Spatial and spatio-temporal methods for public health surveillance

Memoria presentada por
Paula Esther Moraga Serrano
para optar al grado de Doctor

Dirigida por
Dr. Francisco Montes Suay
Dr. Francisco Martínez Ruiz

Departament d'Estadística i Investigació Operativa
Universitat de València

Mayo 2012

Francisco Montes Suay, Catedrático de Estadística e Investigación Operativa del Departament d'Estadística i Investigació Operativa de la Universitat de València, Francisco Martínez Ruiz, Profesor Asociado del Departament d'Estadística i Investigació Operativa de la Universitat de València,

CERTIFICAN que la presente memoria, *Métodos espaciales y espacio-temporales para la vigilancia en salud pública*, ha sido realizada bajo su dirección en el marco del Programa de Doctorado *Estadística i Optimització* (Código 130E) del Departament d'Estadística i Investigació Operativa por Paula Esther Moraga Serrano y constituye su Tesis para optar al Grado de Doctor en Matemáticas.

Y para que conste, en cumplimiento de la legislación vigente, la presenta ante la Facultat de Matemàtiques de la Universitat de València, a 28 de mayo de 2012.

Francisco Montes Suay, Catedratico of Statistics and Operations Research of the Department of Statistics and Operations Research at University of Valencia, Francisco Martínez Ruiz, Associate Professor of the Department of Statistics and Operations Research at University of Valencia,

CERTIFY that this memory, *Spatial and spatio-temporal methods for public health surveillance*, has been developed under their supervision at the Department of Statistics and Operations Research by Paula Esther Moraga Serrano and constitute her thesis work for the degree of Doctor of Philosophy in Mathematics.

And to meet the current legislation, she hands it to the Faculty of Mathematics at University of Valencia, on May 28th 2012.

Francisco Montes Suay

Francisco Martínez Ruiz

Acknowledgments

First, I wish to give my sincere gratitude to my supervisor, Paco Montes, for his continuous support, guidance, constructive criticism and understanding in all the time of research and writing of this thesis. I appreciate his wise advice always guiding me in the right direction. Without his help and support, I could not have lived the extraordinary experiences I have had during this time.

I would like to express my sincere thanks to my co-supervisor, Paco Martínez, for his help and suggestions during the preparation of this thesis.

It would have been impossible to write this thesis without the help and guidance of Professors Martin Kulldorff, Andrew Lawson, Al Ozonoff and Jorge Mateu. They helped me come up with the thesis topics and guided me during their development. I would like to express to them my deep and sincere gratitude for their useful suggestions and comments, and for their time and effort in reviewing part of this work.

I take this opportunity to sincerely acknowledge Ibercaja Foundation and ”la Caixa” Foundation for giving me the opportunity to study, research, and live for two years in the United States. I am deeply indebted to the committee members of ”la Caixa” for the confidence they showed me by granting me a fellowship to perform my studies at Harvard. Their support and kind assistance helped me live an unforgettable experience.

I would like to thank Ángel Corberán for his generous support, advice and help with my fellowship applications.

I thank the Department of Population and Medicine and the MIDAS group at Harvard Medical School and Harvard Pilgrim Health Care for their generous help during my visit.

I thank the Division of Biostatistics and Epidemiology for their assistance during my visit at Medical University of South Carolina.

I am very grateful to John Stelling and everyone at the WHO Collaborating Centre for Surveillance of Antimicrobial Resistance at Brigham & Women’s Hospital in Boston, and the Collaborative Group WHONET-Argentina. I am much indebted to John for his valuable guidance and support.

I thank all the professors, classmates and everyone at Harvard School of Public Health and the Department of Biostatistics. It is hard to imagine a more stimulating and encouraging academic environment.

I am thankful to everyone at the Harvard-Brazil Collaborative Public Health Field Course. I was fortunate to have the opportunity to participate and learn on diverse exciting projects as well to meet amazing people and live wonderful experiences.

I would like to thank my mentors Virgilio Gómez Rubio and Barry Rowlingston, for their guidance and help during my Google Summer of Code experience.

I wish to express my gratitude to Bruno Ciancio and everyone at the European Center for Disease Control of Prevention at Stockholm for their guidance and support during my stay.

I thank the many professors, advisors, classmates, colleagues, roommates and friends I met in Cañete, Paiporta, Valencia, Mainz, la Coruña, Charleston, Boston, Brazil and Stockholm, and during my work in Openfinance, Institut Valencià d'Estadística, Registro de Tumores Infantiles and Ayuntamiento de Valencia. I thank them for their care, encouragement and for sharing experiences and knowledge during those times.

I have been extraordinarily lucky to live with Leah, Vilunya, Nadia and Kira during my time in Boston. They welcomed and made feel at home since the first day. They always treated me as one more in the family and shared their lives and traditions with me. Katherine Yih and her family offered me their care since my first days. Thank you all for your joy, enthusiasm, generosity, care, help and for being such wonderful persons.

Todo esto no hubiera sido posible sin el cariño y apoyo de mi familia. Gracias a todos por estar siempre a mi lado, y en especial a mi hermana Maritere por su apoyo, sus consejos y por preocuparse siempre por mí. Gracias a Pepe por todo su apoyo, su confianza en mí y su gran ayuda en la realización de esta tesis. Por último y más importante, gracias mis padres, Bernardo y Antonia, por todo lo que me han enseñado y por todo el amor que me han dado. Para ellos todo lo que hago es motivo de orgullo, y a ellos especialmente les dedico esta tesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Public health surveillance provides information to identify public health problems and respond appropriately when they occur. This information is crucial to prevent and control a variety of health conditions such as infectious diseases, chronic diseases, injuries, or health-related behaviors. Quality surveillance is needed to understand the true health status of the population and to guide the use of limited public health resources. Under inadequate surveillance systems, leaders are grossly misinformed and may lose opportunities for the application of early prevention and control measures. In these situations, it is possible the resurgence of previously eradicated diseases or the uncontrolled global spread of diseases as in the case of HIV/AIDS (M'ikanatha et al., 2007). Surveillance involves four main integrated activities: the collection of health data, the analyses of the data, their interpretation, and the timely dissemination of the results to those responsible to respond to a population's health needs. Surveillance systems capture spatial, temporal and person characteristics on health outcomes. Incidence and mortality rates quantify the size of the health problem in a given population and provide the basis for initiating disease control measures and evaluating their effectiveness. Temporal trends and demographic and ethnic group comparisons can provide important clues as to disease etiology.

The increased availability of geographically geocoded health and population data, and the development of geographic information systems (GIS) and software for geocoding addresses, have facilitated the ascent of the investigations of spatial and spatio-temporal variations of disease. John Snow's cholera-outbreak investigation in London in 1854 provides one of the most famous examples of spatial analysis. Snow used a spot map to illustrate how cholera deaths appeared to be clustered around a public water pump. The assessment of the spatial pattern of the cholera cases was important in identifying the source of the infection and gave support to the theory of cholera transmission through drinking water. There is a wide range of spatial and spatio-temporal methods that can be applied as a surveillance tool including disease mapping, clustering, and geographic correlation studies. Many of these methods may be used for highlighting areas at high risk, detecting significant disease clusters in space and time, early detection of epidemics, assessing disease risk in relation to a putative source, and identifying disease risk factors. Unfortunately, naive use of the statistical methods can be highly misleading. Therefore, a thorough understanding of potential problems such as changes in case definitions and completeness issues are critical to the analysis of the data and interpretation of the findings.

Over the past few decades, surveillance has undergone considerable development. Certain activities have contributed to the advance of public health surveillance (Brookmeyer and Stroup, 2004). These include technological innovations such as real-time on-line monitoring and advances in GIS, the development of new statistical methods and computational tools to apply them, and more effective use of electronic media and other tools of communications that facilitate dissemination of surveillance information for public health practice. Also, public health surveillance has changed in response to new public health concerns, such as bioterrorist events and relatively new diseases and epidemics, such as severe acute respiratory syndrome (SARS). As public health needs change and new tools and increased computational capacity of computers become available, statistical methods for disease surveillance must continue to evolve to improve the quality of the analyses, and the interpretation and display of the results in the most useful form and appropriate time-frame to meet the interests of policymakers and stakeholders. The aim of this thesis is to propose new techniques for helping public health surveillance practice. In particular, we focus in spatial and spatio-temporal methods that can help deal with missing data (Chapter 4), model the correlated heterogeneity in disease mapping (Chapter 5), detect disease clusters (Chapter 6), and elucidate spatial variations in temporal trends (Chapter 7).

We begin with an overview of public health surveillance and spatial data. Chapter 2 provides an introduction in surveillance systems, as well as a review of statistical methods that have been applied for disease surveillance. Methods considered include computation of rates, temporal trends, detection of clusters and outbreaks, and disease mapping. Spatial data are generally classified in three major types: lattice data, geostatistical data, and point patterns. Chapter 3 is devoted to the review of their basic characteristics and analysis methods. The goal of these two chapters is to provide some ground to the concepts and statistical methods used in surveillance that can help the development of the subsequent chapters.

Chapters 4-7 are based on particular questions of interest. Chapter 4 discusses the problem of missing data and focuses in the analysis of the all-cause and pneumonia and influenza (P&I) mortality data in the United States. National estimates of the all-cause and P&I mortality burden derived from these data treat all missing values as zero counts. The effect of this methodological decision is to bias estimates downward and produce underestimates of the true mortality burden. To evaluate the impact of this treatment of missing data on national estimates, we propose a regression-based procedure that utilizes relevant information to impute missing values and thus produce a more accurate estimate of mortality. We considered and evaluated several model specifications to predict weekly death counts by city, calendar week, calendar year, and age group. We cross-validated these models and calculated revised all-cause and P&I mortality estimates by imputing the missing data and estimating P&I excess mortality using the regression approach recommended by CDC (Serfling, 1963). We then compared these estimates with an without imputation to understand the impact on national estimates of this treatment of unreported data.

Chapter 5 is devoted to disease mapping. Hierarchical Bayesian models involving conditional autoregression (CAR) components are commonly used in disease mapping (Besag et al., 1991). An alternative model to the proper or improper CAR is the Gaussian component mixture (GCM) model (Langford et al., 1999). A review of CAR and GCM models is provided in univariate settings where only one disease is considered, and also in multivariate situations where in addition to the spatial dependence between regions, the dependence among multiple diseases is analyzed. A performance comparison between models using a set of simulated data is reported. Moreover, GCM and CAR models are applied for estimating the relative risk of low birth weight in Georgia, U.S., in the year 2000.

Detection of disease clusters is an important tool in epidemiology that can help to identify risk factors associated with the disease and in understanding its etiology. In Chapter 6 we propose a method for the detection of spatial clusters where the locations of a set of cases and a set of controls are available. The method is based on local indicators of spatial association functions (LISA) (Anselin, 1995), particularly on the development of a local version of the product density, which is a second-order characteristic of spatial point processes. The behaviour of the method is evaluated and compared with Kulldorff's spatial scan statistic (Kulldorff and Nagarwalla, 1995) by means of a simulation study. Both methods are applied to detecting spatial clusters of kidney disease in the city of Valencia, Spain, in the year 2008.

Methods for the assessment of spatial variations in temporal trends (SVTT) are important tools for disease surveillance which can help governments to formulate programmes to prevent diseases, and measure the progress, impact, and efficacy of preventive efforts already in operation. The linear SVTT method is designed to detect areas with unusual different disease linear trends (Kulldorff, 2010). In some situations, however, the bad fit of its estimation trend procedure can lead to wrong conclusions. In Chapter 7, the quadratic SVTT method is proposed as alternative of the linear SVTT method. A performance comparison between the linear and quadratic methods using a set of simulated data is provided to help illustrate their respective properties. The quadratic method is applied to detect unusual different cervical cancer trends in white women in the United States, over the period 1969 to

1995.

Finally, in Chapter 8, a discussion about the methods proposed, directions for future research lines, and some final remarks about public health surveillance are presented.

Chapter 2

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 in the surveillance chain is the application of these data to prevention and control.” (Thacker and Berkelman, 1988).

Public health surveillance information is essential to monitor the health status of a population, and allow managers to respond quickly to a population's health needs. For example, such information can be used to:

- Estimate the magnitude of a health problem
- Portray the natural history of a disease
- Determine the distribution and spread of a health event
- Detect epidemics or new syndromes
- Monitor changes in infectious agents
- Detect changes in health practices and behaviors
- Generate and evaluate hypotheses, and stimulate public health research
- Planning public health actions and use of resources
- Project future needs
- Evaluate control and prevention measures

Comprehensive surveillance systems and appropriate statistical methods are essential for disease surveillance practice. Surveillance systems collect data for a variety of health conditions, including chronic diseases, birth defects, injuries, nosocomial or hospital-acquired infections, emergency room visits and health behaviors. Conducting surveillance requires four basic activities: collection, analysis, interpretation and dissemination of health data. In the collection process, it is important to determine whether the total population in a public health jurisdiction is under surveillance (population-based) or if surveillance will occur at a group of facilities or sentinel sites. Mandatory disease notification is the primary method of collecting disease information. Other surveillance methods include vital records, sentinel surveillance, surveys, registries, or syndromic surveillance.

Analysis of data provides important information on which to base action. These are useful to documenting the magnitude of a health event, and describing it in terms of the personal characteristics of those at risk, and the place and timing of occurrence. A key factor in the selection of methods of analysis is the objective of surveillance. This could be outbreak detection, trend monitoring, comparison between different populations, detection of unusual aggregations of cases, or identification of factors associated with the spatial distribution of the disease. A careful interpretation of the results allow public health officers to allocate resources efficiently and targeting populations for education or preventive programs.

2.1 Surveillance systems

A number of surveillance systems are used routinely by public health departments at local and national levels. The process of data collection can be passive or active. Data in passive surveillance are reported in such a way that the receiving agency waits for data reports to be sent in by health care providers or laboratories. On the other hand, active surveillance occurs when the health department requests information about conditions or diseases to identify possible cases. An active surveillance system provides stimulus to health care workers in the form of individual feedback or other incentives. Although passive surveillance may suffer from incomplete data due to underreporting, compromised accuracy and show selection bias depending on the source of reports or laboratory specimens, it is seen in standard systems that report notifiable diseases to a public health department because it requires substantially less time and resources than active surveillance. Since active surveillance can produce early, timely and complete information, it is especially useful when there is a need to identify all cases, for example during disease outbreaks. Many sources of data can be used for public health surveillance, including the following (Dicker et al., 2006):

- Vital statistics which consists on records of birth, death, marriage, and divorce.

These records are a critical component for public health practice. For example, mortality rate has long been used as indicator of overall population health. Also, the monitoring of preterm birth is important since it is a risk factor for a variety of adverse health outcomes.

- Notifiable disease reports which inform of conditions established by law and are made on the basis of the symptoms alone. Because reports are made without waiting for laboratory confirmation, they do not necessarily indicate the presence of the disease itself.
- Laboratory data for surveillance for many diseases, including diseases caused by virus, bacteria and other pathogens.
- Hospital discharge records which typically include demographic data, diagnoses, operative procedures and length of stay, and exclude personal information which could identify individuals.
- Surveys that sample the health status of citizens representative of the whole population. These are especially useful for monitoring chronic diseases and health-related behaviors.
- Sentinel systems which consist of a pre-arranged sample of reporting sites that collect all cases of a certain condition, such as influenza or certain bacterial infections among children. Sentinel surveillance is a good way to use limited resources to indicate trends in the entire population.
- Data on indicators of disease or of disease potential. For example, zoonotic diseases surveillance which involves the detection of animals infected with diseases that can be transmitted to humans. Or environmental surveillance which focuses on the detection of contamination, radiation, or other conditions in nature that might favor animal populations that may be reservoirs or vectors of disease.
- Registries used for particular conditions, such as cancer and birth defects.
- Adverse health events which may detect potential safety problems of approved drugs and other therapeutic agents.
- Syndromic surveillance systems which use clinical information about disease signs and symptoms that might be suggestive of disease. These systems are especially useful for the detection of adverse effects at the earliest possible time, possibly even before disease diagnoses can be confirmed through unmistakable signs or laboratory confirmation.

Surveillance systems may suffer from limitations that compromise their usefulness. Underreporting, lack of representativeness, lack of timeliness, and inconsistency of case definitions may sometimes be present. In these situations, methodology must be carefully developed and results interpreted, and it is necessary to have a good understanding of the strengths and weaknesses of the data collection methods, the reporting process, and the changes in the surveillance system and practices.

Systems should be regularly evaluated to promote the best use of public health resources, and ensure that problems of public health importance are being monitored efficiently and effectively. The evaluation of surveillance systems should assess whether a system is serving a useful public health function and is meeting the system's objectives, and should include recommendations for improving quality and efficiency (MMWR, 2001). The evaluation should involve an assessment of system attributes, including simplicity, flexibility, data quality, acceptability, sensitivity, positive predictive value (PPV), representativeness, timeliness, and stability. Because each surveillance system is unique and attributes that are important to one system might be less important to another, the evaluation must consider those characteristics of the system that are of the highest priority to achieve its intended purpose and objectives.

2.2 Analysis and interpretation

The process of analysis and interpretation of surveillance data encompasses a broad variety of system designs, analytic methods, modes of presentation, and interpretive uses (Lee et al., 2010). In general, descriptive methods are the basis of routine reporting of surveillance data. These focus on the observed patterns in the data and might also seek to compare the relative occurrence of disease in different subgroups. More specialized hypotheses are explored using inferential methods. The aim of these methods is to make statistical conclusions about the patterns or outcomes of disease. A thorough understanding of the underlying data, the data collection process, and the analytical methods used are critical to the interpretation of the findings. To avoid artifactual inferences, care must be given to changes in reporting procedures or case definitions, misdiagnosis, delay in reporting and increase in reporting due to improved awareness, better laboratory tests or new diagnostic procedures (CDC, 2010).

Surveillance data are examined by characteristics of time, place, and person. The most commonly collected person characteristics are age and gender. The examination of the distribution of health events by these characteristics provides information about the burden of disease in different populations and can reveal important disease trends. Depending on the health problem, other characteristics such as race, ethnicity, occupation, socio-economic status, recent hospitalization, sexual orientation or immunization status can provide information useful for disease control and prevention. Unfortunately, these data are less consistently available for analysis. Time factors may include the time of year, day, progression over time, or the speed of development of disease. Analysis of surveillance data by time is usually conducted to describe the distribution of cases over time, characterize trends, compare the number of cases reported in a particular time period of interest to the number of cases reported during preceding time periods, and detect changes in disease incidence and disease outbreaks. Health departments usually analyze surveillance data by cities, counties, states, or other geographic areas. Maps are helpful in showing the geographic distribution of cases, to facilitate recognition of spatial associations in the data, detect clusters and assessing the geographic relationships of risk factors and disease risk.

2.2.1 Rates

The analysis of disease in a population begins by addressing the occurrence of a particular health event in a particular population over a particular time. Crude rates are used when a summary measure is needed and it is not necessary or desirable to adjust for other factors (DOH, 2010). A crude rate is calculated by dividing the total number of events in a specified time period by the total number of individuals in the population who are at risk for these events and multiplying by a constant, such as 100,000. Much of public health assessment involves describing the health status of a given population over time or comparing health events across populations. In making these comparisons, we need to account for the fact that the occurrence of many health conditions depends on risk factors such as age. Standardization offers a mechanism to adjust rates and remove the effect of known confounding factors. There are two types of standardization methods: direct and indirect.

Directly age-standardized rates provide, for each population, an indicator in terms of the overall rate that would have occurred in an arbitrary external or standard population if it had the age-specific rates of the observed population. The standardized rate is a weighted average of stratum specific rates of the populations to be compared, where the weights represent the relative age distribution of the standard population. Thus, the directly age-standardized rate can be expressed as follows:

$$\sum_{i=1}^{m} r_i \frac{n_i^{(s)}}{\sum_i n_i^{(s)}},$$

where m is the number of age groups, r_i is the rate in age group i in the observed population, and $n_i^{(s)}$ is the population in the ith age group of the standard population. Since all the directly age-standardized rates are based on the same set of weights, they can be readily compared to each other and to the standard population. However, it is important to note that since they are based on an external standard population, they do not reflect the absolute frequency of the event in a population and should be used only for the purpose of comparison, the interest lie on the relative ranking of the rates.

An alternative approach is the indirect standardization. This method is generally used when the number of events is relatively small and the age-specific rates are unstable but the age-specific rates are available for a standard population. Indirect standardization uses the age-specific rates from a standard population (usually the relevant national or regional population) and applies them to the population distribution of the observed population to calculate the total number of events one would expect if the observed population behaved the way the standard population behaved. The number of cases expected can be expressed as

$$E = \sum_{i=1}^{m} r_i^{(s)} n_i,$$

where $r_i^{(s)}$ is the rate in age group i in the standard population, and n_i is the population in stratum i of the observed population. The standardized mortality ratio or SMR is then calculated as the ratio of the total number of observed events, Y, to the total number of expected events:

$$SMR = \frac{Y}{E}.$$

When applied to incidence data it is commonly known as the standardized incidence ratio or SIR. Ratios greater than 1 indicate more cases observed in the observed population than expected from the standard population. In some instances the SMR is multiplied by the overall crude rate of the standard population and presented as an indirectly standardized rate. Unlike directly standardized rates, indirect standardized rates of each observed population are based on its own set of weights. This makes the comparison of indirect standardized rates problematic. In fact, the only comparisons that are always possible are comparisons between each observed population and the standard population.

2.2.2 Trends

Public health agencies have a long tradition of monitoring disease incidence and mortality rates over time and across demographic subgroups. Studies of disease trends are typically used to determine whether incidence and mortality have increased or decreased over time, and to assess the speed with which the changes have occurred. Trends in incidence or mortality rates over time are often characterized by the annual percent change (APC). This approach assumes that the rates change at a constant percentage of the rate of the previous year. Specifically, the APC is estimated by fitting first a linear model where the logarithm of the yearly rates is regressed on time. That is,

$$log(r_t) = \beta_0 + \beta_1 t,$$

where, t denotes year and r_t represent the observed rates. Then, the APC from year t to year $t+1$ is calculated using a transformation of the slope of the trend line:

$$\begin{aligned} \text{APC} &= 100 \times \frac{r_{t+1} - r_t}{r_t} = 100 \times \frac{exp(\beta_0 + \beta_1(t+1)) - exp(\beta_0 + \beta_1 t)}{exp(\beta_0 + \beta_1 t)} \\ &= 100 \times (exp(\beta_1) - 1). \end{aligned}$$

For example, an APC of 1% indicates that the annual rate is increasing on average by 1% a year. That is, if the estimated APC is 1% and the rate is 50,000 per 100,000 in year y, the rate is 50,000 × 1.01 per 100,000 = 50,500 per 100,000 in $y+1$, and 50,500 × 1.01 per 100,000 = 51,005 per 100,000 in $y+2$. A negative APC describes a decreasing trend whereas a positive APC describes an increasing trend.

The APC is easy to calculate and interpret as a measure of disease trends over short time periods (Li et al., 2008). However, for long time periods, it is not always reasonable to expect that a single APC can accurately characterize the trend over the entire time period of interest. In some situations, the linearity of rates on the logarithmic scale, implying a constant rate of change, may not apply over the entire series of data (Clegg et al., 2009). For example, disease rates may drop sharply for a period of several years, drop gradually for several years after that, and then rise gradually for the next several years. When the assumption of a constant rate of change does not hold over the entire time interval, the trend may be described using the average annual percent change (AAPC). The AAPC is derived by first finding the underlying joinpoint model that best fits the data. This allows the determination of when and how often the APC changes. The model for the observations $(t_1, r_1), \ldots, (t_n, r_n)$ where the r_i's denote the rates observed at the time points $t_1 \leq \ldots \leq t_n$ is written as

$$E[r|t] = \beta_0 + \beta_1 t + \delta_1 (t - \tau_1)^+ + \ldots + \delta_k (t - \tau_k)^+,$$

where k is the unknown number of joinpoints, the τ_k's are the unknown joinpoints and $a^+ = a$ for $a > 0$ and 0 otherwise (Kim et al., 2000). The optimal number of joinpoints may be selected using the permutation test which uses a sequence of permutation tests to determine the true number of joinpoints, or the Bayesian Information Criterion (BIC) which finds the model with the best fit by penalizing the cost of extra parameters. The joinpoint model is fitted using joined log-linear segments, so each segment have an associated APC. The APC of segment i is

$$\text{APC}_i = 100 \times (exp(\beta_i) - 1).$$

The AAPC is computed as a weighted average of the slope coefficients for each segment, β_i, with the weights w_i equal to the length of each segment over the interval. Specifically,

$$\text{AAPC} = 100 \times \left(exp\left(\frac{\sum w_i \beta_i}{\sum w_i} \right) - 1 \right).$$

Thus, whereas the APC's for each joinpoint segment provides a complete characterization of the trend over time, the AAPC provides a summary measure of the trend over a fixed interval.

2.2.3 Cluster detection

Conceptually, a cluster is an unusual collection of cases which are close to each other, in the sense that it differs from what would be expected in the absence of exposure to risk factors or to the effects of transmission, and after accounting for the heterogeneous density of the at risk population (Waller and Gotway, 2004). Cluster analysis has attracted great interest in the field of public health, and various techniques have been developed for evaluating whether the incidence of a disease shows a particular tendency to group together. The study of potential clusters allows an evaluation of the possible relationship between a disease and risk factors such as environmental sources of contamination, genetic factors or socio-economic factors. Disease clustering is classified into temporal, spatial, and space-time clustering. Whereas temporal or spatial clustering evaluates whether cases tend to be located close to each other in time or space, respectively, space-time clustering examines the question of whether cases that are close in space are also close in time and vice versa, after adjusting for purely spatial and purely temporal clustering (Pfeiffer et al., 2008). To investigate clustering, many different tests have been proposed for different purposes (Kulldorff, 2006; Tango, 2010). The different methods are classified as either global or local. Global clustering methods are used to assess whether a global tendency for the disease to group together is apparent throughout the study region but do not identify the location of clusters. This type of methods are appropriate, for example, for finding evidence of whether a disease is infectious or not. Local methods are used to detect the locations and extent of clusters, and can be further divided into focused and non-focused tests. Non-focused tests identify the location of all likely clusters in the study region, whereas focused tests investigate whether there is aggregation around a pre-determined source of risk.

The random labeling procedure offers an approach for assessing clustering in case-control studies where the data consist of n locations of all known cases of a disease within a given geographical region over a specified time-period, together with the location m of a set of controls, defined to be a random sample of the population at risk. The null hypothesis of no disease clustering is equivalent to the cases being an independent random sample from the superposition of the cases and controls. That is, conditional on the $n + m$ locations, all possible labelings into n cases and m controls are equally likely. To test this random labeling hypothesis Cuzick and Edwards (1990) proposed a test statistic based on the nearest-neighbor properties in case-control point data. Specifically, they consider a test statistic which represents the number of k nearest-neighbors of cases that are also cases. The rank of the test statistic based on the data observed among the values based on the randomly labeled data allows calculation of the p-value associated to the test. The result can be sensitive to the choice of k, possibly indicating the scale of any clustering observed. In case-control studies we are dealing with a bivariate process, cases and controls corresponding to events of type 1 and 2, respectively. A very useful function for estimating the second-order properties of the process that gave rise to the data is the reduced second moment measure or K-function (Ripley, 1976). The K-function is invariant to the random thinning of the process. From this, it follows that for any bivariate stationary process generated by random labeling of a univariate process into events of type 1 and 2, the difference between K-functions of type 1 and type 2 events at any distance should be 0. This motivates the use of an estimate of the difference between K-functions as the basis for a test of spatial clustering.

The most widely used test for space-time clustering is the Knox test (Knox, 1964). By specifying a spatial and a temporal critical distances, the method determines the number of pairs of cases which are simultaneously close in space and time. A significantly large number would indicate evidence of space-time clustering of the disease. Two principal limitations of the Knox test are its dependence on the subjective choice of the spatial and temporal critical distances, and the bias that will occur if the population increases or decreases with different rates in different geographical areas (Kulldorff and Hjalmars, 1999). Mantel (1967) proposed a test that compares inter-event distances in space and time against a null hypothesis that time and space distances are independent. The sum, across all pairs of cases, of the spatial distances multiplied by the time distances is computed and a transformation is used as test statistic to reduce the effects of large space and time distances. Jacquez (1996), on the other hand, developed a test for space-time interaction which uses the observed number of pairs of cases close in both space and time, where the measure of closeness is defined by the k nearest neighbors. Diggle et al. (1995) extended second-order analysis methods for spatial data, and calculated the difference between the K-function in space and time and the product of the K-function in space and that in time. A significantly positive difference would indicate evidence of spatio-temporal clustering of the disease.

These tests are global tests in that they test for clustering throughout the data without identifying specific clusters. That is, they are designed for evaluating whether cases tend to come in groups or are located close to each other no matter when and where they occur. Methods for the detection of clusters, on the other hand, are designed for detecting and localizing specific clusters and evaluating their significance. The scan statistic (Kulldorff and Nagarwalla, 1995; Kulldorff et al., 1998) has been implemented as a major analytical tool for cluster detection in a spatial, temporal, and space-time setting. A statistical package, `SaTScan`, facilitates its use and can be downloaded from http://www.satscan.org/.

Scan statistics

Scan statistics scan the study region with a huge number of overlapping windows and determine the windows which group together an unusual number of cases, adjusting for multiple testing (Kulldorff et al., 1998). The collection of windows depend on the application. Typically, the spatial version uses circular windows with radius varying continuously from zero to some upper limit such that it does not pass beyond 50% of the at risk population. In the space-time setting, windows are typically defined as cylinders. The circular base defines a geographical area and the height the time period. For each choice of base all choices of the temporal height are considered and vice versa, so that the scanning is done over short and fat cylinders, tall and thin cylinders, and everything in between.

Conditioning on the observed total number of cases C, the scan test statistic S is defined as the maximum likelihood ratio over all possible windows Z. The likelihood ratio S is expressed as

$$S = \frac{max_Z L(Z)}{L_0} = max_Z \frac{L(Z)}{L_0},$$

where $L(Z)$ is the maximum likelihood for window Z, and L_0 is the likelihood function under the null hypothesis which states that the probability of being a case inside Z is equal to the probability of being a case outside Z. The mathematical formulation of S depends on the probability model used. A Poisson model is used for data where the number of events are Poisson distributed, a Bernoulli model for case-control type data (Kulldorff, 1997), an ordinal model for ordinal data, an exponential model for survival data, and a space-time permutation model for looking at space-time interaction clusters when only case data is available (Kulldorff et al., 2005). Let c be the observed number of cases within a window Z. For the Poisson model, let $\mu(Z)$ be the covariate adjusted expected number of cases under the null hypothesis. Then, the ratio $L(Z)/L_0$ for a specific window is

$$\left(\frac{c}{\mu(Z)}\right)^c \left(\frac{C-c}{C-\mu(Z)}\right)^{C-c},$$

if $c > \mu(Z)$, and 1 otherwise.

The window with the maximum likelihood constitutes the most likely cluster, the cluster that is least likely to have occurred by chance. Its statistical significance is obtained through Monte Carlo hypothesis testing (Dwass, 1957). Thus, the previous procedure is repeated for a large number of replicas of data generated under the null hypothesis, say R, and their respective test statistics are calculated. The test statistic of the observed data is combined with these, and the set of the $R+1$ values are ordered. If M is the rank of the observed test statistic, a p-value equal to $M/(R+1)$ would be obtained.

Apart from the most likely cluster, secondary clusters can also be identified, ordered according to the value of S. There will always be a secondary cluster which is almost identical to the most likely one and with almost the same likelihood, expanding or reducing the size of the initial cluster, but clusters of this type provide little additional information. Normally the option chosen is to show the secondary clusters which do not overlap with the most probable cluster, as they can be of greater interest.

2.2.4 Outbreak detection

New concerns about the possible threat of bioterrorism, the emergence of new infectious diseases, and the increasing availability of electronic health data, have lead to a growing interest in methods of surveillance for the early detection of outbreaks (MMWR, 2004). An outbreak is commonly defined as an increase in cases of disease in time or place that is greater than expected. The number of cases indicating presence of an outbreak vary according to the disease, the population, previous exposure to the disease, and time and place of occurrence. For example, if a condition is rare or has serious public health implications, an outbreak may involve only one case. Methods for the early identification of unusual health events which give some indication of the location and shape of the disturbance, are crucial to allow interventions to control and diffuse the source of the disturbance, inhibiting further spread to the population (Lawson and Kleinman, 2005).

A broad range of statistical techniques have been proposed for outbreak detection. These include regression models such as the Serfling's method (Serfling, 1963), time series approaches such as Box-Jenkins models (Box and Jenkins, 1970) and hidden Markov models (Strat and Carrat, 1999), and techniques inspired by statistical process control such as the Shewhart chart (Shewhart, 1931) the cumulative sum (CUSUM) (Page, 1954), and the exponentially weighted moving average (EWMA) control chart (Roberts, 1959). Surveillance systems that use spatial information are important to enable the detection of small localised disease outbreaks. If this information is not used, the impact of a localized outbreak may be diluted through its combination with global data and outbreaks will be missed. Examples of methods that incorporate spatial information include the spatial CUSUM (Raubertas, 1989), and the space-time permutation scan statistic (Kulldorff et al., 2005). A review of the methods for the detection of disease outbreaks can be found in Unkel et al. (2012). Here, we briefly describe the Serfling's method, which has been extensively used for the detection of influenza epidemics, and the space-time permutation scan statistic, which can be used for prospective outbreak detection of disease when population-at-risk information is unavailable.

Serfling's method

Serfling's method (Serfling, 1963) represents a simple approach for the detection of epidemics of influenza-like syndromes. It is the approach the Centers for Disease Control and Prevention (CDC) use to determine epidemic influenza activity and excess mortality attributed to influenza. Influenza is a serious viral infection that annually causes substantial burden of morbidity and mortality (Molinari et al., 2007). Estimates of influenza-associated deaths are important to determine costs and benefits of influenza prevention and control strategies and in preparing for both seasonal epidemics and future pandemics (Thompson et al., 2009). However, estimating the disease burden of influenza is challenging. Influenza infections are often not laboratory confirmed and deaths from influenza are often due to secondary complications that occur after the primary viral infection, such as pneumonia or worsening of chronic health conditions. Thus, many resulting deaths from influenza are not recorded as such and we must instead estimate the disease burden using statistical models applied to non-specific disease outcomes such as pneumonia and influenza (P&I), respiratory and cardiovascular or all-cause mortality (Newall et al., 2010; Muscatello et al., 2008).

In temperate regions, excess mortality occurring during winter months is associated with pandemic and seasonal epidemics of influenza. Consequently, most measures of influenza-attributable disease burden are estimates based on calculating the number of deaths occurring in excess of the number expected if influenza viruses were not circulating (Brammer et al., 2009). Estimates can be calculated using a variety of mathematical models, one of the most common being Serfling's model. This model uses five years P&I mortality data to calculate a non-epidemic seasonal baseline. The model contains terms for intercept, linear trend, and a pair of harmonic terms to capture the underlying sinusoidal behavior of seasonal influenza. It is formulated as:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 sin\left(\frac{2\pi t}{52}\right) + \beta_4 cos\left(\frac{2\pi t}{52}\right) + \epsilon_t,$$

where Y_t denotes observed weekly P&I in the at week t, β_1 and β_2 represent coefficients associated with temporal trends in deaths, β_3 and β_4 represent coefficients associated with seasonal fluctuations, and ϵ_t is a normally distributed error term.

Since the model estimates the non-epidemic seasonal baseline, the cooler weeks in which influenza circulates more widely in temperate settings are excluded to avoid biased parameter estimates (Ozonoff et al., 2006). The influenza-attributable deaths are determined by subtracting estimated seasonal baseline deaths from the observed deaths. The upper limit of a confidence band around the baseline is used to determine the epidemic threshold for that time period. As soon as this threshold is exceeded, an outbreak is declared.

An alternative method has been to use Poisson regression models. These models usually include coefficients similar to those described for Serfling's models as well as additional terms corresponding to influenza virus circulation such as percentages of laboratory specimens testing positive for influenza A(H3N2), A(H1N1), and B viruses, and respiratory syncytial virus (RSV) activity to explicitly account for increased winter mortality not caused by influenza virus (Thompson et al., 2009).

Space-time permutation scan statistic

The space-time permutation scan statistic (Kulldorff et al., 2005), which uses only case data, and applies minimal assumptions concerning the time and the geographical characteristics of the potential outbreaks, is increasingly used for the early detection of disease outbreaks. For example, this method has been applied to laboratory data by the WHO Collaborating Centre for Surveillance of Antimicrobial Resistance group to detect outbreaks of public health importance in the community and hospital settings (Stelling et al., 2010). The space-time permutation scan statistic does not require population-at-risk data, and can be used when only the number of cases is available. It is, therefore, very useful for surveillance of emergency department visits and pharmacy sales where the catchment area for each hospital/pharmacy is undefined and the population-at-risk is unknown.

In this method, as in the space-time scan statistic for retrospective detection of clusters, a search window is gradually moved across space and time looking for potential outbreaks. The scanning window is a cylinder, whose base and height represent the geographical area and the number of days, respectively, of the potential outbreak. The last day is always included in the cylinder. Suppose c_{zd} is the observed number of cases in area z during day d, and C is the total number of observed cases. Since population-at-risk is not available, the expected cases are calculated using only the cases. Thus, for each area and day, the expected number of cases μ_{zd} is calculated conditioning on the observed marginals as the proportion of all cases that occurred in area z times the total number of cases during day d,

$$\mu_{zd} = \frac{1}{C} \sum_{z} c_{zd} \sum_{d} c_{zd},$$

and, for a given particular cylinder A, the expected number of cases μ_A is equal to:

$$\mu_A = \sum_{(z,d) \in A} \mu_{zd}.$$

When both $\sum_{z\in A} c_{zd}$ and $\sum_{d\in A} c_{zd}$ are small compared to C, the observed number of cases in the cylinder c_A is approximately Poisson distributed with mean μ_A. Based on this approximation, the Poisson generalized likelihood ratio (GLR) is used as a measure of the evidence that cylinder A contains an outbreak:

$$\left(\frac{c_A}{\mu_A}\right)^{c_A}\left(\frac{C-c_A}{C-\mu_A}\right)^{C-c_A}.$$

The cylinder with the maximum GLR constitutes the primary outbreak, that is, the least likely to be a chance occurrence. To evaluate its significance, a large number of random permutations are created by shuffling the dates and times and assigning them to the original set of case locations, and ensuring that both the spatial and temporal marginals are unchanged. The most likely outbreak is then calculated for each simulated dataset and the significance is obtained using Monte Carlo hypothesis testing.

2.2.5 Disease mapping

The mapping of disease risk has a long history in public health surveillance. Disease maps provide a rapid visual summary of spatial information and allow the identification of patterns that may be missed in tabular presentations (Elliott and Wartenberg, 2004). Such maps are crucial for describing the spatial variation of the disease, identifying areas of unusually high risk, formulating etiological hypotheses, and allowing better resource allocation. The aim is to obtain low variance estimates of disease risk within geographic units that are as small as possible. These estimates are generally based on counts of the observed cases and the number of individuals at risk. And, possibly, also on covariate information such as the age distribution, lifestyle, enviromental and genetic factors.

The disease risk is often estimated by the raw standardized mortality (or morbidity) ratio (SMR), calculated as the ratio of observed disease cases versus the number of expected cases. However, the SMRs are often misleading and insufficiently reliable for reporting in areas with small populations. In contrast, model-based approaches enable to borrow information from neighboring areas to improve local estimates, resulting in the smoothing of extreme rates based on small sample sizes (Gelfand et al., 2010). Such approaches are often expressed as hierarchichal Bayesian disease mapping models which are readily implemented via Markov chain Monte Carlo (MCMC) algorithms.

Bayesian disease mapping models treat the disease risks $\{\theta_i\}$, in small areas indexed by i, as random variables and specify a distribution for them. A natural model for disease mapping is the following three level hierarchical model:

$$Y_i \sim Po(E_i \times \theta_i) \quad i = 1, \ldots, n;$$

$$log(\theta_i) \sim p(\cdot|\phi),$$
$$\phi \sim \pi(),$$

where Y_i and E_i are respectively the observed and the expected number of cases of disease in area i, θ_i is the relative risk in area i, $p(\cdot|\phi)$ is an appropriate prior distribution for the $\{\theta_i\}$ and ϕ are hyperparameters with hyperprior distributions $\pi()$ (Lawson, 2009a). The distribution specified for $\{\theta_i\}$ is referred to as the second hierarchical level of the model to distinguish it from the first-level distribution that pertains to the random sampling variability of the observed counts about their local mean. It is at this second level that the spatial dependence between the relative risks is introduced. We term this correlated heterogeneity (CH). To increase flexibility an unstructured exchangeable component that models uncorrelated noise (UH) can be included as well as other terms which may include covariates and/or trend effects. Often a choice is made where either $log(\theta_i)$ is considered to be modeled with a linear predictor, or it is given a distribution that then has correlation built in to it. In the former case, a typical log linear model could take the form:

$$log(\theta_i) = \alpha_0 + UH + CH,$$

where prior distributions are assumed for the α_0, UH and CH components.

These hierarchical models allow straightforward extensions to estimate covariate effects, predict missing data and handle spatio-temporal data and multiple diseases. In the space-time setting, for example, the disease count Y_{ij} observed in the area i and time period j, may be modeled as

$$Y_{ij} \sim Po(E_{ij} \times \theta_{ij}),$$

where θ_{ij} is the risk and E_{ij} is the expected number of cases in the given area and period of time. Then, three groups of components for $log(\theta_{ij})$ can be considered:

$$log(\theta_{ij}) = \alpha_0 + A_i + B_j + C_{ij},$$

where A_i is the spatial group, B_j is the temporal group, and C_{ij} is the space-time interaction group (Lawson, 2009b). For example, in Bernardinelli et al. (1995) these groups are defined as follows: $A_i = \phi_i$, $B_j = \beta t_j$ and $C_{ij} = \delta_i t_j$ where ϕ_i is an area random effect, βt_j is a linear trend term in time t_j, and $\delta_i t_j$ is an interaction random effect between area and time.

2.3 Dissemination of results

It is important to note that the goal of surveillance is not merely to collect data for analysis, but to guide public health policy and action to control and prevent diseases. A key aspect of surveillance practice is, therefore, the proper and timely dissemination of information to those responsible for disease prevention and control. Depending on the circumstances, those should include health care providers, health agencies, government agencies, potentially exposed individuals, vaccine manufacturers, private voluntary organizations, legislators on the health subcommittee, and innumerable others (MMWR, 2001). Report findings should be used consistently and thoughtfully to respond quickly to population's health needs.

Chapter 3

Spatial data

A spatial process in d dimensions is denoted as

$$\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}.$$

Here, Z denotes the attribute we observe, for example, the number of sudden infant deaths or the level of rainfall, and $\mathbf{s}$ refers to the location of the observation. Cressie (1993) distinguishes three basic types of spatial data through characteristics of the domain D:

- Lattice data: The domain D is fixed (of regular or irregular shape) and partitioned into a finite number of areal units with well-defined boundaries. Examples of lattice data are attributes collected by ZIP code, census tract, or remotely sensed data reported by pixels.

- Geostatistical data: The domain D is a continuous, fixed set. By continuous we mean that $\mathbf{s}$ varies continuously over D and therefore $Z(\mathbf{s})$ can be observed everywhere within D. By fixed we mean that the points in D are non-stochastic. It is important to note that the continuity only refers to the domain, and the attribute Z can be continuous or discrete. Examples of this type of data are air pollution or rainfall measured at several monitoring sites.

- Point patterns: Unlike geostatistical and lattice data, the domain D in point patterns is random. Its index set gives the locations of random events that are the spatial point pattern. $Z(\mathbf{s})$ may be equal to 1 $\forall \mathbf{s} \in D$, indicating occurrence of the event, or random, giving some additional information.

In what follows, we present the basic characteristics and analysis methods for each of the three types of spatial data.

3.1 Lattice data

Lattice or areal data arise when a fixed domain is partitioned into a finite number of subregions at which outcomes are aggregated. Examples are the number of cancer cases in the provinces of Spain, or the proportion of people living in poverty in the counties of United States. Lattice data may be used for a variety of inferential issues. One of them is the identification of spatial patterns and their strength. If data are spatially correlated, observations in neighboring areas will be more similar than observations in areas that are farther away. Often, in order to draw correct inferences, the smoothing of the data is necessary since observed measurements in small areas present extreme values due to low population sizes or small samples. Other times, it is desired to make predictions for new areal units different from those were data were recorded. For example, we may wish to analyze data at county level that were initially recorded at the zip code level. This is the Modifiable Areal Unit Problem (MAUP).

3.1.1 Spatial proximity matrices

The concept of spatial proximity matrix, W, is useful in the exploration of areal unit data. Given measurements $Y_1, \ldots, Y_n$ associated with areal units $1, 2, \ldots, n$, the (i, j)th element in W, denoted by w_{ij}, spatially connects units i and j in some fashion. W defines a neighborhood structure over the entire study region, and its elements can be viewed as weights. More weight will be associated with j's closer to i than those farther away from i.

The simplest neighborhood definition is provided by the binary matrix where $w_{ij} = 1$ if regions i and j share some common boundary, perhaps a vertex, and $w_{ij} = 0$ otherwise. Customarily, w_{ii} is set to 0 for $i = 1, \ldots, n$. Note that this choice of proximity measure result in a symmetric spatial proximity matrix. Many other possibilities of spatial proximity can be considered. For instance, we may expand the idea of neighborhood to include regions that are close, but not necessarily adjacent. Thus, we could use $w_{ij} = 1$ for all i and j within a specified distance, or, for a given i, $w_{ij} = 1$ if j is one of the m nearest neighbors of i. The weight w_{ij} can also be defined as the inverse distance between units. Alternatively, we may want to adjust for the total number of neighbors in each region and use a standardized matrix with entries $w_{std,i,j} = w_{ij} / \sum_{j=1}^{n} w_{ij}$. Note that this matrix is not symmetric in most situations where the regions are irregularly shaped.

3.1.2 Measures of spatial association

Global indexes of spatial autocorrelation summarize the degree to which similar observations tend to occur near each other over the entire study area. Two standard statistics that are used to measure the global degree of spatial association in areal data are Moran's I and Geary's C. Moran's I statistic (Moran, 1950) takes the form

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \overline{Y})(Y_j - \overline{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \overline{Y})^2},$$

where n is the number of regions, Y_i is the observed value of the variable of interest in region i, $\overline{Y}$ is the mean of Y_i, $i = 1, \ldots, n$, and w_i is a measure of the spatial proximity between region i and j. It should be noted that, unlike a traditional correlation coefficient, Moran's I statistic is not exactly supported on the interval $[-1, 1]$. Positive values indicate a positive spatial autocorrelation. This occurs when neighboring regions tend to have similar values. Negative values indicate a negative spatial autocorrelation, regions that are close to one another tend to have different values. Finally, values near zero indicate an absence of spatial pattern.

Under the null model where the Y_i are i.i.d., I is asymptotically normally distributed with mean and variance equal to

$$E[I] = \frac{-1}{n-1},$$

and

$$Var[I] = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2},$$

where

$$S_0 = \sum_{i \neq j} w_{ij},\ S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2, \text{ and } S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik} \right)^2.$$

Thus, when the number of regions n is sufficiently large, I has a normal distribution and we can decide whether any given pattern deviates significantly from a random pattern by comparing the z-score $z = (I - E[I])/(Var[I])^{1/2}$ to a standard normal distribution. Randomization is an alternative approach to judge the significance of any observed value of I. This method reassigns the observed regional values among the n fixed regions, providing a randomization distribution. If the observed value of I lies in the tails of this distribution, the assumption of independence among the observations is rejected.

Geary's C statistic (Geary, 1954) is written as

$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \overline{Y})^2}.$$

Geary's C is never negative and ranges between 0 to 2. Values of Geary's C close to 0 denote positive autocorrelation and values close to 2 indicate negative correlation.

Values near 1 denote no spatial autocorrelation. The expected value of Geary's C under the null hypothesis of spatial independence is 1. Under the asymptotically normality assumption the variance is

$$\begin{aligned} Var[C] \; = \; & \left\{(n-1)\left(\sum_i\sum_j w_{ij} + \sum_i\left(\sum_j w_{ij}\left(\sum_i w_{ij}-1\right)\right)/2\right)\right. \\ & \left. -\frac{(\sum_i\sum_j w_{ij})^2}{2}\right\} \times \left\{(n+1)\left(\sum_i\sum_j w_{ij}/2\right)^2\right\}. \end{aligned}$$

There is often interest in providing a local measure of similarity between each region's associated value and those of nearby regions. Local Indicators of Spatial Association (LISA) (Anselin, 1995) are designed to provide an indication of the extent of significant spatial clustering of similar values around each observation. A desirable property is that the sum of the LISA's values across all regions, be equal to a multiple of the global indicator of spatial association. As a result, global statistics may be decomposed into a set of local statistics and most LISAs are defined as local versions of well-known global indexes. One of the most popular LISAs is the local version of Moran's I. For the ith region, it is defined as

$$I_i = \frac{n(Y_i - \overline{Y})}{\sum_j (Y_j - \overline{Y})^2} \sum_j w_{ij}(Y_j - \overline{Y}).$$

Note that the global Moran's I is proportional to the sum of the local Moran's I obtained for all subregions:

$$I = \frac{1}{\sum_{i\neq j} w_{ij}} \sum_i I_i.$$

Typically, the values of the LISAs are mapped to indicate the location of areas with comparatively high or low local association with neighboring areas. A high value for I_i suggests that the area is surrounded by areas with similar values. Such an area is part of a cluster of high observations, low observations, or moderate observations. A low value for I_i indicates that the area is surrounded by areas with dissimilar values. Such an area is an outlier. The observation in area i is different from most or all of the observations of its neighbors. To interpret the local Moran's indexes, maps of p-values associated with the probability of exceeding the observed value of each regional LISA are necessary. These p-values, regardless of the presence or absence of global spatial association, may be obtained by a simulation process with a conditional randomization approach. In this approach, the observed value Y_i at region i is fixed, and the remaining $n-1$ values are randomly reassigned over the other regions.

3.1.3 Bayesian inference

Lattice data are often modeled using Bayesian hierarchical models, which allow complete flexibility in how the estimates borrow strength across neighboring units, and hence improve estimation and prediction of the underlying model features (Ma et al., 2006). In a Bayesian approach to statistical analysis, a probability distribution $f(\mathbf{y}|\boldsymbol{\theta})$, called likelihood, is specified for the observed data $\mathbf{y} = (y_1, \ldots, y_n)$, given a vector of unknown parameters $\boldsymbol{\theta}$. Then, a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ is assigned to $\boldsymbol{\theta}$, where $\boldsymbol{\eta}$ is a vector of hyperparameters. The prior distribution for $\boldsymbol{\theta}$ represents the knowledge about $\boldsymbol{\theta}$ before obtaining the data $\mathbf{y}$. If $\boldsymbol{\eta}$ is not known, a fully Bayesian approach would specify a hyperprior distribution for $\boldsymbol{\eta}$. Alternatively, an empirical Bayes approach might be used, by which an estimate $\hat{\boldsymbol{\eta}}$ is used as if $\boldsymbol{\eta}$ were known. Assuming that $\boldsymbol{\eta}$ is known, inference concerning $\boldsymbol{\theta}$ is based on the posterior distribution of $\boldsymbol{\theta}$, which is defined from Bayes' Theorem as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

The denominator $p(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ defines the marginal likelihood of the data $\mathbf{y}$. This is free of $\boldsymbol{\theta}$ and may be set to a scaling constant which does not impact the shape of the posterior distribution. Thus, the posterior distribution is often expressed as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

One principal difficulty in applying Bayesian methods is the calculation of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$, which usually involves high-dimensional integration that is generally not tractable in closed form. Thus, even when the likelihood and the prior distribution have closed-form expressions, the posterior distribution may not. Markov chain Monte Carlo (MCMC) methods are used for solving this problem. MCMC methods work by generating a sample of values $\{\boldsymbol{\theta}^{(g)}, g = 1, \ldots, G\}$ from a convergent Markov chain whose stationary distribution is the posterior, $p(\boldsymbol{\theta}|\mathbf{y})$. From these samples, empirical summaries of the $\boldsymbol{\theta}^{(g)}$ values may be used to summarize the posterior distribution of the parameters of interest. For example, we might use the sample mean to estimate the posterior mean,

$$\widehat{E(\theta_i|\mathbf{y})} = \frac{1}{G}\sum_{i=1}^{G}\theta_i^{(g)},$$

and the sample variance to estimate the sample variance,

$$\widehat{Var(\theta_i|}\mathbf{y}) = \frac{1}{G-1}\sum_{i=1}^{G}(\theta_i^{(g)} - \widehat{E(\theta_i|\mathbf{y})})^2.$$

A variety of MCMC methods have been proposed, the most common of which is the Gibbs sampler (Geman and Geman, 1984). Suppose our model contains k parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. To implement the Gibbs sampler, it is necessary that samples can be generated from each of the full conditional distributions $\{p(\theta_i|\boldsymbol{\theta}_{j\neq i}, \mathbf{y}), i = 1, \ldots, k\}$. Starting with $(\theta_2^{(0)}, \ldots, \theta_k^{(0)})$, the algorithm repeats for $t = 1, \ldots, T$:

$$\text{Draw } \theta_1^{(t)} \text{ from } p(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)}, \mathbf{y})$$

$$\text{Draw } \theta_2^{(t)} \text{ from } p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)}, \mathbf{y})$$

$$\vdots$$

$$\text{Draw } \theta_k^{(t)} \text{ from } p(\theta_k|\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{(k-1)}^{t}, \mathbf{y})$$

The k-tuple obtained at iteration t, $(\theta_1^t, \ldots, \theta_k^t)$, converges to a draw from the joint posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. Thus, for t sufficiently large, (say, bigger than t_0), the sequence $\{\boldsymbol{\theta}^{(t)}, t = t_0 + 1, \ldots, T\}$ is a sample from the true posterior.

MCMC methods require the use of diagnostics to decide when the sampling chains have reached the stationary distribution, that is, the posterior distribution. One easy way to see if the chain has converged is to examine the traceplot which is a plot of the parameter value at each iteration against the iteration number, and see how well the chain is mixing or moving around the parameter space. Sample autocorrelations are also useful since they can inform whether the algorithm will be slow to explore the entire posterior distribution and this will impede convergence. The Geweke diagnostic (Geweke, 1992) takes the first and last part of the chain and compares the means of both parts, to see if the two parts are from the same distribution. It is also common to run a small number m of parallel chains, initialized at different starting locations, and examine the traceplots to see if there is a point after which all chains seem to overlap. Among the most popular diagnostic is that of Gelman and Rubin (1992). Here, m chains are run for $2N$ iterations each and convergence is monitored by the estimated scale reduction factor which is defined as

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{N-1}{N} + \frac{m+1}{mN}\frac{B}{W}\right)},$$

where B/N is the variance between the means from the m parallel chains, and W is the average of the m within-chain variances. Once convergence is reached, variation within the chains and variation between the chains should coincide, so $\sqrt{\hat{R}}$ should approximately equal one.

MCMC algorithms often produce correlated samples of parameters. Thinning a chain by taking systematic samples at every kth iteration is a common approach to produce chains with less autocorrelation. Although MacEachern and Berliner (1994) show that such a thinning gives less accurate estimates than the complete chain, this approach is still useful in long runs which need great storage capacity.

3.1.4 Hierarchical modeling

Observed counts of disease cases Y_i within spatial regions $i = 1, \dots, n$ are often modeled as binomial or multinomial given the population at risk. For relatively rare events, a common statistical practice is to use a Poisson approximation. Thus, a hierarchical mixed effects model for the count data is used where

$$Y_i \sim Po(\mu_i),$$

$$log(\mu_i) = log(E_i) + \mathbf{x_i}'\boldsymbol{\beta} + u_i + v_i,\ i = 1, \dots, n.$$

Here, E_i represent the internally standardized expected counts, $\mathbf{x_i}$ denote known region specific covariates, and u_i and v_i are correlated and uncorrelated random effects, respectively, that account for extra-Poisson variability in the observed data. Usually, the focus of interest is modeling the true underlying relative risk, θ_i, which is expressed as

$$\theta_i = \frac{\mu_i}{E_i} = exp(\mathbf{x_i}'\boldsymbol{\beta} + u_i + v_i),\ i = 1, \dots, n.$$

Bayesian modelling requires specification of prior distributions for the random effects. For modeling the clustering component, a common practice is to use a spatial correlation structure where the estimation of the risk in any area depends on the neighboring areas. The conditional autoregressive (CAR) distribution proposed by Besag et al. (1991) is typically used. The CAR model smoothes the data according to a certain neighborhood structure specified in a proximity matrix W and is expressed as follows:

$$u_i|\mathbf{u_{-i}} \sim N(\ \overline{u}_{\delta_i}, \frac{\sigma_u^2}{n_{\delta_i}}),$$

where $\overline{u}_{\delta_i} = {n_{\delta_i}}^{-1} \sum_{j \in \delta_i} u_j$ and δ_i denotes the set of labels of the neighbors of area i. Hence, u_i has a normal distribution with conditional mean given by the average of the neighboring u_j's and conditional variance inversely proportional to the number of neighbors n_{δ_i}. The uncorrelated heterogeneity may be modeled as independent and identically distributed normal variables with mean zero and variance σ_v^2,

$$v_i \sim N(0, \sigma_v^2).$$

In a full Bayesian analysis, prior distributions are specified for $\boldsymbol{\beta}$ and for the parameters σ_u^2 and σ_v^2 which control the variability of u and v. Then, Bayesian estimation of the parameters is proceed via MCMC methods.

Once fit, hierarchical Bayesian models may be assessed via the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). The DIC is a criterion based on a trade-off between the fit of the data to the model and the complexity of the model. For a likelihood $f(\mathbf{y}|\boldsymbol{\theta})$, the posterior distribution of the deviance is defined as

$$D(\boldsymbol{\theta}) = -2 \log f(\mathbf{y}|\boldsymbol{\theta}).$$

The fit of a model can be summarized by the posterior mean of the deviance

$$\overline{D} = E_{\theta|y}[D],$$

and its complexity may be measured by the effective number of parameters, p_D, which is expressed as the posterior mean of the deviance minus the deviance of the posterior expected parameter estimates,

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\boldsymbol{\theta}]) = \overline{D} - D(\bar{\boldsymbol{\theta}}).$$

The Deviance information criterion (DIC) is defined as

$$DIC = \overline{D} + p_D = 2\overline{D} - D(\bar{\boldsymbol{\theta}}),$$

with smaller values of DIC indicating a better model.

3.1.5 Issues with lattice data

Spatial analyses of aggregated data are subject to the Misaligned Data Problem (MIDP), which occurs when the spatial data are analyzed at a scale different from that at which they were originally collected (Banerjee et al., 2004). In some cases, the purpose might be merely to obtain the spatial distribution of one variable at a new level of spatial aggregation. In other cases, we might wish to relate one variable to another variables that are available at different spatial scales. An example of this scenario is where we want to determine whether the risk of an adverse outcome provided at zip level is related to exposure to an environmental pollutant measured at a network of stations, adjusting for population at risk and other demographic information which are available at county level.

The Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984) is a problem that has long been identified in the analysis of aggregated data, whereby conclusions may change if one aggregates the same underlying data to a new level of spatial aggregation. The MAUP consists of two interrelated effects. The first effect is the scale or aggregation effect. It concerns the different inferences obtained when the same data is grouped into increasingly larger areas. The second effect is the grouping or zoning effect. This effect considers the variability in results due to alternative formations of the areas leading to differences in area shape at the same or similar scales.

Ecological studies are characterized by being based on aggregated data (Robinson, 1950). Such studies contain the potential for ecological fallacy, which occurs when estimated associations obtained from analysis of variables measured at aggregated level lead to conclusions different from analysis based on the same variables measured at the individual level. The ecological inference problem can be viewed as a special case of the MAUP. The resulting bias, called ecological bias, is comprised of two effects analogous to the aggregation and zoning effects in the MAUP. These are the aggregation bias due to the grouping of individuals, and the specification bias due to the differential distribution of confounding variables created by grouping. (Gotway and Young, 2002).

3.2 Geostatistical data

When analyzing patterns of disease, we may wish to study potential disease risk factors. Many of these risk factors are exposure variables representing spatially continuous phenomenons but measured only at particular sites. For example, exposure variables may represent the level of a pollutant observed at several monitoring stations, or the density of mosquitos responsible of disease transmission measured with traps at different locations (Waller and Gotway, 2004). Suppose $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ are observations of a spatial exposure variable Z at the spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$. The data are assumed to be a partial realization of a random process

$$\{Z(\mathbf{s}) : \mathbf{s} \in D\},$$

where D is a fixed subset of $\mathbb{R}^d$ and the spatial index $\mathbf{s}$ varies continuously throughout D. For practical reasons the process $Z(\cdot)$ can only be observed at a finite set of locations. Based upon this partial realization, we seek to infer the characteristics of the spatial process that gives rise to the data observed, such as the mean and variability of the process. These characteristics are useful for the prediction of the process at unobserved locations and the construction of spatially continuous surfaces for attribute values.

3.2.1 Stationarity and Variogram

A random process $Z(\cdot)$ is said to be strictly stationary if for any set of locations $\mathbf{s}_i$, $i = 1, \ldots, N$, and any $\mathbf{h} \in \mathbb{R}^d$, the distribution of $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ is the same as that of $\{Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_n + \mathbf{h})\}$. A less restrictive condition is given by the second-order stationarity. Under this condition, the process has a constant mean,

$$E[Z(\mathbf{s})] = \mu, \forall \mathbf{s} \in D,$$

and the covariances depend only on the differences between locations,

$$Cov(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})) = C(\mathbf{h}),\ \forall \mathbf{s} \in D,\ \forall \mathbf{h} \in \mathbb{R}^d.$$

In addition, if the covariances are functions only of the distances between locations and not of the directions, the process is called isotropic. If not, it is anisotropic. A process is said to be intrinsically stationary if in addition to the constant mean assumption it also satisfies

$$Var[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)] = 2\gamma(\mathbf{s}_i - \mathbf{s}_j),\ \forall \mathbf{s}_i, \mathbf{s}_j.$$

The function $2\gamma(\cdot)$ is known as the variogram and $\gamma(\cdot)$ as the semivariogram. Under the assumption of intrinsic stationarity, the constant-mean assumption implies

$$2\gamma(\mathbf{h}) = Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = E[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2],$$

and the semivariogram can be easily estimated based on the method-of-moments:

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s_i}) - Z(\mathbf{s_j}))^2,$$

where $|N(\mathbf{h})|$ denotes the number of distinct pairs in $N(\mathbf{h}) = \{(\mathbf{s_i}, \mathbf{s_j}) : \mathbf{s_i} - \mathbf{s_j} = \mathbf{h};\ i, j = 1, \ldots, n\}$ (Cressie, 1993).

A plot of the empirical semivariogram against the separation distance conveys important information about the continuity and spatial variability of the process (see Figure 3.1). Often, at relatively short distances, the semivariogram is small, but tends to increase with distance, indicating that observations in close proximity tend to be more alike than those farther apart. Then, at a large separation distance referred to as the range, the semivariogram levels off to a nearly constant value referred to as the sill. Thus, the empirical semivariogram indicates that spatial dependence decays with distance within the range, and observations are spatially uncorrelated beyond the range, this reflected by a near constant variance. If there is a discontinuity or vertical jump at the origin, the process has nugget effect. This effect is often due to measurement error, but can also indicate a spatially discontinuous process.

After the empirical semivariogram has been estimated, a valid theoretical semivariogram model is fitted to it using visual assessment and statistical algorithms such as weighted least squares. Three conditions are necessary for a semivariogram model to be valid. Namely, $\gamma(\mathbf{0}) = 0$, $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})\ \forall \mathbf{h}$, and $\gamma(\cdot)$ is a conditionally negative definite function. That is, $\sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$ for all n, all $\mathbf{s}_1, \ldots, \mathbf{s}_n$

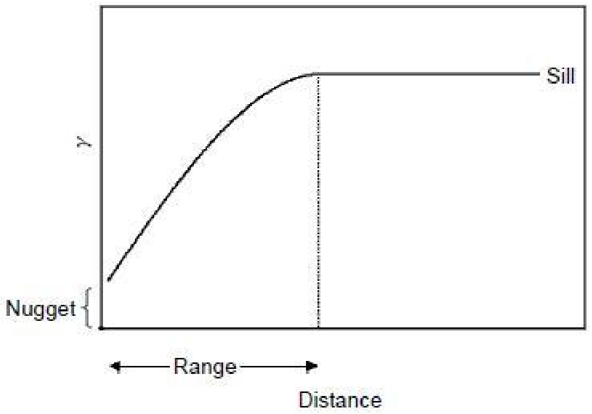


Figure 3.1: Typical semivariogram.

and all $a_1, \ldots, a_n$ such that $\sum_{i=1}^{n} a_i = 0$. A large number of semivariogram models have been used, the most popular being the spherical, exponential, gaussian, Matérn and power models (Gelfand et al., 2010). For example, the exponential model is defined as

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} 0, & \mathbf{h} = 0, \\ c_0 + c_1[1 - exp(-||\mathbf{h}||/a)], & \mathbf{h} \neq 0, \end{cases}$$

where $a > 0$, $c_0 \geq 0$ is the nugget, $c_1 \geq 0$ is the range, and $c_0{+}c_1$ is the sill.

3.2.2 Kriging

Kriging (Matheron, 1963) is a spatial prediction method that can give predictions of unknown values of a random process. Under several assumptions, these predictions are best linear unbiased estimators. Kriging depends on the second-order properties and the variogram of the process (van Beers and Kleijnen, 2003). There are several types of Kriging differing by underlying assumptions and analytical goals. For example, the Ordinary Kriging gives a linear prediction assuming a constant unknown mean, Universal Kriging can be used for data with a non-stationary mean structure, and Cokriging refers to multivariate linear prediction in which one or more interrelated spatial processes are incorporated.

Here, we describe the Ordinary Kriging in which it is assumed

$$Z(\mathbf{s}) = \mu + \delta(\mathbf{s}), \mathbf{s} \in D,$$

where $\mu \in \mathbb{R}$ is unknown and $\delta(\cdot)$ is a zero-mean intrinsically stationary random process with variogram $2\gamma(\cdot)$. Suppose that we have observed data $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$, and want to predict the value of the process at an arbitrary location $\mathbf{s}_0 \in D$. The Ordinary Kriging estimator at $\mathbf{s}_0$ is defined as the linear unbiased estimator

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)$$

of $Z(\mathbf{s}_0)$ that minimizes the mean square predictor error defined as $E[(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2]$. Here, the weights $\lambda_1, \ldots, \lambda_n$ are determined among all the linear predictors satisfying the properties of unbiasedness and minimum mean-squared prediction error. For unbiasedness $E[\hat{Z}(\mathbf{s}_0)] = E[Z(\mathbf{s}_0)]$, or equivalently $\sum_{i=1}^{n} \lambda_i = 1$. To minimize the mean-squared prediction error subject to the unbiasedness constraint the method of Lagrange multipliers is used. Thus, the method finds $\lambda_1, \ldots, \lambda_n$ and a Lagrange multiplier m, that minimize the following objective function:

$$\phi(\lambda_1, \ldots, \lambda_n, m) = E[(\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - Z(\mathbf{s}_0))^2] - 2m(\sum_{i=1}^{n} \lambda_i - 1).$$

After some manipulations, this function can be expressed as

$$\phi(\lambda_1, \ldots, \lambda_n, m) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{s}_i, \mathbf{s}_j) + 2\sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0, \mathbf{s}_i) - 2m(\sum_{i=1}^{n} \lambda_i - 1).$$

Differentiating ϕ with respect $\lambda_1, \ldots, \lambda_n$ and m, and equating the result to zero leads to the linear system

$$\mathbf{\Gamma_0} \boldsymbol{\lambda_0} = \boldsymbol{\gamma_0},$$

where

$$\boldsymbol{\lambda_0} = (\lambda_1, \ldots, \lambda_n, m)',$$

$$\boldsymbol{\gamma_0} = [\gamma(\mathbf{s}_0 - \mathbf{s}_1), \ldots, \gamma(\mathbf{s}_0 - \mathbf{s}_n), 1],$$

and $\mathbf{\Gamma_0}$ is a symmetric $(n+1) \times (n+1)$ matrix with elements as follows:

$$\mathbf{\Gamma_0} = \left\{ \begin{array}{rl} \gamma(\mathbf{s}_i - \mathbf{s}_j), & i = 1, \ldots, n; j = 1, \ldots, n, \\ 1, & i = n+1; j = 1, \ldots, n, \\ & j = n+1; i = 1, \ldots, n, \\ 0, & i = n+1; j = n+1. \end{array} \right.$$

The Ordinary Kriging coefficients can then be determined solving the linear system so that

$$\boldsymbol{\lambda_0} = \mathbf{\Gamma_0^{-1}} \boldsymbol{\gamma_0}.$$

Putting the optimal weights into the expression of this functional one can see that the minimum of the prediction error also known as the kriging variance is given by

$$\sigma^2(\mathbf{s}_0) = \boldsymbol{\lambda_0'} \boldsymbol{\gamma_0} = \boldsymbol{\gamma_0'} \mathbf{\Gamma_0}^{-1} \boldsymbol{\gamma_0}.$$

This expression is a measure of the uncertainty in the prediction of $Z(\mathbf{s}_0)$.

3.3 Point patterns

Point processes are stochastic models that describe the locations of interesting events and possibly some additional information. A point process is written as $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, where D is random. In unmarked processes $Z(\mathbf{s})$ equal 1 $\forall \mathbf{s} \in D$ indicating occurrence of the event. In marked process $Z(\mathbf{s})$ is random giving some information such as the type of event. In general, D is contained in $\mathbb{R}^d$, where $d = 2$ or $d = 3$. A point pattern is a collection of points $\{\mathbf{s}_i \in D : i = 1, \ldots, n\}$ and is typically interpreted as a realization of a point process. A point process is denoted by N. We write $N(A)$ for the random variable which represents the number of events in the region $A \subset D$.

A primary goal in the analysis of point processes is to identify patterns in the distribution of an observed set of locations, as well as the estimation of the density of events across the region of study. In public health data, it is often of interest the identification of aggregated patterns which may reflect unusual clusters of cases of a particular disease. Analysis are also concerned about the correlation between the cases of a particular disease and spatial covariates, such as environmental exposures, and the relationships between different point processes, such the cases and controls of a disease.

3.3.1 First- and second-order properties

The first- and second-order properties of a point process are useful for understanding important aspects of the behavior of the process. The (first-order) intensity function, $\lambda(\cdot)$, describes the way in which the mean value of the process varies across space, whereas the second-order intensity function, $\lambda_2(\cdot,\cdot)$, describes the spatial dependence in the process. The intensity function is defined as

$$\lambda(\mathbf{s}) = \lim_{|d\mathbf{s}| \to 0} \frac{E[N(d\mathbf{s})]}{|d\mathbf{s}|},$$

where $d\mathbf{s}$ denotes an infinitesimal area centered at $\mathbf{s}$. $\lambda(\mathbf{s})$ is the mean number of events per unit area at the point $\mathbf{s}$. The second-order intensity function measures the covariance between values of the process at different regions. This is defined as

$$\lambda_2(\mathbf{s}_i, \mathbf{s}_j) = \lim_{|d\mathbf{s}_i| \to 0, |d\mathbf{s}_j| \to 0} \frac{E[N(d\mathbf{s}_i)N(d\mathbf{s}_j)]}{|d\mathbf{s}_i||d\mathbf{s}_j|}.$$

The concepts of stationarity and isotropy provide a starting place for modeling spatial point processes. We say that a process is stationary if the intensity is constant over the study area, $\lambda(\mathbf{s}) = \lambda$, and, in addition, the second-order intensity depends only on event location differences, $\lambda_2(\mathbf{s_i}, \mathbf{s_j}) \equiv \lambda_2(\mathbf{s_i} - \mathbf{s_j})$. If the process is furthermore isotropic, the second-order intensity depends only on distance,

$\lambda_2(\mathbf{s_i}, \mathbf{s_j}) \equiv \lambda_2(||\mathbf{s_i} - \mathbf{s_j}||)$. In other words, a process is said to be stationary when it is invariant to translation, and isotropic when it is also invariant to rotation.

3.3.2 Complete Spatial Randomness

Point processes provide models for point patterns. The simplest theoretical model is that of complete spatial randomness (CSR), in which an event is equally likely to occur at any location within the study area, regardless of the locations of other events. That is, events distribute uniformly and independently across the study area. The stochastic representation of CSR is the homogeneous Poisson process (HPP), which is characterized by the two following properties:

1. The number of events in any region $A \subset D$ follows a Poisson distribution with mean $\lambda|A|$, where λ denotes the constant intensity function of the process, and $|A|$ the area of A. That is, $N(A) \sim Po(\lambda|A|)$.

2. If $A_1, \ldots, A_k$ are k disjoint regions of D, then the number of points $N(A_1), \ldots,$ $N(A_k)$ are independent random variables, for an arbitrary k.

Most processes achieve a certain deviation from the CSR in some fashion. However, CSR plays a central role in many investigations because it operates as a dividing hypothesis between regular and clustered patterns (Diggle, 1983). We distinguish random, regular and clustered patterns on the basis of the average distance between an event and its nearest neighbor. In a clustered pattern, this distance is smaller than the same distance in a random pattern, whereas in a regular pattern this distance is larger than expected under randomness.

The most elementary test to contrast the CSR hypothesis is the χ^2 test based on quadrat counts. Suppose the study region is partitioned into r rows and c columns which define rc non-overlapping subregions or quadrats of equal area. Under the null hypothesis of CSR, the number of events in quadrat ij, n_{ij}, are independent Poisson random variables with the same expected value $\bar{n} = n/(rc)$. The Pearson Chi-square statistic is defined as

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \bar{n})^2}{\bar{n}},$$

or alternatively as

$$X^2 = (rc - 1)s^2/\bar{n},$$

where s^2 is the sample variance of the rc quadrat counts. This statistic is an index of departure from CSR. If the pattern is CSR, then the ratio of sample variance and sample mean should be approximately 1. Under CSR, X^2 follows approximately a χ^2 distribution with $rc - 1$ degrees of freedom provided that the expected number of events per quadrat exceeds 1 and $rc > 6$. The choice of the shape and the number of quadrats in the χ^2 test is a subjective element that can influence the result. Test statistics based on distances together with Monte Carlo tests eliminate this subjectiveness.

The nearest neighbor distribution refers to the distance from a randomly chosen event to the nearest other event. If the observed pattern has n events and d_i denotes the distance from the ith event to the nearest other event, then the Empirical Distribution Function (EDF) is given by

$$\hat{G}(t) = n^{-1}\#(d_i \leq t).$$

The point to nearest event distribution considers the distance between a randomly chosen location and the nearest event. Using distances e_i from each of m sample points to the nearest of the n events, the EDF is

$$\hat{F}(t) = m^{-1}\#(e_i \leq t).$$

The estimates of the functions together with confidence envelopes constructed from Monte Carlo simulation under the HPP hypothesis can be used to test CSR.

3.3.3 K-function

For stationary and isotropic point processes with intensity λ, the K-function (Ripley, 1976) is defined as

$$K(t) = \lambda^{-1}E[N(b(s,t)\backslash\{\mathbf{s}\}) : \mathbf{s} \in N],\ t > 0,$$

where $b(\mathbf{s}, t)$ is the disc with center $\mathbf{s}$ and radius t. $K(t)$ provides an interpretable measure of the spatial dependency structure in the point process. In particular, if a denotes the area of region D, $\lambda^2 aK(t)$ is the expected number of ordered pairs of points in region D with pairwise distance less than or equal to t. For a homogeneous process with no spatial dependence we expect $K(t) = \pi t^2$. Under regularity $K(t) < \pi t^2$ and under clustering $K(t) > \pi t^2$. An estimate of the K-function from the data $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is given by

$$\widehat{K}(t) = \frac{1}{\lambda^2 a}\sum_{i=1}^{n}\sum_{j\neq i} w_{ij} I(d_{ij} \leq t),$$

where $I(\cdot)$ is the indicator function, and $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$ is the Euclidean distance between the points $\mathbf{s}_i$ and $\mathbf{s}_j$. Here, w_{ij} is used for edge-correction and denotes the reciprocal of proportion of the circle centered on $\mathbf{s}_i$ and with radius d_{ij} which is contained in D. To complete the estimate we need to replace the unknown intensity λ with an estimate $\hat{\lambda}$.

3.3.4 Estimating the intensity function

For a stationary point process, the intensity is constant and for an observed pattern of n points, the natural estimator is the observed number of events per unit area: $\hat{\lambda} = n/|D|$. For nonstationary processes, a common method to estimate the spatially varying intensity function involves kernel density estimation (Silverman, 1986). Usually, kernel estimation methods focus on estimating the probability density function $f(\cdot)$ rather than the intensity function $\lambda(\cdot)$. The density function defines the probability of observing an event at a location $\mathbf{s}$ and integrates to one across the area of study. In contrast, the intensity function provides the number of events expected per unit area at location $\mathbf{s}$ and integrates to the overall mean number of events per unit area. As a result, the density and intensity functions are proportional

$$\lambda(\mathbf{s}) = f(\mathbf{s}) \int_D \lambda(\mathbf{u}) d\mathbf{u},$$

and the relative spatial pattern in densities and intensities are the same.

A kernel estimator of the density function $f(\cdot)$ at the location $\mathbf{s}$ based on the observations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ takes the form

$$\widehat{f}(\mathbf{s}) = \frac{1}{h^2} \sum_{i=1}^{n} k\left(\frac{\mathbf{s} - \mathbf{s}_i}{h}\right),$$

where $k()$ is a radially symmetric bivariate probability density function known as kernel function, and h is a smoothing parameter known as bandwidth. Although the form of the kernel weakly influences the estimates, the bandwidth can have a big impact. Thus, small values of h can result in estimated densities that are too spiky, whereas large values provide smoother surfaces that may ignore local characteristics of the densities. A typical choice for the kernel function might be the quartic kernel, expressed as

$$k(\mathbf{s}) = \begin{cases} \frac{3}{\pi}(1 - \mathbf{s}'\mathbf{s})^2 & \text{if } \mathbf{s}'\mathbf{s} \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding estimated density is

$$\widehat{f}(\mathbf{s}) = \sum_{||\mathbf{s}-\mathbf{s}_i|| \leq h} \frac{3}{\pi h^2} k \left(1 - \frac{||\mathbf{s} - \mathbf{s}_i||^2}{h^2}\right)^2.$$

In practice, exploratory analyses that consider several values of bandwidths may be carried out to determine, somewhat subjectively, an appropriate bandwidth value. Other criterions involve the minimization of the mean integrated squared error between the estimate and the true density (Wand and Jones, 1995). Edge effects tend to distort the kernel estimates close to the boundary of the region since events near the boundary have fewer local neighbors than events in the interior. One way to deal with this problem is to modify the kernel estimate by dividing by the following edge-correction term:

$$p_h(\mathbf{s}) = \int_D h^{-2} k \left(\frac{\mathbf{s} - \mathbf{u}}{h} \right) d\mathbf{u},$$

which represents the volume under the scaled kernel centered on $\mathbf{s}$ which lies inside the study region (Gatrell et al., 1996).

3.3.5 Estimating the ratio of intensity functions

In many applications, the goal of the analysis is the comparison of spatial variation in disease risk between two groups. For example, we may wish to detect differences between the spatial pattern observed in the cases of a particular disease, and the spatial pattern observed in a set of controls reflecting the population at risk. Consider the locations of all the n_1 cases of a disease, and the n_2 locations of a set of controls. The cases and the controls can be assumed to be realizations of two Poisson processes with intensity functions $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$, respectively. Each intensity function is proportional to its associated density. Specifically,

$$f_1(\mathbf{s}) = \lambda_1(\mathbf{s}) / \int_D \lambda_1(\mathbf{u}) d\mathbf{u}, \text{ and } f_2(\mathbf{s}) = \lambda_2(\mathbf{s}) / \int_D \lambda_2(\mathbf{u}) d\mathbf{u},$$

where f_1 denotes the density of the cases, and f_2 denotes the density of the controls.

Kelsall and Diggle (1995) suggest using the logarithm of the ratio of the two spatial densities,

$$r(\mathbf{s}) = log\{f_1(\mathbf{s})/f_2(\mathbf{s})\},$$

to investigate the relation between the spatial patterns of the cases and controls. Note that $r(\mathbf{s})$ and the logarithm of the ratio of intensity functions only differ in an additive constant that does not depend on $\mathbf{s}$:

$$r(\mathbf{s}) = log\{\lambda_1(\mathbf{s})/\lambda_2(\mathbf{s})\} - log \left\{ \int_D \lambda_1(\mathbf{u}) d\mathbf{u} / \int_D \lambda_2(\mathbf{u}) d\mathbf{u} \right\},$$

and hence, the two functions contain identical information regarding the spatial variation in risk. To implement this approach, we estimate the ratio $r(\mathbf{s})$ by the ratio of the kernel density estimates,

$$\widehat{r}(\mathbf{s}) = log\{\widehat{f_1}(\mathbf{s})/\widehat{f_2}(\mathbf{s})\}.$$

Mapping $\widehat{r}(\mathbf{s})$ provides a spatial picture of the areas where $\widehat{r}(\mathbf{s}) > 0$ and where $\widehat{r}(\mathbf{s}) < 0$ indicating that the probability of observing cases rather than controls is more or less likely, respectively. The assessment of these local deviations from zero may indicate the presence of local clusters. Their significance is assessed via Monte Carlo analysis under a random labeling hypothesis. Sometimes, it is also of interest the investigation of overall clustering. Here, the global null hypothesis reflects constant relative risk,

$$H_0 : r(\mathbf{s}) = 0, \ \forall \mathbf{s} \in D.$$

In this situation the spatial densities of cases and controls may vary across the study area but are always in the same relative proportion. Inference is based on the statistic

$$\int_D \widehat{r}(\mathbf{u})^2 d\mathbf{u},$$

which summarizes all deviations between the case and control intensities across the entire study area (Waller and Gotway, 2004).

Chapter 4

Model-based estimation of missing values in mortality data

In any surveillance system, the processes of detection, confirmation and reporting of cases are critical activities which greatly influence the capacity to accurately evaluate the impact of a given disease in a population (Michel et al., 2000). Ideally, surveillance data is timely, complete, and of high quality. In practice, surveillance data are often delayed, incomplete, or unreported. A thorough understanding of the limitations of surveillance data is necessary to inform policy decisions and support efforts to monitor and control disease.

Missing data are a common problem in surveillance which can be a serious impediment for data analysis. Analyzing data while naively ignoring missing data sources can result in biased estimates of disease burden and invalid inference to the larger population. Little and Rubin (1987) established a general statistical framework for classifying missing data mechanisms. Let $\boldsymbol{Y}^*$ denote the complete set of measurements which would have been obtained were there no missing values, and partition this set into $\boldsymbol{Y}^* = (\boldsymbol{Y}^{(o)}, \boldsymbol{Y}^{(m)})$ with $\boldsymbol{Y}^{(o)}$ denoting the measurements actually observed and $\boldsymbol{Y}^{(m)}$ those that are missing. Let $\boldsymbol{R}$ denote a set of indicator variables, denoting which elements of $\boldsymbol{Y}^*$ fall into $\boldsymbol{Y}^{(o)}$ and which into $\boldsymbol{Y}^{(m)}$. A probability model for the missing value mechanism defines the probability distribution of $\boldsymbol{R}$ conditional on $\boldsymbol{Y}^* = (\boldsymbol{Y}^{(o)}, \boldsymbol{Y}^{(m)})$. Data are missing completely at random (MCAR) if $\boldsymbol{R}$ is independent of both $\boldsymbol{Y}^{(o)}$ and $\boldsymbol{Y}^{(m)}$, that is, if there are no systematic differences between the observed and the missing values. Data are missing at random (MAR) if $\boldsymbol{R}$ is independent of $\boldsymbol{Y}^{(m)}$ or, in other words, if the probability a variable is missing depends only on available information. Finally, data are missing not at random (MNAR) if $\boldsymbol{R}$ is dependent on $\boldsymbol{Y}^{(m)}$, that is, if the probability that an observation is missing depends on information that is not observed, like the value of the observation itself.

Common approaches to handle missing data challenges include ad-hoc adjustments, weighting methods, multiple imputation, and model-based approaches. Frequently, ad-hoc approaches use complete or available observations, impute missing values using the mean of the the known observations, or include as new variables indicators of missingness and their interactions. Analyses based on these methods, however, may be biased when missing data are not MCAR. When it is plausible that missing data are MAR, such biases can be overcome using other techniques. Weighting (Carpenter et al., 2006) is a simple approach for making the observations included in the analysis representative of the original sample. In this approach, a model for the probability of missingness is fit, and the inverse of these probabilities are used as weights for the complete observations. Weighting is useful to correct for bias, but it still has disadvantages in terms of efficiency since not all observations are included in the study (Horton and Kleinman, 2007). Multiple imputation (Rubin, 1976) is a three-step approach that allows observations with incomplete data to be included in the analyses. Each missing value is replaced with several imputed values, each of which is predicted from a slightly different model and also reflects sampling variability. This process results in the creation of a number of completed datasets that are analyzed using complete-data methods. Finally, the results are combined across datasets. The key issue is the appropriate specification of the imputation model to avoid potential for bias. In model-based procedures a model for the partially missing data is defined and inferences are based on the likelihood under that model, with parameters estimated by procedures such the EM algorithm (Dempster et al., 1977). This procedure presents several advantages over other methods like flexibility and availability of large sample estimates of variance, which take into account incompleteness in the data.

National estimates of the all-cause and P&I mortality burden derived from these data treat all missing values as zero counts. The effect of this methodological decision is to bias estimates downward and produce underestimates of the true mortality burden, although the extent of underestimation has not been studied systematically. To address this issue, we propose a regression-based procedure that utilizes relevant information to impute missing values and thus produce a more accurate estimate of mortality. Our imputation approach uses Poisson regression to model weekly death counts by city, calendar week, calendar year, and age group. We describe the full regression model below. In cases where the full model may require information from other missing observations, we eliminate the corresponding predictors from the model and refit a new regression model using a subset of the complete set of predictors.

The outline of the chapter is as follows. First, we describe the all-cause and P&I mortality data from the 122 Cities Mortality Reporting System (122 CMRS). Next we describe our imputation model and consider several different model specifications. We then evaluate each model with the all-cause data using a cross-validation approach, and select the set of predictors that offer a combination of performance and robustness. Finally in Section 4.4 we use the model thus selected to compute revised national all-cause and P&I mortality, and P&I excess mortality estimates, and conclude by comparing these revised estimates to those estimates calculated without imputation.

4.1 Data

Our data consist of weekly all-cause and P&I deaths during the period 1962-2004 from the 122 Cities Mortality Reporting System (122 CMRS) operated by the Centers for Disease Control and Prevention (CDC). This is one of several component systems used for influenza surveillance in the United States and has been continuously operated by CDC since 1962, currently by the Influenza Division of the National Center for Immunization and Respiratory Disease (NCIRD). Participating cities comprise roughly 25% of the U.S. population, with a heavier geographic representation in large- and mid-sized urban areas. Each city reports weekly counts of all-cause, influenza-related, and pneumonia-related deaths. CDC uses the ratio of P&I deaths to all-cause deaths as its primary indicator for determining epidemic-associated activity. These data are the timeliest publicly available source of influenza mortality data in the U.S., with typical reporting lags of 2-3 weeks. The CDC collects data from participating cities and reports weekly deaths of all-cause and P&I mortality stratified to the following seven age groups: less than 28 days, 28 days - 1 year, 1-14 years, 15-24 years, 25-44 years, 45-64 years, and 65 years and over.

We analyzed 2,236 weeks of data from 122 cities stratified across seven age groups, for a total of 1,909,544 observations. Of these, 64,085 (3.36%) are missing in the all-cause mortality data, i.e. deaths for that week and age group were unreported to the CDC from the relevant city. And 64,153 (3.36%) are missing in the P&I mortality data. Table 4.1 summarizes the extent of missing data across all participating cities. We note that in the all-cause data, 110 of 122 (90.16%) of cities have less than 2% missing values. The remaining 12 cities (9.84%) have missing data ranging from 2% to 82%. In the P&I data, 88.52% of cities have less than 2% missing values, and the rest of cities have missing values ranging from 2% to 82%. Figure 4.1 shows the distribution of missing data across study years, calendar weeks, and age groups in the all-cause and P&I mortality data. Year 1965 and week 52 have the highest percentage of missing values (0.14% and 0.41%, respectively). The percentage of missing values is very similar for all age groups and ranges 0.42% to 0.49%.

% missing data	[0.75-1)	[1-2)	[2-5)	[5-10)	[10-15)	[15-70)	[70-75)	[75-82)
All-cause	69	41	5	2	1	0	2	2
P&I	67	41	7	2	1	0	2	2

Table 4.1: Distribution of cities according with their percentages of missing data in the all-cause and P&I mortality data.

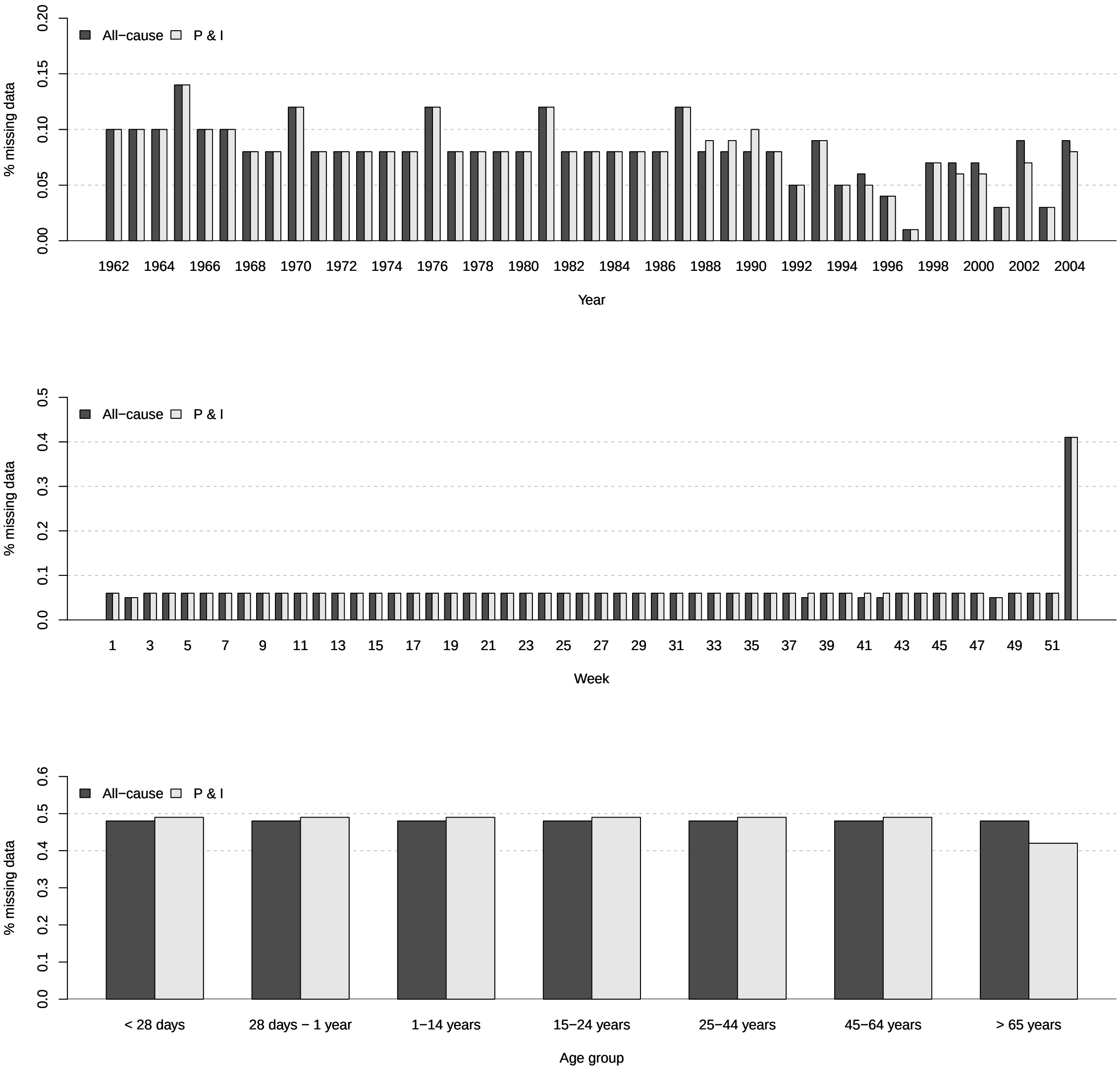


Figure 4.1: Percentages of missing data in each year, week, and age group in the all-cause and P&I mortality data.

4.2 Methods

Let $Y_{w,y,ag,c}$ denote the number of deaths in week $w \in \{1, \ldots, n_W\}$, year $y \in \{1, \ldots, n_Y\}$, age group $ag \in \{1, \ldots, n_{AG}\}$ and city $c \in \{1, \ldots, n_C\}$. For the 122 CMRS data, $n_W = 52$, $n_Y = 43$, $n_{AG} = 7$ and $n_C = 122$. As potential predictors to impute a missing value, we considered nearby weeks, years, age groups, or cities. By 'nearby' we intend the natural meaning in context, which might be temporal (the preceding or following week or year), spatial (geographically close cities), or categorical (next older or younger age group). To make this concrete, suppose there was no reported mortality count from Boston MA for 15-24 year olds during week 24 of year 1999. As predictors to impute the missing value, we might consider the same week and year but use mortality of 25-44 year olds, or mortality from nearby Cambridge MA or Somerville MA. We might also consider mortality in the same age group from weeks 23 or 25 of 1999, or perhaps from week 24 of the years 1998 or 2000. In order to keep the number of predictors manageable, we specified a maximum number of neighboring weeks, years, age groups, and cities. We denote the bound on each of these parameters by m_W, m_Y, m_{AG}, and m_C respectively.

Different choices of m_W, m_Y, m_{AG} and m_C can result in different estimates. In Section 4.3 we present comparisons for several choices when applied to the all-cause data set. Associated to these values are the sets of indexes I_W, I_Y, I_{AG} and I_C that are defined as follows: $I_W = \{-m_W, \ldots, -1, 1, \ldots, m_W\}$ if $m_W \geq 1$, $I_W = \emptyset$ otherwise, $I_Y = \{-m_Y, \ldots, -1, 1, \ldots, m_Y\}$ if $m_Y \geq 1$, $I_Y = \emptyset$ otherwise, $I_{AG} = \{-m_{AG}, \ldots, -1, 1, \ldots, m_{AG}\}$ if $m_{AG} \geq 1$, $I_{AG} = \emptyset$ otherwise, and $I_C = \{1, \ldots, m_C\}$ if $m_C \geq 1$, $I_C = \emptyset$ otherwise.

Suppose the value $Y_{w,y,ag,c}$ is missing. We can estimate this value on the log scale with a linear combination of the known number of deaths in the closest weeks, years, age groups and cities:

$$\begin{aligned}
log(Y_{w,y,ag,c}) &= \widehat{\beta^0} + \sum_{i \in I_W} \widehat{\beta^{W+i}} I(w+i; w, y, ag, c) Y^*_{w+i,y,ag,c} \\
&\quad + \sum_{i \in I_Y} \widehat{\beta^{Y+i}} I(y+i; w, y, ag, c) Y_{w,y+i,ag,c} \\
&\quad + \sum_{i \in I_{AG}} \widehat{\beta^{AG+i}} I(ag+i; w, y, ag, c) Y_{w,y,ag+i,c} \\
&\quad + \sum_{i \in I_C} \widehat{\beta^{C_i}} I(c_i; w, y, ag, c) Y_{w,y,ag,c_i}.
\end{aligned}$$

Here, c_i denotes the i^{th} closest city to c. $Y^*_{w+i,y,ag,c}$ is the number of deaths in the week w plus i more weeks if $i > 0$, or less $-i$ weeks if $i < 0$. Therefore, it is possible $Y^*_{w+i,y,ag,c}$ represents the number of deaths in a year different from y. For example, if $w = 52$, $Y^*_{w+1,y,ag,s} = Y_{1,y+1,ag,s}$, and if $w = 1$, $Y^*_{w-1,y,ag,s} = Y_{52,y-1,ag,s}$. More generally, $\forall i \in I_W$,

$$Y^*_{w+i,y,ag,c} = \begin{cases} Y_{w+i,y,ag,c} & \text{if } (i > 0 \text{ and } i > 52 - w \text{) or } (i < 0 \text{ and } -i \geq w), \\ Y_{i-(52-w),y+1,ag,c} & \text{if } i > 0 \text{ and } i > 52 - w, \\ Y_{52-(-i-w),y-1,ag,c} & \text{if } i < 0 \text{ and } -i \geq w. \end{cases}$$

$I(w+i; w, y, ag, c)$ is a binary variable that indicates whether $Y^*_{w+i,y,ag,c}$ is missing ($I(w+i; w, y, ag, c)$=0) or known ($I(w+i; w, y, ag, c)$=1). $I(y+i; w, y, ag, c)$, $I(ag+i; w, y, ag, c)$, $I(c_i; w, y, ag, c)$ are defined analogously and indicate whether the values of year $y + i$, age group $ag + i$, and city c_i are missing or not. Thus, if the number of deaths corresponding to some weeks, years, age groups or cities are missing, the associated binary variables would be 0, and they would not be taken into account in the estimation of the missing value. Likewise, if any of m_W, m_Y, m_{AG} or m_C is equal to 0, the number of deaths corresponding to weeks, years, age groups or cities would not be taken into account in the estimation since the corresponding sets of indexes would be empty. Note also that for some i, $Y^*_{w+i,y,ag,c}$, $Y_{w,y+i,ag,c}$, $Y_{w,y,ag+i,c}$ or Y_{w,y,ag,c_i} could not exist. For example, if $y = n_Y$, $Y_{w,y+1,ag,c}$ does not exist. For the estimation of $Y_{w,y,ag,s}$, the terms corresponding to non existent values will be eliminated in the formula of the regression.

The $\widehat{\beta}$'s needed for the estimation of the missing deaths in the age group ag and the city c in a given year and week, are obtained from the fit of the following Poisson regression model,

$$\begin{aligned} log(E[\boldsymbol{Y_{ag,c}}])) \;=\; & \beta^0 + \sum_{i \in I_W} \beta^{W+i} I(w+i; w, y, ag, c) \boldsymbol{Y_{ag,c,(w+i)}} \\ & + \sum_{i \in I_Y} \beta^{Y+i} I(y+i; w, y, ag, c) \boldsymbol{Y_{ag,c,(y+i)}} \\ & + \sum_{i \in I_{AG}} \beta^{AG+i} I(ag+i; w, y, ag, c) \boldsymbol{Y_{ag,c,(ag+i)}} \\ & + \sum_{i \in I_C} \beta^{C_i} I(c_i; w, y, ag, c) \boldsymbol{Y_{ag,c,(c_i)}}, \end{aligned}$$

where $\boldsymbol{Y_{ag,c}}$ is the response vector which contains all deaths in city c and age group ag, in years 1 to n_Y, and weeks 1 to n_W. And the vectors $\boldsymbol{Y_{ag,c,(w+i)}}$, $i \in I_W$, $\boldsymbol{Y_{ag,c,(y+i)}}$, $i \in I_Y$, $\boldsymbol{Y_{ag,c,(ag+i)}}$, $i \in I_{AG}$, and $\boldsymbol{Y_{ag,c,(c_i)}}$, $i \in I_C$, are the covariates used in the regression. Given that $Y_{w,y,ag,c}$ is the j^{th} component of $\boldsymbol{Y_{ag,c}}$, the j^{th} components of $\boldsymbol{Y_{ag,c,(w+i)}}$, $\boldsymbol{Y_{ag,c,(y+i)}}$, $\boldsymbol{Y_{ag,c,(ag+i)}}$ and $\boldsymbol{Y_{ag,c,(c_i)}}$ are $Y^*_{w+i,y,ag,c}$, $Y_{w,y+i,ag,c}$, $Y_{w,y,ag+i,c}$ and Y_{w,y,ag,c_i} respectively.

4.3 Model evaluation

We specify bounds for the model predictors as follows:

$$0 \leq m_W, m_Y, m_{AG} \leq 5$$

$$m_C = 0 \text{ or } 3$$

Using these bounds, we consider 25 different model specifications to estimate missing values in the all-cause data. These models differ in the subset of predictors specified in the regression model. The details of the various specifications are shown in Table 4.2. The regression model performed very poorly in cities with more than 70% missing values and those cities were excluded in the following computations. To evaluate goodness-of-fit, we cross-validate by estimating known values in the all-cause data set using each one of the models, then computing prediction errors, i.e. differences between the true and imputed number of deaths. We then computed the mean and the variance of the errors overall and stratified by age group. We considered lower absolute values of the mean and the variance of the errors as indicators of superior model performance. We summarize the results of these evaluations in tables 4.3 and 4.4.

Models 1, 2 and 3, which use only information from prior or upcoming years but no other predictors, showed higher mean errors compared to all other models. Within a given set of predictors, variances tend to decrease as we increase the bounds m_Y, m_W, m_{AG} and m_C and thus incorporate more information. We also note that models with only one type of predictor (e.g. only years or only age groups) generally have larger variances than models that utilize multiple types of neighboring data. Considering only models 11, 14, 17, 20 and 23, we observe that the model with generally smaller error variances is model 11. Considering small mean errors, models 23 performed the best. For overall performance across several age groups, we selected model 11. This model uses information from the 5 previous and 5 following years, weeks and age groups, and the 3 closest cities. With this model, the error variance obtained is the minimum of all models for age groups 25-44 years and 45-64 years, and for all age groups combined. The variance of the rest of age groups are very close to the minimum variances obtained across all models. When we compare models 11, 12 and 13, we note they have similar error means and variances. Variances improve only slightly as we increase the maximum number of years, weeks and age groups included in the regression model.

4.4 Revised national estimates

We described our model evaluation process and our eventual model selection in the previous section. We then impute the missing all-cause and P&I data using the same

	Years	Weeks	Age groups	Cities
	m_Y	m_W	m_{AG}	m_C
Model 1	5	0	0	0
Model 2	4	0	0	0
Model 3	3	0	0	0
Model 4	0	5	0	0
Model 5	0	4	0	0
Model 6	0	3	0	0
Model 7	0	0	5	0
Model 8	0	0	4	0
Model 9	0	0	3	0
Model 10	0	0	0	3
Model 11	5	5	5	3
Model 12	4	4	4	3
Model 13	3	3	3	3
Model 14	5	5	5	0
Model 15	4	4	4	0
Model 16	3	3	3	0
Model 17	5	5	0	3
Model 18	4	4	0	3
Model 19	3	3	0	3
Model 20	5	0	5	3
Model 21	4	0	4	3
Model 22	3	0	3	3
Model 23	0	5	5	3
Model 24	0	4	4	3
Model 25	0	3	3	3

Table 4.2: Model specifications to estimate missing values in the all-cause deaths during 1962-2004. Maximum number of previous and following years, weeks, age groups and maximum number of closest cities used in each model.

	M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8	M. 9	M. 10	M. 11	M. 12	M. 13
	years			weeks			age			cities	years, weeks, age, cities		
< 28 days	-0.16	-0.17	-0.2	-0.01	-0.01	-0.01	**0**	-0.02	0	0	-0.09	-0.09	-0.09
28 days-1 year	0.08	0.08	0.06	-0.01	-0.01	-0.01	**0**	0	-0.97	0	-0.03	-0.02	-0.02
1-14 years	**0**	0	0	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0	-0.01	-0.02	-0.02
15-24 years	-0.08	-0.07	-0.06	-0.01	-0.01	-0.01	**0**	0	0	0	-0.05	-0.04	-0.03
25-44 years	-0.07	-0.06	-0.07	**-0.01**	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.05	-0.05	-0.06
45-64 years	-0.09	-0.06	-0.06	-0.02	**0**	-0.02	-0.04	-0.04	-0.05	0.02	-0.03	-0.05	-0.01
> 65 years	-0.59	-0.51	-0.42	-0.03	**0.02**	-0.02	-0.08	-0.08	-0.08	-0.02	-0.52	-0.46	-0.42
All age groups	-0.13	-0.12	-0.11	-0.01	**0**	-0.01	-0.02	-0.02	-0.16	0	-0.11	-0.1	-0.09

	M. 14	M. 15	M. 16	M. 17	M. 18	M. 19	M. 20	M. 21	M. 22	M. 23	M. 24	M. 25
	years, weeks, age groups			years, weeks, cities			years, age, cities			weeks, age, cities		
< 28 days	-0.1	-0.1	-0.11	-0.09	-0.09	-0.1	-0.1	-0.11	-0.12	-0.01	0	0
28 days-1 year	-0.02	-0.01	-0.01	-0.04	-0.04	-0.04	0.01	0.02	0	-0.01	-0.01	-0.01
1-14 years	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
15-24 years	-0.05	-0.04	-0.04	-0.05	-0.05	-0.04	-0.06	-0.05	-0.04	-0.01	-0.01	-0.01
25-44 years	-0.05	-0.04	-0.05	-0.05	-0.04	-0.05	-0.06	-0.06	-0.07	-0.02	-0.01	-0.02
45-64 years	-0.05	-0.07	-0.03	-0.12	-0.08	-0.09	0.01	-0.02	0.02	-0.02	-0.01	-0.01
> 65 years	-0.49	-0.44	-0.41	-0.47	-0.36	-0.35	-0.56	-0.55	-0.45	-0.04	-0.03	-0.05
All age groups	-0.11	-0.1	-0.1	-0.12	-0.1	-0.1	-0.11	-0.11	-0.09	-0.02	-0.01	-0.01

Table 4.3: Expectation of the prediction errors of the different models proposed to estimate all-cause deaths during 1962-2004.

	M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8	M. 9	M. 10	M. 11	M. 12	M. 13
	years			weeks			age			cities	years, weeks, age, cities		
< 28 days	6.56	6.66	6.87	6.33	6.47	6.83	8	92.7	8.07	7.94	5.64	10.37	**5.48**
28 days-1 year	7	11.56	181.27	2.75	2.82	2.99	5.87	5.96	237947.68	6.45	3.25	3.22	4.05
1-14 years	1.83	1.86	1.86	1.76	1.76	1.8	10.37	10.37	11.85	2.21	2.36	4.25	4.49
15-24 years	4.1	4.14	4.31	3.54	3.6	3.66	4.81	4.81	4.81	5.65	3.16	3.19	3.27
25-44 years	23.44	23.55	23.66	20.17	20.46	21.39	28.22	28.22	28.67	35.6	**14.4**	14.71	14.75
45-64 years	91.27	98.14	99.72	85.73	86.21	90.71	211.47	209.16	297.15	138.57	**53.65**	57.1	57.78
> 65 years	549.67	585.85	593.17	496.28	495.98	515.08	1366.83	1352.18	1399.63	664.29	283.66	**282.24**	285.93
All age groups	97.74	104.57	130.14	88.08	88.19	91.78	233.65	243.34	34242.55	122.96	**52.33**	53.6	53.7
	M. 14	M. 15	M. 16	M. 17	M. 18	M. 19	M. 20	M. 21	M. 22	M. 23	M. 24	M. 25	
	years, weeks, age			years, weeks, cities			years, age, cities			weeks, age, cities			
< 28 days	5.64	10.94	5.53	5.6	5.63	5.71	6.1	11.9	6.22	6.02	8.12	6.74	
28 days-1 year	2.87	2.82	4.07	3.59	4.2	6.68	4.07	4.16	23.82	**2.55**	2.57	4.49	
1-14 years	2.48	4.83	5.02	1.72	**1.71**	1.72	2.55	3.51	3.73	6.47	6.32	7.25	
15-24 years	**3.14**	3.17	3.23	3.49	3.53	3.58	3.51	3.53	3.61	3.22	3.27	3.35	
25-44 years	14.46	14.7	14.97	18.52	18.65	18.85	17.14	17.17	17.25	14.95	15.21	15.26	
45-64 years	53.94	57.43	58.69	73.02	75.1	76.42	61.01	63.76	64.51	61.77	65.12	69.11	
> 65 years	302.84	304.5	315.1	408.43	417.36	423.04	300.77	296.58	302.16	290.87	298.83	301.88	
All age groups	55.08	56.93	58.1	73.5	75.18	76.58	56.48	57.26	60.21	55.12	57.06	58.3	

Table 4.4: Variance of the prediction errors of the different models proposed to estimate all-cause deaths during 1962-2004.

model (model 11) and compare the reported estimates to those after imputation. We do not include deaths corresponding to cities with more than 70% missing values. National estimates are presented in Figure 4.2. For each of the years, we obtain higher mortality estimates imputing missing values than assuming they are zero counts. We observe markedly differences in the all-cause mortality patterns obtained with these two procedures. Specifically, the mortality obtained imputing the missing data shows an increase of mortality in years 1969 to 1970, 1986 to 1987, and 2003 to 2004, whereas the mortality using zero counts shows a decrease in these periods of time. Likewise, in years 1993 to 1994, and 1998 to 1999 there is a decrease in mortality obtained with the model and an increase using zero counts. Also, from 1997 to 2004 we obtain higher mortality using the model than using using zero counts. We can also observe differences in the patterns of P&I mortality. The trend over time is very similar using both procedures but higher mortality burden is observed imputing missing values in years where the percentage of missing values is greater than 0.10% (1965, 1970, 1976, 1981 and 1987), and years 1993, 1998 and 2004.

We estimated P&I excess mortality in seasons 1972/1973 to 2003/2004 using the Serfling methods preferred by 122 CMRS researchers and described in Serfling (1963). In Table 4.5 and Figure 4.3 we present side-by-side the excess deaths estimated using the data with missing values imputed, and using the data treating missing values as zero counts. Some differences can be observed. Using our model excess deaths are not estimated in season 1994/1995 whereas 57.62 excess deaths are estimated replacing the missing values by zeros. The biggest difference is observed in season 2003/2004 where 403.13 less excess deaths are estimated using the imputation model.

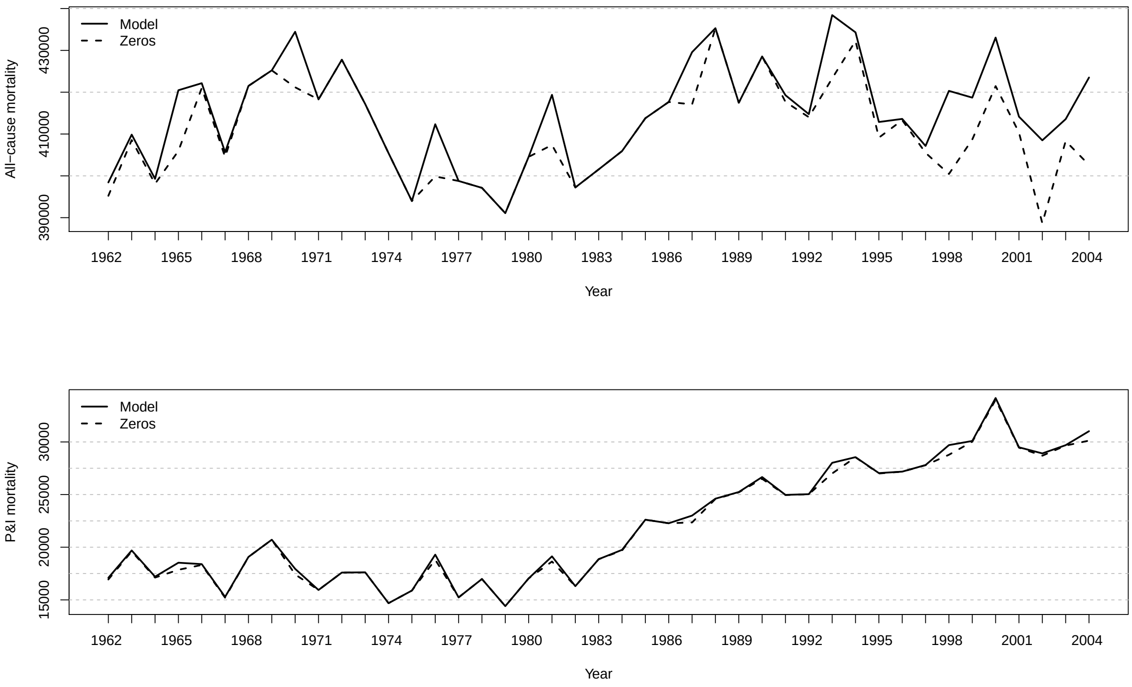


Figure 4.2: National estimates of all-cause (above) and P&I (below) mortality during 1962 to 2004. Zeros represents deaths obtained when replacing missing data by zeros. Model represents deaths obtained when replacing missing values by our regression model estimates.

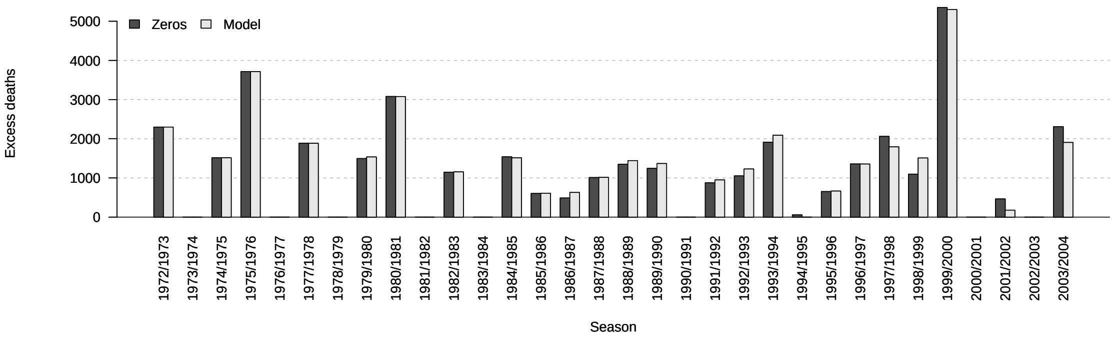


Figure 4.3: Seasonal P&I excess deaths obtained with Serfling's method during 1972/1973 to 2003/2004. Zeros represents excess deaths obtained when replacing missing data by zeros. Model represents excess deaths obtained when replacing missing values by our regression model estimates.

	Zeros	Model
1972/1973	2298.72	2298.43
1973/1974	0	0
1974/1975	1515.35	1514.59
1975/1976	3714.91	3713.92
1976/1977	0	0
1977/1978	1886.42	1884.86
1978/1979	0	0
1979/1980	1494.94	1536.21
1980/1981	3080.12	3077.72
1981/1982	0	0
1982/1983	1147.66	1154.83
1983/1984	0	0
1984/1985	1539.57	1512.67
1985/1986	604.1	609.36
1986/1987	488.04	630.29
1987/1988	1010.54	1014.23
1988/1989	1350.98	1441.97
1989/1990	1244.93	1368.43
1990/1991	0	0
1991/1992	877.09	949.63
1992/1993	1052.82	1228.98
1993/1994	1911.91	2090.32
1994/1995	57.62	0
1995/1996	652.77	665.56
1996/1997	1357.17	1355.06
1997/1998	2061.06	1793.38
1998/1999	1097.3	1509.64
1999/2000	5350.84	5299.97
2000/2001	0	0
2001/2002	464.12	177.16
2002/2003	0	0
2003/2004	2309.66	1906.53

Table 4.5: Seasonal P&I excess deaths obtained with Serfling's method during 1972/1973 to 2003/2004. Zeros represents excess deaths obtained when replacing missing data by zeros. Model represents excess deaths obtained when replacing missing values by our regression model estimates.

Chapter 5

Gaussian component mixtures and CAR models in Bayesian disease mapping

The representation and analysis of disease incidence or mortality data has become established as a basic tool in the analysis of regional public health data. The growing interest in the distribution of certain diseases among epidemiologists and medical geographers together with the advancements of computational techniques has led to substantial advances in disease mapping (Ugarte et al., 2006; Lawson and Banerjee, 2010; Lawson, 2009a). Disease maps provide a rapid visual summary of complex geographic information and may identify subtle patterns in the data that are missed in tabular presentations.

The importance of spatial dependences in the data and the underlying process of interest have long been recognized by scientists. The application of traditional covariance-based spatial statistical models is inappropriate or computationally inefficient in many problems. In recent years, the application of Bayesian hierarchical spatial and spatio-temporal models have become increasingly popular since the advances in computational techniques, such as Markov chain Monte Carlo (MCMC) methods. Modeling spatial interactions that arise in spatially referenced data is commonly done by incorporating the spatial dependence into the covariance structure via an autoregressive model (Besag et al., 1991). In the case of irregular lattice data, a common autoregressive model used is the conditional autoregressive (CAR) model. This model produces spatial dependence in the covariance structure as a function of a neighbor matrix and often a fixed unknown spatial correlation parameter. It is also possible to use other structures which can mimic spatial correlation effects, such as, for example, a Gaussian component mixture (GCM). In this chapter, we compare the performance of GCM and CAR structures both in univariate and multivariate extensions where multiple diseases are analyzed.

The outline of this chapter is as follows. First, we review the basics of the the CAR and the GCM structures. Then a simulation study is made where we fit several models that incorporate these structures to different data sets simulated from a variety of models for the true underlying risk. Components of the simulated models are correlated heterogeneity (CH), uncorrelated heterogeneity (UH), trend and covariates. We use both real and simulated covariates, some of them with values that produce patterns of increased risk, to encompass a wide range of situations that can be found in real settings. In terms of goodness-of-fit we examine the posterior summaries such as the mean of the parameters, the mean square error (MSE), and also display maps that show the geographic patterns of the parameter values. Based on these results, the comparison between the models is made. In Section 5.2 the properties of Multivariate CAR and Multivariate GCM distributions are derived. A simulation study is carried out to assess the performance of each one of the approaches. We consider three diseases and generate different datasets of the true risk, choosing different CH and UH components to reflect different types of dependence between regions and among diseases. Next, we fit to each dataset models which incorporate Multivariate CAR and Multivariate GCM components. The performance of the models is assessed by means of the MSE and the parameter estimates obtained with each model. Finally, in Section 5.3, we model the CH of a real data set using GCM and CAR structures, and compare the results obtained with each one. Specifically, we estimate the relative risk of low birth weight in Georgia, U.S., in the year 2000 adjusting for the median household income and the percentage of poverty.

5.1 Bayesian disease mapping models

A common Bayesian model for disease mapping is the following three level hierarchical model:

$$Y_i \sim Po(E_i \times \theta_i) \quad i = 1, \ldots, n;$$

$$log(\theta_i) \sim p(\cdot|\phi),$$

$$\phi \sim \pi(),$$

where Y_i and E_i are respectively the observed and the expected number of cases of disease in area i, θ_i is the relative risk in area i, $p(\cdot|\phi)$ is an appropriate prior distribution for the $\{\theta_i\}$ and ϕ are hyperparameters with hyperprior distributions $\pi()$ (Lawson, 2009a). A correlated heterogeneity (CH) component is introduced at the second hierarchical level to model the spatial dependence between the relative risks. Also, an unstructured exchangeable component that models uncorrelated noise (UH) can be included as well as other terms such as covariates and trend effects.

5.1.1 Conditional Autoregressive (CAR) models

To model CH, a Gaussian Markov random field prior distribution is most commonly used in disease mapping. These models are usually specified by a set of area-specific spatially correlated Gaussian random effects

$$u_i, \quad i = 1, \ldots, n;$$

where n is the number of areas in the study region. Their joint distribution is expressed as follows:

$$\mathbf{u} \sim MVN(\mu, v\Sigma),$$

where $\mathbf{u'} = (u_1, \ldots, u_n)$, μ is the mean vector, $v > 0$ controls the overall variability of the u_i and Σ is a positive definite matrix. This joint multivariate Gaussian model can also be expressed in the form of a set of conditional distributions by writing the between-area covariance matrix in the following form:

$$v\Sigma = v(I - \gamma C)^{-1} D,$$

where I is the identity matrix, D is a diagonal matrix with elements D_{ii} proportional to the conditional variance of $u_i|u_j$, C is a weight matrix with elements C_{ij} reflecting spatial association between areas i and j, and γ controls the overall strength of the spatial dependence. Thus, the series of conditional distributions may be written as

$$u_i|\mathbf{u_{-i}} \sim N(\mu_i + \sum_{j=1}^{n} \gamma C_{ij}(u_j - \mu_i), vD_{ii}).$$

Various constraints are needed on the values of C, D and γ in order to ensure that Σ is symmetric positive definite: Σ is only symmetric if $C_{ij}D_{jj} = C_{ji}D_{ii}$, $Var(u_i|u_j) = vD_{ii} > 0$ so D_{ii} must be > 0. To ensure Σ is positive definite, γ must lie between γ_{min} and γ_{max} where γ_{min}^{-1} and γ_{max}^{-1} are the smallest and largest eigenvalues of $D^{-1/2}CD^{-1/2}$. In practice, we often expect positive spatial dependence, so constrain the prior for γ to be between 0 and γ_{max}. $\gamma = 0$ implies no spatial dependence.

Besag et al. (1991) proposed an intrinsic version of this CAR model in which the covariance matrix Σ is not positive definite. Their model corresponds to choosing $C_{ij} = {n_{\delta_i}}^{-1}$ if areas i and j are adjacent and $C_{ij} = 0$ otherwise (with C_{ii} also set to 0) and $D_{ii} = {n_{\delta_i}}^{-1}$. Moreover, $\gamma = \gamma_{max}$ which turns out to always be 1 with this particular choice of C_{ij} and D_{ii}. Here n_{δ_i} is the number of areas which are adjacent to area i. This leads to the following model for the conditional distribution

$$u_i|\mathbf{u_{-i}} \sim N(\ \overline{u}_{\delta_i}, \frac{v}{n_{\delta_i}}),$$

where $\overline{u}_{\delta_i} = \dfrac{1}{n_{\delta_i}} \sum\limits_{j \in \delta_i} u_j$ and δ_i denotes the set of labels of the neighbors of area i. Hence u_i has a normal distribution with conditional mean given by the average of the neighboring u_j's and conditional variance inversely proportional to the number of neighbors n_{δ_i}.

5.1.2 Gaussian component Mixture (GCM) models

Another structure to model spatial correlation can be considered (Langford et al., 1999). The Gaussian component mixture (GCM) proposed consists of random effects u_i, $i = 1, \dots, n$ which are defined as:

$$u_i = \sum_{j}^{n_{\delta_i}} w_{ij}{}^{*} \lambda_j, \text{ where}$$

$$\lambda_j \overset{ind}{\sim} N(0, {\tau_\lambda}^{-1}) \quad \text{and} \quad w_{ij}{}^{*} = \frac{w_{ij}}{\sum_j^{n_{\delta_i}} w_{ij}}.$$

Here τ_λ is the precision of the normal distributions and $w_{ij} \geq 0 \; \forall i, j \in \{1, \dots, n\}$. If we denote $w_i = \sum\limits_{j}^{n_{\delta_i}} w_{ij}$, u_i can be written as

$$u_i = \frac{1}{w_i} \sum_{j}^{n_{\delta_i}} w_{ij} \lambda_j.$$

Here note that

$$E(u_i) = 0 \text{ and } Var(u_i) = \frac{{\tau_\lambda}^{-1}}{{w_i}^2} \sum_{j}^{n_{\delta_i}} {w_{ij}}^2,$$

and therefore this makes the variability of the field locally variable as in the case of the CAR model. The covariance is expressed as

$$Cov(u_i, u_k) = \frac{{\tau_\lambda}^{-1}}{w_i w_k} \sum_{h}^{n_{\delta_{ik}}} w_{ih} w_{kh},$$

where $n_{\delta_{ik}}$ is the number in common in the two neighborhoods and λ_h is the effect for those in common. If we define $W_{ik} = \sum\limits_{h}^{n_{\delta_{ik}}} w_{ih} w_{kh}$, it is straightforward to find the spatial correlation of the field as

$$\rho_{ik} = \frac{W_{ik}}{(\sum_j^{n_{\delta_i}} {w_{ij}}^2 \sum_l^{n_{\delta_k}} {w_{kl}}^2)^{1/2}}.$$

Weights are defined in such a way that pairs of locations that are close to each other have a high weight associated, whereas pairs of locations further apart have a low weight. This may be achieved using some form of distance decay function or contiguity based spatial weights. Usually weights for i are chosen so that $w_{ij} = 1$ $\forall j \in \delta_i$ and therefore $w_i = n_{\delta_i}$. Another option is to choose $w_{ij} = (n_{\delta_i})^{-1}$ $\forall j \in \delta_i$ and then $w_i = 1$. With both choices, u_i is expressed as the following

$$u_i = \frac{1}{n_{\delta_i}} \sum_{j}^{n_{\delta_i}} \lambda_j.$$

And we obtain

$$E(u_i) = 0, \quad Var(u_i) = \frac{{\tau_\lambda}^{-1} n_{\delta_i}}{{n_{\delta_i}}^2} = \frac{{\tau_\lambda}^{-1}}{n_{\delta_i}},$$

$$Cov(u_i, u_k) = \frac{{\tau_\lambda}^{-1} n_{\delta_{ik}}}{n_{\delta_i} n_{\delta_k}} \quad \text{and} \quad \rho_{ik} = \frac{n_{\delta_{ik}}}{(n_{\delta_i} n_{\delta_k})^{1/2}}.$$

This yields a positive correlation on the range (0,1). Then $\rho_{ik} = 1$ when the regions overlapping are the same and 0 when there is no overlap. Also note that the correlations are short range in that completely separate (non-overlapping) neighborhoods have 0 correlation.

It should be noted that a variant of this model can be defined within the CAR model paradigm where

$$u_i | \mathbf{u_{-i}} \sim N(\gamma \sum_{j}^{n} C_{ij} u_j, v D_{ii}).$$

Assuming $\mu_i = 0$ and in the spatial association matrix C, we define $C_{ij} = n_{\delta_{ik}}/n_{\delta_i}$ with $C_{ii} = 0$ and $D_{ii} = 1/n_{\delta_i}$; then this allows different association based on overlap of neighborhoods. Also then the partial squared correlation is defined as $\gamma^2 C_{ij} C_{ji}$ (Stern and Cressie, 1999) and so for $\gamma = 1$,

$$\rho_{ik} = \frac{n_{\delta_{ik}}}{(n_{\delta_i} n_{\delta_k})^{1/2}}.$$

However this formulation does not lead to a simple neighborhood mean structure in the conditional distribution $[u_i | \mathbf{u_{-i}}]$ and so we do not pursue this here.

5.1.3 Simulation study

We would like to evaluate the behavior of the GCM in comparison to the intrinsic CAR as a model for CH. To this end, we set up a simulation study to compare the CAR and the GCM structures for modeling CH. We want to examine whether one is better at recovering the true spatial variation or if there is only slightly differences between them. To do so we simulate data sets of count distributions from a number

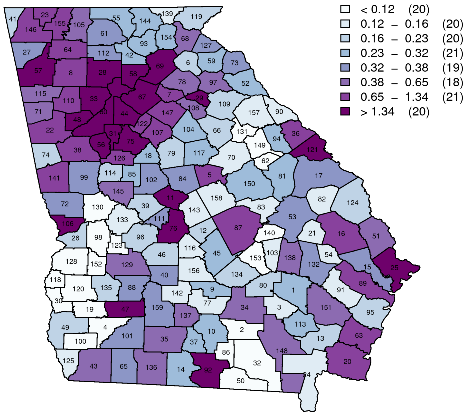


Figure 5.1: Map of the expected asthma counts in childhood in Georgia.

of different models. Some of the models have a CH component that is specified either with a CAR or with a GCM component. Two models are fitted to each simulated data set, one where the CH is modeled with a CAR component and other where a GCM component is used. To evaluate the merits of each of the fitted models we use a variety of criteria. We calculate the MSE and examine the patterns of the fitted parameters, and we observe the differences between the real parameter values and the estimated ones for biases.

To carry out the study we need to choose both a suitable map of small areas to simulate relative risks within, and a set of fixed expected cases for these small areas. The spatial structure we decide to use throughout the simulations is that of the 159 counties in the state of Georgia, U.S. It was felt important to use a real data set of expected cases for the mapped area instead of a simulated one. Although this choice has a disadvantage in that the study is relatively specific to the data, it does allow us to work close to real data. The set of fixed expected counts are chosen as the expected asthma mortality cases in childhood in the year 2000. Figure 5.1 shows the distribution of the expected counts in the map.

Simulated models

We simulate the observed cases from a Poisson distribution

$$Y_i \sim Po(E_i \times \theta_i) \quad i = 1, \ldots, n;$$

where n is the number of counties, Y_i and E_i are respectively the total number of observed and expected cases in i, and θ_i is the relative risk in i. Our purpose is to study the performance of GCM and CAR structures at modeling disease risk in a wide range of situations that can appear in real settings and, therefore, θ_i values are simulated from several models that produce different simulate count distributions in the map.

The first model that we consider is the so called convolution model for count data which takes into account uncorrelated and spatially correlated random effects. The method is usually adopted in disease mapping to obtain reliable estimates of the relative risk in those areas where the low number of observations makes the rough relative risk estimates unstable (where rough means a relative risk estimated using only observations and target population data at a given area level). We adopt two convolution models that differ in the way CH u is simulated. We call GCM the model where u is generated with a GCM structure and CAR the model which uses a CAR structure. To observe the performance of GCM and CAR structures when we try to assess covariate effects, we also need to use models in which we incorporate the effect of a covariate. To this end, we choose models GCM and CAR with covariates which are specified in the same way that GCM and CAR models adding a covariate term. It is also used as a model one with two covariate effects given by the coordinates of the counties. This model is called Trend and it tries to model the disease risk that is thought to be determined by a geographic trend. Finally, we use the model Trend with UH, which incorporates to the Trend model a term that models UH. Thus, the models used for the simulation of $\{\theta_i\}$ are the following:

- GCM or CAR: $\theta_i = exp(\alpha + u_i + v_i)$,
- GCM or CAR with covariate: $\theta_i = exp(\alpha + \alpha_{cov}\ cov_i + u_i + v_i)$,
- Trend: $\theta_i = exp(\alpha + \beta_1(x_i - \bar{x}) + \beta_2(y_i - \bar{y}))$,
- Trend with UH: $\theta_i = exp(\alpha + \beta_1(x_i - \bar{x}) + \beta_2(y_i - \bar{y}) + v_i)$,

where v is the UH, u is the CH, x and y are the centroid coordinates of the counties and cov is a covariate. The component v is simulated in all models as $v_i \sim N(0, {sd_v}^2)$, with a sd_v value fixed for the standard deviation. In GCM models the CH component is simulated with a GCM structure

$$u_i = \frac{1}{n_{\delta_i}} \sum_{j \in \delta_j} \lambda_j, \quad \lambda_i \sim N(0, {sd_\lambda}^2).$$

and in CAR models with an improper CAR structure

$$u_i|\mathbf{u_{-i}} \sim N(\overline{u}_{\delta_i}, \frac{(sd_{CAR})^2}{n_{\delta_i}}).$$

To see the performance of GCM and CAR structures when they are used to fit simulated data with covariate effects, we simulate models GCM and CAR with different covariates. Specifically, we use 15 covariates, one real and the rest simulated.

The real one is cov_{pop} and it is defined as the standardized population density of Georgia counties:

$$(cov_{pop})_i = \frac{pop_i - \overline{pop}}{(n^{-1}\sum_{i=1}^{n}(pop_i - \overline{pop})^2)^{1/2}},$$

where pop_i is the population density of county i, $i = 1, \ldots, n$, and $\overline{pop} = n^{-1}\sum_{i=1}^{n} pop_i$. The other covariates are simulated. The first is cov_n and for $i = 1, 2, \ldots, n$ it is simulated from a normal distribution with mean reflecting a geographic trend: $(cov_n)_i \sim N(0.5x_i + 0.5y_i, 0.5^2)$, where x and y are the vectors of centroid coordinates of the counties. The following is called cov_{nc} and has the same values as cov_n except for two groups of contiguous counties where the value is increased by 1. Specifically $(cov_{nc})_i = (cov_n)_i + 1$, if $i \in H_0$, the set of counties made up of county 44 and its nine nearest neighbors and county 159 and its nine nearest neighbors, and $(cov_{nc})_i = (cov_n)_i$ otherwise. Thus, this covariate reflects a trend pattern and two clusters of high values.

We are also interested to see if clusters and different degrees of high covariate values has some effect in the performance of GCM and CAR structures. To see this the rest of simulated covariates are three sets of covariates, cov_{1k}, cov_{2k} and cov_{3k} with $k \in \{1, 2, 3, 4\}$. Each set of covariates are generated in such a way that there is a group of counties with high values that form a pattern of increased risk in the map as k increases, while the rest of values are random generated and constant across the set of covariates. For $k \in \{1, 2, 3, 4\}$ cov_{1k} is a covariate with five isolated counties of high values that represent five hot spots, cov_{2k} has one big cluster of high values and cov_{3k} contains a single big cluster on the left of the map. To simulate these covariates we generate first an auxiliary covariate cov_0, as $(cov_0)_i \sim N(0, 0.1^2)$ $\forall i$, and define $m_1 = 0.5$, $m_2 = 1$, $m_3 = 1.5$ and $m_4 = 2$. We also define the following sets of counties that refer to the counties with high covariate values in each set of covariates:

- H_1: this set is made up of the five hot spots counties in cov_{1k}. We choose them as five isolated single counties with expected counts approximately equal to the 10th, 30th, 50th, 70th and 90th percentiles of the distribution of the expected counts,

- H_2: this set refers to the cluster in cov_{2k}. It consists of a group of eighteen contiguous counties with center county 143,

- H_3: it is used in the definition of cov_{3k}. It is made up of sixty contiguous counties on the left side of the map.

Then, for $i = 1, 2, \ldots, n$ and $k \in \{1, 2, 3, 4\}$, the covariates are simulated as follows: $(cov_{1k})_i \sim N(m_k, 0)$ if $i \in H_1$ and $(cov_{1k})_i = (cov_0)_i$ otherwise, $(cov_{2k})_i \sim N(m_k, 0.1^2)$ if $i \in H_2$ and $(cov_{2k})_i = (cov_0)_i$ otherwise, and $(cov_{3k})_i \sim N(m_k, 0.1^2)$

if $i \in H_3$ and $(cov_{3k})_i = (cov_0)_i$ otherwise. Note that for $j \in \{1, 2, 3\}$ and $k \in \{1, 2, 3, 4\}$, cov_{jk} values in sets H_j are simulated from normal distributions with larger mean as k increases. And therefore these covariates show patterns of increased values as k increases. Maps of Georgia with the values of the covariates represented are shown in Figure 5.2.

We simulate different data sets of count distributions using the above presented models and different values for the parameters α, α_{cov}, β_1, β_2, sd_λ, sd_{CAR}, sd_v and the different covariates. The parameter specification is made taking into account that the simulated risk θ has to take values that are plausible in real settings. Thus, we tune the parameter values in order to yield a relative risk approximately between 0.5 and 4 for most counties and avoiding extreme values. This procedure is especially important in models with covariates where some counties can have a high covariate value. If the parameters are not carefully chosen in these situations, for some counties the models can yield extreme relative risk that are impossible in real settings. The combination of model and parameter values chosen is presented in Table 5.1. We simulate 500 data sets from each of the combinations to have stable results.

Fitted models

Six different hierarchical models are used to fit the simulated data sets, each assuming a Poisson likelihood for the first level and a different structure for the logarithm of the risk in the second level. The fitted models GCM, CAR, GCM with covariate and CAR with covariate are the same that the models with the same name that are used to simulate count distributions. Therefore, in GCM and CAR models $log(\theta)$ is modeled as a sum of an intercept and two random effects, one for CH and another for UH. Models GCM and CAR with covariate incorporate also the effect of a covariate. The CH effect is modeled with a GCM component in GCM models and with a CAR component in CAR models. The other two fitted models are GCM with trend and CAR with trend. These models express $log(\theta)$ as a sum of an intercept, a trend, and UH and CH effects. The trend is modeled incorporating the effect of the coordinates of the counties centroids (x and y). CH is modeled in model GCM with trend with a GCM structure, and in model CAR with trend with a CAR structure. Moreover, prior distributions are adopted for the hyperparameters of the models. In all models we choose, for the intercept, the covariate effect, the trend effects, and the UH, a normal distribution with mean 0 and standard deviation simulated from $U(0, 5)$. The prior distributions for the standard deviation of the CH components GCM and CAR is chosen as a $U(0, 5)$. If we denote the index m as the number of covariates, and set $m = 0$ in GCM or CAR models, and $m = 2$, $(cov_1)_i = (x_i - \bar{x})$ and $(cov_2)_i = (y_i - \bar{y})$, $i = 1, \ldots, n$ in GCM or CAR with trend models, the specification of the models is as follows:

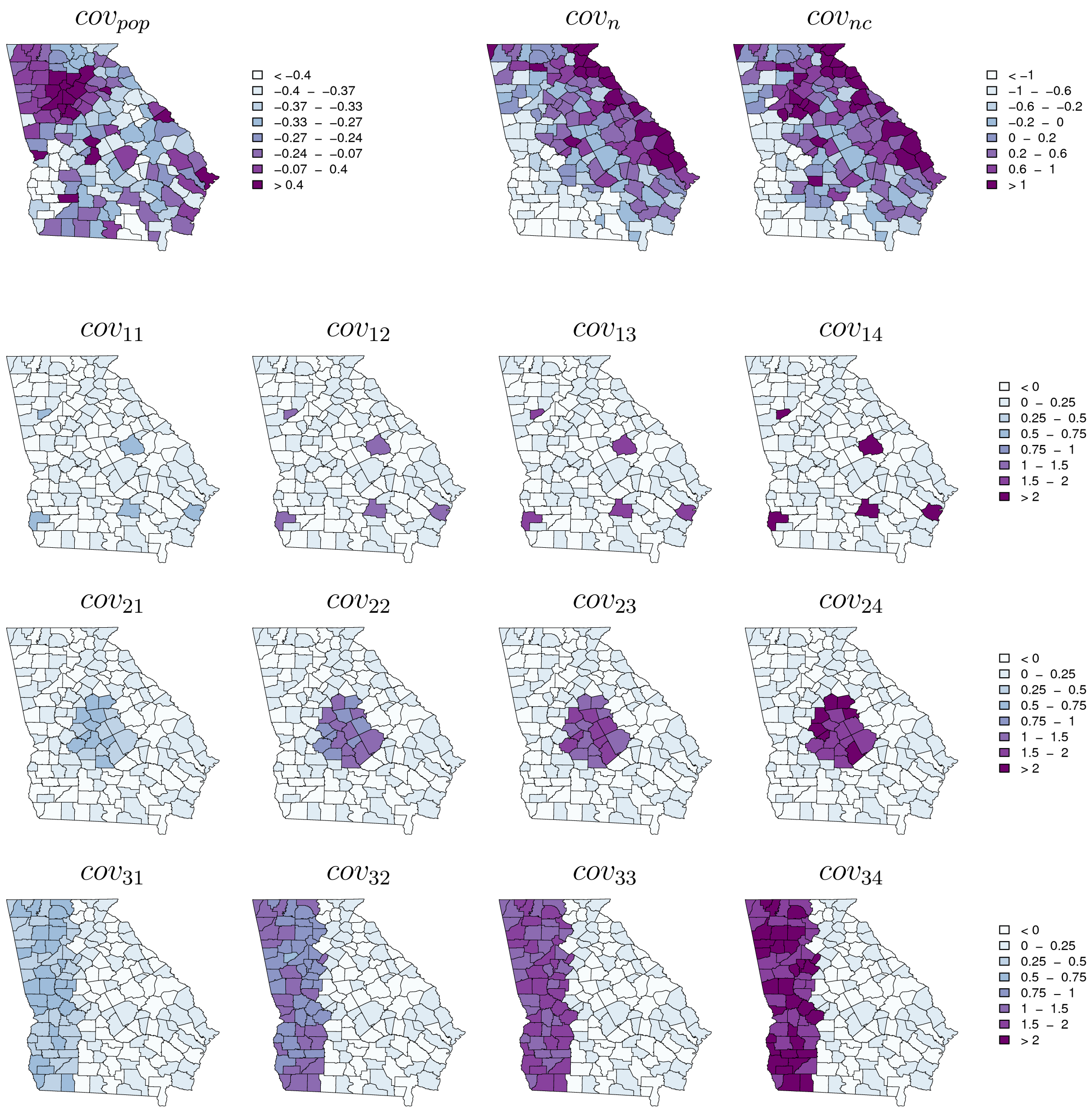


Figure 5.2: Maps of the covariates assumed in the simulated models.

Simulated model	α	α_{cov}	β_1	β_2	sd_λ	sd_{CAR}	sd_v	cov
GCM	0.1	-	-	-	1	-	1	-
	0.1	-	-	-	2	-	1	-
	0.01	-	-	-	1	-	1	-
	-0.1	-	-	-	0.25	-	1	-
CAR	0.1	-	-	-	-	1	1	-
	0.1	-	-	-	-	2	1	-
	0.1	-	-	-	-	1	2	-
	0.01	-	-	-	-	0.75	1	-
	-0.1	-	-	-	-	0.25	1	-
GCM with covariate	0.01	0.01	-	-	1	-	1	cov_{pop}
	-0.1	0.01	-	-	0.25	-	1	cov_{pop}
	0.1	0.5	-	-	0.25	-	1	cov_n
	0.1	0.5	-	-	0.25	-	1	cov_{nc}
	0.1	1	-	-	0.5	-	1	cov_{jk}; $j = 1, 2, 3$; $k = 1, 2, 3, 4$
CAR with covariate	0.01	0.01	-	-	-	0.75	1	cov_{pop}
	-0.1	0.01	-	-	-	0.25	1	cov_{pop}
	0.1	0.5	-	-	-	0.25	1	cov_n
	0.1	0.5	-	-	-	0.25	1	cov_{nc}
	0.1	1	-	-	-	0.5	1	cov_{jk}; $j = 1, 2, 3$; $k = 1, 2, 3, 4$
Trend	0.01	-	1	1	-	-	-	-
Trend with UH	0.01	-	1	1	-	-	1	-

Table 5.1: Parameter specification and covariates used in simulated models.

GCM, GCM with covariates or GCM with trend

$$Y_i \sim Po(E_i \times \theta_i)$$

$$log(\theta_i) = \alpha + \textstyle\sum_{l=1}^{m} \alpha_{cov_l} \times (cov_l)_i + v_i + u_i$$

$$\alpha \sim N(0, \tau_\alpha^{-1}),\ \alpha_{cov_l} \sim N(0, \tau_{\alpha_{cov_l}}^{-1})\ \forall l,\ v_i \sim N(0, \tau_v^{-1}),\ \lambda_i \sim N(0, \tau_\lambda^{-1}),\ u_i = \frac{1}{n_{\delta_i}} \sum_{j \in \delta_i} \lambda_j$$

$$\tau_\alpha = sd_\alpha^{-2},\ \tau_{\alpha_{cov_l}} = sd_{\alpha_{cov_l}}^{-2}\ \forall l,\ \tau_v = sd_v^{-2},\ \tau_\lambda = sd_\lambda^{-2}$$

$$sd_\alpha \sim U(0,5),\ sd_{\alpha_{cov_l}} \sim U(0,5)\ \forall l,\ sd_v \sim U(0,5),\ sd_\lambda \sim U(0,5)$$

CAR, CAR with covariates or CAR with trend

$$Y_i \sim Po(E_i \times \theta_i)$$

$$log(\theta_i) = \alpha + \textstyle\sum_{l=1}^{m} \alpha_{cov_l} \times (cov_l)_i + v_i + u_i$$

$$\alpha \sim N(0, \tau_\alpha^{-1}),\ \alpha_{cov_l} \sim N(0, \tau_{\alpha_{cov_l}}^{-1})\ \forall l,\ v_i \sim N(0, \tau_v^{-1}),\ u_i | u_{-i} \sim N(\overline{u}_{\delta_i}, \frac{\tau_u^{-1}}{n_{\delta_i}})$$

$$\tau_\alpha = sd_\alpha^{-2},\ \tau_{\alpha_{cov_l}} = sd_{\alpha_{cov_l}}^{-2},\ \forall l,\ \tau_v = sd_v^{-2},\ \tau_u = sd_u^{-2}$$

$$sd_\alpha \sim U(0,5),\ sd_{\alpha_{cov_l}} \sim U(0,5)\ \forall l,\ sd_v \sim U(0,5),\ sd_u \sim U(0,5)$$

The purpose of the study is to compare the performance of GCM and CAR structures in modeling the relative risk. Therefore, two different models, one where CH is modeled with GCM and another where CH is modeled with CAR, are fitted to each one of the simulated models (see Table 5.2). Specifically GCM and CAR are fitted to simulated models GCM and CAR, GCM with covariate and CAR with covariate are fitted to simulated models GCM with covariate and CAR with covariate, and GCM with trend and CAR with trend are fitted to simulated models Trend and Trend with UH. The fitting is made using `WinBUGS` (Lunn et al., 2000). Two MCMC chains are run for each model and data set until they achieve convergence. In particular the chains are run for a total of 80000 iterations, the first 30000 of which are discarded as burn-in and with a thin parameter of 50. In total 2000 iterations are saved.

Results

The performance of each fitted model is assessed by examining the posterior mean of the parameters averaged over the replicate data sets, and the MSE of each of the parameters. For a parameter s in a county i, $i = 1, 2, \ldots, n$, we obtain their

Fitted / Simulated	GCM	CAR	GCM with covariate	CAR with covariate	GCM with trend	CAR with trend
GCM	√	√				
CAR	√	√				
GCM with covariate			√	√		
CAR with covariate			√	√		
Trend					√	√
Trend with UH					√	√

Table 5.2: Models fitted to each one of the simulated models.

posterior mean averaged over the $R = 500$ replicate data sets computing

$$\frac{1}{R}\sum_{r=1}^{R} \ (s_i^{fitted})^r,$$

where $(s_i^{fitted})^r$ is the fitted value of $(s_i^{real})^r$, the simulated value of s in county i and replication r. The MSE of parameter s averages the squared differences between fitted and simulated values over locations in the map and replicate datasets,

$$MSE[s] = \frac{1}{R}\sum_{r=1}^{R} \ MSE^r[s] \ = \frac{1}{R}\sum_{r=1}^{R} \frac{1}{n}\sum_{i=1}^{n}\Big((s_i^{real})^r - (s_i^{fitted})^r\Big)^2.$$

We compare these values for each of the models, and examine the distribution patterns of the fitted values of the relative risk and the CH and UH components.

Tables 5.3 and 5.4 present a summary of the results obtained when we fit the data simulated with some of the models specified in Table 5.1. The data is fitted with two models (one with a GCM and another with a CAR structure). For each parameter of the simulated models, these tables show their MSE and their posterior mean averaged over the replicate data sets calculated with each of the fitted models. For each simulated model we mark in bold the value obtained with the fitted model that performs better. In MSE tables we mark the values that represent lower MSE, and in posterior mean tables the values closer to the real ones.

We identify some situations where the GCM model gives a slightly better fit than CAR. If data are simulated with GCM model with $\alpha = 0.1$, $sd_v = 1$ and $sd_u = 1$, fitted model GCM produces smaller MSE for all parameters and parameter estimates closer to the real ones. Moreover, fitted model GCM gives smaller MSE and better fit of α and sd_u if the data are simulated with the following models: GCM with $\alpha = -0.1$, $sd_v = 1$ and $sd_u = 0.25$, CAR with $\alpha = 0.1$, $sd_v = 2$ and $sd_u = 1$, and CAR with $\alpha = -0.1$, $sd_v = 1$ and $sd_u = 0.25$. We have not identified any situation where CAR performs better than GCM. In all situations GCM does better than CAR either producing smaller MSE or giving parameter estimates closer to the real ones.

We observe that when we fit data simulated with GCM, GCM performs better than CAR. We obtain lower MSE for α and u and lower MSE for θ in all the scenarios but the one simulated with $\alpha = 0.01$. Moreover, fitted α values are closer to the real ones. On the other hand, when the models are used to fit data simulated with CAR it is not clear which model, GCM or CAR, does better. We see that MSE of α is lower with GCM. But MSE of the rest of parameters and fitted values of α, sd_v and sd_u are better with GCM in some situations and with CAR in others. Regarding the results obtained when the simulated models are Trend and Trend with UH, we observe that GCM does better in θ, α and v, while CAR does better in the trend effects β_1 and β_2.

If we observe the results obtained for the simulated models GCM and CAR with covariates cov_{pop}, cov_n and cov_{nc}, we observe that in all situations the GCM estimates the spatial effect and the intercept α better. When the fitted model is GCM, MSE of u and α is lower and fitted values of α are closer to the real ones. CAR, on the contrary, does better in θ, sd_v and sd_u. Regarding α_{cov}, we see that GCM does better when the simulated models are CAR. We also observe that sd_v and sd_u are better fitted with CAR.

Regarding the results obtained when we fit the data simulated with models GCM and CAR with cov_{jk}, $j = 1, 2, 3$ and $k = 1, 2, 3, 4$, we see that GCM does better in the spatial effect u and in the intercept α, and CAR does better in the relative risk θ. MSE of the UH v is lower with CAR if the simulated model is GCM. When the simulated model is CAR, fitted models GCM and CAR do better in the half of situations. We also see that across the simulated models GCM, fitted model GCM produces lower MSE of α_{cov}. If the simulated models are CAR, MSE of α_{cov} is sometimes better with GCM and other times better with CAR. Regarding the fitted values, we see that GCM does better estimating sd_v and CAR does better estimating sd_u. Moreover, α is better fitted with GCM if the simulated model is GCM and better fitted with CAR if the simulated model is CAR. The fitted values of the covariate effect are estimated better with GCM in half of the situations and with CAR in the other half of situations.

For all simulated models, we have produced maps of the posterior mean of the relative risk, and the CH and UH effects obtained with each fitted model, average across data sets. Across many simulated models, we have found that the distribution patterns of the parameters are very similar with fitted models with GCM and CAR structures. Figure 5.3 depicts maps of the average across data sets posterior mean relative risk, the CH and UH components obtained with fitted models GCM and CAR with covariate when fitting the data sets simulated with GCM with cov_{11}, $\alpha = 0.1$, $\alpha_{cov} = 1$, $sd_\lambda = 0.5$ and $sd_v = 1$. We can see in Table 5.3, that for this model GCM does better in spatial effect. The MSE of CH component u obtained with GCM model is smaller than the MSE obtained with CAR model. CAR does slightly better on θ and on the UH component v. In Table 5.3, we can see that MSE of θ and MSE of v are slightly lower with the fitted CAR model. Maps in Figure 5.3 show only little differences between fitted values with each model.

Conclusions

The similarity of the results obtained using GCM and CAR suggests that the two models are not so different. The theoretical similarity between them can be checked comparing the correlation of two regions for the GCM model, to the correlation to those arising from the improper CAR model. As indicated in Section 5.1.2, the correlation between two random effects u_i and u_k, $i = 1, \dots, n$ of a GCM structure with weights $w_{il} = 1, \forall i$ and $\forall l \in \delta_i$, is given by

$$Cor_{GCM}(u_i, u_k) = \frac{n_{\delta_{ik}}}{(n_{\delta_i} n_{\delta_k})^{1/2}}.$$

The correlation between u_i and u_k for the improper CAR can be approximated by the following expression (Assunçao and Krainski, 2009),

$$Cor_{CAR}(u_i, u_k) \approx \frac{1}{(n_{\delta_i} n_{\delta_k})^{1/2}} \frac{a_{ik} + \sum_l a_{il} a_{lk} / n_{\delta_l}}{(1 + 1/n_{\delta_i} \sum_l a_{il} a_{li} / n_{\delta_l})^{1/2} (1 + 1/n_{\delta_k} \sum_l a_{kl} a_{lk} / n_{\delta_l})^{1/2}},$$

where $a_{ik} = 1$ if areas i and k are neighbors, and 0 otherwise. We note that if areas i and k are not neighbors and have no common neighbors, $n_{\delta_{ik}}$=0, $a_{ik} = 0$ and $\sum_l a_{il} a_{lk} / n_{\delta_l} = 0$. Then, both for GCM and CAR structures the correlation between the two areas' effects is 0. We also see that for GCM the correlation between u_i and u_k is positively associated with the number of common neighbors, and negatively associated with the number of neighbors of each area. For the CAR structure the correlation expression is not so simple but a similar relationship holds. $Cor_{CAR}(u_i, u_k)$ is positively associated with a weighted sum of the common areas of i and k with weights equal to the number of neighbors of each of the common areas ($\sum_l a_{il} a_{lk} / n_{\delta_l}$). Also, $Cor_{CAR}(u_i, u_k)$ is negatively associated with a weighted sum of the number of neighbors of area i ($\sum_l a_{il} a_{li} / n_{\delta_l}$) and area k ($\sum_l a_{kl} a_{lk} / n_{\delta_l}$). Hence, the general pattern of great similarity of the two models can be explained by the two similar correlations. We illustrate numerically this similarity by computing $Cor_{GCM}(u_i, u_k)$ and the approximation to $Cor_{CAR}(u_i, u_k)$ for each pair of counties in Georgia and summarizing the differences between them. We observe that 89.87% of the correlations are within 1% of each other, and 90.10% of them are within 5%. Moreover, the proportions of correlations within 20% and 40% of each other are 96.28% and 99.74% respectively. The maximum difference between correlations is 60.78%.

Simulated model					MSE θ		MSE α		MSE α_{cov}		MSE v		MSE u	
	α	α_{cov}	sd_v	sd_u	GCM	CAR	GCM	CAR	GCM	CAR	GCM	CAR	GCM	CAR
GCM	0.1	-	1	1	**3.714**	3.715	**0.026**	0.028	-	-	**0.712**	0.717	**0.213**	0.229
GCM	0.1	-	1	2	**7.488**	7.516	**0.051**	0.058	-	-	0.798	**0.791**	**0.63**	0.652
GCM	0.01	-	1	1	3.203	**3.191**	**0.028**	0.031	-	-	0.728	**0.724**	**0.219**	0.229
GCM	-0.1	-	1	0.25	**2.105**	2.113	**0.029**	0.03	-	-	**0.726**	0.733	**0.051**	0.069
CAR	0.1	-	1	1	4.359	**4.32**	**0.027**	0.027	-	-	0.758	**0.745**	0.317	**0.303**
CAR	0.1	-	1	2	10.976	**10.905**	**0.039**	0.059	-	-	0.938	**0.856**	0.909	**0.803**
CAR	0.1	-	2	1	**40.247**	40.542	**0.052**	0.07	-	-	**1.785**	1.808	**0.359**	0.425
CAR	0.01	-	1	0.75	2.98	**2.969**	**0.025**	0.025	-	-	0.736	**0.733**	**0.198**	0.199
CAR	-0.1	-	1	0.25	**2.139**	2.141	**0.025**	0.027	-	-	**0.734**	0.738	**0.061**	0.08
GCM with cov_{pop}	0.01	0.01	1	1	**3.271**	3.272	**0.032**	0.034	0.01	**0.009**	**0.735**	0.735	**0.214**	0.227
GCM with cov_n	0.1	0.5	1	0.25	3.456	**3.418**	**0.025**	0.026	**0.038**	0.042	0.704	**0.697**	**0.062**	0.064
GCM with cov_{11}	0.1	1	1	0.5	3.048	**3.021**	**0.025**	0.026	**0.77**	0.787	0.704	**0.701**	**0.085**	0.1
GCM with cov_{22}	0.1	1	1	0.5	4.037	**3.992**	**0.024**	0.026	**0.153**	0.159	0.69	**0.687**	**0.098**	0.104
GCM with cov_{34}	0.1	1	1	0.5	**17.913**	17.921	0.03	**0.029**	**0.013**	0.014	**0.573**	0.577	**0.074**	0.087
CAR with cov_{pop}	-0.1	0.01	1	0.25	2.083	**2.071**	**0.031**	0.032	0.007	**0.007**	0.742	**0.732**	**0.074**	0.074
CAR with cov_{nc}	0.1	0.5	1	0.25	3.56	**3.524**	**0.026**	0.027	**0.031**	0.035	0.694	**0.688**	**0.062**	0.067
CAR with cov_{12}	0.1	1	1	0.5	3.483	**3.464**	**0.024**	0.025	0.416	**0.413**	0.708	**0.703**	**0.122**	0.124
CAR with cov_{23}	0.1	1	1	0.5	5.903	**5.867**	**0.024**	0.025	**0.052**	0.054	**0.682**	0.684	**0.116**	0.127
CAR with cov_{34}	0.1	1	1	0.5	**15.971**	15.988	**0.028**	0.029	0.016	**0.015**	0.576	**0.576**	**0.1**	0.103

Table 5.3: MSE for the parameters of the simulated models GCM, CAR and GCM and CAR with covariates cov_{pop}, cov_n, cov_{nc}, cov_{11}, cov_{22}, cov_{34}, cov_{12} and cov_{23}, when they are fitted with models GCM, CAR and GCM and CAR with covariate. The first column specifies the simulated models with parameter values listed in columns called α, α_{cov}, sd_v and sd_u, where sd_u represents sd_λ in GCM models and sd_{CAR} in CAR models. The rest of the columns refer to the MSE of θ, α, α_{cov}, v and u.

	Simulated model				Fitted α		Fitted α_{cov}		Fitted sd_v		Fitted sd_u	
	α	α_{cov}	sd_v	sd_u	GCM	CAR	GCM	CAR	GCM	CAR	GCM	CAR
GCM	0.1	-	1	1	**0.062**	0.06	-	-	**0.932**	0.902	**1.005**	0.836
GCM	0.1	-	1	2	**0.045**	0.043	-	-	1.055	**1.01**	**1.498**	1.276
GCM	0.01	-	1	1	**−0.033**	−0.037	-	-	**0.922**	0.9	**1.025**	0.854
GCM	−0.1	-	1	0.25	**−0.113**	−0.114	-	-	**0.878**	0.841	0.799	**0.703**
CAR	0.1	-	1	1	0.043	**0.05**	-	-	**1.002**	0.96	1.168	**1.021**
CAR	0.1	-	1	2	**−0.004**	−0.03	-	-	1.281	**1.202**	**2.111**	1.843
CAR	0.1	-	2	1	**0.03**	0.001	-	-	**2.01**	1.949	1.387	**1.354**
CAR	0.01	-	1	0.75	−0.035	**−0.03**	-	-	**0.944**	0.906	0.991	**0.872**
CAR	−0.1	-	1	0.25	**−0.118**	−0.12	-	-	**0.878**	0.836	0.835	**0.743**
GCM with cov_{pop}	0.01	0.01	1	1	**−0.016**	−0.025	0.005	**0.008**	0.915	**0.916**	**1.025**	0.826
GCM with cov_n	0.1	0.5	1	0.25	**0.048**	0.048	**0.461**	0.456	0.884	**0.894**	0.876	**0.684**
GCM with cov_{11}	0.1	1	1	0.5	**0.06**	0.055	**0.595**	0.586	**0.896**	0.88	0.842	**0.717**
GCM with cov_{22}	0.1	1	1	0.5	**0.056**	0.053	0.909	**0.912**	**0.88**	0.878	0.895	**0.726**
GCM with cov_{34}	0.1	1	1	0.5	0.063	**0.068**	1.008	**0.998**	**0.946**	0.926	0.769	**0.654**
CAR with cov_{pop}	-0.1	0.01	1	0.25	**−0.111**	−0.121	**0.012**	0.015	0.834	**0.85**	0.919	**0.696**
CAR with cov_{nc}	0.1	0.5	1	0.25	**0.056**	0.053	**0.513**	0.514	0.887	**0.896**	0.83	**0.662**
CAR with cov_{12}	0.1	1	1	0.5	0.053	**0.054**	**0.699**	0.693	**0.895**	0.892	0.954	**0.769**
CAR with cov_{23}	0.1	1	1	0.5	0.053	**0.057**	**0.985**	0.971	**0.925**	0.892	0.905	**0.799**
CAR with cov_{34}	0.1	1	1	0.5	0.116	**0.096**	0.937	**0.963**	**0.975**	0.956	0.784	**0.656**

Table 5.4: Fitted values for the parameters of the simulated models GCM, CAR and GCM and CAR with covariates cov_{pop}, cov_n, cov_{nc}, cov_{11}, cov_{22}, cov_{34}, cov_{12} and cov_{23}, when they are fitted with models GCM, CAR and GCM and CAR with covariate. The first column specifies the simulated models with parameter values listed in columns called α, α_{cov}, sd_v and sd_u, where sd_u represents sd_λ in GCM models and sd_{CAR} in CAR models. The rest of the columns show the fitted values of α, α_{cov}, sd_v and sd_u.

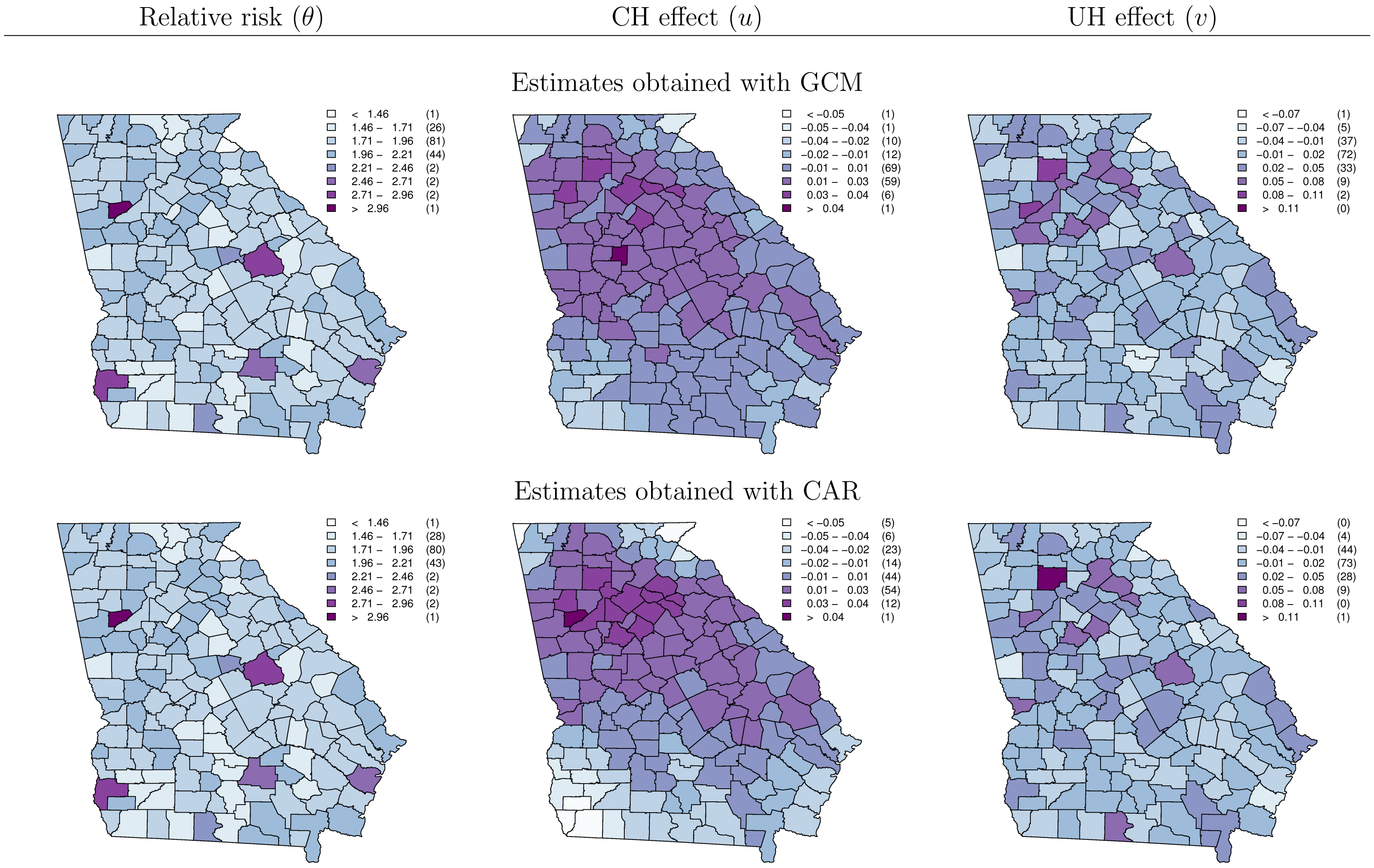


Figure 5.3: Averaged posterior mean of the relative risk, the mean CH effect and the mean UH effect, for GCM and CAR with covariate models when fitting the data sets simulated with GCM with covariate cov_{11}, $\alpha = 0.1$, $\alpha_{cov} = 1$, $sd_\lambda = 0.5$ and $sd_v = 1$.

5.2 Multiple disease mapping models

Often it is appropriate to consider the analysis of multiple diseases using multivariate spatial modeling. A number of diseases may share the same set of (spatially distributed) risk factors, or are linked by etiology, a common risk factor, or an affected organ. Moreover, the presence of one disease might encourage or inhibit the presence of another over a region. When we have information on $p \geq 2$ diseases over the same regions, an obvious first choice would be to use p separate univariate models. However, because correlation across diseases may occur we may need a multivariate spatial model to properly analyze this kind of data. This will permit modeling of dependence among those diseases while maintaining spatial dependence between regions. Identifying similar patterns in geographical variation of related diseases in a multivariate way may provide more convincing evidence for any real clustering in the underlying risk than would be available from the analysis of any single disease separately. Several multivariate areal models have been proposed to date, any of which could be applied to multiple disease mapping. Let Y_{ik} be the observed number of cases of disease k in region i, $i = 1, \ldots, n$, $k = 1, \ldots, p$, and let E_{ik} be the expected number of cases for the same disease in this same region. As in the univariate case, the Y_{ik} are thought of as random variables, while the E_{ik} are thought of as fixed and known. For the first level of the hierarchical model, we assume Y_{ik} are independent of each other such that

$$Y_{ik} \sim Po(E_{ik} \times \theta_{ik}) \quad i = 1, \ldots, n; \quad k = 1, \ldots, p.$$

In the second level, the relative risks are assigned a prior distribution $p(\cdot|\gamma)$ with hyperprior distribution $\pi()$.

$$log(\theta_{ik}) \sim p(\cdot|\gamma),$$
$$\gamma \sim \pi().$$

For example it can be assumed $log(\theta_{ik}) = \mathbf{x'_{ik}}\beta_\mathbf{k} + \phi_{ik}$ where the $\mathbf{x_{ik}}$ are explanatory, region-level spatial covariates for disease k having parameter coefficients β_j, and ϕ_{ik} are random effects (Jin et al., 2007).

5.2.1 Multivariate Conditionally Autoregressive (MCAR) models

Carlin and Banerjee (2003) and Gelfand and Vounatsou (2003) generalized the univariate CAR to a joint model for the random effects ϕ_{ik} under a separability assumption, which permits modeling of correlation among the p diseases while maintaining spatial dependence across space. The multivariate CAR (MCAR) can be viewed as a conditionally specified probability model for interactions between space and an attribute of interest. It acknowledges dependence between the diseases as well as dependence across space (Zhang et al., 2010). The multivariate intrinsic autoregressive (MIAR) distribution is an important special case of the MCAR. We give a review of a MICAR model illustrating in the case of $p = 2$ diseases.

Let areal random effects corresponding to the two diseases be $\mathbf{\Phi} = (\boldsymbol{\phi}_1', \boldsymbol{\phi}_2')$, where $\boldsymbol{\phi}_1' = (\phi_{11}, \ldots, \phi_{n1})$, $\boldsymbol{\phi}_2' = (\phi_{12}, \ldots, \phi_{n2})$, and n is the number of areal units. Under the MIAR model, the multivariate joint distribution is defined as

$$p(\mathbf{\Phi}) \propto exp\{-1/2\mathbf{\Phi}'[\Lambda \otimes (D - W)]\mathbf{\Phi}\},$$

where W is a proximity matrix whose elements w_{ij} measure the closeness of each pair of areas i and j, D is a diagonal matrix with i^{th} diagonal element equal to $\sum_j w_{ij}$, Λ is $2\times$ 2 and positive definite, and $\otimes$ denotes the Kronecker product. This corresponds to the conditional distribution

$$\begin{pmatrix}\phi_{i1}\\ \phi_{i2}\end{pmatrix} |\boldsymbol{\phi}_{-(i1,i2)} \sim N\left(\begin{pmatrix}\bar{\phi}_{i1}\\ \bar{\phi}_{i2}\end{pmatrix}, (\sum_j w_{ij}\Lambda)^{-1}\right),$$

where $\boldsymbol{\phi}_{-(i1,i2)}$ stands for the collection of all ϕ_{ij} except ϕ_{i1} and ϕ_{i2}. Let $\bar{\phi}_{i1} = \sum_j w_{ij}\phi_{j1}/\sum_j w_{ij}$ and $\bar{\phi}_{i2} = \sum_j w_{ij}\phi_{j2}/\sum_j w_{ij}$, the averages of the random effects for area i's neighbors specific to variables 1 and 2, respectively. It can be seen that Λ serves as a scaled conditional precision for (ϕ_{i1}, ϕ_{i2}) where $\sum_j w_{ij}$ is a scale parameter. Areas with more neighbors have higher precision. Since Λ is common for all areas $i = 1, \ldots, n$, it controls the conditional precision for each pair of variables at the same site averaged over all areas (Ma and Carlin, 2007).

5.2.2 Multivariate Gaussian component Mixture (MGCM) models

The Multivariate Gaussian component mixture (MGCM) consists of random effects u_{ik}, $i = 1, \ldots, n$; $k = 1, \ldots, p$ which are defined as

$$u_{ik} = \mathbf{w_i}^{*\prime}\lambda_{\mathbf{ik}},$$

where $\mathbf{w_i}^{*\prime} = (w_{i1}{}^*, \ldots, w_{in_{\delta_i}}{}^*)$ and $\lambda_{\mathbf{ik}}{}' = (\lambda_{1k}, \ldots, \lambda_{n_{\delta_i}k})$. The components of $\mathbf{w_i}^*$ are $w_{il}{}^* = \dfrac{w_{il}}{\sum_l^{n_{\delta_i}} w_{il}}$ with $w_{il} \geq 0\ \forall i, l$. In addition, $(\lambda_{l1}, \ldots, \lambda_{lp})' \overset{ind}{\sim} N(\mathbf{0}, \Sigma)\ \forall l$, with Σ being the covariance matrix of the normal distribution. If, as in the univariate Gaussian component mixture, we denote $w_i = \sum_l^{n_{\delta_i}} w_{il}$, u_{ik} can be written as

$$u_{ik} = \sum_l^{n_{\delta_i}} w_{il}{}^*\lambda_{lk} = \frac{1}{w_i}\sum_l^{n_{\delta_i}} w_{il}\lambda_{lk}.$$

We can see that $E(u_{ik}) = 0$. The expressions of the covariance between effects are the following,

$$Cov(u_{ik_1}, u_{jk_2}) = \begin{cases} \dfrac{S_{k_1k_2}}{w_i^2} \displaystyle\sum_l^{n_{\delta_i}} w_{il}^2 & \text{if } i = j, \\ \dfrac{S_{k_1k_2}}{w_i w_j} \displaystyle\sum_h^{n_{\delta_{ij}}} w_{ih} w_{jh} & \text{if } i \neq j, \end{cases}$$

where $n_{\delta_{ij}}$ is the number in common in the two neighborhoods and $S_{k_1k_2}$ is the covariance between common areas h for disease k_1 and k_2. If we define $W_{ij} = \displaystyle\sum_h^{n_{\delta_{ij}}} w_{ih} w_{jh}$, it is straightforward to find the spatial correlation of the field as

$$\rho_{ik_1,jk_2} = \frac{S_{k_1k_2} W_{ij}}{(S_{k_1k_1})^{1/2} (S_{k_2k_2})^{1/2} (\sum_l^{n_{\delta_i}} {w_{il}}^2 \sum_m^{n_{\delta_j}} {w_{jm}}^2)^{1/2}}.$$

5.2.3 Multivariate simulation study

To see the performance and compare MGCM and MCAR structures, we carry out a simulation study. It is based on the spatial layout of the $n = 159$ counties in the state of Georgia. We assume $p = 3$ diseases and use a conditionally independent Poisson likelihood to generate Y_{ik}, the observed count for disease k in area i, where $i = 1, \dots, n$ and $k = 1, \dots, p$. Thus,

$$Y_{ik} \sim Po(E_{ik} \times \theta_{ik}),$$

where E_{ik} and θ_{ik} denotes the expected count and the relative risk for disease k and area i respectively. For the computation of the expected counts, we use the observed counts of three ambulatory care-sensitive conditions: angina, asthma, and chronic obstructive pulmonary disease (COPD), as well as the population for each county in the state of Georgia in the year 2007. These data are provided by the Online Analytical Statistical Information System (OASIS; http://oasis.state.ga.us/). The expected counts for disease k in each area i is set equal to the disease rate in the state multiplied by the population in the county i, i.e. $E_{ik} = pop_i \sum_i Y_{ik} / \sum_i pop_i$, where pop_i is the total population in area i, and Y_{ik} is the count for disease k in area i. Figure 5.4 displays the expected counts for each disease in the counties of Georgia.

We generate data using six different models that encompass different ranges of relative risk variability. With each of them we generate the relative risk θ_{ik} with the following expression

$$\theta_{ik} = exp(\alpha_k + u_{ik} + v_{ik}).$$

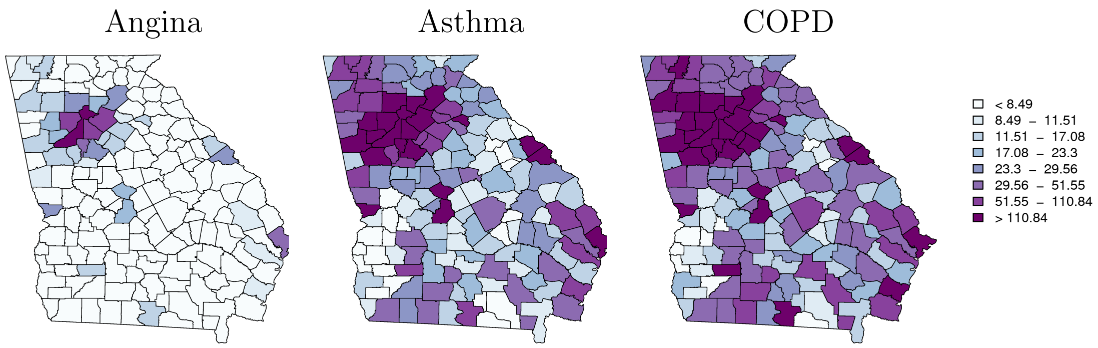


Figure 5.4: Maps of the expected counts of angina, asthma and COPD used in the simulation study.

To make sure the values of the simulated risks are sensible, that is, non-extreme and between 0.5 and 4 for most counties, we need to choose the values of α_k, u_{ik} and v_{ik} carefully. The α_k's are being fixed constant set to $\alpha_1 = 0.03$, $\alpha_2 = 0.01$ and $\alpha_3 = -0.1$. The uncorrelated component is generated with

$$v_{ik} \sim N(0, {sd_{v_k}}^2),$$

where $sd_{v_1} = 0.7$, $sd_{v_2} = 0.65$ and $sd_{v_3} = 0.6$. To generate the correlated component u_{ik} we use two different distributions. Specifically, we use an MGCM distribution for three of the models simulated, that is,

$$u_{ik} = \frac{1}{n_{\delta_i}} \sum_{l \in \delta_i} \lambda_{lk}, \text{ where } \begin{pmatrix} \lambda_{l1} \\ \lambda_{l2} \\ \lambda_{l3} \end{pmatrix} \overset{ind}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right).$$

And a MCAR distribution with covariance matrix equal to Σ for the other three models. Both in the models simulated with MGCM and MCAR distributions, we use the following three covariance matrices,

$$\Sigma_1 = \begin{pmatrix} 0.010 & 0.006 & 0.016 \\ 0.006 & 0.016 & 0.005 \\ 0.016 & 0.005 & 0.057 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.079 & 0.030 & 0.004 \\ 0.030 & 0.034 & 0.010 \\ 0.004 & 0.010 & 0.089 \end{pmatrix}, \text{ and}$$

$$\Sigma_3 = \begin{pmatrix} 0.040 & 0.032 & 0.027 \\ 0.032 & 0.044 & 0.027 \\ 0.027 & 0.027 & 0.023 \end{pmatrix}.$$

These are positive definite matrices chosen in such a way that $exp(\alpha_k + u_{ik} + v_{ik})$ yield plausible relative risk values, and with different correlations that originate different degrees of correlation between diseases. Specifically, the correlations for Σ_1

are ρ_{12}=0.44, ρ_{13}=0.68, and ρ_{23}=0.16. For Σ_2, ρ_{12}=0.59, ρ_{13}=0.05, and ρ_{23}=0.19. And for Σ_3, ρ_{12}=0.76, ρ_{13}=0.90, and ρ_{23}=0.85. Depending on the model and the covariance matrix used to generate the correlated heterogeneity, the simulated models are named MGCM Σ_1, MGCM Σ_2 and MGCM Σ_3, MCAR Σ_1, MCAR Σ_2 and MCAR Σ_3.

We use two models, MGCM and MCAR, to fit the generated data sets. Both of them express the risk as a sum of an intercept, an UH random effect and a CH random effect. In the MGCM model the CH random effect is modeled with an MGCM component. In MCAR, the CH is modeled with an MCAR structure. Specifically, they are specified as follows:

MGCM

$$Y_{ik} \sim Po(E_{ik} \times \theta_{ik}),\;\; log(\theta_{ik}) = \alpha_k + v_{ik} + u_{ik}$$

$$\alpha_k \sim N(0, \tau_{\alpha_k}^{-1}),\; v_{ik} \sim N(0, \tau_{v_k}^{-1}),\; u_{ik} = \frac{1}{n_{\delta_i}} \sum_{l \in \delta_i} \lambda_{lk},\; \begin{pmatrix} \lambda_{l1} \\ \lambda_{l2} \\ \lambda_{l3} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Omega^{-1}\right)$$

$$\tau_{\alpha_k} = sd_{\alpha_k}^{-2},\; \tau_{v_k} = sd_{v_k}^{-2},\; sd_{\alpha_k} \sim U(0,5),\; sd_{v_k} \sim U(0,5)$$

$$\Omega \sim Wishart(R,3),\; R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

MCAR

$$Y_{ik} \sim Po(E_{ik} \times \theta_{ik}),\;\; log(\theta_{ik}) = \alpha_k + v_{ik} + u_{ik}$$

$$\alpha_k \sim N(0, \tau_{\alpha_k}^{-1}),\; v_{ik} \sim N(0, \tau_{v_k}^{-1}),\; \begin{pmatrix} u_{i1} \\ u_{i2} \\ u_{i3} \end{pmatrix} |\boldsymbol{u}_{-(i1,i2,i3)} \sim N\left(\begin{pmatrix} \bar{u}_{i1} \\ \bar{u}_{i2} \\ \bar{u}_{i3} \end{pmatrix}, \Omega^{-1}\right)$$

$$\tau_{\alpha_k} = sd_{\alpha_k}^{-2},\; \tau_{v_k} = sd_{v_k}^{-2},\; sd_{\alpha_k} \sim U(0,5),\; sd_{v_k} \sim U(0,5)$$

$$\Omega \sim Wishart(R,3),\; R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

To compare the MGCM and MCAR structures, we generate 300 data sets with each of these six simulated models and we fit the models MGCM and MCAR to each one. Here, the number of data sets is restricted to 300 since more computation time is needed to fit multivariate models. The fitting is carried out using `WinBUGS`. Convergence was achieved running two MCMC chains of 80000 iterations for each model and data set. For each chain a burn-in of 30000 and a thinning rate of 50 iterations are used. In total, 2000 iterations are kept.

Results

To assess the performance of MGCM and MCAR models we compute the MSE and the estimates of the parameters with each model. In addition to this, we represent in maps the averaged posterior mean of the relative risk for each disease. The results obtained are summarized in tables 5.5 and 5.6, and Figure 5.5. They indicate a similar fit of MGCM and MCAR models.

Table 5.5 shows the MSE and the fitted values for the parameters of the simulated models MGCM Σ_1, MGCM Σ_2, MGCM Σ_3, MCAR Σ_1, MCAR Σ_2 and MCAR Σ_3, when they are fitted with models MGCM and MCAR. We observe that for all the simulated models, the MSE of all the parameters obtained with MGCM and MCAR are very similar. They are always lower with the fitted model MGCM except when the model simulated is MCAR Σ_2 and θ_1 is estimated. In this case, the MSE of θ_1 is 0.351 with the fitted model MGCM and 0.346 with the fitted model MCAR.

None of the fitted models allow us to obtain precise estimates of the real parameters $\alpha_1 = 0.03$, $\alpha_2 = 0.01$ and $\alpha_3 = -0.1$ (see Table 5.6). If the simulated models are MGCM Σ_1, MGCM Σ_2 and MGCM Σ_3, with both MGCM and MCAR models the estimated α_1 ranges from 0.0173 to 0.0239, α_2 from 0.004 to 0.01, and α_3 from -0.0974 to -0.0897. If the simulated models are MCAR Σ_1, MCAR Σ_2 and MCAR Σ_3 with both fitted models, the estimated α_1 ranges from 0.0138 to 0.0266, α_2 from -0.0021 to 0.0055, and α_3 from -0.0993 to -0.0875.

In the simulations, we use $sd_{v_1} = 0.7$, $sd_{v_2} = 0.65$ and $sd_{v_3} = 0.6$. Both the estimates obtained with fitted models MGCM and MCAR underestimate these true values. For each of the simulated models, we observe that the estimates obtained are closer to the real ones with fitted model MGCM than with fitted model MCAR. For the simulated MGCM, we obtain estimate of sd_{v_1} between 0.6477 and 0.6768, estimate of sd_{v_2} from 0.6099 to 0.6224, and values of sd_{v_3} from 0.5601 to 0.5753. For the simulated MCAR, the estimated sd_{v_1} ranges from 0.6476 to 0.6689, sd_{v_2} from 0.6089 to 0.624, and sd_{v_3} from 0.5556 to 0.5865.

Figure 5.5 depicts maps of the average across data sets posterior mean relative risk obtained with MGCM and MCAR models when fitting data sets generated with MCAR Σ_1. They show the same pattern of risk with both fitted models.

5.3 Estimation of risk of low birth weight in Georgia

In this section we estimate the risk of low birth weight (LBW), defined as babies weighing less than 2500 grams at birth, during the year 2000 in the 159 counties of Georgia, United States. We are interested in seeing in a real example the differences in the estimates and the goodness of fit obtained using GCM and CAR components

Models		MSE											
Simulated	Fitted	θ_1	θ_2	θ_3	α_1	α_2	α_3	v_1	v_2	v_3	u_1	u_2	u_3
MGCM Σ_1	MGCM	0.325	0.069	0.041	0.004	0.002	0.003	0.202	0.068	0.059	0.016	0.014	0.019
	MCAR	0.329	0.073	0.043	0.004	0.003	0.003	0.206	0.072	0.064	0.024	0.019	0.025
MGCM Σ_2	MGCM	0.331	0.067	0.041	0.004	0.003	0.003	0.205	0.07	0.065	0.025	0.017	0.024
	MCAR	0.332	0.069	0.043	0.004	0.003	0.003	0.208	0.074	0.069	0.032	0.021	0.03
MGCM Σ_3	MGCM	0.324	0.067	0.04	0.004	0.003	0.003	0.2	0.071	0.056	0.019	0.019	0.014
	MCAR	0.332	0.07	0.042	0.005	0.003	0.003	0.205	0.074	0.061	0.027	0.023	0.02
MCAR Σ_1	MGCM	0.326	0.07	0.042	0.004	0.002	0.002	0.199	0.068	0.062	0.014	0.014	0.022
	MCAR	0.335	0.071	0.043	0.005	0.002	0.002	0.203	0.072	0.067	0.022	0.019	0.027
MCAR Σ_2	MGCM	0.351	0.067	0.045	0.004	0.002	0.002	0.204	0.07	0.07	0.029	0.017	0.032
	MCAR	0.346	0.07	0.046	0.004	0.003	0.002	0.206	0.073	0.073	0.036	0.021	0.035
MCAR Σ_3	MGCM	0.321	0.067	0.042	0.005	0.002	0.003	0.201	0.068	0.056	0.016	0.015	0.014
	MCAR	0.333	0.07	0.043	0.005	0.002	0.003	0.206	0.072	0.062	0.024	0.019	0.02

Table 5.5: MSE for the parameters of the simulated models MGCM Σ_1, MGCM Σ_2, MGCM Σ_3, MCAR Σ_1, MCAR Σ_2 and MCAR Σ_3, when they are fitted with models MGCM and MCAR. For $k = 1, 2, 3$, θ_k refers to the vector $(\theta_{1k}, \ldots, \theta_{nk})'$, v_k to the vector $(v_{1k}, \ldots, v_{nk})'$, and u_k to the vector $(u_{1k}, \ldots, u_{nk})'$.

Models		Fitted values					
Simulated	Fitted	α_1	α_2	α_3	sd_{v1}	sd_{v2}	sd_{v3}
MGCM Σ_1	MGCM	0.0239	0.0065	-0.0875	0.6599	0.6224	0.5731
	MCAR	0.0208	0.0091	-0.0955	0.6477	0.6134	0.5607
MGCM Σ_2	MGCM	0.0171	0.01	-0.0897	0.6768	0.6205	0.5753
	MCAR	0.0173	0.0086	-0.0974	0.6637	0.6136	0.562
MGCM Σ_3	MGCM	0.0175	0.0041	-0.0897	0.6726	0.6191	0.5703
	MCAR	0.0183	0.006	-0.0956	0.6594	0.6099	0.5601
MCAR Σ_1	MGCM	0.0204	0.0125	-0.0963	0.6689	0.6209	0.5795
	MCAR	0.0207	0.0117	-0.0993	0.6556	0.6137	0.5609
MCAR Σ_2	MGCM	0.0138	-0.0021	-0.0942	0.6718	0.624	0.5865
	MCAR	0.0161	0.0013	-0.0945	0.66	0.6187	0.5664
MCAR Σ_3	MGCM	0.0266	0.0055	-0.0875	0.6595	0.6148	0.5706
	MCAR	0.0222	0.0045	-0.0941	0.6476	0.6089	0.5556

Table 5.6: Fitted values for the parameters of the simulated models MGCM Σ_1, MGCM Σ_2, MGCM Σ_3, MCAR Σ_1, MCAR Σ_2 and MCAR Σ_3, when they are fitted with models MGCM and MCAR.

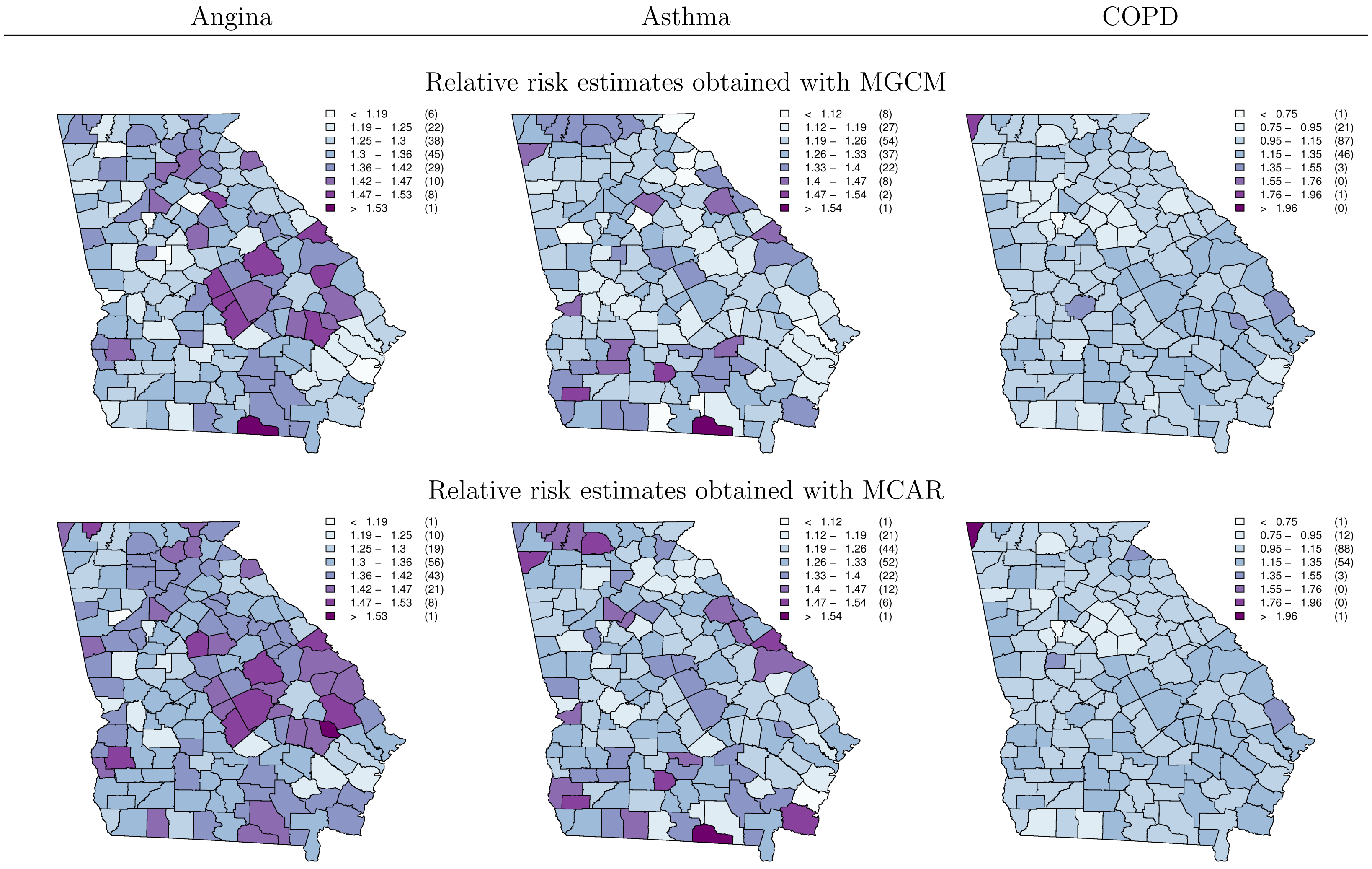


Figure 5.5: Averaged posterior mean of the relative risk for MGCM and MCAR models when fitting the data sets simulated with model MCAR Σ_1.

for modeling the CH. To this end, we fit the models GCM with two covariates and CAR with two covariates which model the CH with a GCM and a CAR component respectively. They model the number of LBW with a Poisson distribution with mean the relative risk times the expected number of LBW. The logarithm of the relative risk in each county i, θ_i, is expressed as a sum of an intercept (α), the effect of two covariates cov_1 and cov_2, and random effects for the CH (u_i) and for the UH (v_i). Specifically, $log(\theta_i) = \alpha + \beta_1 cov_1 + \beta_2 cov_2 + u_i + v_i$. Normal distributions with mean 0 and standard deviations simulated from an $U(0, 5)$ are used as prior distributions of the intercept, the covariates effects, and the UH component. The prior distributions for the standard deviation of the GCM and CAR components are chosen from an $U(0, 5)$.

The number of LBW and the number of total births in Georgia for the year 2000 were obtained from the Online Analytical Statistical Information System (OASIS; http://oasis.state.ga.us/). In the models we adjusted for the effect of the standardized median household income (cov_1) and the percentage of people living in poverty (cov_2) in the year 2000. They are both indicators of poor health outcomes and often used together in deprivation indices. Information about these two covariates were obtained from the US census (http://www.census.gov). The expected number of LBW in each county were computed multiplying the total births in each county by the overall rate of LBW in Georgia, that is, the total LBW divided by the total births.

A summary of the observed and expected LBW and the covariates included in the model is shown in Table 5.7. The average count of LBW events in Georgia is 71.8 with a standard deviation equal to 163.04. The minimum and maximum LBW is 2 and 1326 respectively. We observe values of the standardized median household income between -1.34 and 4.04. The mean percentage of poverty is 16.11 with values situated between 3.8 and 28.6. Model fitting was carried out using `WinBUGS`. Convergence was obtained running two MCMC chains of 70000 iterations with a burn-in period of 20000 and a thin parameter of 50.

5.3.1 Results

Estimates and 95% credible intervals of the overall LBW relative risk ($\widehat{\alpha}$), and the coefficients of the standardized median household income ($\widehat{\beta}_1$) and the percentage of poverty ($\widehat{\beta}_2$) were obtained with fitted models GCM and CAR with two covariates. Results are shown in Table 5.8. With both models we observe a negative, although not significant association between the mean income and LBW. With GCM we obtain $\widehat{\beta}_1 = -0.013$, and with CAR $\widehat{\beta}_1 = -0.019$. We also observe a positive association between poverty and LBW. With GCM $\widehat{\beta}_2 = 0.027$, and with CAR $\widehat{\beta}_2 = 0.025$.

A summary of the standardized incidence ratio and the relative risks obtained

	Standardized income	Poverty	Y	E	SIR	RR	
						GCM	CAR
Mean	0	16.11	71.8	71.8	1.07	1.08	1.08
S.D.	1	5.37	163.04	160.44	0.32	0.19	0.19
Minimum	-1.34	3.8	2	1.81	0.26	0.72	0.7
1st quartile	-0.69	12	14	13.46	0.85	0.93	0.94
Median	-0.26	16.4	27	25.03	1.05	1.08	1.09
3rd quartile	0.42	20.05	61.5	58.64	1.26	1.21	1.21
Maximum	4.04	28.6	1326	1167	2.06	1.61	1.65

Table 5.7: Descriptive statistics of covariates standardized median household income and percentage of poverty, observed (Y) and expected (E) LBW, standardized incidence ratio (SIR), and relative risks (RR) estimated with fitted models GCM with two covariates and CAR with two covariates.

	$\hat{\alpha}$ (95% C.I.)	$\hat{\beta}_1$ (95% C.I.)	$\hat{\beta}_2$ (95% C.I.)
GCM	-0.372 (-0.647, -0.073)	-0.013 (-0.091, 0.061)	0.027 (0.007, 0.043)
CAR	-0.349 (-0.527, -0.067)	-0.019 (-0.083, 0.037)	0.025 (0.007, 0.037)

Table 5.8: Estimates and 95% credible intervals (C.I.) of the overall LBW relative risk ($\widehat{\alpha}$), and the coefficients of the standardized median household income ($\widehat{\beta_1}$) and the percentage of poverty ($\widehat{\beta_2}$) obtained with fitted models GCM and CAR with two covariates.

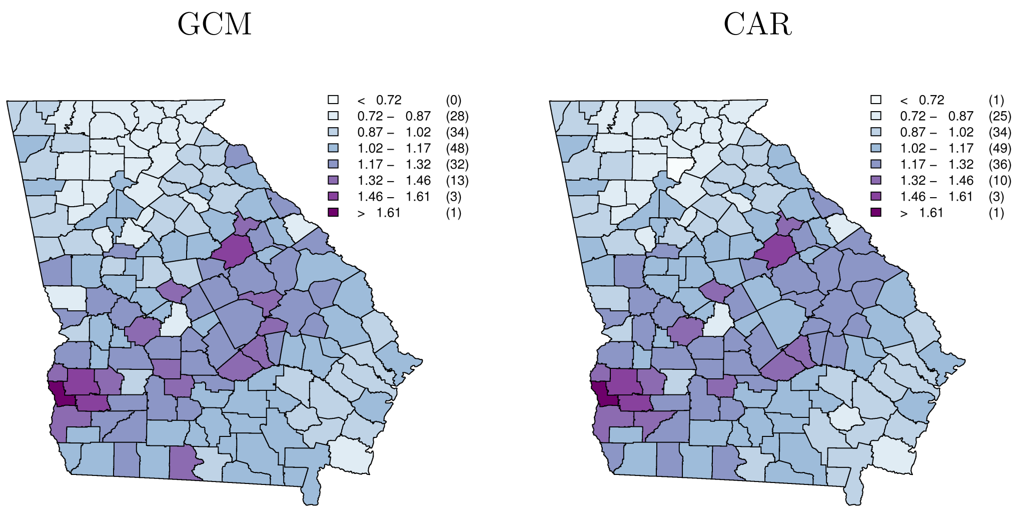


Figure 5.6: Posterior mean relative risk of LBW obtained with fitted models GCM and CAR with two covariates.

with the models are represented in Table 5.7. Figure 5.6 displays the posterior expectations of the relative risks for both fitted models. We can see that the estimates and the spatial pattern of the relative risks obtained with model GCM are very similar to the ones obtained using the CAR model. With both models we obtain a mean relative risk of 1.08 with standard deviation equal to 0.19. Approximately the values range from 0.7 and 1.6. In Figure 5.6, we can see that the North-West and South-East regions of Georgia have lower relative risk than the rest of the state.

The goodness of fit of the models is assessed by means of the deviance information criterion (DIC) (Spiegelhalter et al., 2002). In this example, we note that the deviance, the effective number of parameters, and the DIC are very similar with both models. Specifically, we obtain a slightly lower deviance and DIC for the CAR model (991.57 and 1042.91), compared to the GCM model (994.61 and 1045.22). The effective number of parameters is a little lower with GCM model (50.61), compared to CAR model (51.33). Overall, both GCM and CAR with two covariates provide a similar model for the relative risk.

Chapter 6

Detection of disease clusters with LISA functions

In this chapter we present a method for the detection of spatial clusters of a particular disease with point data where temporal information is not available. More specifically, we focus on case-control studies where the data consist of n locations of all known cases of a disease within a given geographical region over a specified time-period, together with the location m of a set of controls, defined to be a random sample of the population at risk. There are certain advantages and disadvantages to studying a disease using point data (Lawson, 2006). One negative is that there may be no relation between the location assigned to the individual and the etiology of the disease. Most of the time the location assigned to cases is the place of residence and it may be that the disease is related to risk factors which arise in other places where the individual spends part of his time. Moreover, the exact location of the individual is not always available for various reasons, such as confidentiality. On the other hand, these type of data provide detailed spatial information which can be lost when aggregated. Thus the data aggregated are an approximation of the point data. Therefore, whenever point data are available, it is recommended that spatial information does not get lost and that it is analysed at this level of resolution. The method we present is based on the second order properties of a spatial point process. In particular, the local properties (LISA functions) of the product density function are used. These functions have already been used for studying the spatial structure of a point process (Cressie and Collins, 2001), and for detecting features in images with noise (Mateu et al., 2007).

The structure of the chapter is as follows. First of all the concepts of product density and LISA functions are presented. Then the LISA method for detecting spatial clusters is explained. In Section 6.3 an evaluation of the LISA method is made through a simulation study. The power, sensitivity, specificity and type I error of the LISA method are calculated when it is applied to distinct situations where the cluster shape, the sample size, the cluster size and the density of cases are different. The same values are calculated with the scan method and the results obtained with each method are compared. Finally, the LISA and scan methods are applied to a real case where the existence and location of kidney disease clusters during 2008 are studied in the municipality of Valencia, Spain.

6.1 Product density function and LISA functions

We consider a spatial point process N which is observed in a region $A \subset \mathbb{R}^2$ with area $|A| = a$. For stationary and isotropic point processes with intensity λ, the K function Ripley (1976) is defined as

$$K(t) = \lambda^{-1} E[N(b(s,t) \backslash \{s\}) : s \in N], \ t > 0,$$

where $b(s,t)$ is the disc with center s and radius t. $K(t)$ provides an interpretable measure of the spatial dependency structure in the point process. In particular, $\lambda^2 a K(t)$ is the expected number of ordered pairs of points in region A with pairwise distance less than or equal to t. An estimate of the K function from the data $\{x_1, x_2, \ldots, x_{N(A)}\}$ observed in A is given by

$$\widehat{K}(t) = \frac{1}{\lambda^2 a} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(d_{ij} \leq t),$$

where $I(\cdot)$ is the indicator function, $n = N(A)$ and $d_{ij} = ||x_i - x_j||$ is the Euclidean distance between the points x_i and x_j. Here w_{ij} is used for edge-correction and denotes the reciprocal of proportion of the circle centered on x_i and with radius d_{ij} which is contained in A. To complete the estimate we need to replace the unknown intensity λ with an estimate, say $\hat{\lambda} = (n-1)/a$. The final estimate of $K(t)$ is therefore

$$\widehat{K}(t) = (n-1)^{-2} a \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(d_{ij} \leq t).$$

The K function is a cumulative function and the investigation into its differential leads to the product density function defined as

$$\rho(t) \equiv \frac{\lambda^2 K'(t)}{2\pi t}.$$

When estimated from the data the empirical product density function, $\hat{\rho}(t)$, provides a description of the density of interevent distances among the observed locations. High values for $\hat{\rho}(t)$ for small distances t indicate the aggregation of points. On the other hand, small values for small distances t indicate a situation of inhibition. An estimator of $\lambda^2 K'(t)$ can be expressed as

$$\widehat{\lambda^2 K'(t)} \equiv a^{-1} \sum_{i=1}^{n} \sum_{j \neq i} f_\epsilon(||x_i - x_j|| - t), \quad t > \epsilon > 0, \tag{6.1}$$

where f_ϵ is a kernel function and ϵ is the bandwidth. f_ϵ provides weights which determine the contribution of each interevent distance to the density estimator in the distance t. Increasing the value of ϵ increases the bias but smoothes the estimate. The estimation $\widehat{\rho_\epsilon}$ of the product density function has the expression

$$\widehat{\rho_\epsilon}(t) = \frac{1}{2\pi t a} \sum_{i=1}^{n} \sum_{j \neq i} f_\epsilon(||x_i - x_j|| - t), \quad t > \epsilon > 0.$$

A corrected version of this estimation which takes into account the edge effect is given by

$$\widehat{\rho_\epsilon}(t) = \frac{1}{2\pi t a} \sum_{i=1}^{n} \sum_{j \neq i} \frac{2\pi ||x_i - x_j||}{|\partial b(x_i, ||x_i - x_j||) \cap A|} f_\epsilon(||x_i - x_j|| - t), \quad t > \epsilon > 0,$$

where $|\partial b(x_i, ||x_i - x_j||) \cap A|$ denotes the perimeter length of the disc with its center x_i and radius $||x_i - x_j||$ contained in region A.

Both, the K function and the product density function provide a global measure of the covariance structure by adding up the contributions of each point observed in the process. We now consider the individual contributions to the estimated functions which are analogous to the local statistics described by Anselin (1995), called local indicators of spatial association (LISA). A product density LISA function, $\rho^{(i)}$, indicates the contribution of case x_i to the global estimation ρ and can provide a description of the data structure Cressie and Collins (2001).

A LISA product density function can be constructed in the same way as the global estimator. We begin by considering local characteristics of the K function. We define

$$\{\lambda K(t)\}^{(i)} \equiv E[N(b(x_i, t) \backslash \{x_i\}) : x_i \in N], \; t > 0$$

as the expected number of points which are found at a distance less than or equal to t from x_i. A kernel estimator of $\{\lambda K'(t)\}^{(i)}$ is

$$\{\widehat{\lambda K'(t)}\}^{(i)} = \sum_{j \neq i} f_\epsilon(||x_i - x_j|| - t), \quad t > \epsilon > 0,$$

and for a homogenous Poisson process $\hat{\lambda} = (n-1)/a$ is an unbiased estimator of λ. Thus,

$$\hat{\lambda}\{\widehat{\lambda K'(t)}\}^{(i)} = (n-1)a^{-1} \sum_{j \neq i} f_\epsilon(||x_i - x_j|| - t), \quad t > \epsilon > 0,$$

provides a kernel estimator for $\{\lambda^2 K'(t)\}^{(i)}$.

By analogy with the formulation of the estimation of the global product density, a local version of the product density function is given by

$$\widehat{\rho_\epsilon}^{(i)}(t) = \frac{n-1}{2\pi t a} \sum_{j \neq i} f_\epsilon(||x_i - x_j|| - t), \quad t > \epsilon > 0.$$

The corrected version of $\widehat{\rho_\epsilon}^{(i)}$ is

$$\widehat{\rho_\epsilon}^{(i)}(t) = \frac{n-1}{2\pi t a} \sum_{j \neq i} \frac{2\pi ||x_i - x_j||}{|\partial b(x_i, ||x_i - x_j||) \cap A|} f_\epsilon(||x_i - x_j|| - t), \quad t > \epsilon > 0.$$

For a fixed t, $\widehat{\rho_\epsilon}(i)(t)$ satisfies the definition of LISA statistic presented by Anselin (1995), given that the sum of individual product density LISA functions is proportional to the global function in t,

$$\widehat{\rho_\epsilon}(t) = \frac{1}{n-1} \sum_{i=1}^{n} \widehat{\rho_\epsilon}^{(i)}(t).$$

6.2 LISA method for detecting clusters

We propose a method for cluster detection in situations where a set of cases and a set of controls for a certain disease is available. The method calculates the LISA functions for each case and compares them with the LISA functions that would be obtained under the null hypothesis that the cases are a random sample of the total set of cases and controls. Through a Monte Carlo procedure it can be checked, for each case, whether the difference between the LISA obtained with the data and the LISA under the null hypothesis is significant or not. The method identifies the cases which have significant associated differences as cases belonging to an aggregation zone. The stages in the method are explained as follows:

1. The first step consists of calculating, for each case, values for the LISA function by involving only the rest of the cases. The intensity of the processes of cases and controls can be very different in different subregions of region under study depending on several factors, for example population at risk. Therefore, for each case we choose a different vector of distances for which the LISA function will be calculated. For a given case this vector is made up of 100 values. The smallest distance, d_1, is chosen as that which exists between the case and the nearest point, and the greatest distance, d_{100}, is chosen in such a way that the circle with its center in the case and radius d_{100} contains $x\%$ of the points. The remaining distances are chosen between those taking increments equal

to $(d_{100} - d_1)/99$. Here x is chosen subjectively, taking into account that we want to calculate the LISA functions for distances at which local interactions between points are expected to operate, and not where environmental gradients are expected to vary. For the calculation of the LISA functions we need to choose a kernel function and the values of the bandwidths. It is widely known in nonparametric and point process literature that the key point in kernel-based estimation is the choice of the bandwidth parameter, and not the type of kernel function used. Thus, we make use here of the Epanechnikov kernel, as one of the most widely used functions in point process literature. It is given by

$$f_\epsilon(r) = \begin{cases} \frac{3}{4\epsilon}\left(1 - \frac{r^2}{\epsilon^2}\right), & \text{if} - \epsilon \leq r \leq \epsilon; \\ \quad 0, & \text{otherwise} \end{cases}$$

where $\epsilon > 0$ is the bandwidth. We decide to choose an adaptative bandwidth that takes into account the distribution of all the points included in the different subregions where LISA functions are computed. Thus, for each case i we consider A_i, the convex hull of the points included in the circle with its center in the case and radius the greatest distance where LISA is computed, that is to say the value of d_{100} associated with case i. Then for each case i we choose a bandwidth ϵ_i guided by the results in Collins (1995) and Fiksel (1988), namely $\epsilon_i = (5^{1/2}/10)|A_i|^{1/2}n_i^{-1/2}$, where $|A_i|$ is the area of A_i and n_i the number of points included in it.

2. Next, approximations to the LISA function values for each case under the null hypothesis are calculated. For a given case, a set of points is considered to be formed by all the controls and all the cases, except for the one considered. Then a pattern is generated, randomly labeling each point as a case or a control such that the number of cases and controls continues to be the same. Afterwards, the case considered is added to the pattern obtained and the LISA function is calculated for the new cases, with the same set of distances and the same bandwidth used in the calculation of the function with the original data. This procedure is repeated R times and the averages of the LISA values obtained at each iteration are calculated, which are taken as the LISA function of the case under the null hypothesis.

3. In the following stage the difference, dif_0, between the LISA calculated with the original data and the LISA calculated under the null hypothesis is obtained for each case. Given that the objective is to detect aggregation zones, to calculate this difference only the values of the functions in the distances where the LISA function obtained with the original data is higher than the LISA obtained under the null hypothesis are taken into account. Thus, the difference associated to a case is calculated as the average of the positive values of the

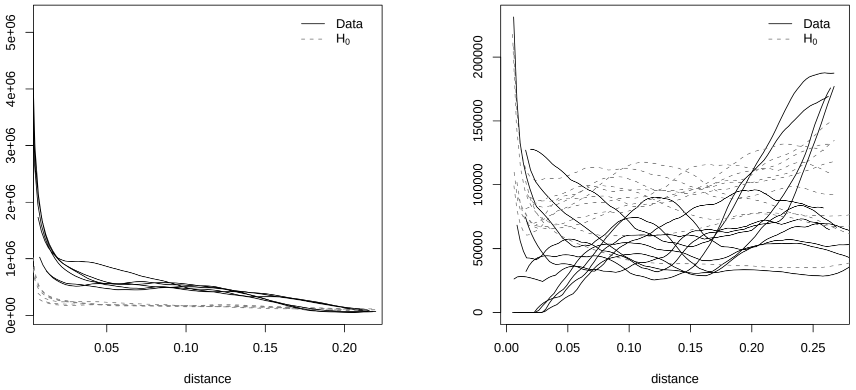


Figure 6.1: LISA functions obtained with original data and under the null hypothesis for 10 cases located in a cluster (left) and for another 10 not located in a cluster (right).

differences between the LISA function with the original data and the LISA function under the null hypothesis in each distance. Figure 6.1 shows both LISA functions for 10 cases located in a cluster (left) and for another 10 not located in a cluster (right). The graphs in Figure 6.1 come from one of the scenarios of the simulation study described in Section 6.3.

4. Finally, Monte Carlo is applied to see the significance of the difference in each case. For each one of the cases a vector of differences under the null hypothesis $\{dif_1, dif_2, \ldots, dif_R\}$ is constructed. For $j \in \{1, \ldots, R\}$, dif_j is obtained with the difference, in the same way as previously explained, between L^j and LM^j, where L^j is the LISA function of the case calculated under the null hypothesis in the replica j and $LM^j(t) = \frac{1}{R}\sum_{k=0,k\neq j}^{R} L^k(t)$, with L^0 being the LISA function of the case calculated at Step 1 with the original data. Then pos_0, the position occupied by the original difference of the case calculated at Step 3, dif_0, in the set $\{dif_0, dif_1, dif_2, \ldots, dif_R\}$ ordered from larger to smaller, is calculated. The significance of dif_0 is calculated by applying the Bonferroni correction for multiple comparisons. In this way, if a value of significance equal to α is taken, and $pos_0/(R+1) < \alpha/(\#\text{ cases})$, dif_0 is significant and the case is marked as belonging to a zone where aggregation exists Diggle (2003).

The results of the method are the differences associated to each case, which indicate the distance between the LISA obtained with the original data and the LISA under the null hypothesis, and the significance of each difference. To show these results, the set of cases is considered as a marked point pattern where the marks are the values of the differences. In this way it is possible to represent the contour lines of the values of the differences in a graph and moreover mark the cases whose differences are significant.

6.3 Evaluation of LISA method through a simulation study

To evaluate the method we apply it in several situations which allow us to assess its performance both in the presence and absence of spatial clusters. In each situation we calculate the type I error, power, sensitivity and specificity. We also calculate these values with the scan method for the Bernoulli model, and compare the results obtained with each one. The scan method is well documented in Kulldorff and Nagarwalla (1995). It enables the location of clusters and the evaluation of their significance taking into account the correction for multiple tests.

To evaluate the LISA method when clusters exist, we define the window under study as the unit square and the cluster zone as a circle with its center in $(0.3; 0.6)$ and with fixed radius equal to 0.1. A set of cases and controls are generated in the window under study such that the density of cases is greater within the cluster than outside and bearing in mind the different simulation scenarios. The density of cases outside the cluster is set at 10%. The different scenarios are created by varying the sample size, cluster size and the density of cases within the cluster, taking into account the following:

- Three sample sizes have been used, 400, 1000 and 2000, in which cases and controls are included. For cluster size, 3%, 6% and 10% of the sample size is used. Both the points within the cluster and those outside it are randomly generated within the corresponding boundaries.

- The density of cases within the cluster is 2, 3 and 6 times the density of cases outside the cluster. Thus, 10% of the points outside the cluster and 20%, 30% or 60% of the points within the cluster are randomly labeled as cases. The remaining points are the controls.

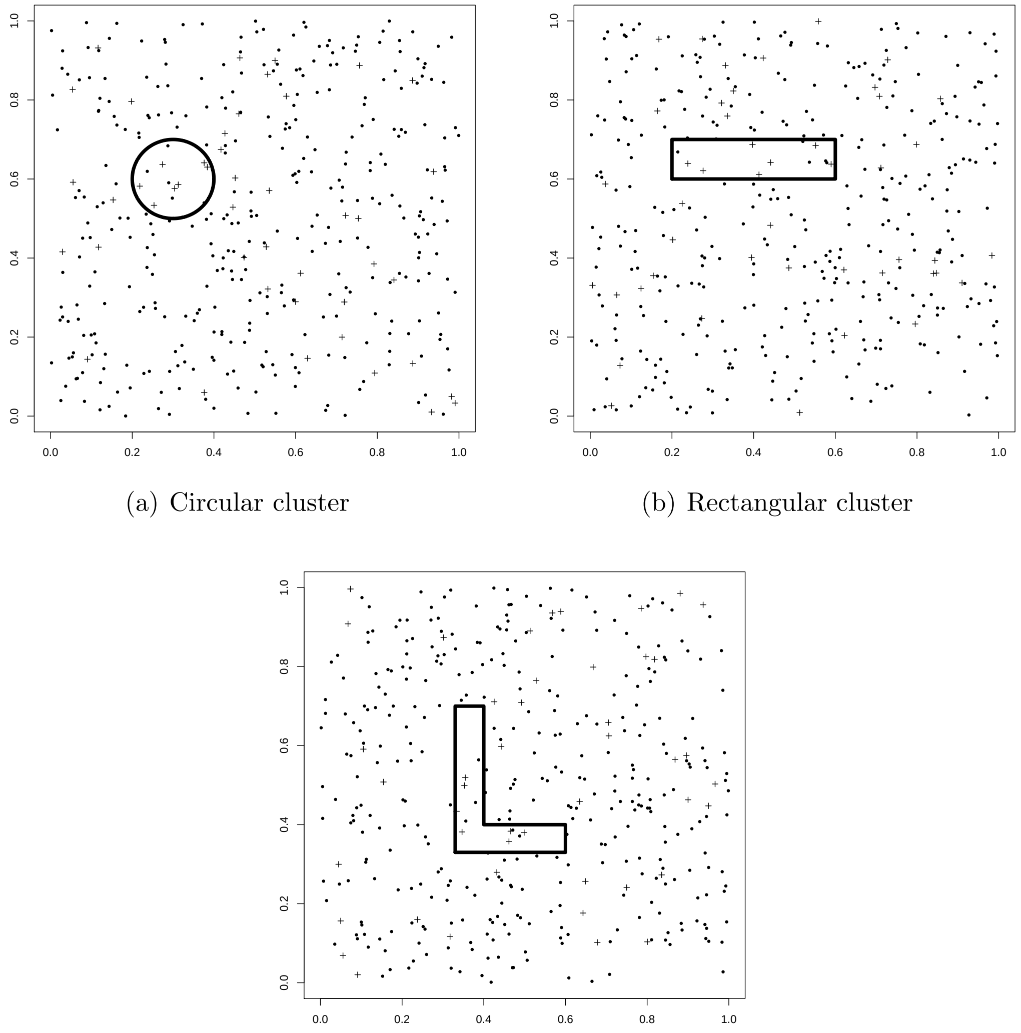
For each of the resulting 27 scenarios, 100 patterns of points are generated and the methods are applied with a value of $\alpha = 0.05$. For the application of the LISA method, we subjectively choose the value of x in step 1 of Section 6.2 to be equal to 20. For each scenario the power, sensitivity and specificity of the methods are calculated. Power is defined as the proportion of the 100 patterns in which the null hypothesis is rejected, this rejection occurring when the p-value associated with the LISA function of one case is smaller than α. Sensitivity is the proportion of cases belonging to the cluster which have been correctly detected by the method. Specificity is the proportion of cases not belonging to the cluster which have been correctly detected by the method. These two values are obtained as the average values obtained in the 100 patterns.

This procedure is repeated for another two forms of the cluster, one rectangle with a clear predominance of one dimension over the other and two orthogonal rectangles which give rise to an L-shaped structure. The vertices in the rectangle are $(0.2; 0.6)$, $(0.6; 0.6)$, $(0.6; 0.7)$ and $(0.2; 0.7)$, and the L-shape is obtained by joining $(0.33; 0.33)$, $(0.6; 0.33)$, $(0.6; 0.4)$, $(0.4; 0.4)$, $(0.4; 0.7)$ and $(0.33; 0.7)$. Figure 6.2 shows 3 of the 27 possible simulation scenarios in which the 3 forms that the cluster can take are represented.

We are also interested in comparing the behavior of the methods in terms of type I error. To this end we simulate point patterns under the null hypothesis of no clusters, and apply the methods to obtain estimates of the true significance level used. In particular we consider the unit square as the study window, and the same circle, rectangle and L-shape structures defined before. In each of these settings we simulate 9 point patterns under the null hypothesis, using samples sizes equal to 400, 1000 and 2000, and number of points inside the structure equal to 3%, 6% and 10% of the sample size. To be consistent with the null hypothesis we label the points as cases and controls in such a way that the densities of cases inside and outside the structure are equal. Specifically we randomly label 10% of the points as cases both inside and outside. Next we generate 100 point patterns for each situation and apply the methods with $\alpha = 0.05$. The type I error probability is defined as the proportion of 100 patterns in which the null hypothesis is rejected.

6.3.1 Results

Tables 6.1-6.3 show the power, sensitivity and specificity of the LISA and scan methods for each one of the cluster shapes, and for the distinct sample sizes, cluster sizes and multiplicities. They also show type I error probabilities in the situations where no clusters exist. It can be seen that the two methods are more powerful in the circular zone of the cluster, followed by the rectangular form and the L-shape. Moreover, for all the shapes, the power of the LISA method is larger than the scan method. As far as sensitivity is concerned, we see that the LISA method is larger than the scan method in most situations. The specificity of the two methods is similar, except in situations in which the multiplicity is 6 and the cluster size is 6 or 10, where LISA has less specificity than scan. With the LISA method therefore, the null hypothesis is rejected more often although it does not always detect the cluster zone exactly. Moreover, if the multiplicity or the cluster size increases, the sensitivity of each method also increases. We can see that in all situations type I error probabilities obtained with scan method are close to the true level of significance 0.05, whereas the ones obtained with LISA are greater than this value if sample size is 1000 or 2000.



(a) Circular cluster

(b) Rectangular cluster

(c) L-shaped cluster

Figure 6.2: Three possible simulation scenarios in which the 3 shapes that the cluster can take are represented. In each scenario cases (+) and controls (·) have been generated using a sample size of 400, a cluster size equal to 3% of the sample size and with a density of cases within the cluster 6 times the density of the cases outside.

		Type I error		Power						Sensitivity						Specificity					
		LISA	Scan	LISA			Scan			LISA			Scan			LISA			Scan		
Size		Multiplicity		Multiplicity			Multiplicity			Multiplicity			Multiplicity			Multiplicity			Multiplicity		
Sample	Cluster	1	1	2	3	6	2	3	6	2	3	6	2	3	6	2	3	6	2	3	6
400	3%	0.02	0.05	0.20	0.26	0.91	0.02	0.08	0.62	0	0.02	0.49	0.02	0.07	0.53	0.99	0.99	0.97	0.99	0.99	0.96
	6%	0.06	0.08	0.20	0.52	1	0	0.14	1	0.02	0.10	0.89	0	0.12	0.9	0.99	0.99	0.94	1	0.96	0.95
	10%	0.02	0.05	0.25	0.93	1	0.05	0.55	1	0.04	0.28	0.98	0.04	0.44	0.94	0.99	0.98	0.91	0.98	0.90	0.97
1000	3%	0.13	0.02	0.43	0.69	1	0.07	0.15	1	0.01	0.15	0.93	0.05	0.12	0.88	0.99	0.99	0.96	0.98	0.98	0.98
	6%	0.15	0.06	0.57	1	1	0.11	0.82	1	0.06	0.53	1	0.06	0.68	0.94	0.99	0.98	0.89	0.98	0.95	0.99
	10%	0.13	0.05	0.89	1	1	0.15	1	1	0.18	0.80	1	0.11	0.91	0.97	0.98	0.96	0.84	0.97	0.95	0.99
2000	3%	0.29	0.05	0.75	1	1	0.08	0.61	1	0.06	0.47	0.99	0.03	0.48	0.92	0.99	0.98	0.93	0.99	0.99	0.99
	6%	0.25	0.05	0.97	1	1	0.34	1	1	0.21	0.86	1	0.22	0.89	0.96	0.98	0.95	0.85	0.97	0.98	0.99
	10%	0.18	0.02	1	1	1	0.60	1	1	0.47	0.96	1	0.45	0.94	0.98	0.98	0.92	0.81	0.96	0.98	0.99

Table 6.1: Type I error, power, sensitivity and specificity of the LISA and scan methods when the structure is circular in shape. Values for sample sizes equal to 400, 1000 and 2000, structure sizes equal to 3%, 6% and 10% of the sample size, and multiple of the density of cases outside the structure equal to 1, 2, 3 and 6.

		Type I error		Power						Sensitivity						Specificity					
		LISA	Scan	LISA			Scan			LISA			Scan			LISA			Scan		
Size		Multiplicity		Multiplicity			Multiplicity			Multiplicity			Multiplicity			Multiplicity			Multiplicity		
Sample	Cluster	1	1	2	3	6	2	3	6	2	3	6	2	3	6	2	3	6	2	3	6
400	3%	0.08	0.05	0.17	0.23	0.64	0.05	0.04	0.36	0.01	0.04	0.19	0	0.01	0.27	0.99	0.99	0.98	0.99	0.99	0.95
	6%	0.06	0.07	0.23	0.44	1	0.06	0.12	0.98	0	0.07	0.69	0.04	0.07	0.7	0.99	0.99	0.95	0.98	0.98	0.90
	10%	0.05	0.04	0.25	0.72	1	0.07	0.35	1	0.02	0.19	0.89	0.04	0.26	0.84	0.99	0.98	0.92	0.97	0.93	0.88
1000	3%	0.17	0.01	0.37	0.63	1	0.01	0.14	0.98	0.02	0.14	0.71	0	0.06	0.64	0.99	0.98	0.97	1	0.98	0.94
	6%	0.11	0.07	0.57	0.92	1	0.10	0.44	1	0.06	0.30	0.95	0.04	0.27	0.85	0.99	0.98	0.91	0.99	0.96	0.92
	10%	0.17	0.03	0.72	1	1	0.13	0.92	1	0.10	0.56	0.99	0.08	0.71	0.91	0.99	0.96	0.86	0.98	0.89	0.91
2000	3%	0.29	0.03	0.63	0.97	1	0.08	0.40	1	0.05	0.27	0.92	0.03	0.23	0.77	0.99	0.99	0.94	1	0.98	0.95
	6%	0.19	0.04	0.91	1	1	0.19	0.93	1	0.16	0.60	0.99	0.10	0.64	0.86	0.99	0.97	0.87	0.99	0.93	0.93
	10%	0.29	0.02	1	1	1	0.43	1	1	0.27	0.83	1	0.24	0.80	0.94	0.98	0.93	0.81	0.96	0.91	0.91

Table 6.2: Type I error, power, sensitivity and specificity of the LISA and scan methods when the structure has a rectangular form. Values for sample sizes equal to 400, 1000 and 2000, structure sizes equal to 3%, 6% and 10% of the sample size, and multiple of the density of cases outside the structure equal to 1, 2, 3 and 6.

		Type I error		Power						Sensitivity						Specificity					
		LISA	Scan	LISA			Scan			LISA			Scan			LISA			Scan		
Size		Multiplicity		Multiplicity			Multiplicity			Multiplicity			Multiplicity			Multiplicity			Multiplicity		
Sample	Cluster	1	1	2	3	6	2	3	6	2	3	6	2	3	6	2	3	6	2	3	6
400	3%	0.02	0.04	0.14	0.25	0.57	0.07	0.07	0.29	0.01	0.02	0.19	0.02	0.02	0.21	0.99	0.99	0.98	0.98	0.98	0.94
	6%	0.06	0.09	0.26	0.41	0.99	0.03	0.07	0.88	0.03	0.08	0.52	0.01	0.05	0.65	0.99	0.99	0.95	0.99	0.98	0.86
	10%	0.03	0.02	0.23	0.70	1	0.05	0.32	1	0.02	0.14	0.79	0.02	0.25	0.85	0.99	0.98	0.92	0.99	0.93	0.84
1000	3%	0.13	0.04	0.35	0.54	1	0.06	0.07	0.90	0.01	0.08	0.54	0	0.03	0.57	0.99	0.99	0.97	0.99	0.99	0.92
	6%	0.12	0.04	0.52	0.95	1	0.09	0.38	1	0.05	0.25	0.90	0.05	0.27	0.84	0.99	0.98	0.92	0.98	0.93	0.90
	10%	0.08	0.05	0.70	1	1	0.18	0.83	1	0.09	0.43	0.96	0.09	0.58	0.91	0.99	0.97	0.87	0.97	0.90	0.89
2000	3%	0.26	0.05	0.69	0.94	1	0.11	0.21	1	0.04	0.18	0.84	0.03	0.11	0.72	0.99	0.99	0.94	0.99	0.97	0.93
	6%	0.34	0.03	0.94	1	1	0.15	0.81	1	0.12	0.50	0.98	0.10	0.50	0.87	0.99	0.96	0.88	0.98	0.93	0.92
	10%	0.32	0.02	1	1	1	0.33	1	1	0.21	0.72	0.99	0.21	0.79	0.92	0.98	0.94	0.84	0.95	0.90	0.90

Table 6.3: Type I error, power, sensitivity and specificity of the LISA and scan methods when the structure has an L-shape. Values for sample sizes equal to 400, 1000 and 2000, structure sizes equal to 3%, 6% and 10% of the sample size, and multiple of the density of cases outside the structure equal to 1, 2, 3 or 6.

We now examine the behavior of the methods in each one of the structure shapes. In all of them, and when the sample size is 400, the methods are not very sensitive in situations with a multiplicity of 3 and a cluster size of 3% or 6%, or for all 3 cluster sizes when the multiplicity is 2. In these situations the methods are not capable of detecting clusters, which is why we avoid making comparisons.

Circular shape.- When the sample size is 400 the minimum and maximum sensitivities are 0.28 and 0.98 for the LISA method, and 0.44 and 0.94 for the scan method. The specificity values for both methods range from 0.9 to 0.99 and are very similar, except when the cluster size is 10% and the multiplicity is 3. In this situation the specificities of LISA and scan are 0.98 and 0.90 respectively. With both methods type I error probabilities range from 0.2 to 0.8. The probabilities obtained with LISA method are lower than those obtained with scan method. If the sample size is 1000, LISA is more sensitive than scan except when the multiplicity is 2 and the cluster size is 3%, and when the multiplicity is 3 and the cluster size is 6% or 10%. The specificity of LISA is better when the multiplicities are 2 or 3. If the sample size is 2000, the LISA method is more sensitive if multiplicity is 6. In the rest of the situations the sensitivity of both methods is similar. On the other hand, scan has more specificity for multiplicities 3 and 6. When sample sizes are 1000 and 2000, type I error probabilities are greater with LISA method than with scan method. Namely, probabilities obtained with LISA range from 0.13 to 0.15 when sample size is 1000, and from 0.18 to 0.29 when sample size is 2000. With scan method these probabilities are all in the range 0.02 to 0.06.

Rectangular shape.- If the sample size is 400, the LISA sensitivity values oscillate between 0.19 and 0.89 and those of scan between 0.26 and 0.84. Specificity for the LISA and scan methods ranges from 0.92 to 0.98 and from 0.88 to 0.95, respectively. The type I error probabilities with each method are not very different, the minimum probability is 0.04 and the maximum 0.08. If the sample size is 1000, the LISA sensitivities are larger except when the cluster size is 10% and multiplicity is 3. The specificity of the LISA method is also larger, except for multiplicity 2 and cluster size 3%, and multiplicity 6 and cluster size equal to 6% or 10%. Type I error probabilities with LISA range from 0.11 to 0.17, and with scan from 0.01 to 0.07. The LISA method has larger sensitivity values when the sample size is 2000, except when multiplicity is 3 and cluster size is 6%. Values of specificity are very similar, except in situations where multiplicity is 6 and cluster size is 6% or 10%, where the differences can become 0.10, to the detriment of the LISA method. The LISA method yields type I error probabilities much higher than scan method. Namely in the range 0.19 to 0.29 for LISA, and 0.02 to 0.04 for scan method.

L shape.- When the sample size is 400, the LISA method has more specificity although scan works better in sensitivity. Type I error probabilities are similar with both methods and range from 0.02 to 0.09. For a sample size equal to 1000, the LISA method is more sensitive than the scan method in all cases, except when multiplicity is 3 and cluster size is 6% or 10%, and when multiplicity is 6 and cluster size is 3%. The specificity values are also better, except when multiplicity is 6 and cluster size is 10%, where the difference is 0.02. With LISA method type I error probabilities are higher than those obtained with scan. When the sample size is 2000, the LISA method is more sensitive than the scan method, except when multiplicity is 3 and cluster size 10%. Specificity is also better in all cases except when multiplicity is 6 and cluster size is 6% or 10%. In those cases, the maximum difference of specificity is ≤ 0.06. The type I error probabilities obtained with LISA are very high, they range from 0.26 to 0.34. Those obtained with scan range from 0.02 to 0.05.

We note that in the majority of situations where multiplicity is 6 and the cluster size is 6% or 10%, the specificity of LISA is lower than that of scan. The explanation for this comes from the fact that many of the cases situated outside the cluster, but sufficiently near to its border, have significant associated differences. There are many small distances between those cases and the others due to the large number of cases which are inside the cluster. Thus their LISA functions are very different from their LISA function approximations obtained under the null hypothesis. The method, therefore, does not inform us of the exact borders of the cluster, but of their approximate location.

Regarding type I error probabilities, we observe that LISA method does not perform well across different situations. When sample size is 400 type I error probabilities are close to 0.05. However, when sample size is 1000 and 2000, LISA method overstates the true significance whereas scan method yields estimations very close to the real value. Moreover type I error probabilities obtained with LISA increase as the number of points increases.

6.4 Detection of clusters of kidney disease in Valencia

Next we use the LISA method and the scan method to evaluate the existence and location of kidney disease clusters in the city of Valencia, Spain, during 2008. A case-control study is proposed where the data consist of the locations of individuals with kidney disease (cases) and a set of locations obtained from the population at risk (controls), which allows the heterogeneity of the population to be taken into account.

# cases	25-29	30-34	35-39	40-44	45-49	50-54	55-59
Men	0	1	4	3	0	3	5
Women	1	0	1	1	0	3	1

# cases	60-64	65-69	70-74	75-79	80-84	85-89	90-94
Men	8	3	15	10	7	1	0
Women	1	2	4	4	6	3	1

Table 6.4: Number of kidney disease cases in each of the strata sex-age group.

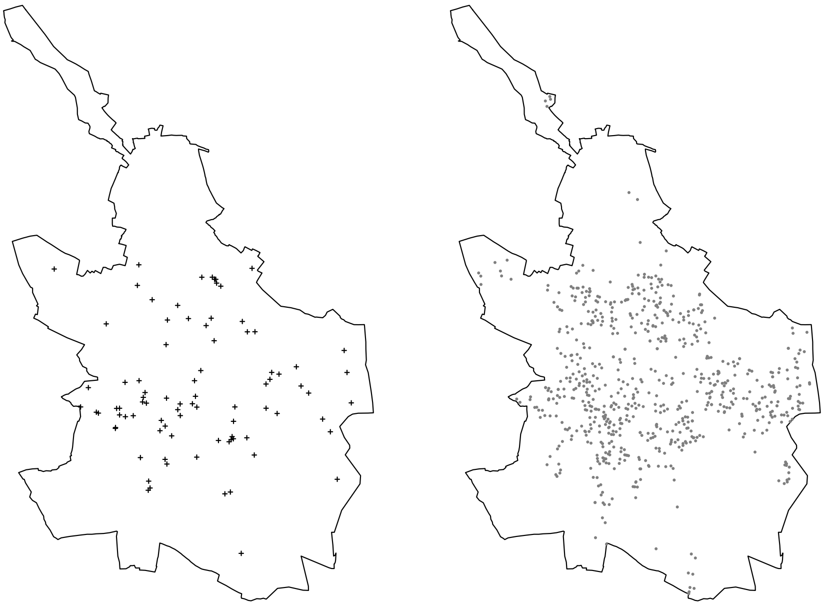
6.4.1 Data

The geographical area of study is the city of Valencia. The set of cases consists of the 88 people diagnosed with kidney disease in 2008 and was collected by the Conselleria de Sanitat of Valencia. For each case we know their address, sex and age. There were 60 men with ages ranging from 31 to 86, and 28 women with ages ranging from 28 to 90. The sex and age distribution of the cases are shown in Table 6.4. The control set is formed by 704 locations obtained as a random sample drawn from the population at risk. For each case, 8 controls are sampled choosing 8 residential addresses of people living in Valencia in 2008, excluding the cases, and with the same sex and age as the case. Controls were provided by Valencia City Council. For the application of the statistical methods, the geocoding of the addresses to UTM coordinates was carried out. Figure 6.3 shows two maps of Valencia with the locations of the cases and the controls.

6.4.2 Results

We apply the LISA and scan methods to the set of cases and controls, obtaining the following results. With the LISA method and using a value of $\alpha = 0.05$ and $x = 20$, four cases with significant associated differences are obtained. These cases are located in the west and north of Valencia. The results of the method are presented in the first two maps of Figure 6.4. Figure 6.4(a) shows the locations of cases with significant associated differences, that is, the cases which the method identifies as belonging to an aggregation zone. Figure 6.4(b) shows the contour lines of the values of the differences associated to each case. We observe that the detected cases are located in two of the zones where the differences are larger.

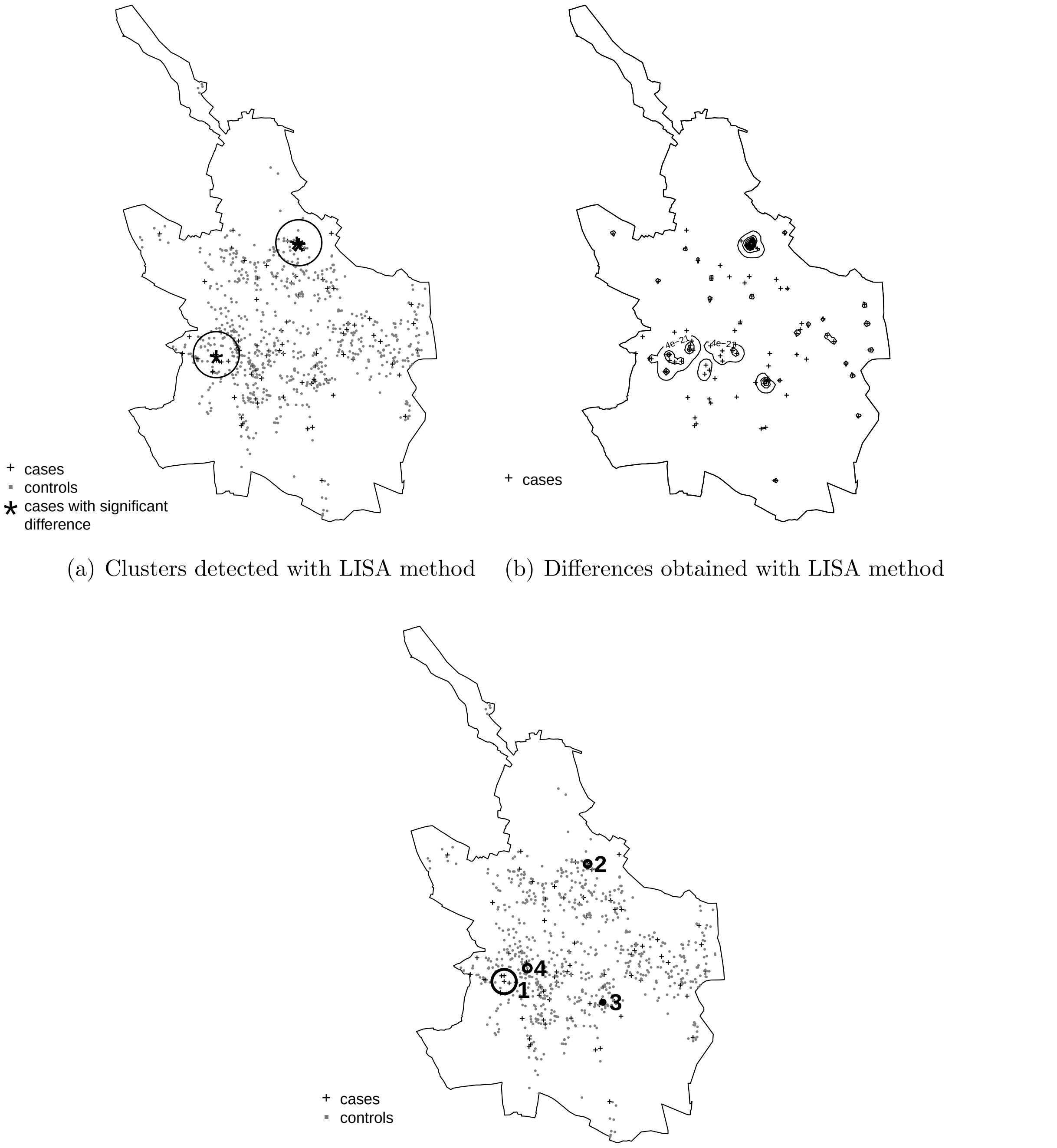
The clusters detected with the scan method with the Bernoulli model and 999 simulations are shown in Figure 6.4(c). The most likely cluster, labeled as 1 in the figure, has a p-value equal to 0.026 and a population of 10 individuals, 7 of them cases. The ratio between observed and expected cases (SIR) is 6.397. The method also highlights three secondary clusters which do not overlap with the most likely



(a) Cases

(b) Controls

Figure 6.3: Locations of the cases and the controls of kidney disease in Valencia during 2008.



(a) Clusters detected with LISA method (b) Differences obtained with LISA method

(c) Clusters detected with scan method

Figure 6.4: Results obtained with LISA and scan methods. First map highlights the cases detected as belonging to clusters with the LISA method, and the second map shows the contour lines for the differences associated to each case. In the third map the zones detected by the scan method as being possible clusters are shown. Only zone 1 is significant.

cluster and which are not statistical significant. These clusters are labeled as 2, 3 and 4 and have p-values equal to 0.244, 0.572 and 0.960 respectively.

On comparing the results obtained it can be seen that the aggregation zones identified with each method are located in the same areas. The two zones where LISA detects cases with significant differences are located in the same areas where the most likely cluster and first secondary cluster detected with the scan method are located. Furthermore, the significant and non significant clusters detected with scan are situated in the zones where the differences obtained with LISA are larger.

Chapter 7

Spatial variations in temporal trends (SVTT)

The monitoring of disease occurrence provides fundamental baseline information for the development of research programs into the aetiology of the disease, as well as for the planning and evaluation of public health interventions. It is well known that disease incidence and mortality change over time, and using for example disease registry data, such temporal trends are closely monitored as they provide important clues concerning the success or failure of disease prevention and control strategies. Now, it is possible that these temporal trends are different in different geographical regions, and knowledge about such differences may provide clues as to where disease prevention and control measures are successfully implemented, or where new unknown health hazards have emerged. To facilitate investigations of emerging spatial patterns and temporal trends of disease risks, several spatio-temporal models for disease mapping have been proposed, either based on a parametric description of time trends, on independent risk estimates for every period, on autoregressive approaches, or on the definition of the joint covariance matrix for all the periods as a Kronecker product of matrices (Martínez-Beneito et al., 2008; Waller et al., 1997; Xia and Carlin, 1998; Bernardinelli et al., 1995). The scan statistics for spatial variations in temporal trends (SVTT) are designed for the detection of clusters of areas with unusual different temporal trends. The linear SVTT method (Kulldorff, 2010) is based on scan statistics and uses a Poisson regression with time as independent variable to estimate the disease trend. This type of estimation makes impossible to detect points in time where the tendency changes and gives low power in some situations. In this chapter we review the linear SVTT method and propose a new one, the quadratic SVTT method, that allows a more flexible trend estimation and increases the power of detection in some situations where the linear method fails to detect existing clusters.

The structure of the chapter is as follows. First, we review the basics of the linear SVTT method. In Section 7.2 the quadratic SVTT method is introduced and the advantages that it presents over the linear method are discussed. Then, a simulation study is carried out where the quadratic and linear methods are applied to different data sets simulated from a variety of models for the true underlying disease trends. Next, the quadratic method is applied to find areas with unusual cervical cancer trends in white females in the United States over the period 1969 to 1995. Finally, in Section 7.5, a evaluation study is conducted where the power, sensitivity, positive predictive value (PPV) and type I error probabilities of linear and quadratic methods are calculated when they are applied to several simulated scenarios.

7.1 Spatial Variations in Temporal Trends

The spatial, temporal and spatio-temporal scan statistics are used to detect spatial, temporal and spatio-temporal disease clusters respectively (Kulldorff et al., 1998). Specifically, they are designed to highlight areas where the number of cases are significantly greater than expected. Unlike these scan statistics, the statistics for spatial variations in temporal trends (SVTT) are used for the detection and inference of any zone with exceptionally different temporal trend. The SVTT method considers a fixed temporal period of interest where the disease trend is assessed. The method gradually scans a spatial window centered in each location and with different sizes, and the disease trend in each of the windows is estimated. The window with the trend that most differs from the rest of trends is considered as having unusual different trend and its statistical significance is assessed.

The SVTT linear method is a special case that can be used for the detection and inference of any zone with exceptionally increasing or decreasing linear trend. Here, the trend estimation is done using a Poisson regression with time as independent variable, the time changing population size as offset, and the number of events as dependent variable. These trends are then used to adjust the expected number of cases for each location and time, where the adjustment will be different inside and outside the window due to the different estimated trends. With the new expected counts, the likelihood for this window is calculated, and the maximum likelihood over all windows is found. This maximum is then compared with the maximum likelihoods from a large set of random data sets. Since the interest is only in the difference in the trends between areas, and not the overall trend, the analysis is conditioned on the latter. This is done in the randomization step, by not randomizing the observed times. Instead, for each time, a spatial location according the background population size at that time is randomized.

7.2 Quadratic SVTT method

The trend estimation procedure in the linear SVTT method can lead to wrong conclusions in some situations. For example, consider that a particular disease has constant trend across the study region except in one area where it has a parabolic trend. Here, the linear SVTT would estimate the parabolic trend as a constant trend, and the area with different trend would not be identified. To overcome the bad performance of the linear SVTT method in situations like this, we propose the quadratic SVTT method. Basically, the quadratic method is a modification of the linear method where the trend estimation procedure is changed. Specifically, a new explanatory variable, time squared, is added to the regression model used in the linear method. This modification provides better estimates of the real trends, and increases the power of detection in situations where the linear method fails to detect existing clusters. The steps of the quadratic method are the same that those performed in the linear method, with the only difference in the way the trends are estimated. First, a huge number of windows are constructed. Then, for each window, the trends inside and outside are estimated and the likelihood is computed. After that, the most likely cluster, defined as the window with maximum likelihood, is picked up and their p-value is obtained using Monte Carlo hypothesis testing. The rest of this section gives specific details about the estimation of the trends, the hypothesis test considered, the test statistic and the procedure to assess its significance.

A log-linear model is used to estimate the trend over time of the disease risk, defined as the number of observed cases over the number of expected cases. In the model, the observation time and the observation time squared are used as explanatory variables. These variables are denoted as X_1 and X_2, respectively. Thus, for $t = 1, \ldots, T$, the model is specified as

$$log(\mu_t/m_t) = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t},$$

or equivalently as

$$\mu_t = m_t exp(\beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t}),$$

where μ_t and m_t denote the mean number of observed cases and the number of expected cases, respectively, at time t. Let y_t denote the number of cases at time t and let $\boldsymbol{x_t'} = (1, X_{1t}, X_{2t})$ and $\boldsymbol{\beta'} = (\beta_0, \beta_1, \beta_2)$. The likelihood expression of the Poisson model is

$$L(\boldsymbol{\beta}|y_1, \ldots, y_T) = \prod_{t=1}^{T} \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!} = \prod_{t=1}^{T} \frac{exp\big(-m_t exp(\boldsymbol{x_t'\beta})\big)\big\{m_t exp(\boldsymbol{x_t'\beta})\big\}^{y_t}}{y_t!},$$

The vector of parameters $\boldsymbol{\beta}$ is estimated by maximizing the log likelihood

$$l = log(L) = \sum_{t=1}^{T} -m_t exp(\boldsymbol{x_t'\beta}) + \sum_{t=1}^{T} y_t log(m_t) + \sum_{t=1}^{T} y_t \boldsymbol{x_t'\beta} - \sum_{t=1}^{T} log(y_t!).$$

via the Newton-Raphson algorithm.

For each of the windows constructed, we estimate the global trend and the trends inside and outside the window, and perform a hypothesis test to assess whether the inside trend is different from the outside trend. Specifically, we test the following hypothesis:

$$H_0\text{: } \beta_1^{in} = \beta_1^{out} \text{ and } \beta_2^{in} = \beta_2^{out} \text{ vs. } H_a\text{: } \beta_1^{in} \neq \beta_1^{out} \text{ or } \beta_2^{in} \neq \beta_2^{out},$$

where $\{\beta_j^{in}, j = 0, 1, 2\}$ and $\{\beta_j^{out}, j = 0, 1, 2\}$ are the regression coefficients of the trend inside and outside z, respectively. For testing the hypothesis, the log likelihood ratio (LLR) of each window is computed. Let $\{\beta_j^{g}, j = 0, 1, 2\}$ be the regression coefficients of the global trend. For time $t = 1, \ldots, T$, let y_t^{in} and y_t^{out} be the observed cases inside and outside z, and m_t^{in} and m_t^{out} the expected cases inside and outside z. The LLR of window z takes the form

$$LLR(z) = l_a(z) - l_0(z) = (l1_a(z) + l2_a(z)) - (l1_0(z) + l2_0(z)),$$

where

$$\begin{aligned}
l1_a(z) &= \sum_{t=1}^{T} -m_t^{in} exp(\boldsymbol{x_t'\beta^{in}}) + \sum_{t=1}^{T} y_t^{in} log(m_t^{in}) \\
&\quad + \sum_{t=1}^{T} y_t^{in} \boldsymbol{x_t'\beta^{in}} - \sum_{t=1}^{T} log(y_t^{in}!), \\
l2_a(z) &= \sum_{t=1}^{T} -m_t^{out} exp(\boldsymbol{x_t'\beta^{out}}) + \sum_{t=1}^{T} y_t^{out} log(m_t^{out}) \\
&\quad + \sum_{t=1}^{T} y_t^{out} \boldsymbol{x_t'\beta^{out}} - \sum_{t=1}^{T} log(y_t^{out}!), \\
l1_0(z) &= \sum_{t=1}^{T} -m_t^{in} exp(\boldsymbol{x_t'\beta^{g\ in}}) + \sum_{t=1}^{T} y_t^{in} log(m_t^{in}) \\
&\quad + \sum_{t=1}^{T} y_t^{in} \boldsymbol{x_t'\beta^{g\ in}} - \sum_{t=1}^{T} log(y_t^{in}!), \\
l2_0(z) &= \sum_{t=1}^{T} -m_t^{out} exp(\boldsymbol{x_t'\beta^{g\ out}}) + \sum_{t=1}^{T} y_t^{out} log(m_t^{out}) \\
&\quad + \sum_{t=1}^{T} y_t^{out} \boldsymbol{x_t'\beta^{g\ out}} - \sum_{t=1}^{T} log(y_t^{out}!),
\end{aligned}$$

and

$$\boldsymbol{\beta^{in}} = (\beta_0^{in}, \beta_1^{in}, \beta_2^{in})', \ \boldsymbol{\beta^{out}} = (\beta_0^{out}, \beta_1^{out}, \beta_2^{out})',$$

$$\boldsymbol{\beta}^{\boldsymbol{g\ in}} = (\beta_0^{g\ in}, \beta_1^g, \beta_2^g)',\ \boldsymbol{\beta}^{\boldsymbol{g\ out}} = (\beta_0^{g\ out}, \beta_1^g, \beta_2^g)',$$

$$\beta_0^{g\ in} = log\left(\sum_{t=1}^{T} y_t^{in}\right) - log\left(\sum_{t=1}^{T} m_t^{in} exp(\beta_1^g X_{1t} + \beta_1^g X_{2t})\right),$$

$$\beta_0^{g\ out} = log\left(\sum_{t=1}^{T} y_t^{out}\right) - log\left(\sum_{t=1}^{T} m_t^{out} exp(\beta_1^g X_{1t} + \beta_1^g X_{2t})\right).$$

The test statistic is defined as the maximum LLR over all windows. Its statistical significance is obtained through Monte Carlo hypothesis testing by generating a large number of random data sets generated under the null hypothesis.

7.3 Simple examples using simulated data

We are interested in seeing the performance of the SVTT quadratic method when detecting geographical variations in trends with different shapes. Also, we want to see whether the method erroneously detects areas with different trends when they actually do not exist. To this end, we simulate different datasets with and without areas with different trends, and see the performance of the quadratic method when it is applied to them. The datasets are simulated in such a way to contain areas with different increasing and decreasing trends, and areas with trends of different shapes, as linear and parabolic shapes. Moreover, we also apply the linear method in all of these situations and compare the quadratic and linear methods results.

The geographical region used to carry out the simulation is the state of New Mexico. New Mexico is divided in 32 counties which are represented in Figure 7.1. The period of time considered in the simulation is from 1980 to 2000, a total of 21 years. Given this region and this fixed period of time, we simulate seven situations setting the disease trend of one county being different from the rest. Then, we apply the quadratic and linear methods to each of the simulated data, and for each method we observe which areas, if any, have been detected as having unusual different trend, and examine their inside and outside trends estimates.

The specification of the scenarios is as follows: In all situations we work with a constant population equal to 10,000 in each county and year. We define variable time t as $t = yr - 1980 + 1$, where $yr \in \{1980, \ldots, 2000\}$. Therefore $t \in \{1, \ldots, 21\}$. For each county, we generate the number of observed cases in each time t. We choose county 21 as the county with different trend. In this county the number of observed cases is simulated using a expression different from the rest of counties. In the first six scenarios the observed cases are generated as a function of time. The expressions of the observed counts inside and outside the county with different trend are shown

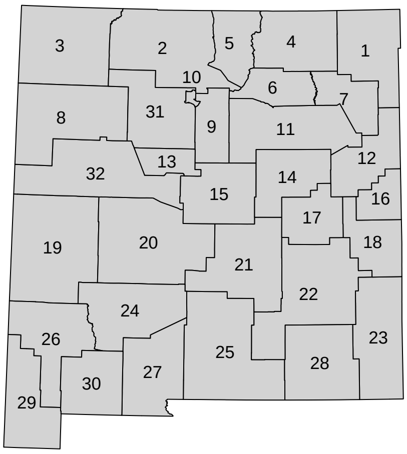


Figure 7.1: Locations of New Mexico state counties.

in Table 7.1. The observed counts in scenario 7 have been set fixed to take a trend with a particular curve shape shown in Figure 7.2. In scenario 1, the trend is the same in all counties. In situations 2, 3 and 7, the county 21 has a increasing trend different from the rest of counties. And in situations 4, 5 and 6, the observed counts are specified as the exponential function of a quadratic function of time.

The results obtained are presented in Figure 7.2. The figure shows for each scenario, the trends generated inside and outside county 21. These trends represent the risk of disease in each time, calculated as the number of observed cases divided by the number of expected cases. Moreover, in situations where the methods identify groups of counties with unusual different trends, plots of the estimated trends inside and outside these groups of counties are depicted. We observe that the quadratic method works well in all situations whereas the linear method fails in situation 5. In this situation all counties have a constant trend except county 21 which have a trend with parabolic shape. This county is detected with the quadratic method but not with the linear method which estimates trend in county 21 constant as in the rest of counties. Scenario 1 has the same trend in all counties, and both methods work good as they do not find any area with unusual different trend.

7.4 Application to cervical cancer mortality in the United States

Cervical cancer is a highly preventable and curable disease if detected early. Important strategies to reduce its risk include screening with the Papanicolaou (Pap) and human papillomavirus (HPV) tests, as well as prevention of HPV infection with

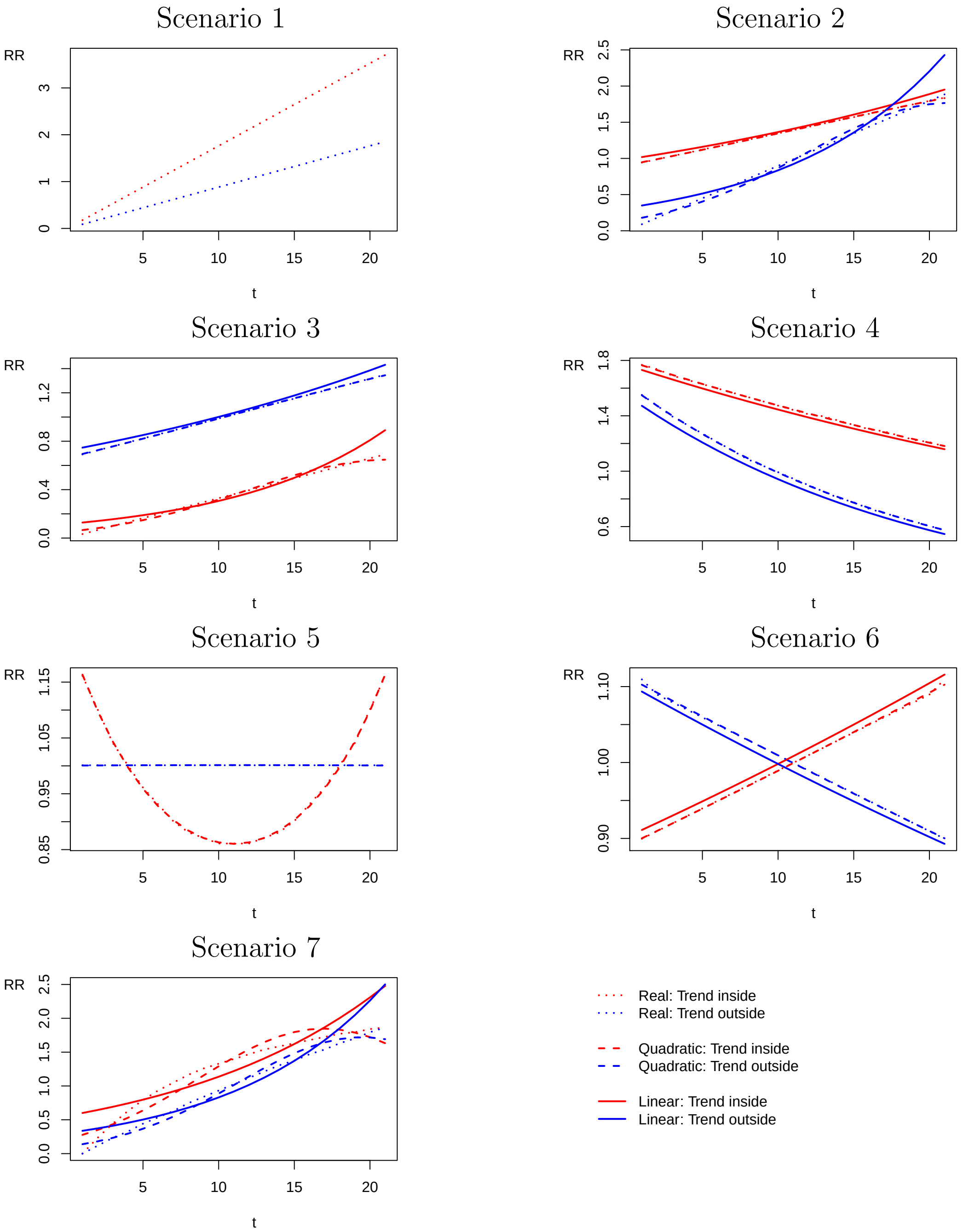


Figure 7.2: Real trends simulated inside and outside the area with unusual different trend. Estimated trends inside and outside the area obtained with the quadratic and linear methods. Estimated trends are not presented in situations where the method does not detect any area with unusual different trend.

Scenario	# cases observed inside the area with different trend	# cases observed outside the area with different trend
1	$2t$	t
2	$20 + t$	$2t$
3	t	$20 + t$
4	$100\ exp(0.15 - 0.02t + 0t^2)$	$100\ exp(0.05 - 0.05t + 0t^2)$
5	$100\ exp(0.213 - 0.066t + 0.003t^2)$	$100\ exp(0 + 0t + 0t^2)$
6	$100\ exp(-0.11 + 0.01t + 0t^2)$	$100\ exp(0.11 - 0.01t + 0t^2)$

Table 7.1: For scenarios 1 to 6, functions of time used to generate the number of cases inside and outside the area with different trend at each time $t = 1, \ldots, 21$.

the HPV vaccine. Although cervical cancer incidence and mortality rates have declined approximately 50 percent in the United States over the past three decades, the disease remains a serious health threat. The National Cancer Institute estimates cervical cancer accounts for 2.5 percent of all cancers that afflict women in the United States, and about 13,500 cases are diagnosed each year.

This section examines the spatial variations in temporal trends occurred in cervical cancer mortality in white women living in the United States during the period 1969 to 1995. For each of the counties of Unites States and D.C., the population and deaths of cervical cancer of white women are available as aggregated counts in periods of three years, from 1969 until 1995, and by 4-years age groups.

We use the quadratic SVTT method to detect groups of counties with unusual different temporal trends. The method identifies 5 of such groups of which only 3 are significant. For each of the detected groups, Table 7.2 shows its population, observed and expected deaths, risk, LLR and p-value. Moreover, parameter estimates of the trend inside and outside the groups are also shown. The groups of counties detected are labeled with the numbers 1 to 5 according to their significance, with the group 1 being the most significant and the group 5 the less significant. Figure 7.3 represents the significant detected groups of counties in the United States map and their temporal trends. We observe that the area with the most significant different trend corresponds to a group of counties located in New York, New Jersey, Pennsylvania and Delaware. In this area the trend decreases until 1991 and then is approximately constant. The second area detected comprises a big part of the east of United States, except the counties located in the east coast and the state of Florida. Here, the trend is more decreasing than the global trend. The third group corresponds to south California. In this area we observe a decreasing trend until 1985 and then a constant trend.

Cluster	Population	Obs.	Exp.	Risk	Trend Inside			Trend Outside			LLR	P-value
					β_0	β_1	β_2	β_0	β_1	β_2		
1	27, 404, 723	9, 459	10, 422.77	0.90	0.366	-0.158	0.010	0.723	-0.217	0.012	55.31	0.01
2	146, 291, 125	53, 784	50, 975.66	1.11	0.772	-0.216	0.011	0.588	-0.200	0.012	40.27	0.01
3	18, 878, 767	6, 191	6, 082.28	1.02	0.582	-0.187	0.012	0.692	-0.212	0.011	18.22	0.02
4	3, 832, 163	1, 033	1, 269.45	0.81	0.633	-0.210	0.006	0.686	-0.210	0.011	9.59	0.87
5	5, 004, 532	966	1, 375.49	0.70	0.612	-0.394	0.032	0.686	-0.209	0.011	9.27	0.92

Table 7.2: Population, observed and expected deaths, risk, LLR and p-value of the detected groups of counties with unusual different cervical cancer trend in white women over the period 1969 to 1995. Parameter estimates of the trends inside and outside of each group of counties.

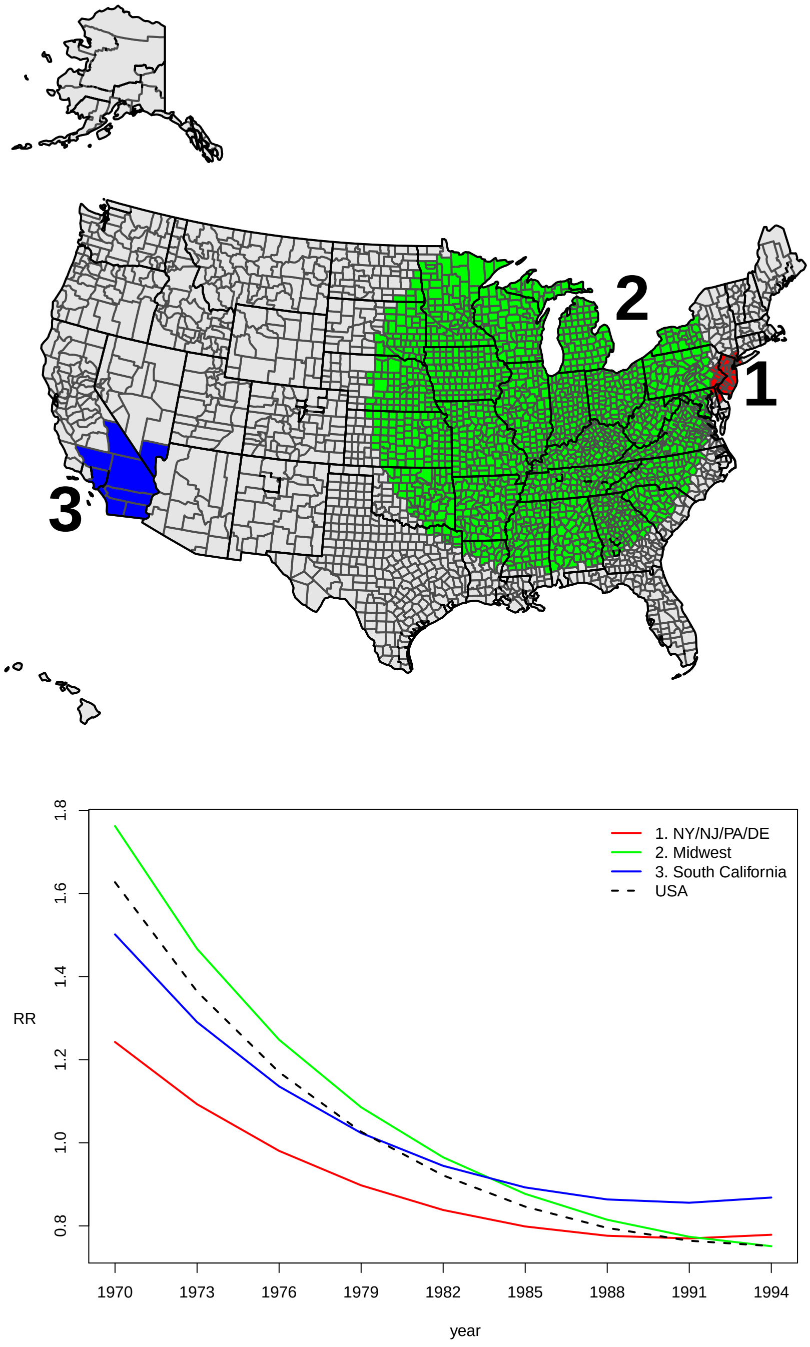


Figure 7.3: Temporal trends in areas with unusual different cervical cancer trends in white women over the period 1969 to 1995.

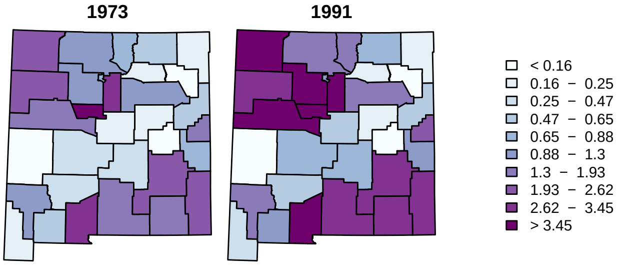


Figure 7.4: Expected brain cancer cases in New Mexico in years 1973 and 1991.

7.5 Power evaluation

In this section we perform a simulation study to evaluate the quadratic SVTT method and to compare it with the linear SVTT method. We consider the 32 counties of New Mexico and simulate several situations varying the period of time of study and the population, observed and expected disease cases in each of the counties and times. The observed counts are generated in such a way that there is a particular zone with a trend different from the trends of the rest of the region of study. Both the quadratic and the linear methods are applied to the simulated data, and the power, sensitivity and PPV of the methods are computed. In order to see the performance of the methods at detecting geographic variations in temporal trends both when these variations occur in single counties, and when they occur in a group of counties, we set up two different situations:

- **One single county with different trend.-** In the first situation there is only one single county, namely county 15, with a different temporal trend in the region of study. In this case we use a temporal window of 21 years, from 1980 to 2000. In each county and year, the underlying population is set constant to $10,000$ implying also an equal number of expected counts. We set the expected counts equal to 100 in each county and year.

- **Group of counties with different trend.-** In the second situation there is a group of contiguous counties with a different temporal trend, namely counties 15, 20 and 21. Here the temporal window is 19 years, from 1973 to 1991. The expected counts of the disease are set equal to the expected counts of brain cancer calculated using the population of the whole state of New Mexico as standard population. Figure 7.4 represents the expected counts in years 1973 and 1991.

We want to assess the performance of the methods when they are applied to detect variations in temporal trends of different shapes. We expect the quadratic method, unlike the linear, to pick up variations when the shape of the unusual trend is parabolic. Furthermore, we want to see whether the methods performance is comparable at detecting variations in linear trends. For each of the two situations described above, we set up 37 scenarios where different temporal trends are simulated. In the scenarios corresponding to the first situation, the temporal trends are constant in every county different from county 15. On the other hand, in scenarios corresponding to the second situation, the temporal trends are constant in every county different from counties 15, 20 and 21. In counties with nonconstant trend, the temporal trends are simulated as follows: In the first 12 scenarios the simulated trends have a parabolic shape in logarithmic scale, 6 with a minimum and 6 with a maximum. In the following 12 scenarios the trends have a linear shape in logarithmic scale, 6 increasing and 6 decreasing. And in the last 12 scenarios the trend has a curve shape, 6 increasing and 6 decreasing. Moreover, in order to compute the type I error probabilities, we set one scenario where all the counties have the same constant trend risk.

We simulate 1,000 data sets of disease observed counts in each county and year. For a given county the observed number of cases in time t, Y_t, is simulated from a Poisson distribution with mean $E_t \times \theta_t$

$$Y_t \sim Po(E_t \times \theta_t),$$

where E_t and θ_t are the expected count and the relative risk in time t respectively. In counties with parabolic and linear shape, θ_t is defined as

$$\theta_t = exp(\beta_0 + \beta_1 t + \beta_2 t^2),$$

where the $\beta's$ are chosen to make the trend to have a parabolic or linear shape. In the scenarios where the trend is linear, β_2 is equal to 0. In counties with curve shape the values θ_t for each t are obtained with a nonparametric procedure. In counties where trend is constant, $\theta_t = 1 \; \forall t$. Figure 7.5 shows the different trend shapes in county 15 in the first situation.

We apply both the quadratic and the linear methods to the simulated data sets in each scenario with a significance level $\alpha = 0.05$. In the situations where there is a county with different risk trend, we calculate the power, the sensitivity and the PPV obtained with each method. In the situations where all the counties have the same trend, we calculate the type I error probabilities. Power is defined as the proportion of the 1,000 data sets in which the null hypothesis is rejected. This rejection occurs when the p-value associated to any county or groups of counties with unusual different temporal trend detected by the method is smaller than α. Sensitivity is the number of counties that have different trend and have been correctly detected divided by the total number of counties that have different trend. The PPV is the number of counties which have different trend and have been correctly detected by the method divided by the total number of counties detected as having different trend. These two values are obtained as the average values obtained in the 1,000 patterns. In the situation in which all counties have the same risk trend we calculate the type I error probability as the proportion of the 1,000 data sets in which the null hypothesis is rejected. The results obtained allow us to evaluate and compare both methods. These are presented in tables 7.3-7.5.

- **One single county with different trend.-** The results obtained in situations with one single county with different trend are the following. We observe that in scenarios where the risk trend has a parabola shape the linear method does not work well and the power, sensitivity and PPV are very low. For this method we obtain values of power ≤ 0.06, sensitivity ≤ 0.03 and PPV ≤ 0.004. However, the quadratic method does better at detecting the area with different trend and the power, sensitivity and PPV values obtained are higher. The lowest power and sensitivity are 0.07 and 0.03, respectively and they are obtained in the scenarios with the parabolas with less curvature. In the scenarios with biggest parabola curvatures we obtain power and sensitivity values equal to 1. The PPV ranges from 0.011 to 0.90. We observe that the power and sensitivity of the quadratic method decrease as the parabola becomes flatter.

  In scenarios with linear trend we see that the power and sensitivity of the quadratic method are a little lower than the values obtained with the linear method. The greatest differences between linear and quadratic methods are 0.12 in power and also 0.12 in sensitivity. In the quadratic method, the power ranges from 0.07 to 1, and the sensitivity from 0.03 to 1. In the linear method, power range 0.08 to 1 and sensitivity 0.04 to 1. For both methods, power and sensitivity decrease as the slope of the trend line decreases. We obtain higher PPV with the quadratic method when β_1 is equal to -0.03, -0.025, 0.025 and 0.03.

  In situations with curve shape trend, we observe that power and sensitivity are 0.999 or 1 in all cases for both linear and quadratic methods. The PPV obtained with the quadratic method are slightly higher than those obtained with the linear. They range 0.83 to 0.89 with the quadratic method, and 0.71 to 0.86 with the linear method.

  We observe that type I error probabilities of both quadratic and linear methods equal to 0.05, the significance level specified.

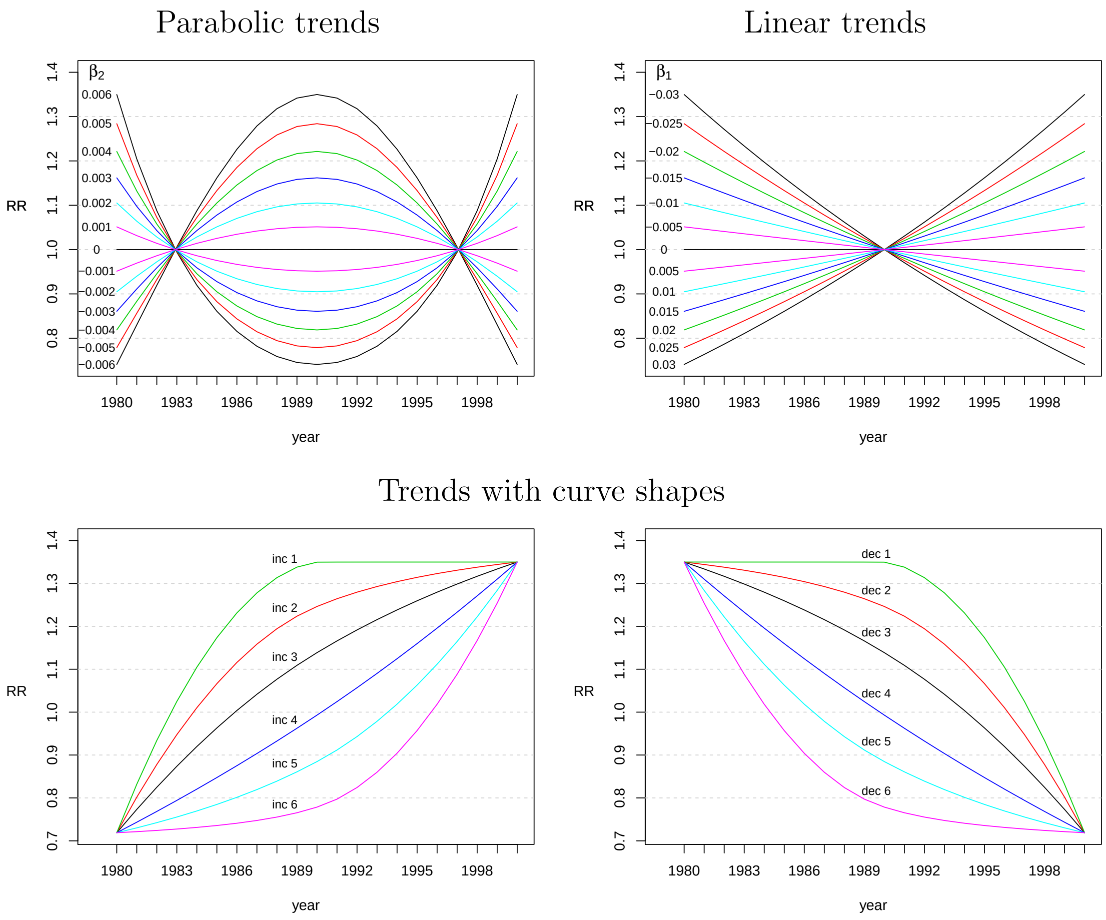
- **Group of counties with different trend.-** In situations where there is a group of counties with different trend we obtain the following results. When the trend has a parabolic shape the quadratic method has higher sensitivity and PPV than the linear method. The power is also higher in all cases except when β_2 is equal to 0.001 or -0.001. In both cases the power difference is 0.01.

We observe that, as in the situation with only one county with different trend, the power, sensitivity and PPV obtained with the quadratic method increases as the curvature of the parabola increases. We observe also that the linear and quadratic methods give similar results for parabolas with low curvature.

When the trend has a linear shape, the power and sensitivity of the linear method are better than those of the quadratic method. For the quadratic method power ranges from 0.09 to 1, and sensitivity 0.02 to 0.98. For the linear method, power ranges from 0.11 to 1 and sensitivity 0.03 to 0.99. PPV is also higher for the linear method in all situations except when β_1 is equal to -0.03, -0.025, 0.025 and 0.03.

In situations with curve shape trend we obtain power and sensitivity values greater than 0.94 in all situations for both methods. The PPV of the quadratic method ranges from 0.86 to 0.94, and the PPV of the linear method from 0.79 to 0.91. PPV values obtained with the quadratic method are slightly higher than those obtained with the linear method.

The type I error probabilities are 0.07 and 0.10 with the quadratic and linear methods, respectively.



Parabolic trends

Linear trends

Trends with curve shapes

Figure 7.5: Simulated trends of the disease risk in situations where only one county has a different temporal trend.

	One county with different temporal trend						Group of counties with different temporal trend					
	Power		Sensitivity		PPV		Power		Sensitivity		PPV	
β_2	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear
0.006	1	0.05	1	0.02	0.85	0.003	1	0.14	0.96	0.04	0.89	0.04
0.005	0.998	0.06	0.998	0.02	0.89	0.002	0.99	0.11	0.86	0.03	0.86	0.03
0.004	0.99	0.06	0.99	0.03	0.90	0.004	0.83	0.12	0.59	0.03	0.69	0.02
0.003	0.73	0.05	0.72	0.01	0.63	0.001	0.48	0.09	0.28	0.03	0.35	0.01
0.002	0.26	0.05	0.23	0.02	0.17	0.003	0.20	0.08	0.09	0.02	0.08	0.01
0.001	0.07	0.06	0.03	0.02	0.01	0.003	0.10	0.11	0.03	0.03	0.02	0.01
0	0.05(*)	0.05(*)	n/a	n/a	n/a	n/a	0.07(*)	0.10(*)	n/a	n/a	n/a	n/a
-0.001	0.08	0.06	0.04	0.02	0.01	0.002	0.09	0.10	0.03	0.03	0.02	0.01
-0.002	0.27	0.06	0.24	0.01	0.17	0.001	0.20	0.1	0.10	0.02	0.11	0.01
-0.003	0.76	0.05	0.74	0.01	0.65	0.002	0.54	0.09	0.32	0.02	0.42	0.01
-0.004	0.98	0.04	0.98	0.01	0.88	0.001	0.89	0.10	0.65	0.02	0.77	0.01
-0.005	1	0.05	1	0.01	0.90	0.001	0.995	0.10	0.89	0.02	0.90	0.01
-0.006	1	0.05	1	0.01	0.84	0.002	1	0.09	0.98	0.02	0.91	0.01

Table 7.3: Scenarios where the trend is of parabolic shape. Power, sensitivity and PPV obtained with the quadratic and linear methods in scenarios where the area with different temporal trend is made up of one single county (left), and where the area is made up of a group of counties (right). (*) refers to type I error probabilities obtained in the scenario with equal risk trend in all counties.

	One county with different temporal trend						Group of counties with different temporal trend					
	Power		Sensitivity		PPV		Power		Sensitivity		PPV	
β_1	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear
0.03	1	1	1	1	0.85	0.81	1	1	0.98	0.99	0.92	0.91
0.025	1	1	1	1	0.90	0.88	0.997	1	0.91	0.95	0.91	0.90
0.02	0.96	0.98	0.96	0.98	0.84	0.87	0.91	0.97	0.70	0.81	0.79	0.86
0.015	0.62	0.74	0.60	0.72	0.49	0.59	0.58	0.74	0.35	0.48	0.46	0.61
0.01	0.22	0.30	0.19	0.26	0.13	0.18	0.28	0.38	0.13	0.21	0.15	0.24
0.005	0.07	0.08	0.03	0.04	0.01	0.01	0.11	0.13	0.04	0.05	0.03	0.04
0	$0.05^{(*)}$	$0.05^{(*)}$	n/a	n/a	n/a	n/a	$0.07^{(*)}$	$0.10^{(*)}$	n/a	n/a	n/a	n/a
-0.005	0.09	0.09	0.05	0.05	0.02	0.02	0.09	0.11	0.02	0.03	0.01	0.02
-0.01	0.21	0.30	0.17	0.27	0.12	0.19	0.17	0.27	0.07	0.12	0.07	0.13
-0.015	0.63	0.73	0.61	0.72	0.53	0.60	0.48	0.64	0.27	0.38	0.35	0.45
-0.02	0.94	0.97	0.94	0.97	0.85	0.85	0.86	0.93	0.60	0.70	0.71	0.74
-0.025	0.999	1	0.999	1	0.91	0.89	0.99	0.999	0.87	0.92	0.85	0.82
-0.03	1	1	1	1	0.85	0.80	0.999	1	0.97	0.99	0.87	0.82

Table 7.4: Scenarios where the trend is linear. Power, sensitivity and PPV obtained with the quadratic and linear methods in scenarios where the area with different temporal trend is made up of one single county (left), and where the area is made up of a group of counties (right). (*) refers to type I error probabilities obtained in the scenario with equal risk trend in all counties.

	One county with different temporal trend						Group of counties with different temporal trend					
	Power		Sensitivity		PPV		Power		Sensitivity		PPV	
	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear
inc 1	1	1	1	1	0.84	0.79	1	1	0.991	0.995	0.92	0.89
inc 2	1	1	1	1	0.86	0.83	1	1	0.99	0.98	0.92	0.88
inc 3	1	0.999	1	0.999	0.88	0.85	1	0.999	0.99	0.94	0.89	0.84
inc 4	1	1	1	1	0.86	0.80	1	1	0.996	0.998	0.93	0.90
inc 5	1	1	1	1	0.89	0.84	1	1	0.996	0.998	0.94	0.90
inc 6	1	1	1	1	0.87	0.86	1	1	0.998	0.998	0.92	0.91
dec 1	1	1	1	1	0.87	0.81	1	1	0.996	0.996	0.89	0.79
dec 2	1	0.999	1	0.999	0.83	0.81	1	1	0.995	0.992	0.88	0.81
dec 3	1	0.999	1	0.999	0.87	0.84	1	1	0.998	0.98	0.89	0.83
dec 4	1	1	1	1	0.87	0.82	1	1	0.995	0.997	0.87	0.79
dec 5	1	1	1	1	0.86	0.82	1	1	0.995	0.996	0.86	0.81
dec 6	1	1	1	1	0.87	0.86	1	1	0.99	0.99	0.87	0.83

Table 7.5: Scenarios where the trend is of curve shape. Power, sensitivity and PPV obtained with the quadratic and linear methods in scenarios where the area with different temporal trend is made up of one single county (left), and where the area is made up of a group of counties (right).

Chapter 8

Conclusions and future research lines

8.1 Conclusions

Proper statistical methodology is needed for public health surveillance. Over the past decade, the use of spatial and spatio-temporal techniques has become increasingly common through advances in computing and the availability of geocoded databases. In this thesis, we addressed various aspects related to the analysis of geocoded health data and developed statistical methods that can be used for helping surveillance practice.

Model-based estimation of missing values in mortality data

Specifically, in Chapter 4, we presented a model to estimate missing values in the all-cause and P&I mortality data from the 122 CMRS operated by the CDC. This work was developed under the supervision of Al Ozonoff at Harvard School of Public Health, and as a result, one paper is in progress. We proposed different specifications which modeled weekly death counts by city, calendar week, calendar year, and age group, and the best model was selected assessing the expectation and variance of the prediction errors of all models using a cross-validation approach. The selected model uses information from the 5 previous and 5 following years, weeks and age groups, and the 3 closest cities. The model performs very poorly in cities with a large number of missing values and we did not include in our computations cities with more than 70% missing values. We computed the mortality burden and excess of deaths using data where missing values were imputed with the model proposed and using data where missing values were assumed to be zero counts. Comparison of the results obtained with both procedures reflects some differences in the mortality burden and trends over time. These results lead us to conclude that some approach has to be taken in order to handle missing data. Otherwise, analyses could produce inaccurate estimates and incorrect conclusions. The model we presented is just one of possible alternatives to impute missing values, but it is a useful approach to estimate missing data and improve inferences in situations with a moderate number of missing values.

Gaussian component mixtures (GCM) and CAR models in Bayesian disease mapping

In Chapter 5, we presented two possible structures to model correlated heterogeneity in hierarchical Bayesian models: GCM and CAR models. We explained their respective properties both in univariate and multivariate situations where the risk of multiple diseases was modeled. We carried out a simulation study in both situations to compare the behavior of these structures when they were used to fit several simulated models that encompass a wide range of situations that can be encountered in real settings. The results of the univariate study suggest that in most situations GCM and CAR have similar properties for modeling. We observed that across many models the patterns of relative risk and UH and CH effects plotted were very similar, and the goodness-of-fit criteria indicated a similar fitting to the data. One of the objectives of the study was to analyze the performance of the different fitted models at estimating covariate effects. We observed that different covariate patterns yield different results, and that across some covariates GCM and CAR produce similar estimates. Similarly, the multivariate study results indicate a comparable performance of MGCM and MCAR, with only slight differences in the relative risk patterns and the goodness-of-fit. This work can also be found in Moraga and Lawson (2012).

CAR structures within hierarchical models are enormously popular for small area estimation, and offer a robust and flexible mechanism for modeling correlated heterogeneity. However, this work show they are certainly not the only models nor necessarily the optimal for this type of data. The simulation studies carried out show that GCM and CAR have a comparable performance in a wide range of situations, and also identify some types of situations where GCM gain advantage over CAR models. These results lead us to conclude that GCM models are a good alternative to CAR models.

Detection of disease clusters with LISA functions

The method for the detection of clusters in case-control studies presented in Chapter 6 can be found in Moraga and Montes (2011). The method is based on local indicators of spatial association (LISA) functions of the product density function of the point pattern of cases, and highlights the cases which form part of an agglomeration zone. It has the advantage that it does not need *a priori* specification of the shape of the cluster and is capable, therefore, of detecting clusters of any shape.

The results of the method are dependent on the sample size, the cluster size and the density of cases inside the cluster. It is important to point out that a considerable number of controls are necessary for the estimates of the LISA functions to be good and for the method to function. A major disadvantage of the LISA method when comparing with the scan method, is that type I error is high, specially when the number of cases is high. We conclude that LISA method is very useful in situations where clusters exist, since it is capable of detecting clusters of any shape and with high sensitivity and specificity.

Spatial variations in temporal trends (SVTT)

In Chapter 7 the quadratic SVTT method for the detection of areas with unusual different disease temporal trends has been presented. This method was developed under the supervision of Martin Kulldorff at Harvard Medical School, and has been implemented in the `SaTScan` software, www.satscan.org. The quadratic SVTT method is based on the spatial scan statistics, and is a modified version of the linear SVTT method where the trend estimation procedure is changed. Through several examples, we observed that the quadratic method is able to detect areas with unusual different trends in situations where the linear method fails, and also gives better estimates of the real trends. A simulation study was carried out to compare the performance of the quadratic and linear methods. The methods were applied to detect areas with different temporal trends on simulated datasets generated in such a way to encompass a wide range of situations that can be encountered in real settings. The results indicate a comparable performance of both methods in situations where the trends are linear. However, when the trends are not linear, the quadratic method is better than the linear method and presents higher sensitivity, specificity and PPV. The quadratic method is very useful for the evaluation of the control and prevention measures in progress. When applied to detect areas with unusual different trends of cervical cancer in white women in the United States over the period from 1969 to 1995, it was observed that mortality is decreasing overall but not in the same way in all areas. Specifically, three areas of interest where the behavior of the risk trend is significantly different from the rest were highlighted.

8.2 Future research lines

Detection of spatio-temporal clusters with LISA functions

While spatial clusters of disease are of great interest, spatio-temporal clusters are also important and can be encountered often. The spatial LISA method described in Chapter 6 can be easily adapted to look for spatio-temporal clusters by considering cases and controls as three-dimensional points with coordinates representing the geographic locations and dates of the individuals, and computing distances in $\mathbb{R}^3$. We will adapt the spatial LISA method to the spatio-temporal setting. To do that, we first must transform the study region and the points in the dataset so that they are in a coordinate system where distances between three-dimensional points make sense. Specifically, we should have a coordinate system where geographic and temporal distances mean the same, in terms of proximity and remoteness of the points. After carrying out the transformation of the study region and the dataset, the steps of the spatio-temporal method will be the same as the steps of the spatial method but considering points and distances in $\mathbb{R}^3$.

In this work we will carry out a simulation study to evaluate the performance of the LISA method when it is applied to detect spatio-temporal clusters. Although many disease clusters may have a regular shape, it is very common to find irregularly shaped clusters arising in real situations, such as disease concentrations along rivers and oceans shores, transport ways or plumes of air pollution. Moreover, some clusters relate to a fixed point in space, whereas others allow the spatial focus to move with time (Duczmal and Assunçao, 2004; Duczmal et al., 2006; Biggeri et al., 1996; Katsouyanni et al., 1991; Xu et al., 1989). In the simulation study it will be important to use cluster zones with similar characteristics as the ones that could be encountered in real situations. Thus, some of the cluster zones constructed will have irregular shapes or grow or move over time.

Model-based detection of disease clusters

The existing techniques for detecting disease clusters have several limitations: they can not deal with some types of data, do not adequately model situations with excess of zeros, or do not address confounding and bias. Therefore, a model-based approach would be of interest in order to explore disease incidence to potential risk factors. We would like to develop a method to detect and evaluate clusters in spatial, temporal, and spatio-temporal settings that extends the existing techniques and overcomes their limitations. To this end, we will use a flexible method that uses Generalized Linear Models (GLMs) and allows the incorporation of information about covariates and random effects. Specifically, the method will consider a large number of potential clusters in the study region and will construct dummy variables associated to each one. The collection of potential clusters that the method will test is not composed of all possible configurations of aggregated zones in the study region. In the spatial setting, for example, clusters will be defined as circles or ellipses of geographic size between zero and an upper limit defined either as a percent of the population at risk or in terms of the geographical spread of the disease. Then, for each one of the potential clusters, a model relating the incidence, mortality or other measure of the disease and the dummy variable will be fitted. Selection of the best models will be done using general techniques of model selection such as AIC.

The models selected will indicate the most likely clusters and their significance. Since models can be specified using different distributions for the response variable and can include fixed and random effects, this method offers several advantages over the existing ones. Specifically, different types of outcomes such as Poisson or Binomial can be considered. Confounding can be addressed including covariates in the model. Moreover, the incorporation of random effects makes possible to deal with overdispersion and measurement bias.

This method will be implemented in a new `R` package called `DClusterm` using techniques that are described in Zhang and Lin (2009b), Zhang and Glaz (2008) (GLMs), Zhang and Lin (2009a) (mixed models), and Gómez-Rubio and López-Quílez (2010) (zero-inflated models). This work was initiated during the Google Summer of Code 2011 program where I participated under the supervision of Virgilio Gómez-Rubio and Barry Rowlingston.

Applied work

We are also working in two applied projects to understand the dynamics of measles in Europe, and the incidence of leptospirosis in a slum of Salvador de Bahia, Brazil. These analyses will help stakeholders to make decisions to allocate resources and target interventions in an effective way. The European Centre of Disease Prevention and Control (ECDC) aims at strengthening Europe's defences against infectious diseases. With the eradication of measles as objective, one of its projects studies the spatio-temporal distribution of measles in Europe. Disease mapping, temporal trends and detection of clusters methods are helping us to assess the distribution of the disease.

The project about leptospirosis is taking place in Pau da Lima, a slum neighborhood that grew at the outskirts of urban Salvador de Bahia. Pau da Lima hosts a densely populated community which is characterized by a high annual incidence of leptospirosis. Since 2003, Fiocruz's Centro de Pesquisa Gonçalo Moniz has been leading community-based cohort studies on lepstospirosis to understand the transmission dynamics of leptospirosis within this community. In the CHICAS group at Lancaster University, UK, we are helping with the analysis of the data. Spatial and spatio-temporal methods are being useful to identify the enviromental and social determinants of infection, as well as to understand the unexplained spatio-temporal structure of the process.

8.3 Final remarks

Merely collecting and analysing health data has little impact. However, successful surveillance programs also disseminate results to inform findings and strategies to the target audience (M'ikanatha et al., 2007). A particular strength of the spatial and spatio-temporal techniques is that outputs can be visualized through maps, greatly facilitating effective communication (Pfeiffer et al., 2008). Despite clear advances in some aspects of surveillance, there is still an urgent need to effectively address ethical, policy and legal concerns related to the readily access to detailed information of individual persons that might constrain access to data of potential public health importance (Teutsch and Thacker, 1995). Attention to these issues as well as further development of appropriate analytical methods, and advancement in collection and dissemination tools, can facilitate the effective planning, implementation, and growth of a variety of public health surveillance programs.

Bibliography

Anselin, L. (1995). Local indicators of spatial association - lisa. *Geographical Analysis*, 27:93–115.

Assunçao, R. and Krainski, E. (2009). Neighborhood dependence in bayesian spatial models. *Biometrical Journal*, 51:851–869.

Banerjee, S., Carlin, B. P., and Gelfand, A. G. (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman & Hall/CRC.

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.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathemathics*, 43:1–59.

Biggeri, A., Barbone, F., Lagazio, C., Bovenzi, M., and Stanta, G. (1996). Air pollution and lung cancer in trieste: spatial analysis of risk as a function of distance from sources. *Environmental Health Perspectives*, 104:750–754.

Box, G. E. P. and Jenkins, G. M. (1970). Time series analysis: Forecasting and control. *Holden-Day.*

Brammer, L., Budd, A., and Cox, N. (2009). Seasonal and pandemic influenza surveillance considerations: Mortality surveillance. *Influenza and Other Respiratory Viruses*, 3(2):51–58.

Brookmeyer, R. and Stroup, D. F. (2004). *Monitoring the health of populations. Statistical Principles & Methods for Public Health Surveillance.* Oxford University Press, New York.

Carlin, B. P. and Banerjee, S. (2003). Hierarchical multivariate car models for spatio-temporally correlated survival data. In *in Bayesian Statistics 7*, pages 45–63. University Press.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169:571–584.

CDC (2010). Principles of epidemiology in public health practice. Centers for Disease Control and Prevention (CDC), available at `http://www.cdc.gov/training/products/ss1000/ss1000-ol.pdf`.

Clegg, L. X., Hankey, B. F., Tiwari, R., Feuer, E. J., and Edwards, B. K. (2009). Estimating average annual per cent change in trend analysis. *Statistics in Medicine*, 28:3670–3682.

Collins, L. B. (1995). *Inter-event Distance Methods for the Statistical Analysis of Spatial Point Processes.* PhD thesis, Department of Statistics, University of Chicago.

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

Cressie, N. A. C. and Collins, L. B. (2001). Analysis of spatial point patterns using bundles of product density lisa functions. *Journal of Agricultural, Biological, and Environmental Statistics*, 6:118–135.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–22.

Dicker, R., Coronado, F., Koo, D., and Parrish, R. G. (2006). *Principles of Epidemiology in Public Health Practice, 3rd Edition.* Centers for Disease Control and Prevention (CDC), Atlanta.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns, 2nd editon.* Edward Arnold, London.

Diggle, P. J., Chetwynd, A. G., Haggkvist, R., and Morris, S. E. (1995). Second order analysis of space-time clustering. *Statistical Methods in Medical Research*, 4:124–136.

DOH (2010). Guidelines for using and developing rates for public health assessment. Washington State Department of Health, available at `http://www.doh.wa.gov/data/guidelines/guidelines.htm`.

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

Duczmal, L., Kulldorff, M., and Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease clusters. *Journal of Computational and Graphical Statistics*, 15:428–442.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28:181–187.

Elliott, P. and Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives*, 112:998–1006.

Fiksel, T. (1988). Edge-corrected density estimators for point processes. *Statistics*, 19:67–75.

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

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *Incorporated Statistician*, 5:115–145.

Gelfand, A. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4:11–25.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., editors (2010). *Handbook of Spatial Statistics*. Chapman & Hall/CRC, Boca Raton, FL.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulations using multiple sequences. *Statistical Science*, 7:457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741.

Geweke, J. (1992). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.* In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), Bayesian Statistics 4. Oxford University Press, New York.

Gómez-Rubio, V. and López-Quílez, A. (2010). Statistical methods for the geographical analysis of rare diseases. *Adv Exp Med Biol*, 686:151–171.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.

Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat.*, 61:79–90.

Jacquez, G. M. (1996). A k nearest neighbor test for space-time interaction. *Statistics in Medicine*, 15:1935–1949.

Jin, X., Carlin, B. P., and Banerjee, S. (2007). Order-free coregionalized areal data models with application to multiple disease mapping. *Journal of the Royal Statistical Society*, 69:817–838.

Katsouyanni, K., Trichopoulos, D., Kalandidi, A., Tomos, P., and Riboli, E. (1991). A case-control study of air pollution and tobacco smoking in lung cancer among women in athens. *Preventive Medicine*, 20:271–280.

Kelsall, J. E. and Diggle, P. J. (1995). Kernel estimation of relative risk. *Bernoulli*, (1)1/2:003–016.

Kim, H. J., Fay, M. P., Feuer, E. J., and Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, 19:335–351.

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

Kulldorff, M. (1997). A spatial scan statistic. *Commun. Statist.-Theory Meth.*, 26(1):1481–1496.

Kulldorff, M. (2006). Tests of spatial randomness adjusted for an inhomogeneity: A general framework. *Journal of the American Statistical Association*, 101(475):1289–1305.

Kulldorff, M. (2010). *SaTScan User Guide v9.0.* Martin Kulldorff and Information Management Services Inc., http://www.satscan.org.

Kulldorff, M., Athas, W. F., Feuer, E. J., Miller, B. A., and Key, C. R. (1998). Evaluating cluster alarms: A space-time scan statistic and brain cancer in los alamos, new mexico. *American Journal of Public Health*, 88(9):1377–1380.

Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R., and Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS Med*, 2(3):216–224.

Kulldorff, M. and Hjalmars, U. (1999). The knox method and other tests for space-time interaction. *Biometrics*, 55:544–552.

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

Langford, I. H., Leyland, A. H., Rasbash, J., and Goldstein, H. (1999). Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 48:253–268.

Lawson, A. B. (2006). *Statistical Methods in Spatial Epidemiology.* Wiley, New York.

Lawson, A. B. (2009a). *Bayesian Disease Mapping: hierarchical modeling in spatial epidemiology.* CRC Press, New York.

Lawson, A. B. (2009b). *Bayesian Disease Mapping: Hierarchical Modeling In Spatial Epidemiology.* Chapman & Hall/CRC.

Lawson, A. B. and Banerjee, S. (2010). Bayesian spatial analysis. In Fotheringham, S. and Rogerson, P., editors, *Handbook of Spatial Analysis*, chapter 9. Sage, New York.

Lawson, A. B. and Kleinman, K. (2005). *Spatial and Syndromic Surveillance for Public Health.* John Wiley & Sons, England.

Lee, L. M., Teutsch, S. M., Thacker, S. B., and Louis, M. E. S. (2010). *Principles and Practice of Public Health Surveillance.* Oxford University Press, New York.

Li, Y., Tiwari, R. C., and Zou, Z. (2008). An age-stratified poisson model for comparing trends in cancer rates across overlapping regions. *Biom J.*, 50(4):608–619.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data.* Wiley, New York.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Ma, H. and Carlin, B. P. (2007). Bayesian multivariate areal wombling for multiple disease boundary analysis. *Bayesian Analysis*, 2:281–302.

Ma, H., Virnig, B. A., and Carlin, B. P. (2006). Spatial methods in areal administrative data analysis. *Italian Journal of Public Health*, 3:94–104.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the gibbs sampler. *The American Statistician*, 48:188–190.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27:209–220.

Martínez-Beneito, M. A., López-Quílez, A., and Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27:2874–2889.

Mateu, J., Lorenzo, G., and Porcu, E. (2007). Detecting features in spatial point processes with clutter via local indicators of spatial association. *Journal of Computational and Graphical Statistics*, 16(4):1–23.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.

Michel, P., Wilson, J. B., Martin, S. W., Clarke, R. C., McEwen, S. A., and Gyles, C. L. (2000). Estimation of the under-reporting rate for the surveillance of escherichia coli o157:h7 cases in ontario, canada. *Epidemiology and Infection*, 125:35–45.

M'ikanatha, N. M., Lynfield, R., Beneden, C. A. V., and de Valk, H., editors (2007). *Infectious Disease Surveillance*. Wiley-Blackwell, Malden, MA.

MMWR (2001). Updated guidelines for evaluating public health surveillance systems. *Morbidity and Mortality Weekly Report (MMWR)*, 50(RR13). Available at `http://www.cdc.gov/mmwr/`.

MMWR (2004). Framework for evaluating public health surveillance systems for early detection of outbreaks. *Morbidity and Mortality Weekly Report (MMWR)*, 53(RR5). Available at `http://www.cdc.gov/mmwr/`.

Molinari, N. A., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., and Bridges, C. B. (2007). The annual impact of seasonal influenza in the us: measuring disease burden and costs. *Vaccine*, 25(27):5086–5096.

Moraga, P. and Lawson, A. B. (2012). Gaussian component mixtures and car models in bayesian disease mapping. *Computational Statistics & Data Analysis*, 56:1417–1433.

Moraga, P. and Montes, F. (2011). Detection of spatial disease clusters with lisa functions. *Statistics in Medicine*, 30:1057–1071.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37:17–23.

Muscatello, D. J., Morton, P. M., Evans, I., and Gilmour, R. (2008). Prospective surveillance of excess mortality due to influenza in new south wales: feasibility and statistical approach. *Communicable Diseases Intelligence*, 32(4):435–442.

Newall, A. T., Viboud, C., and Wood, J. G. (2010). Influenza-attributable mortality in australians aged more than 50 years: a comparison of different modelling approaches. *Epidemiol. Infect.*, 138(6):836–842.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem.* Geobooks, Norwich, U.K.

Ozonoff, A., Sukpraprut, S., and Sebastiani, P. (2006). Modeling seasonality of influenza with hidden markov models. In *Proceedings of the American Statistical Association, Section on Statistics in Defense and National Security.*

Page, P. E. (1954). Continuous inspection schemes. *Biometrika*, 41:100–115.

Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., and Clements, A. C. (2008). *Spatial analysis in epidemiology.* Oxford University Press.

Raubertas, R. F. (1989). An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Statistics in Medicine*, 8:267–271.

Ripley, B. D. (1976). The second order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1:239–250.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–357.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–590.

Serfling, R. E. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports*, 78:494–506.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product.* Van Nostrand Reinhold, Princeton.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC, Boca Raton, FL.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and der Linde, A. V. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64:583–616.

Stelling, J., Yih, W. K., Galas, M., Kulldorff, M., Pichel, M., Terragno, R., Tuduri, E., Espetxe, S., Binsztein, N., O'Brien, T. F., Platt, R., and WHONET-Argentina, C. G. (2010). Automated use of whonet and satscan to detect outbreaks of shigella spp. using antimicrobial resistance phenotypes. *Epidemiol Infect.*, 138(6):873–883.

Stern, H. S. and Cressie, N. A. C. (1999). Inference for extremes in disease mapping. In Lawson, A. B., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J. F., and Bertollini, R., editors, *Disease Mapping and Risk Assessment for Public Health*, chapter 5. Wiley, New York.

Strat, Y. L. and Carrat, F. (1999). Monitoring epidemiologic surveillance data using hidden markov models. *Statistics in Medicine*, 18:3463–3478.

Tango, T. (2010). *Statistical methods for disease clustering.* Springer.

Teutsch, S. M. and Thacker, S. B. (1995). Planning a public health surveillance system. *Epidemiol Bull.*, 16(1):1–6.

Thacker, S. B. and Berkelman, R. L. (1988). Public health surveillance in the united states. *Epidemiol Rev.*, 10:164–90.

Thompson, W. W., Weintraub, E., Dhankhar, P., Cheng, P. Y., Brammer, L., Meltzer, M. I., Bresee, J. S., and Shay, D. K. (2009). Estimates of us influenza-associated deaths made using four different methods. *Influenza and Other Respiratory Viruses*, 3:37–49.

Ugarte, M. D., Ibáñez, B., and Militino, A. F. (2006). Modelling risks in disease mapping. *Statistical Methods in Medical Research*, 15:21–35.

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C., and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175:49–82.

van Beers, W. C. M. and Kleijnen, J. P. C. (2003). Kriging for interpolation in random simulation. *Journal of the Operational Research Society*, 54:255–262.

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.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data.* Wiley, New York.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall/CRC, Boca Raton, FL.

Xia, H. and Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: mapping ohio lung cancer mortality. *Statistics in Medicine*, 17:2025–2043.

Xu, Z. Y., Blot, W. J., and Xiao, H. P. (1989). Smoking air pollution and the high rates of lung cancer in shenyang, china. *Journal of the National Institute of Cancer*, 81:1800–1806.

Zhang, Y., Hodges, J. S., and Banerjee, S. (2010). Smoothed anova with spatial effects as a competitor to mcar in multivariate spatial smoothing.

Zhang, Z. and Glaz, J. (2008). Bayesian variable window scan statistics. *Journal of Statistical Planning and Inference*, 138(11):3561–3567.

Zhang, Z. and Lin, G. (2009a). Cluster detection based on spatial associations and iterated residuals in generalized linear mixed models. *Biometrics*, 65(2):353–60.

Zhang, Z. and Lin, G. (2009b). Spatial scan statistics in loglinear models. *Computational Statistics and Data Analysis*, 53(8):2851–2858.

Resumen

Objetivos y metodología

La vigilancia en salud pública proporciona información útil para identificar los problemas de salud pública y dar una respuesta apropiada cuando éstos ocurren. Esta información es crucial para prevenir y controlar una variedad de condiciones de salud como enfermedades infecciosas, enfermedades crónicas, daños en la salud, y comportamientos relacionados con la salud. Una vigilancia de calidad es necesaria para entender el verdadero estatus de salud en la población y guiar el uso de recursos limitados. Bajo sistemas de vigilancia inadecuados, los responsables están desinformados y pueden perder oportunidades para la aplicación temprana de medidas de prevención y control. En estas situaciones, es posible el resurgimiento de enfermedades previamente erradicadas o la extensión incontrolada de enfermedades como en el caso del VIH/SIDA (M'ikanatha et al., 2007). La vigilancia en salud pública implica cuatro principales actividades integradas: la recopilación de datos de salud, el análisis de los datos, su interpretación y la diseminación oportuna de los resultados a los responsables de responder a las necesidades de salud de la población. Los sistemas de vigilancia capturan características espaciales, temporales y personales de los datos de salud. Las tasas de incidencia y mortalidad cuantifican el tamaño de los problemas de salud en una población dada y proporcionan las bases para iniciar medidas de control de enfermedades y evaluar su efectividad. Las tendencias temporales y las comparaciones entre grupos demográficos y étnicos pueden proporcionar importantes claves sobre la etiología de la enfermedad.

El aumento de la disponibilidad de datos de salud y de población geográficamente georreferenciados, así como el desarrollo de los sistemas de información geográfica (SIG), han facilitado el aumento de investigaciones sobre las variaciones espaciales y espacio-temporales de enfermedades. La investigación del brote de cólera en Londres en 1854 por John Snow proporciona uno de los ejemplos más famosos de análisis espacial. Snow utilizó un mapa para ilustrar como las muertes por cólera parecían situarse alrededor de una fuente pública. La evaluación del patrón espacial de los casos de cólera fue importante para identificar la fuente de infección y apoyó la teoría de transmisión de cólera a través de agua contaminada. Existe un amplio rango de métodos espaciales y espacio-temporales que pueden aplicarse como herramienta de vigilancia incluyendo mapas de enfermedades, clustering y estudios de correlación geográfica. Muchos de estos métodos pueden ser usados para destacar áreas con un alto riesgo, detectar clusters de enfermedades, la detección temprana de epidemias, evaluar el riesgo de enfermedad en relación a una fuente putativa, e identificar factores de riesgo de enfermedades. Desafortunadamente, un mal uso de los métodos estadísticos puede ser altamente engañoso. Por tanto, un profundo entendimiento de los posibles problemas como cambios en la definición de los casos y cuestiones relativas a la integridad de los datos son críticos en el análisis de los datos y la interpretación de los resultados.

Durante las últimas décadas, la vigilancia en salud pública ha sufrido un desarrollo considerable. Entre las actividades que han contribuido a su avance están las innovaciones tecnológicas como la monitorización en tiempo real y avances en SIG, el desarrollo de nuevos métodos estadísticos y herramientas computacionales para aplicarlos, y el uso más efectivo de medios electrónicos y otras herramientas de comunicación que facilitan la diseminación de la información (Brookmeyer and Stroup, 2004). Asimismo, la vigilancia en salud pública ha cambiado en respuesta a nuevas preocupaciones, como acciones de terrorismo biológico o las enfermedades y epidemias relativamente nuevas, como el síndrome respiratorio agudo severo (SARS). A medida que cambios en la salud pública se hacen necesarios y existen nuevas herramientas y las capacidades computacionales aumentan, los métodos estadísticos para la vigilancia de enfermedades deben continuar evolucionando para mejorar la calidad de los análisis, y la interpretación y visualización de los resultados en la forma más útil y en el marco de tiempo apropiado para satisfacer los intereses de quienes elaboran y toman decisiones. El objetivo de esta tesis es proponer nuevas técnicas para ayudar a la práctica de la vigilancia en salud pública. En particular, nos centramos en métodos espaciales y espacio-temporales que puedan ayudar en temas como la existencia de datos faltantes (Capítulo 4), modelizar la heterogeneidad correlacionada en mapas de enfermedades (Capítulo 5), detectar clusters de enfermedades (Capítulo 6), y elucidar variaciones espaciales en tendencias temporales (Capítulo 7).

Comenzamos con una visión general de la vigilancia en salud pública y de los datos espaciales. El Capítulo 2 proporciona una introducción a los sistemas de vigilancia, así como una revisión de los métodos estadísticos que han sido aplicados en la vigilancia de enfermedades. Los métodos considerados incluyen el cálculo de tasas, tendencias temporales, detección de clusters y brotes, y mapas de enfermedades. Generalmente, los datos espaciales se clasifican en tres grandes tipos: retículo, geoestadísticos y patrones puntuales. El Capítulo 3 está dedicado a la revisión de sus características básicas y los métodos de análisis. El objetivo de estos dos capítulos es proporcionar una base de los conceptos y métodos estadísticos usados en vigilancia que pueden ayudar en el desarrollo de los siguientes capítulos.

Los capítulos 4-7 están basados en cuestiones de interés particulares. El Capítulo 4 trata el problema de datos faltantes y se centra en el análisis de datos de mortalidad de todas las causas y de neumonía y gripe (P&I) en los Estados Unidos. Estimaciones nacionales de la mortalidad de todas las causas y de P&I derivadas de estos datos tratan todos los valores faltantes como ceros. El efecto de esta decisión metodológica es sesgar las estimaciones y subestimar la verdadera mortalidad. Para evaluar el impacto de este tratamiento de valores faltantes en las estimaciones nacionales, proponemos un procedimiento basado en una regresión que utiliza información relevante para imputar los valores faltantes y así producir una estimación más precisa de la mortalidad. Consideramos y evaluamos varios modelos que predicen el número de muertes semanales por ciudad, semana de calendario, año de calendario y grupo de edad. Crosvalidamos estos modelos y calculamos estimaciones revisadas de la mortalidad por todas las causas y por P&I imputando los valores faltantes y estimamos el exceso de mortalidad de P&I usando un procedimiento de regresión recomendado por el CDC (Serfling, 1963). Por último, comparamos las estimaciones con y sin imputación para entender el impacto de este tratamiento de datos no disponibles en las estimaciones nacionales.

El Capítulo 5 está dedicado a mapas de enfermedades. Los modelos jerárquicos Bayesianos que utilizan componentes condicionales autorregresivas (CAR) son comúnmente usados en mapas de enfermedades (Besag et al., 1991). Un modelo alternativo a las componentes CAR propia o impropia es el modelo mixtura de componentes Gaussianas (GCM) (Langford et al., 1999). En este capítulo llevamos a cabo una revisión de los modelos CAR y GCM en escenarios univariables donde sólo se considera una enfermedad, y también en situaciones multivariables donde además de la dependencia espacial entre regiones, se analiza la dependencia entre múltiples enfermedades. Mostramos una comparación del comportamiento de los modelos usando un conjunto de datos simulados. Además, utilizamos los modelos GCM y CAR para estimar el riesgo relativo de bajo peso al nacer en Georgia, Estados Unidos, en el año 2000.

La detección de clusters de enfermedades es una importante herramienta en epidemiología que puede ayudar a identificar factores de riesgo y a entender su etiología. En el Capítulo 6 proponemos un método para la detección de clusters espaciales donde se disponen de las localizaciones de un conjunto de casos y un conjunto de controles. El método está basado en funciones indicadores locales de asociación espacial (LISA) (Anselin, 1995), particularmente en el desarrollo de una versión local de la densidad producto, una característica de segundo orden de los procesos puntuales espaciales. El comportamiento del método es evaluado y comparado con el estadístico de escaneo espacial de Kulldorff (Kulldorff and Nagarwalla, 1995) por medio de un estudio de simulación. Ambos métodos se aplican para detectar clusters espaciales de enfermedades renales en la ciudad de Valencia, España, en el año 2008.

Los métodos para la evaluación de las variaciones espaciales en tendencias temporales (SVTT) son importantes herramientas para la vigilancia de enfermedades que pueden ayudar a los gobiernos a formular programas para prevenir enfermedades, y medir el progreso, impacto y eficacia de esfuerzos preventivos ya en operación. El método lineal SVTT está diseñado para detectar áreas con tendencias lineales de enfermedades inusualmente diferentes (Kulldorff, 2010). En algunas situaciones, sin embargo, el mal ajuste del procedimiento de estimación de tendencias puede llevar a conclusiones equivocadas. En el Capítulo 7, proponemos el método cuadrático SVTT como alternativa al método lineal SVTT. Llevamos a cabo una comparación entre los métodos lineal y cuadrático usando un conjunto de datos simulados para ayudar a ilustrar sus respectivas propiedades. Por último, aplicamos el método cuadrático para detectar tendencias de cáncer cervical inusualmente diferentes en mujeres blancas en los Estados Unidos, durante el periodo 1969 a 1995.

Conclusiones

Una metodología estadística apropiada es necesaria para la vigilancia en salud pública. Durante la pasada década, el uso de técnicas espaciales y espacio-temporales ha sido cada vez más común gracias a los avances en computación y a la disponibilidad de bases de datos georreferenciadas. En esta tesis tratamos varios aspectos relacionados con el análisis de datos de salud georreferenciados y desarrollamos métodos estadísticos que pueden ser usados para ayudar en la práctica de la vigilancia.

Estimación de valores faltantes en datos de mortalidad basada en modelos

En el Capítulo 4 presentamos un modelo para estimar los valores faltantes en los datos de mortalidad debidos a cualquier causa y de neumonía y gripe (P&I) de las 122 ciudades operadas por los Centros de Control y Prevención de Enfermedades (CDC). Este trabajo fue desarrollado bajo la supervisión de Al Ozonoff en la Harvard School of Public Health, y como resultado un artículo está en preparación. En este capítulo proponemos diferentes especificaciones que modelizan el número de muertes semanales por ciudad, semana y año de calendario y grupo de edad, y seleccionamos el mejor modelo evaluando la media y la varianza de los errores de predicción de los modelos mediante un procedimiento de validación cruzada. El modelo seleccionado usa información de los cinco anteriores y los cinco posteriores años, semanas y grupos de edad, y de las tres ciudades más cercanas. En nuestros cálculos no incluimos ciudades con más de 70% de valores faltantes ya que el modelo funciona mal en ciudades con un gran número de valores faltantes. Calculamos la mortalidad y el exceso de muertes usando un conjunto de datos donde los valores faltantes han sido imputados con el modelo propuesto, y un conjunto de datos en el que los valores faltantes se asumen como ceros. La comparación de los resultados obtenidos con ambos procedimientos refleja algunas diferencias en la mortalidad y en las tendencias temporales. Estos resultados nos llevan a concluir que es necesario algún procedimiento para tratar los datos faltantes y no basta con imputarles un valor cero, porque los análisis podrían producir estimaciones imprecisas y conclusiones incorrectas. El modelo presentado es simplemente una de las posibles alternativas para imputar los valores faltantes, pero es un procedimiento útil para mejorar las inferencias en situaciones con un número moderado de valores faltantes.

Mixtura de componentes Gaussianas (GCM) y modelos CAR en mapas Bayesianos de enfermedades

En el Capítulo 5, presentamos dos posibles estructuras para modelizar la heterogeneidad correlada en modelos jerárquicos Bayesianos: los modelos GCM y CAR. Explicamos sus respectivas propiedades en situaciones univariantes y en situaciones multivariantes en las que se modeliza el riesgo de múltiples enfermedades. Llevamos a cabo un estudio de simulación en ambas situaciones para comparar el comportamiento de estas estructuras cuando se usan para ajustar varios modelos simulados que abarcan un amplio rango de situaciones que pueden encontrarse en la realidad. Los resultados del estudio univariante sugieren que en la mayoría de las situaciones las estructuras GCM y CAR modelizan de forma similar. Observamos que en muchos modelos los patrones del riesgo relativo y de los efectos UH y CH son muy similares, y que los criterios de bondad del ajuste indican un ajuste similar de los datos. Uno de los objetivos del estudio perseguía analizar el comportamiento de los diferentes modelos ajustados al estimar los efectos de las covariables. Concluimos que diferentes patrones de covariables proporcionan diferentes resultados, y que para algunas covariables GCM y CAR producen estimaciones similares. Del mismo modo, los resultados del estudio multivariable indican un comportamiento comparable de las estructuras MGCM y MCAR, con tan sólo pequeñas diferencias en los patrones de riesgo relativo y la bondad del ajuste. Este trabajo puede encontrase también en Moraga and Lawson (2012).

Las estructuras CAR son muy populares en modelos jerárquicos para la estimación en áreas pequeñas, y ofrecen un mecanismo robusto y flexible para modelizar la heterogeneidad correlada. Sin embargo, este trabajo muestra que no son los únicos modelos ni necesariamente los óptimos para este tipo de datos. Los estudios de simulación llevados a cabo muestran que las estructuras GCM y CAR tienen un comportamiento comparable en un amplio rango de situaciones, e incluso identifican algunos tipos de situaciones donde las estructuras GCM aventajan a las CAR. Estos resultados nos llevan a concluir que los modelos GCM son una buena alternativa a los modelos CAR.

Detección de clusters de enfermedad con funciones LISA

El método para la detección de clusters en estudios caso-control presentado en el Capítulo 6 puede encontrarse en Moraga and Montes (2011). El método se basa en funciones indicadores locales de asociación espacial (LISA) de la densidad producto del patrón puntual de los casos, y detecta los casos que forman parte de una zona de aglomeración. Tiene la ventaja que no necesita especificación *a priori* de la forma del cluster y es capaz, por tanto, de detectar clusters de cualquier forma. Los resultados del método dependen del tamaño muestral, el tamaño del cluster y de la densidad de casos dentro del cluster. Es importante destacar que es necesario un número considerable de controles para que las estimaciones de las funciones LISA sean buenas y el método funcione. Una importante desventaja del método LISA cuando lo comparamos con el método de escaneo es que el error de tipo I puede ser grande cuando el número de casos es grande. Concluimos que el método LISA es muy útil en situaciones donde existen clusters ya que es capaz de detectar clusters de cualquier forma y con una alta sensibilidad y especificidad.

Variaciones espaciales en tendencias temporales (SVTT)

En el Capítulo 7 presentamos el método SVTT cuadrático para la detección de áreas con tendencias temporales de enfermedades inusualmente diferentes. Este método ha sido desarrollado bajo la supervisión de Martin Kulldorff en la Harvard Medical School, y ha sido implementado en el software `SaTScan`, www.satscan.org. El método SVTT cuadrático está basado en estadísticos de escaneo espacial, y es una versión modificada del método SVTT lineal porque se cambia el procedimiento de estimación de la tendencia. Usando varios ejemplos observamos que el método cuadrático es capaz de detectar áreas con tendencias diferentes en situaciones en las que el método lineal falla, produciendo también mejores estimaciones de las tendencias reales. Se ha llevado a cabo un estudio de simulación para comparar el comportamiento de los métodos cuadrático y lineal. Los métodos se han aplicado a la detección de áreas con tendencias temporales diferentes en conjuntos de datos simulados, generados de tal manera que abarquen un amplio rango de situaciones que pueden encontrarse en la realidad. Los resultados indican un comportamiento comparable de ambos métodos en situaciones con tendencias lineales. Sin embargo, cuando las tendencias no son lineales, el método cuadrático es mejor que el método lineal y presenta mayor sensibilidad, especificidad y valores predictivos positivos (PPV). El método cuadrático es muy útil para la evaluación de las medidas de control y prevención. Cuando se aplica para detectar áreas con tendencias inusualmente diferentes de cáncer cervical en mujeres blancas de Estados Unidos durante el periodo de 1969 a 1995, se observa que globalmente la mortalidad disminuye pero no de la misma manera en todas las áreas. Específicamente, se detectan tres áreas de interés donde el comportamiento de la tendencia del riesgo es significativamente diferente del resto.

Futuras líneas de investigación

Detección de clusters espacio-temporales con funciones LISA

No solo los clusters espaciales de enfermedades son de gran interés, también los clusters espacio-temporales son importantes y pueden presentarse a menudo. El método LISA espacial descrito en el Capítulo 6 puede ser fácilmente adaptado para buscar clusters espacio-temporales si consideramos los casos y los controles como puntos tridimensionales con coordenadas representando las localizaciones y las fechas asociadas a cada individuo, y calculando las distancias en $\mathbb{R}^3$. Adaptaremos el método espacial LISA al marco espacio-temporal. Para ello, debemos transformar primero la región de estudio y los puntos del conjunto de datos de tal manera que estén en un sistema de coordenadas en el que las distancias entre puntos tridimensionales tengan sentido. Específicamente, deberemos tener un sistema de coordenadas donde las distancias geográficas y temporales signifiquen lo mismo, en términos de proximidad y lejanía de los puntos. Después de llevar a cabo la transformación de la región de estudio y del conjunto de datos, los pasos del método espacio-temporal serán los mismos que los pasos del método espacial pero considerando los puntos y las distancias en $\mathbb{R}^3$.

En este trabajo llevaremos a cabo un estudio de simulación para evaluar el comportamiento del método LISA cuando es aplicado para detectar clusters espacio-temporales. Aunque muchos clusters de enfermedades pueden tener una forma regular, es muy común encontrar clusters con forma irregular en situaciones reales, como concentraciones de enfermedad a lo largo de las orillas de ríos y océanos, vías de transporte o focos de contaminación atmosférica. Además, algunos clusters se identifican en un punto fijo en el espacio, mientras que en otros el foco espacial se desplaza en el tiempo (Duczmal and Assunçao, 2004; Duczmal et al., 2006; Biggeri et al., 1996; Katsouyanni et al., 1991; Xu et al., 1989). En el estudio de simulación será importante usar zonas de cluster con características similares a las zonas que puedan encontrarse en situaciones reales. Así, algunas de las zonas de cluster construidas deberán tener formas irregulares o crecer o desplazarse en el tiempo.

Detección de clusters de enfermedades basada en modelos

Las técnicas existentes para la detección de clusters de enfermedades tienen varias limitaciones: no pueden tratar algunos tipos de datos, no modelizan adecuadamente situaciones con exceso de ceros, o no tratan factores de confusión o el sesgo. Por lo tanto, un procedimiento basado en modelos sería de interés para explorar la incidencia de la enfermedad y los posibles factores de riesgo. Desarrollaremos un método para detectar y evaluar clusters espaciales, temporales, y espacio-temporales que extienda las técnicas existentes y supere sus limitaciones. Para este fin, usaremos un método flexible basado en Modelos Lineales Generalizados (GLMs) que permita la incorporación de información sobre covariables y efectos aleatorios. Específicamente, el método deberá considerar un gran número de posibles clusters en la región de estudio y construir variables binarias asociadas con cada uno. La colección de posibles clusters que el método contrastará no está compuesta de todas las configuraciones posibles de zonas agregadas en la región de estudio. En el marco espacial, por ejemplo, los clusters serán definidos como círculos o elipses de tamaño geográfico entre cero y un límite superior definido en base a un porcentaje de la población en riesgo o en términos de la extensión geográfica de la enfermedad. Después, para cada uno de los posibles clusters, se ajustará un modelo relacionando la incidencia, mortalidad u otra medida de la enfermedad y la variable binaria. La selección de los mejores modelos se hará usando técnicas generales de selección de modelos como el AIC. Los modelos seleccionados indicarán los clusters más probables y su significación. Este método ofrece varias ventajas sobre los existentes ya que los modelos pueden especificarse usando diferentes distribuciones para la variable respuesta y pueden incluir efectos fijos y aleatorios. Específicamente, pueden considerarse diferentes tipos de variables como Poisson o Binomial. Los factores de confusión puede tratarse incluyendo covariables en el modelo. Además, la incorporación de efectos aleatorios hace posible tratar la sobredispersión y el sesgo de la medida.

Este método será implementado en un nuevo paquete de `R` llamado `DClusterm` usando técnicas que son descritas en Zhang and Lin (2009b), Zhang and Glaz (2008) (GLMs), Zhang and Lin (2009a) (modelos mixtos), y Gómez-Rubio and López-Quílez (2010) (modelos inflados de ceros). Este trabajo fue iniciado durante el programa Google Summer of Code 2011 en el cual participé bajo la supervisión de Virgilio Gómez-Rubio y Barry Rowlingston.

Trabajo aplicado

También estamos trabajando en dos proyectos aplicados que buscan entender la dinámica del sarampión en Europa y la incidencia de la leptospirosis en una favela de Salvador de Bahía, Brasil. Estos análisis ayudaran a los responsables a tomar decisiones para asignar recursos y dirigir las intervenciones de una manera efectiva. El Centro Europeo de Prevención y Control de Enfermedades (ECDC) busca fortalecer las defensas de Europa contra las enfermedades infecciosas. Con la erradicación del sarampión como objetivo, uno de sus proyectos estudia la distribución espacio-temporal del sarampión en Europa. Mapas de enfermedades, tendencias temporales y métodos para la detección de clusters están siendo de utilidad para evaluar la distribución de la enfermedad.

El proyecto sobre leptospirosis se sitúa en Pau da Lima, un barrio pobre que creció a las afueras de la parte urbana de Salvador de Bahía. Pau da Lima alberga una comunidad densamente poblada caracterizada por un alta incidencia anual de leptospirosis. Desde 2003, el Centro de Pesquisa Gonçalo Moniz, Fiocruz, efectúa estudios de cohorte sobre lepstospirosis para entender la dinámica de transmisión de la enfermedad en esta comunidad. En el grupo CHICAS de la Universidad de Lancaster se está colaborando con el análisis de los datos. Métodos espaciales y espacio-temporales están siendo de utilidad para identificar los determinantes medioambientales y sociales de la infección, así como para entender la estructura espacio-temporal latente del proceso.

Comentarios finales

La simple recopilación y el análisis de datos de salud tiene poco impacto. Sin embargo, programas de vigilancia efectivos permiten difundir los resultados e informan sobre los descubrimientos y estrategias (M'ikanatha et al., 2007). Una ventaja particular de las técnicas espaciales y espacio-temporales es que los resultados pueden visualizarse a través de mapas, facilitando una comunicación efectiva (Pfeiffer et al., 2008). A pesar de claros avances en algunos aspectos de la vigilancia, todavía existe una necesidad urgente de tratar eficazmente cuestiones éticas, políticas y legales relacionadas con el fácil acceso a información detallada de personas individuales que puede restringir el acceso a datos de importancia para la salud pública (Teutsch and Thacker, 1995). La atención a estos asuntos, así como el desarrollo de métodos analíticos apropiados y el avance de las herramientas de recopilación y diseminación, pueden facilitar la planificación, la implementación y el crecimiento de programas de vigilancia en salud pública.