}(F_n(q_1) - F(q_1), \ldots, F_n(q_k) - F(q_k))$ is asymptotically multivariate normal with covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{a,b}\}$, with

$$\begin{aligned}\sigma_{a,b} = {} & E[I(d(\mathbf{x}_1, \mathbf{x}_2) \le q_a, d(\mathbf{x}_1, \mathbf{x}_3) \le q_b)] \\ & - EI(d(\mathbf{x}_1, \mathbf{x}_2) \le q_a) EI(d(\mathbf{x}_1, \mathbf{x}_3) \le q_b).\end{aligned}$$

A number of standard test statistics can be used to evaluate the distance between $F_n(\cdot)$ and $F(\cdot)$ for hypothesis testing, but the lack of independence among observed distances between individuals precludes the use of standard statistics without using appropriate modifications.

The noted asymptotic normality suggests the following statistic to measure the distance between $F_n(\mathbf{q})$ and $F(\mathbf{q})$:

$$\tilde{M}(F_n(\mathbf{q}), F(\mathbf{q})) = (F_n(\mathbf{q}) - F(\mathbf{q}))^{\mathrm{T}} \boldsymbol{\Sigma}^{-} (F_n(\mathbf{q}) - F(\mathbf{q})),$$

a Mahalanobis-like statistic, where $\boldsymbol{\Sigma}^{-}$ is a generalized inverse (see Rao and Mitra, 1971) of the covariance matrix of the vector $F_n(\mathbf{q})$. For definiteness we use the Moore–Penrose generalized inverse. In applications we typically use an estimator of $\tilde{M}$: consider the quadratic form

$$M(F_n(\mathbf{q}), F(\mathbf{q})) = (F_n(\mathbf{q}) - F(\mathbf{q}))^{\mathrm{T}} \mathbf{S}^{-} (F_n(\mathbf{q}) - F(\mathbf{q})),$$

where $\mathbf{S}$ is the estimated covariance matrix, obtained by generating repeated samples of size n from an assumed null spatial distribution of the individuals over the region of interest. To calculate $\mathbf{S}$ we could also take repeated samples from historic data, if available. We note that the M statistic can also be computed when the data consists of counts recorded at a finite number of fixed locations (see Bonetti and Pagano, 2004a), with minor modifications. If these fixed locations are a result of a discretization of the individuals addresses, there is the possibility of a loss of power to detect deviations from the null geographic distribution.

An alternative definition of M can be given in terms not of the cumulative distribution function, but of its first differences at the subsequent bin counts along the distance axis. The ecdf and the cdf of D are therefore summarized by the observed proportions o_j and the expected probabilities $e_j = j/k$ within each of the bins, with $j = 1, \ldots, k$. The variance–covariance matrix in that case needs to be modified in the obvious manner, since the first differences are a linear combination of the values of the cumulative distribution functions.

As an alternative, a consistent estimator for the variance–covariance matrix $\boldsymbol{\Sigma}$ can also be constructed. The covariance matrix can be estimated consistently by the terms

$$\widehat{\sigma}_{a,b} = 4\left\{\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} h(\mathbf{X}_i,\mathbf{X}_j,\mathbf{X}_k;q_a,q_b) - \left[\frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} I(d(\mathbf{X}_i,\mathbf{X}_j)\le q_a)\right] \times \left[\frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} I(d(\mathbf{X}_i,\mathbf{X}_j)\le q_b)\right]\right\},$$

where

$$h(\mathbf{X}_i,\mathbf{X}_j,\mathbf{X}_k;q_a,q_b) = 6^{-1}\sum_{\rho}[I(d(\mathbf{X}_{\rho_1},\mathbf{X}_{\rho_2})\le q_a, d(\mathbf{X}_{\rho_1},\mathbf{X}_{\rho_3})\le q_b)]$$

is the symmetrized kernel computed over the collection $\rho = \{(\rho_1,\rho_2,\rho_3)\}$ of the six permutations of the indices (i,j,k) (see Bonetti and Pagano, 2004b). In the calculation of this estimator, for efficiency the triple sum should be implemented as a single loop by making use of (fast) matrix multiplications for the inner sums.

8.3.2.1 Example

As an example, consider points uniformly distributed on the unit square $[0,1]\times[0,1]$. The distribution of the interpoint distance between two such points is as described in Bartlett (1964). The approximate quantiles at probabilities (0.2, 0.4, 0.6, 0.8, 1) from that distribution are (0.2912, 0.4435, 0.5891, 0.7573, 1.4142). Using these as cutoff values, consider the empirical estimator of that cdf $F_n(q_h)$ at the deciles $q_h, h = 1,\ldots,5$. Note that $q_5 = 2^{1/2}$ is the largest possible interpoint distance on the unit square, and that the cumulative distribution function is always equal to one for that value, so that consideration of $F_n(d_h)$ at $d_h, h = 1,\ldots,4$ suffices.

Table 8.1 shows the asymptotic variance–covariance matrix ($\boldsymbol{\Sigma}^*$) of $n^{1/2}(F_n(d_1)-F(d_1),\ldots,F_n(d_4)-F(d_4))$, as estimated from 3000 samples of size 5000.

We then considered four sample sizes n = 100, 250, 500, and 1000. For each sample size we computed the estimator of the variance–covariance matrix $\boldsymbol{\Sigma}$ one hundred times, as described above. On the left-hand side of Table 8.2 we report, for each sample size and for each element of the matrix, the relative bias of the variance–covariance matrix estimator, computed as the difference between the average of the 100 matrices and $\boldsymbol{\Sigma}^*$, divided by $\boldsymbol{\Sigma}^*$. On the right-hand side of Table 8.2 we report, also for each sample size and for each element of the matrix, the coefficient of variation relative to $\boldsymbol{\Sigma}^*$, that is, the ratio between the standard deviation of each term as computed from the 100 matrices and $\boldsymbol{\Sigma}^*$.

The variance–covariance matrix estimator appears to be centered satisfactorily at the true (as estimated by $\boldsymbol{\Sigma}^*$) variance–covariance matrix of the ecdf of

Table 8.1 Estimated variance–covariance matrix $\boldsymbol{\Sigma}$ of $\sqrt{n}$ times the interpoint distance ecdf. The matrix is based on 3000 samples of size 5000.

	d_1	d_2	d_3	d_4
d_1	0.011	0.022	0.029	0.027
d_2	0.022	0.051	0.068	0.058
d_3	0.029	0.068	0.092	0.077
d_4	0.027	0.058	0.077	0.060

Table 8.2 Relative bias and coefficient of variation (relative to $\boldsymbol{\Sigma}^*$ in Table 8.1) of the estimator of $\boldsymbol{\Sigma}$.

Relative bias

$n = 100$

	d_1	d_2	d_3	d_4
d_1	−0.06	−0.05	−0.04	−0.03
d_2	−0.05	−0.05	−0.04	−0.02
d_3	−0.04	−0.04	−0.03	−0.02
d_4	−0.03	−0.02	−0.02	0.00

$n = 250$

	d_1	d_2	d_3	d_4
d_1	0.02	0.02	0.02	0.01
d_2	0.02	0.01	0.01	0.01
d_3	0.02	0.01	0.00	0.00
d_4	0.01	0.01	0.00	0.00

$n = 500$

	d_1	d_2	d_3	d_4
d_1	0.05	0.05	0.04	0.04
d_2	0.05	0.03	0.02	0.02
d_3	0.04	0.02	0.02	0.02
d_4	0.04	0.02	0.02	0.01

$n = 1000$

	d_1	d_2	d_3	d_4
d_1	0.04	0.04	0.03	0.03
d_2	0.04	0.03	0.02	0.02
d_3	0.03	0.02	0.02	0.02
d_4	0.03	0.02	0.02	0.02

Coefficient of variation

$n = 100$

	d_1	d_2	d_3	d_4
d_1	0.60	0.45	0.35	0.24
d_2	0.45	0.28	0.19	0.13
d_3	0.35	0.19	0.12	0.08
d_4	0.24	0.13	0.08	0.11

$n = 250$

	d_1	d_2	d_3	d_4
d_1	0.35	0.29	0.24	0.16
d_2	0.29	0.19	0.13	0.09
d_3	0.24	0.13	0.08	0.05
d_4	0.16	0.09	0.05	0.07

$n = 500$

	d_1	d_2	d_3	d_4
d_1	0.20	0.17	0.14	0.09
d_2	0.17	0.11	0.07	0.05
d_3	0.14	0.07	0.04	0.03
d_4	0.09	0.05	0.03	0.05

$n = 1000$

	d_1	d_2	d_3	d_4
d_1	0.13	0.11	0.09	0.06
d_2	0.11	0.07	0.05	0.03
d_3	0.09	0.05	0.03	0.02
d_4	0.06	0.03	0.02	0.03

the interpoint distance. The relative bias of the estimator is reassuringly small (less than or equal to 6 %) even for the smaller values of n. The variance of the estimator is such that the relative standard errors only fall below 20 % when the sample size n is at least equal to 500. Lastly, it should also be noted there tends to be more bias and variability in the estimation of the variances and covariances that involve the cdf evaluated at small distances compared to larger distances.

8.4 SPATIO-TEMPORAL SURVEILLANCE: AN EXAMPLE

Although the focus of this chapter is on spatial methods, we may also consider the temporal aspect of a surveillance data stream, as well as methodology that integrates the spatial and temporal information for the purposes of surveillance. This integrated approach is often referred to as spatio-temporal surveillance. In this section we illustrate the spatial methods described above with a real data set, and then continue our example with this data set to illustrate the utility of temporal and spatio-temporal methodology. To simplify the exposition we only consider the day-to-day behavior of the system and ignore any memory from one day to the next. Clearly, a real system would have memory beyond a single day (see Reis *et al.*, 2003).

The data set that we use to illustrate these methods was collected by a large health provider in Massachusetts. As patients arrive for emergency care, their cases are geocoded (typically the residential or billing address of the patient), and this information is centralized electronically on a daily basis. In the interest of anonymity, in this exposition the spatial data provided has been aggregated by census tract, with jittering to further conceal the true locations of the individual patients involved. We consider a subset of these electronic data, consisting of upper respiratory infections (URIs) arriving at emergency and urgent care departments for this provider between the dates of January 1, 1996 and October 30, 1999, a stretch of 1399 days or nearly four years of daily data.

This data stream thus provides the temporal patterns of disease in the form of the number of cases arriving each day, as well as the spatial patterns of disease produced as the locations of patients change over time. For all further analysis, we have divided the data into three groups according to the day of the week: weekends and holidays, days after weekends and holidays, and the remaining days in the week. This was necessary because some of the locations provided were closed on weekends and holidays, leading to a stratification of case volume and spatial patterns on different days of the week.

Since there were no known bioterrorism attacks in Massachusetts during the period of study, for the purpose of evaluating methods, we chose to augment the real data with simulated clusters. To this end, we created three new data sets. For two, we chose six adjacent census tracts in close proximity and added one additional URI per day per tract, for a total of six additional cases per day. For

brevity, we call such a simulated signal a 'cluster'. The two data sets contain clusters centered around census tracts labeled 477 and 179 respectively, and we refer to the corresponding data sets accordingly. In a third round of simulations, we added both clusters of six cases, for a total of 12 additional cases (six each in the two separate locations; see Figures 8.2 and 8.6).

8.4.1 Temporal Component

Before consideration of the spatial and spatio-temporal surveillance of the Massachusetts data, we briefly describe an approach to the temporal surveillance of such data. Rather than describe the variety of methods available (see Chapter 2), we simply describe the modeling approach that we have taken with these particular data.

Let $N(t)$ denote the daily case volume of URIs across the entire study area, $1 \leq t \leq 1399$. The time series $N(t)$ shows several trends which make modeling challenging. Both the mean and variance of $N(t)$ have strong seasonal and day-of-week variation (see Figures 8.3 and 8.4). Closure of some locations on weekends and holidays further complicates modeling and analysis.

In order to construct a model for $N(t)$ we first used standard linear regression methods to fit a deterministic component for the mean expected case count. This is essentially the approach described in Brookmeier and Stroup (2004,

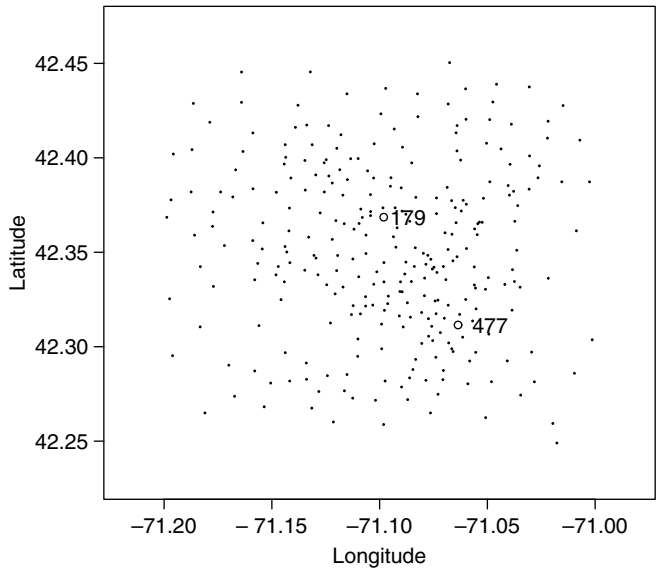


Figure 8.2 Locations of the census tracts with superimposed clusters, relative to the remaining census tracts in the Massachusetts data set.

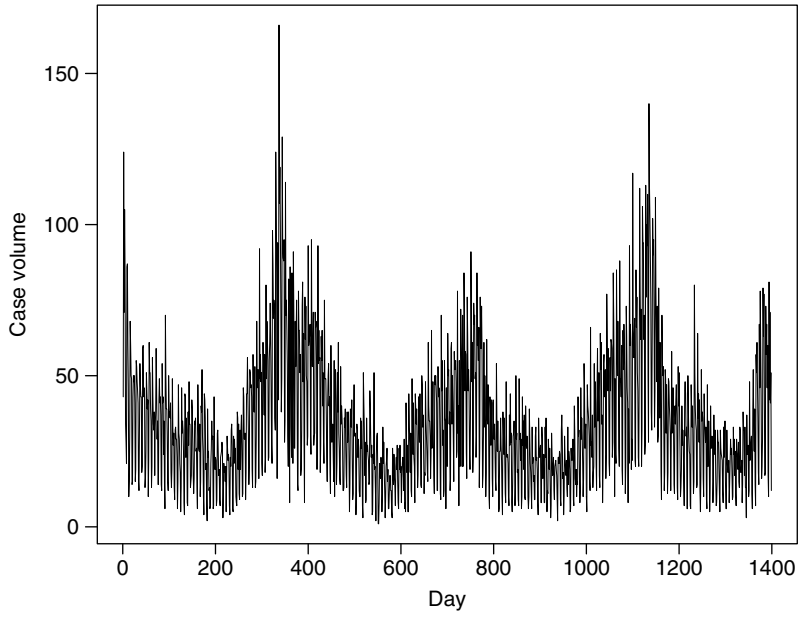


Figure 8.3 The time series $N(t)$ exhibits a seasonal pattern in addition to occasional sharp increases in the winter months.

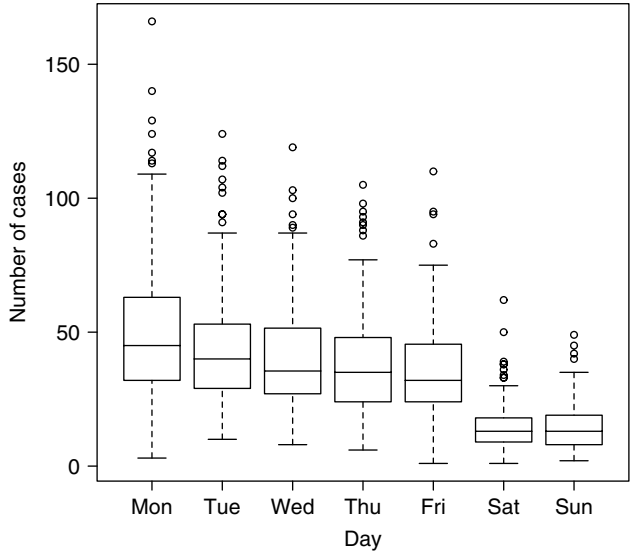


Figure 8.4 Number of cases by day of week.

pp. 203–231). The linear model included several harmonic terms for the characteristic seasonal effect on URIs, as well several indicator variables corresponding to identifiable day-of-week effects. An additional indicator for the months of December through February (the well-known 'flu season') was included to account for the frequent excess of cases in these months. The day-of-week variation is exhibited in both first and second moments, so after subtracting the fitted values from the observed data we divided by the daily standard error in order to standardize the residuals. Denote by $\eta(t)$ the time series constructed from each resulting data point; we can think of $\eta(t)$ as a standardized residual departure from the baseline mean.

The residuals $\eta(t)$ are characterized by a high degree of autocorrelation. Our goal is to model the residuals, resulting in a predicted value for $N(t)$ that can be compared to the observed value. Taking a simple approach, we used a first-order autoregression (AR(1)) to model the autocorrelation. After inclusion of the autoregressive terms the standard deviation of the residuals was reduced by nearly 10 % from 0.923 to 0.838. Thus the full model is:

$$\begin{aligned} N(t) \equiv\ & \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{365}\right) + \alpha_2 \sin\left(\frac{2\pi}{365}\right) + \alpha_3 I(\text{wkend}) \\ & + \alpha_4 I(\text{Monday}) + \alpha_5 \cos\left(\frac{2\pi}{30}\right) + \alpha_6 \sin\left(\frac{2\pi}{30}\right) \\ & + \alpha_7 I(\text{flu season}) + \text{interaction terms} + \in(t), \\ \eta(t) \equiv\ & \frac{\in(t)}{\sigma} = \beta\eta(t-1) + \xi(t). \end{aligned}$$

Thus we can view $N(t)$ as a test statistic for temporal surveillance, where we consider any observed N falling in a critical region to raise an alarm.

8.4.2 Bivariate Test Statistic

In order to fully utilize the available information, we consider using a bivariate test statistic, composed of the two statistics, the M statistic calculated daily and $N(t)$, described above. In an abuse of notation we refer to their daily product as NM, dropping the reference to time, t.

$N(t)$ allows us to calculate a residual value for the number of cases arriving, based on the time series prediction for that day, and the residuals are distributed approximately normally. Simultaneously, $\log(NM)$ can be used to evaluate the deviation of the spatial distribution of cases from normalcy. Theoretical research (Bonetti and Pagano, 2004a) shows that asymptotically, NM follows a χ^2 distribution with degrees of freedom not dependent on N. This has two immediate consequences. First, $\log(NM)$ and N are asymptotically independent. Second, the log of a χ^2 variable is approximately normal, therefore $\log(NM)$ is approximately normal as well. Due to the independence of $N(t)$ and NM, standard techniques from multivariate analysis are applicable for construction of an elliptical

or other appropriately chosen rejection region for a bivariate normal population at prespecified alpha level that we can use to test for deviations from normalcy.

Another approach we can use is to consider bivariate values in the event of a bioterrorist attack; in this case there is an optimal discriminator (the quadratic classification rule) between two bivariate normal populations (Johnson and Wichern, 2002, Section 11.3) in order to decide if an attack has occurred. This rule defines a classification boundary via a quadratic form (defined by means and covariances of the training set populations) in order to assign new observations to one of the existing populations. The two populations in this case would be the bivariate distribution under the null, and the modeled bivariate distribution under the alternative of a biological attack. The classification rule is a quadratic form that, given $\eta(t)$ and $\log(NM)$ on a particular day, assigns this observation to either the null or alternative population. In Figure 8.5 we illustrate a typical case of the null and alternative populations, together with the boundary of the discriminator. This rule minimizes the expected error of misclassification. The false positive rate can be controlled by shifting the quadratic boundary appropriately, as determined via simulation or resampling of the historical record.

8.4.3 Power Calculations

With each of 1399 days of data, we added a simulated cluster to each day and compared the power of temporal, spatial, and spatio-temporal statistics to

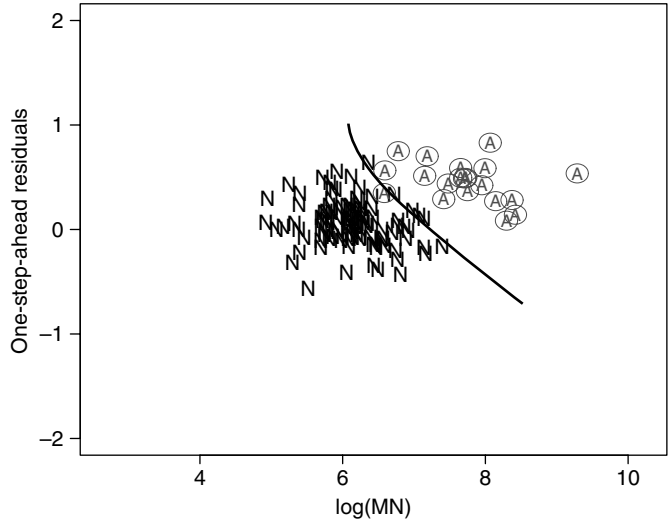


Figure 8.5 Subset of the null (labeled N) and alternative (circled A) populations used to train the quadratic discriminator. A portion of the classification boundary is also shown.

detect such a signal. Power calculations were performed separately for each of the three daily categories, since prediction and behavior differ within each of these categories. We define power as the ratio of daily detections to the total number of days observed. Using the statistics discussed above, we calculated power based on the simulated disease signal in our three constructed data sets.

For the univariate test statistic based on time series modeling, we calculated power to detect a temporal signal in the data. Using the first 1096 days (three full years) to train the model, power was calculated to detect an additional 6, 9, or 12 cases added to the case counts of the final 303 days of data. Results are shown in Table 8.3 (these results are not stratified by location since the statistic depends only on the number of cases and not the spatial locations). We see that the power to detect a disturbance increases as the size of the disturbance increases, as it should.

For the three data sets of clustered data, we calculated values of the two test statistics on each of the 1399 days available (the code for MEET was provided by Toshiro Tango) and compared the value of the statistics to their respective distributions under the null hypothesis of no clustering. These null distributions are determined using resampling methods with the unaltered historical data. Results are shown in Table 8.4.

When considering a bivariate statistic we generate a training sample based on a modeled signal consisting of six cases near location 179, superimposed on each of the first 1096 days of data. The temporal test statistic is N, and for the spatial statistic we choose $\log(NM)$. Here, the transformation of the test statistic M to $\log(NM)$ standardizes the distribution for differing numbers of cases, and the logarithm gives a statistic that is roughly normally distributed.

Following this approach, we generate two distinct bivariate normal populations of values, consisting of η residuals together with $\log(NM)$ calculations for 1096 days of null and alternative training data. For the simulated clusters in the final 303 days of data, we calculate the corresponding bivariate test statistic and use the quadratic classification rule to place each day's simulated cluster into the null (no signal) population or the alternative (signal present) population. Power in this case is the number of clusters classified in the alternative over total number of observations. Results are shown in Table 8.4.

The power of the univariate statistic $N(t)$ which detects deviations from the predicted number of cases on a daily basis illustrates the difficulties of time series modeling for public health surveillance. Rather than rely on a simple

Table 8.3 Power to detect temporal clusters.

	Hol./wkends 94 days	Wkdays 165 days	Day after hol. 44 days	Overall weighted average
$N+6$	0.266	0.248	0.250	0.254
$N+9$	0.479	0.315	0.318	0.366
$N+12$	0.755	0.467	0.364	0.541

Table 8.4 Power to detect various cluster models. Group 1 refers to the cluster centered at tract 179, with an additional case added to tracts 179, 182, 183, 184, 191, and 192. Group 2 refers to the cluster centered at tract 477, with an additional case added to tracts 477, 478, 479, 480, 482, and 484.

		Wkends/hol. 438 days	Wkdays 749 days	Day after hol. 212 days	Overall weighted average
Group 1	MEET	0.813	0.194	0.099	0.373
$N+6$	M statistic	0.495	0.362	0.250	0.387
	Bivariate statistic	0.585	0.394	0.227	0.429
Group 2	MEET	0.769	0.085	0.066	0.296
$N+6$	M statistic	0.475	0.295	0.222	0.340
	Bivariate statistic	0.543	0.358	0.250	0.399
Groups 1 & 2	MEET	0.986	0.427	0.226	0.571
$N+12$	M statistic	0.568	0.430	0.325	0.457
	Bivariate statistic	0.904	0.606	0.386	0.667

autoregression, results could be improved by considering a multivariate periodic autoregression (Pagano, 1978). For both the MEET and M statistics, power is consistently higher for weekends and holidays than for other types of days. On weekends and holidays, mean case volume is much lower at the clinics. This leads to a higher signal-to-noise ratio in the simulated data and thus a more detectable spatial aberration when a cluster of fixed magnitude (6 or 12 cases) is added to the data. The MEET is especially sensitive to this type of aberration, as adding one case to a region where the expected number of cases is minuscule greatly inflates the statistic. Although the MEET has especially high power on weekends and holidays, the power of the MEET statistic declines much more rapidly than M as the case volume increases.

Both spatial statistics perform quite well in detecting the simultaneous 179/477 clusters. Superior performance on data sets containing multiple clusters is a characteristic typically shared by distance-based methods of cluster detection as compared to other spatial methods (Kulldorff *et al.*, 2003; Ozonoff *et al.*, 2004).

The bivariate statistic shows promise for an effective use of available data. The power results show that for these simulated clusters, the bivariate approach outperforms the use of purely temporal or purely spatial information.

8.5 LOCATING CLUSTERS

Having decided that the M statistic indicates that there is a deviation from the null distribution, the next step is to determine whether this deviation is caused by an exogenous cluster, or clusters, of individuals, and to locate this cluster or clusters.

There are not too many principled guides in the literature for locating clusters, especially if there is more than one cluster (see Lawson and Denison, 2002, for discussion of small-area data). Fortunately, the M statistic related methods based on distances suggest a natural method for locating clusters.

Here we concentrate on the spatial location problem, leaving the time component to later studies. Let $\{s_i\}_1^n$ be the locations of the individuals, and $\mathbf{D} = (d_{ij})$ be the $n \times n$ distance matrix where d_{ij} is the distance between s_i and s_j. In Figure 8.6 we see a distribution of points in a plane with two clusters of points superimposed. We presume that the null hypothesis about the distribution of $\{s_i\}_1^n$ has been rejected, and we now search to locate the exogenous cluster or clusters that presumably were the reason for rejection.

Consider each row of the matrix of distances, $\mathbf{D}$. Fixing on row i, the $d_{ij}, j = 1, \ldots, n$, are a sample of independent distances from the point s_i. From the null distribution, either theoretically or via Monte Carlo, we can determine what the null distribution of points from s_i should be. Then we can compare this distribution with the observed distribution of distances from s_i, and presumably will be able to discern the points s_j that belong to an exogenous cluster. Of course, it is too much to hope that for a single i we will pick these s_j with any confidence, but if we gather information from all the s_i, one at a time, we can use the aggregate information to identify the clusters. This is the intuitive description of the method we use.

Choose a row $i, i = 1, \ldots, n$, and determine the null distribution of the distances from s_i. This may have to be achieved by resampling points from the null distribution of points. Having determined this distribution, then, for a fixed

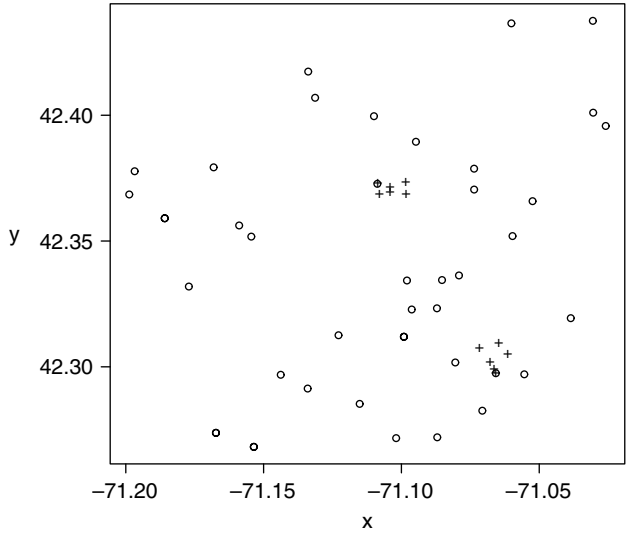


Figure 8.6 Typical distribution of cases for the Massachusetts data set. Superimposed clusters are denoted by a '+'.

integer $k > 1$, determine the k equispaced quantiles for the distribution, and hence create k equiprobable bins to receive the d_{ij}. The d_{ij} associated with the s_j in exogenous clusters will give rise to bins with excessive counts. These s_j from exogenous clusters will have a similar impact for other i, and so a record can be kept of the s_j which appear in bins that are oversubscribed, as we consider each row i.

To aggregate over the rows, consider a scoring system. For each row i, let $\text{score}(i,\ j) = 1$ if s_j belongs to an oversubscribed bin. Then with each point s_j associate the

$$\text{score}(j) = \sum_{i=1}^{n} \text{score}(i, j).$$

Subsequently look at these scores to determine which ones are too large. These are the ones that can be tagged as belonging to the exogenous clusters.

The binning process described above is a traditional way of determining goodness of fit. One of its disadvantages is that the underlying distribution of points is continuous, whereas the binning is inherently discrete. This may manifest itself in points which are in an exogenous cluster but, because of the discrete character of the bins, fall just next door to an oversubscribed bin. To overcome this effect of discretization, we compromise by defining a score function which takes the value one for an oversubscribed bin, and the value 0.50 for the bins on either side of the oversubscribed bin. If the oversubscribed bin is on a boundary (either it is the first or last bin) then it will only have one neighbor.

This scoring system is, of course, one of an infinite number of scoring systems one can devise.

The only remaining unknown is the definition of what we mean by an oversubscribed bin. For a fixed i we can consider the n distances d_{ij} as a sample of independent and identically distributed variables. Thus the counts of the numbers falling into the k bins can be considered as the realization of a multinomial distribution of size n with equiprobable cells, each with probability $1/k$. We can determine a bin to be oversubscribed if the number of distances in the bin exceeds n/k by two standard deviations. Other cutoffs can be entertained.

The last step is then to determine how large $\text{score}(j)$ must be before we declare s_j to be a location within a cluster. A cutoff can be determined via Monte Carlo methods.

This method is exemplified below.

We now return to the example taken from data from a large health care provider in Massachusetts. We wish to show the efficacy of the above method in locating clusters in the data. Again, we subset the data by the day of the week (weekends/holidays, day after weekend/holiday, and weekdays).

As a measure of the adequacy of this method, we borrow from methods used in diagnostic testing and report estimates of sensitivity and specificity. Suppose our method tags b regions as a comprising cluster or clusters. In our setting,

we define sensitivity as the probability of detecting the regions that actually constitute the cluster(s). Specificity is defined as the probability that the regions that are not tagged are not in the cluster(s).

Table 8.5 provides a summary of the results of this method applied to the Massachusetts data. Here we give results for detecting the three cluster models (region 179, 477, and a cluster in 179 and 477 simultaneously) for the three different types of days (weekends and holidays, weekdays, and days after holidays). We consider three different significance levels for determining the cutoff for the scores: 0.05, 0.10 and 0.15. Increasing the significance has the effect of increasing sensitivity and decreasing specificity. However, specificity remains high in all scenarios.

In the surveillance setting, we would often be satisfied with detecting at least part of the cluster. Therefore, we can imagine a much more forgiving definition of sensitivity as the probability of detecting at least one of the cluster regions. In other words, we are not concerned that we detect all of the regions in the cluster, as health professionals alerted by the alarm would likely fan out from investigating that region to surrounding areas that would likely comprise the cluster. Were we to use this as a measure of efficacy, the method would undoubtedly appear even better. The last column probably best approximates the ubiquitous 95 % specificity.

On the other hand, a high specificity is also desirable. A typical system may require a decision each day and missing an outbreak might lead to disastrous consequences; but, by the same token, too many false positive alarms might lull the analyst into treating the system with skepticism and subsequently missing a valid alarm. It is thus comforting to see the high specificities in Table 8.5.

Table 8.5 Sensitivities and specificities for locating the clusters with the Massachusetts data set. Results are given with the cutoff for the score being determined as the 95th, 90th or 85th percentiles of the empirical distribution of the scores.

	95th		90th		85th	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
Group 1, $N+6$						
Hol./wkends	0.76	0.99	0.84	0.99	0.89	0.99
Wkdays	0.61	0.98	0.73	0.96	0.89	0.95
Day after hol.	0.59	0.97	0.68	0.95	0.79	0.94
Group 2, $N+6$						
Hol./wkends	0.77	0.99	0.84	0.99	0.88	0.98
Wkdays	0.60	0.98	0.75	0.96	0.82	0.98
Day after hol.	0.63	0.97	0.72	0.95	0.81	0.94
Group 1 and 2, $N+12$						
Hol./wkends	0.45	0.99	0.61	0.99	0.76	0.98
Wkdays	0.45	0.98	0.62	0.96	0.72	0.95
Day after hol.	0.48	0.97	0.63	0.95	0.73	0.93

8.6 CONCLUSION

Spatial surveillance has as it goal the recognition of deviations from the 'normal' distribution of events in a region. We have shown the utility of distance methods in achieving the stated objectives of spatial surveillance. Distance methods are characterized as statistical methods that utilize the distances between events in detecting aberrations in spatial behavior.

Two statistics are immediately applicable for use in spatial surveillance. The MEET statistic is widely recognized in contemporary literature and practice. It is only applicable to aggregated data, such that the data consists of counts of events within each spatial region. The statistic has been shown to have good power, especially when case volume is low relative to the increase in the number of cases attributed to a cluster.

Much of the focus of this chapter has been on the M statistic. This statistic seeks to detect changes in the distribution of the interpoint distances by considering a statistic that is similar to the Mahalanobis distance. This can be easily applied to data streams with either aggregated or exact location information.

The M statistic can further be extended to incorporate temporal trends in the data stream, as exemplified by the bivariate statistic illustrated above. Not surprisingly, this leads to an increase in power for the detection of anomalies in the data, as one would expect such a disturbance to represent an increase in case volume, as well as a disturbance in the spatial distribution.

Spatial surveillance requires not only an alarm to be sounded when a disturbance has occurred, but also some indication of the location and shape of the disturbance to facilitate further investigation and efficient methods to control and diffuse the source of the disturbance, inhibiting further spread to the population. We have shown an effective method for locating the source of the signal causing an alarm with the M statistic. Such methods are crucial to the success and efficacy of a surveillance system.

Distance methods are a natural tool for spatial surveillance. The issues presented by this problem require methods that incorporate information extending beyond a simple mean or other traditional statistics that are often employed when confronted with data on the real line. The increased dimensionality and correlation of the data call for methods that can distinguish between normal and abnormal behavior for an infinite number of scenarios that are not easily characterized or classified. Distance methods appear to have the potential to capture the complexity of a spatial distribution. Further, statistics such as M allow for incorporation of additional information into the data stream, such as temporal trends.

It would be unfair to fail to recognize the many difficulties that arise when working with distance data. As has been illustrated above, these methods still require much refinement and further research. Working with the dependencies intrinsic in interpoint distances is complicated and requires further rigorous investigation. Much of these complexities can be circumvented in practical

implementation via resampling methods. However, to better understand, generalize, and optimize these statistics, greater theoretical understanding is needed. It has also been shown that theoretical developments can lead to an increase in efficiency and decreases in computation time. This is exemplified by the estimation of the variance–covariance matrix for the M statistic.

We advocate the use and further development of distance methods in spatial surveillance. These methods have been shown to be effective and complementary when compared to quadrat methods, such as the spatial scan statistic. Further, the M statistic has great promise in detecting spatial aberrations that extend beyond simple circular clusters. Ideally, a surveillance system would make use of multiple statistical methods, coupled with vigilant and timely epidemiological investigation of alarms raised by these automated methods.

ACKNOWLEDGMENTS

This research was partially funded by grants from the National Institutes of Health, RO1AI28076, and the National Library of Medicine, RO1LM007677.

9

Multivariate Surveillance

Christian Sonesson and Marianne Frisén

9.1 INTRODUCTION

There are several important multivariate surveillance problems in public health. Public health surveillance is defined by Thacker (2000) to be the ongoing systematic collection, analysis, interpretation, and dissemination of health data for the purpose of preventing and controlling disease, injury, and other health problems. Spatial public health surveillance is multivariate since several locations are involved. There are many important spatial surveillance situations. One concerns the detection of an increased incidence of a disease when there is a spatial dependency between locations (Lawson, 2004), another the departure from spatial independence due to disease clusters (Kulldorff, 2001). The case of child leukemia has been the subject of several retrospective studies. The spatial information in the data is a vital component for the construction of efficient surveillance methods. For example, a local change or a spreading disease will not be detected unless the spatial information is used properly.

In both national and international programs data which are multivariate for reasons other than spatial are routinely collected in several areas of public health. One example is syndromic surveillance where several symptoms are studied (see Lober *et al.*, 2003). Further examples of different public health surveillance data sources are given by Stroup *et al.* (2004). The increased interest in surveillance methodology in the USA in the wake of the 9/11 terrorist attack is notable. Several new types of data are now being collected, such as nurse hotline calls, poison center calls and over-the-counter sales. Since the data collected involve several related variables, this calls for multivariate surveillance techniques. New types of surveillance systems have also been developed recently. These include EARS, ESSENCE, RODS and WSARE. A review of some systems is given by Lober *et al.* (2002).

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

To be able to conduct effective public health surveillance, efficient statistical techniques must be used. Most of the work on multivariate surveillance can be found in the general statistical literature or in journals aimed at industrial applications, although the number of publications in journals on public health is increasing rapidly. We will focus on some principal approaches taken for the construction of multivariate surveillance methods. These general approaches can be applied in a public health context and do not depend on the distributional properties of the process in focus. Note specially that both continuous and discrete data can be mixed. Even though all the general approaches described are valid for all types of processes, some are more natural to use for mixed data and published formulas are often expressed in terms of specific processes. Reviews on multivariate surveillance methods can be found in, for example, Alt (1985), Basseville and Nikiforov (1993), Wierda (1994), Lowry and Montgomery (1995), and Ryan (2000).

In Section 9.2 we will give some specifications. In Section 9.3 different approaches to the construction of multivariate surveillance methods are described and exemplified. We will give several examples of multivariate surveillance methods applied in diverse public health areas, especially spatial ones. Several types of multivariate counterparts to the univariate Shewhart, exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) methods have been proposed. In most previous work, a multivariate normal distribution has been assumed for the vector of random variables. This assumption can be questioned in many public health settings, although it might be argued that at least the marginal densities can be approximated by a normal distribution due to large sample approximations. The justification of proposed methods hinges on the distributional assumption in some, but not all, cases. In Section 9.4 we discuss the special challenge in evaluating multivariate surveillance methods and also the concept of optimality in a multivariate setting. We make some concluding remarks in Section 9.5.

9.2 SPECIFICATIONS

The term 'statistical surveillance' was defined in Chapter 3, and we state it here again for the multivariate case. We denoted by $\mathbf{Y} = \{\mathbf{Y}(t),\ t = 1, 2, \dots\}$ the multivariate process under surveillance. At each time point, t, a p-variate vector $\mathbf{Y}(t) = \left(Y_1(t),\ Y_2(t), \dots, Y_p(t)\right)^{\mathrm{T}}$ of variables is observed. The components of the vector might be, for example, the number cases of a certain disease in p different areas. When the process behaves as earlier and no change has occurred $\mathbf{Y}(t)$ has a certain distribution, for example, with a certain mean vector $\boldsymbol{\mu}_0$ and a certain covariance matrix $\boldsymbol{\Sigma}_Y$. The purpose of the surveillance method is to detect a deviation to a changed state as soon as possible in order to take preventive actions. If we denote the current time point by s, we want to determine whether or not a change in the distribution of $\mathbf{Y}$ has occurred

before s, that is, to discriminate between the events $\{\tau \leq s\}$ and $\{\tau > s\}$, where τ denotes the time point of the change. In order to do so, we can use all available observations of the process $\mathbf{Y}_s = \{\mathbf{Y}(t),\ t \leq s\}$ to form an alarm statistic denoted by $p(\mathbf{Y}_s)$. The surveillance method triggers an alarm, indicating that the change has happened in the process, at the first point in time when $p(\mathbf{Y}_s)$ exceeds an alarm limit $G(s)$.

9.3 APPROACHES TO MULTIVARIATE SURVEILLANCE

We will discuss several different approaches to the construction of multivariate surveillance methods. A discussion in connection with the CUSUM method can be found in Chapter 6. First we describe some common techniques to reduce dimensionality. Then we describe the approach of scalar accumulation where the components of the vector of observations are transformed into a scalar statistic for each time point before the accumulation over time. We also describe the approach in which parallel univariate surveillance methods are used for each component variable and an alarm is triggered if any of the univariate methods triggers an alarm in accordance with the union–intersection principle. Another approach is the vector accumulation approach. Here the alarm statistics of the parallel surveillance methods for each component variable are combined to form a general alarm statistic. In both the vector and the scalar accumulation approaches, usually the correlations between the variables are used in the transformation. The final approach is to jointly handle the multivariate nature (e.g. spatial) of the observational vector and the different time points simultaneously while aiming to satisfy some global optimality criterion. The notation adopted in the literature for the various methods is not uniform, but here we have followed the one most commonly used.

9.3.1 Reduction of Dimensionality

The detection ability of a multivariate surveillance method deteriorates as the number of variables increases (see Runger *et al.*, 1999). This is true unless some structure focuses the detection ability. One way to reduce the dimensionality is to consider the principal components instead of the original variables, as proposed and discussed by, among others, Jackson (1985), Mastrangelo *et al.* (1996), Kourti and MacGregor (1996) and Scranton *et al.* (1996). The principal components consist of linear combinations of the original variables. One attractive feature of the principal components is the orthogonality among them. However, as pointed out in Lowry and Montgomery (1995), unless the principal components have clear interpretations it might be difficult to draw any reasonable conclusions from a surveillance method based on them. In Runger (1996) an alternative transformation, using so-called U^2 statistics, was introduced to allow

the practitioner to choose the subspace of interest (see also Runger *et al.*, 1999). Another alternative is to use projection pursuit (Ngai and Zhang, 2001; Chan and Zhang, 2001). After reducing the dimensionality any of the approaches for multivariate surveillance described below can be used. Hawkins (1991) suggested a regression adjustment for each of the individual variables and also suggested how these could be used by the different approaches depending on whether the aim is general, group or individual conclusions. Rosolowski and Schmid (2003) use the Mahalanobis distance to reduce the dimensionality of the statistic, thus expressing the distance from the target of the mean and the autocorrelation in a multivariate time series.

9.3.2 Reduction to One Scalar Statistic for Each Time

The most radical and most commonly used reduction of the dimension is to summarize the components for each time point into one statistic. One natural reduction when dealing with multivariate normal variables is to use the T^2 statistic suggested by Hotelling (1947). The Hotelling T^2 statistic is defined as

$$T^2(t) = (\mathbf{Y}(t) - \boldsymbol{\mu}_0(t))^{\mathrm{T}} \mathbf{S}_{\mathbf{Y}(t)}^{-1} (\mathbf{Y}(t) - \boldsymbol{\mu}_0(t)),$$

where the sample covariance matrix $\mathbf{S}_{\mathbf{Y}(t)}$ is used to estimate $\boldsymbol{\Sigma}_Y$. When $\boldsymbol{\Sigma}_Y$ is regarded as known, the statistic has a χ^2 distribution and is referred to as the χ^2 statistic. Regression and other linear weighting as a reduction method is discussed by Healy (1987), Hawkins (1991), and Kourti and MacGregor (1996). Since the observations at each time point consist of a vector, we can first transform the vector from the current time point into a scalar statistic, which we accumulate over time. In Sullivan and Jones (2002) this is referred to as 'scalar accumulation'. One example is when, for each time point, a statistic representing the important aspects of the spatial pattern is constructed from a purely spatial analysis. This statistic is then used in a surveillance method. The reduction to a univariate variable can be followed by univariate monitoring of any kind.

Originally, the Hotelling T^2 statistic was used in a Shewhart approach, and this is sometimes referred to as the Hotelling T^2 control chart. An alarm is triggered as soon as the statistic $T^2(t)$ is large enough. When $\boldsymbol{\Sigma}_Y$ is regarded as known, one can apply the same procedure to the χ^2 statistic. How to choose alarm limits for these methods is discussed by Lowry and Montgomery (1995). For multivariate binomial data, Lu *et al.* (1998) proposed a method based on a weighted sum of the number of adverse units for each of the components for each time point. The surveillance method used for this weighted sum was a Shewhart one. Other applications of Shewhart methods are described by Runger (1996) and Kang and Albin (2000). Note that over time there is no accumulation of the observation vectors if the Shewhart method is used. Only

the observations taken at the current time point are used. The lack of use of previous observations by the Shewhart method applied to the $T^2(t)$ statistic (or any other derived statistic) in the multivariate case will result in an ineffective method for detecting small and moderate changes in the process, which is also the case when using the Shewhart method in the univariate setting.

To achieve a more efficient method, all previous observations should be used in the alarm statistic. There are many suggestions of combinations of different dimension reductions with different monitoring methods. Crosier (1988) suggested calculating the Hotelling T variable (the square root of $T^2(t)$) and using this as the variable in a univariate CUSUM method, making it a scalar accumulation method. In this case the alarm statistic used is $S_t = \max(0,\ S_{t-1} + T(t) - k)$, where $S_0 \geq 0$ and $k > 0$, which is combined with a constant alarm limit.

Rogerson (1997) uses the scalar accumulation approach for spatial surveillance of count data. A modified version of the Tango statistic (see Tango, 1995) is calculated for each time point. This statistic is implemented in the CUSUM method to detect changes in the clustering tendency from an otherwise random spatial pattern. The method was evaluated both using simulations and by application to data on Burkitt's lymphoma in Uganda. In Rogerson (2001) the same approach was used for monitoring point patterns, but here instead a local Knox statistic (see also Knox, 1964) for space-time interaction was calculated for each time point. The surveillance methodology used was again the CUSUM method.

In public health, nonparametric methods, which do not rely on the assumption of normally distributed variables, are of special interest. A nonparametric scalar accumulation approach was used in Liu (1995), where the observation vector for a specific time point was reduced to a univariate index describing the distance to the center in the multivariate distribution. Here, ranks were used to get rid of the dependency on the distributional properties of the observation vector. Several methods were discussed for the surveillance step, including the CUSUM method. Yeh *et al.* (2003) suggested a transformation of multivariate data at each time to a distribution percentile and scalar accumulation versions of the EWMA method were suggested for detection of changes in the mean as well as in the covariance.

9.3.3 Parallel Surveillance

If one uses one univariate surveillance method for each of the individual components in parallel, one can combine these into a single surveillance procedure in several ways. These are referred to as combined univariate methods or parallel methods. The most common one is to signal an alarm if any of the univariate methods signals. This is thus a combination of methods using the union–intersection principle for multiple inference problems. Sometimes the Bonferroni method is used to control a false alarm error (see Alt, 1985). General

references about parallel methods include Woodall and Ncube (1985), Hawkins (1991), Pignatiello and Runger (1990), Yashchin (1994), and Timm (1996). The parallel approach is the one most commonly used in public health settings and also in general (see Stoumbos *et al.*, 2000).

For spatial disease surveillance with data from several different locations, Raubertas (1989) suggested a parallel version of the Poisson CUSUM method, where each location is monitored by one Poisson CUSUM. To account for the positive spatial correlation between nearby locations, the author suggested pooling within neighborhood observations. To detect emerging clusters of a disease parallel Shewhart methods were recently used by Kleinman *et al.* (2004), where estimates of the expected counts in each region were based on a generalized linear mixed model. Kulldorff (2001) introduced the space-time scan statistic (see also Chapter 7), which is an extension of the spatial scan statistic suggested in Kulldorff (1997). The space-time scan statistic scans different circular areas in space which are extended backwards in time to form space-time cylinders. The scanning is performed for a number of center points in space. All possible cylinders represent possible clusters in space-time with elevated incidence of a disease. An alarm statistic is formed by considering the cylinder which deviates most from space-time independence. This represents a parallel method taken over all possible space-time clusters represented by the cylinders. A discussion of the use of additional data, such as covariates, is found in Burkom (2003).

A form of scan statistic has been proposed by Kulldorff *et al.* (2003) for the case where the independent variable can be structured hierarchically. In this case the spatial distance is exchanged for the distance within the hierarchical structure. This scan statistic is used in Kulldorff *et al.* (2003) for an analysis of the relationship between occupation and death from silicosis. The hierarchical structure in this case consists of the classification of occupations according to the Classified Index of Industries and Occupations.

Several methods suitable for the type of nonnormally distributed data common in public health have been suggested. A generalized linear model was used in Skinner *et al.* (2003) to model independent multivariate Poisson counts. Deviations from the model were monitored by parallel Shewhart methods. For detection of an increased incidence of congenital malformations, a multivariate version of the sets method using data of malformations from multiple sources has been proposed (Chen, 1978; Chen *et al.*, 1982). In Stroup *et al.* (1988) a parallel method was used to detect excess deaths from pneumonia and influenza using monthly data from five different age groups. Multiple time series techniques were used to construct simultaneous forecasts which were compared to the actual data. Large deviations were considered evidence of an excess. In Steiner *et al.* (1999) paired binary results from surgical outcomes were monitored using a parallel method of two individual CUSUM methods based on the outcome variables. However, to be able to detect also small simultaneous changes in both outcome variables, the method was complemented with a third alternative, which signals an alarm if both individual CUSUM statistics are above a

lower alarm limit at the same time. The addition of the combined rule is in the same spirit as the vector accumulation methods presented in Section 9.3.4. Parallel CUSUM methods were also used by Marshall *et al.* (2004) to monitor the performance in general hospitals in the UK. The false alarms were controlled by using the false discovery rate from Benjamini and Hochberg (1995). For evaluation of the detection ability, the probability of successful detection (see Chapter 3) was used.

Several implementations of parallel methods can be found in newly developed health surveillance systems. The Early Aberration Reporting System (EARS) of the Centers for Disease Control and Prevention (CDC) is one example. Hutwagner *et al.* (2003) apply CUSUM methods to laboratory-based salmonella serotype data. There are over 2000 different serotypes and one CUSUM method is used for each serotype. Another type of surveillance method included in EARS is the 'historical limits method' of Stroup *et al.* (1993). This is a Shewhart method currently applied to case data from nine diseases (hepatitis A, hepatitis B, hepatitis C/non-A/non-B, legionellosis, meningococcal infections, measles, mumps, pertussis, and rubella). The method compares the number of cases reported in the last month to the average of the proceeding 5 years. The Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE) system has been developed by the US Department of Defense Global Emerging Infections System. It monitors more than 100 primary care and emergency clinics in the Washington, DC, area and is described by Pavlin (2003) and Burkom (2003). An extension is the ESSENCE II system described in Burkom (2003) and Lombardo *et al.* (2003). New types of data, including over-the-counter sales and absenteeism, are collected to aid in an earlier detection of a disease outbreak. Several detection methods are used, including a modified EWMA method run at county level, making it a parallel method over different counties. Also included are some of the methods in the EARS system as well as modified versions of the scan statistic for spatial clustering analysis. The cluster analysis is performed for one derived statistic for each of seven different syndrome groups, again using the parallel approach, this time over the different syndrome groups. The National Bioterrorism Syndromic Surveillance Demonstration Program described by Platt *et al.* (2003) is a nationwide US program under the auspices of the CDC and several other organizations. Part of the program is based on the ESSENCE system. One feature is a parallel method across different areas under surveillance, where in each area signals will be issued. In Chapter 10 and Wong *et al.* (2003) the WSARE (What's Strange About Recent Events) system is described. The purpose is to search a database of emergency department cases for anomalous patterns by comparing the events of the current day with the events which occurred exactly 5, 6, 7, and 8 weeks earlier. The method scans the database forming 2×2 contingency tables containing as columns the current day and previous days respectively and rows describing some characteristic of the cases in the database. A null hypothesis of independence of row and column attributes is tested using a χ^2 test or

Fisher's exact test, yielding a score for each of the row–column combinations. A p-value is obtained from the score via a randomization test. A determination of which of the p-values are significant is done by using the false discovery rate. The number of false positives in relation to the total number of tests is then controlled. In spirit the WSARE algorithm corresponds to a parallel surveillance method although it is presented in a hypothesis testing framework. Another surveillance system is the Real-time Outbreak and Disease Surveillance (RODS) system, which uses as data the main complaints collected (in free-text form) by nurses at the emergency departments during patient registration. A Bayesian classifier is used to classify the symptoms into any of eight different categories. Parallel surveillance methods are used for the different categories in different spatial regions (see Tsui *et al.*, 2003). Surveillance of large databases is also of interest in pharmacovigilance, where several different adverse drug reactions are recorded together with the drugs the patients have taken. Several new methods have been proposed under the rubrics of Bayesian data mining (DuMouchel, 1999) and Bayesian neural networks (Bate *et al.*, 1998; Orre *et al.*, 2000). What they have in common is the parallel approach over different drug–reaction combinations. Other examples of Bayesian methods can be found in Chapters 11 and 12.

9.3.4 Vector Accumulation Methods

In the parallel approach we accumulate the information of each component and decide at each time point to sound an alarm if any of the alarm statistics exceeds a limit. Another way to use the accumulated information on each component is to transform the vector of componentwise alarm statistics into a scalar alarm statistic and sound an alarm if this statistic exceeds a limit. This is naturally referred to as 'vector accumulation'.

Lowry *et al.* (1992) proposed a multivariate extension of the univariate EWMA method, which is referred to as MEWMA. This multivariate EWMA method uses a vector of univariate EWMA statistics $\mathbf{Z}(t) = \mathbf{\Lambda}\mathbf{Y}(t) + (\mathbf{I} - \mathbf{\Lambda})\mathbf{Z}(t-1)$, where $\mathbf{Z}(0) = \mathbf{0}$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$. An alarm is triggered at $t_A = \min\{t;\ \mathbf{Z}(t)^T \mathbf{\Sigma}_{\mathbf{Z}(t)}^{-1} \mathbf{Z}(t) > L\}$ for some alarm limit, L. In effect, MEWMA is just the Hotelling T^2 control chart applied to univariate EWMA statistics instead of the original data from only the current time point and is thus a vector accumulation method. Note that if we use $\lambda_1 = \ldots = \lambda_p = 1$ only the data from the current time point is used. In Lowry *et al.* (1992) the MEWMA method was shown to be an improvement over the Shewhart method applied to the χ^2 statistic. Optimal designs of a multivariate EWMA method with respect to the weighting parameters can be found in Lowry *et al.* (1992). Methods for computing the average run length of multivariate EWMA methods have been given in Runger and Prabhu (1996) using a Markov chain approach. In Bodden and Rigdon (1999) an integral approach is used. Instead of applying the multivariate EWMA directly to the original data, a dimension reduction transformation can first

be used. In Scranton *et al.* (1996) a subset of the principal components of the original process variables was used which, for the cases studied, improved the performance compared to the version with all process variables. The same approach, but on U-transformed data, was used in Runger *et al.* (1999). The application to the simultaneous monitoring of the mean and the variance is treated in Domangue and Patch (1991).

One natural way to construct a multivariate version of the CUSUM method would be to proceed as in the MEWMA case and construct the Hotelling T^2 control chart applied to univariate CUSUM statistics for the individual variables. For the pth variable, the CUSUM statistic for time point t would be $S_{t,p} = \max(0,\ S_{t-1,p} + Y_p(t) - k_p)$, where $S_{0,p} \geq 0$ and $k_p > 0$. The time point of an alarm for this procedure would be written as $t_A = \min\{t;\ \mathbf{S}(t)^T \boldsymbol{\Sigma}_{\mathbf{S}(t)}^{-1} \mathbf{S}(t) > L\}$, where $\mathbf{S}(t)$ is a vector containing the individual CUSUM statistics. One important feature of such a method is the lower barrier of each of the univariate CUSUM statistics (assuming we are interested in a positive change). No such method seems to have been suggested previously in the literature. Instead other approaches to construct a multivariate CUSUM have been taken. Crosier (1988) suggested the MCUSUM method where a statistic consisting of univariate CUSUMs for each component is used. This is similar to the MEWMA statistic, which corresponds to a vector accumulation method. However, the way the components are used is not the same. The suggestion by Crosier was to use the following recursive formula:

$$\mathbf{S}_t = \begin{cases} \mathbf{0}, & \text{if } C_t \leq k, \\ (\mathbf{S}_{t-1} + \mathbf{Y}(t) - \boldsymbol{\mu}_0(t))(1 - k/C_t), & \text{if } C_t > k, \end{cases}$$

where $\mathbf{S}_0 = \mathbf{0}$, $k > 0$, and

$$C_t = \{(\mathbf{S}_{t-1} + \mathbf{Y}(t) - \boldsymbol{\mu}_0(t))^T \boldsymbol{\Sigma}_{\mathbf{Y}(t)}^{-1} (\mathbf{S}_{t-1} + \mathbf{Y}(t) - \boldsymbol{\mu}_0(t))\}^{1/2}.$$

The time of an alarm is then given by $t_A = \min\{t;\ \mathbf{S}_t^T \boldsymbol{\Sigma}_{\mathbf{Y}(t)}^{-1} \mathbf{S}_t > L\}$ for some chosen alarm limit L.

An alternative way to construct a vector-accumulating multivariate CUSUM is given by Pignatiello and Runger (1990). The methods use different weighting of the variables. One important feature of these two methods is that the characteristic zero return of the CUSUM technique is made in a way which is suitable when all the components change at the same time point.

An example of a rank-based CUSUM method can be found in Qiu and Hawkins (2001), where the detection in a multivariate Poisson distribution is used as an example. Further developments can be found in Qiu and Hawkins (2003).

For spatial surveillance with data consisting of regional counts, Rogerson and Yamada (2004) constructed a multivariate CUSUM with a covariance structure corresponding to the spatial autocorrelation between regions. The method was compared to the use of parallel CUSUMs for each region. Changes in only a small number of regions were best detected using the parallel approach, while the opposite was true if the change occurred in many regions.

9.3.5 Simultaneous Solution

In this section we will approach time and space simultaneously in the analysis and aim at a total optimal surveillance. This total optimality is not guaranteed by the approaches described in the sections above. These start with a reduction either in time or space (or other multivariate setting). Sometimes a sufficient reduction will result in such a separation of the spatial and the temporal components. However, such a sufficient separation is not always available. The use of the sufficient statistic implies that no information is lost. Wessman (1998) proved that when all the variables change at the same time a sufficient reduction to univariate surveillance exists.

Healy (1987) analyzed the case of simultaneous change in a specified way for all the variables. He derived the CUSUM method both for a change in location and a change in the covariance. The results are univariate CUSUMs for a function of the variables. For detection of a change in location the solution is a linear combination of the individual variables. For detection of a change in covariance it is the T^2 statistic. The CUSUM method is minimax optimal. Thus, the multivariate methods of Healy (1987) are simultaneously minimax optimal for the specified direction when all variables change at the same time. Additional discussions of minimax optimality can be found in Lai (1998) and Lai and Shan (1999).

Another way to achieve a simultaneously optimal solution is by applying the full likelihood ratio method presented in Chapter 3. This can be used if the event to be detected is specified. The full likelihood ratio method was used in the case of clustering in a spatial log-linear model on a fixed lattice by Järpe (1999). The sufficient reduction of the spatio-temporal pattern resulted in one statistic to be calculated for each time point. This statistic was only based on the (spatial) observations for that particular time point. However, in some cases the sufficient reduction to a univariate statistic for each time point involves observations from different time points. An example is Järpe (2001), where the optimal method of detection of an increased radiation level was derived. In that application the shift process spread spatially with time. Wessman (1999) examined the case of different change points.

9.4 EVALUATION OF THE PROPERTIES OF MULTIVARIATE SURVEILLANCE METHODS

As pointed out by Kaufmann *et al.* (1997), a delay of one day in detection of and response to an epidemic due to a bioterrorist attack could result in the loss of thousands of lives and millions of dollars. This is one example of the importance of using surveillance methods that are as efficient as possible, if not optimal. Optimality is hard to achieve and even hard to define in the multivariate surveillance case, as described by Frisén (2003). For a specified type

of change in the process, general theory can be used to guide in the direction of optimal methods according to the different optimality criteria presented in Chapter 3 on optimal surveillance. Consider, for example, the case when we measure disease incidence in several different areas. If we restrict our attention to a global increase in the incidence in all areas occurring at the same time, the multivariate situation is easily reduced to a univariate one. Then we can proceed as in the univariate case and derive optimal methods. However, for many applications the specification of a global change is too restrictive. The problem is how to determine which type of change to focus on and which not to. The method derived according to the specification of a global change will be excellent in detecting a global change but will not be capable of detecting, for example, small clusters of increased incidence. On the other hand, if focusing on detecting all kinds of clusters which can occur, the detection ability of the surveillance method for a specific type of cluster will be small. A fundamental question in multivariate surveillance is what type of changes to focus on. In a public health setting this issue is crucial. Syndromic surveillance is one example of almost infinitely many scenarios for a changed pattern which would be of interest to detect. One way to focus the attention is to consider some type of dimension reduction transformation (Hawkins, 1991; Runger, 1996).

Timeliness in detection is of extreme interest in public health surveillance, and in Wagner *et al.* (2001), Sonesson and Bock (2003), Mostashari and Hartman (2003), and Aylin *et al.* (2003) it was pointed out that measures other than those traditionally used such as the sensitivity, specificity and predictive value positive are important. Guidelines for the evaluation of surveillance systems are being developed by the CDC and its collaborators (see Sosin, 2003). In the draft guidelines it is suggested that surveillance systems are to be evaluated using naturally occurring outbreaks as well as simulations. This approach to different means of evaluation can also be found regarding the evaluation of the ESSENCE system (see Burkom, 2003; Lombardo *et al.*, 2003).

To evaluate the timeliness, different measures such as the average run length, the conditional expected delay, the expected delay and the probability of successful detection (see Chapter 3) can also be used in a multivariate setting. However, the measures of evaluation have to be more precisely defined as the change in the multivariate case can be in several different directions at different times.

ARL^1 is the most commonly used measure of the detection ability also in the multivariate case. In some cases the evaluation of ARL^1 is performed for the case where all variables, or a known subset, change. Then the results of Wessman (1998) can be used to reduce the problem to a univariate one. This fact seems to have been overlooked in literature. If the comparisons between methods are made with ARL for the case with the same change point (the first) for all components, it implies that a better result should be obtained for the reduction to a univariate surveillance than for vector accumulation. There are several claims to the contrary in the literature. However, these claims are based

on different principles for the choice for parameters in the methods compared, which makes the numerical comparisons difficult to interpret.

For use in a public health setting it is not entirely satisfactory to restrict the evaluation of the methods to changes occurring at the same time as the surveillance starts, since the detection ability depends on when the change occurs. Methods with short delay times for detection of early changes are often slow in detecting changes occurring later. There are several examples of this in the univariate case (Frisén and Wessman, 1999; Sonesson, 2003) as well as in the multivariate case. Qiu and Hawkins (2001) take this into account and evaluate their nonparametric CUSUM by the conditional expected delay $\mathrm{CED}(t) = \mathrm{E}[t_A - \tau | t_A \geq \tau = t]$ for the case when the change occurs immediately, CED(1), but also for the case when the (common) change occurs after 500 time points, CED(500). Large differences in detection ability, with respect to the time point of the change, were found. Other examples of evaluations of later changes can be found in Ngai and Zhang (2001) and Sullivan and Jones (2002). It is important to note that the relative performances between methods are strongly affected by the measurement of evaluation. The ranking of the various types of moving average methods studied might be upset, even reversed, if comparing early and late changes. In the multivariate setting different change points for different variables are an additional complication. Thus it is always recommended that a thorough evaluation, involving changes of different types occurring at different time points, is performed using several types of evaluations.

Wessman (1999) suggested a generalization of the ARL measure to allow for the possibility of different change points for different variables. He analyzes the effect of using the marginal distribution, as by the parallel methods, or the joint distribution of the variables, as by methods which reduce to one statistic for each time. The latter approach is very sensitive to the correlations between the variables and correlations between the change points for the different variables.

9.5 CONCLUDING DISCUSSION

Multivariate surveillance is a diverse and difficult area which deals with the detection of changed patterns in multidimensional data. In public health, the nature of the data is often high-dimensional and collected into huge databases. The construction of surveillance methods is challenging from many points of view, including practical, computational as well as purely statistical. It is encouraging to see the increasing interest in this area of public health. This includes practical issues involving the collection of new types of data, computational ones such as the implementation of automated methods in large-scale surveillance databases, and the statistical theory on which to base the surveillance methods. In this chapter we have focused on the statistical aspects of the multivariate surveillance problem.

We have given a description of methods, characterizing them as scalar accumulating, parallel, vector accumulating or simultaneous. Many methods first reduce

the dimension by using transformed variables, such as principal components, and then use one of the approaches for multivariate surveillance. However, it should be understood that there is no sharp limit between several of these categories. For example, what is regarded as a dimension reduction transformation and what is thought to be a scalar accumulation, sometimes overlap. Fuchs and Benjamini (1994) suggest multivariate profile charts which in the same diagram demonstrate both the overall multivariate surveillance and individual ones and thus combine two of the approaches. Several of the public health systems suggested have flavors of several of the approaches referred to.

It is important that the type of change we aim to detect is well specified. The more specifically the type of change is stated the better the ability of the surveillance to detect this change. Hauck *et al.* (1999) describes different scenarios for how the change might influence variables and the relation between these. One way to focus the detection ability is by the specification of a loss function depending on how important changes in different directions are. Mohebbi and Havre (1989) use weights from a linear loss function instead of the covariance for the reduction to a univariate statistic which is monitored by a univariate CUSUM. Tsui and Woodall (1993) use a nonlinear loss function and a vector accumulation method. This method is named MLEWMA. For some methods the detection ability depends only on a noncentrality parameter which measures the magnitude of the multidimensional change. These methods are known as directionally invariant. However, this is not necessarily a good property, since in many cases one is more interested in detecting a certain type of change. Preferably, this specification should be motivated by the application. However, 'automatic ways' have also been suggested.

One might raise the question: 'What multivariate surveillance method is the best one to use in applications?' A concise answer to such a general question is, however, not possible to give. Different methods are suitable for different applications. Consider, for example, the syndromic surveillance system established by the New York City Department of Health and Mental Hygiene described in Heffernan *et al.* (2004). In this system emergency department visits are categorized into any of the syndromes: common cold, sepsis, respiratory, diarrhea, fever, rash, asthma or vomiting. Respiratory and fever syndromes are identified as being of particular interest for bioterrorism surveillance. Since an anthrax attack will lead to a simultaneous increase in both respiratory and fever syndromes one should use a reduction to a univariate surveillance method. For a general surveillance of any bioterrorist attack leading to either respiratory or gastrointestinal syndromes, which are not supposed to occur simultaneously, one might instead prefer to use parallel methods of the two types. One advantage with parallel methods is that the medical interpretation of alarms will be clearer.

One important problem is the identification of why an alarm was raised. This is especially so when no narrow specification of the change to be detected is made. A simple example is the inability of the Hotelling T^2 control chart to distinguish between a change in the mean vector from a change in the covariance structure. When using a Hotelling T^2 control chart, Mason *et al.* (1995) provided a

general approach involving a decomposition of the T^2 statistic into independent components. Among other suggestions is principal component analysis; see Pignatiello and Runger (1990), Kourti and MacGregor (1996) and Maravelakis *et al.* (2002). The steps of epidemiological investigation following an alarm triggered by a surveillance method are described by Pavlin (2003). The investigation consists of confirmation of the existence of an outbreak, verifications of diagnosis, estimation of the number of cases, development and evaluation of a hypothesis about the outbreak, and finally implementation of control measures and communication of the findings to other public health practitioners. Several of these steps can be assisted by the alarm system itself and the data collected. An example is a spatially restricted outbreak. Information about where the increased number of cases is located will assist the epidemiological investigation. A multivariate surveillance method does not provide immediate answers to such a question if the information is pooled across different areas. The importance of knowledge about where to concentrate the effort after an alarm indicating a bioterrorist attack is discussed by Mostashari and Hartman (2003).

PART III

Database Mining and Bayesian Methods

10

Bayesian Network Approaches to Detection

Weng-Keen Wong and Andrew W. Moore

10.1 INTRODUCTION

Multidimensional data with a temporal component is available from numerous biosurveillance sources. We would like to tackle the problem of early disease outbreak detection from such sources. Consider a situation in which we have a database of emergency department (ED) cases from several hospitals in a city. Each record in this database contains information about the individual who was admitted to the ED. This information includes fields such as age, gender, symptoms exhibited, home location, work location, and time admitted. (To maintain patient confidentiality, personal identifying information, such as patient names, addresses, and identification numbers, was not in the data set used in this research.) Clearly, when an epidemic sweeps through a region, there will be extreme perturbations in the number of ED visits. While these dramatic upswings are easily noticed during the late stages of an epidemic, the challenge is to detect the outbreak during its early stages and mitigate its effects.

Although we have posed our problem in an anomaly detection framework, the majority of anomaly detection algorithms are inappropriate for this domain. Typical anomaly detection identifies individual records that have a rare attribute or rare combination of attributes. As an example, suppose we apply a traditional anomaly detection technique to ED records. We might then find an unusual record such as a patient who was over 100 years old living in a sparsely populated region. This outlier is not at all indicative of a disease outbreak. Outlier detection succeeds in finding data points that are rare based on the underlying density, but these data points are treated in isolation from each other. Early

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

epidemic detection, on the other hand, hinges on identifying anomalous groups, which we will refer to as *anomalous patterns*. Specifically, we want to know if the recent proportion of a group with specific characteristics is anomalous based on what the proportion is normally. Traditional outlier detection will likely return isolated irregularities that are insignificant to the early detection system.

An alternate approach would be to monitor aggregate daily counts of either a single attribute or a combination of attributes. For instance, we could monitor the daily number of people appearing in the ED with respiratory problems. This approach converts multivariate surveillance data into a univariate time series. Many different algorithms can then be used to monitor this univariate surveillance data, including methods from statistical quality control such as CUSUM and EWMA (which are discussed in Chapter 2 of this book), time series models (Box and Jenkins, 1976), and regression techniques (Serfling, 1963; Farrington *et al.*, 1996; Lazarus *et al.*, 2002). This technique works well if we know a priori which disease to monitor since we can improve the timeliness of detection by monitoring specific features of the disease. For example, if we were vigilant against an anthrax attack, we can concentrate our efforts on ED cases involving respiratory problems. In our situation, we need to perform nonspecific disease monitoring because we do not know what disease to expect, particularly in the case of a bioterrorist attack. As a result, instead of monitoring health care data for predefined patterns, we resort to detecting any significant anomalous patterns in the multivariate ED data. Furthermore, by taking a multivariate approach that inspects all available attributes in the data, particularly the temporal, spatial, demographic, and symptom-related features, we hope to improve the timeliness of detection.

Contrast set mining (Bay and Pazzani, 1999) has the same flavor as the approach we take except it finds rules with more than two components using a pruning algorithm to reduce the exponential search space. This optimization prunes away rules whose counts are too small to yield a valid chi-square test. In addition, multiple hypothesis testing problems are addressed using a Bonferroni correction in contrast set mining, while we use a randomisation test.

10.2 ASSOCIATION RULES

Our approach to early disease outbreak detection uses a rule-based anomaly pattern detector called What's Strange About Recent Events (WSARE: Wong *et al.*, 2002, 2003). WSARE operates on discrete, multidimensional data sets with a temporal component. This algorithm compares recent data against a baseline distribution with the aim of finding rules that summarize significant patterns of anomalies. Each rule takes the form $X_i = V_i^j$, where X_i is the ith feature and V_i^j is the jth value of that feature. Multiple components are joined together by a logical AND. For example, a two-component rule would be Gender = Male AND Home Region = NW. It is helpful to think of the rules as

SQL SELECT queries. They characterize a subset of the data having records with attributes matching the components of the rule. WSARE finds those subsets whose proportions have changed the most between recent data and the baseline.

Another problem facing detection systems is determining the baseline distribution. This distribution is usually obtained from a period of time in the past when no epidemics are known to have happened. However, determining this distribution for early disease outbreak detection is extremely difficult due to the different trends present in health care data. Seasonal variations in weather and temperature can dramatically alter the distribution of such data. For example, the flu season typically occurs during mid-winter, resulting in an increase in ED cases involving cough and fever symptoms. Disease outbreak detectors intended to detect epidemics such as SARS, West Nile virus and anthrax are not interested in detecting the onset of the flu season and would be thrown off by it. Day-of-week variations make up another periodic trend. Figure 10.1, which is taken from Goldenberg *et al.* (2002), clearly shows the periodic elements in cough syrup and liquid decongestant sales.

Choosing the wrong baseline distribution can have dire consequences for an early detection system. Consider once again a database of ED records. Suppose we are presently in the middle of the flu season and our goal is to detect anthrax, not an influenza outbreak. Anthrax initially causes symptoms similar to those of influenza. If we choose the baseline distribution to be outside of the current flu season, then a comparison with recent data will trigger many false anthrax

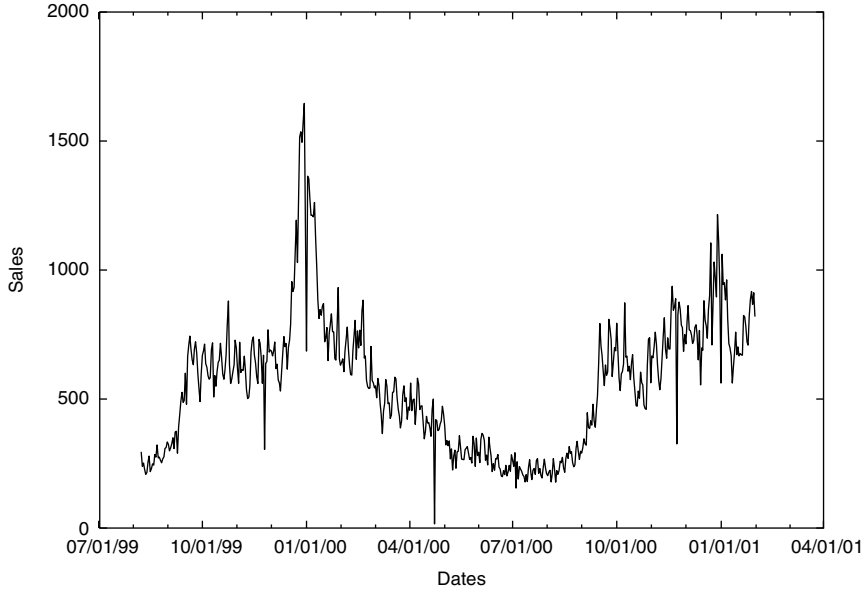


Figure 10.1 Cough syrup and liquid decongestant sales (from Goldenberg *et al.*, 2002).

alerts due to the flu cases. Conversely, suppose we are not in the middle of the flu season and that we obtain the baseline distribution from the previous year's influenza outbreak. The system would now consider high counts of flu-like symptoms to be normal. If an anthrax attack occurs, it would be detected at a very late stage.

There are clearly tradeoffs when defining this baseline distribution. At one extreme, we would like to capture any current trends in the data. One solution would be to use only the most recent data, such as data from the previous day. This approach, however, makes the algorithm susceptible to outliers that may only occur in a short but recent time period. On the other hand, we would like the baseline to be accurate and robust against outliers. We could use data from all previous years to establish the baseline. This choice would smooth out trends in the data and likely raise alarms for events that are due to periodic trends.

We propose building a Bayesian network to represent the joint distribution of the baseline. From this joint distribution, we represent the baseline distributions from the conditional distributions formed by conditioning on what we term *environmental attributes*. These features are precisely those attributes that account for trends in the data, such as the season, the current flu level, and the day of week.

10.3 WSARE

At this point we will provide an overview of the WSARE algorithm. WSARE operates on a daily basis, in which, for each day, the algorithm treats records from the past 24 hours as recent events. Using historical data beyond the past 24 hours, WSARE then creates a baseline distribution which is assumed to capture the usual behavior of the system being monitored under the environmental conditions of the current day. Once the baseline distribution has been created, the algorithm considers all possible one- and two-component rules over events occurring on the current day. The rules are scored with a scoring function that assigns high scores to rules corresponding to subsets of data that have unusual proportions when compared against the baseline distribution. The rule with the highest score for the day has its p-value calculated using a randomization test. If this p-value is lower than a specified threshold, an alert is raised.

10.3.1 Creating the Baseline Distribution

Learning the baseline distribution involves taking all records prior to the past 24 hours and building a Bayesian network from this subset. During the structure learning, we differentiate between environmental attributes, which are features that cause trends in the data, and *response attributes*, which are the remaining

features. The environmental attributes are specified by the user based on the user's knowledge of the problem domain. If there are any latent environmental attributes that are not accounted for in this model, the detection algorithm may have some difficulties. However, as will be described later, WSARE was able to overcome some hidden environmental attributes in our simulator.

The network structure is learned from categorical data using an efficient structure search algorithm called optimal reinsertion (Moore and Wong, 2003) based on ADTrees (Moore and Lee, 1998). Environmental attributes in the structure are prevented from having parents because we are not interested in predicting their distributions, but rather, we want to use them to predict the distributions of the response attributes. The structure search also exploits this constraint by avoiding search paths that assign parents to the environmental attributes.

We have often referred to environmental attributes as attributes that cause periodic trends. Environmental attributes, however, can also include any source of information that accounts for recent changes in the data. Incorporating such knowledge into the Bayesian network can aid in detecting anomalies other than the ones we already know about. For example, suppose we detect that a botulism outbreak has occurred and we would still like to be on alert for any anthrax releases. We can add 'Botulism Outbreak' as an environmental attribute to the network and supplement the current data with information about the botulism outbreak.

Once the Bayesian network is learned, we have a joint probability distribution for the data. We would like to produce a conditional probability distribution, which is formed by conditioning on the values of the environmental attributes. Suppose that today is February 21, 2003. If the environmental attributes were Season and Day of Week, then we would set Season = Winter and Day of Week = Monday. Let the response attributes in this example be $X_1, \ldots, X_n$. We can then obtain the probability distribution $P(X_1, \ldots, X_n | \text{Season} = \text{Winter}, \text{DayofWeek} = \text{Monday})$ from the Bayesian network. For simplicity, we represent the conditional distribution as a data set formed by sampling 10 000 records from the Bayesian network conditioned on the environmental attributes. The size of this sampled data set has to be large enough to ensure that samples with rare combinations of attributes will be present, hence the choice of 10 000 records. We will refer to this sampled data set as DB_{baseline}. The data set corresponding to the records from the past 24 hours of the current day will be named DB_{recent}.

One of the benefits of using a Bayesian network to represent the baseline distribution is the Bayesian network's generalization capability, which can be used to predict the probability of a situation that may not have been encountered in the past. As an example, suppose we would like to estimate the joint probability $P(X_1, \ldots, X_n |$ Season = Spring, Weather = Snow, Day of Week = Monday). Furthermore, suppose that we have no ED records from snowy spring Mondays in our historical data. However, we do have ED records obtained during snow days,

during Mondays, and during spring. The Bayesian network is able to generalize from such data to produce an estimate of what the joint probability would be even though no records from snowy spring Mondays existed in the training data.

10.3.2 Finding the Best One-Component Rule

After the baseline distribution is generated, WSARE proceeds by finding the best one-component rule. This step requires exhaustively searching over all feature–value combinations and scoring them. The scoring mechanism establishes a 2×2 contingency table for each rule. Suppose the rule is Respiratory Syndrome = True. We set up a contingency table as shown in Table 10.1 with the cells containing counts for records matching and not matching the rule for both data sets DB_{recent} and DB_{baseline}. Let C_{recent} be the count for DB_{recent} and C_{baseline} be that for DB_{baseline}.

The score of a rule is determined through a hypothesis test in which the null hypothesis is the independence of the row and column attributes of the 2×2 contingency table. In effect, the hypothesis test measures the difference between the counts for the recent period and those for the baseline. This test produces a p-value that determines the significance of the anomalies found by the rule. This p-value will be referred to as the *score* in order to distinguish it from the corrected p-values used below. We use the chi-square test whenever its assumptions are not violated. Since we are searching for anomalies, the counts in the contingency table are frequently small numbers and we resort to using Fisher's exact test (Good, 2000). Running Fisher's exact test on Table 10.1 yields a score of 0.025 939, which indicates that C_{recent} for the rule Home Location = NW is anomalous when compared to that of C_{baseline}.

10.3.3 Two-Component Rules

At this point, the best one-component rule for a particular day has been found. We will refer to the best one-component rule for day i as BR_i^1. The algorithm then attempts to find the best two-component rule for the day by adding on one extra component to BR_i^1. This extra component is determined by supplementing BR_i^1 with all possible feature–value pairs, except for the one already present in BR_i^1, and selecting the resulting two-component rule with the best score. We resort to building rules in this greedy manner in order to reduce the computational

Table 10.1 A sample 2×2 contingency table.

	C_{recent}	C_{baseline}
Home Location = NW	6	496
Home Location $\neq$ NW	40	9504

cost of an exhaustive search. Scoring is performed in exactly the same manner as before, except that the counts C_{recent} and C_{baseline} are calculated by counting the records that match the two-component rule. The best two-component rule for day i is subsequently found and we will refer to it as BR_i^2

BR_i^2, however, may not be an improvement over BR_i^1. We need to perform further hypothesis tests to determine if the presence of either component has a significant effect. This can be accomplished by determining the scores of having each component through Fisher's exact test. If we label BR_i^2's components as C_0 and C_1, then the two 2×2 contingency tables for Fisher's exact tests are as shown in Tables 10.2 and 10.3.

Once we have the scores for both tables, we need to determine if they are significant or not. We used the standard α value of 0.05 and considered a score to be significant if it was less than or equal to α. If the scores for the two tables were both significant, then the presence of both components had an effect. As a result, the best rule overall for day i is BR_i^2. On the other hand, if any one of the scores was not significant, then the best rule overall for day i is BR_i^1.

This greedy approach is an approximation that provides a compromise between accuracy and computational complexity. Without this greedy approach, searching for the best n-component rule through an exhaustive search will require searching over a number of rules that grows exponentially as n increases. With the greedy approximation, the rule search space only grows linearly in terms of n. Even with only a maximum of two components in a rule, we found empirically that the greedy search was 30 times faster than the exhaustive search. On the other hand, certain rules will not be found by this greedy approach, such as a two-component rule in which

Table 10.2 First 2×2 contingency table for a two-component rule.

Records from recent matching C_0 and C_1	Records from baseline matching C_0 and C_1
Records from recent matching C_1 and differing on C_0	Records from baseline matching C_1 and differing on C_0

Table 10.3 Second 2×2 contingency table for a two-component rule.

Records from recent matching C_0 and C_1	Records from baseline matching C_0 and C_1
Records from recent matching C_0 and differing on C_1	Records from baseline matching C_0 and differing on C_1

neither component is found during the single best scoring component stage. Wong (2004) compared the greedy and exhaustive searches on data from our simulator described in Section 10.4.1, and the results of WSARE on an activity monitor operating characteristic (AMOC) curve show no significant differences.

10.3.4 Obtaining the *p*-Value for Each Rule

The score produced by the previous step cannot be accepted at face value as a *p*-value because of a multiple hypothesis testing problem. Suppose we follow the standard practice of rejecting the null hypothesis when the *p*-value is less than α, where $\alpha = 0.05$. When only one hypothesis test is performed, the probability of making a false discovery under the null hypothesis would be $\alpha = 0.05$. On the other hand, if we perform 1000 hypothesis tests, one for each possible rule under consideration, then the probability of making a false discovery could be as bad as $1-(1-0.05)^{1000} \approx 1$, which is much greater than 0.05 (Miller *et al.*, 2001).

We need to add an adjustment for the multiple hypothesis tests. This problem can be addressed using a Bonferroni correction (Bonferroni, 1936), but this approach can be unnecessarily conservative. Instead, we use a randomization test (Good, 2000) in which the date is assumed to be independent of the other features. In this test, the non-date features of both DB_{recent} and DB_{baseline} remain the same but the dates are shuffled between the two data sets, resulting in two newly randomized data sets RDB_{recent} and RDB_{baseline}, respectively. RDB_{recent} and RDB_{baseline} will now both contain records with dates from the original recent period and from the baseline period. The procedure is described below.

Let $UCP =$ uncompensated $p-$value (i.e., the score as defined above)
For $j = 1$ to 1000
  Let $DB = DB_{\text{recent}} \cup DB_{\text{baseline}}$
  Produce RDB^j_{recent} and RDB^j_{baseline} from DB
  Let $RDB^j = RDB^j_{\text{recent}} \cup RDB^j_{\text{baseline}}$
  Let $BR^j =$ Best rule on RDB^j
  Let $UCP^j =$ uncompensated $p-$value of BR^j on RDB^j

Let the compensated *p*-value of *BR* be *CPV*, that is,

$$CPV = \frac{\text{No. of Rand Tests in which } UCP^j < UCP}{\text{No. of Rand Tests}}.$$

CPV estimates the chance of seeing an uncompensated *p*-value as good as *UCP* if in fact there was no relationship between the date and the other features.

10.4 EVALUATION

10.4.1 The Simulator

Validation of early outbreak detection algorithms is generally a difficult task for two main reasons. First of all, ground truth must be established by marking the start and end of the outbreak periods in the data. Determining the ground truth is a time-consuming process since a group of epidemiologists must manually inspect the data and come to a consensus on the outbreak periods. As a result, labelled outbreak data is scarce. Secondly, in order to make statistically significant conclusions about a detection algorithm, a large number of outbreak data sets are needed, each of which must contain a real outbreak. Due to these limitations, the best available option is to create a simulator which approximates the effects of an epidemic on a population. We evaluated WSARE on a small-scale city simulator. Although this simulator is not entirely realistic, the extremely noisy data sets produced by the simulator are still a challenge for any detection algorithms. Further experimental results, including examples of output from applying WSARE to Pittsburgh-area ED data, can be found in Wong *et al.* (2003).

Our city consists of nine regions, each of which contains a different sized population, ranging from 100 people in the smallest area to 600 people in the largest section. We ran the simulation for a two-year period from January 1, 2002 to December 31, 2003. The environment of the city is not static, with weather, flu levels, and food conditions in the city changing from day to day. Flu levels are typically low in the spring and summer but start to climb during the fall. We made the flu season strike in winter, resulting in the highest flu levels during the year. Weather, which only takes on the values hot or cold, is as expected for the four seasons, with the additional feature that it has a good chance of remaining the same as it was yesterday. Each region has a food condition of good or bad. A bad food condition facilitates the outbreak of food poisoning in the area.

We implemented this city simulation using a Bayesian network, as shown in Figure 10.2. We will use the convention that any nodes shaded black in the Bayes network are set by the system and do not have their values generated probabilistically. Due to space limitations, instead of showing 18 separate nodes for the current and previous food conditions of each region, we summarize

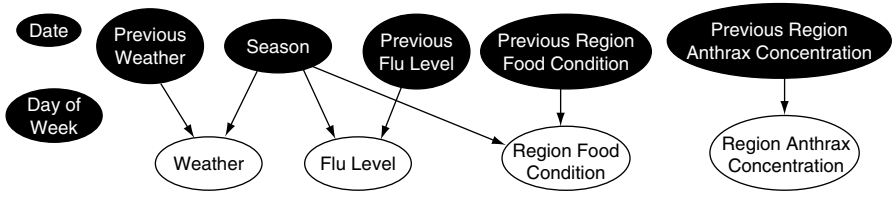


Figure 10.2 City Status Bayesian network.

them using the generic nodes Region Food Condition and Previous Region Food Condition, respectively. This same space-saving technique is used for the current and previous region anthrax concentrations. Most of the nodes in this Bayesian network have an arity of two to three values. For each day, after the black nodes have their values set, the values for the white nodes are sampled from the Bayesian network. These records are stored in the City Status (CS) data set. The simulated anthrax release was selected for a random date during a specified time period. One of the nine regions is chosen randomly for the location of the simulated release. On the date of the release, the Region Anthrax Concentration node is set to high. The anthrax concentration remains high for the affected region for a randomly chosen length of time.

The second Bayesian network used in our simulation produces individual health care cases. Figure 10.3 depicts the Patient Status (PS) network. On each day, for each person in each region, we sample the individual's values from this network. The black nodes first have their values assigned from either a pre-existing table of values or from the CS data set record for the current day. The white nodes are then sampled from the PS network. Each individual's health profile for the day is thus generated. Nodes such as Flu Level, Day of Week, Season, Weather, Region Grassiness, and Region Food Condition are intended to represent environmental variables that affect the upswings and

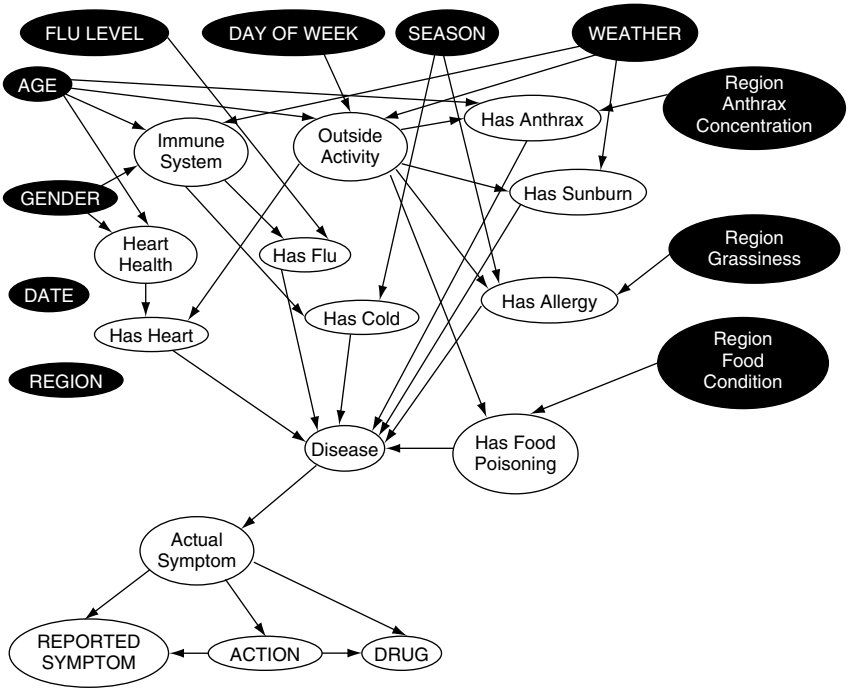


Figure 10.3 Patient Status Bayesian network.

downswings of a disease. Some of these environmental variables will be hidden from the detection algorithm. The Region Grassiness node indicates the amount of pollen in the air and thus affects the allergies of a patient. The Disease node indicates the status of each person in the simulation. A person is either healthy or they can have, in order of precedence, allergies, a cold, sunburn, the flu, food poisoning, heart problems, or anthrax. If an individual has more than one disease, the Disease node picks the disease with the highest precedence. A sick individual then exhibits one of the following symptoms: none, respiratory problems, nausea, or a rash. The actual symptom associated with a person may not necessarily be the same as the symptom that is reported to health officials. Actions available to a sick person include doing nothing, buying medication, going to the ED, or being absent from work or school. As with the CS network, the arities for each node in the PS network are small, ranging from two to four values. If the patient performs any action other than doing nothing, the patient's health care case is added to the PS data set. Only the attributes in Figure 10.3 labeled with upper-case letters are recorded, resulting in a great deal of information being hidden from the detection algorithm, including some latent environmental attributes. The number of cases generated daily by the PS network is typically in the range of 30 to 50 records. Table 10.4 contains two examples of records in the PS data set.

10.4.2 Algorithms

We ran four detection algorithms on 100 different PS data sets. Of the four detection algorithms, three were variations on WSARE which we will describe below. Each data set was generated for a two-year period, beginning on January 1, 2002 and ending December 31, 2003. The detection algorithms trained on data from the first year until the current day, while the second year was used for evaluation. The anthrax release was randomly chosen in the period between January 1, 2003 and December 31, 2003.

Table 10.4 Examples of two records in the Patient Status data set.

Region	NW	N
Age	Child	Senior
Gender	Female	Male
Flu Level	High	None
Day of Week	Weekday	Weekday
Weather	Cold	Hot
Season	Winter	Summer
Action	Absent	ED visit
Reported Symptom	Nausea	Rash
Drug	None	None
Date	Jan-01-2002	Jun-21-2002

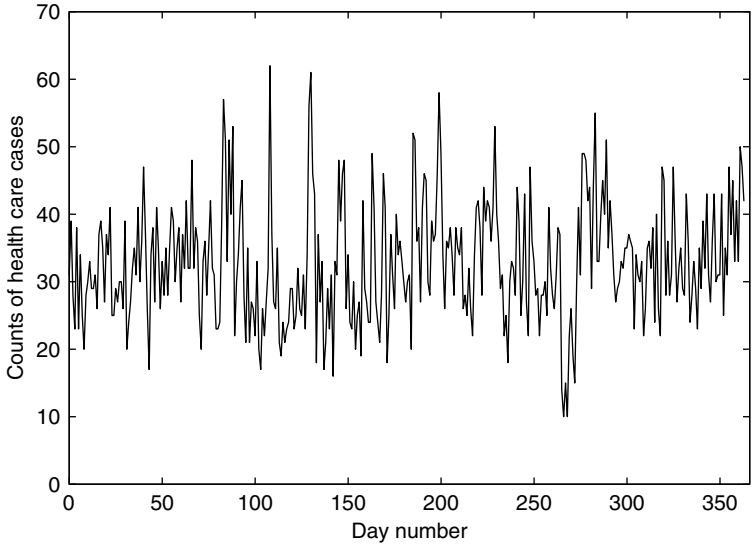


Figure 10.4 Daily counts of health care data.

We tried to simulate anthrax attacks that are not trivially detectable. Figure 10.4 plots the total count of health care cases on each day during the evaluation period. A naive detection algorithm would assume that the highest peak in this graph would be the date of the anthrax release. However, the anthrax release for Figure 10.4 occurred on day 276.

10.4.2.1 Moving average

The first algorithm that we used was a moving average algorithm that predicted the count for the current day as the average of counts from the previous 7 days. The window of 7 days was intended to capture any recent trends that might have appeared in the data. An alarm level was generated by fitting a Gaussian to data prior to the current day and obtaining a p-value for the current day's count. The mean for the Gaussian was calculated using data from 7 days before the current day, while the standard deviation was obtained using data from 28 days prior to the current day.

10.4.2.2 WSARE 2.0

WSARE 2.0 is a version of WSARE that uses raw historical data as the baseline distribution. The baseline distribution was obtained using records from 7, 14, 21, and 28 days before the current day. The attributes used by WSARE 2.5 and 3.0 (both of which are described below) as environmental attributes were

ignored by WSARE 2.0. If these attributes were not ignored, WSARE 2.0 would report many trivial anomalies. For instance, suppose the environmental attribute Day of Week = Sunday for the current day. If this attribute were not ignored, WSARE 2.0 would notice that 100 % of the records for the current day had Day of Week = Sunday while only 14.3 % of records in the baseline data set matched this rule.

10.4.2.3 WSARE 2.5

Instead of building a Bayesian network over the past data, WSARE 2.5 simply builds a baseline from all records prior to the current period with their environmental attributes equal to the current day's. In our simulator, we used the environmental attributes Flu Level, Season, Day of Week and Weather. To clarify this algorithm, suppose that for the current day we have the following values of these environmental attributes: Flu Level = high, Season = winter, Day of Week = weekday and Weather = cold. Then DB_{baseline} would contain only records before the current period with environmental attributes having exactly these values. It is possible that no such records exist in the past with exactly this combination of environmental attributes. If there are fewer than five records in the past that match, WSARE 2.5 cannot make an informed decision when comparing the current day to the baseline and simply reports nothing for the current day.

10.4.2.4 WSARE 3.0

WSARE 3.0 uses the same environmental attributes as WSARE 2.5 but builds a Bayesian network for all data from January 1, 2002 to the begining of the current day. We hypothesized that WSARE 3.0 will detect the simulated anthrax outbreak sooner than WSARE 2.5 because 3.0 can handle the cases where there are no records corresponding to the current day's combination of environmental attributes. The Bayesian network is able to generalize from days that do not match today precisely, producing an estimate of the desired conditional distribution. For efficiency reasons, we allowed WSARE 3.0 to learn a network structure from scratch once every 30 days. On intermediate days, WSARE 3.0 simply updates the parameters of the previously learned network without altering its structure.

10.5 RESULTS

Our evaluation criteria examine the algorithms' performance on an AMOC curve (Fawcett and Provost, 1999). The AMOC curve in Figure 10.5 plots detection time versus false positives over alarm thresholds ranging from 0 to 0.2

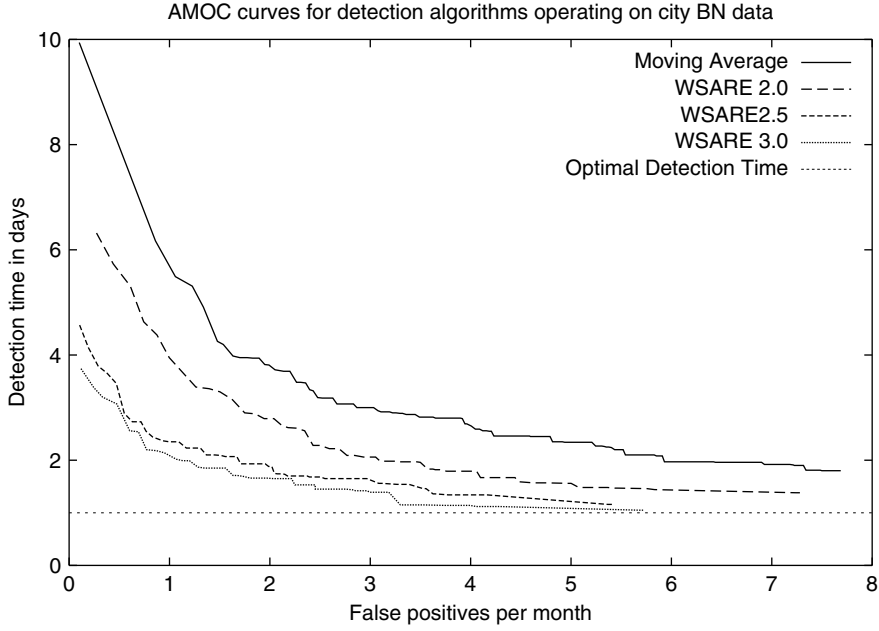


Figure 10.5 Asymptotic behavior of algorithms for simulated data.

in 0.001 increments. The lower alarm thresholds yield lower false positives and higher detection times, while the converse is true with higher alarm thresholds. Figure 10.5 fills in the lines to display the asymptotic behavior of the algorithms. The optimal detection time is 1 day, as shown by the dotted line at the bottom of the graph. We add a one-day delay to all detection times to simulate reality where current data is only available after a 24-hour delay. Any alert occurring before the start of the simulated anthrax attack is treated as a false positive. Detection time is calculated as the first alert raised after the release date. If no alerts are raised after the release, the detection time is set to 14 days.

Figure 10.6 plots the receiver operating (ROC) curve for the algorithms. Producing an ROC curve requires an explanation of the criteria for true positives, false positives, true negatives, and false negatives. For each of the 100 data sets, we have the date of the outbreak along with its duration, where the duration is defined as the number of consecutive days after the outbreak in which at least one reported event is due to anthrax. False positives are defined as any alerts raised before the outbreak date, where an alert is a p-value that is less than or equal to the alarm threshold. True negatives are the number of non-alerts before the outbreak date. Before considering false negatives and true positives, we need to define the outbreak period. If the outbreak begins on day i, then the end of the outbreak is on day i + duration. Let the outbreak period be $[i, i + \text{duration})$. The number of true positives is 1 if an alert is raised during the outbreak period and 0 if the outbreak is not detected. Finally, we

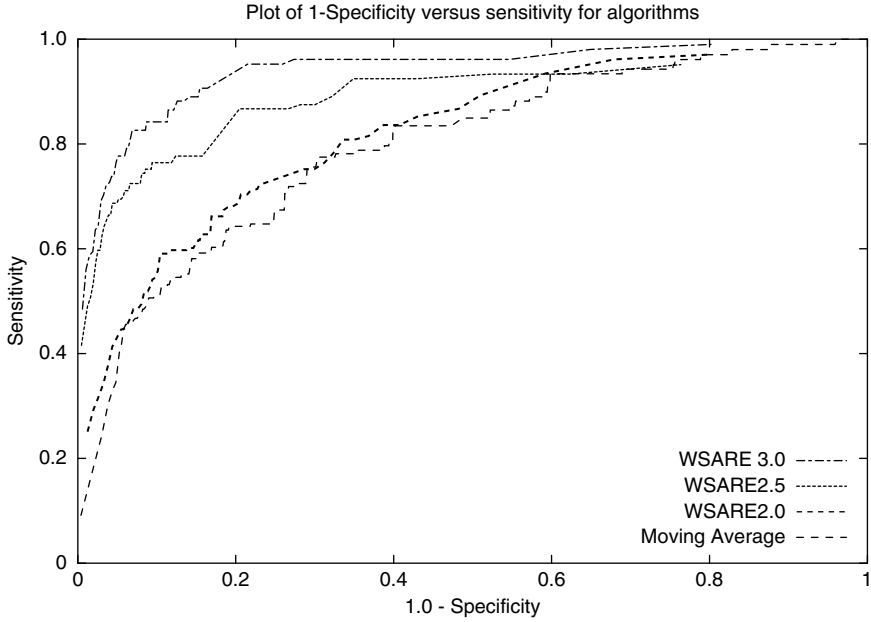


Figure 10.6 ROC curves for the algorithms.

count false negatives as the number of days beginning with the outbreak date and ending on the first alert during the outbreak period. If the outbreak is not detected, the number of false negatives is set to be the outbreak duration.

Figures 10.7–10.10 depict the histograms of detection times at one false positive per month for the four algorithms listed. On the x-axis the detection times range from 0 to 14 days, while the y-axis shows how many of the 100 simulation files fall into a histogram bin. The detection times at 14 days correspond to simulation files in which the detection algorithm did not detect the outbreak. Since we do not have an analytic form of the AMOC curve, we searched for the closest alarm threshold between 0 and 0.2 using 0.0001 increments that would yield one false positive per month for each algorithm. These alarm threholds were 0.0017, 0.0149, 0.0298, and 0.0224 for the moving average algorithm, WSARE 2.0, WSARE 2.5, and WSARE 3.0, respectively.

From the results of our simulation, WSARE 2.5 and WSARE 3.0 outperform the other algorithms in terms of the AMOC curve and the ROC curve. Both the moving average algorithm and WSARE 2.0 were thrown off by the periodic trends present in the PS data. We had previously proposed that WSARE 3.0 would have a better detection time than WSARE 2.5 due to the Bayesian network's ability to produce a conditional distribution for a combination of environmental attributes that may not exist in the past data. After checking the simulation results for which WSARE 3.0 outperforms WSARE 2.5, we conclude

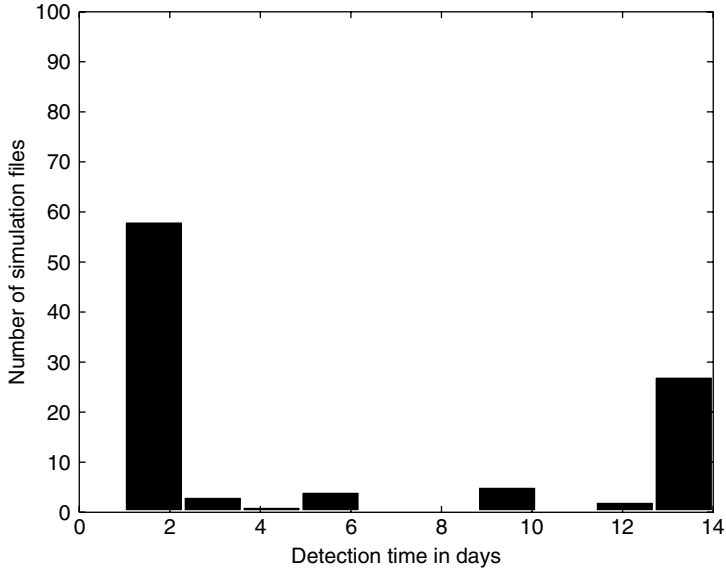


Figure 10.7 Moving average histogram of detection times at one false positive per month.

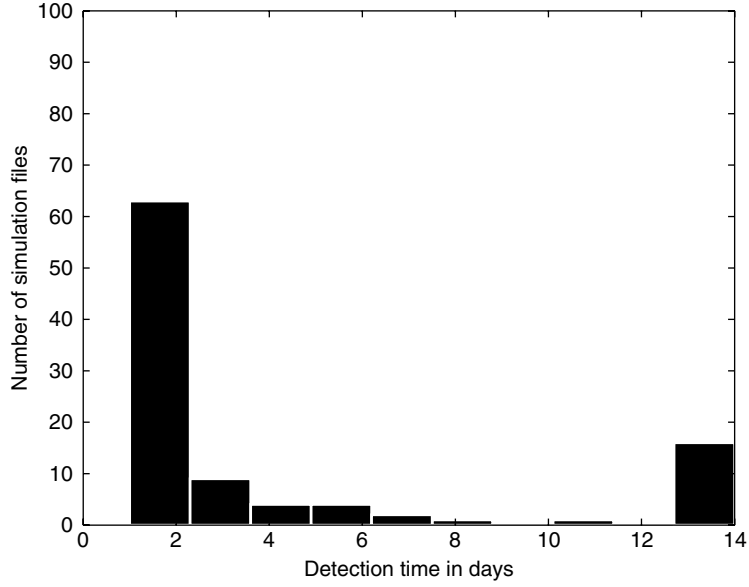


Figure 10.8 WSARE 2.0 histogram of detection times at one false positive per month.

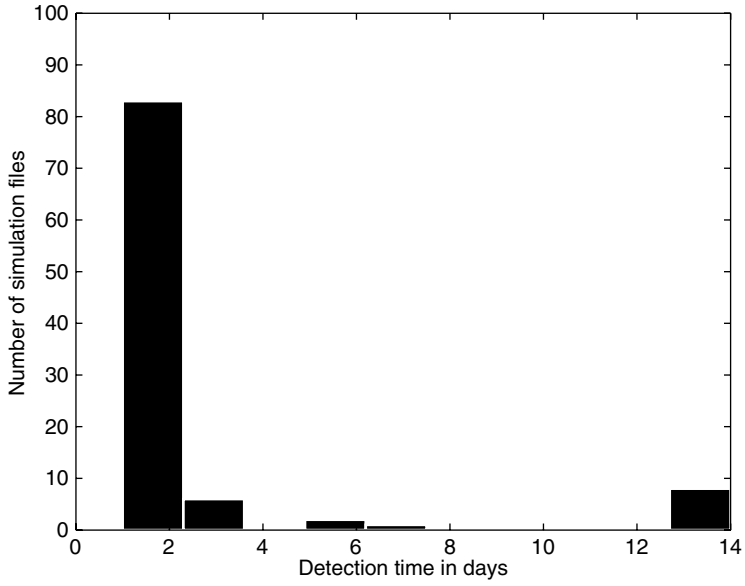


Figure 10.9 WSARE 2.5 histogram of detection times at one false positive per month.

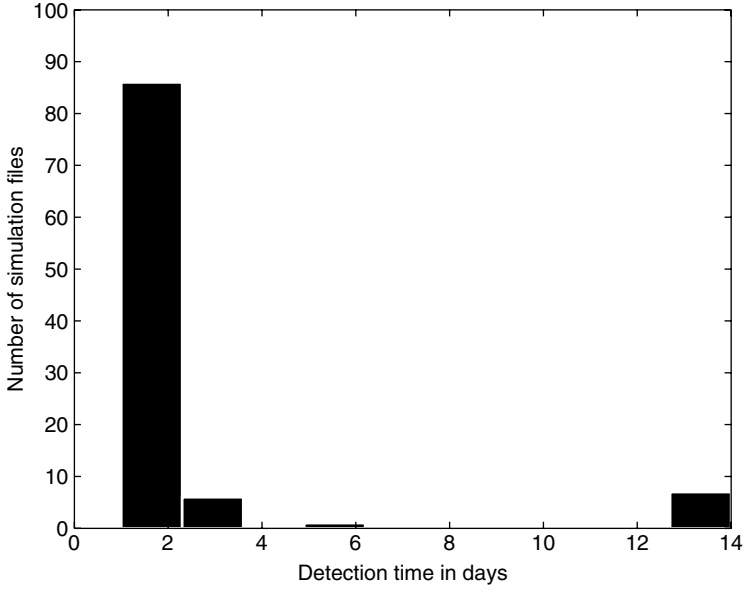


Figure 10.10 WSARE 3.0 histogram of detection times at one false positive per month.

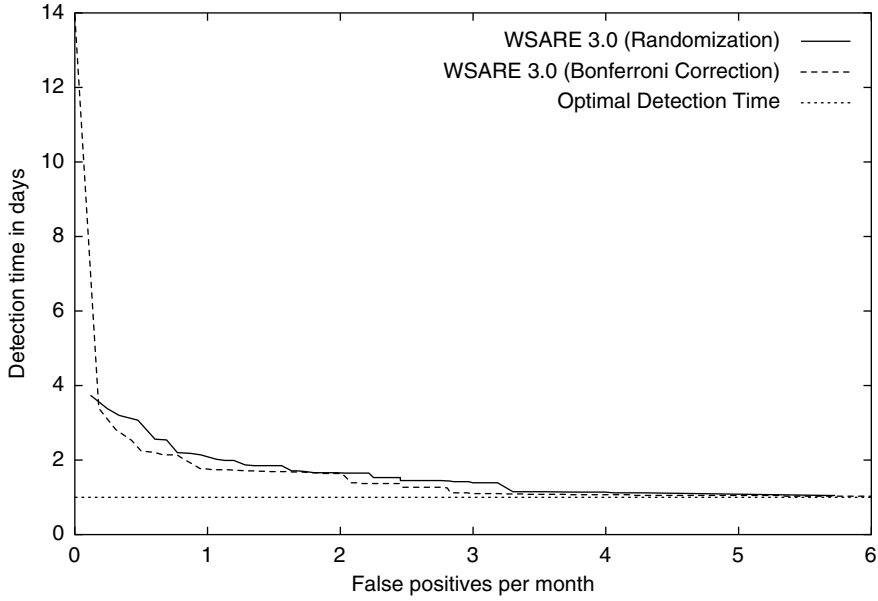


Figure 10.11 WSARE 3.0 with a randomization test versus WSARE 3.0 with a Bonferroni correction.

that in some cases our proposition was true. In others, the *p*-values estimated by WSARE 2.5 were not as low as those of version 3.0. The baseline distribution of WSARE 2.5 is likely not as accurate as the baseline of WSARE 3.0 due to smoothing performed by the Bayesian network. The false positives found by WSARE 2.5 and WSARE 3.0 are likely due to other nonanthrax illnesses that were not accounted for in the Bayesian network. Had we explicitly added a Region Food Condition environmental attribute to the Bayesian network, this additional information would likely have reduced the false positive count.

In Figure 10.11 we evaluated how WSARE 3.0 would be affected if a Bonferroni correction were used in the place of the randomization test. From the simulated data results, the Bonferroni correction appears to be a slight improvement over the randomization test. We suspect that as the number of attributes in the data increases, thereby increasing the number of hypothesis tests, the Bonferroni correction becomes more conservative and less effective than the randomization test.

10.6 CONCLUSION

Even with multiple periodic trends and other nonanthrax illnesses present in the simulated data, WSARE 3.0 has been shown to be successful at detecting anomalous patterns that are indicative of an anthrax release. WSARE 3.0

outperformed all the other algorithms evaluated. For a false positive rate of one per month, WSARE 3.0 detects the simulated anthrax release about 2 days earlier than WSARE 2.0 and about 6 hours earlier than WSARE 2.5. In addition, the false positive rate for WSARE 3.0 could have been reduced even further if more environmental attributes capturing the current state of the system had been added to the Bayesian network.

11

Efficient Scan Statistic Computations

Daniel B. Neill and Andrew W. Moore

11.1 INTRODUCTION

One of the core goals of data mining is to discover patterns and relationships in data. In many applications, however, it is important not only to discover patterns, but also to distinguish those patterns that are *significant* from those that are likely to have occurred by chance. This is particularly important in epidemiological applications, where a rise in the number of disease cases in a region may or may not be indicative of an emerging epidemic. In order to decide whether further investigation is necessary, epidemiologists must know not only the location of a possible outbreak, but also some measure of the likelihood that an outbreak is occurring in that region. More generally, we are interested in spatial data mining problems where the goal is detection of *overdensities*: spatial regions with high scores according to some density measure. The density measure can be as simple as the count (e.g. number of disease cases, or units of cough medication sold) in a given area, or can adjust for quantities such as the underlying population. In addition to discovering these high-density regions, we must perform statistical testing in order to determine whether the regions are significant. As discussed above, a major application is in detecting clusters of disease cases, for purposes ranging from detection of bioterrorism (e.g. anthrax attacks) to identifying environmental risk factors for diseases such as childhood leukemia (Openshaw *et al.*, 1988; Waller *et al.*, 1994; Kulldorff and Nagarwalla, 1995). Kulldorff (1999) discusses many other applications, including mining astronomical data (e.g. identifying star clusters), military reconnaissance, and medical imaging.

We consider the case in which data has been aggregated to a uniform, two-dimensional grid. Let G be an $N \times N$ grid of squares, where each square $s_{ij} \in G$

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

is associated with a *count* c_{ij} and an underlying *population* p_{ij}. For example, a square's count may be the number of disease cases in that geographical location in a given time period, while its population may be the total number of people 'at risk' for the disease. Our goal is to search over all rectangular regions $S \subseteq G$, and find the region S^* with the highest *density* according to a density measure D: $S^* = \arg\max_S D(S)$. We use the abbreviations *mdr* for the maximum density region S^*, and *mrd* for the maximum region density $D(S^*)$, throughout. We will also find the statistical significance (p-value) of this region by randomization testing, as described below.

The density $D(S)$ of a region S can be an arbitrary function of the total count of the region, $C(S) = \sum_S c_{ij}$, and the total population of the region, $P(S) = \sum_S p_{ij}$. Thus we will often write $D(C, P)$, where C and P are the count and population of the region under consideration. It is important to note that, while the term 'density' is typically understood to mean the ratio of count to population, we use the term in a much broader sense, to denote a class of *density functions* D which includes the 'standard' density function $D_1(C, P) = C/P$. For our purposes, we assume that the density function D satisfies the following three properties:

(1) For a fixed population, density increases monotonically with count

$$\frac{\partial D}{\partial C}(C, P) \geq 0, \qquad \text{for all } (C, P).$$

(2) For a fixed count, density decreases monotonically with population

$$\frac{\partial D}{\partial P}(C, P) \leq 0, \qquad \text{for all } (C, P).$$

(3) For a fixed ratio C/P, density increases monotonically with population

$$\frac{\partial D}{\partial P}(C, P) + \frac{C}{P}\frac{\partial D}{\partial C}(C, P) \geq 0, \qquad \text{for all } (C, P).$$

The first two properties state that an overdensity is present when a large count occurs in a small population. In the case of a uniform population distribution, the population of a region is proportional to its area, and thus an overdensity is present when a large count occurs in a small area. The third property states, in essence, that an overdensity is more significant when the underlying population is large. This is true because smaller populations P will typically have higher variance in densities C/P. For example, assuming that counts are Poisson distributed with means proportional to P, the variance of C/P is proportional to $P/P^2 = 1/P$. We also allow D to remain constant as population increases for a fixed ratio C/P, thus including the standard density function $D_1 = C/P$; we do not, however, allow functions where D decreases in this case. We will also make one more assumption involving the second partials of D; this fourth property is

not strictly necessary but makes our computation easier, eliminating the need to check for local maxima of the density function. A large class of functions satisfy all four properties, including Kulldorff's spatial scan statistic, discussed in detail below.

11.1.1 The Spatial Scan Statistic

A nonmonotonic density measure which is of great interest to epidemiologists is Kulldorff's *spatial scan statistic*, first presented in Kulldorff (1997), and also discussed in Chapter 7 of this collection. This statistic, which we denote by D_K, is in common use for finding significant spatial clusters of disease cases, which are often indicative of an emerging outbreak. Kulldorff's statistic assumes that counts c_{ij} are generated by an inhomogeneous Poisson process, that is, a Poisson process with spatially varying parameter. Thus each c_{ij} is assumed to be generated independently from a Poisson distribution with mean qp_{ij}, where q is the underlying 'disease rate' (or expected value of C/P) and p_{ij} is the population of that square. We then calculate the log of the likelihood ratio of two possibilities: that the disease rate q is higher in the region than outside the region, and that the disease rate is identical inside and outside the region. For a region with count C and population P, in a grid with total count C_{tot} and population P_{tot}, we can calculate

$$D_K = C \log \frac{C}{P} + (C_{\text{tot}} - C) \log \frac{C_{\text{tot}} - C}{P_{\text{tot}} - P} - C_{\text{tot}} \log \frac{C_{\text{tot}}}{P_{\text{tot}}}$$

if $C/P > C_{\text{tot}}/P_{\text{tot}}$, and $D_K = 0$ otherwise. Kulldorff (1997) proved that the spatial scan statistic is *individually most powerful* for finding a single significant region of elevated disease rate: for a fixed false positive rate, and for a given set of regions tested, it is more likely to detect the overdensity than any other test statistic.

11.1.2 Randomization Testing

Once we have found the mdr of grid G according to our density measure, we must still determine the significance of this region. Since the exact distribution of the test statistic is only known in special cases (such as density $= C/P$, with a uniform underlying population), in general we must find the region's p-value by randomization. To do so, we run a large number R of random replications, where a replica has the same underlying populations p_{ij} as G, but assumes a uniform disease rate $q_{\text{rep}} = C_{\text{tot}}(G)/P_{\text{tot}}(G)$ for all squares. For each replica G', we first generate all counts c_{ij} randomly from an inhomogeneous Poisson distribution with mean $q_{\text{rep}}p_{ij}$, then compute the mrd of G' and compare this to mrd(G). The p-value for the mdr is computed to be $(R_{\text{beat}} + 1)/(R + 1)$, where R_{beat} is the number of replicas G' with mrd(G') $\geq$ mrd(G), and R is the total number of

replications. If this p-value is less than 0.05, we can conclude that the discovered region is significant (unlikely to have occurred by chance) and is thus a 'spatial overdensity'. If the test fails, we have still discovered the maximum density region of G, but there is not sufficient evidence that this is an overdensity.

11.1.3 The Naive Approach

The simplest method of finding the mdr is to compute the density of all rectangular regions of sizes $k_1 \times k_2$, where $k_{\min} \leq k_1, k_2 \leq k_{\max}$. (We use $k_{\min} = 3$ and $k_{\max} = N$ throughout.) For an $N \times N$ grid, there are a total of $(N - k_1 + 1)$ $(N - k_2 + 1)$ regions of each size $k_1 \times k_2$, and thus a total of $O(N^4)$ regions to examine. We can compute the density of any rectangular region S in $O(1)$, by first finding the count $C(S)$ and population $P(S)$, then applying our density measure $D(C, P)$. (The count and population can be found in constant time by using a precomputed matrix of cumulative counts; then we can compute a region's count by adding/subtracting at most four cumulative counts, and similarly for populations.) This technique allows us to compute the mdr of an $N \times N$ grid G in $O(N^4)$ time. However, significance testing by randomization also requires us to find the mrd for each replica G', and compare this to mrd(G). Since calculation of the mrd takes $O(N^4)$ time for each replica, the total complexity is $O(RN^4)$, and R is typically large (we assume $R = 1000$). As discussed in Neill and Moore (2004b), several tricks may be used to speed up this procedure for cases where there is no significant spatial overdensity, but these do not help in cases when an overdensity is found. In general, the $O(N^4)$ complexity of the naive approach makes it infeasible for even moderately sized grids: we estimate a runtime of 45 days for a 256×256 grid on our test system, which is clearly far too slow for real-time detection of disease outbreaks.

While one alternative would be to search for an approximate solution using one of the variety of cluster detection algorithms in the literature, we present an algorithm which is exact (always finds the mdr) and yet is much faster than naive search. The key intuition is that, since we only care about finding the mdr, we do not need to search over every single rectangular region: in particular, we do not need to search a set of regions if we can prove (based on other regions we have searched) that none of them can be the mdr. As a simple example, if a given region has a very low count, we may be able to conclude that *no* subregion contained in that region can have a score higher than the mrd, and thus we do not need to actually compute the score of each subregion. These observations suggest a top-down, *branch-and-bound* approach: we maintain the current maximum score of the regions we have searched so far, calculate upper bounds on the scores of subregions contained in a given region, and *prune* regions which cannot contain the mdr. Similarly, when we are searching a replica grid, we only care about whether the mrd of the replica is higher than the mrd of the original grid. Thus we can use the mrd of the

original grid for pruning on the replicas, and can stop searching a replica if we find a region with score higher than this mrd.

11.2 OVERLAP-MULTIRESOLUTION PARTITIONING

Our top-down approach to cluster detection can be thought of as a multiresolution search of the space under consideration: we search first at coarse resolutions (large regions), then at successively finer resolutions (smaller regions) as necessary. This suggests that a hierarchical, space-partitioning data structure such as kd-trees (Preparata and Shamos, 1985), mrkd-trees (Deng and Moore, 1995), or quadtrees (Samet, 1990) may be useful in speeding up our search. However, our desire for an exact solution makes it difficult to apply these data structures to our problem. In a kd-tree, each spatial region is recursively partitioned into two disjoint 'child' regions, each of which can then be further subdivided. The difficulty, however, is that many subregions of the parent are not contained entirely in either child, but overlap partially with each. Thus, in addition to recursively searching each child for the mdr, we must also search over all of these 'shared' regions at each level of the tree. Since there are $O(N^4)$ shared regions even at the top level of the tree (i.e. regions partially overlapping both halves of grid G), an exhaustive search over all such regions is too computationally expensive, and thus a different partitioning approach is necessary.

An initial step toward our partitioning can be seen by considering two divisions of a rectangular spatial region S: first, into its left and right halves (which we denote by S_1 and S_2), and second, into its top and bottom halves (which we denote by S_3 and S_4). Assuming that S has size $k_1 \times k_2$, this means that S_1 and S_2 have size $\frac{1}{2}k_1 \times k_2$, and S_3 and S_4 have size $k_1 \times \frac{1}{2}k_2$. Considering these four (overlapping) halves, we can show that any subregion of S either is contained entirely in (at least) one of $S_1, \ldots, S_4$, or contains the centroid of S. Thus one possibility would be to search S by exhaustively searching all regions containing its centroid, then recursing the search on its four 'children' $S_1, \ldots, S_4$. Again, there are $O(N^4)$ 'shared' regions at the top level of the tree (i.e. regions containing the centroid of grid G), so an exhaustive search is infeasible.

Our solution, as in our previous work (Neill and Moore, 2004a, 2004b), is a partitioning approach in which adjacent regions partially overlap, a technique we call 'overlap-multiresolution partitioning', or 'overlap-multires' for short. Again we consider the division of S into its left, right, top, and bottom 'children'. However, while in the discussion above each child contained exactly half the area of S, now we let each child contain *more* than half the area. We again assume that region S has size $k_1 \times k_2$, and we choose fractions $f_1, f_2 > \frac{1}{2}$. Then S_1 and S_2 have size $f_1 k_1 \times k_2$, and S_3 and S_4 have size $k_1 \times f_2 k_2$. This partitioning (for $f_1 = f_2 = \frac{3}{4}$) is illustrated in Figure 11.1. Note that there is a region S_C common to all four children; we call this region the *center* of S. The size of S_C is $(2f_1 - 1)k_1 \times (2f_2 - 1)k_2$, and thus the center has non-zero area. When we

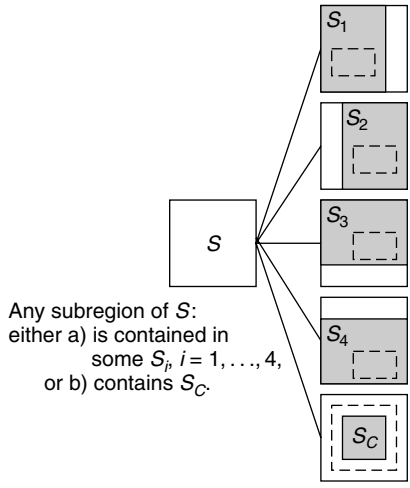


Figure 11.1 Overlap-multires partitioning of region S.

partition region S in this manner, it can be proved that any subregion of S either is contained entirely in (at least) one of $S_1, \ldots, S_4$, or contains the center region S_C. Figure 11.1 illustrates each of these possibilities.

Now we can search S by recursively searching $S_1, \ldots, S_4$, then searching all of the regions contained in S which contain the center S_C. Unfortunately, at the top level there are still $O(N^4)$ regions contained in grid G which contain its center G_C. However, since we know that each such region contains the large region G_C, we can place very tight bounds on the score of these regions, often allowing us to prune most or all of them. (We discuss how these bounds are calculated in the following subsection.) Thus the basic outline of our search procedure (ignoring pruning, for the moment) is:

```
overlap-search(S)
{
 call base-case-search(S)
 define child regions S_1...S_4, center S_C as above
 call overlap-search(S_i) for i=1...4
 for all S' such that S' is contained in S and contains S_C,
   call base-case-search(S')
}
```

Now we consider how to select the fractions f_1 and f_2 for each call of overlap-search, and characterize the resulting set Φ of regions S on which overlap-search(S) is called. Regions $S \in \Phi$ are called *gridded regions*, and regions $S \notin \Phi$ are called *outer regions*. For simplicity, we assume that the grid G is square, and that its size N is a power of 2. We begin the search by calling overlap-search(G). Then for each recursive call to overlap-search(S), where the size of

S is $k_1 \times k_2$, we set $f_1 = \frac{3}{4}$ if $k_1 = 2^r$ for some integer r, and $f_1 = \frac{2}{3}$ if $k_1 = 3 \times 2^r$ for some integer r. We define f_2 identically in terms of k_2, and then the child regions $S_1, \ldots, S_4$ and the center region S_C are defined in terms of f_1 and f_2 as above. This choice of f_1 and f_2 has the useful property that all gridded regions have sizes 2^r or 3×2^r for some integer r. For instance, if the original grid G has size 64×64, then the children of G will be of sizes 64×48 and 48×64, and the grandchildren of G will be of sizes 64×32, 48×48, and 32×64. This process can be repeated recursively down to regions of size $k_{\min} \times k_{\min}$, forming a structure that we call an *overlap-kd tree*. The first two levels of the overlap-kd tree are shown in Figure 11.2. Note that even though grid G has four child regions, and each of its child regions has four children, G has only 10 (not 16) distinct grandchildren, several of which are the child of multiple regions.

Our overlap-kd tree has several nice properties, which we present here without proof. First, for every rectangular region $S \subseteq G$, either S is a gridded region (contained in the overlap-kd tree), or there exists a unique gridded region S' such that S is an outer region of S' (i.e. S is contained in S', and contains the center region of S'). This means that, if overlap-search is called exactly once for each gridded region, and no pruning is done, then base-case-search will be called exactly once for every rectangular region $S \subseteq G$. In practice, we will prune many regions, so base-case-search will be called *at most once* for every rectangular region, and every region will be either searched or pruned. The second nice property of our overlap-kd tree is that the total number of gridded regions $|\Phi|$ is $O((N \log N)^2)$ rather than $O(N^4)$. This implies that, if we are able to prune (almost) all outer regions, we can find the mdr of an $N \times N$ grid in $O((N \log N)^2)$ time. In fact, we may not even need to search all gridded regions, so in many cases the search will be even faster.

Before we consider how to calculate score bounds and use them for pruning, we must first deal with an essential issue in searching overlap-kd trees. Since a child region may have multiple parents, how do we ensure that each gridded

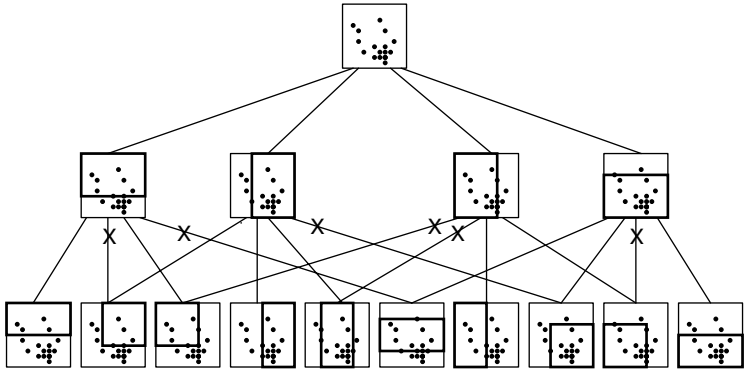


Figure 11.2 The first two levels of the overlap-kd tree. Each node represents a gridded region (denoted by a thick square) of the entire data set (thin square and dots).

region is examined only once, rather than being called recursively by each parent? One simple answer is to keep a hash table of the regions we have examined, and only call overlap-search(S) if region S has not already been examined. The disadvantage of this approach is that it requires space proportional to the number of gridded regions, $O((N \log N)^2)$, and spends a substantial amount of time doing hash queries and updates. A more elegant solution is what we call *lazy expansion*: rather than calling overlap-search(S_i) on all four children of a region S, we selectively expand only certain children at each stage, in such a way that there is exactly one path from the root of the overlap-kd tree to any node of the tree. One such scheme is shown in Figure 11.2: if the path between a parent and child is marked with an X, lazy expansion does not make that recursive call. No extra space is needed by this method; instead, a simple set of rules is used to decide which children of a node to expand. A child is expanded if it has no other parents, or if the parent node has the highest *priority* of all the child's parents. We give parents with lower aspect ratios priority over parents with higher aspect ratios: for example, a 48×48 parent would have priority over a 64×32 parent if the two share a 48×32 child. This rule allows us to perform variants of the search where regions with very high aspect ratios are not included; an extreme case would be to search only for squares, as in our previous work (Neill and Moore, 2004a). Within an aspect ratio, we fix an arbitrary priority ordering. Since we maintain the property that every node is accessible from the root, the correctness of our algorithm is maintained: every gridded region will be examined (if no pruning is done), and thus every region $S \subseteq G$ will be either searched or pruned.

11.2.1 Score Bounds

We now consider which regions can be *pruned* (discarded without searching) during our multiresolution search procedure. First, given some region S, we must calculate an upper bound on the scores $D(S')$ for regions $S' \subset S$. More precisely, we are interested in two upper bounds: a bound on the score of *all* subregions $S' \subset S$, and a bound on the score of the *outer* subregions of S (those regions contained in S and containing its center S_C). If the first bound is less than or equal to the mrd, we can prune region S completely; we do not need to search any (gridded or outer) subregion of S. If only the second bound is less than or equal to the mrd, we do not need to search the outer subregions of S, but we must recursively call overlap-search on the gridded children of S. If both bounds are greater than the mrd, we must both recursively call overlap-search and search the outer regions.

The calculation of these bounds involves a series of subcalculations and geometric proofs that are beyond the scope of this chapter. Full details are provided in (Neill and Moore, 2004b).

11.3 RESULTS

We first describe results with artificially generated grids and then real-world case data. An artificial grid is generated from a set of parameters $(N, k_1, k_2, \mu, \sigma, q', q'')$ as follows. The grid generator first creates an $N \times N$ grid, and randomly selects a $k_1 \times k_2$ 'test region'. Then the population of each square is chosen randomly from a normal distribution with mean μ and standard deviation σ (populations less than zero are set to zero). Finally, the count of each square is chosen randomly from a Poisson distribution with parameter qp_{ij}, where $q = q'$ inside the test region and $q = q''$ outside the test region.

For all our simulated tests, we used grid size $N = 256$, and a background disease rate of $q'' = 0.001$. We tested for three different combinations of test region parameters $(k_1 \times k_2, q')$: $(7 \times 9, 0.01)$, $(11 \times 5, 0.002)$, and $(4 \times 3, 0.002)$. These represent the cases of an extremely dense disease cluster, and large and small disease clusters which are significant but not extremely dense. We also ran a fourth test where no disease cluster was present, and thus $q = 0.001$ everywhere. We used three different population distributions for testing: the 'standard' distribution ($\mu = 10^4, \sigma = 10^3$), and two types of 'highly varying' populations. For the 'city' distribution, we randomly selected a 10×10 'city region': square populations were generated with $\mu = 5 \times 10^4$ and $\sigma = 5 \times 10^3$ inside the city, and $\mu = 10^4$ and $\sigma = 10^3$ outside the city. For the 'high-σ' distribution, we generated all square populations with $\mu = 10^4$ and $\sigma = 5 \times 10^3$. For each combination of test region parameters and population distribution, runtimes were averaged over 20 random trials. We also ran an additional 90 trials (for a total of 110) to test accuracy, confirming that the algorithm found the mdr in all cases. We also recorded the average number of regions examined; for our algorithm, this includes calculation of score bounds as well as scores of individual regions. Separate results are presented for the original grid and for each replica; for a large number of random replications ($R = 1000$) the results per replica dominate, since total runtime is $t_{\text{orig}} + R(t_{\text{rep}})$ to search the original grid and perform randomization testing. See Table 11.1 for results.

Table 11.1 Performance of algorithm, simulated data sets, $N = 256$. For each data set, we give the time in seconds to search the original grid and each replica grid, as well as the number of regions searched. The speedup is the ratio of runtimes of the naive and fast approaches.

test	method	sec/orig	speedup	sec/rep	speedup	regions (orig)	regions (rep)
all	naive	3864.00	×1	3864.00	×1	1.03B	1.03B
7×9, 0.01	fast	5.47	×706	1.68	×2300	100K	1.20K
11×5, 0.002	fast	21.72	×178	12.43	×311	1.03M	196K
4×3, 0.002	fast	42.96	×90	40.57	×95	2.59M	1.87M
no region	fast	189.68	×20	110.25	×35	27.4M	12.7M

Our first observation was that the runtime and number of regions searched were not significantly affected by the underlying population distribution; typically the three results differed by only 5–10 %, and in many cases test regions were found *faster* for the highly varying distributions than the standard distribution. Thus Table 11.1, rather than presenting separate results for each population distribution, presents the average performance over all three population distributions for each test. This result demonstrates the robustness of the algorithm to highly nonuniform populations; this is very different than our previous work (Neill and Moore, 2004a), where the algorithm was severely slowed by highly varying populations. The algorithm achieved average speedups ranging from $35\times$ (for no test region) to $2300\times$ (for an extremely dense test region) as compared to the naive approach. We note that, for the case of no test region, it is typically not necessary to run more than 10–20 randomizations before concluding with high probability that the discovered region is not significant. For example, if four or more of the first 10 replicas beat the original grid, we know that this result will only occur 0.1% of the time if the region is significant, so we can safely assume that the region is not significant. Thus our true average 'worst-case' results will be closer to the $95\times$ speedup on small, significant (but not extremely dense) test regions. Since the naive approach requires approximately 45 days for a 256×256 grid with $R = 1000$, this suggests that our algorithm can complete the same task in less than 12 hours.

We now discuss the performance of the algorithm on various real-world datasets. Our first test set was a database of (anonymized) emergency department (ED) data collected from Western Pennsylvania hospitals in the period 1999–2002. This dataset contained a total of 630 000 records, each representing a single ED visit and giving the latitude and longitude of the patient's home location to the nearest 0.005 degrees. These locations were mapped to three grid sizes: $N = 128$, 256, and 512. For each grid, we tested for spatial clustering of 'recent' disease cases: the 'count' of a square was the number of ED visits in that square in the last 2 months, and the 'population' of a square was the total number of ED visits in that square. See Figure 11.3 for a picture of this data set, including the highest scoring region. For each of these grids, our fast algorithm found the same, statistically significant region (p-value 1/1001) as the naive approach. The major difference, of course, was in runtime and number of regions searched (see Table 11.2). Our algorithm found the mdr of the original grids $22–24\times$ faster than the naive approach; however, much faster performance was achieved when searching the replica grids. The algorithm achieved speedups increasing from $450\times$ to $4700\times$ as grid size increased from 128 to 512.

Our second test set was a nationwide database of retail sales of over-the-counter cough and cold medication. Sales figures were reported by zip code; the data covered 5000 zip codes across the USA, with highest coverage in the Northeast. In this case, our goal was to see if the spatial distribution of sales on a given day (February 14, 2004) was significantly different than the spatial

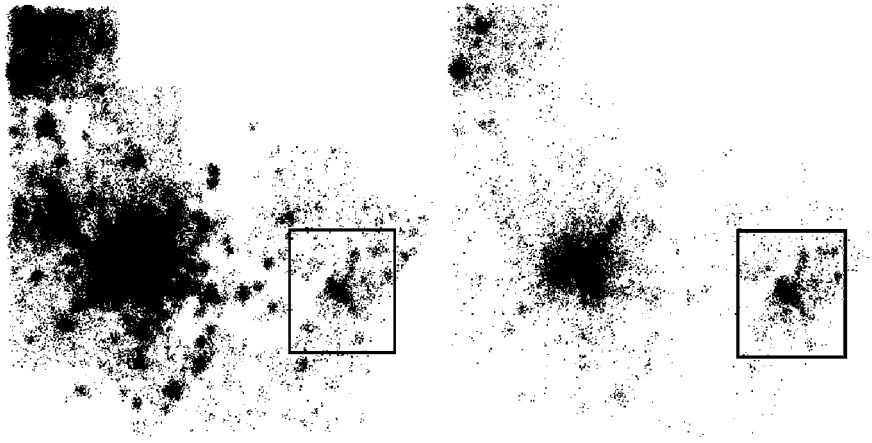


Figure 11.3 Emergency department data set. The left-hand picture shows the 'population' distribution and the right-hand picture shows the 'counts'. The winning region is shown as a rectangle.

Table 11.2 Performance of algorithm, real-world data sets. For each data set, we give the time in seconds to search the original grid and each replica grid, as well as the number of regions searched. The speedup is the ratio of runtimes of the naive and fast approaches.

test	method	sec/orig	speedup	sec/rep	speedup	regions (orig)	regions (rep)
ED	naive	72	×1	68.0	×1	62.0M	62.0M
($N = 128$)	fast	3	×24	0.15	×453	5.12M	15.9K
ED	naive	1 207	×1	1 185.0	×1	1.03B	1.03B
($N = 256$)	fast	55	×22	1.2	×988	95.9M	74.7K
ED	naive	19 146	×1	18 921.0	×1	16.8B	16.8B
($N = 512$)	fast	854	×22	4.0	×4730	1.51B	120K
national OTC	naive	71	×1	77.0	×1	62.0M	62.0M
($N = 128$)	fast	2	×36	0.8	×96	682K	200K
national OTC	naive	1 166	×1	1 232.0	×1	1.03B	1.03B
($N = 256$)	fast	14	×96	2.8	×440	3.24M	497K
regional OTC	naive	78	×1	79.0	×1	62.0M	62.0M
($N = 128$)	fast	2	×39	0.6	×132	783K	101K
regional OTC	naive	1 334	×1	1 330.0	×1	1.03B	1.03B
($N = 256$)	fast	13	×103	1.8	×739	3.10M	168K

distribution of sales a week before (February 7), and to identify a significant cluster of increased sales if one exists. (Note that the statistic adjusts for increases or decreases in the total number of sales; clusters are only detected if there is spatial variation in the amount of increase/decrease.) Thus we used the sales on February 7 as our underlying population distribution, and the sales

on February 14 as our count distribution. Slight modifications to Kulldorff's statistic were necessary to deal with regions with zero population and nonzero count (i.e. sales on February 14 but not February 7). We created four grids from this data, two using all of the national data, and two using only data from the Northeast (where a greater proportion of zip codes report sales data). For both 'national' and 'regional' over-the-counter data, we created grids of sizes $N = 128$ and $N = 256$, converting each zip code's centroid to a latitude and longitude. For each of these four grids, our algorithm found the same statistically significant region (p-value $1/1001$) as the naive approach, and achieved speedups of 96–$132\times$ on the 128×128 grids and 440–$739\times$ on the 256×256 grids (Table 11.2).

Thus the algorithm found the maximum density region in all of our simulated and real-world trials, while achieving speedups of at least $20\times$ (and typically much larger) as compared to the naive approach.

11.3.1 Comparison to SaTScan

It is difficult to evaluate the computational speed of an algorithm in isolation, and thus a comparison to other techniques in the literature is necessary. We note, however, that none of the prior algorithmic work on scan statistics allows for the detection of *elongated* clusters; the detection of compact clusters (e.g. circles or squares) is a significantly easier computational task, since there is one less degree of freedom to search over. Thus the most accurate comparison is to the obvious technique of naively searching all rectangles; this comparison was done in the previous section. However, since no available software actually uses this 'naive rectangles' approach, we feel that a comparison to other techniques (though inexact at best) will be useful.

In particular, we focus on Martin Kulldorff's SaTScan software (www.satscan.org), also discussed by Kulldorff in Chapter 7 of this collection. SaTScan represents the current state of the art in cluster detection, and is widely used in the epidemiological community. We emphasize that this is not an 'apples to apples' comparison: because of the inexactness of this comparison and the inherent differences between the two methods of cluster detection, it is difficult to draw general conclusions. In particular, there are three main differences between the methods. First, as noted above, our algorithm searches for elongated clusters (in particular, axis-aligned rectangles), while SaTScan searches for compact clusters (in particular, circles). Thus (assuming that M is the number of distinct spatial locations) our algorithm must search over the $O(M^4)$ possible rectangles, while SaTScan must search over the $O(M^3)$ possible circles. Second, neither our algorithm nor SaTScan actually searches over 'all' of the regions of the given type (rectangles or circles). SaTScan searches only circles centered at one of the data points, reducing the search space to $O(M^2)$ regions. Our method, on the other hand, aggregates the data points to a uniform

$N \times N$ grid, and searches over the $O(N^4)$ gridded rectangular regions. Thus our method's runtime is a function of the grid resolution N, while SaTScan's runtime is a function of the number of spatially distinct data points M. If each data point truly represents cases occurring at that precise spatial location, we are losing some precision by aggregating points to a grid; however, this loss of precision is minimal for high grid resolutions N. Also, in cases where data points are derived from regions (e.g. representing a census tract or zip code by a point mass at the center of that region) then the assumption of discrete data points is itself somewhat inexact. Finally, both our method and SaTScan use clever computational techniques to speed up performance: our pruning method allows us to search only a small subset of the $O(N^4)$ gridded rectangular regions, while obtaining the same results as if we had searched all of these regions. SaTScan, though it does not use pruning to speed up the search (and thus, must actually search over all of the $O(M^2)$ regions), uses an 'incremental addition' technique which allows searching in constant time per region. (We also achieve constant search time per region, using the 'cumulative counts' trick noted in Section 11.1.3.)

As a simple comparison, we ran both our method and SaTScan on the emergency department data set discussed above. This data set consisted of 630 000 records, of which the last 60 000 (recent data) were used as 'counts' and the entire data set was used as population. Since many records corresponded to identical spatial locations, this gave us approximately $M = 17\,000$ distinct spatial locations. We ran both our method and SaTScan on this dataset, using the same system (Pentium 4, 1800 MHz processor, 1 GB RAM) for each. For all runs, we used 999 Monte Carlo replications. Our system found the most significant rectangular region in 11 minutes for a 128×128 grid and 81 minutes for a 256×256 grid, computing a p-value of 1/1000 in each case. SaTScan ran out of memory and thus was unable to find the most significant circular region for this data set; in comparison, our method requires very little memory (less than 50 MB for grid sizes up to 256×256). Thus we instead ran SaTScan on one-tenth of the data (60 000 records, 10 000 used as 'count'), containing $M = 8400$ distinct spatial locations. In this case, SaTScan found the most significant circular region in 4 hours; this suggests that (given sufficient memory) it would find the most significant circular region for the entire data set in approximately 16.5 hours.

We note that, for the smaller data set, both methods found very similar spatial regions. SaTScan found a circle with center coordinates (40.34° N latitude, 79.82° W longitude) and diameter 18.58 km, with $C = 2458$, $P = 8443$, and a score (log-likelihood ratio) of 413.56. For a 128×128 grid size, our method found a rectangle with almost the same centroid (40.32° N latitude, 79.82° W longitude), and size 23.6×17.2 km. This slightly larger region had $C = 2599, P = 9013$, and a score of 429.85. In this case, the most significant rectangular region has a low aspect ratio, so, as expected, the region and score are similar to that found by SaTScan. If, on the other hand, the most significant

rectangular region has a high aspect ratio, we would expect our algorithm to find a region with significantly higher score.

We emphasize again that this comparison between our method and SaTScan is both preliminary (testing only on a small sample of data sets) as well as inexact (because of the differences between the algorithms discussed above). Thus we do not attempt to draw any general conclusions about the relative speeds of the two methods; we note only that our 'fast spatial scan' is able to find elongated clusters in times comparable to (and in at least some cases, significantly faster than) the detection of compact clusters by SaTScan. Since SaTScan is in wide use in the epidemiological community, this demonstrates that the runtime of our method is sufficiently fast to be useful for the detection of significant spatial clusters.

11.4 CONCLUSIONS AND FUTURE WORK

We have presented a fast multiresolution partitioning algorithm for detection of significant spatial overdensities, and demonstrated that this method results in significant (20–2000×) speedups on real and artificially generated data sets. We are currently applying this algorithm to national-level hospital and pharmacy data, attempting to detect disease outbreaks based on statistically significant changes in the spatial clustering of disease cases. Our eventual goal is the automatic real-time detection of outbreaks, and application of a fast partitioning method using the techniques presented here may allow us to achieve this difficult goal.

12

Bayesian Data Mining for Health Surveillance

David Madigan

12.1 INTRODUCTION

Data mining concerns the extraction of useful knowledge from data. Statistical tools and ideas obviously lie at the core of data mining, but since data mining usually (but not always) focuses on larger-scale data repositories, computing issues come to the fore. Data mining textbooks, by contrast with statistics textbooks, describe algorithms and pay careful attention to issues of feasibility and scale (see, for example, the outstanding text of Hand *et al.*, 2001). By 'data mining for health surveillance' I mean applications of data mining to health-related, observational, timestamped data. Many applications will additionally concern spatially referenced data. 'Surveillance' performs ongoing monitoring of such data and discriminates between normal conditions and anomolous conditions of one sort or another.

The Bayesian approach to statistical analysis and data mining computes conditional probability distributions of quantities of interest (such as future observables) given the observed data. Bayesian analyses usually begin with a *full probability model* – a joint probability distribution for all the observable and unobservable quantities under study – and then use Bayes' theorem to compute the requisite conditional probability distributions. In fact, the theorem prescribes the basis for statistical learning in the probabilistic framework. Computing is the big issue confronting a data miner working in the Bayesian framework. The computations required by Bayes' theorem can be demanding, especially with large data sets. In fact, widespread application of Bayesian data analysis methods has only occurred in the last decade or so, having had to wait for computing power as well as breakthroughs in simulation technology. Barriers still exist for truly large-scale applications.

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

The primary advantages of the Bayesian approach are its conceptual simplicity and the common-sense interpretation of Bayesian outputs. The ability to incorporate prior knowledge can also be a boon. Many data mining applications provide copious data, but for models with thousands if not millions of dimensions or parameters, a limited amount of prior knowledge, often in the form of prior exchangeability information, can sharpen inferences considerably. Perhaps more commonly though, the available data simply swamp whatever prior knowledge is available, and the precise specification of the prior becomes irrelevant.

This chapter explores one particular Bayesian approach to surveillance. Specifically, I consider multivariate temporal probabilistic models. Some variables in the model represent the true state of the world at a particular time (and possibly place). Other variables correspond to observables. The state variables are unobservable and for these we might or might not posit such values as 'normal' and 'abnormal', or 'normal', 'flu season', and 'epidemic'. With a broad enough definition, the term 'hidden Markov model' (HMM) coincides with the class of models this paper discusses. I use the language of graphical models to describe these models and focus on Markov chain Monte Carlo (MCMC) to learn the models from data and to make inferences. In keeping with the data mining theme, I focus on models that do not require reference or nonepidemic data, but instead work with all available historical data.

In what follows I consider simple applications and relatively simple models. Application of the kinds of semi-latent Bayesian methods I discuss here to more realistic problems, while conceptually straightforward, present computational challenges. I will return to this at the end.

12.2 PROBABILISTIC GRAPHICAL MODELS

The use of graphs to represent statistical models has a rich history dating back at least to the 1920s (Wright, 1921). Recently, probabilistic graphical models have emerged as an important class of models and have impacted fields such as data mining, causal analysis, and statistical learning. A probabilistic graphical model is a multivariate probabilistic model that uses a graph to represent a set of conditional independences. The vertices of the graph represent the random variables of the model and the edges encode the conditional independences. In general, each missing edge corresponds to a conditional independence. Graphs with different types of edges – directed, undirected, or both – lead to different classes of probabilistic models. In what follows we will only consider acyclic directed models, also known as *Bayesian networks*. This is somewhat of a misnomer since there is nothing Bayesian *per se* about Bayesian networks.

Spiegelhalter and Lauritzen (1990) presented a Bayesian analysis of acyclic directed probabilistic graphical models, and this topic continues to attract research attention. Here we sketch the basic framework with a stylized version of a real epidemiological application.

In Norway, the Medical Birth Registry (MBR) gathers data nationwide on congenital malformations such as Down's syndrome. The primary purpose of

the MBR is to track prevalences over time and identify abnormal trends. The data, however, are subject to a variety of errors, and epidemiologists have built statistical models to make inference about true prevalences. For Down's syndrome, such a model includes three dichotomous random variables: the reported Down's syndrome status, R, the true Down's syndrome status, S, and the maternal age, A, where age is dichotomized at 40, say.

Figure 12.1 displays a possibly reasonable model for these variables. This acyclic directed graph represents the assumption that the reported status and the maternal age are conditionally independent given the true status. The joint distribution of the three variables factors accordingly:

$$\Pr(A, S, R) = \Pr(A)\Pr(S \mid A)\Pr(R \mid S). \tag{12.1}$$

This factorization features a term for every vertex, the term being the conditional density of the vertex given its parents. In general, this factorization implies that vertices (more correctly, the random variables corresponding to vertices) are conditionally independent of their nondescendants given their parents (Lauritzen *et al.*, 1990).

The specification of the joint distribution of A, S, and R in (12.1) requires five parameters – $\Pr(R \mid S), \Pr(R \mid \overline{S}), \Pr(S \mid A), \Pr(S \mid \overline{A})$, and $\Pr(A)$ – where $\overline{A}$ denotes maternal age less than 40 and $\overline{S}$ denotes the absence of Down's syndrome. Once these probabilities are specified, the calculation of specific conditional probabilities such as $\Pr(R \mid A)$ can proceed via a series of local calculations without storing the full joint distribution (Dawid, 1992).

To facilitate Bayesian learning for the five parameters, Spiegelhalter and Lauritzen (1990) and Cooper and Herskovits (1992) make two key assumptions that greatly simplify subsequent analysis. First, they assume that the parameters are independent a priori. Figure 12.2 embodies this assumption. For instance,

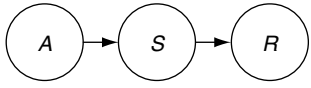


Figure 12.1 Down's syndrome: an acyclic directed graphical model.

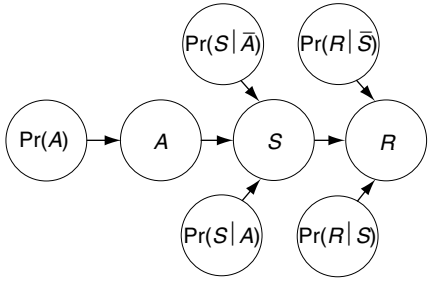


Figure 12.2 Down's syndrome: an acyclic directed Bayesian graphical model.

$\Pr(S \mid A)$ in Figure 12.2 has no parents. Therefore, it is marginally independent of, for instance, $\Pr(A)$, since this is not a descendant of $\Pr(S \mid A)$. Second, they assume that each of the probabilities has a beta distribution (or Dirichlet distribution for categorical variables with more than two levels). This assumption results in closed-form expressions for posterior distributions and for the marginal likelihood. Unforunately, the models I consider below do not yield closed-form estimates and I resort to MCMC via the BUGS language and software.

BUGS is a useful tool for Bayesian data mining. The UK Medical Research Council at Cambridge has developed BUGS over the last decade. The program is avaiable free of charge from: http://www.mrc-bsu.cam.ac.uk/bugs/. There are versions for Unix, DOS, and Windows (WinBUGS). The BUGS manual (Spiegelhalter *et al.*, 1999) describes BUGS:

> BUGS is a computer program that carries out Bayesian inference on statistical problems using Gibbs sampling. BUGS assumes a Bayesian or full probability model, in which all quantities are treated as random variables. The model consists of a defined joint distribution over all unobserved (parameters and missing data) and observed quantities (the data); we then need to condition on the data in order to obtain a posterior distribution over the parameters and unobserved data. Marginalising over this posterior distribution in order to obtain inferences on the main quantities of interest is carried out using a Monte Carlo approach to numerical integration (Gibbs sampling).

Several authors have described MCMC algorithms for HMMs. I refer the reader to Scott (2002) for an excellent overview.

12.3 HIDDEN MARKOV MODELS FOR SURVEILLANCE: ILLUSTRATIVE EXAMPLES

Le Strat and Carrat (1999) pioneered the use of hidden Markov models for surveillance, albeit from a non-Bayesian perspective. Their first application concerned surveillance of a univariate inluenza-like illness (ILI) time series. The French Sentinelles Network provided the data. This is a national surveillance system that includes about 1% of all general practitioners in France. The Network defines ILI as the combination of a sudden fever of at least 39 °C with respiratory signs and myalgia. Figure 12.3 shows weekly ILI incidence rates for the same period that Le Strat and Carrat analyzed. The website at http://www.b3e.jussieu.fr/sentiweb publishes these data. The figure shows two dynamics – a low-level dynamic with incidence rates that vary according to a seasonal pattern (the nonepidemic pattern) and a high-level dynamic in which the incidence rate increases sharply at irregular intervals (the epidemic dynamic). Le Strat and Carrat state their primary interest as the timing of ILI epidemics. They define an epidemic as 'the occurence of a number of cases of a disease, in a given period of time and in a given population, that exceeds the expected number'. They continue:

> This definition thus assumes a mixture of two (or more) dynamics – one for the 'expected' number of cases, another for the 'excess' cases. Hidden Markov models

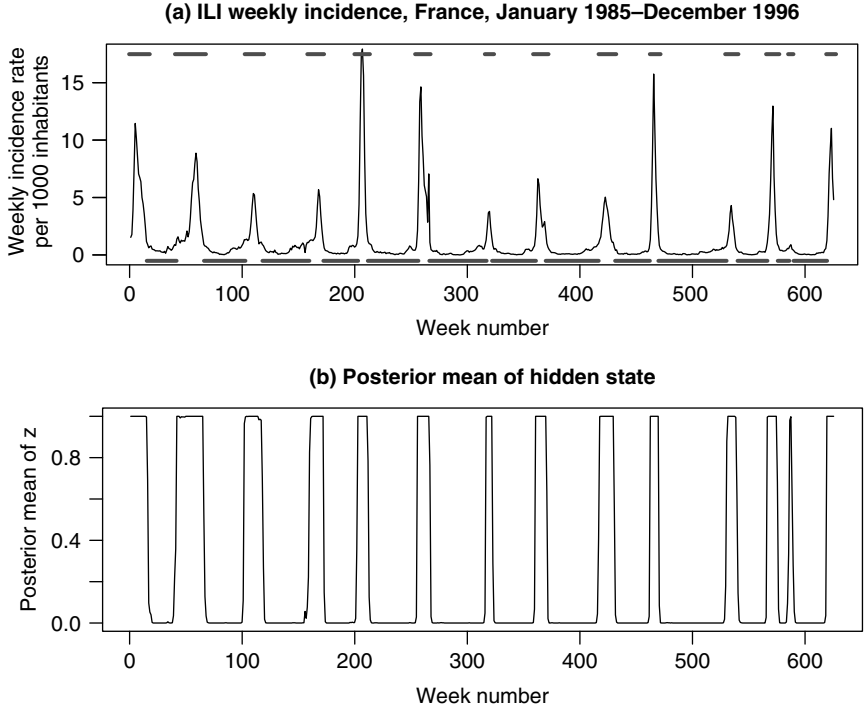


Figure 12.3 The French ILI data. (a) Incidence rates per 1000 inhabitants. (b) Posterior mean of the hidden state from a Gaussian two-state HMM. The horizontal line segments in (a) correspond to time periods where the posterior mean of the hidden state exceeds 0.5.

provide the most natural way of making inferences about such phenomena, by assigning different probability distrubutions to the two dynamics.

Hidden Markov models represent a subclass of the more general graphical models, and Figure 12.4 presents an acyclic directed graphical model for the standard HMM. Each pair of vertices represents a mixture model, that is, a pair $(z_t, y_t), t = 1, \ldots, n$, with $z_t \in \{1, \ldots, K\}$ and $y_t | z_t \sim f_{z_t}(y_t)$. Generally the

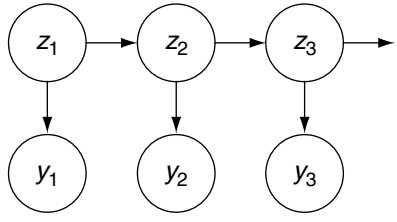


Figure 12.4 Standard hidden Markov model.

y_t are observed and the z_t are not. For $t \in \{1, \ldots, n\}$, y_t is conditionally independent of all the remaining variables conditional on z_t. Marginally, the z_t form a first-order Markov chain. The y_t can be multivariate. The z_t could also, in principle, be multivariate and/or continuous, but for now we will only consider the univariate categorical case. Fields such as speech processing, finance, and bioinformatics make extensive use of this general model structure – see, for example, MacDonald and Zucchini (1997) and the references therein. Robert *et al.* (1993) presented a Bayesian analysis of the standard HMM. Robert *et al.* (2000) extended this analysis using a reversible jump algorithm to make inference about K.

For the French ILI data, Le Strat and Carrat consider models with different values of K but focus primarily on the case $K = 2$. For the conditional distribution of y_t given z_t, they posit a Gaussian model with state-dependent mean, μ_j, and precision, τ_j, $j = 1, \ldots, K$. Serfling's cyclic regression method is a widely used surveillance method that accounts for seasonality and trend (Serfling, 1963). Le Strat and Carrat incorporate Serfling's method via a model for μ_j:

$$\mu_j(t) = \gamma_j + \beta_j t + \delta_j \cos\left(\frac{2\pi t}{r}\right) + \epsilon_j \left(\frac{2\pi t}{r}\right).$$

Here β_j represents a state-specific trend and δ_j and ϵ_j are state specific parameters associated with an r-period seasonality.

We present a Bayesian analysis of the Le Strat and Carrat model. This requires prior distributions for the various parameters which we specify as follows for $j = 1, \ldots, K$:

$$\begin{aligned}
\tau_j &\sim \text{dgamma}(0.001, 0.001),\\
\gamma_j &\sim \text{dnorm}(0.0, 10^{-6}),\\
\beta_j &\sim \text{dnorm}(0.0, 10^{-6}),\\
\delta_j &\sim \text{dnorm}(0.0, 10^{-6}),\\
\epsilon_j &\sim \text{dnorm}(0.0, 10^{-6}),\\
p_{j,1:K} &\sim \text{ddirch}(\alpha[1:K])
\end{aligned}$$

Here 'dnorm(μ, τ)' represents a normal distribution with mean μ and precision τ, 'dgamma(α, β)' represents a gamma distribution with mean α/β and precision β^2/α, and 'ddirch' represents a Dirichlet distribution. $p[j, 1:K]$ denotes the vector of K transition probabilities from state $j, j = 1, \ldots, K$, and $\alpha[1:K]$ is a user-specified hyperparameter vector for $p[j, 1:K]$'s Dirichlet prior. In our analyses, α is always a vector of K ones. These various choices for the prior distributions reflect minimal prior knowledge.

Figure 12.3(b) shows the posterior mean of the state variable z for each time point. These MCMC results used 11 000 iterations and discarded the first 1000.

These results closely mirror those of Le Strat and Carrat's Figure 2. Their non-Bayesian analysis used the Viterbi algorithm to estimate the most likely state sequence. The hidden state variable clearly picks out the elevated periods in the original series. Figure 12.5 shows the corresponding posterior distributions of the parameters. The posterior variances differ sharply between the two hidden states.

Le Strat and Carrat presented a second application concerning monthly poliomyelitis cases in the USA from January 1970 to December 1983. These data are from http://www.maths.monash.edu.au/~hyndman/tseries. Here we model the observed counts via mixtures of Poisson distributions with state-specific parameter $\lambda_j(t)$ given by

$$\log(\lambda_j(t)) = \gamma_j + \beta_j t + \delta_j \cos\left(\frac{2\pi t}{r}\right) + \epsilon_j \left(\frac{2\pi t}{r}\right).$$

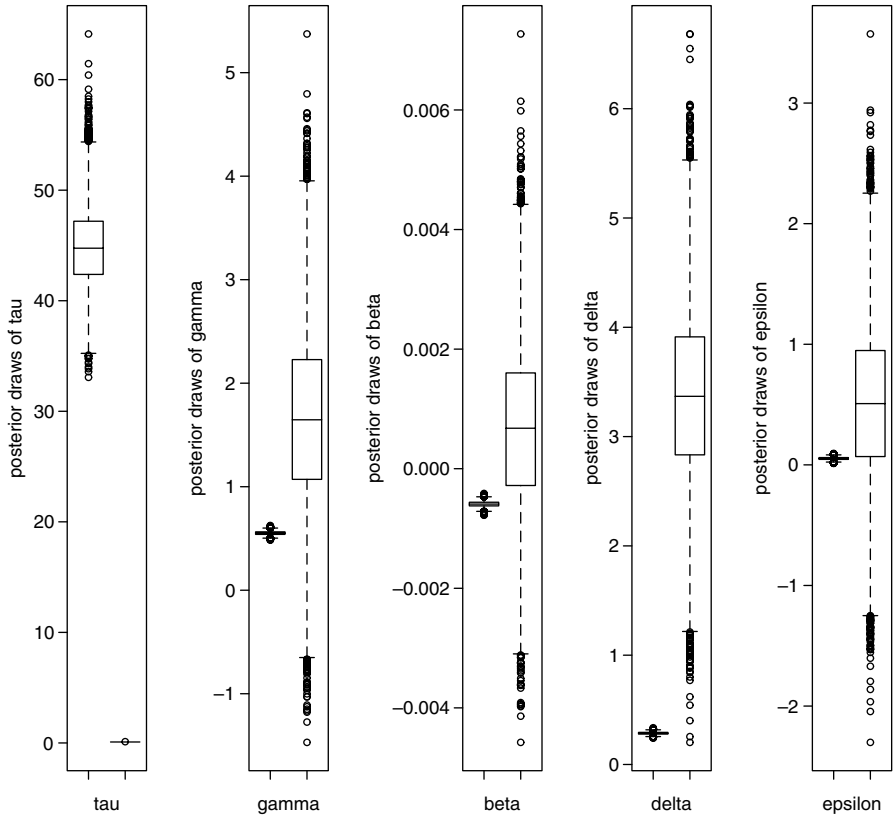


Figure 12.5 Posterior distributions for the Le Strat and Carrat model. The right (left) boxplot in each panel represents the posterior distribution conditional on the hidden state equal to one (zero).

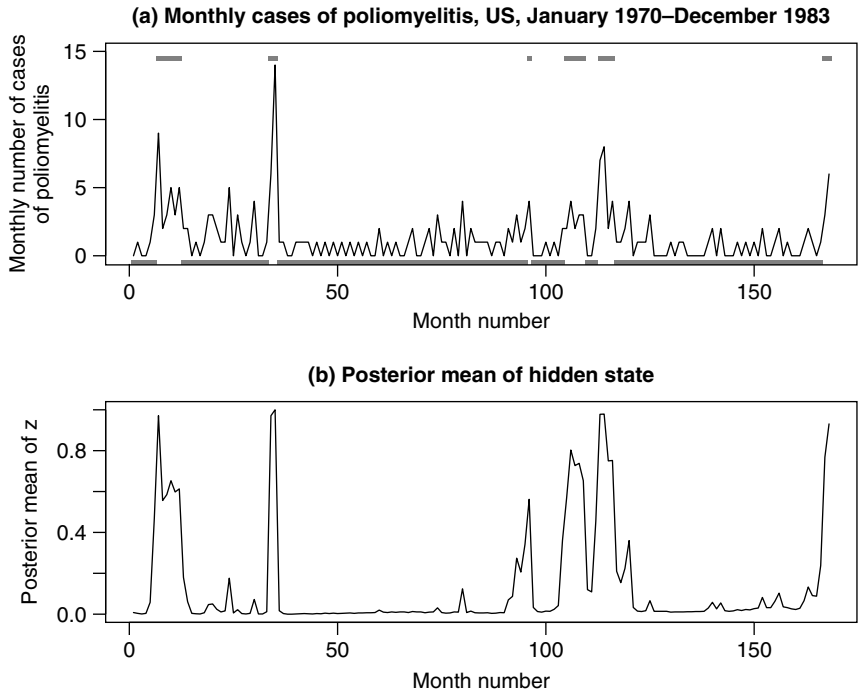


Figure 12.6 The US poliomyelitis data. (a) Total reported monthly cases. (b) Posterior mean of the hidden state from a Poisson two-state HMM. The horizontal line segments in (a) correspond to time periods where the posterior mean of the hidden state exceeds 0.5.

Figure 12.6 shows the data and the corresponding state variable posterior means.

12.4 HIDDEN MARKOV MODELS FOR SURVEILLANCE: FURTHER EXPLORATION

The Bayesian approach to HMM analysis combined with MCMC provides a highly flexible framework for model exploration. This section looks at several extensions of the basic model.

12.4.1 Beyond Normality

Returning to the ILI example, Rath *et al.* (2003) note the discordance of the nonnegativity of incidence rates and the Gaussian assumption. They propose, analyze, and defend a model with an exponential distribution for the

nonepidemic incidence rates. Here I consider in addition a more general gamma model and a lognormal model.

To compare competing models, Le Strat and Carrat use the Bayesian information criterion (BIC) with a penalty term that counts the number of free parameters in the model but ignores the hidden state variable. In my Bayesian MCMC context, model scores that use MCMC output provide a more convenient alternative. One possibility is to use MCMC to compute the marginal likelihood, and I discuss this further below. Another possibility focuses on appropriately penalized simulated log-likelihoods. Scott *et al.* (2004), for instance, consider the posterior distribution of penalized log-likelihood values produced by the MCMC sampler. They use the same penalty term as Le Strat and Carrat. Spiegelhalter *et al.* (2002) introduced the deviance information criterion (DIC),

$$DIC = \overline{D(\boldsymbol{\theta})} + p_D$$

where $\overline{D(\boldsymbol{\theta})}$ is the average deviance (i.e., minus twice the log-likelihood) with respect to the posterior distribution of the parameters, $\boldsymbol{\theta}$, and p_D is the effective number of parameters for the model. Celeux *et al.* (2003) discuss DICs for hidden Markov and other mixture models. Their analysis suggests that 'complete DICs', where

$$\overline{D(\boldsymbol{\theta})} = -2E_{\boldsymbol{\theta},z}[\log p(y, z|\boldsymbol{\theta})|y],$$

outperform other DICs, and this is the approach I adopt here. Concerning the p_D term, I adopt the suggestion of Gelman *et al.* (2003, p. 182) and set p_D equal to one-half the variance of the simulated deviances. I refer to the resulting model score as DIC_C^V. Table 12.1 shows the results for different distributional assumptions in the two-state $K = 2$ model.

This analysis suggests that a lognormal distribution provides a better fit than the other three. However, several modeling directions await exploration. For example, my analysis assumes that the various state-dependent regression parameters $(\gamma_j, \beta_j, \delta_j, \epsilon_j)$ remain constant over time. Furthermore, I have only considered models where the distribution of the observed values is the same for the epidemic and nonepidemic states. The flexibility of the graphical modeling framework combined with MCMC renders such explorations relatively straightforward.

Table 12.1 Model scores for different distributional assumptions.

Epidemic state	Nonepidemic state	$\overline{D(\boldsymbol{\theta})}$	s.d.$(D(\boldsymbol{\theta}))$	DIC_C^V
Gaussian	Gaussian	629.0	10.6	685.5
Lognormal	Lognormal	122.3	25.2	440.3
Gamma	Gamma	405.4	15.3	522.4
Exponential	Exponential	476.0	17.6	630.9

12.4.2 How Many Hidden States?

The DIC_C^V score can also help evaluate a modest number of different choices for the size of the hidden state space. Table 12.2 shows results for the ILI data and the Gaussian model. The three-state model provides the best score. Le Strat and Carrat's BIC score favored a five-state model.

In the Le Strat and Carrat surveillance context, models with more than two states present interpretational challenges and both Le Strat and Carrat (1999) and Rath *et al.* (2003) focused exclusively on the two-state model. Figure 12.7 shows the output of the three-state model where I sorted the three states according to the posterior means of $\mu_j, j = 1, 2, 3$. Arguably, the third state does a better job than the higher state in the two-state model. In particular, the three-state model does not classify the blip around week 588 as 'epidemic'. I return to the model interpretation issues at the end.

12.4.3 Label Switching

Note that the HMM likelihood is invariant under abitrary permutations of the state labels. As a consequence, when two hidden states are similar to each other, MCMC draws can swap the labels of these two states. In our analyses, we imposed the constraint that $\mu_k > \mu_j$ when $k > j$, and subsequent analyses showed no evidence of label switching. However, this constraint does alter the basic model, and we refer the reader to Scott (2002) for a discussion of this issue.

12.4.4 Multivariate Extensions

For univariate time series, the preceding machinery seems rather excessive. In a multivariate setting, however, the HMM approach becomes more interesting. Figure 12.8, for example, shows a particular model for three-dimensional observations. Here the first two components of y depend directly on the hidden state, while the third component depends indirectly on the hidden state and

Table 12.2 Model scores for different numbers of hidden states.

K	$\overline{D(\boldsymbol{\theta})}$	s.d.$(D(\boldsymbol{\theta}))$	DIC_C^V
1	2729.0	3.0	2733.5
2	629.0	10.6	685.5
3	210.8	20.7	424.0
4	122.5	30.6	589.8
5	122.3	35.0	734.1
6	51.9	43.1	980.3

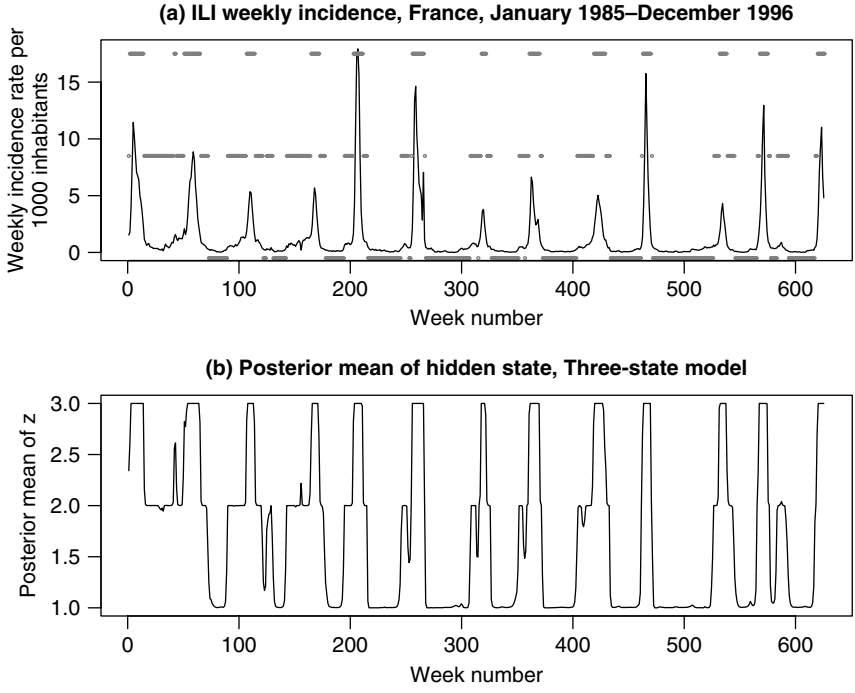


Figure 12.7 The French ILI data. (a) Incidence rates per 1000 inhabitants. (b) Posterior mean of the hidden state from a Gaussian three-state HMM. The horizontal line segments in (a) correspond to the three different states.

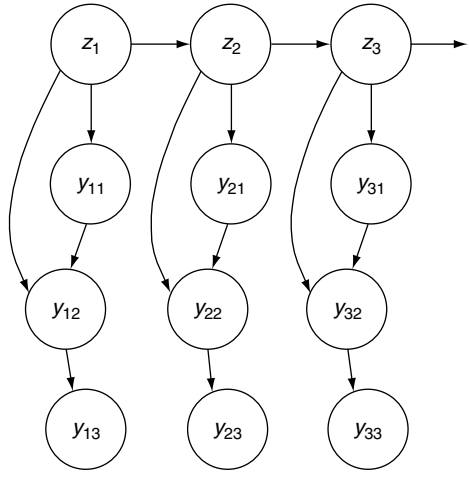


Figure 12.8 A multivariate hidden Markov model.

directly on the second component of y. For example, the first component of y might correspond to school absenteeism, the second to emergency room respiratory syndrome counts, and the third to unit sales of over-the-counter cough medications.

In addition to the modeling issues I discussed above, multivariate models raise issues such as model space exploration, observations measured at different times, and lagged observations. In the next section, I explore one particular issue concerning random observation times.

12.5 RANDOM OBSERVATION TIME HIDDEN MARKOV MODELS

The standard HMM assumes that observations arrive at equally spaced timepoints. However, in many surveillance applications, observations arrive at random times. For example, several US public health departments now carry out routine surveillance of sales of over-the-counter medications and also chief complaints at emergency rooms. Both data sources feature irregularly spaced obsverations. Furthermore, the elapsed times between observations themselves carry useful information.

More specifically, I consider a situation in which observations y_t arrive at random, state-dependent times. Let δ_t denote the elapsed time between y_{t-1} and y_t. Figure 12.9 presents one possible model. One instantiation of this model makes the following distributional assumptions. First, I model the observed y_t's as zero-mean normals:

$$[y_t|z_t = i] \sim \text{dnorm } (0, \sigma_i^2), \quad i = 0, \ldots, K-1.$$

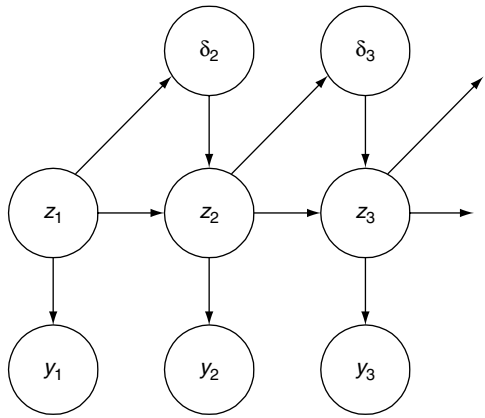


Figure 12.9 Random observation time hidden Markov model.

Second, I assume that the δ_t are integer-valued and geometrically distributed:

$$[\delta_t | z_{t-1} = i] \sim \text{geometric } (p_i), \quad i = 0, \ldots, K-1.$$

Third, I assume that the z_t arise from an underlying continuous-time Markov chain that we denote Z. I have chosen the first two specific assumptions to demonstrate the methodology; alternative assumptions lead to similar derivations.

We initially consider the case where $K = 2$. Assume that Z stays in state 0 for an exponential amount of time having mean $1/\lambda_0$ before switching to state 1. Similarly, assume that Z stays in state 1 for an exponential amount of time having mean $1/\lambda_1$ before switching to state 0. Standard continuous-time Markov chain theory (see Ross, 1989, p. 263) then provides the following probabilities:

$$p(z_t = 0 | z_{t-1} = 0, \delta_t = t) = \frac{\lambda_1}{\lambda_0 + \lambda_1} + \frac{\lambda_0}{\lambda_0 + \lambda_1} \exp^{-(\lambda_0 + \lambda_1)t},$$

$$p(z_t = 0 | z_{t-1} = 1, \delta_t = t) = \frac{\lambda_1}{\lambda_0 + \lambda_1} - \frac{\lambda_1}{\lambda_0 + \lambda_1} \exp^{-(\lambda_0 + \lambda_1)t}.$$

For $K > 2$, I restrict attention to the case where Z is a birth and death process with reflecting boundaries. That is, when Z is in state $i, i \in \{1, \ldots, K-2\}$, the only possible transitions are to state $i-1$ or to state $i+1$. From state 0, Z can only transition to state 1. Similarly, from state $K-1$, Z can only transition to state $K-2$. Denote by $\mathbf{Q}$ the infinitesimal generator of Z. $\mathbf{Q}$ is a $K \times K$ matrix with elements $q_{ij}, i = 0, \ldots, K-1,\ j = 0, \ldots, K-1$, defined as follows:

$$\begin{aligned}
q_{00} &= -\lambda_0, \\
q_{01} &= \lambda_0, \\
q_{K-1,K-1} &= -\lambda_{K-1}, \\
q_{K-1,K-2} &= \lambda_{K-1}, \\
q_{i,i} &= -\lambda_i, \\
q_{i,i+1} &= \beta_i \lambda_i, \\
q_{i,i-1} &= (1 - \beta_i)\lambda_i = \delta_i \lambda_i,
\end{aligned}$$

with $0 < \beta_i < 1$ for $1 \leq i \leq K-2$. Thus, starting in state i, Z sojourns there for a duration that is exponentially distributed with parameter λ_i. The process then jumps to state $j = i+1$ with probability β_i or to state $j = i-1$ with probability $1 - \beta_i$; the sojourn time in state j is exponentially distributed with parameter λ_j, and so on (Karlin and Taylor, 1975, p. 134).

The matrix of transition probabilities $p(z_t = j | z_{t-1} = i, \delta_t = t)$ is given by $\exp(t\mathbf{Q})$. Karlin and Taylor (1975, p. 152) and Bhattacharya and Waymire (1990, p. 315) describe an approach to the computation of these transition probabilities

via an eigendecomposition of $\mathbf{Q}$. Specifically, denote by $(\pi_0, \ldots, \pi_{K-1})$ the limiting probabilities associated with Z. These are given by:

$$\pi_1 = \frac{\lambda_0}{\lambda_1 \delta_1} \pi_0,$$
$$\pi_j = \left(\frac{\lambda_0}{\lambda_j}\right) \frac{\beta_1 \beta_2 \cdots \beta_{K-2}}{\delta_1 \delta_2 \cdots \delta_{K-2}} \pi_0, \qquad 2 \leq j \leq K-1,$$
$$\pi_0 = 1 - \sum_{i=1}^{K-1} \pi_i.$$

$\mathbf{Q}$ has K real eigenvalues $\alpha_0, \ldots, \alpha_{K-1}$ and corresponding eigenvectors $\mathbf{x}_0, \ldots, \mathbf{x}_{K-1} \in \mathbb{R}^K$. Then

$$p(z_t = j | z_{t-1} = i, \delta_t = t) = \sum_{l=0}^{K-1} x_{li} e^{t\alpha_l} x_{lj} \pi_j.$$

As an alternative to this spectral approach, Ross (1989, p. 286) suggests two approximation methods for the transition probabilities. In situations where the eigendecomposition is burdensome, Ross's approach could provide a viable alternative.

Along standard lines for Bayesian graphical models I assume that the joint density of the variables mentioned so far factors as follows:

$$[\mathbf{z}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\sigma}] = [\mathbf{p}][\boldsymbol{\lambda}][\boldsymbol{\beta}][\boldsymbol{\sigma}] \left\{ \prod_{i=1}^{n} [y_i | z_i, \boldsymbol{\sigma}] \right\} \times \left\{ \prod_{i=2}^{n} [z_i | z_{i-1}, \delta_i, \boldsymbol{\lambda}, \boldsymbol{\beta}][\delta_i | z_{i-1}, \mathbf{p}] \right\},$$

where $[\cdot]$ denotes a probability density. Prior distributions for $\mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\beta}$, and $\boldsymbol{\sigma}$ complete the model specification:

$$\left[\sigma_i^2\right] \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2), \qquad i = 0, \ldots, K-1,$$
$$[p_i] \sim \text{Beta}(\alpha_p, \beta_p), \qquad i = 0, \ldots, K-1,$$
$$[\lambda_i] \sim \Gamma(\alpha_\lambda, \beta_\lambda), \qquad i = 0, \ldots, K-1,$$
$$[\beta_i] \sim \text{Beta}(\alpha_\beta, \beta_\beta), \qquad i = 1, \ldots, K-2.$$

An MCMC procedure samples from each of the following conditional densities in turn:

$$[z_t | -] \propto [z_t | z_{t-1}, \delta_t, \lambda_0, \lambda_1] \left[z_{t+1} | z_t, \delta_{t+1}, \lambda_0, \lambda_1\right] \times \left[\delta_{t+1} | z_t, p_0, p_1\right] [y_t | z_t, \sigma_0, \sigma_1], \tag{12.2}$$

$$\left[\sigma_i^2 | -\right] \propto [\sigma_i] \prod_{\substack{t=1 \\ z_t = i}}^{n} [y_t | z_t, \sigma_i], \tag{12.3}$$

$$[p_i|-] \propto [p_i] \sum_{\substack{t=2 \\ z_{t-1}=i}}^{n} [\delta_t|z_{t-1}, p_i], \tag{12.4}$$

$$[\boldsymbol{\lambda}, \boldsymbol{\beta}|-] \propto [\boldsymbol{\lambda}][\boldsymbol{\beta}] \prod_{t=2}^{n} [z_t|z_{t-1}, \delta_t, \boldsymbol{\lambda}, \boldsymbol{\beta}], \tag{12.5}$$

where, for instance, $[z_t|-]$ denotes the conditional density of z_t given all the other unknowns, as well as $\mathbf{y}$ and $\boldsymbol{\delta}$.

For modest K, sampling from (12.2) is trivial, requiring K calculations and normalization step.

Since I chose the convenient scaled inverse-χ^2 prior distribution for $\boldsymbol{\sigma}$, (12.3) is available in closed form as

$$\left[\sigma_i^2|-\right] \sim \text{Inv-}\chi^2 \left(\nu_0 + n_i, \frac{\nu_0 \sigma_0^2 + n_i v_i}{\nu_0 + n_i} \right), \qquad i = 0, \ldots, K-1,$$

where $n_i = \sum I(z_t = i)$ and $v_i = n_i^{-1} \sum y_t^2 I(z_t = i)$.

A closed form for (12.4) also exists:

$$[p_i|-] \sim \text{Beta} \left(\alpha_p + \sum_{t=2}^{n} I(z_{t-1} = i), \right.$$
$$\left. \beta_p + \sum_{t=2}^{n} \delta_t I(z_{t-1} = i) - \sum_{t=2}^{n} I(z_{t-1} = i) \right), \qquad i = 0, \ldots, K-1.$$

I use a Metropolis step to sample values of $\boldsymbol{\Lambda} \equiv (\boldsymbol{\lambda}, \boldsymbol{\beta})$. Specifically, let $q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}')$ denote the proposal density for the transition from $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda}'$. This proposal density may depend on any or all of $\mathbf{z}, \mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\sigma}$ and $\mathbf{p}$. Then accept $\boldsymbol{\Lambda}'$ with probability

$$\alpha(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}'|\mathbf{z}, \boldsymbol{\delta}) = \min \left\{ 1, \frac{[\boldsymbol{\Lambda}'] \prod_{t=2}^{n} [z_t|z_{t-1}, \delta_t, \boldsymbol{\Lambda}']}{[\boldsymbol{\Lambda}] \prod_{t=2}^{n} [z_t|z_{t-1}, \delta_t, \boldsymbol{\Lambda}]} \times \frac{q(\boldsymbol{\Lambda}', \boldsymbol{\Lambda})}{q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}')} \right\}.$$

An independence Metropolis sampler for (12.5) is straightforward to implement and is the approach I adopt. For $i = 0, \ldots, K-1$, sample a candidate λ_i' uniformly in the interval $[\lambda_i - c_\lambda, \lambda_i + c_\lambda]$ where c_λ is chosen experimentally. Similarly (when $K > 2$), sample a candidate β_i' uniformly in the interval $[\beta_i - c_\beta, \beta_i + c_\beta]$ where c_β is chosen experimentally In this case $q(\boldsymbol{\Lambda}', \boldsymbol{\Lambda}) = q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') = (2c_\lambda)^{-K}(2c_\beta)^{-(K-2)}$.

Here, in contrast to the DIC approach above, I describe a method to compute the the marginal likelihood associated with specific values of K. For each candidate value of k of K, we wish to calculate the associated marginal likelihood, which we denote $m_k(\mathbf{y}, \boldsymbol{\delta})$. Chib and Jeliazkov (2001) describe

an ingenious approach for calculating marginal likelihoods from Metropolis–Hastings output. First, note the identity

$$m_k(\mathbf{y}, \boldsymbol{\delta}) = \frac{[\mathbf{y}, \boldsymbol{\delta}|\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*][\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*]}{[\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}]}, \tag{12.6}$$

where $\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*$, and $\mathbf{p}^*$ are arbitrary values of $\boldsymbol{\Lambda}, \boldsymbol{\sigma}$ and $\mathbf{p}$ respectively. I will compute each of three terms on the right-hand side of (12.6) separately.

$[\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*]$ requries an evaluation of the prior densities for $\boldsymbol{\Lambda}, \boldsymbol{\sigma}$ and $\mathbf{p}$ and is straightforward.

$[\mathbf{y}, \boldsymbol{\delta}|\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*]$ is not available directly, but we use an MCMC approach. From the Markov properties of the model, and letting $\boldsymbol{\theta}^* = \{\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*\}$, I derive the following:

$$\begin{aligned}
[\mathbf{y}, \boldsymbol{\delta}|\boldsymbol{\theta}^*] &= \int [\mathbf{y}, \boldsymbol{\delta}|\boldsymbol{\theta}^*, \mathbf{z}][\mathbf{z}|\boldsymbol{\theta}^*] d\mathbf{z} \\
&= \int [\mathbf{y}|\boldsymbol{\theta}^*, \mathbf{z}][\boldsymbol{\delta}|\boldsymbol{\theta}^*, \mathbf{z}][\mathbf{z}|\boldsymbol{\theta}^*] d\mathbf{z} \\
&= \int \left\{ \prod_{i=1}^{n} [y_i|z_i, \boldsymbol{\theta}^*] \right\} \left\{ \prod_{i=2}^{n} [\delta_i|z_i, z_{i-1}, \boldsymbol{\theta}^*] \right\} [\mathbf{z}|\boldsymbol{\theta}^*] d\mathbf{z} \\
&= \int \left\{ \prod_{i=1}^{n} [y_i|z_i, \boldsymbol{\theta}^*] \right\} \left\{ \prod_{i=2}^{n} [\delta_i|z_{i-1}, \boldsymbol{\theta}^*] \times [z_i|\delta_i, z_{i-1}, \boldsymbol{\theta}^*] \right\} d\mathbf{z}.
\end{aligned}$$

In particular, we have used the following facts:

$$\begin{aligned}
&\mathbf{y} \coprod \boldsymbol{\delta}|\mathbf{z}, \boldsymbol{\theta}^*, \\
&y_i \coprod z_{-i}, y_{-i}|z_i, \boldsymbol{\theta}^*, \\
&\delta_i \coprod \delta_{-i}, z_1, \ldots, z_{i-2}, z_{i+1}, \ldots, z_n|z_i, z_{i-1}, \boldsymbol{\theta}^*, \\
&z_i \coprod z_1, \ldots, z_{i-2}|z_{i-1}, \boldsymbol{\theta}^*
\end{aligned}$$

where $\coprod$ denotes conditional independence.
The MCMC procedure draws $\mathbf{z}^j, j = 1, \ldots, m$, as follows. First, set $z_1^j = z_1$. Then, for $i = 2, \ldots, n$, draw z_i^j from $[z_i|\delta_i, z_{i-1}^j, \boldsymbol{\theta}^*]$. Then for large m,

$$[\mathbf{y}, \boldsymbol{\delta}|\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*] \approx \frac{1}{m} \sum_{j=1}^{m} \left\{ \prod_{i=1}^{n} [y_i|z_i^j, \boldsymbol{\theta}^*] \right\} \left\{ \prod_{i=2}^{n} [\delta_i|z_{i-1}^j, \boldsymbol{\theta}^*] \right\}.$$

For the final term, $[\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}]$, note that

$$[\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}] = [\boldsymbol{\Lambda}^*|\mathbf{y}, \boldsymbol{\delta}][\boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*].$$

By multiplying both sides of the identity

$$\alpha(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*|\mathbf{z}, \boldsymbol{\delta}) q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*) [\boldsymbol{\Lambda}|\mathbf{z}, \mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\sigma}, p] = \alpha(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}|\mathbf{z}, \boldsymbol{\delta}) q(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}) [\boldsymbol{\Lambda}^*|\mathbf{z}, \mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\sigma}, \mathbf{p}]$$

by $[\boldsymbol{\sigma}, \mathbf{p}, \mathbf{z}|\mathbf{y}, \boldsymbol{\delta}]$ and integrating with respect to $(\boldsymbol{\Lambda}, \boldsymbol{\sigma}, \mathbf{p}, \mathbf{z})$, Chib and Jeliazkov (2001) show that

$$[\boldsymbol{\Lambda}^*|\mathbf{y}, \boldsymbol{\delta}] = \frac{E_1\{\alpha(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*|\mathbf{z}, \boldsymbol{\delta})q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*)\}}{E_2\{\alpha(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}|\mathbf{z}, \boldsymbol{\delta})\}} \tag{12.7}$$

where the numerator expectation E_1 is with respect to $[\boldsymbol{\Lambda}, \boldsymbol{\sigma}, \mathbf{p}, \mathbf{z}|\mathbf{y}, \boldsymbol{\delta}]$ and the denominator expectation E_2 is with respect to $[\boldsymbol{\sigma}, \mathbf{p}, \mathbf{z}|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*] \times q(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda})$. We can estimate each of the expectations in (12.7) via Monte Carlo.

To estimate the numerator, we take the draws $\{\boldsymbol{\Lambda}^j, \boldsymbol{\sigma}^j, \mathbf{p}^j, \mathbf{z}^j\}_{j=1}^m$ from the full MCMC run and average the quantity $\alpha(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*|\mathbf{z}, \boldsymbol{\delta})q(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*)$. The expectation in the denominator of (12.7) conditions on $\boldsymbol{\Lambda}^*$. Following Chib and Jeliazkov (2001), continue the MCMC simulation for a further m' iterations with the three conditional densities

$$[\boldsymbol{\sigma}|\mathbf{p}, \mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*, \mathbf{z}], [\mathbf{p}|\boldsymbol{\sigma}, \mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*, \mathbf{z}] \text{ and } [\mathbf{z}|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*, \boldsymbol{\sigma}, \mathbf{p}].$$

At each iteration of this reduced run, given the values $(\boldsymbol{\sigma}^j, \mathbf{p}^j, \mathbf{x}^j)$, draw $\boldsymbol{\Lambda}^j$ from $q(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda})$. Now $(\boldsymbol{\Lambda}^j, \boldsymbol{\sigma}^j, \mathbf{p}^j, \mathbf{x}^j)$ is a draw from $[\boldsymbol{\sigma}, \mathbf{p}, \mathbf{z}|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*] \times q(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda})$ and the marginal ordinate can be estimated as

$$[\boldsymbol{\Lambda}^*|\mathbf{y}, \boldsymbol{\delta}] \approx \frac{m^{-1}\sum_{j=1}^m \alpha(\boldsymbol{\Lambda}^j, \boldsymbol{\Lambda}^*|\mathbf{z}^j, \boldsymbol{\delta})q(\boldsymbol{\Lambda}^j, \boldsymbol{\Lambda}^*)}{m'^{-1}\sum_{k=1}^{m'} \alpha(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}^k|\mathbf{z}^k, \boldsymbol{\delta})}.$$

The final step is to estimate $[\boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*]$. Note that the values $\mathbf{z}^k$ from the reduced run are marginally from $[\mathbf{z}|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*]$. Thus, we have

$$[\boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*] \approx m'^{-1}\sum_{k=1}^{m'}[\boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*, \mathbf{z}^k].$$

Since $\boldsymbol{\sigma} \perp\!\!\!\perp \mathbf{p}|\{\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}, \mathbf{z}\}$, $[\boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*, \mathbf{z}^k]$ is available in closed-form as the product of the two densities (12.3) and (12.4).

To summarize, we represent the marginal likelihood as

$$m_k(\mathbf{y}, \boldsymbol{\delta}) = \frac{[\mathbf{y}, \boldsymbol{\delta}|\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*][\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*]}{[\boldsymbol{\Lambda}^*|\mathbf{y}, \boldsymbol{\delta}][\boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*]}.$$

We estimate $[\mathbf{y}, \boldsymbol{\delta}|\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*]$ via Monte Carlo. We compute $[\boldsymbol{\Lambda}^*, \boldsymbol{\sigma}^*, \mathbf{p}^*]$ directly. We estimate $[\boldsymbol{\Lambda}^*|\mathbf{y}, \boldsymbol{\delta}]$ reusing the draws from the full MCMC run as well as draws from a 'reduced run' that continues a version of the original MCMC run. Finally, we estimate $[\boldsymbol{\sigma}^*, \mathbf{p}^*|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\Lambda}^*]$ reusing the reduced run draws.

A Perl implementation of this algorithm is available from the author.

12.6 INTERPRETATION OF HIDDEN MARKOV MODELS FOR SURVEILLANCE

In an HMM, the observations arise from a mixture model. The hidden state essentially acts as a switch between the mixture components and also deals with temporal dependence. This begs a couple of questions: Why should the hidden state neatly line up with 'epidemic' and 'nonepidemic' states of the world? What if a HMM with more than two states provides a significantly better fit?

One alternative approach eschews any semantic interpretation of the hidden state and focuses instead on prospective anomaly detection. Consider a K-state HMM and focus on a particular week, say week w of each year. The HMM can provide the posterior distribution of the hidden state in week w for the previous years. Suppose there are n previous years and denote these posterior distributions by $f_1^w, \ldots, f_n^w$. This set of distributions captures the typical historical behavior of week w and might, for example, include an early onset of flu season in addition to normally low levels. The HMM can also provide the posterior distribution of the week-w hidden state for the current year, say f_c^w. If f_c^w differs

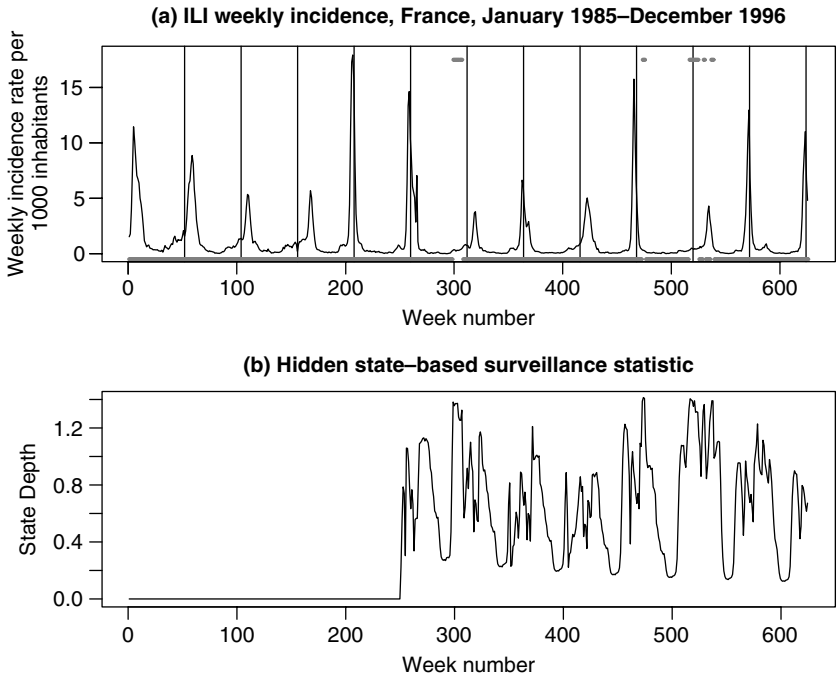


Figure 12.10 The French ILI data. (a) Incidence rates per 1000 inhabitants. (b) Hidden state-based surveillance statistic starting in week 250. The horizontal line segments in (a) correspond to time periods where the surveillance statistic exceeds 1.25.

significantly from $f_1^w, \ldots, f_n^w$, then there is evidence that the current week is anomalous.

Figure 12.10 shows an analysis using the three-state Gaussian model. Figure 12.10(b) shows a surveillance statistic that simply computes the average Euclidean distance between f_c^w and $\{f_1^w, \ldots, f_n^w\}$ starting in week 250. This statistic shows somewhat higher levels around week 520. The upper plot has a vertical line at week 520 and at the corresponding week in each of the other years. It seems that the statistic is identifying a late flu season.

12.7 DISCUSSION

This chapter has examined Bayesian hidden Markov models as a data mining method for surveillance. Multivariate surveillance provides the motivation for this work, but our univariate analyses already raise some interesting issues. Surveillance of the hidden state distribution seems promising. A key extension will incorporate a spatial component in the hidden layer of thes model.

ACKNOWLEDGMENTS

US National Science Foundation grants support my research. I am grateful to David Talaga for the discussions that lead to the random observation time HMM work and to an anonymous reviewer for helpful comments.

13

Advanced Modeling for Surveillance: Clustering of Relative Risk Changes

Andrew B. Lawson

13.1 INTRODUCTION

Syndromic surveillance of disease spread has, as a fundamental component, the assessment of the spatial association of incident cases. Not only is there a need to be able to assess whether a single map of disease for a particular time interval displays 'unusual aggregations' of disease, it is also important to be able to assess whether these aggregations have a spatially distinct pattern which is in itself unusual. These effects can be assessed by the use of statistical techniques.

13.2 CLUSTER CONCEPTS

The definition of a cluster or unusual aggregation of incidence can be wide ranging. Knox (1989) identifies a cluster as '*a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance*'. This is a definition which relies on statistical significance to assess clustering without recourse to restrictive requirements about shape or extent of the cluster to be found. If specific shapes are to be expected, and only these are to be detected, then a different, more restrictive definition would arise. This might be the case where an infectious agent displayed a characteristic spread pattern and it was to be distinguished from other types of unusual aggregations. However, without prior knowledge of such forms, it would be

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

important to be able to detect *any* unusual aggregations, whatever their form. This would generally be true in public health surveillance in the bioterrorism context where there is a strong chance that novel insults are introduced whose spatial distribution has not been examined before.

One disadvantage of the hypothesis testing approaches to cluster detection is their need to define alternative hypotheses that restrict cluster form. For example, SaTScan (http://www.satscan.org) tends to detect circular clusters as it defines the scanning window as a circular region. Equally the Besag and Newell test defines essentially circular zones for testing (see Chapter 4 of this volume for a review). A narrow definition of cluster form can lead to a lack of sensitivity in this approach. For the purposes of surveillance in real time there is a need for cluster detection methods to be able to detect unusual aggregations of disease quickly, without strong restrictions placed on the form of the cluster to be detected. A model-based approach should allow this flexibility.

13.3 CLUSTER MODELING

Usually we define two forms of data available for cluster studies: case event data and count data. These data types correspond to individual-level and aggregate-level analyses, respectively. Case event data usually consist of residential addresses of incident cases of a disease, whereas count data are usually collected within predefined small areas (regions such as city blocks, census tracts, or zip codes) and hence are spatial aggregations of the case event data. For conventional spatial cluster analysis the usual aim is to analyze a disease incidence map for 'unusual' aggregations. To achieve this aim there are a number of approaches which can be employed. Case event data will be considered initially, and then counts.

13.3.1 Spatial Modeling of Case Event Data

Case event data can usually be considered to form a point process, albeit one which arises within a heterogeneous population. Models assumed for such data are often heterogeneous Poisson process models. In these models the local intensity (density) of cases at a spatial location $\mathbf{s}$ is governed by a first-order intensity function $\lambda(\mathbf{s})$. This function can describe the localized variation in disease risk as well as long-range trend in risk. Usually the population at risk of the disease locally must be included in the model. This is reasonable as it is to be expected that higher concentrations of susceptible population will lead to higher incidence rates. This effect is often captured by the definition:

$$\lambda(\mathbf{s}) = \lambda_0(\mathbf{s})\theta(\mathbf{s})$$

where $\lambda_0(\mathbf{s})$ is a function of the 'at risk' population and $\theta(\mathbf{s})$ is a relative risk function measuring the excess risk experienced locally. Often the aim is to

model $\theta(\mathbf{s})$. The function $\lambda_0(\mathbf{s})$ is a nuisance function and can be conditioned out or estimated nonparametrically (Lawson, 2001, pp. 42–44). If the spatial distribution of cases of a control disease is available then it is possible to condition out the $\lambda_0(\mathbf{s})$ function from the analysis. A control disease may be any disease which is matched to the age–sex risk structure of the case disease but unaffected by the excess risk agent of interest. For example, in the bioterrorism attack situation, an anthrax insult may affect respiratory or dermal symptoms but may not affect intestinal symptoms. Hence the spatial distribution of short-latency intestinal disease might be a suitable control. The matching of control diseases can be very difficult and their use is controversial. Resort can be made to estimation of $\lambda_0(\mathbf{s})$ based on the known at-risk population distribution of the local area (Lawson and Williams, 1994).

Modeling $\theta(\mathbf{s})$ can be considered in a variety of ways. First, if a nonparametric estimate of $\theta(\mathbf{s})$ is required, then it is possible to form the ratio

$$\widehat{\theta}(\mathbf{s}) = \frac{\widehat{\lambda}(\mathbf{s})}{\widehat{\lambda}_0(\mathbf{s})}, \tag{13.1}$$

where $\widehat{\lambda}(\mathbf{s})$ is estimated from the case distribution (usually via density estimation) and $\widehat{\lambda}_0(\mathbf{s})$ is estimated from the control distribution (via density estimation). If only expected rates are available for the study region then $\widehat{\lambda}_0(\mathbf{s})$ may be obtained by nonparametric regression. Areas of excess risk on the resulting map of $\widehat{\theta}(\mathbf{s})$ can be assessed for their significance using a Monte Carlo procedure which leads to a *p*-value surface (Lawson, 2001, p. 67). Areas of risk which may be deemed 'significant' will appear above, say, the 0.05 contour level on the *p*-value surface map. Figures 13.1 and 13.2 display the case event and control disease distributions for larynx (case) and lung cancer (control) in a study of an environmental hazard. Figure 13.3 displays the density ratio (extraction map) of the cases to the control densities for the spatio-temporal frame. The areas of elevated risk (*clustering*) are clearly displayed in the southeast of the map. A *p*-value surface could also be constructed and areas assessed for significance.

While extraction mapping can be instructive when exploring spatial clustering, it has limitations. First, areas of elevated risk must be tested for significance and the use of *p*-value surfaces has not so far been evaluated in terms of power for the detection of clusters of different kinds. The main concern about these methods is the fact that the density estimates used for the smoothing of the case or control disease are controlled by smoothing parameters, and these can be varied to yield different results. Lawson and Williams (1994) noted that the use of smoothed expected rates yields different inferences compared to control disease estimates. Optimal smoothing can be achieved by cross-validation but it is not clear whether these optimal results are appropriate for spatial disease distributions. For further information on these methods the reader is referred to Lawson (2001 Chapter 5), and Kelsall and Diggle (1995), and for the inclusion

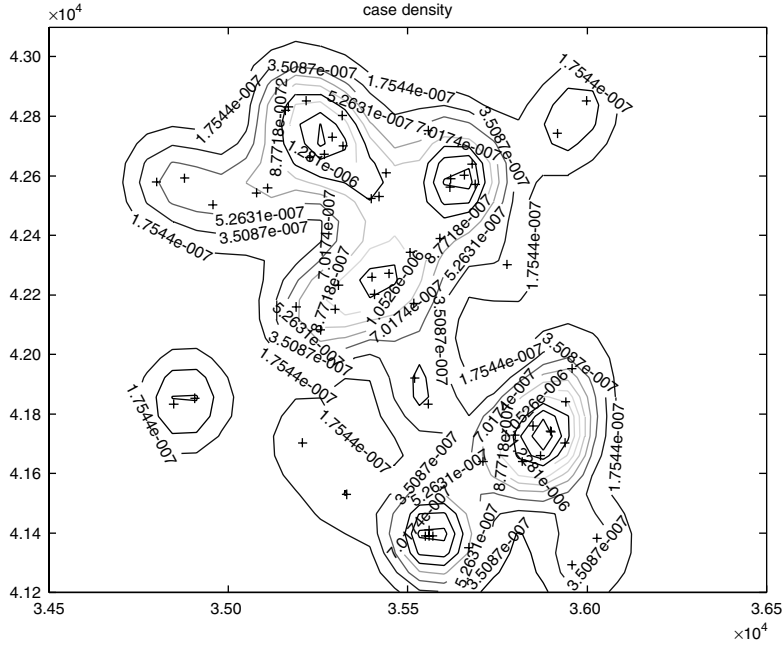


Figure 13.1 Larynx cancer case event distribution for a fixed time and spatial frame. The density estimate contour is superimposed.

of covariates to Kelsall and Diggle (1998). In general, the use of nonparametric clustering methods is limited to exploratory examination and is prone to sensitivity to the smoothing method used.

The types of clusters found using these methods are known as *hotspot* clusters and are the type that are likely to be of interest in a biosurveillance context. The term *hotspot* arises from the fact that these clusters are arbitrarily defined areas of significant excess risk.

To move beyond the arbitrary smoothing and restrictive null and alternative hypotheses imposed by many hypothesis tests, a modeling approach is often useful. A modeling approach to clustering is useful when a variety of cluster forms are possible and also the inclusion of covariates may be required.

There are a range of possible approaches to this type of modeling. One possibility is to consider a hidden process of cluster centers which describe the clustering tendency. This arises from considering mixture models for risk.

13.3.1.1 Hidden mixtures

There are different approaches to mixture modeling in this area. One approach involves the idea of marginal mixtures where the local intensity of cases consists of weighted components. For example, Fernández and Green (2002) proposed

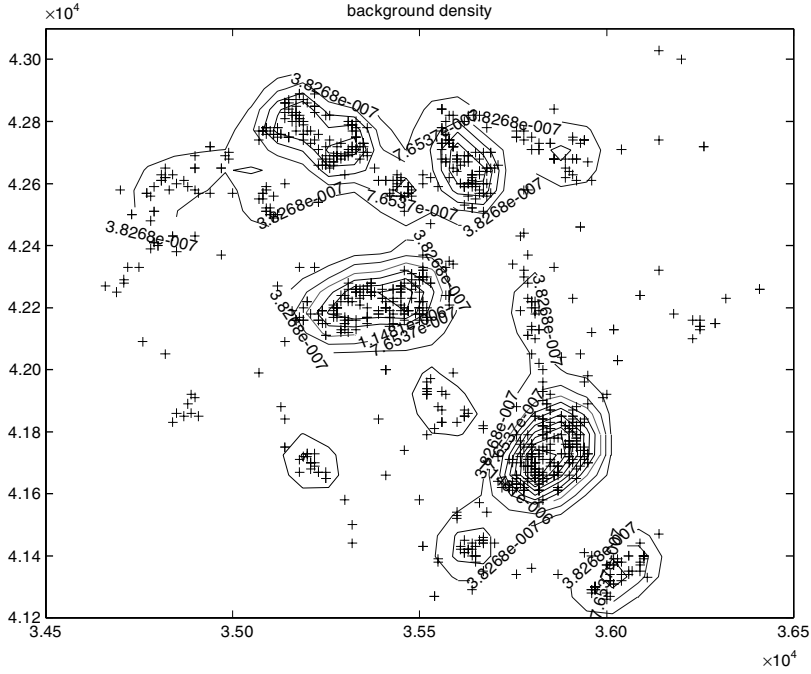


Figure 13.2 Lung cancer control distribution for the same space-time period as Figure 13.1. Density estimate contours superimposed.

a model where the local risk is defined marginally. In essence, this form of modeling ignores the spatial structure of the risk in the mixture. A logical extension of the marginal mixture into spatial problems is the use spatial mixtures where the locations of centers (mixture component locations) is spatially defined (Lawson 1996, 1995). The hidden center process must be estimated. In general, these methods rely on the use of reversible jump Markov chain Monte Carlo (MCMC) for implementation. Unfortunately, the computational complexity and degree of tuning required by these methods is likely to prohibit their extensive use within a surveillance context where near-real-time modeling is required.

13.3.1.2 Mixed effect models

It is possible to use mixed effect models where clustering is not specifically modeled but is determined from the residual process. For example, for a case event point process $\{\mathbf{s}_i, i = 1, \ldots, m\}$ observed within an area T modeled by a heterogeneous Poisson process with first-order intensity $\lambda(\mathbf{s})$, the log-likelihood conditional on m can be defined with a log-linear link to a linear predictor

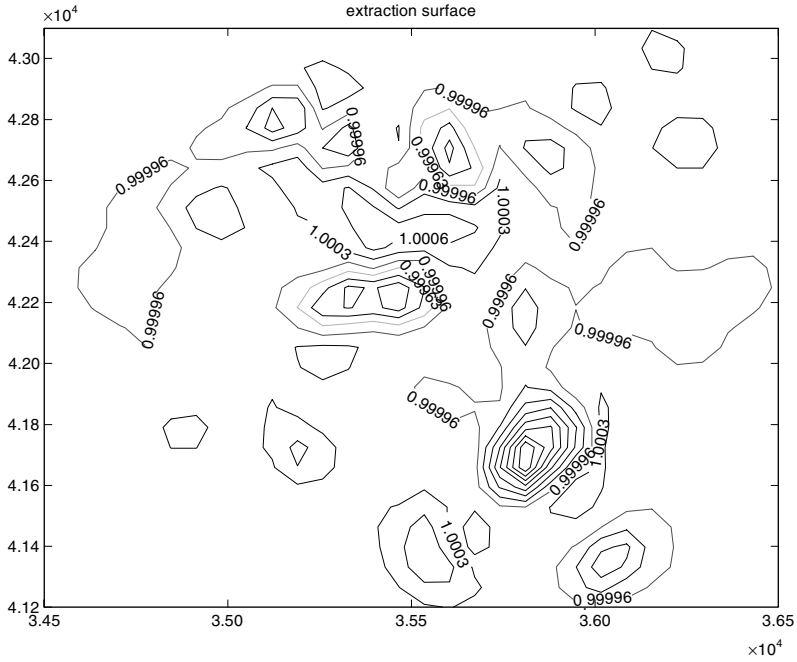


Figure 13.3 Density ratio map (extraction map) of larynx to lung cancer shown in Figures 13.1 and 13.2. Contour levels close to 1.0 represent null excess risk.

$\eta_i = \mathbf{x}_i^{\mathrm{T}}\beta$, where $\mathbf{x}_i^{\mathrm{T}}$ is the ith row of the $m \times p$ matrix of covariates and β is a $p \times 1$ vector of parameters, as

$$\lambda(\mathbf{s}_i) = \lambda_0(\mathbf{s}_i)\exp\{\eta_i\}, \tag{13.2}$$

$$l(\underline{s};\beta) = \sum_{i=1}^{m}\log\lambda(\mathbf{s}_i) - m\log\Lambda$$

where $\Lambda = \int_T \lambda(\mathbf{u})d\mathbf{u}$. The $\boldsymbol{\beta}$ parameters can be estimated by maximum likelihood and use of standard statistical packages is possible (Berman and Turner, 1992; Lawson, 1992; Baddeley and Turner, 2000). The estimated surface of intensity $\widehat{\lambda}(\mathbf{s}_i) = \widehat{\lambda}_0(\mathbf{s}_i)\exp\{\widehat{\eta}_i\}$ depends on the the $\boldsymbol{\beta}$ estimates as well as on the background estimate $\widehat{\lambda}_0(\mathbf{s}_i)$. This background can be estimated from a control disease or other population 'at risk' surrogate. Estimates of $\boldsymbol{\beta}$ will be sensitive to the estimation of $\lambda_0(\mathbf{s}_i)$. An alternative conditional likelihood can be derived which avoids this problem (Diggle and Rowlingson, 1994). Residuals can be defined for this point process (Lawson, 1993) whereby the saturated estimate of $\lambda(\mathbf{s}_i)$ is compared to the model-based estimate. Define this residual as $r(\mathbf{s}_i) = \widehat{\lambda}_s(\mathbf{s}_i) - \widehat{\lambda}(\mathbf{s}_i)$. Areas of significant excess risk could be isolated from the $r(\mathbf{s}_i)$ surface, assuming that the covariates in x are spatial. A Monte Carlo p-value

surface could be computed using a variety of possible simulation approaches. One simple approach is to generate 99 sets of pseudo case event data simulated under a parametric bootstrap from the estimated model with $\widehat{\lambda}(\mathbf{s}_i)$. A set of 99 saturated estimate surfaces $(\widehat{\lambda}^*_s(\mathbf{s}_i))$ can yield a set of 99 pseudo residual surfaces, $r^*(\mathbf{s}_i)$, and then the pointwise Monte Carlo ranking of $r(\mathbf{s}_i)$ among the $\{r^*(\mathbf{s}_i)\}$ will yield local probability estimates of excess risk behavior. This approach would allow the inclusion of (spatial) covariates within the formulation. However any correlation between the covariate and risk excess will be unidentified.

The above approach can be taken a step further by the inclusion of random effects in the intensity specification (13.2). In general, it is possible to extend the basic point process model to one with a stochastic intensity function (a Cox process). Recent research has placed emphasis on the assumption that the intensity is a realization of a Gaussian random field (log-Gaussian Cox process: see Møller and Waagepetersen, 2002). This type of model leads to extra (correlated) variation in the local intensity of cases. In terms of clustering this usually implies that there is an overall clustering tendency for the data and locally there are areas of similar clustering of cases. If the correlation of the random field is strong then there will be large areas of like intensity. However, if the correlation is weak then a relatively variable intensity will arise. In the limit, uncorrelated heterogeneity would remain. In an approximate Bayesian setting, consider the likelihood for a heterogeneous Poisson process where the intensity $\lambda(\mathbf{s}_i)$ has a prior distribution which is multivariate normal. Any realization of the $\lambda(\mathbf{s}_i)$ will be multivariate normally distributed for a spatial Gaussian process (Ripley, 1988). Hence the hierarchy can approximately be viewed as

$$\mathbf{s}_i \sim PP(\lambda(\mathbf{s}_i)), \tag{13.3}$$

$$\lambda(\mathbf{s}_i) \sim MVN(\mu(\mathbf{s}_i), \mathbf{C}), \tag{13.4}$$

where $\mathbf{C}$ is a covariance matrix that controls the degree of spatial correlation in the field. The elements of $\mathbf{C}$ could be parametrically specified by $c(\lambda(\mathbf{s}_i), \lambda(\mathbf{s}_j)) \equiv c_{ij} = \tau \exp(-\alpha d_{ij})$. Here, d_{ij} is the distance between the ijth sites, τ describes the variance at zero distance and α describes the strength of correlation at d_{ij} distance. If no distance effect exists then only an uncorrelated heterogeneity term would be included. Estimation in this type of model will lead to the recovery of a smooth surface of risk which depends for its smoothness on the α parameter. An issue arises here as to where the clustering is to be measured. If clustering is modeled then clustering effects will appear in the model and any model residual (for a well-fitting model) should not contain any cluster artifacts. On the other hand, a parsimonious model for the nonclustering component could be fitted and the residual examined for excess risk. Inclusion of a global spatial correlation component (as defined in (13.4) above) in such models may lead to complications as the clustering behavior may be absorbed by the correlation parameter within the fitted model. In addition, the smoothing used could reduce the ability

to observe clusters. In short, global smoothing, via the use of correlation prior distributions, could lead to smearing of risk and ultimately removal of clusters. This issue is discussed further in Section 13.4.

Alternative models for clustering focus on the cluster form itself and their locations. These can be local models or can be global in that parameters control the overall form of the clustering field. Hidden process models (Lawson and Clark, 1999; Lawson and Denison, 2002, Chapters 4, 5, 14) seek to estimate a process of cluster centers underlying the case events. This process has unknown locations and number of clusters. In these models the random effects are replaced by random objects (centers). Computational complexity of fitting these models (e.g. the need for specially tuned MCMC) and others (Gangnon and Clayton, 2000; Green and Richardson, 2002) tend to limit their usefulness especially when surveillance focuses on the speed of analysis. Instead a more computationally efficient proposal has been made where a cluster spread parameter is allowed to vary spatially across a map. This parameter controls the local size of clusters. It is assumed to have a spatial correlation prior within a hierarchical Bayesian model (Hossain and Lawson, 2005).

13.3.2 Spatial Modeling of Count Data

The spatial modeling of count data follows in parallel with that of case event data. (See Chapters 4 and 5 in this volume for discussion of these models.) Counts arise from aggregation of case events and so it is natural to consider a Poisson distribution for the counts in small areas. Often the assumption is made that

$$y_i \sim \text{Pois}(e_i\theta_i),$$

where e_i is an expected count within the small area and θ_i is the relative risk within the same area. Here θ_i is modeled as for $\theta(\mathbf{s})$ in the case event situation. Often a log-linear form is assumed for $\log\ \theta_i$, that is, $\log\theta_i = \eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, where $\mathbf{x}_i^{\mathrm{T}}$ is the ith row of the $m \times p$ matrix of covariates and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. The analysis of excess relative risk can proceed with the inclusion of relevant covariates or confounders (such as deprivation indices). Often a summary measure is computed from the observed count and expected count in each small area as a crude estimate of relative risk: the standardized incidence ratio, $\widehat{\theta}_i = y_i/e_i$. This measure can be mapped for the purposes of exploratory assessment of excess risk (as for extraction mapping in the case event situation (13.1)). The instability of this ratio estimate is well known and use of the standardized incidence ratio is limited by this feature.

Extensive development of Bayesian models for relative risk has been witnessed in the last 10 years. This has mainly taken the form of log-linear modeling with both spatial and nonspatial (uncorrelated) random effects. That is,

$\log \theta_i = \eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + v_i + u_i$, where v_i and u_i are separate random effects which have prior distributions. These are examples of generalized linear mixed models (GLMMs). For these models full Bayesian analysis with posterior sampling is now commonplace (Besag *et al.*, 1991) due to the availability of software (WinBUGS). A variant of this approach has been proposed where a full spatial Gaussian process prior (with parametric covariance function specification) was proposed (see Diggle *et al.*, 1998; Wikle, 2002). The term *clustering* has been applied to the correlated term in this formulation (usually defined to be u_i). This is unfortunate as the resulting smooth surface, while correlated, has removed much of the clustering evidence. On the other hand, if a parsimonious GLMM were chosen without correlated heterogeneity and the residual from the fitted model examined for excess risk then this could be useful and computationally efficient approach in a surveillance context. Thus a parametric bootstrap could be used to assess clustering. The stages of this approach would be as follows:

(1) Fit a GLMM model as

$$y_i \sim \text{Pois}(e_i\theta_i), \qquad \text{with} \log \theta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + v_i \text{ and } v_i \sim N(0, \tau_v),$$

where v_i is uncorrelated heterogeneity and τ_v is the variance of the v effect, and any confounders (such as deprivation) are included within $\mathbf{x}_i^{\mathrm{T}}$.

(2) Compute

$$\widehat{r}_i = y_i - \widehat{y}_i = y_i - e_i \exp\{\mathbf{x}_i^{\mathrm{T}}\widehat{\beta} + \widehat{v}_i\}.$$

(3) Generate 99 sets of synthetic data $\{y_{ij}^*\}, j = 1, \ldots, 99$, from the fitted model (i.e. simulate from a Poisson distribution with mean $e_i \exp\{\mathbf{x}_i^{\mathrm{T}}\widehat{\beta} + \widehat{v}_i\}$).

(4) Compute $\widehat{r}_{ij} = y_{ij}^* - \widehat{y}_i$.

(5) Rank the $\widehat{r}_i$ amongst the $\widehat{r}_{ij}$. Denote the rank by R_i.

(6) Compute the p-value for the residual as $p_i = 1 - R_i/100$.

(7) Areas of the map with $p_i < C$ would be regarded as significant excess risk. C can be chosen at an appropriate critical level such as 0.05 or 0.01.

The interpretation in step 7 depends on a good model fit and it would be assumed that the normal variation in risk is accounted for in the fitted linear predictor ($\exp\{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{v}_i\}$). An alternative approach, in a Bayesian model, would be to use the predictive distribution to provide a probability statement about the observed data.

13.3.3 Spatio-Temporal Modeling of Case and Count Data

Spatio-temporal (ST) modeling is of fundamental importance in a surveillance context. Usually surveillance has an implicit temporal component and sequential estimation is a key concept. The question 'What is strange about recent

events?' implies that we are looking for a temporal change and need to evaluate whether it is significant. However, little work has been developed in the area of spatio-temporal modeling within the surveillance context.

Most work on surveillance methodology has evolved in temporal applications. These are reviewed by Farrington and Andrews (2004) and in Chapters 2 and 3 of this volume.

For case events a recent proposal has been made by Diggle *et al.* (2004) where an ST point process model is employed with an estimated population at risk defined from a period prior to the infection. In this case the data are incident cases of GP-reported disease and the focus is on the probability of a new case at location $\mathbf{s}$ and time t. The overall intensity in space-time is

$$\lambda(\mathbf{s}, t) = \lambda_0(\mathbf{s}, t)\rho(\mathbf{s}, t).$$

As described above, the excess risk is characterized by

$$\log \rho(\mathbf{s}, t) = d(\mathbf{s}, t)'\boldsymbol{\beta} + W(\mathbf{s}, t).$$

In this case the $d(\mathbf{s}, t)'\boldsymbol{\beta}$ term is a linear predictor containing the usual explanatory risk factor covariates, and $W(\mathbf{s}, t)$ represents anomalous variation. This is a random component which varies in space and time. They assume a log-Gaussian Cox process with covariate extension. The correlations in space and time are assumed to be exponential and separable. Essentially the model fitting is akin to that used for spatial modeling where a random effect (field) is fitted for current data and a p-value surface computed for current data. Posterior probability of excess risk is computed via Monte Carlo. The approach described uses historical data to estimate an (assumed to be) time-constant spatial background, $\widehat{\lambda}_0(\mathbf{s}, t) = \widehat{\lambda}_0(\mathbf{s})$. This may be a gross assumption given that epidemics could be at different stages at different times in the surveillance exercise. In addition, the spatial correlation in the $W(\mathbf{s}, t)$ could be smoothed out by inappropriate choice of correlation parameters. No attempt is made to assess the ability of this method to correctly signal new alarms in the data.

In the same paper, a count data model is proposed which assumes a binomial count distribution for cases (y_{it}) out of a population (n_{it}) in space-time units, that is,

$$y_{it} \sim bin(n_{it}, \pi_{it}), \qquad i = 1, \ldots, m, t = 1, \ldots, T,$$

and the probability of a case has a logit link to a linear predictor composed of fixed covariate effects and random effects representing time, season, and spatial components: $\eta_{it} = \log \mathrm{it}(\pi_{it}) = T_t + D_t + U_i$. Here T_t is a temporal effect, D_t is a 7-day seasonal effect and U_i is a spatial effect. Suitable prior distributions

are assumed for the effects. An autoregressive term based on the number of previous cases is also considered. The model is fitted to the series of events in space-time. Although a natural extension of the Poisson process in space-time, a Poisson distribution with expected counts precalculated was not assumed for the model. No attempt is made to allow for intermediate clustering in time and no quantification of alarm behavior is made. Both the above models allow for spatial and temporal dependence and are extensions of the spatial models described above. They include smoothing terms in both time and space and assume that estimated relative risk under the model will yield evidence for unusual behavior. To what degree this reduces the chance of finding clusters or different change points is unclear. An alternative cluster modeling approach has been proposed in space-time by Clark and Lawson (2002), where clusters are identified within a hidden cluster process. Although a local approach, it required complex computational procedures.

An alternative approach in space-time is to consider, at each time point, what should be expected in the spatial unit or location and to consider a residual effect as evidence for clustering or unusual behavior. This has not been examined so far for case events, but has been considered for count data. A GLMM for surveillance data has been recently proposed by Kleinman *et al.* (2004). In that work a basic logistic linear model with covariates and extra variation is fitted and the probability of a someone being a case on any day is estimated ($\widehat{p}_{it}$). This probability is estimated from previous time periods only. The tail probability from a binomial distribution is computed and used to assess the null hypothesis that the historical data model is adequate. The analysis is applied to HMO coverage in Massachusetts for lower respiratory infection syndrome. See also Chapter 5 of this volume.

Some alternative approaches can be conceived for the analysis of clustering in space-time. First, it is often useful to consider Poisson count models with expected rates when small-area data are available for time periods. In this case we have

$$y_{it} \sim \text{Pois}(e_{it}\theta_{it}),$$

and we are interested in how unusual y_{it+1} is. As in the above case, we could fit a model to historical data and then we could compare our data with the model expectation from the historical data. In this case we could examine whether

$$E(y_{it+1}) = e_{it}\widehat{\theta}_{it}.$$

To do this we could examine residuals of various kinds. The simplest residual could be

$$\widehat{r}_{it+1} = y_{it+1} - e_{it}\widehat{\theta}_{it}, \tag{13.5}$$

and this could be tested as before using a parametric bootstrap. Another possibility, which turns out to be better at detecting sharp changepoints, is to compute what are dubbed *surveillance* residuals (Lawson *et al.*, 2004). These residuals utilize a Bayesian posterior predictive distribution to derive predictive values for the current data based on $E(y_{it+1}|\theta_{it})$. These residuals can be computed easily from posterior sample output. In general, there is an issue about whether a model should contain the current data in its estimation or not. For example, we could calculate (13.5) from the current data instead:

$$\widehat{r^*}_{it+1} = y_{it+1} - e_{it}\widehat{\theta}_{it+1}.$$

In general, it is probably beneficial to use historical data unless considerable drift has occurred. However, this also raises the issue of what historical period one should use and to what degree a model should be allowed to follow changes in the data. For example, if you use a shorter historical period then you will track short range changes more, whereas if you extend the period you will get less change in the historical estimates (they will be smoothed out). Short-term tracking will be anticonservative and longer-term the opposite. The question of whether you should absorb changes after they occur must also be faced. For example, if a change point is signalled at day 3, should I include day 3 in the historical data for day 4? If I do, I will have adapted to the new level. Is this appropriate?

Another issue is that most models described above include covariates. What happens if my choice of covariates is poor and my model does not fit well, or my model progressively fits worse? Do I make on-line adjustments or do I ignore goodness of fit? These appear to remain open questions. Finally in this section, it is appropriate to raise the issue of variety of detectable effects. Often what may be important in surveillance is the ability to detect a variety of changes as they occur. One might want to detect unusual individual unit changes (as described above), or the location of clusters, or cluster changes in time, or even gradual changes in time. In space-time a variety of features might be of interest. One particular issue is the spread of existing clusters to new areas. Figure 13.4 depicts a simulation of cluster spread that would be difficult to model with conventional random effect models.

It is surmised that locally adaptive methods must be developed to deal with this type of change. An example of the use of directional derivatives in this context is given in Lawson *et al.* (2004) and Clark and Lawson (2005). A different approach to multiple feature detection/monitoring in space-time disease maps is given by Lawson (2004).

In the next section, I examine the issue of multivariate or syndromic surveillance for clustering.

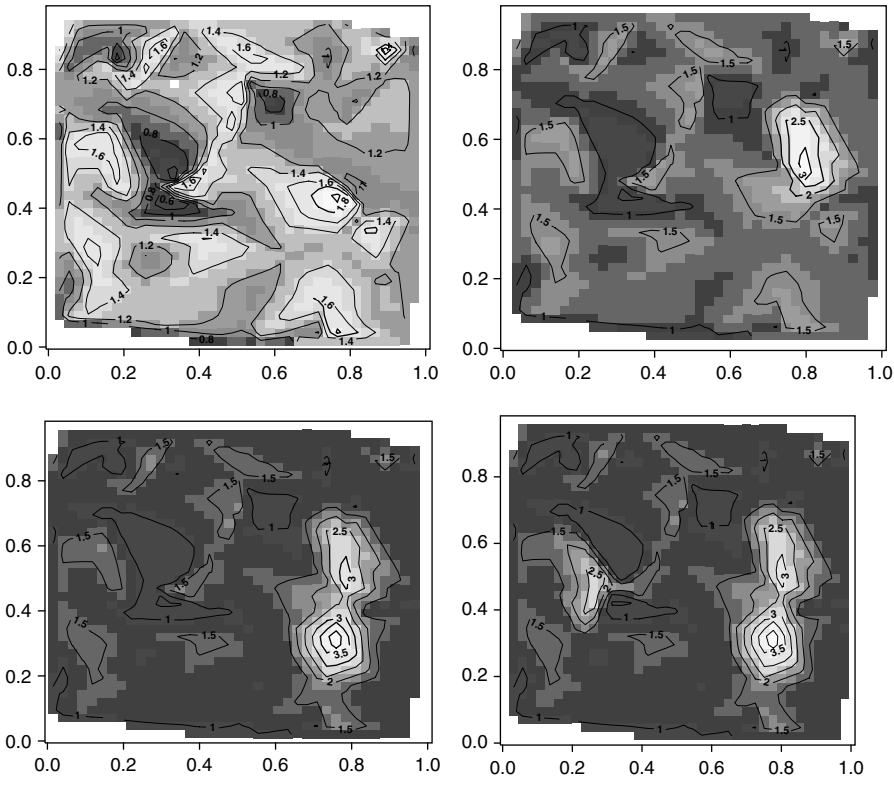


Figure 13.4 Rowwise from top: sequence of four simulated time periods with progressive introduction of elevated relative risk. The first period is simulated with 100 uniform locations and log-normal relative risk. Note the cluster spread between period 2 and 3.

13.4 SYNDROMIC CLUSTER ASSESSMENT

In real public health surveillance situations you may want to make population-based interventions based on information from multiple data sources. These sources are often linked, and for interventions to be efficient there should be an appreciation of the linkages and their meaning. A simple example would be the evidence of raised numbers of visits to emergency rooms for lower respiratory tract infections in old people, and also raised incidence of dermal conditions in old people reported to general practitioners. Both these pieces of evidence support a potential atmospheric insult (which of course could relate to a bioterrorism incident) and may be linked by a common cause. This suggests that we need to consider many streams of data to find associations. Further, if we want to include mapped data we must also consider multiple space-time analyses.

This suggests that we might explore data mining techniques for suitable methods (see Wong *et al.*, 2002; and Chapter 10 in this volume).

Besides multiple disease monitoring (vector monitoring), we might also be concerned with ways to ensure *early* detection of an effect by using ancillary information. This ancillary information could be health-related but not necessarily disease-specific. In fact anything which might suggest that an outbreak of disease is occurring in a population could be used, for example: pharmaceutical sales, job absenteeism, or school absenteeism. Breakdowns of sales would be useful to monitor for specific etiologies. However, indicators can be nonspecific and so the exact etiology may not be estimable at such an early stage. Another set of information that could be useful is data on early symptoms of a disease in the population that could indicate the inception of an outbreak. Often these symptoms are also nonspecific and also not usually recorded by GPs. For example, if GPs recorded patients reporting cough, wheezing, chest infections, and related symptoms then at the population level it might be possible to make early detections of important outbreaks. Syndromic surveillance is discussed more fully in Chapters 1, 2, and 3 in this volume. The multivariate formulation is also given in Chapter 9.

If we assume that clustered data are to be detected in space-time then we can define a Bayesian model formulation that is as follows. Define y_{it} as the current data (counts usually) for the ith monitored site (could be a small area or address) and y_{iT} is the cumulative data on the disease up to and including time t. A parameter vector $\boldsymbol{\theta}$ is defined. Syndromic variables are also available: x_{it} is one such variable and $\mathbf{s}_{it}$ is the vector of syndromic variables.

Define the complete data and ancillary (syndromic) vector as

$$\mathbf{D}_{it} = \begin{cases} y_{it} \\ x_{i1t} \\ x_{i2t} \\ x_{i3t} \\ \dots \end{cases} = \begin{cases} y_{it} \\ \mathbf{s}_{it} \end{cases}$$

An example of a typical syndromic situation is described in Figure 13.5. All reported cases of gastrointestinal disease are of interest, and the syndromic variable is over-the-counter pharmacy sales. This example has been kindly supplied by Victoria Edge of Health Canada.

13.4.1 The Bayesian Posterior Distribution

First of all, it must be considered whether we are interested in the joint behavior of $\mathbf{s}_{iT}$ and y_{it}, or simply interested in y_{it} given we observed values of $\mathbf{s}_{iT}$. If we are simply interested in early detection it may be more natural to adopt the second conditional definition. For example, we might ask what is the probability of a

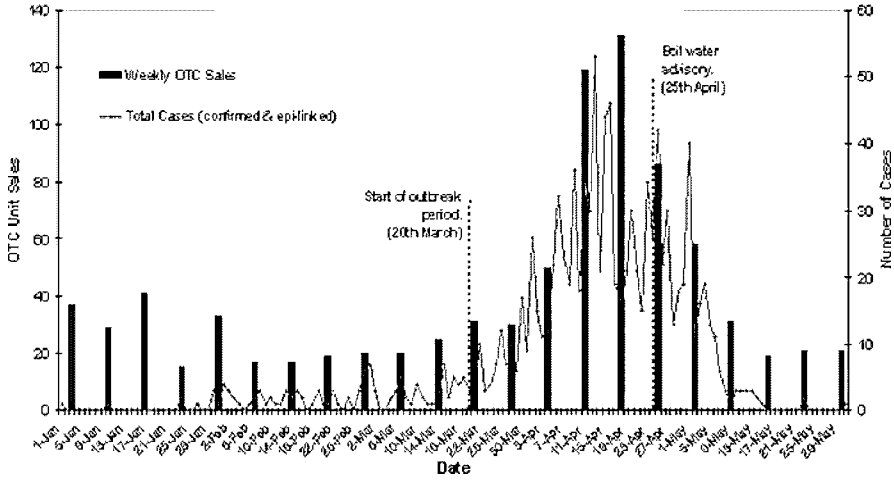


Figure 13.5 A comparison of weekly aggregate unit sales of over-the-counter antidiarrheals and antinauseants with the epidemic curve, from January to May 2001 in The Battlefords, Saskatchewan. The epidemic curve indicates the total number of isolate-confirmed cases and epidemiologically linked cases by reported onset date.

count of anthrax y_{it} given that we saw the vector of syndromic variables $\mathbf{s}_{iT}$. On the other hand, let $\mathbf{s}_{it}$ be the vector of other diseases. If we are interested in monitoring multiple diseases, then we might want to ask what is the probability that disease 1 has count y_{it}, disease 2 has x_{i1t}, and disease 3 has x_{i2t} together.

13.4.1.1 Conditioning on $\mathbf{s}_{iT}$

In a Bayesian model conditioning on $\mathbf{s}_{iT}$, the posterior distribution can be identified as

$$P(\boldsymbol{\theta}|y_{iT}, \mathbf{s}_{iT}) \propto f(y_{it}|\boldsymbol{\theta}, x_{iT})P(\boldsymbol{\theta}|y_{iT-1}, \mathbf{s}_{iT-1}),$$

where $P(\boldsymbol{\theta}|y_{iT-1}, \mathbf{s}_{iT-1})$ is the posterior distribution up to and including time $T-1$. It is assumed that $\boldsymbol{\theta}$ is not time-dependent nor depends on the spatial configuration. The equivalent (posterior) predictive distribution is given by

$$P(y_{it}|y_{iT-1}) = \int f(y_{it}|\boldsymbol{\theta}, \mathbf{s}_{iT})P(\boldsymbol{\theta}|y_{iT-1}, \mathbf{s}_{iT-1})d\boldsymbol{\theta}.$$

Within an MCMC sampler this can be approximated via

$$\approx \frac{1}{G}\sum_{g=1}^{G} f(y_{it}|\boldsymbol{\theta}_{T-1}^{g}, \mathbf{s}_{iT}),$$

where $\boldsymbol{\theta}_{T-1}^{g}$ is the sampled parameter vector for the gth iteration from the posterior at $T-1$. This is called recursive Bayesian learning. This implies that we can predict the data at the (i,t)th point given $\mathbf{s}_{iT}$ and the value of the average posterior-sampled $\boldsymbol{\theta}$ vector. In the case of cluster modeling we would specify a likelihood model for the space-time behavior of the clusters or clustering and then substitute this for $f(y_{it}|\boldsymbol{\theta},\mathbf{s}_{iT})$. For example, assuming y_{it} is a count in a small area at time t, then we could assume

$$y_{it} \sim \text{Pois}(e_{it}\theta_{it}),$$
$$\log\theta_{it} = \alpha_0 + \nu_i + u_i + \gamma_t + \eta_{it} + f(\mathbf{s}_{iT})$$

where $\nu_i + u_i + \gamma_t + \eta_{it}$ are spatial, temporal and space-time component random effects and $f(\mathbf{s}_{iT})$ is a function of the history and current values of the syndromic variables. We could proceed by testing for jumps in the variances of the random effects (as in Lawson, 2004). This will also allow us to assess whether there are changes in the spatial, temporal or localized spatio-temporal components. In addition, we can test whether dependence on syndromic variables is found. In this case, if we assume dependence only on the last estimated value, the MCMC sampler would yield a predictive estimate computed as

$$\approx \frac{1}{G}\sum_{g=1}^{G}\text{Pois}\{e_{it}\boldsymbol{\theta}_{it-1}^{g}(\mathbf{s}_{it-1})\}.$$

13.4.1.2 Unconditional multivariate Version

$\mathbf{D}_t$ is a vector of count data and syndromic variables at time t. We assume that discrete variables are monitored only. The posterior given the evolution up to and including t is

$$P(\boldsymbol{\theta}|\mathbf{D}_T) \propto f(\mathbf{D}_t|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{D}_{T-1}),$$

where $f(\mathbf{D}_t|\boldsymbol{\theta})$ is the new data likelihood which could include correlations between elements (which could be *maps* or *time series*).
The associated predictive distribution is given by

$$P(\mathbf{D}_t|\mathbf{D}_{T-1}) = \int f(\mathbf{D}_t|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{D}_{T-1})d\boldsymbol{\theta},$$

where $P(\boldsymbol{\theta}|\mathbf{D}_{T-1}) = f(\mathbf{D}_{t-1}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{D}_{T-2})$.

13.5 BAYESIAN VERSION OF THE OPTIMAL SURVEILLANCE ALARM FUNCTION

It should also be noted that there is an optimal surveillance methodology that is based on likelihood or posterior ratios (see Chapters 3 and 9 of this volume; Frisén, 2003; Sonesson and Bock, 2003). These methods assume a definition of *optimal* to be fast detection time for an effect of interest.

Define a frequentist alarm function for the current time (s) as

$$P(\boldsymbol{x}_s) = \sum_{k=1}^{s} \pi_k \prod_{u=k}^{s} \frac{f(\boldsymbol{x}(u)|\mu')}{f(\boldsymbol{x}(u)|\mu^0)} \Big/ \sum_{k=1}^{s} \pi_k,$$

where $\boldsymbol{x}_s$ represents data at s and $f(\boldsymbol{x}(u))$ is the likelihood.
Here the function is designed to detect any change (of μ^0 to μ') on the range $k = 1, \ldots, s$. π_k is the probability of a jump at k given there has not been one before. Often for discrete times the geometric distribution is used for π_k. A Bayesian version of this would have

$$P(\boldsymbol{x}_s) = \sum_{k=1}^{s} h(k) \frac{\prod_{u=k}^{s} f(\boldsymbol{x}(u)|\mu') g(\mu'|u)}{\prod_{u=k}^{s} f(\boldsymbol{x}(u)|\mu^0) g(\mu^0|u)} \Big/ \sum_{l=1}^{s} h(l).$$

Here $h(k)$ is the probability of a jump at k, and $g(\mu'|u)$ is the conditional prior distribution of the new μ value given the time u. Note that for an alarm which is simply concerned with the jump at the present time, s (and only then), the alarm function simplifies down to the Bayes factor:

$$BF = \frac{f(\boldsymbol{x}(s)|\mu') g(\mu'|s)}{f(\boldsymbol{x}(s)|\mu^0) g(\mu^0|s)}.$$

Otherwise the alarm function is a weighted product of posteriors for the $s-k+1$ time points with weights $w_k = h(k) / \sum_{l=1}^{s} h(l)$.

13.5.1 Clustering and $f(\boldsymbol{x}(u)|\mu)$

The density $f(\boldsymbol{x}(u)|\mu)$ can be defined in different ways depending on the surveillance task. For clustering the density could be a function such as

$$f(\boldsymbol{x}|\boldsymbol{\theta}) \propto H(\boldsymbol{x}; \boldsymbol{x}_{\delta x}, \boldsymbol{\theta})$$

where $H(\boldsymbol{x}; \cdot, \cdot)$ relates the data $\boldsymbol{x}$ to data in a neighborhood of $\boldsymbol{x} : \delta_x$. Local likelihood models could be assumed where, for example, in the count data case, we assume that locally the likelihood depends on a scale parameter defining a neighborhood around a data location. This scale parameter (δ_x) has a correlated

prior distribution. This leads to an estimate of relative risk within the neighborhood. For the case of a single time point (Bayes factor) alarm we have a scaled jump (α) in relative risk (i.e., $\theta'_{\delta_i t} = \alpha\theta^0_{\delta_i t}$)

$$BF_t = \frac{\prod_i (e_{\delta_i t}\theta'_{\delta_i t}) e^{-(e_{\delta_{it}}\theta'_{\delta_{it}})} g(\theta'|t)}{\prod_i (e_{\delta_i t}\theta^0_{\delta_i t}) e^{-(e_{\delta_{it}}\theta^0_{\delta_{it}})} g(\theta^0|t)}$$
$$= \frac{g(\alpha\theta^0|t)}{g(\theta^0|t)} \prod_i \left(\frac{\theta'_{\delta_i t}}{\theta^0_{\delta_i t}}\right) e^{-\sum e_{\delta_{it}}[\theta'_{\delta_{it}} - \theta^0_{\delta_{it}}]}.$$

Values of δ_{it} would have to be estimated for each region as posterior expected values. The alarm is conditional on these quantities. A more complex alarm is found for longer time scales.

13.5.2 A Simple Real-Time Biohazard Model

One scenario where space-time surveillance may be important is when a highly infectious agent is released within an densely populated area such as a city. Anthrax spread and related possible bioterrorism attacks can lead to the consideration of a spatio-temporal surveillance of health conditions. First of all, it is considered that a mobile recording unit (MRU) travels across the study area and records the conditions at some arbitrary spatio-temporal sample point. The track of the MRU can be arbitrary, and can be related to previous surveillance of areas worst affected by the spread of disease/hazard. Figure 13.6 displays a typical sequence of sample locations.

Around the sample point there will be a detection radius, or more generally a spatial detection function. This function will be similar to those used in distance

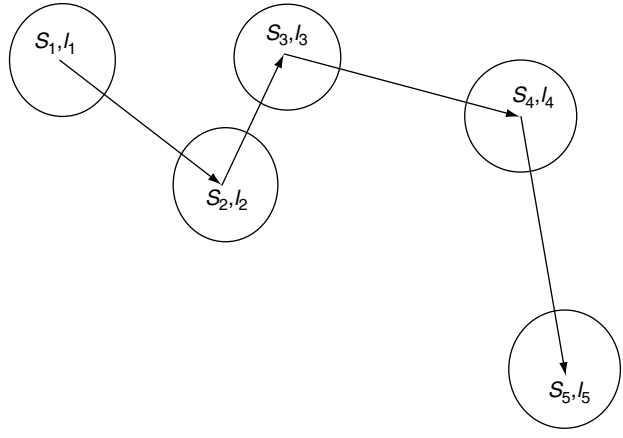


Figure 13.6 Mobile recording unit at five different sampling locations.

sampling (Buckland *et al.*, 2001). This function defines the probability of detecting any cases of disease in a given distance and direction of the sample point. Define the following ingredients: the first-order intensity of the disease infection process is

$$\lambda(\boldsymbol{x}, t) = \rho g(\boldsymbol{x}, t) f(\boldsymbol{x}, t);$$

suitably normalised, this function can be viewed as a probability of becoming infected at space-time location $(\boldsymbol{x}, t)$. The parameter ρ is an overall rate, $g(\boldsymbol{x}, t)$ is the background population intensity which can vary in space and time, and $f(\boldsymbol{x}, t)$ is also a function of space-time but is parameterized with functions of the disease process of interest. These functions could simply be of explanatory covariates or could be of unobserved effects. They could also be functions of the number of already infected individuals existing up to the observation time.

In addition to the these disease modeling features there must also be associated with this intensity a function which determines how easy it will be to detect any cases of disease in the vicinity of the observation point.

This function is referred to as the *detection function*. The joint probability of detection and incidence is

$$d(\mathbf{s}, l, \boldsymbol{x}, t)\lambda(\boldsymbol{x}, t)e^{-\int \lambda\{u,v\}dxdt}.$$

The resulting likelihood, assuming full knowledge of the history of the disease infection process, is given, for any time point p, within area A, by

$$L = \prod_{j=2}^{p} \prod_{k \in (l_{j-1}, l_j)} d(s_j, l_j, x_k, t_k)\lambda(x_k, t_k)e^{-\int_{l_{j-1}}^{l_j} \int_A \lambda\{u,v\}dxdt}.$$

13.5.2.1 Detection functions

A variety of functions could serve as a detection function, depending on the nature of the detection process and the ease with which the disease in question can be detected. For example, it may be that during an attack only visual sightings of cases may be made, and complete ascertainment of the case disease state may be impossible. Hence, a distance effect may come into play in this case.

A radial decline detection probability may be appropriate. These could be, for example, Gaussian in space and exponential in time:

$$d(\mathbf{s}, l, x, t) = \frac{1}{2\pi\kappa\theta}.e^{-\frac{1}{2\kappa}\delta_x^2 - \delta_t/\theta},$$

where $\delta_x = ||\mathbf{s} - \boldsymbol{x}||, \delta_t = l - t, t < l$. This would imply a temporal decay in ability to detect cases. This detection function is separable in space and time and no

interaction is considered. Of course more complex functions could include also interaction between space and time.

Other simpler alternatives could be imagined – for example, uniform probabilities in space and time,

$$d(\mathbf{s}, l, \boldsymbol{x}, t) = \frac{1}{\pi\tau^2}\{I(||\mathbf{s} - \boldsymbol{x}|| < \tau\}\frac{1}{\phi}\{I(||l - t|| < \phi\},$$

where $I(\cdot)$ is the indicator function, $t < l$, τ is a detection radius, and ϕ is a detection span.

There may also be a need to speed up computation by extending the sampling idea expressed above to a situation which was not mobile but. which had a large database which could not be analyzed completely at each time point. In this case a sampling scheme would need to be adopted. In fact, for clustering it may be useful to adopt a adaptive sampling scheme where clustering is analyzed when previous clustering was found in the vicinity. For example, in the case of counts within small areas $\{y_{it}\}$ we could introduce a probability which is dependent on location and time, p_{it} say, that defines how likely an area is to be sampled for analysis. Hence, a model could be developed for cluster analysis where, at observation time t, we would have a posterior distribution $P(\boldsymbol{\theta}_t|y_{it})$. We assume that samples of $\boldsymbol{\theta}_{it-1}$ are available for the regions and write

$$q_i = \frac{p_{it}P(\boldsymbol{\theta}_{t-1}|y_{it})}{\sum_{j=1}^{m} p_{jt}P(\boldsymbol{\theta}_{t-1}|y_{jt})}.$$

Then areas can be drawn with probability q_i. The definition of the prior distribution for p_{it} would be an interesting issue. In the next section I discuss another aspect of shrinking the data analysis problem: computational improvements.

13.6 COMPUTATIONAL ISSUES

Many issues arise with the large volume of potential data that needs to be constantly sifted to allow for optimal surveillance in the sense of coverage of data. Data mining has developed for analysis of large databases, and many of the methods used there can be applied here. Inevitably, there is a need for computational speed-ups, especially if Bayesian methods are used where MCMC is needed.

MCMC sampler speed-ups can be implemented via particle filtration and importance resampling (sequential Monte Carlo), the use of windows (sliding or otherwise), posterior or likelihood approximations and special computational algorithms (e.g. spatial computational speed-ups). In the first case it is possible to apply particle filtration via resampling to clustering problems, and the basic methodology is given in Doucet *et al.* (2001). The use of short time windows will also reduce the computational burden at the expense of long-term effects.

The use of likelihood or posterior approximations would also help, particularly if multivariate normal approximations could be employed, as they could be sampled relatively easily. Finally, special algorithms to speed up spatial computation may be useful (Moore, 1999; Chapter 11 of this volume).

13.7 CONCLUSIONS AND FUTURE DIRECTIONS

In conclusion, there is a wide range of possible models for the inclusion of cluster detection methods in syndromic surveillance systems. In this chapter I have emphasized modeling issues in cluster detection and have not considered hypothesis testing. This is considered in Chapters 7 and 8.

The main issue in the information of the cluster model is how the clusters themselves are represented with the model. At one extreme we could only examine residuals for evidence of clustering and not include any clustering terms in the model. At the other extreme we could explicitly include cluster terms (such as hidden mixture components or cluster variances). If one is little concerned about the form of clusters then residual analysis may suffice. However, if one wants to provide great information about the structure and behavior of the clusters then cluster terms should be used. Note that conventional random effect models are not really *clustering* models but can allow for correlated relative risk. Finally, the need for computational speed-ups within surveillance modeling is going to be important. The use of sampling, filtration, and approximations could all have a part to play in this endeavor.

References

Aerne, L.A., Champ, C.W. and Rigdon, S.E. (1991) Evaluation of control charts under linear trend. *Communications in Statistics. Theory and Methods*, **20**, 3341–3349.

Aitkin, M. (1996a) Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In A. Forcina, G.M. Marchetti, R. Hatzinger, and G. Galmacci (eds), *Statistical Modelling*: Proceedings of the 11th International Workshop, pp. 87–92. Cittá di Castello: Graphos.

Aitkin, M. (1996b) A general maximum likelihood analysis of overdispersion in generalised linear models. *Statistics and Computing*, **6**, 251–262.

Alexander, F.E. and Boyle, P. (eds) (1996) *Methods for Investigating Localized Clustering of Disease*. Lyon: International Agency for Research on Cancer.

Alm, S.E. (1997) On the distribution of the scan statistic of a two dimensional Poisson process. *Advances in Applied Probability*, **29**, 1–16.

Alm, S.E. (1998) Approximations and simulations of the distribution of scan statistics for Poisson processes in higher dimensions. *Extremes*, **1**, 111–126.

Al-Osh, M.A. and Alzaid, A.A. (1987) First-order integer-valued autoregressive (INAR(1)) process. *Journal of Times Series Analysis*, **8**, 261–275.

Alt, F.B. (1985) Multivariate quality control. In N.L. Johnson and S. Kotz (eds) *Encyclopedia of Statistical Science* 6, pp. 110–122. New York: John Wiley & Sons, Inc.

Alt, K.W. and Vach, W. (1991) The reconstruction of 'genetic kinship' in prehistoric burial complexes – problems and statistics. In H.H. Bock and P. Ihm (eds), *Classification, Data Analysis and Knowledge Organization*. Berlin: Springer-Verlag.

Alwan, L.C. (2000) Designing an effective exponential cusum chart without the use of nomographs. *Communications in Statistics*, **29**, 2879–2893.

American Public Health Association (2000) *Control of Communicable Diseases Manual* (J.E. Chin, ed.). Washington, DC: American Public Health Association.

Anderson, N.H. and Titterington, D.M. (1997) Some methods for investigating spatial clustering, with epidemiological applications. *Journal of the Royal Statistical Society, Series* A, **160**, 87–105.

Andersson, E. (2002) Monitoring cyclical processes – a nonparametric approach. *Journal of Applied Statistics*, **29**, 973–990.

Arteaga, C. and Ledolter, J. (1997) Control charts based on order-restricted tests. *Statistics & Probability Letters*, **32**, 1–10.

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman

Aylin, P., Best, N., Bottle, A. and Marshall, C. (2003) Following Shipman: a pilot system for monitoring mortality rates in primary care. *The Lancet*, **362**, 485–491.

Baddeley, A. and Turner, R. (2000) Practical maximum pseudolikelihood for spatial point patterns. *Australia and New Zealand Journal of Statistics*, **42**, 283–322.

Banerjee, S., Wall, M.M. and Carlin, B.P. (2003) Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, **4**, 123–142.

Barbujani, G. (1987) A review of statistical methods for continuous monitoring of malformation frequencies. *European Journal of Epidemiology*, **3**, 67–77.

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

Basseville, M. and Nikiforov, I. (1993) *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ: Prentice Hall.

Bate, A., Lindquist, M., Edwards, I.R., Olsson, S., Orre, R., Lansner, A. and Freitas, R.M.D. (1998) A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, **54**, 315–321.

Bay, S.D. and Pazzani, M.J. (1999) Detecting change in categorical data: mining contrast sets. In S. Chaudhuri and D. Madigan (eds), *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 302–306. New York: Association for Computing Machinery.

Beato Filho, C.C., Assuno, R.M., Silva, B.F., Marinho, F.C., Reis, I.A. and Almeida M.C. (2001) Homicide clusters and drug traffic in Belo Horizonte, Minas Gerais, Brazil from 1995 to 1999. *Cadernos de Saúde Pública*, **17**, 1163–1171. Available at http://www.scielo.br/pdf/csp/v17n5/6324.pdf

Beibel, M. (2000) A note on sequential detection with exponential penalty for the delay. *Annals of Statistics*, **28**, 1696–1701.

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

Bell, C., Gordon, L. and Pollak, M. (1994) An efficient nonparametric detection scheme and its application to surveillance of a Bernoulli process with unknown baseline. In E. Carlstein, H.-G. Müller and D. Siegmund (eds), *Change-Point Problems* IMS Lecture Notes – Monograph Series, vol. 23, pp. 7–27. Hayward, CA: Institute of Mathematical Statistics.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Berke, O. and Grosse Beilage, E. (2003) Spatial relative risk mapping of pseudorabies-seropositive pig herds in an animal-dense region. *Journal of Veterinary Medicine, Series B*, **50**, 322–325.

Berke, O., von Keyserlingk, M., Broll, S. and Kreienbrock, L. (2002) On the distribution of *Echinococcus multilocularis* in red foxes in Lower Saxony: identification of a high risk area by spatial epidemiological cluster analysis. *Berliner und Münchener Tierarztliche Wochenschrift*, **115**, 428–434.

Berman, M. and Turner, T.R. (1992) Approximating point process likelihoods with GLIM. *Applied Statistics*, **41**, 31–38.

Bernardinelli, L., Clayton, D.G., Pascutto, C., Montomoli, C., Ghislandi, M. and Songini, M. (1995) Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, **14**, 2433–2443.

Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory*. Chichester: John Wiley & Sons, Ltd.

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

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Bhattacharya, R. and Waymire, E. (1990) *Stochastic Processes with Applications*. New York: John Wiley and Sons, Inc.

Birnbaum, D. (1984) Analysis of hospital infection surveillance data. *Infection Control*, **5**, 332–338.

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

Bithell, J.F. (1995) The choice of test for detecting raised disease risk near a point source. *Statistics in Medicine*, **14**, 2309–2322.

Bock, R.D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika*, **46**, 443–459.

Bodden, K.M. and Rigdon, S.E. (1999) A program for approximating the in-control ARL for the MEWMA chart. *Journal of Quality Technology*, **31**, 120–123.

Böhning, D., Dietz, E. and Schlattmann, P. (2000) Space-time mixture modelling of public health data. *Statistics in Medicine*, **19**, 2333–2344.

Bonetti, M and Pagano, M. (2004a) The interpoint distance distribution as a descriptor of point patterns, with an application to cluster detection. *Statistics in Medicine*. In press.

Bonetti, M and Pagano, M. (2004b) Parametric estimation of interpoint distance distributions, with an application to biosurveillance data. Unpublished.

Bonferroni, C.E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Borror, C.M., Champ, C.W. and Rigdon, S.E. (1998) Poisson EWMA control charts. *Journal of Quality Technology*, **30**, 352–361.

Borror, C.M., Keats, J.B. and Montgomery, D.C. (2003) Robustness of the time between events CUSUM. *International Journal of Production Research*, **41**, 3435–3444.

Boscoe, F.P., McLaughlin, C., Schymura, M.J. and Kielb, C.L. (2003) Visualization of the spatial scan statistic using nested circles. *Health and Place*, **9**, 273–277.

Box, G.E.P. and Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Boyle, E., Johnson, H., Kelly, A. and McDonnell, R. (2004) Congenital anomalies and proximity to landfill sites. *Irish Medical Journal*, **97**, 16–18.

Breslow, N. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Breslow, N. and Day, N. (1987) *Statistical Methods in Cancer Research, Volume 2: The Design and Analysis of Cohort Studies*. Lyon: International Agency for Research on Cancer.

Brookmeier, R. and Stroup, D. (eds) (2004) *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health*. Oxford: Oxford University Press.

Brookmeyer, R. and Gail, M.H. (1994) *AIDS Epidemiology: A Quantitative Approach*. New York: Oxford University Press.

Brown, S.M., Benneyan, J.C., Theobald, D.A., Sands, K., Hahn, M.T., Potter-Bynoe, G.A., Stelling, J.M., O'Brien, T.F. and Goldmann, D.A. (2002) Binary cumulative sums and moving averages in nosocomial infection cluster detection. *Emerging Infectious Diseases*, **8**, 1426–1432.

Bryant, G. and Monk, P. (2001) *Final Report of the Investigations into the North Leicestershire Cluster of Variant Creutzfeldt–Jakob Disease*. Leicester: Leicestershire NHS Health Authority.

Buehler, J.W. (1998) Surveillance. In K.J. Rothman and S. Greenland (eds), *Modern Epidemiology*, pp. 435–457. Philadelphia: Lippincott Williams and Wilkins Publishers.

Buehler, J.W., Berkelman, R., Hartley, D. and Peters, C. (2003) Syndromic surveillance and bioterrorism-related epidemics. *Emerging Infectious Diseases*, **9**, 1197–1204.

Buntinx, F., Geys, H., Lousbergh, D., Broeders, G., Cloes, E., Dhollander, D., Op De Beeck, L., Vanden Brande, J., Van Waes, A. and Molenberghs, G. (2003) Geographical differences in cancer incidence in the Belgian province of Limburg. *European Journal of Cancer*, **39**, 2058–2072.

Burkom, H.S. (2003) Biosurveillance applying scan statistics with multiple, disparate data sources. *Journal of Urban Health*, **80**, i57–i65.

Cardinal, M., Roy, R. and Lambert, J. (1999). On the application of integer-valued time series models for the analysis of disease incidence. *Statistics in Medicine*, **18**, 2025–2039.

Carlin, B.P. and Banerjee, S. (2003) Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In J.M. Bernardo, M.J. Bayarri, A.P. Dawid, J.O. Berger, D. Heckerman, A.F.M. Smith, and M. West (eds), *Bayesian Statistics*, **7**. Oxford: Oxford University Press.

Carlin, B.P. and Louis, T.A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Carstairs, V. (1981) Small area analysis and health service research. *Community Medicine*, **3**, 131–139.

Celeux, G., Forbes, F., Robert, C. and Titterington, D. (2003) Deviance information criteria for missing data models. Cahiers de mathématiqués de Ceremade 0325, INRIA Rhône-Alpes.

Centers for Disease Control and Prevention (1988) Update: Graphic method for presentation of notifiable disease data – United States 1990. *Morbidity and Mortality Weekly Report*, **40**, 124–125.

Centers for Disease Control and Prevention (2001) Updated guidelines for evaluating public health surveillance systems. *Morbidity and Mortality Weekly Report*, **50**, 1–35.

Centers for Disease Control and Prevention (2004) Emergency preparedness & response: Biological agents/diseases. http://www.bt.cdc.gov/Agent/agentlist-category.asp (accessed August 17, 2004).

Chan, L.K. and Zhang, J. (2001) Cumulative sum control charts for the covariance matrix. *Statistica Sinica*, **11**, 767–790.

Chaput, E.K., Meek, J.I. and Heimer, R. (2002) Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut. *Emerging Infectious Diseases*, **8**, 943–948.

Chen, R. (1978) A surveillance system for congenital abnormalities. *Journal of the American Statistical Association*, **73**, 323–327.

Chen, J. and Glaz, J. (1996) Two dimensional discrete scan statistics. *Statistics and Probability Letters*, **31**, 59–68.

Chen, R., Mantel, N., Connelly, R.R. and Isacson, P. (1982) A monitoring system for chronic diseases. *Methods of Information in Medicine*, **21**, 86–90.

Chen, R., Connelly, R.R. and Mantel, N. (1993) Analysing post-alarm data in a monitoring system, in order to accept or reject the alarm. *Statistics in Medicine*, **12**, 1807–1812.

Chen, R., Iscovich, J. and Goldbourt, U. (1997) Clustering of leukaemia cases in a city in Israel. *Statistics in Medicine*, **16**, 1873–1887.

Chib, S. and Jeliazkov, I. (2001) Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, **96**, 270–281.

Choi, K. and Thacker, S.B. (1981) An evaluation of influenza mortality surveillance, 1962–1979. 1: Time series forecasts of expected pneumonia and influenza deaths. *American Journal of Epidemiology*, **113**, 215–226.

Chu, C.-S.J. (1995) Detecting a shift in GARCH models. *Econometric Reviews*, **14**, 241–266.

Clark, A.B. and Lawson, A.B. (2002) Spatio-temporal cluster modelling of small area health data. In A.B. Lawson and D. Denison (eds), *Spatial Cluster Modelling*, Chapter 14. Boca Raton, FL: CRC Press.

Clark, A.B. and Lawson, A.B. (2005) Surveillance of individual level disease maps. Submitted.

Clayton, D.G. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk. In P. Elliott, J. Cuzick, D. English and R. Stern (eds), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. Oxford: Oxford University Press.

Clayton, D.G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, **43**, 671–691.

Cliff, A.D. and Ord, J.K. (1981) *Spatial Processes: Models and Applications*. London: Pion.

Colorado Department of Public Health and Environment (2002) Analysis of birth defect data in the vicinity of the Redfield plume area in southeastern Denver county: 1989–1999. Colorado Department of Public Health and the Environment (http://www.cdphe.state.co.us/hm/redfield_birthdefects_study.pdf).

Cooper, B. and Lipsitch, M. (2004) The analysis of hospital infection data using hidden Markov models. *Biostatistics*, **5**, 223–237.

Cooper, G. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Costagliola, D., Flahault, A., Galinec, D., Garnerin, P., Menares, J. and Valleron, A.-J. (1991) A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. *American Journal of Public Health*, **81**, 97–99.

Coulston, J.W. and Riitters, K.H. (2003) Geographic analysis of forest health indicators using spatial scan statistics. *Environmental Management*, **31**, 764–773.

Cousens, S., Smith, P.G., Ward, H., Everington, D., Knight, R.S.G., Zeidler, M., Stewart, G., Smith-Bathgate, E.A.B., Macleod, M.A., Mackenzie, J. and Will R.G. (2001) Geographical distribution of variant Creutzfeldt–Jakob disease in Great Britain, 1994–2000. *Lancet*, **357**, 1002–1007.

Couteron, P. and Kokou, K. (1997) Woody vegetation spatial patterns in a semi-arid savanna of Burkina Faso, West Africa. *Plant Ecology*, **132**, 211–227.

Crosier, R.B. (1988) Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, **30**, 291–303.

Crowder, S.V. (1989) Design of exponentially weighted moving average schemes. *Journal of Quality Technology*, **21**, 155–162.

Cuzick, J. and Edwards, R. (1990) Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society, Series B*, **52**, 73–104.

Cuzick, J. and Hills, M. (1991) Clustering and clusters – summary. In G. Draper (ed.), *The Geographical Epidemiology of Childhood Leukaemia and Non-Hodgkin Lymphomas in Great Britain, 1966–1983*, pp. 123–125. London: HMSO.

Daley, D. and Vere-Jones, D. (1988) *An Introduction to the Theory of Point Processes*. New York: Springer-Verlag.

Daniel, W.W. (1995) *Biostatistics: A Foundation for Analysis in the Health Sciences*. New York: John Wiley & Sons, Inc.

Davison, A.C. and Snell, E.J. (1991) Residuals and diagnostics. In D.V. Hinkley, N. Reid and E.J. Snell (eds), *Statistical Theory and Modelling*. London: Chapman & Hall.

Dawid, A. (1992) Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, **2**, 25–36.

Deng, K. and Moore, A.W. (1995) Multiresolution instance-based learning. In *Proceedings of the 14th International Joint Conference on Artifical Intelligence*, pp. 1233–1239. San Mateo, CA: Morgan Kanfmann.

Devine, O. and Louis, T. (1994) A constrained empirical Bayes estimator for incidence rates in areas with small populations. *Statistics in Medicine*, **13**, 1119–1133.

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

Diggle, P. and Rowlingson, B. (1994) A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A*, **157**, 433–440.

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

Diggle, P.J., Tawn, J. and Moyeed, R. (1998) Model-based geostatistics. *Applied Statistics*, **47**, 299–350.

Diggle, P.J., Morris, S. and Wakefield, J.C. (2000) Point-source modelling using matched case-control data. *Biostatistics*, **1**, 1–17.

Diggle, P., Knorr-Held, L., Rowlingson, B., Su, T., Hawtin, P. and Bryant, T. (2004) On-line monitoring of public health surveillance data. In R. Brookmeyer and D. Stroup (eds), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, Chapter 9. Oxford: Oxford University Press.

Dobbertin, M., Baltensweiler, A. and Rigling, D. (2001) Tree mortality in an unmanaged mountain pine (*Pinus mugo* var. *uncinata*) stand in the Swiss National Park impacted by root rot fungi. *Forest Ecology and Management*, **145**, 79–89.

Doherr, M.G., Hett, A.R., Rufenacht, J., Zurbriggen, A. and Heim, D. (2002) Geographical clustering of cases of bovine spongiform encephalopathy (BSE) born in Switzerland after the feed ban. *Veterinary Record*, **151**, 467–472.

Domangue, R. and Patch, S.C. (1991) Some omnibus exponentially weighted moving average statistical process monitoring schemes. *Technometrics*, **33**, 299–313.

Doucet, A., de Freitas, N. and Gordon, N. (eds) (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Du, J.G. and Li, Y. (1991) The integer-valued autoregressive (INAR(p)) model. *Journal of Times Series Analysis*, **12**, 129–142.

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

DuMouchel, W. (1999) Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *American Statistician*, **53**, 177–202.

Dwass, M. (1957) Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, **28**, 181–187.

Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. (eds) (2000) *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press.

Enemark, H.L., Ahrens, P., Juel, C.D., Petersen, E., Petersen, R.F., Andersen, J.S., Lind, P. and Thamsborg, S.M. (2002) Molecular characterization of Danish *Cryptosporidium parvum* isolates. *Parasitology*, **125**, 331–341.

Ewan, W.D. and Kemp, K.W. (1960) Sampling inspection of continous processes with no autocorrelation between successive results. *Biometrika*, **47**, 363–380.

Falconi, F., Ochs, H. and Deplazes, P. (2002) Serological cross-sectional survey of psoroptic sheep scab in Switzerland. *Veterinary Parasitology*, **109**, 119–127.

Fang, Z., Kulldorff, M. and Gregorio, D.I. (2004) Brain cancer in the United States, 1986–95: a geographic analysis. *Neuro-oncology*, **6**, 179–187.

Farrington, P. and Andrews, N. (2004) Outbreak detection: application to infectious disease surveillance. In R. Brookmeyer and D. Stroup (eds), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, Chapter 8. Oxford: Oxford University Press.

Farrington, P., and Beale, A.D. (1998) The detection of outbreaks of infectious disease. In L. Lierl, A.D. Cliff, A. Valleron, P. Farrington, and M. Bull (eds), *Geomed '97*. Stuttgart: B.G. Teubner.

Farrington, C.P. and Andrews, N.J. (2004) Statistical aspects of detecting infectious disease outbreaks. In R. Brookmeyer and D.F. Stroup (eds), *Monitoring the Health of Populations*. Oxford: Oxford University Press.

Farrington, C.P., Andrews, N.J., Beale, A.D. and Catchpole, M.A. (1996) A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society, Series A*, **159**, 547–563.

Fawcett, T. and Provost, F. (1999) Activity monitoring: noticing interesting changes in behavior. In S. Chaudhuri and D. Madigan (eds), *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 53–62. New York: Association for Computing Machinery.

Fernàndez, C. and Green, P.J. (2002) Modelling spatially corelated data via mixtures: a Bayesian approach. *Journal of the Royal Statitsical Society, Series B*, **64**, 805–826.

Fevre, E.M., Coleman, P.G., Odiit, M., Magona, J.W., Welburn, S.C. and Woolhouse, M.E.J. (2001) The origins of a new *Trypanosoma brucei rhodesiense* sleeping sickness outbreak in eastern Uganda. *Lancet*, **358**, 625–628.

Flahault, A., Garnerin, P., Chauvin, P., Farran, N., Saidi, Y., Diaz, C. and Toubiana, L. (1995) Sentinelle traces of an epidemic of acute gastroenteritis in France. *Lancet*, **346**, 162–163.

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

Frisén, M. (1986) Unimodal regression. *The Statistician*, **35**, 479–485.

Frisén, M. (1992) Evaluations of methods for statistical surveillance. *Statistics in Medicine*, **11**, 1489–1502.

Frisén, M. (2000) Statistical surveillance of business cycles. Research Report, Department of Statistics, Göteborg University.

Frisén, M. (2003) Statistical surveillance: optimality and methods. *International Statistical Review*, **71**, 403–434.

Frisén, M. and de Maré, J. (1991) Optimal surveillance. *Biometrika*, **78**, 271–280.

Frisén, M. and Gottlow, M. (2003) Graphical evaluation of statistical surveillance. Research Report 2003:10, Statistical Research Unit, Göteborg University.

Frisén, M. and Wessman, P. (1999) Evaluations of likelihood ratio methods for surveillance. Differences and robustness. *Communications in Statistics. Simulation and Computation*, **28**, 597–622.

Fuchs, C. and Benjamini, Y. (1994) Multivariate profile charts for statistical process control. *Technometrics*, **36**, 182–195.

Gallus, G., Radaelli, G. and Marchi, M. (1991) Poisson approximation to a negative binomial process in the surveillance of rare health events. *Methods of Information in Medicine*, **30**, 206–209.

Gallus, G., Mandelli, C., Marchi, M. and Radaelli, G. (1986) On surveillance methods for congenital malformations. *Statistics in Medicine*, **5**, 565–571.

Gan, F.F. (1991) Monitoring observations generated from a binomial distribution using modified exponentially weighted moving average control charts. *Journal of Statistical Computation and Simulation*, **37**, 45–60.

Gan, F.F. (1992a) Cusum control charts under linear drift. *The Statistician*, **41**, 71–84.

Gan, F.F. (1992b) Exact run length distributions for one-sided exponential CUSUM schemes. *Statistica Sinica*, **2**, 297–312.

Gan, F.F. (1993) An optimal design of EWMA control charts based on median run-length. *Journal of Statistical Computation and Simulation*, **45**, 169–184.

Gan, F.F. (1994) Design of optimal exponential CUSUM control charts. *Journal of Quality Technology*, **26**, 109–124.

Gangnon, R. and Clayton, M. (2000) Bayesian detection and modeling of spatial disease clustering. *Biometrics*, **56**, 922–935.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis* (2nd edn). London: Chapman & Hall.

George, M., Wiklund, L., Aastrup, M., Pousette, J., Thunholm, B., Saldeen, T., Wernroth, L., Zaren, B. and Holmberg, L. (2001) Incidence and geographical distribution of sudden infant death syndrome in relation to content of nitrate in drinking water and groundwater levels. *European Journal of Clinical Investigation*, **31**, 1083–1094.

German, R.R. (2000) Sensitivity and predictive value positive measurements for public health surveillance systems. *Epidemiology*, **11**, 720–727.

Glaz, J. and Balakrishnan, N. (eds) (1999) *Scan Statistics and Applications*. Boston: Birkhäuser.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*. New York: Springer-Verlag.

Glick, B.J. (1979) The spatial autocorrelation of cancer mortality. *Social Science and Medicine*, **13D**, 123–130.

Goldenberg, A., G. Shmueli, R.A. Caruana, and S.E. Fienberg, (2002) Early statistical detection of anthrax outbreaks by tracking over-the-couner medication sales. *Proceedings of the National Academy of Sciences*, **99**, 5237–5249.

Good, P. (2000) *Permutation Tests – A Practical Guide to Resampling Methods for Testing Hypotheses* (2nd edn). New York: Springer-Verlag.

Gordon, L. and Pollak, M. (1997) Average run length to false alarm for surveillance schemes designed with partially specified pre-change distribution. *Annals of Statistics*, **25**, 1284–1310.

Green, P.J. and Richardson, S. (2002) Hidden Markov models and disease mapping. *Journal of the American Statitsical Association*, **97**, 1055–1070.

Green, C., Hoppa, R.D., Young, T.K. and Blanchard, J.F. (2003) Geographic analysis of diabetes prevalence in an urban area. *Social Science and Medicine*, **57**, 551–560.

Gregorio, D.I., Kulldorff, M., Barry, L., Samociuk, H. and Zarfos, K. (2001) Geographic differences in primary therapy for early stage breast cancer. *Annals of Surgical Oncology*, **8**, 844–849.

Gregorio, D.I., Kulldorff, M., Barry, L. and Samociuk, H. (2002) Geographic differences in invasive and in situ breast cancer incidence according to precise geographic coordinates, Connecticut, 1991–1995. *International Journal of Cancer*, **100**, 194–198.

Grimson, R.C. (1991) A versatile test for clustering and a proximity analysis of neurons. *Methods of Information in Medicine*, **30**, 299–303.

Gustavsson, F. (2000) *Adaptive Filtering and Change Detection.* Chichester: John Wiley & Sons, Ltd.

Hamilton, J.D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

Hand, D., Mannila, H. and Smyth, P. (2001) *Principles of Data Mining*. Cambridge, MA: MIT Press.

Hanson, C.E. and Wieczorek, W.F. (2002) Alcohol mortality: a comparison of spatial clustering methods. *Social Science and Medicine*, **55**, 791–802.

Härdle, W. (1991) *Smoothing Techniques: with Implementation in S*. New York: Springer-Verlag.

Harrison, P.J. and Stevens, C.F. (1976) Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society, Series B*, **38**, 205–247.

Hauck, D.J., Runger, G.C. and Montgomery, D.C. (1999) Multivariate statistical process monitoring and diagnosis with grouped regression-adjusted variables. *Communications in Statistics. Simulation and Computation*, **28**, 309–328.

Hawkins, D.M. (1991) Multivariate quality control based on regression-adjusted variables. *Technometrics*, **33**, 61–75.

Hawkins, D.M. and Olwell, D.H. (1998) *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer-Verlag.

Healy, J.D. (1987) A note on multivariate CUSUM procedures. *Technometrics*, **29**, 409–412.

Heffernan, R., Mostashari, F., Das, D., Karpati, A., Kulldorff, M. and Weiss, D. (2004) Syndromic surveillance in public health practice: the New York City emergency department system. *Emerging Infectious Diseases*, **10**, 858–864.

Helfenstein, U. (1986) Box–Jenkins modelling of some viral infectious diseases. *Statistics in Medicine*, **5**, 37–47.

Hill, G.B., Spicer, C.C. and Weatherall, J.A.C. (1968) The computer surveillance of congenital malformations. *British Medical Journal*, **24**, 215–218.

Hjalmars, U., Kulldorff, M., Gustafsson, G. and Nagarwalla, N. (1996) Childhood leukemia in Sweden: using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine*, **15**, 707–715.

Hjalmars, U., Kulldorff, M., Wahlquist, Y. and Lannering, B. (1999) Increased incidence rates but no space-time clustering of childhood malignant brain tumors in Sweden. *Cancer*, **85**, 2077–2090.

Hoar, B.R., Chomel, B.B., Rolfe, D.L., Chang, C.C., Fritz, C.L., Sacks, B.N. and Carpenter, T.E. (2003) Spatial analysis of *Yersinia pestis* and *Bartonella vinsonii* subsp. *berkhoffii* seroprevalence in California coyotes (*Canis latrans*). *Preventive Veterinary Medicine*, **56**, 299–311.

Hossain, M. and Lawson, A.B. (2005) Local likelihood disease clustering: development and evaluation. *Environmental and Ecological statistics* (to appear)

Hotelling, H. (1947) Multivariate quality control. In C. Eisenhart, M.W. Hastay and W.A. Wallis (eds), *Techniques of Statistical Analysis*. New York: McGraw-Hill.

Hsu, C.E., Jacobson, H.E. and Soto Mas, F. (2004) Evaluating the disparity of female breast cancer mortality among racial groups – a spatiotemporal analysis. *International Journal of Health Geographics*, **3**:4.

Huang, L., Kulldorff, M. and Gregorio, D. (2004) A spatial scan statistic for survival data. Manuscript.

Huillard d'Aignaux, J., Cousens, S.N., Delasnerie-Laupretre, N., Brandel, J.P., Salomon, D., Laplanche, J.L., Hauw, J.J. and Alperovitch, A. (2002) Analysis of the geographical distribution of sporadic Creutzfeldt–Jakob disease in France between 1992 and 1998. *International Journal of Epidemiology*, **31**, 490–495.

Hunter, J.S. (1986) The exponentially weighted moving average. *Journal of Quality Technology*, **18**, 203–210.

Hutwagner, L.C., Maloney, E.K., Bean, N.H., Slusker, L. and Martin, S.M. (1997) Using laboratory-based surveillance data for prevention: an algorithm for detecting salmonella outbreaks. *Emerging Infectious Diseases*, **3**, 395–400.

Hutwagner, L., Thompson, W., Seeman, G.M. and Treadwell, T. (2003) The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *Journal of Urban Health – Bulletin of the New York Academy of Medicine*, **80**, i89–i96.

Inskip, H., Beral, V., Fraser, P. and Haskey, P. (1983) Methods for age-adjustment of rates. *Statistics in Medicine*, **2**, 483–493.

Jackson, J.E. (1985) Multivariate quality control. *Communications in Statistics. Theory and Methods*, **14**, 2657–2688.

Järpe, E. (1999) Surveillance of the interaction parameter of the Ising model. *Communications in Statistics. Theory and Methods*, **28**, 3009–3027.

Järpe, E. (2000) On univariate and spatial surveillance. Ph.D thesis, Department of Statistics, Göteborg University.

Järpe, E. (2001) Surveillance, environmental. In A.El-Shaarawi and W.W. Piegorsh (eds), *Encyclopedia of Environmetrics*. Chichester: John Wiley & Sons, Ltd.

Jemal, A., Kulldorff, M., Devesa, S.S., Hayes, R.B. and Fraumeni, J.F. (2002) A geographic analysis of prostate cancer mortality in the United States. *International Journal of Cancer*, **101**, 168–174.

Johnson, R.A. and Wichern, D.W. (2002) *Applied Multivariate Statistical Analysis* (5th edn). Upper Saddle River, NJ: Prentice Hall.

Joly, D.O., Ribic, C.A., Langenberg, J.A., Beheler, K., Batha, C.A., Dhuey, B.J., Rolley, R.E., Bartelt, G., Van Deelen, T.R. and Samual, M.D. (2003) Chronic wasting disease in free-ranging Wisconsin white-tailed deer. *Emerging Infectious Diseases*, **9**, 599–601.

Kafadar, K. and Stroup, D.F. (1992) Analysis of aberrations in public health surveillance data: estimating variances on correlated samples. *Statistics in Medicine*, **11**, 1551–1568.

Källén, B. and Winberg, J. (1969) Multiple malformations studied with a national register of malformations. *Pediatrics*, **44**, 410–417.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, **82**, 35–45.

Kaminski, R.J., Jefferis, E.S. and Chanhatasilpa, C. (2000) A spatial analysis of American police killed in the line of duty. In L.S. Turnbull, E.H. Hendrix and B.D. Dent (eds), *Atlas of Crime: Mapping the Criminal Landscape.* Phoenix, AZ: Oryx Press.

Kang, L. and Albin, S.L. (2000) On-line monitoring when the process yields a linear profile. *Journal of Quality Technology*, **32**, 418–426.

Karlin, S. and Taylor, H. (1975) *First Course in Stochastic Processes* (2nd edn). New York: Academic Press.

Kaufmann, A.F., Meltzer, M.I. and Schmid, G.P. (1997) The economic impact of a bioterrorist attack: Are prevention and postattack intervention programs justifiable? *Emerging Infectious Diseases*, **3**, 83–94.

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

Kelsall, J. and Diggle, P. (1998) Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics*, **47**, 559–573.

Kenett, R. and Pollak, M. (1996) Data-analytic aspects of the Shiryaev–Roberts control chart: surveillance of a non-homogeneous Poisson process. *Journal of Applied Statistics*, **23**, 125–137.

Kharrazi, M., *et al.* (1998) Pregnancy outcomes around the B.K.K. landfill, West Covina, California: Ananalysis by address. California Department of Health Services.

Klassen, A.C., Curriero, F.C., Hong, J.H., Williams, C., Kulldorff, M., Meissner, H.I., Alberg, A. and Ensminger, M. (2004) The role of area-level influences on prostate cancer grade and stage at diagnosis. *Preventive Medicine*, **39**, 441–448.

Kleinman, K., Abrams, A., Kulldorff, M. and Platt, R. (2005) Choosing a denominator for the space-time scan statistic approach to spatial surveillance. *Epidemiology and Infection* (to appear).

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

Kleinman, K., Lazarus, R. and Platt, R. (2004b) Kleinman *et al.* Respond to 'Surveilling Surveillance'. *American Journal of Epidomiology*, **159**, 228.

Knorr-Held, L. (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19**, 2555–2567.

Knorr-Held, L. and Besag, J. (1998) Modelling risk from disease in time and space. *Statistics in Medicine*, **17**, 2045–2060.

Knoth, S. and Schmid, W. (2002) Monitoring the mean and the variance of a stationary process. *Statistica Neerlandica*, **56**, 77–100.

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

Knox, E.G. (1989) Detection of clusters. In P. Elliott (ed.), *Methodology of Enquiries into Disease Clustering*, pp. 17–20. London: Small Area Health Statistics Unit.

Knuesel, R., Segner, H. and Wahli, T. (2003) A survey of viral diseases in farmed and feral salmonids in Switzerland. *Journal of Fish Diseases*, **26**, 167–182.

Kourti, T. and MacGregor, J.F. (1996) Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, **28**, 409–428.

Krieger, A.M., Pollak, M. and Yakir, B. (2003) Surveillance of a simple linear regression. *Journal of the American Statistical Association*, **98**, 456–469.

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

Kulldorff, M. (1998) Statistical methods for spatial epidemiology: tests for randomness. In M. Löytönen and A. Gatrell (eds), *GIS and Health*. London: Taylor & Francis, 49–62.

Kulldorff, M. (1999) Spatial scan statistics: models, calculations and applications. In J. Glaz and N. Balakrishnan (eds), *Scan Statistics and Applications*, pp. 303–322. Boston: Birkhäuser.

Kulldorff, M. (2001) Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A*, **164**, 61–72.

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

Kulldorff, M., Feuer, E.J., Miller, B.A. and Freedman, L.S. (1997) Breast cancer in north-eastern United States: a geographical analysis. *American Journal of Epidemiology*, **146**, 161–170.

Kulldorff, M., Athas, W., Feuer, E., Miller, B. and Key, C. (1998) Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health*, **88**, 1377–1380.

Kulldorff, M., Fang, Z. and Walsh, S.J. (2003) A tree-based scan statistic for database disease surveillance. *Biometrics*, **59**, 323–331.

Kulldorff, M., Tango, T. and Park, P. (2003) Power comparisons for disease clustering tests. *Computational Statistics and Data Analyis*, **42**, 665–684.

Kulldorff, M., Zhang, Z., Hartman, J., Heffernan, R., Huang, L. and Mostashari, F. (2004a) Evaluating disease outbreak detection methods: Benchmark data and power calculations. *Morbidity and Mortality Weekly Report*, **53** (supplement), 144–151.

Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2004b) An elliptic spatial scan statistic. Manuscript.

Lai, T.L. (1995) Sequential changepoint detection in quality-control and dynamical systems. *Journal of the Royal Statistical Society, Series B*, **57**, 613–658.

Lai, T.L. (1998) Information bounds and quick detection of parameter in stochastic systems. *IEEE Transactions on Information Theory*, **44**, 2917–2929.

Lai, T.L. and Shan, Z. (1999) Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Transactions on Automatic Control*, **44**, 952–966.

Latour, A. (1997) The multivariate GINAR(p) process. *Advances in Applied Probability*, **29**, 228–248.

Latour, A. (1998) Existence and stochastic structure of a non-negative integer-valued autoregressive processes. *Journal of Times Series Analysis*, **19**, 439–455.

Lauritzen, S., Dawid, A., Larsen, B. and Leimer, H.-G. (1990) Independence properties of directed Markov fields. *Networks*, **20**, 491–505.

Lawson, A.B. (1992) GLIM and normalising constant models in spatial and directional data analysis. *Computational Statistics and Data Analysis*, **13**, 331–348.

Lawson, A.B. (1993a) A deviance residual for heterogeneous spatial Poisson processes. *Biometrics*, **49**, 889–897.

Lawson, A.B. (1993b) On the analysis of mortality events associated with a pre-specified fixed point. *Journal of the Royal Statistical Society, Series A*, **156**, 363–377.

Lawson, A.B. (1995) Markov chain Monte Carlo methods for putative pollution source problems in environmental epidemiology. *Statistics in Medicine*, **14**, 2473–2486.

Lawson, A.B. (1996) Markov chain Monte Carlo methods for spatial cluster processes. In M.M. Meyer and J.L. Rosenberger (eds), *Computing Science and Statistics: Proceedings of the Interface*, Volume 27, pp. 314–319. Fairfax Station, VA: Interface Foundation of North America.

Lawson, A.B. (2001) *Statistical Methods in Spatial Epidemiology*. New York: John Wiley & Sons, Inc.

Lawson, A.B. (2005) Local likelihood Bayesian cluster modelling for small area health data. Submitted.

Lawson, A.B. (2004) Some issues in the spatio-temporal analysis of public health surveillance data. In R. Brookmeyer and D. Stroup (eds), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, Chapter 11. Oxford: Oxford University Press.

Lawson, A.B., Biggeri, A. and Lagazio, C. (1996) Modelling heterogeneity in discrete spatial data models via MAP and MCMC methods. In A. Forcina, G. Marchetti, R. Hatzinger, and G. Galmacci (eds), Proceedings of the 11th International Workshop on Statistical Modelling, pp. 240–250. Graphos, Citta di Castello.

Lawson, A.B., Biggeri, A. and Dreassi, E. (1999) Edge effects in disease mapping. In A.B. Lawson, D. Böhning, E. Lesaffre, A. Biggeri, J.-F. Viel, and R. Bertollini (eds), Disease Mapping and Risk assessment for Public Health, Chapter 6, pp. 85–97. Wiley.

Lawson, A.B. Böhning, D., Lesaffre, E., Biggeri, A., Viel, J.-F. and Bertollini, R. (1999) *Disease Mapping and Risk Assessment for Public Health*. Chichester: John Wiley & Sons, Ltd.

Lawson, A.B., Browne, W.J. and Vidal-Rodeiro, C.L. (2003) Disease Mapping with WinBUGS and MLwiN. Chichester: John Wiley & Sons, Ltd.

Lawson, A.B. and Clark, A. (1999). Markov chain Monte Carlo methods for clustering in case event and count data in spatial epidemiology. In M.E. Halloran and D. Berry (eds), *Statistics and Epidemiology: Environment and Clinical Trials*, pp. 193–218. New York: Springer-Verlag.

Lawson, A.B., Clark, A.B. and Vidal-Rodeiro, C.L. (2004) Developments in general and syndromic surveillance for small area health data. *Journal of Applied Statistics*, **31**, 963–978.

Lawson, A.B. and Cressie, N. (2000) Spatial statistical methods for environmental epidemiology. In C.R. Rao and P.K. Sen (eds), *Bioenvironmental and Public Health Statistics*, Handbook of Statistics Vol. 18, pp. 357–396. Amsterdam: Elsevier.

Lawson, A.B. and Denison, D. (2002) *Spatial Cluster Modelling*. Boca Raton, FL: Chapman & Hall/CRC.

Lawson, A.B. and Waller, L. (1996) A review of point pattern methods for spatial modelling of events around sources of pollution. *Environmetrics*, **7**, 471–488.

Lawson, A.B. and Williams, F. (1993) Applications of extraction mapping in environmental epidemiology. *Statistics in Medicine*, **12**, 1249–1258.

Lawson, A.B. and Williams, F. (1994) Armadale: a case study in environmental epidemiology. *Journal of the Royal Statistical Society, Series A*, **157**, 285–298.

Lawson, A.B. and Williams, F. (2000) Spatial competing risk models in disease mapping. *Statistics in Medicine*, **19**, 2451–2468.

Lazarus, R., Kleinman, K., Dashevsky, I., DeMaria Jr, A., and Platt, R. (2001) Using automated records for rapid detection of illness syndromes: the example of lower respiratory disease. *BMC Public Health*, **1**, 1–9.

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

Le Strat, Y. and Carrat, F. (1999) Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, **18**, 3463–3478.

Liu, R.Y. (1995) Control charts for multivariate processes. *Journal of the American Statistical Association*, **90**, 1380–1387.

Ljung, G.M. and Box, G.E.P. (1978) On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

Loader, C.R. (1991) Large-deviation approximations to the distribution of scan statistics. *Advances in Applied Probability*, **23**, 751–771.

Lober, W.B., Karras, B.T., Wagner, M.M., Overhage, J.M., Davidson, A.J., Fraser, H., Trigg, L.J., Mandl, K.D., Espino, J.U. and Tsui, F.C. (2002) Roundtable on bioterrorism detection: information system-based surveillance. *Journal of the American Medical Informatics Association*, **9**, 105–115.

Lober, W.B., Trigg, L.J., Karras, B.T., Bliss, D., Ciliberti, J., Stewart, L. and Duchin, J.S. (2003) Syndromic surveillance using automated collection of computerized discharge diagnoses. *Journal of Urban Health*, **80**, i97–i106.

Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Lewis, S.H., Loschen, W., Sari, J., Sniegoski, C., Wojcik, R. and Pavlin, J. (2003)A systems overview of the Electronic Surviellence System for the Early Notification of Community-Based Epidemics (ESSENCE II). *Journal of Urban Health – Bulletin of the New York Academy of Medicine*, **80**, i32–i42.

López-Abente, G., Morales-Piga, A., Bachiller-Corral, F.J., Illera-Martín, O., Martín-Domenech, R. and Abraira, V. (2003) Identification of possible areas of high prevalence of Paget's disease of bone in Spain. *Clinical and Experimental Rheumatology*, **21**, 635–638.

Lowry, C.A. and Montgomery, D.C. (1995) A review of multivariate control charts. *IIE Transactions*, **27**, 800–810.

Lowry, C.A., Woodall, W.H., Champ, C.W. and Rigdon, S.E. (1992) A multivariate exponentially weighted moving average control chart. *Technometrics*, **34**, 46–53.

Lu, X.S., Xie, M., Goh, T.N. and Lai, C.D. (1998) Control chart for multivariate attribute processes. *International Journal of Production Research*, **36**, 3477–3489.

Lucas, J.M. (1985) Counted data CUSUM's. *Technometrics*, **27**, 129–144.

Lucas, J.M. and Crosier, R.B. (1982) Fast initial response for CUSUM quality control schemes: give your CUSUM a head start. *Technometrics*, **24**, 199–205.

Lucas, J.M. and Saccucci, M.S. (1990) Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, **32**, 1–12.

Lui, K.-J. and Rudy, R.K. (1989) An application of a mathematical model to adjust for time lag in case reporting. *Statistics in Medicine*, **8**, 259–262.

MacDonald, L. and Zucchini, W. (1997) *Hidden Markov and Other Models for Discrete-Valued Time Series*. London: Chapman & Hall.

MacEachren, A.M. (1995) *How Maps Work: Representation, Visualisation and Design*. New York: Guilford Press.

Mandl, K.D., Overhage, J.M., Wagner, M.M., Lober, W.B., Sebastiani, P., Mostashari, F., Pavlin, J.A., Gesteland, P.H., Treadwell, T., Koski, E., Hutwagner, L., Buckeridge, D.L., Aller, R.D. and Grannis, S. (2004) Implementing syndromic surveillance: a practical guide informed by the early experience. *Journal of the American Medical Informatics Association*, **11**, 141–150.

Manton, K., Woodbury, M. and Stallard, E. (1981) A variance components approach to categorical data models with heterogeneous mortality rates in North Carolina counties. *Biometrics*, **37**, 259–269.

Maravelakis, P.E., Bersimis, S., Panaretos, J. and Psarakis, S. (2002) Identifying the out of control variable in a multivariate control chart. *Communications in Statistics. Theory and Methods*, **31**, 2391–2408.

Margai, F. and Henry, N. (2003) A community-based assessment of learning disabilities using environmental and contextual risk factors. *Social Science and Medicine*, **56**, 1073–1085.

Marshall, R. (1991) Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics*, **40**, 283–294.

Marshall, C., Best, N., Bottle, A. and Aylin, P. (2004) Statistical issues in the prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society, Series A*, **167**, 541–559.

Mason, R.L., Tracy, N.D. and Young, J.C. (1995) Decomposition of T^2 for multivariate control chart interpretation. *Journal of Quality Technology*, **27**, 99–108.

Mastrangelo, C.M., Runger, G.C. and Montgomery, D.C. (1996) Statistical process monitoring with principal components. *Quality and Reliability Engineering International*, **12**, 203–210.

Mathers, C., Harris, R. and Lancaster, P. (1994) A CUSUM scheme based on the exponential distribution for surveillance of rare congenital malformations. *Australian Journal of Statistics*, **36**, 21–30.

McCulloch, C. and Searle, S. (2001) *Generalized, Linear, and Mixed Models*, New York: John Wiley & Sons, Inc.

Michelozzi, P., Capon, A., Kirchmayer, U., Forastiere, F., Biggeri, A., Barca, A. and Perucci, C.A. (2002) Adult and childhood leukemia near a high-power radio station in Rome, Italy. *American Journal of Epidemiology*, **155**, 1096–1103.

Miller, C.J., C. Genovese, R.C. Nichol, L. Wasserman, A. Connolly, D. Reichart, A. Hopkins, J. Schneider, and A. Moore (2001) Controlling the false discovery rate in astrophysical data analysis. Technical report, Carnegie Mellon University.

Miller, M.A., Gardner, I.A., Kreuder, C., Paradies, D.M., Worcester, K.R., Jessup, D.A., Dodd, E., Harris, M.D., Ames, J.A., Packham, A.E. and Conrad, P.A. (2002) Coastal freshwater runoff is a risk factor for *Toxoplasma gondii* infection of southern sea otters *(Enhydra lutris nereis)*. *International Journal for Parasitology*, **32**, 997–1006.

Miller, M.A., Grigg, M.E., Kreuder, C., James, E.R., Melli, A.C., Crosbie, P.R., Jessup, D.A., Boothroyd, J.C., Brownstein, D. and Conrad, P.A. (2004) An unusual genotype of *Toxoplasma gondii* is common in California sea otters *(Enhydra lutris nereis) and is a cause of mortality. International Journal for Parasitology*, **34**, 275–284.

MMWR (2001) Recognition of illness associated with the intentional release of a biologic agent. *Morbidity and Mortality Weekly Report*, **50**, 893–897.

Mohebbi, C. and Havre, L. (1989) Multivariate control charts: a loss function approach. *Sequential Analysis*, **8**, 253–268.

Møller, J. and Waagepetersen, R.P. (2002) Statitsical inference for Cox processes. In A.B. Lawson and D. Denison (eds), *Spatial Cluster Modelling*, Chapter 3. Boca Raton, FL: CRC Press.

Monmonier, M. (1996) *How to Lie with Maps* (2nd edn). Chicago: University of Chicago Press.

Montgomery, D. (1996) *Introduction to Statistical Quality Control*. New York: John Wiley & Sons, Inc.

Moore, A.W. (1999) Very fast mixture-model-based clustering using multiresolution kd-trees. In M. Kearns and D. Cohn (eds), *Advances in Neural Information Processing Systems*, Volume 10, pp. 543–549. San Francisco: Morgan Kaufmann.

Moore, A. and M.S. Lee (1998) Cached sufficient statistics for efficient machine learning with large datasets. *Journal of Artificial Intelligence Research*, **8**, 67–91.

Moore, A. and Wong, W.-K. (2003) Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. In T. Fawcett and N. Mishra (eds), *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 552–559. Menlo Park, CA: AAAI Press.

Moore, M., Katona, P., Kaplan, J.E., Schonberger, L.B. and Hatch, M.H. (1982) Poliomyelitis in the United States, 1969–1981. *Journal of Infectious Diseases*, **4**, 558–563.

Mostashari, F. and Hartman, J. (2003) Syndromic surveillance: a local perspective. *Journal of Urban Health*, **80**, i1–i7.

Mostashari, F., Kulldorff, M., Hartman, J.J., Miller, J.R. and Kulasekera, V. (2003) Dead bird clustering: a potential early warning system for West Nile virus activity. *Emerging Infectious Diseases*, **9**, 641–646.

Moustakides, G.V. (1986) Optimal stopping-times for detecting changes in distributions. *Annals of Statistics*, **14**, 1379–1387.

Naus, J. (1965a) The distribution of the size of maximum cluster of points on the line. *Journal of the American Statistical Association*, **60**, 532–538.

Naus, J. (1965b) Clustering of random points in two dimensions. *Biometrika*, **52**, 263–267.

Neill, D.B. and Moore, A.W. (2004a) A fast multi-resolution method for detection of significant spatial disease clusters. In S. Thrun, L. Saul, and B. Schölkopf (eds), *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Neill, D.B. and Moore, A.W. (2004b) Rapid detection of significant spatial clusters. In R. Kolhavi, J. Gehrke, W. DuMouchel and J. Ghosh (eds), *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery.

New York State Department of Health (2001) Cancer Surveillance Improvement Initiative. (http://www.health.state.ny.us/nysdoh/cancer/csii/nyscsii2.htm).

Ngai, H.M. and Zhang, J. (2001) Multivariate cumulative sum control charts based on projection pursuit. *Statistica Sinica*, **11**, 747–766.

Nobre, F.F., Monteiro, A.B.S., Telles, P.R. and Williamson, G.D. (2001) Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statistics in Medicine*, **20**, 3051–3069.

Norström, M., Pfeiffer, D.U. and Jarp, J. (2000) A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. *Preventive Veterinary Medicine*, **47**, 107–119.

Olea-Popelka, F.J., Griffin, J.M., Collins, J.D., McGrath, G. and Martin, S.W. (2003) Bovine tuberculosis in badgers in four areas in Ireland: does tuberculosis cluster? *Preventive Veterinary Medicine*, **59**, 103–111.

Openshaw, S., Charlton, M., Craft, A. and Birch, J. (1988) Investigation of leukemia clusters by use of a geographical analysis machine. *Lancet*, **1**, 272–273.

Orre, R., Lansner, A., Bate, A. and Lindquist, M. (2000) Bayesian neural networks with confidence estimations applied to data mining. *Computational Statistics & Data Analysis*, **34**, 473–493.

Ozonoff, A., Bonetti, M., Forsberg, L. and Pagano, M. (2004) Power comparisons for an improved disease clustering test. *Computational Statistics and Data Analyis*, **48**, 679–684

Pagano, M. (1978) On periodic and multiple autoregressions. *Annals of Statistics*, **6**, 1310–1317.

Page, E.S. (1954) Continuous inspection schemes. *Biometrika*, **41**, 100–115.

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

Patil, G.P. and Taillie, C. (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, **11**, 183–197.

Pavlin, J.A. (2003) Investigation of disease outbreaks detected by 'syndromic' surveillance systems. *Journal of Urban Health*, **80**, i107–i114.

Perez, A.M., Ward, M.P., Torres, P. and Ritacco, V. (2002) Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina. *Preventive Veterinary Medicine*, **56**, 63–74.

Petzold, M., Sonesson, C., Bergman, E. and Kieler, H. (2004) Surveillance in longitudinal models. Detection of intrauterine growth restriction. *Biometrics*, **60**, 1025–1033

Pickle, L.W. and Hermann, D.J. (1995) Cognitive aspects of statistical mapping. Technical Report 18, NCHS Office of Research and Methodology, Washington, DC.

Pignatiello, Jr, J.J. and Runger, G.C. (1990) Comparisons of multivariate CUSUM charts. *Journal of Quality Technology*, **22**, 173–186.

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

Pollak, M. and Siegmund, D. (1975) Approximations to the expected sample size of certain sequential tests. *Annals of Statistics*, **3**, 1267–1282.

Pollak, M. and Siegmund, D. (1985) A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrika*, **72**, 267–280.

Poor, V.H. (1998) Quickest detection with exponential penalty for delay. *Annals of Statistics*, **26**, 2179–2205.

Preparata, F.P. and Shamos, M.I. (1985) *Computational Geometry: An Introduction*. New York: Springer-Verlag.

Qiu, P.H. and Hawkins, D. (2001) A rank-based multivariate CUSUM procedure. *Technometrics*, **43**, 120–132.

Qiu, P.H. and Hawkins, D. (2003) A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. *The Statistician*, **52**, 151–164.

Radaelli, G. (1992) Using the Cuscore technique in the surveillance of rare health events. *Journal of Applied Statistics*, **19**, 75–81.

Radaelli, G. (1994) Poisson and negative binomial dynamics for counted data under CUSUM-type chart. *Journal of Applied Statistics*, **21**, 347–356.

Radaelli, G. (1996) Detection of an unknown increase in the rate of a rare event. *Journal of Applied Statistics*, **23**, 105–113.

Rao, C.R. and Mitra, S.K. (1971) *Generalized Inverse of Matrices and its Applications*. Wiley.

Rath, T.M., Carreras, M. and Sebastiani, P. (2003) Automated detection of influenza epidemics with hidden Markov models. In M.R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse, C. Borgelt and F. Pfenning (eds), *Advances in Intelligent Data Analysis V*, pp. 521–531. Berlin: Springer-Verlag.

Raubertas, R.F. (1989) An analysis of disease surveillance data that uses geographic locations of the reporting units. *Statistics in Medicine*, **8**, 267–271.

Reis, B.Y. and Mandl, K.D. (2003) Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*, **3**, 1–11.

Reis, B.Y., Pagano, M. and Mandl, K.D. (2003) Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences of the USA*, **100**, 1961–1965.

Rigau-Perez, J.G., Millard, P.S., Walker, D.R., Deseda, C.C. and Casta-Velez, A. (1999) A deviation bar chart for detecting dengue outbreaks in Puerto Rico. *American Journal of Public Health*, **89**, 374–378.

Ripley, B. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Ripley, B.D. (1976) The second-order analysis of a stationary point process. *Journal of Applied Probability*, **13**, 255–266.

Ripley, B.D. (1981) *Spatial Statistics*. New York: Wiley.

Robert, C., Celeux, G. and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics and Probability Letters*, **16**, 77–83.

Robert, C., Rydén, T. and Titterington, D. (2000). Bayesian inference in hidden Markov models through reversible jump Markov chain Monte Carlo. *Journal of the Royal Statistical Society, Series B*, **62**, 57–75.

Robert, C.P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Roberts, S.W. (1959) Control chart tests based on geometric moving averages. *Technometrics*, **1**, 239–250.

Roberts, S.W. (1966) A comparison of some control chart procedures. *Technometrics*, **8**, 411–430.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988) *Order Restricted Inference*. Chichester: John Wiley & Sons, Ltd.

Roche, L.M., Skinner, R. and Weinstein, R.B. (2002) Use of a geographic information system to identify and characterize areas with high proportions of distant stage breast cancer. *Journal of Public Health Management and Practice*, **8**(2), 26–32.

Rogerson, P. (1997) Surveillance methods for monitoring the development of spatial patterns. *Statistics in Medicine*, **16**, 2081–2093.

Rogerson, P. (2001a) A statistical method for the detection of geographic clustering. *Geographical Analysis*, **33**, 215–227.

Rogerson, P. (2001b) Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society, Series A*, **164**, 87–96.

Rogerson, P. (2004) Formulas for the design of cusum quality control charts. Submitted for publication.

Rogerson, P. and Sun, Y. (2001) Spatial monitoring of geographic patterns: an application to crime analysis. *Computers, Environment, and Urban Systems*, **25**, 539–556.

Rogerson, P. and Yamada, I. (2004a) Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report*, **53** (supplement), 79–85.

Rogerson, P. and Yamada, I. (2004b) Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine*, **23**, 2195–2214.

Rosolowski, M. and Schmid, W. (2003) EWMA charts for monitoring the mean and the autocovariances of stationary Gaussian processes. *Sequential Analysis*, **22**, 257–285.

Ross, S. (1989) *Introduction to Probability Models* (4th edn). San Diego, CA: Academic Press.

Ross, A. and Davis, S. (1990) Point pattern analysis of the spatial proximity of residences prior to diagnosis of persons with Hodgkin's disease. *American Journal of Epidemiology*, **132**, 53–62.

Rossi, G., Lampugnani, L. and Marchi, M. (1999) An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine*, **18**, 2111–2122.

Runger, G.C. (1996) Projections and the U^2 chart for multivariate statistical process control. *Journal of Quality Technology*, **28**, 313–319.

Runger, G.C. and Prabhu, S.S. (1996) A Markov chain model for the multivariate exponentially weighted moving averages control chart. *Journal of the American Statistical Association*, **91**, 1701–1706.

Runger, G.C., Keats, J.B., Montgomery, D.C. and Scranton, R.D. (1999) Improving the performance of the multivariate exponentially weighted moving average control chart. *Quality and Reliability Engineering International*, **15**, 161–166.

Ryan, T.P. (2000) *Statistical Methods for Quality Improvement*. New York: John Wiley & Sons, Inc.

Sabel, C.E., Boyle, P.J., Löytönen, M., Gatrell, A.C., Jokelainen, M., Flowerdew, R. and Maasilta P. (2003) Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death. *American Journal of Epidemiology*, **157**, 898–905.

Samet, H. (1990) *The Design and Analysis of Spatial Data Structures*. Reading, MA: Addison-Wesley.

Sankoh, O.A., Ye, Y., Sauerborn, R., Muller, O. and Becher, H. (2001) Clustering of childhood mortality in rural Burkina Faso. *International Journal of Epidemiology*, **30**, 485–492.

Sauders, B.D., Fortes, E.D., Morse, D.L., Dumas, N., Kiehlbauch, J.A., Schukken, Y., Hibbs, J.R. and Wiedmann, M. (2003) Molecular subtyping to detect human listeriosis clusters. *Emerging Infectious Diseases*, **9**, 672–680.

Schmid, W. and Schöne, A. (1997) Some properties of the EWMA control chart in the presence of autocorrelation. *Annals of Statistics*, **25**, 1277–1283.

Schnell, D., Zaidi, A. and Reynolds, G. (1989) A time series analysis of gonorrhea surveillance data. *Statistics in Medicine*, **8**, 343–352.

Scott, S. (2002) Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337–351.

Scott, S., James, G. and Sugar, C. (2004) Hidden Markov models for longitudinal comparisons. Technical report (http://www-rcf.usc.edu/∼ sls/hmm_hs_v2.pdf).

Scranton, R.D., Runger, G.C., Keats, J.B. and Montgomery, D.C. (1996) Efficient shift detection using multivariate exponentially-weighted moving average control charts and principal components. *Quality and Reliability Engineering International*, **12**, 165–171.

Serfling, R.E. (1963) Methods for current statistical analysis of excess pneumonia–influenza deaths. *Public Health Reports*, **78**, 494–506.

Sheehan, T.J., Gershman, S.T., MacDougal, L., Danley, R., Mroszczyk, M., Sorensen, A.M. and Kulldorff, M. (2000) Geographical surveillance of breast cancer screening by tracts, towns and zip codes. *Journal of Public Health Management and Practice*, **6**, 48–57.

Sheehan, T.J., DeChello, L.M., Kulldorff, M., Gregorio, D.I., Gershman, S. and Mroszczyk, M. (2004) The geographic distribution of breast cancer incidence in Massachusetts 1988–1997, adjusted for covariates. *International Journal of Health Geographics*, **3**:17.

Shewhart, W.A. (1931) *Economic Control of Quality of Manufactured Product*. London: Macmillan.

Shiryaev, A.N. (1963) On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, **8**, 22–46.

Siegmund, D. (1985) *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer-Verlag.

Siegmund, D. and Venkatraman, E.S. (1995) Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Annals of Statistics*, **23**, 255–271.

Siegmund, D.O. and Worsley, K.J. (1995) Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Annals of Statistics*, **23**, 608–639.

Silverman, B.W. (1976) Limit theorems for dissociated random variables. *Advances in Applied Probability*, **8**, 806–819.

Skinner, K.R., Montgomery, D.C. and Runger, G.C. (2003) Process monitoring for multiple count data using generalized linear model-based control charts. *International Journal of Production Research*, **41**, 1167–1180.

Smith, A.F. and West, M. (1983) Monitoring renal transplants: an application of the multiprocess Kalman filter. *Biometrics*, **39**, 867–878.

Smith, K.L., DeVos, V., Bryden, H., Price, L.B., Hugh-Jones, M.E. and Keim, P. (2000) Bacillus anthracis diversity in Kruger National Park. *Journal of Clinical Microbiology*, **38**, 3780–3784.

Snow, J. (1854) *On the Mode of Communication of Cholera* (2nd edn). London: Churchill Livingstone.

Sonesson, C. (2003) Evaluations of some exponentially weighted moving average methods. *Journal of Applied Statistics*, **30**, 1115–1133.

Sonesson, C. and Bock, D. (2003) A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society, Series A*, **166**, 5–21.

Song, C. and Kulldorff, M. (2003) Power evaluation of disease clustering tests. *International Journal of Health Geographics*, **2**:9.

Sosin, D.M. (2003) Draft framework for evaluating syndromic surveillance systems. *Journal of Urban Health*, **80**, i8–i13.

Spiegelhalter, D. and Lauritzen, S. (1990) Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579–605.

Spiegelhalter, D.J., Best, N.G., Gilks, W.R. and Inskip, H. (1996) Hepatitis B: a case study in MCMC methods. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

Spiegelhalter, D., Thomas, A. and Best, N. (1999) WinBUGS version 1.2 user manual. Technical report, MRC Biostatistics Unit.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583–640.

Srivastava, M.S. and Wu, Y. (1993) Comparison of EWMA, CUSUM and Shiryaev–Roberts procedures for detecting a shift in the mean. *Annals of Statistics*, **21**, 645–670.

Steiner, S.H., Cook, R.J. and Farewell, V.T. (1999) Monitoring paired binary surgical outcomes using cumulative sum charts. *Statistics in Medicine*, **18**, 69–86.

Stern, H.S. and Cressie, N.A.C. (2000) Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, **19**, 2377–2397.

Stern, L. and Lightfoot, D. (1999) Automated outbreak detection: a quantitative retrospective analysis. *Epidemiology and Infection*, **122**, 103–110.

Steutel, F.W. and van Harn, K. (1979) Discrete analogues of self-decomposability and stability. *Annals of Probability*, **7**, 893–899.

Stone, R.A. (1998) Investigation of excess environmental risk around putative sources: statistical problems and a proposed test. *Statistics in Medicine*, **7**, 649–660.

Stoumbos, Z.G., Reynolds Jr, M.R., Ryan, T.P. and Woodall, W.H. (2000) The state of statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association*, **95**, 992–998.

Stroup, D.F. and Thacker, S.B. (1993) A Bayesian approach to the detection of aberrations in public-health surveillance data. *Epidemiology*, **4**, 435–443.

Stroup, D.F., Thacker, S.B. and Herndon, J.L. (1988) Application of multiple time-series analysis to the estimation of pneumonia and influenza mortality by age 1962–1983. *Statistics in Medicine*, **7**, 1045–1059.

Stroup, D.F., Williamson, G.D., Herndon, J.L. and Karon, J.M. (1989) Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine*, **8**, 323–329.

Stroup, D.F., Wharton, M., Kafadar, K. and Dean, A.G. (1993). Evaluation of a method for detecting aberrations in public health surveillance data. *American Journal of Epidemiology*, **137**, 373–380.

Stroup, D.F., Williamson, D.F., Dean, A.D., Haddad, S., Basha, M. and Rapose, W. (1994) *Statistical Software for Public Health Surveillance*. Atlanta, GA: Centers for Disease Control and Prevention.

Stroup, D.F., Brookmeyer, R. and Kalsbeek, W.D. (2004) Public health surveillance in action: a framework. In R. Brookmeyer and D.F. Stroup (eds), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, pp. 1–35. Oxford: Oxford University Press.

Sudakin, D.L., Horowitz, Z. and Giffin, S. (2002) Regional variation in the incidence of symptomatic pesticide exposures: applications of geographic information systems. *Journal of Toxicology – Clinical Toxicology*, **40**, 767–773.

Sullivan, J.H. and Jones, L.A. (2002) A self-starting control chart for multivariate individual observations. *Technometrics*, **44**, 24–33.

Sun, D., Tsutakawa, R. and Kim, H. (2000) Spatio-temporal interaction with disease mapping. *Statistics in Medicine*, **19**, 2015–2035.

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

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

Teutsch, S.M. and Churchill, R.E. (2000) *Principles and Practice of Public Health Surveillance*, (2nd edition). Oxford: Oxford University Press.

Thacker, S. (1994) 'Historical development', in S. Teutsch and R. Churchill (eds), *Principles and Practice of Public Health Surveillance*. Oxford: Oxford University Press.

Thacker, S.B. (2000) Historical Development. In S.M. Teutsch and R.E. Churchill (eds), *Principles and Practice of Public Health Surveillance* (2nd edn), pp. 1–16. New York: Oxford University Press.

Thacker, S.B. and Berkelman, R.L. (1988) Public health surveillance in the United States. *Epidemiol Reviews*, **10**, 164–190.

Thacker, S.B., Choi, K. and Brachman, P.S. (1983) The surveillance of infectious diseases. *Journal of the American Medical Association*, **249**, 1181–1185.

Thomas, A.J. and Carlin, B.P. (2003) Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. *Statistics in Medicine*, **22**, 113–127.

Timm, N.H. (1996) Multivariate quality control using finite intersection tests. *Journal of Quality Technology*, **28**, 233–243.

Tsui, K.L. and Woodall, W.H. (1993) Multivariate control charts based on loss functions. *Sequential Analysis*, **12**, 79–92.

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

Tsutakawa, R. (1988) Mixed model for analysing geographic variability in mortality rates. *Journal of the American Statistical Association*, **83**, 37–42.

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

United States Department of Agriculture (2001) West Nile virus in equids in the Northeastern United States in 2000. USDA, APHIS, Veterinary Services (http://www.aphis.usda.gov/vs/ceah/wnvreport.pdf).

VanBrackle, L. and Williamson, G.D. (1999) A study of the average run length characteristics of the National Notifiable Diseases Surveillance System. *Statistics in Medicine*, **18**, 3309–3319.

VanEenwyk, J., Bensley, L., McBride, D., Hoskins, R., Solet, D., McKeeman Brown, A., Topiwala, H., Richter, A. and Clark, R. (1999) Addressing community health concerns around SeaTac Airport: Second Report. Washington State Department of Health (http://www.doh.wa.gov/EHSPHL/Epidemiology/NICE/publications/Seatac_Report2.pdf).

Vardeman, S. and Ray, D. (1985) Average run lengths for CUSUM schemes when observations are exponentially distributed. *Technometrics*, **27**, 145–150.

Vidal-Rodeiro, C.L. and Lawson, A.B. (2005) An evaluation of the edge effects in disease map modelling. *Computational Statistics and Data Analysis*. To appear.

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

Wagner, M., Tsui, F.C., Espino, J.U., Dato, V.M., Sittig, D.F., Caruana, R.A., McGinnis, L.F., Deerfield, D.W., Druzdzel, M.J. and Friedsma, D.B. (2001) The emerging science of very early detection of disease outbreaks. *Journal of Public Health Management Practice*, **7**, 51–59.

Wallenstein, S. (1980) A test for detection of clustering over time. *American Journal of Epidemiology*, **111**, 367–72.

Waller, L.A., Turnbull, B.W., Clark, L.C. and Nasca, P. (1992) Chronic disease surveillance and testing of clustering of disease and exposure. *Environmetrics*, **3**, 281–300.

Waller, L., Turnbull, B., Clark, L. and Nasca, P. (1994) Spatial pattern analyses to detect rare disease clusters. In N. Lange, L. Ryan, L. Billiard, D. Brillinger, L. Conquest, and J. Greenhouse (eds), *Case Studies in Biometry*, pp. 3–24. New York: John Wiley & Sons, Inc.

Waller, L.A., Carlin, B.P., Xia, H. and Gelfand, A.E. (1997) Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, **92**, 607–617.

Walsh, S.J. and DeChello, L.M. (2001) Geographical variation in mortality from systemic lupus erythematosus in the United States. *Lupus*, **10**, 637–646.

Walsh, S.J. and Fenster, J.R. (1997) Geographical clustering of mortality from systemic sclerosis in the southeastern United States, 1981–90. *Journal of Rheumatology*, **24**, 2348–2352.

Walter, S.D. (1993) Visual and statistical assessment of spatial clustering in mapped data. *Statistics in Medicine*, **12**, 1275–1291.

Walter, S.D. (1994) A simple test for spatial pattern in regional health data. *Statistics in Medicine*, **13**, 1037–1044.

Ward, M.P. (2001) Blowfly strike in sheep flocks as an example of the use of a time-space scan statistic to control confounding. *Preventive Veterinary Medicine*, **49**, 61–69.

Ward, M.P. (2002) Clustering of reported cases of leptospirosis among dogs in the United States and Canada. *Preventive Veterinary Medicine*, **56**, 215–226.

Watier, L., Richardson, S. and Hubert, B. (1991) A time series construction of an alert threshold with application to S. bovismorbificans in France. *Statistics in Medicine*, **10**, 1493–1509.

Weatherall, J.A.C. and Haskey, J.C. (1976) Surveillance of malformations. *British Medical Journal*, **32**, 39–44.

Weinstock, M.A. (1981) A generalized scan statistic test for the detection of clusters. *International Journal of Epidemiology*, **10**, 289–293.

Wessman, P. (1998) Some principles for surveillance adopted for multivariate processes with a common change point. *Communications in Statistics. Theory and Methods*, **27**, 1143–1161.

Wessman, P. (1999) Studies on the surveillance of univariate and multivariate processes. Ph.D. thesis, Department of Statistics, Göteborg University.

West, M. and Harrison, J. (1989) *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

Wetherill, G.B. and Brown, D.W. (1991) *Statistical Process Control*: Theory and Practice London: Chapman & Hall.

Wharton, M., Price, W., Hoesly, F., Woolard, D., White, K., Greene, C. and McNabb, S. (1993) Evaluation of a method for detecting outbreaks of diseases in 6 states. *American Journal of Preventive Medicine*, **9**, 45–49.

White, C.H. and Keats, J.B. (1996) ARLs and higher order run length moments for Poisson CUSUM. *Journal of Quality Technology*, **28**, 363–369.

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

Wierda, S.J. (1994) Multivariate statistical process control: recent results and directions for future research. *Statistica Neerlandica*, **48**, 147–168.

Wikle, C.K. (2002) Spatial modelling of count data: A case study in modelling breeding bird survey data on large spatial domains. In A.B. Lawson and D.G.T. Denison (eds), *Spatial Cluster Modelling*, Chapter 11. Boca Raton, FL: CRC Press.

Williamson, G.D. and Hudson, G.W. (1999) A monitoring system for detecting aberrations in public health surveillance reports. *Statistics in Medicine*, **18**, 3283–3298.

Wolter, C. (1987) Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen's sets technique. *Methods of Information in Medicine*, **26**, 215–219.

Wong, W.-K. (2004) Data mining for early disease outbreak detection. Ph.D. thesis, Carnegie Mellon University.

Wong, W.-K., A.W. Moore, G. Cooper, and M. Wagner (2002) Rule-based anomaly pattern detection for detecting disease outbreaks. In K. Ford (ed.), *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pp. 217–223. Menlo Park, CA: AAAI Press.

Wong, W.-K., A. Moore, G. Cooper, and M. Wagner (2003) Bayesian network anomaly pattern detection for disease outbreaks. In T. Fawcett and N. Mishra (eds), *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 808–815. Menlo Park, CA: AAAI Press.

Wong, W.K., Moore, A., Cooper, G. and Wagner, M. (2003) WSARE: What's strange about recent events? *Journal of Urban Health*, **80**, i66–i75.

Woodall, W.H. (1997) Control charts based on attribute data: bibliography and review. *Journal of Quality Technology*, **29**, 172–183.

Woodall, W.H. and Montgomery, D.C. (1999) Research issues and ideas in statistical process control. *Journal of Quality Technology*, **31**, 376–386.

Woodall, W.H. and Ncube, M.M. (1985) Multivariate CUSUM quality-control procedures. *Technometrics*, **27**, 285–292.

Wright, S. (1921) Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.

Xia, H. and Carlin, B.P. (1998) Spatio-temporal models with errors in co-variates: Mapping ohio lung cancer mortality. *Statistics in Medicine*, **17**, 2025–2043.

Yashchin, E. (1993) Statistical control schemes – methods, applications and generalizations. *International Statistical Review*, **61**, 41–66.

Yashchin, E. (1994) Monitoring variance components. *Technometrics*, **36**, 379–393.

Yeh, A.B., Lin, D.K.J., Zhou, H.H. and Venkataramani, C. (2003) A multivariate exponentially weighted moving average control chart for monitoring process variability. *Journal of Applied Statistics*, **30**, 507–536.

Yiannakoulias, N., Rowe, B.H., Svenson, L.W., Schopflocher, D.P., Kelly, K., Voaklander, D.C. (2003) Zones of prevention: the geography of fall injuries in the elderly. *Social Science and Medicine*, **57**, 2065–2073.

Zaidi, A., Schnell, D. and Reynolds, G. (1989) Time series analysis of syphillis surveillance data. *Statistics in Medicine*, **8**, 353–362.

Zeger, S.L. (1988) A regression model for time series of counts. *Biometrika*, **75**, 621–629.

Zhu, L. and Carlin, B.P. (2000) Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, **19**, 2265–2278.

Index

Spatial and Syndromic Surveillance for Public Health Edited by A.B. Lawson and K. Kleinman