Space-time surveillance of emerging infectious disease

by

Colin Robertson
M.Sc., University of Victoria, 2007
B.A., Simon Fraser University, 2002

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Geography

Supervisory Committee

Space-time surveillance of emerging infectious disease

by

Colin Robertson
M.Sc., University of Victoria, 2007
B.A., Simon Fraser University, 2002

Supervisory Committee

Dr. Trisalyn Nelson, Supervisor
(Department of Geography)

Dr. Alec Ostry, Departmental Member
(Department of Geography)

Dr. Farouk Nathoo, Outside Member
(Department of Mathematics & Statistics)

Dr. Craig Stephen, Additional Member
(Faculty of Veterinary Medicine, University of Calgary)

Abstract

Supervisory Committee

Dr. Trisalyn Nelson, Supervisor
(Department of Geography)

Dr. Alec Ostry, Departmental Member
(Department of Geography)

Dr. Farouk Nathoo, Outside Member
(Department of Mathematics & Statistics)

Dr. Craig Stephen, Additional Member
(Faculty of Veterinary Medicine, University of Calgary)

ABSTRACT

Emerging diseases are an increasingly important public health problem. This research investigates space-time disease surveillance for emerging infectious diseases. A system was developed in Sri Lanka monitoring clinical diagnoses in cattle, poultry and buffalo. Veterinarians submitted surveys using mobile phones and GPS. This surveillance system proved to be both feasible and acceptable and provided timely information on animal health patterns in Sri Lanka. A critical review of software and methods for space-time disease surveillance provides guidance on the selection and implementation of appropriate analytic methods for surveillance data. For the data collected in this research, a hidden Markov model is developed which estimates region-specific prevalence estimates after controlling for sentinel-level factors. The use of cluster detection methods in surveillance research is demonstrated using data from an outbreak of suspected leptospirosis in Sri Lanka in 2005-2009.

Table of Contents

List of Tables

List of Figures

Acknowledgments

First and foremost I want to thank my advisor, Trisalyn Nelson. It has been my honor to be her first Ph.D. student. Over the five years that Trisalyn has supervised me, she has been a superb mentor. I appreciate all her contributions of energy, time, ideas, and funding to make my Ph.D. and my entire time here at UVIC productive, stimulating and most of all, enjoyable. I have grown and learned a lot over these five years and I am extremely grateful to Trisalyn for all she has contributed.

My PhD work would not have been possible without Dr. Craig Stephen. Very few graduate students are given the opportunity to work on a large multidisciplinary project and play as large a role as I have been able to. Thanks also to committee members Dr. Farouk Nathoo and Dr. Alec Ostry. I am very thankful for the help of Professor Andrew Lawson and Dr. Ying MacNab on the review of methods for space-time disease surveillance.

Working in Sri Lanka was an experience that I will always remember fondly. There are too many people to thank by name that provided support of all kinds and made my time there both fun and productive. First I would like to thank Sam Daniels. Countless times, the project was saved with a simple phone call or visit from Sam and I thank him for everything that his work has contributed to my research. Thanks to Preeni Abeynayake and Indra Abeygunawardena for always providing everything I needed in Sri Lanka, and for welcoming a Canadian geographer into the halls of the Faculty of Veterinary Science at the University of Peradeniya. Finally I'd like to extend a special thank you to Suraj Gunawardana. He worked tirelessly, often travelling all over Sri Lanka with us, and in the process became a good friend.

I would like to thank all current and former members of the SPAR lab for both making the lab a fun and stimulating place, and putting up with me when I wasn't in the best of moods! Carson Farmer, Jed Long, Mary Smulders, Ben Stewart, Nick Gralewicz, Katheryn Morrison, Jessica Fritterer and Jack Teng have all been fantastic labmates through the years.

Thanks to many friends, especially Douglas Braun, Chris Pasztor, Christy Lightowlers and Kate Sawford. Finally I'd like to thank Erin Hegan sincerely, for being a source of joy throughout my PhD.

Last, I would like to thank my whole family for all their love and encouragement, and especially my parents who have been unwavering cheerleaders and supporters of my academic life. Without them I would not have been able to finish. Thank you.

Colin Robertson

University of Victoria
Dec 2010

Co-Authorship Statement

All the manuscripts (Chapters 2-6) were co-authored, and the following outlines each of the authors' contributions, as well as the doctoral candidate, Colin Robertson:

Chapter 2: Colin Robertson and Kate Sawford identified and designed the research program, performed the research, analyzed the data, and prepared the manuscript. Sam Daniels, Craig Stephen and Trisalyn Nelson aided in the preparation of the manuscript with comments, edits and advice on content.
published in Emerging Infectious Diseases

Chapter 3: Colin Robertson designed the review, performed the review, and prepared the manuscript. Trisalyn Nelson, Andrew Lawson, and Ying MacNab aided in the preparation of the manuscript with comments, edits and advice on content.
published in Spatial and Spatio-temporal Epidemiology

Chapter 4: Colin Robertson designed the review, performed the review, and prepared the manuscript. Trisalyn Nelson aided in the preparation of the manuscript with comments, edits and advice on content.
published in International Journal of Health Geographics

Chapter 5: Colin Robertson identified and designed the research program, performed the research, analyzed the data, and prepared the manuscript. Kate Sawford, Craig Stephen and Trisalyn Nelson aided in the preparation of the manuscript with comments, edits and advice on content.

Chapter 6: Colin Robertson identified and designed the research program, performed the research, analyzed the data, and prepared the manuscript. Trisalyn Nelson and Craig Stephen aided in the preparation of the manuscript with comments, edits and advice on content.

Chapter 1: Introduction to space-time disease surveillance for emerging diseases

1.1 Introduction

Infectious diseases are of increasing importance due to the emergence of new pathogens (Daszak et al. 2000) and the persistence and resurgence of older diseases (Keeling and Gilligan 2000). Over 15 million people die each year due to infectious diseases (Morens et al. 2004). Many infectious diseases have recently emerged or expanded their prevalence and geographic range. Developing the capacity to predict a newly emerging infectious disease (EID) is increasingly becoming an imperative for the global public health community. The primary means of limiting morbidity, mortality and socio-economic impacts due to EID is disease surveillance, defined by the World Health Organization (WHO) as the ongoing systematic collection, collation, analysis and interpretation of data and the dissemination of information to those who need to know in order for action to be taken (World Health Organization 2007). Analysis of disease-related data, as part of surveillance, serves a number of public health functions, one of which is detecting the presence of unusual health outcome events (Wagner et al. 2006). Detection of unusual patterns of disease, and the ecological and socioeconomic conditions that contribute to these patterns can facilitate timely detection and prediction of outbreaks, and support timely interventions that may slow down or help contain an epidemic of an infectious disease. Whereas many surveillance systems aggregate data to examine only the temporal pattern of cases, incorporating both spatially and temporally explicit analysis methodologies and novel sources of spatial data may facilitate the

detection of unusual disease events, and contribute to a greater understanding of disease emergence processes.

Geographers have long been interested in analyzing patterns of disease across space (Light 1944; Banks 1955; Haggett 1992). Many of the processes giving rise to EIDs are themselves traditional areas of geographic inquiry: urbanization, land-use change, international mobility and travel patterns, physical changes in climate, ocean salinity, and species distributions (Morse 1995; Haggett 1994). The defining characteristic of disease emergence is change (Buchanan et al. 2006). Geographical approaches such as geographical information systems (GIS), remote sensing, and spatial analysis, are becoming widely used in the study and analysis of disease patterns. Spatial representation of disease cases, risks, and exposures may enhance our understanding of specific processes such as the basic reproductive rate or the nature of the transmission cycle (e.g., Odiit et al. 2006; Lai et al. 2004), facilitate more timely outbreak detection (e.g., Lawson and Kleinman 2005), and improve design and evaluation of control strategies (e.g., Morrison et al. 1998). In particular, these approaches are useful for infectious disease epidemiology concerned with the spread of new and persistent diseases in space and time.

This dissertation investigates space-time disease surveillance for EIDs. As the world becomes increasingly interdependent, changing ecologies are creating new opportunities for novel infections (Haggett 1994). Proof of this lies in the growing number of EIDs over the last four decades (Jones et al. 2008). The role geographic analysis can play in detecting, understanding, and forming response to EIDs is a central focus of this research.

Opportunities to monitor changes in variables related to disease emergence are greater now than they ever have been in the past. There are many sources of data tracking environmental and socioeconomic processes, human and animal movements, often in different formats, at varied spatial and temporal scales. Satellite imagery for example, has been widely employed in spatial studies of disease pattern (e.g., Bogh et al. 2007, Odiit et al. 2006). There are also increasing ways data are collected and integrated into automated systems. Cell phones can be made to keep a record of people's movements through physical and cultural landscapes that affect health outcomes (Wiehe et al. 2008). Sales data recording purchases of health-related products are being incorporated into early-warning, pre-diagnostic surveillance systems (Edge et al. 2006). Website queries are being used to monitor trends in influenza (Hulth et al. 2009; Ginsberg et al. 2009). The pace of development of surveillance systems has been staggering. According to one estimate, by as early as 2003 over 100 surveillance systems at the state and municipal level were operating in the United States alone (Buehler et al. 2003), and the number is likely much larger today. However, there remains debate as to the success of newly developed approaches to disease surveillance (Reingold 2003; Stoto et al. 2004). Evaluation of EID surveillance systems is also notoriously difficult to implement (Vrbova et al. 2009).

Disease surveillance has evolved to become a critical component of public health infrastructures. However, when applied to EIDs, a number of important challenges arise. Epidemiological investigations are traditionally inspired by disease surveillance based on case reporting and/or laboratory testing (Teutsch and Churchill 2000). However, with surveillance for EIDs the unit of analysis is often an indicator of risk – rather than actual

cases of disease. In the absence of an actual emergence, detailed response plans are often lacking (e.g., Britch et al. 2007). In general, approaches to surveillance that monitor indicators of disease risk are highly sensitive but not very specific (Fricker Jr. and Rolka 2006). In systems tracking pre-diagnostic data, false alarms are a major issue, which hampers interpretation of "signals". It seems that when systems rely on increasingly sophisticated statistical methods and data sources further removed from the pathogen, the plausibility of action declines (Fearnly 2008). The challenge is therefore to develop acceptable and useful EID surveillance systems that are based on sound statistical methods and can be incorporated into existing public/animal health infrastructures. This dissertation will help promote a critical understanding of EID surveillance methodology to help address these issues.

An additional consideration when building EID surveillance is the social, environmental, and economic contexts within which pathogens emerge. Zoonotic EIDs frequently occur in developing and/or tropical countries (Jones et al. 2008). EID surveillance in resource-limited settings incurs many additional challenges. Technological infrastructure is often lacking – making electronic surveillance networks difficult or unfeasible to implement. Further, traditional laboratory surveillance facilities are often poorly developed or non-existent, making validation data difficult to obtain. Building surveillance systems within lower-resource settings requires careful consideration of these and many other factors.

Through this research and with these challenges in mind, I develop an understanding of space-time disease surveillance for emerging infectious disease from three perspectives: theoretical, methodological, and applied. Table 1.1 presents a

summary of the specific challenges related to EID surveillance that are investigated in this dissertation, and how these challenges are approached.

Ultimately, the fundamental question tackled in this dissertation is:

Can you build a surveillance system to detect an emerging disease in a resource-limited setting by taking advantages of new communications technology and advances in geographic analysis?

This research and the larger project (Veterinary public health as part of the global response to emerging diseases. Building a sustainable model in Sri Lanka with extension to South and Southeast Asia, Teasdale-Corti Global Health Research Partnership Program) of which it is part, addressed these questions by observing two salient features of emerging diseases, they occur disproportionately in developing countries and are usually of animal origin (Jones et al. 2008).

These two facts formed the basis for the system developed in this thesis, called the Infectious Disease Surveillance and Analysis System (IDSAS). This prototype system, developed in Sri Lanka in coordination with the Department of Animal Production and Health, tracked syndromes and clinical diagnoses made by veterinarians. Chapter 2 reports on the development, implementation, and experience of building this system in Sri Lanka. Working within a resource-limited setting provided a number of challenges. Mobile-phone based surveys were employed by participating veterinarians to record data on animal health indicators not traditionally tracked in Sri Lanka. The objectives of the overall IDSAS project was to develop an operational understanding of deploying mobile surveillance in a resource-limited setting, assess the feasibility and acceptability of such a system, and promote capacity building related to emerging

infectious diseases among veterinarians in Sri Lanka. This chapter highlights lessons learned building surveillance capacity in Sri Lanka, of which fundamental aspects include the importance of political support for surveillance methods amongst users and stakeholders. Over 5500 clinical diagnoses reports were received by system during 2009.

Analysis of surveillance data from novel populations incurs many challenges. In Chapters 3 and 4 I review methods and software for surveillance data analysis. The motivation for these chapters was to develop a thorough understanding of quantitative surveillance methodology, and ultimately develop an approach to analyze the data gathered from IDSAS. The review of methods considers both testing and modelling-based approaches, and reviews them in terms of system scale, scope, function, technical considerations, and disease characteristics. Much of the statistical and computer science literature developing surveillance methods consider methodology solely in terms of algorithm performance (i.e., sensitivity, specificity, timeliness). Yet in practice, new data sources require an understanding of the baseline patterns before any outbreak detection can be attempted. And more importantly, the objectives of specific surveillance systems vary greatly. Method selection, if it is to be of practical surveillance value, needs to consider multiple factors related to implementation, sustainability, and use. A secondary aspect of surveillance methods that is rarely discussed in the literature is the availability of software. Chapter 4 considers this problem in detail. Four software packages are reviewed and issues related to data preprocessing, methods, technical issues, analytic output, and user facility are discussed.

Analysis of IDSAS data makes up the major contribution of Chapter 5. As noted earlier, one of the principal objectives of IDSAS was to develop baseline patterns for the

submissions to the system as well as specific diseases. One approach to thinking about data contributed by participatory disease sentinels, such as those providing data to IDSAS, is as the result of multiple data-generating processes. The observed data can be thought of as the result of two interdependent processes: the sentinel process itself, and the disease process. This is a fundamental recognition of the fact that people contributing data will vary according to a host factors that may impact how they submit data. Similar to how pharmaceutical sales data exhibit day-of-the-week variations that result from people's shopping tendencies, data in IDSAS contain non-disease related variations due to the participating veterinarians. This is true of all systems dependent on user-generated data. A two-step modelling procedure is developed to determine the influence of sentinel process covariates on data submissions, and then the effect of weather variables on disease. This is done using a Bayesian hidden markov modelling approach that yields region specific baseline estimates of prevalence for four commonly reported suspected diagnoses in the IDSAS system: Mastitis, Babesiosis, Ephemeral Fever, and Milk Fever. The Poisson hidden markov model developed in Chapter 5 is shown to provide a convenient and robust way to model novel surveillance data and may have utility in other types of volunteered geographic information systems (Goodchild 2007).

Chapter 6 demonstrates how geographical analysis can be incorporated into surveillance, and generates new knowledge on the epidemiology of an emerging disease, leptospirosis, of Sri Lankan (Agampodi et al. 2009), and international significance (Bharti et al. 2003). A large outbreak of suspected leptospirosis occurred in Sri Lanka in 2008. The number of notified cases increased to over 7000 in 2008, from historical average of 1000-2000. Chapter 6 considers the spatial epidemiology of this outbreak,

investigating the pattern of cases over all of Sri Lanka from 2005-2009. Relationships between risk and covariates are analyzed using regression, space-time scan statistics are employed to detect clusters of space and time, and a new approach for adjusting relative risk estimates with cluster analysis is developed. What emerges from this analysis is the changing epidemiology of a serious public health issue in Sri Lanka. During the endemic period, suspected cases of leptospirosis risk were associated with average distance to rivers and the proportion of small farms in each area. During the outbreak period, a positive association between outbreak locations and population density was detected. Rainfall analysis in four locations revealed a two-month lag between rainfall and notified leptospirosis cases, though this effect was not as evident during the outbreak. The analysis suggests a shift in transmission dynamics in 2008. The role of rainfall in the outbreak requires further investigation, though the relationship seems more complex than within-month correlation. The role of animals other than rats in the maintenance and transmission of Leptospires in Sri Lanka remains an area for further investigation.

The analysis in Chapter 6 provides evidence of change in the pattern of leptospirosis cases in Sri Lanka, and geographical factors that partially explain the change in pattern. Incorporating spatial risk factors within an analytical surveillance framework is demonstrated as feasible and informative for public health decision-making and guiding further studies. In the context of surveillance for emerging diseases, the analysis provides an example of what is possible when disparate data are integrated and geographic and temporal patterns are examined.

The key contributions of this dissertation are summarized in Chapter 7, along with a discussion of limitations of the research, and directions for future work. The

interdependencies and changing ecologies giving rise to emerging diseases are unlikely to abate. This dissertation explores how space-time disease surveillance in the developing world is situated to improve our understanding and ability to detect and respond to emerging diseases in the future.

Table 1.1 Challenges of emerging infectious disease surveillance and how they are addressed in this research.

Chapters	Key Challenges Addressed	Approach Taken
	Theoretical	
2	Building EID under resource constraints	Development of mobile-phone based surveillance system
	Methodological	
3, 4, 5	Developing a critical understanding space-time methods for surveillance data analysis Establishing normal patterns in novel surveillance data sources Determining disease sentinel effects in surveillance data	Review of methods based on contextual factors associated with implementation Review of available software Developing a hidden markov model for frontline disease sentinel data
	Applied	
5, 6	Understanding distribution and patterns of commonly reported cattle diagnoses Understanding the distribution of suspected leptospirosis cases in Sri Lanka	Modelling reported diagnoses and establishing region specific estimates of the average weekly number of cases Space-time cluster analysis and logistic regression modelling with spatial covariates

Chapter 2: Implementing Mobile Phone-Based Early Warning in Lower Resource Settings: Lessons learned from building infectious disease surveillance capacity in Sri Lanka

2.1 Abstract

With many emerging infectious diseases arising first in animals in low and middle income countries, surveillance of animal health in these areas may be important for forecasting emerging disease risks to humans. We present an overview of the implementation of a mobile-phone based frontline surveillance system developed in Sri Lanka. Field veterinary surgeons reported animal health information using mobile phones. Submissions increased steadily over nine months, with almost 4000 interactions between veterinarians and the animal population received by the system. Development of human resources and increased communication between local stakeholders were instrumental for successful implementation. The primary lesson taken from this experience is that mobile phone-based surveillance of animal populations is both acceptable and feasible in lower resource settings, however any system implementation plan must take into consideration the time it takes to garner support for novel surveillance methodologies amongst users and stakeholders.

2.2 Introduction

Emerging infectious diseases (EIDs) in animals and people are being identified more frequently than ever before, many in low income tropical countries, and this trend is expected to continue (Greger 2007). Approximately 75 percent of EIDs in people are estimated to have come from animals (Greger 2007), so there is much interest in the utility of animal health surveillance for prediction of human health risks (Rabinowitz et

al. 2008; Rabinowitz 2009; Halliday et al. 2007; Rabinowitz et al. 2005). The Canary Database, an online database named after the canary in the coalmine analogy, demonstrates the broad interest in this idea, containing over 1600 articles related to animal sentinels of zoonotic, environmental, and toxic effects on human health (Canary Database). In practice however, establishing links between animal and human health data has been difficult because data collected in animal and human health surveillance systems are collected at different resolutions, scales, and for different purposes. Human health surveillance is often based on aggregated diagnoses data obtained from standardized electronic medical records. Animal health surveillance systems vary widely (Doherr 2000). Where electronic veterinary records are kept, data can be extracted to central databases and analyzed. However, in lower resource settings electronic recording of veterinary services is often not feasible.

In many human health projects in resource challenged areas, mobile technologies have emerged as a promising solution for collecting, transmitting and analyzing human health information in a timely fashion (Bernabe-Ortiz et al. 2008; Missinou et al. 2005; Diero et al. 2006; Shirima et al. 2007). In Peru, a mobile phone-based surveillance system has been used for early detection of infectious disease outbreaks in the Peruvian Navy (Stoto et al. 2008). In Africa, the Satellife project has been employing mobile data collection devices for over two decades in human health surveys, and currently a project is underway using mobile phones and wireless technology in disease surveillance in Uganda (Mobile Active 2008). Many UN health and development projects in Africa now employ mobile phones for collection of field data (Vital Wave Consulting 2009).

However, the authors are not aware of any examples of mobile phone-based disease surveillance to support an animal-based EID system in the developing world.

In response to these challenges we have developed the Infectious Disease Surveillance and Analysis System (IDSAS), a mobile phone-based surveillance system targeted at animal populations in lower resource settings. A pilot version of this system was implemented in January 2009 in partnership with the Department of Animal Production and Health (DAPH) in Sri Lanka. The objective of this system is to collect animal health information from field veterinary surgeons (FVSs) in a timely fashion in order to establish baseline patterns in animal health. By establishing baseline patterns in animal health conditions via regular electronic surveillance, we aim to build capacity to detect changes that may facilitate early detection of changing EID risks. Here we describe design and implementation of the system, present preliminary data on submission patterns, provide examples of some of the data that is being collected, and discuss obstacles and opportunities encountered during the first nine months of operation. The objective of this paper is to highlight and generalize some of the lessons learned during the planning and implementation of IDSAS in Sri Lanka.

2.3 Material and Methods

2.3.1 Delivery of veterinary services in Sri Lanka

The provision of veterinary services in Sri Lanka is largely carried out by the DAPH, a national-level body responsible for control of livestock diseases, livestock research, animal breeding, and education in animal husbandry. Delivery of veterinary services is implemented through provincial level DAPH councils and field offices. Provinces are made up of districts, which are further divided into divisional secretariat

(DS) divisions. Each DS division is assigned a FVS who is responsible for providing animal health services within that division.

2.3.2 System structure

Forty FVSs were recruited to pilot IDSAS in four districts in separate provinces. The districts (Nuwara Eliya, Anuradhapura, Matara, Ratnapura) were selected to capture variation in livestock practices, climate, and environment (Figure 2.1).

Capacity for electronic collection and submission of data was developed in IDSAS to decrease the existing time from detection to reporting of animal health events as compared to the existing method using mailed written reports. Internet access is limited in many parts of Sri Lanka but the cellular phone network is extensive. Mobile phones, namely Palm Centro smartphones, were used as the data collection platform. Animal health surveys were developed using EpiSurveyor, a free and open-source software package developed for collection of public health data (www.datadyne.org). EpiSurveyor has been used extensively for human health data collection in Africa.

Surveys could be filled out in remote areas without cellular service and transmitted when the user was back in an area of reception. Decoupling data collection from transmission-capable locations greatly expanded the geographical range of the surveillance system. The location of each survey was also collected with global positioning system (GPS) software and an external receiver connected to the phone via Bluetooth. FVSs collected data throughout the course of their daily working activities (clinic and farm visits). Survey and GPS data were encoded and transmitted to a central database via email at the end of each day. A schematic overview of IDSAS is presented in Figure 2.2.

2.3.3 Information structure

The pilot study was restricted to cases in chickens, cattle, and buffalo. Every time a FVS visited a farm or saw a case in clinic involving one of these species they completed a survey within EpiSurveyor and recorded the location (for farm visits). While we aimed for daily submissions, our minimum target submission rate was two surveys, per FVS, per week. This was based on an estimate of the number of cases in chickens, cattle and buffalo seen on average by individual FVSs and work-related disruptions that could interfere with data submission (training, sick days, holidays etc.).

The first draft of the survey was based on the Alberta Veterinary Surveillance Network's Veterinary Practice Surveillance initiative (Government of Alberta 2010). In the second stage, the survey was reviewed with a number of FVSs and government employees within the DAPH to ensure it was applicable to veterinary practice in Sri Lanka. The majority of questions were single answer, multiple choice-type questions, though additional comments were allowed in a free-text field. The survey was designed to minimize the time required to fill out each survey, reduce the number of data entry errors, and permit simpler and automated data analysis.

Data for each case included: date, location, type of operation, nature of visit (routine/non-routine), age and sex of affected animals, number on farm, number affected, clinical syndrome, clinical diagnosis, laboratory testing if applicable, and other species on the premises. A survey could contain up to three cases if all three species were present on a farm. FVSs selected from clinical syndromes outlined in Table 2.1. Within EpiSurveyor each syndromic grouping was linked to a list of clinical diagnoses.

2.3.4 Reporting and Data Analysis

Data reported here represent the experience of IDSAS from January 1, 2009 through September 30, 2009. Weekly surveillance reports were disseminated to project partners containing a list of cases. These reports documented the following details pertaining to each case submitted during the previous week: date, species, reported syndrome, suspected clinical diagnosis, number of animals affected, number of animals on farm, number of dead animals, and a flag indicating whether samples were submitted to a laboratory.

2.3.4.1 Data completeness and submission patterns

Measures of data completeness used for IDSAS at the planning and early implementation stages follow the guidelines set out by Lescano et al (2008). In the planning stage it is important to assess the burden placed on data collectors to determine if data can be collected with existing resources. The IDSAS data collection procedure involved separate software programs for animal health surveys and GPS data collection. These data were linked via a common identifier entered by FVSs at the time of survey completion. To explore the linkage between survey and GPS data, we report completeness for surveys, GPS points, and linked survey-GPS records. We also report the percentage of surveys with a linked GPS point. As FVSs work six days of the week, we expect a day-of-the-week effect and therefore examine variation in survey submission by day of the week. Finally, we examine weekly submission counts to determine temporal patterns. We fit a linear trend model to the weekly counts to determine the average change in submissions per week.

2.3.4.2 Statistical Surveillance

Digital storage of data that otherwise might not be captured allows more sophisticated statistical analysis. To demonstrate how the IDSAS database could be used in an outbreak detection context, we present an example of statistical surveillance using the total number of weekly surveys submitted by participating FVSs as an indicator for unusual animal health events. We use these data in a prospective temporal surveillance cumulative sum (CUSUM) statistic implemented in the statistical software package R (Höhle 2007). The CUSUM measures accumulations of extra variance in a sequential framework, and alarms are signaled when the statistic exceeds a specified threshold. Parameters are required for the expected value, the reference value *k*, and the alarm threshold *h*. We estimated values for *k* and *h* based on an expected false positive rate of one every 52 weeks, to detect a change two standard deviations above the reference value. We evaluated two baseline scenarios: the mean of the first 14-week period, and a set value of 100 surveys per week. Analysis was carried out weekly beginning at week 14 until the end of the study period.

2.3.4.3 Caseload and case profile

The distribution of cases seen is presented broken down by species and district. We also present the frequency of the five most commonly reported syndromes for each species.

2.3.4.3 Assessing system implementation

The experience of implementing IDSAS provides lessons for future surveillance projects in lower resource settings. We synthesize some of the key lessons learned during this phase of IDSAS based on technical, financial, political, and ethical/societal/cultural considerations (Chretien et al. 2008).

2.4 Results

2.4.1 Data completeness and submission patterns

IDSAS was operational for 273 days. During this period, 3981 unique surveys were submitted to the system by participating FVSs. This corresponds to approximately 99 surveys per FVS over a 9-month period (11 per month), above our intended submission target minimum of 2 submissions per FVS per week. During this period, 96% of days had at least one conducted survey. The total number of unique GPS points submitted was 1650. Of these, 1172 (71%) were linked to an associated survey. Of the total days under surveillance, 76% had GPS data collected, and 64% had both GPS and survey data recorded. Informal discussions with many FVSs revealed that it took about one minute to complete an animal health survey, and one minute to collect a GPS point once IDSAS had been in place for 6 months.

Temporal patterns in submissions are presented in Figure 2.3. In general, there was an increasing overall trend. The linear trend model revealed a significant weekly increase in submissions of 1.65% ($p < 0.001$, $R^2 = 0.31$). The trend was also characterized by large variation (coefficient of variation = 3.01), with a large drop (39 surveys) in submissions in week 14. Day-of-the-week variation was present in submissions as expected, with weekly survey counts totaling 306 on Saturdays and 326 on Sundays, while during the week totals ranged from 515 to 695.

2.4.2 Statistical surveillance

Based on parameters described above, reference value *k* was estimated at 2.6 and the threshold value *h* was 4.1. Using week 14 as a baseline, 84 weekly visits were expected, which in the CUSUM analysis flagged an alarm at week 26 and weeks 30 through to the

end of the study period (week 38). Using the expected value of 100 weekly visits, alarms were signalled from weeks 31 through 38.

2.4.3 Caseload and case profile

Out of 3981 surveys submitted during the 9 months of operation, 3150 cases were reported (i.e., reported an animal health issue). The majority (83%) of cases were seen in cattle, followed by chickens and buffalo (Table 2.2). These were mostly from an area known to contain a large number of cattle dairy operations. Production-related syndromes were the most commonly reported across all species, with decreased feed intake/milk production most prevalent in cattle and buffalo, and decreased egg production/weight gain/appetite in chickens (Figure 2.4). In buffalo, markedly higher gastrointestinal and lameness submissions were noted relative to other syndrome groupings. Gastrointestinal signs were common in Anuradhapura across all species. Cases in chickens were found predominantly in Ratnapura, where there is a large number of poultry operations. The syndrome profiles for chickens were similar across all districts (Figure 2.4).

2.4.4 Alerts identified by IDSAS

There was one instance in which suspected cases of 'Black quarter' (*Clostridium chauvoei)* were identified at the time of review of the weekly report. As the DAPH was made aware of the cases shortly after they were identified by the FVS they were able to confirm that the FVS collected tissue samples for diagnostic testing. This increased information flow would not have been possible under the DAPH surveillance program as written reports of suspected cases from FVSs are received on a monthly basis and each must be reviewed individually to identify suspected cases of a particular disease of interest. Additional statistical alerts generated by analysis could be evaluated, as part of

the objective of IDSAS is to establish the baseline caseload burden in areas under surveillance.

2.4.5 Assessing system implementation (Table 2.3)

2.4.5.1 Technical considerations

Technical barriers were a major challenge during implementation of IDSAS. The system introduced new data collection requirements for FVSs. Using cell phones for data collection required training and ongoing technical support.

2.4.5.2 Financial considerations

The main costs of the system were associated with data collection hardware. Each phone and GPS extension set cost approximately 500 CAD. This cost may have been reduced if phones were available for purchase locally. Proprietary software options with different hardware requirements were available but rejected as recurring licensing costs could not be sustained while hardware was a one-time expense. Though data plans are an ongoing cost, the size of files generated by IDSAS is typically less than one kilobyte. The cost of data transmission per user per month in Sri Lanka is less than five dollars CAD. Investments in hardware and human resources for data collection can be quickly recouped as these resources are extendible to many other fields in which the Sri Lankan government is involved (e.g., human epidemiology, environmental assessment, disaster planning).

2.4.5.3 Political considerations

Political support has been the most important factor in the successful implementation and operation of IDSAS. Animal health reporting standards set by the World Animal Health Organization (OIE) require member countries to report on a suite of animal diseases. The

introduction of a new surveillance system as part of a research project resulted in initial confusion about how such a system could fit within existing surveillance networks. A major challenge in the implementation of IDSAS was drawing the distinction between IDSAS as a research project and the national animal disease reporting system of the DAPH. Negotiating this challenge was possible with support from key figures in the government and the University of Peradeniya.

2.4.5.4 Ethical, societal, and cultural considerations

During the design and early implementation of IDSAS concerns around privacy and data security were addressed promptly as they arose. No information pertaining to animal owners was collected. No personal identifiers from FVSs were linked to survey submissions.

2.5 Discussion

IDSAS has been developed based on the premise that monitoring animal health can provide information for early warning of EIDs and changing disease patterns. Preliminary results presented here demonstrate significant enhancement of existing technological infrastructure. Equipping FVSs with the necessary means of communication enables timely case submission, and the skills to make use of these tools has helped to build further capacity in animal health surveillance. Weekly reports document increased knowledge and information flow between Sri Lankan animal health stakeholders. Finally, through IDSAS significant progress has been made toward establishing baseline patterns of suspected diagnoses and syndromes in cattle, buffalo, and chickens.

Uptake of IDSAS over its initial 9 months of operation resulted in data generation on almost 4000 interactions between FVSs and the animal population. Increasing use of

IDSAS over time is also illustrated by a positive linear trend in submissions. Statistical surveillance of the number of surveys submitted by FVSs revealed that an upward shift in submissions occurred around week 30. The overall trend is likely due to FVSs gaining competency with the technology while the shift is likely due to a combination of reduced number of submissions in weeks 14-16 related to training and examinations and the final stages of the civil war in weeks 19-21, followed by retraining in week 23. The alarms signaled by the CUSUM analysis illustrate the importance of modeling the expected value when using surveillance statistics.

The distribution of cases highlights one of the challenges with this type of data, and indeed many types of surveillance data, and that is how to interpret variability in cases in the absence of data on the population at risk. The high number of cattle cases in Nuwara Eliya was expected given prior knowledge of the large number of milk-producing cattle in that region. Yet the distribution of cases would only be expected to reflect the true disease burden in the population if the likelihood of a veterinarian seeing a case in a given species were proportional to the underlying disease distribution in the 3 species in each area. For example, in Nuwara Eliya, cattle raisers might be more inclined to call their veterinarian in the event of a sick cow compared to a sick chicken. The solution to this problem, if the aim is to establish a predictive, prospective disease surveillance system, is establishing normal patterns of case submission for the population. For this to be realized this system (and others) must be maintained over a period of time within the same geographical areas.

One of the barriers to implementation of IDSAS in its current form is the cost of hardware and the need for a server administrator. However, since the pilot project in Sri

Lanka a new version of EpiSurveyor has been released. A number of important changes have been made: the software now runs on a wide range of standard mobile phones; data can be uploaded to servers administered by datadyne.org as well as analyzed on the phones themselves; and GPS data can be collected within Episurveyor. These changes drastically reduce the costs of implementing mobile surveillance: the cost per mobile phone unit reduces substantially and there is no need for governments to purchase and administer their own database.

At this time the DAPH has decided to incorporate IDSAS into its ongoing disease surveillance efforts and the system is being run on two parallel servers, one at the DAPH and the original server that hosts IDSAS. After this transition period the system will continue to run only on the DAPH server and may be modified to suit additional surveillance priorities (e.g., goats, swine). The DAPH will not be providing incentives to FVSs for participation. It would be valuable to solicit further FVS review once the system has been transitioned, and to monitor submissions long term.

Beyond the data collected by IDSAS to date, this research demonstrates that, through developing social capital and technological capacity, novel surveillance methods can be implemented that are feasible and acceptable in lower resource settings. These considerations are supplemented with lessons for planning and implementation of surveillance systems. It is hoped that by disseminating the results of this initiative other governments will be able to tailor IDSAS to their particular animal health surveillance needs. The collaboration and relationships established in this project should yield further benefits through technical training and pooling of human and physical resources for sustaining and promoting veterinary public health in Sri Lanka. Additionally, the

advantages of electronic health surveillance using mobile data collection afforded by IDSAS are immediately known to important administrative figures that can affect change in other areas of animal and human health policy and planning.

2.6 Conclusions

Developing surveillance capacity in Sri Lanka has generated valuable human resources and relationships that, when coupled with technology, may be the key to early detection of EIDs. FVSs are developing a valuable technological skill set for remote data collection. The data collected from IDSAS offers DAPH stakeholders and FVSs a new perspective on disease within the animal population, creating new opportunities for dialogue and mutual understanding. Increased communication, through training, surveillance reporting, and regular meetings, has been an important aspect of improving veterinary public health awareness and is a key result of the IDSAS project. Social capital, though difficult to measure, is an important precursor to successful surveillance in the developing world (Ndiaye 2003). It is important to note that it takes a significant amount of time to build social capital under any circumstances. Future development of similar surveillance programs should take this temporal component of project development into consideration, helping to ensure that new initiatives gain momentum over time. FVSs have indispensable local knowledge about animals in their division. Leveraging this awareness via regular electronic surveillance is a first step towards formalizing this knowledge store to improve surveillance for EIDs in Sri Lanka.

2.7 Acknowledgements

This project was funded in part by the Teasdale-Corti Global Health Partnership and the National Sciences and Engineering Research Council of Canada. IDSAS represents the

culmination of a collaborative effort between stakeholders at the DAPH, the University of Peradeniya Faculty of Veterinary Science, and provincial levels of government in Sri Lanka. We would like to thank Drs. Swarnalatha Podimanike Herath (Director General of the DAPH), Hitihami Mudiyanselage Amarasiri Chandrasoma (Director in Animal Health at the DAPH), Jeewaranga Dharmawardana (former Director of the Veterinary Research Institute), Ravi Bandara, and the provincial directors for their invaluable efforts. We would also like to thank Drs. Preeni Abanayake and Indra Shiyamila Abegunawardana for their input and assistance. We would like to draw particular attention to the efforts of Dr. Walimunige Suraj Niroshan Gunawardana, the research assistant involved in development and implementation of IDSAS, whose ongoing efforts have been invaluable in maintaining support and enthusiasm for the project. Lastly we would like to thank all of the participating FVSs for their ongoing submissions, input, and patience during the implementation process.

Table 2.1. Syndrome groupings used in animal health surveys in the Infectious Disease Surveillance and Analysis System.

Species	Syndrome groupings
Buffalo and Cattle	• abortion/birth defect • ambulatory lameness • decreased feed intake/milk production • gastrointestinal signs • neurological signs • recumbency • peripheral edema/misc swelling • reproduction/obstetrics problems • respiratory • skin/ocular/mammary • sudden or unexplained death • urologic • vesicular/ulcerative • other
Poultry	• ambulatory • decreased egg production/decreased weight gain/decreased appetite • neurological/recumbent • peripheral edema/misc swelling • respiratory • skin/ocular • sudden or unexplained death • other

Table 2.2. Total number of cases in cattle, buffalo, and chickens in each of the four study districts covered by the Infectious Disease Surveillance and Analysis System from January 1, 2009 to September 30, 2009.

District	Cattle cases	Buffalo cases	Chickens cases	Total
Ratnapura	548	106	146	800
Matara	388	62	55	505
Nuwara Eliya	1095	16	11	1122
Anuradhapura	596	70	57	723
Total	2627	254	269	3150

Table 2.3. Lessons learned for planning and implementing surveillance systems in settings.

Considerations for surveillance in lower resource settings	IDSAS Experience	Generalized lessons
Technical	Cell phones permitted timely collection and transmission of data to the surveillance system. Touch screen interfaces were new technology for FVSs.	Utilizing familiar technologies such as basic cell phones will minimize training time. Cell phones enable timely data collection and transmission.
	Ongoing training was essential. A local research assistant made training more effective, in particular because FVSs could learn the system in their first language.	Developing local expertise at the project outset is invaluable for ensuring sustained technical and logistical support.
Financial	The hardware required for data collection was relatively low-cost, but much higher compared to hardware available in Sri Lanka. Importing cell phones for the project was challenging.	Where possible use hardware that is locally available.
	Open-source software was used where possible, eliminating licensing as a recurring cost but requiring more training and technical skills to maintain.	Open-source software options should be selected over proprietary options to reduce costs and generate technological capacity.
Political	External funding covered the initial hardware and software costs.	Obtaining external financial support to cover the initial investment required will make implementation more feasible.
	Support at the provincial level was critical for engagement of FVSs.	Garnering support at all levels of government is critical at the early implementation phase.
	Engagement of key political stakeholders was essential to alleviate fears around the potential for harm of novel types of surveillance data.	Early in the design process it is important to discern what the outputs of the system will be and their added value.
Ethical, societal and cultural	Government officials were initially concerned about data security.	Build appropriate data security into all components of the system.
	It was late in the implementation phase when government stakeholders recognized the potential for additional data uses.	Examples of additional uses of data collected will generate support for new surveillance initiatives.
	At the onset of the project, FVSs were skeptical about the usefulness of data generated by IDSAS but over time envisaged not only how the outputs could be used in disease surveillance but in informing their daily veterinary duties	Adoption of novel surveillance methodology requires user acceptance in addition to new technical skills. Time and experience will allow this transition to occur.
	Many farms are geographically isolated making access to FVSs difficult.	The quality and quantity of data from surveillance systems is impacted by the ability of an animal owner to access animal health services.

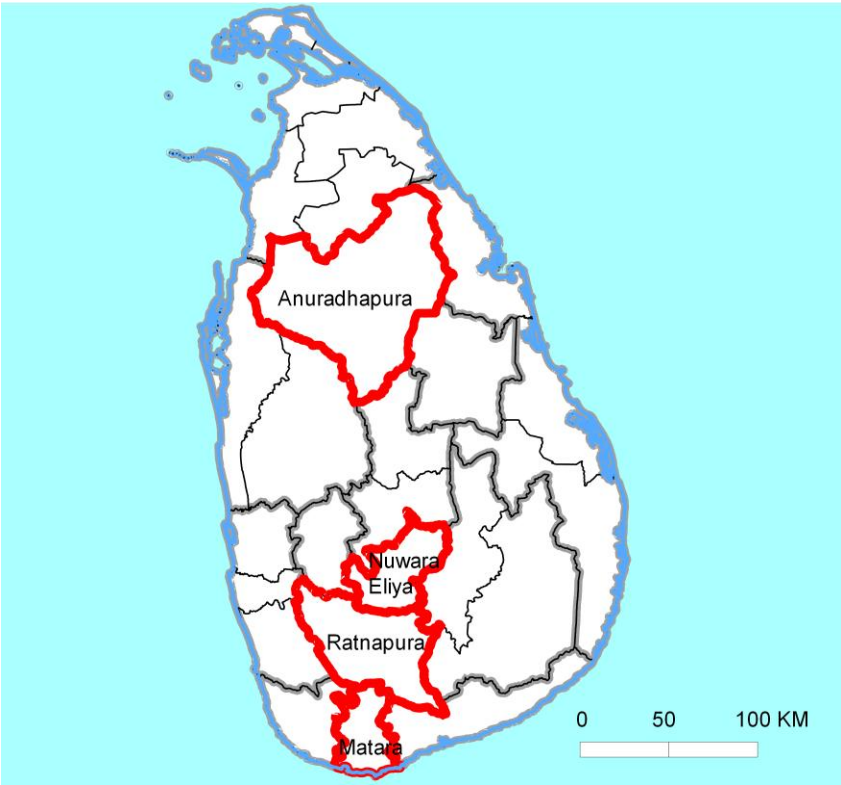


Figure 2.1 Study districts (red) where field veterinarians participating in the Infectious Disease Surveillance and Analysis System collect data on animal health seen during their daily working activities.

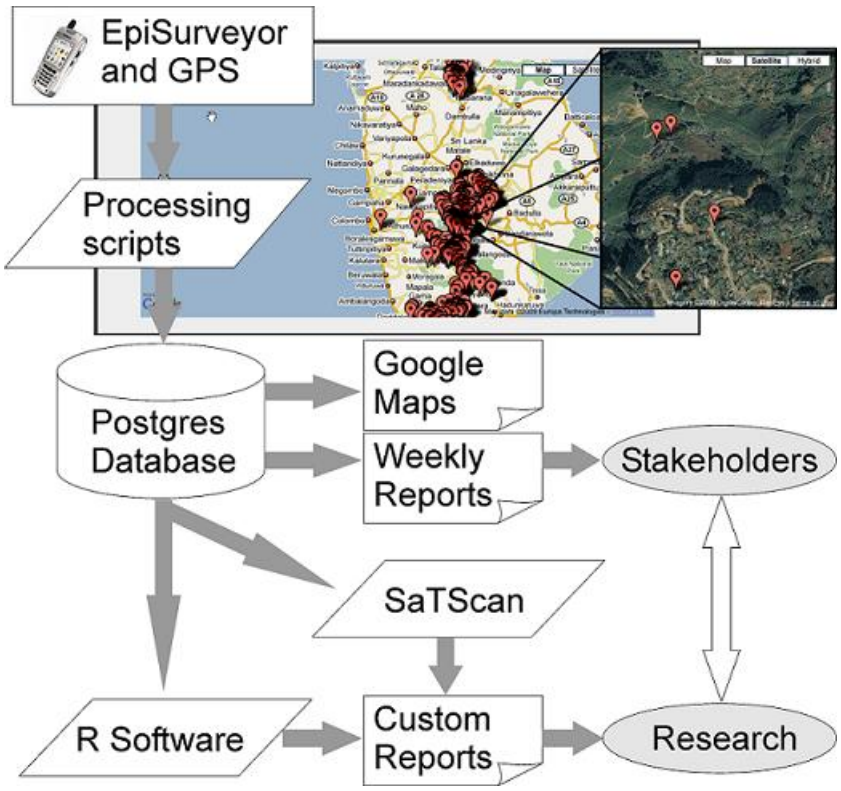


Figure 2.2 Schematic overview of the major components of the Infectious Disease Surveillance and Analysis System.

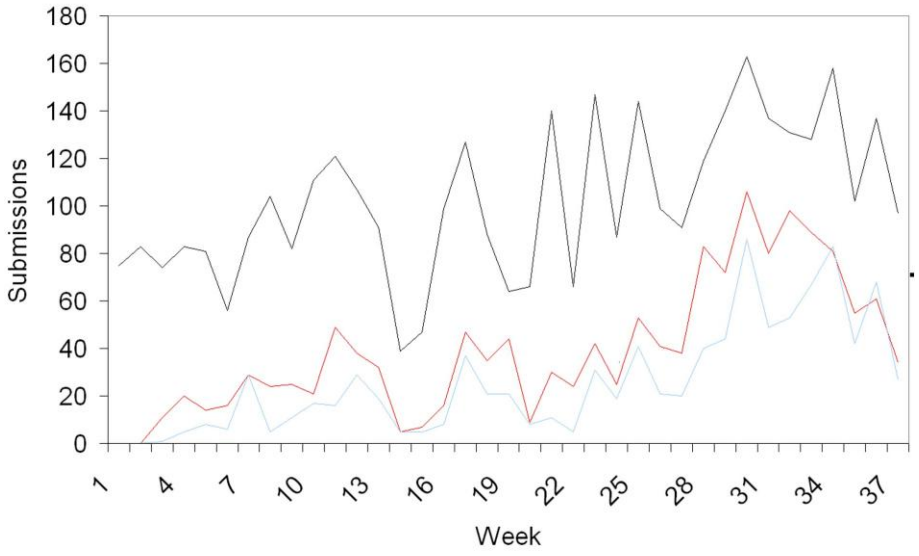
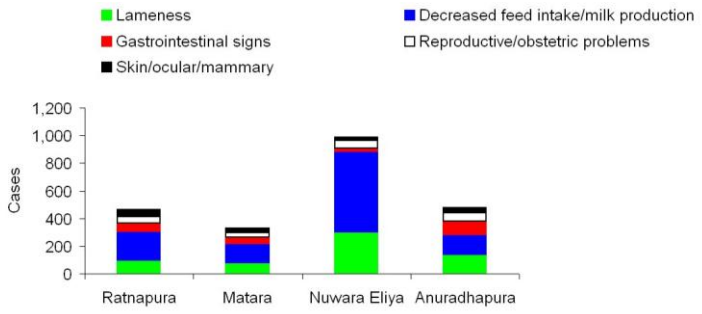
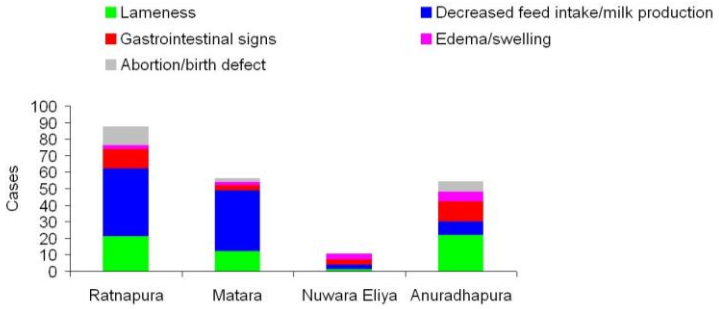


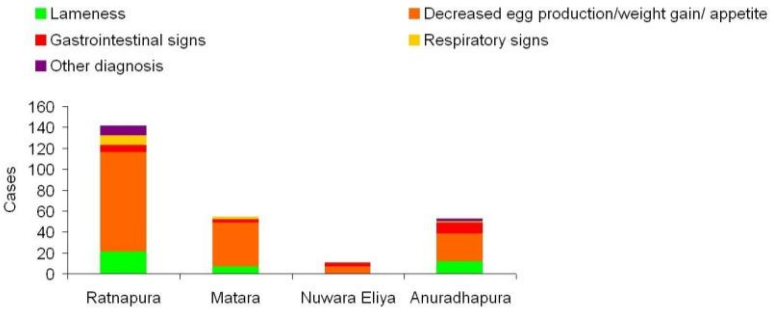
Figure 2.3 Number of surveys (black), GPS points (red) and linked survey-GPS (blue) submissions to Infectious Disease Surveillance and Analysis System from January 1, 2009 to September 30, 2009.



a)



b)



c)

Figure 2.4 Frequency of syndrome groups seen by field veterinarians in (a) cattle, (b) buffalo, and (c) chickens in each of the four study districts part of the Infectious Disease Surveillance and Analysis System from January 1, 2009 to September 30, 2009.

Chapter 3: Review of methods for space-time disease surveillance

3.1 Abstract

A review of some methods for analysis of space-time disease surveillance data is presented. Increasingly, surveillance systems are capturing spatial and temporal data on disease and health outcomes in a variety of public health contexts. A vast and growing suite of methods exists for detection of outbreaks and trends in surveillance data and the selection of appropriate methods in a given surveillance context is not always clear. While most reviews of methods focus on algorithm performance, in practice, a variety of factors determine what methods are appropriate for surveillance. In this review, we focus on the role of contextual factors such as scale, scope, surveillance objective, disease characteristics, and technical issues in relation to commonly used approaches to surveillance. Methods are classified as testing-based or model-based approaches. Reviewing methods in the context of factors other than algorithm performance highlights important aspects of implementing and selecting appropriate disease surveillance methods.

3.2 Introduction

Early detection of unusual health events can enable coordinated response and control activities such as travel restrictions, movement bans on animals, and distribution of prophylactics to susceptible members of the population. The experience with Severe Acute Respiratory Syndrome (SARS), which emerged in southern China in late 2002 and spread to over 30 countries in 8 months, indicates the importance of early detection (Banos and Lacasa 2007). Disease surveillance is the principal tool used by the public

health community to understand and manage the spread of diseases, and is defined by the World Health Organization as the ongoing systematic collection, collation, analysis and interpretation of data and dissemination of information in order for action to be taken (World Health Organization 2007). Surveillance systems serve a variety of public health functions (e.g., outbreak detection, control planning) by integrating data representing human and/or animal health with statistical methods (Diggle et al. 2003), visualization tools (Moore et al. 2008), and increasingly, linkage with other geographic datasets within a GIS (Odiit et al. 2006).

Surveillance systems can be designed to meet a number of public health objectives and each system has different requirements in terms of data, methodology and implementation. Outbreak detection is the intended function of many surveillance systems. In syndromic surveillance systems, early-warning signals are provided by analysis of pre-diagnostic data that may be indicative of people's care-seeking behaviour during the early stages of an outbreak. In contrast, systems designed to monitor food and water-borne (e.g., cholera) pathogens are designed for case detection, where one case may trigger a response from public health workers. Similarly, where eradication of a disease in an area is a public health objective, surveillance may be designed primarily for case detection. Alternatively, where a target disease is endemic to an area, perhaps with seasonal variation in incidence, such as rabies, monitoring space-time trends may be the primary surveillance objective (Childs et al. 2000).

Surveillance systems differ with respect to a number of qualities which we term *contextual factors*. For evaluation of surveillance systems, this is well known, as the evaluative framework set out by the Centre for Disease Control and Prevention (CDC)

encompasses assessment of simplicity, flexibility, data quality, acceptability, sensitivity, predictive value positive, representativeness, timeliness, and stability (Buehler et al. 2004). Selection of appropriate methods for space-time disease surveillance should consider system-specific factors indicative of the context under which they will be used (Table 1). These factors serve as the axes along which we will review methods for space-time disease surveillance.

There has been rapid expansion in the development of automated disease surveillance systems. Following the 2001 bioterrorism attacks in the United States, there was expanded interest and funding for the development of electronic surveillance networks capable of detecting a bioterrorist attack. Many of these were designed to monitor data that precede diagnoses of a disease (i.e., syndromic surveillance). By May 2003 there were an estimated 100 syndromic surveillance systems in development throughout the U.S. (Buehler et al. 2003). Due to the noisy nature of syndromic data, these systems rely heavily on advanced statistical methods for anomaly detection. As data being monitored in syndromic systems precede diagnoses they contain a signal that is further removed from the pathogen than traditional disease surveillance, so in addition to having potential for early warning, there is also greater risk of false alarms (i.e., mistakenly signalling an outbreak) (Stoto et al. 2004).

One example is a national surveillance system called BioSense developed by the CDC in the United States. BioSense is designed to support early detection and situational awareness for bioterrorism attacks and other events of public health concern (Bradley et al. 2005). Data sources used in BioSense include Veterinary Affairs and Department of Defense facilities, private hospitals, national laboratories, and state surveillance and

healthcare systems. The broad mandate and national scope of the system necessitated the use of general statistical methods insensitive to widely varying types, quality, consistency and volume of data. Two methods used in BioSense are a generalized linear mixed-model which estimates counts of syndrome cases based on location, day of the week and effects due to seasonal variation and holidays. Counts are estimated weekly for each syndrome-location combination. A second temporal surveillance approach computed for each syndrome under surveillance is a cumulative sum of counts where events are flagged as unusual if the observed count is two standard deviations above the moving average. The selection of surveillance methods in BioSense considered factors associated with heterogeneity of data sources and data volume among others.

Another example is provided by a state-level disease surveillance system developed for Massachusetts called the Automated Epidemiological Geotemporal Integrated Surveillance (AEGIS) system, where both time-series modelling and spatial and space-time scan statistics are used (Reis et al. 2007). The modular design of the system allowed for 'plug-in' capacity so that functionality already implemented in other software (i.e., SaTScan) could be leveraged. In AEGIS, daily visit data from 12 emergency department facilities are collected and analyzed. The reduced data volume and greater standardization enable more advanced space-time methods to be used as well as tighter integration with the system's communication and alerting functions (Reis et al. 2007).

Decisions on method selection and utilization are based on a variety of factors, yet most reviews of statistical methods for surveillance data compare and describe algorithms from a purely statistical or computational perspective (e.g., Buckeridge et al. 2005;

Sonesson and Bock 2003; Yan et al. 2006). The selection of statistical approaches to surveillance for implementation as part of a national surveillance system is greatly impacted by design constraints due to scalability, data quality and data volume whereas the use of surveillance data for a standalone analysis by a local public health worker may be more impacted by software availability, learning curve, and interpretability. Selection of appropriate statistical methods is key to enabling a surveillance system to meet its objectives.

A frequently cited concern of surveillance systems is how to evaluate whether they are meeting their objectives (Reingold 2003; Sosin and DeThomasis 2004). A framework for evaluation developed by the CDC considers outbreak detection a function of timeliness, validity, and data quality (Buehler et al. 2004). The degree to which these factors contribute to system effectiveness may vary for different surveillance systems, especially where objectives and system experiences differ. For example, newly developed systems in developing countries may place a greater emphasis on evaluating data quality and representativeness, as little is known about the features of the data streams at early stages of implementation (Lescanso et al. 2008). Algorithm performance is usually measured by sensitivity, specificity and timeliness. Sensitivity is the probability of an alarm given an outbreak, and specificity is the probability of no alarm when there is no outbreak. Timeliness is measured in number of time units to detection, and has been a focus of systems developed for early outbreak detection (Wagner et al. 2001). The importance of each of these measures of performance need to be evaluated in the light of the system's *contextual factors* outlined in Table 1.

Our goal in this review of approaches to space-time disease surveillance is to synthesize major surveillance methods in a way that will focus on the feasibility of implementation and highlight contrasts between different methods. First, we aim to place methods in the context of some key aspects of practical implementation. Second, we aim to highlight how methods of space-time disease surveillance relate to different surveillance contexts. Disease surveillance serves a number of public health functions under varying scenarios and methods need to be tailored and suited to particular contexts. Finally, we provide guidance to public health practitioners in understanding methods of space-time disease surveillance. We limit our focus to methods that use data encoded with both spatial and temporal information.

This paper is organized as follows. The next section describes different statistical approaches to space-time disease surveillance with respect to the contextual factors outlined in Table 1. Methods are categorized as statistical tests, models, and emerging research areas which includes mostly new and experimental approaches. We conclude with a summary and brief discussion of our review.

3.3 Space-Time Disease Surveillance Methods

Methods for space-time disease surveillance can address a surveillance objective in a variety of ways. Most methods assume a study area made up of smaller, non-overlapping sub-regions where cases of disease are being monitored. The variable under surveillance is the count of the number of cases. In retrospective analysis, the data are fixed and methods are used to determine whether an outbreak occurred during the study period, or characterize the spatial-temporal trends in disease over the course of the study period (Marshall 1991). In the prospective scenario, the objective is to determine whether any

single sub-region or collection of sub-regions is undergoing an outbreak (currently), and analysis occurs in an automated, sequential fashion as data accumulate over time. Prospective methods require special consideration as data do not form a fixed sample from which to make inferences (Sonneson and Bock 2003). Parallel surveillance methodologies compute a test statistic separately for each sub-region and signal an alarm if any of sub-regions are significantly anomalous (Figure 3.1a). While in vector accumulation methods, test statistics in a parallel surveillance setting are combined to form one general alarm statistic (Figure 3.1b). Conversely, a scalar accumulation approach computes one statistic over all sub-regions for each time period (Frisen and Sonesson 2005) (Figure 3.1c). For example, Rogerson (1997) used the Tango (1995) statistic to monitor changes in spatial point patterns.

3.3.1 Statistical tests

Statistical tests in space-time disease surveillance generally seek to determine whether disease incidence in a spatially and temporally defined subset is unusual compared to the incidence in the study region as a whole. Thus, this class of methods is designed to detect clusters of disease in space and time, and suit surveillance systems designed for outbreak detection. Most spatial cluster detection methods such as the Geographical Analysis Machine (Openshaw et al. 1987), density estimation (Blithell 1989; Lawson and Williams 1993), Turnbull's method (Turnbull et al.1990), the Besag and Newell (1991) test, spatial autocorrelation methods such as the Gi* (Getis and Ord 1992), and LISAs (Anselin 1995), and the spatial scan statistic (Kulldorff and Nagarwalla 1995) are types of statistical tests. The development of methods for space-time cluster detection naturally evolved from these purely spatial methods. We can stratify methods in the statistical test

class into three types: tests for space-time interaction, cumulative sum methods, and scan statistics.

3.3.1.1 Tests for space-time interaction

Space-time interaction of disease indicates that the cases cluster such that nearby cases in space occur at about the same time. The form of the null hypotheses is usually conditioned on population, and can factor in risk covariates such as age, occupation, and ethnicity. Detecting the presence of space-time interaction can be a step towards determining characteristics of infection processes for new or poorly understood diseases (Aldstadt 2007). Additionally, non-infectious diseases exhibiting space-time interaction may suggest the presence of an additional causative agent, such as a point source of contamination and/or pollution or an underlying environmental variable. These methods require fixed samples of space-time data representing cases of disease.

All tests for space-time interaction consider the number of cases of disease that are related in space-time, and compare this to an expectation under a null hypothesis of no interaction (Kulldorff and Hjalmars 1999). The Knox test (1964) uses a simple test statistic which is the number of case pairs close both in space and in time. This count is compared to the null expectation conditional on the number of pairs close only in space, and the number of pairs close only in time; i.e. the times of occurrence of the cases are independent of case location. A major shortcoming of the Knox (1964) method is that the definition of "closeness" is arbitrary. Mantel's (1967) test addresses this by summing across all possible space-time pairs, while Diggle et al. (1995) identify clustering at discrete distance bands in the space-time K function. For infectious diseases, it is likely that near space-time pairs are of greater importance, so Mantel suggests a reciprocal

transformation such that distant pairs are weighted less than near pairs. The Mantel test can in fact be used to test for association between any two distance matrices, and is often used by ecologists to test for interaction between space and another distance variable such as genetic similarity (Legendre and Fortin 1989).

The reciprocal transformation used in the Mantel statistics assumes a distance decay effect. While this may be appropriate for infectious diseases, for non-infectious diseases or diseases about which little is known, this assumed functional form of disease clustering may be inappropriate. A different approach is taken by Jacquez (1996) where relations in space and time are defined by a nearest neighbour relation rather than Euclidean distance. Here, the test statistic is defined by the number of case pairs that are *k* nearest neighbours in both space and time. When space-time interaction is present, the test statistic is large. Another method for testing an infectious aetiology hypothesis given by Pike and Smith (1974), assesses clustering of cases relative to another control disease, though selection of appropriate controls can be difficult.

The scale of the disease surveillance context can impact the selection of space-time interaction tests because these tests are sensitive to changes in the underlying population at risk (population shift bias). Therefore, large temporal scales will be more likely to exhibit changes in population structure and introduce population shift bias. An unbiased version of the Knox test given by Kulldorff and Hjalmars (1999) accounts for this by adjusting the statistic by the space-time interaction inherent in the background population. Changes in background population over time can be incorporated into all space-time interaction tests using a significance test based on permutations conditioned

on population changes. However, this obviously requires data on the population over time which may not always be easy to obtain.

Space-time interaction tests are univariate and therefore only suitable for testing cases of a single disease. Consideration of multiple host diseases is possible, though there is no mechanism to test for interaction or relationships between different host species. Another major consideration is the function of the surveillance system or analytic objective. Interaction tests can only report the presence or absence of space-time interaction. They give no information about the spatial and temporal trends in cases, nor consider naturally occurring background heterogeneity. A final point is that these tests use case data, and therefore require geo-coded singular event data, making these methods unsuitable when disease data are aggregated to administrative units.

3.3.1.2 Cumulative sum (CUSUM) methods

Cumulative sum methods for space-time surveillance developed out of traditional statistical surveillance applications such as quality control monitoring of industrial manufacturing processes. In CUSUM analysis, the objective is to detect a change in an underlying process. In application to disease surveillance, the data are in the form of case counts for sub-regions of a larger study area. A running sum of deviations is recalculated at each time period. For a given sub-region, a count y_t of cases at time t is monitored as follows

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

where S_t is the cumulative sum alarm statistic, k is a parameter which represents the expected count, so that observed counts in exceedence of k are accumulated. At each time period, an alarm is signaled if S_t is greater than a threshold parameter h. If a CUSUM is

run long enough, false alarms will occur as exceedences are incrementally accumulated. The false-positive rate is controlled by the expected time it takes for a false alarm to be signalled, termed the in-control average run length, denoted ARL_0. The ARL_0 is directly related to the threshold value for *h*, which can be difficult to specify in practice. High values of *h* yield long ARL_0 and vice versa. In practice, approximations are used to determine a value for *h* for a chosen ARL_0 (Siegmund 1985), though this remains a key issue in CUSUM methods.

The basic univariate CUSUM in (1) can be extended to incorporate the spatial aspect of surveillance data. In this sense, CUSUM is a temporal statistical framework around which a space-time statistical test can be built. In an initial spatial extension, Rogerson (1997) coupled the (global) Tango statistic (1995) for spatial clustering in a CUSUM framework. For a point pattern of cases of disease, compute the spatial statistic, and use this value of the statistic to condition the expected value at the next time period. Observed and expected values are used to derive a *z*-score which is then monitored as a CUSUM (Rogerson 2005a). One scalar approach taken by Rogerson (2005b) is to monitor only the most unexpected value, or peak, of each time period as a Gumbel variate (Gumbel distribution is used as a statistical distribution for extreme values). An additional approach is to compute a univariate CUSUM in a parallel surveillance framework (Woodall and Ncube 1985). Here the threshold parameter *h* must be adjusted to account for the multiple tests occurring across the study area. Yet this approach takes no account of spatial relationships between sub-regions (i.e., spatial autocorrelation).

CUSUM surveillance of multiple sub-regions can be considered a multivariate problem where a vector of differences between the observed and expected counts for

each sub-region is accumulated. Spatial relationships between sub-regions can be incorporated by explicitly modelling the variance-covariance matrix. Rogerson and Yamada (2004) demonstrate this approach by monitoring a scalar variable representing the multivariate distance of the accumulated differences between observed and expected over all sub-regions. This is modelled as

$$MC1_t = \max(0, \| S_t \| - kn_t) \qquad (2)$$

where $\|S_t\| = \sqrt{S_t^t \sum^{-1} S_t}$, and $\sum$ is a variance-covariance matrix capturing spatial dependence, and S_t is a *2* x *p* vector of differences between observed and expected cases of disease in time *t* for each *p* sub-region (Rogerson and Yamada 2004).

CUSUM methods are attractive for prospective disease surveillance because they offer a temporal statistical framework within which spatial statistics can be integrated. They therefore overcome one of the limitations of traditional spatial analysis applied to surveillance in that repeated testing over time (and space) can be corrected for. A full description of the inferential properties of the CUSUM framework is given by Rogerson (2005). These methods are therefore most appropriate for long temporal scales, especially when historical data are used to estimate the baseline. Multivariate CUSUM given by Rogerson and Yamada (2004) is for a singular disease over multiple sub-regions, but could be used to monitor multiple diseases over multiple sub-regions. This may be most applicable in a syndromic surveillance application. The simplicity of univariate CUSUM makes training and technical expertise less of a factor than the multivariate case. Multivariate CUSUM is also more difficult to interpret and specification of the threshold parameter requires simulation experimentation or a large temporal extent from which to establish a baseline.

3.3.1.3 Scan statistics

Scan statistics developed originally for temporal clustering by Naus (1965) test whether cases of disease in a temporally defined subset exceed the expectation given a null hypothesis of no outbreak. The length of the temporal window is varied systematically in order to detect outbreaks of different lengths. This approach was first extended to spatial cluster detection in the Geographical Analysis Machine (Openshaw et al. 1987). The spatial approach looks for clusters by scanning over a map of cases of disease using circular search areas of varying radii. Kulldorff and Nagaralla (1995) refined spatial scanning with the development of the spatial scan statistic which adjusts for the multiple testing of many circular search areas. The spatial scan statistic overcomes the multiple-testing problem (common to many local spatial analysis methods) by taking the most likely cluster defined by maximizing the likelihood that the cases within the search area are part of a cluster compared to the rest of the study area. Significance testing for this one cluster is then assessed via monte carlo randomization. Secondary clusters can be assessed in the same way and ranked by p-value.

In Kuldorff (2001), the spatial scan statistic is extended to space-time, such that cylindrical search areas are used where the spatial search area is defined by cylinder radius, and the temporal search area is defined by cylinder height. In prospective analysis, candidate cylinders are limited to those that start at any time during the study period and end at the current time period (i.e., alive clusters). Significance is determined through randomization and comparing random permutations to the likelihood ratio maximizing cylinder in the observed data. An additional consideration to take account of multiple hypothesis testing over time (correlated sequential tests) is given by including previously

tested cylinders (which may be currently ‘dead’) in the randomization procedure (Kulldorff 2001).

The space-time scan statistic (Kulldorff 2001) approaches the surveillance problem in a novel way and aptly handles some key shortcomings of other local methods (multiple testing, locating clusters, pre-specifying cluster size). However, a limitation is that the expectation is conditional on an accurate representation of the underlying population at risk, data which may be hard to obtain. In long-term space-time surveillance scenarios, accurate population estimates between decennial censuses are rare or must be interpolated. In syndromic applications, where cases are affected by unknown variations in care-seeking behaviours, the raw population numbers may not accurately reflect the at-risk population. In Kulldorff et al (2005), the expected value for each unit under surveillance is estimated from historical case data rather than population data. Generating the expected value from the history of the process under surveillance is most suitable for real-time prospective surveillance contexts where the current state of the process is of interest. This extension allows the application of the space-time scan statistic in a wider range of surveillance applications.

A remaining limitation of the cylindrical space-time scan statistic is the use of circular search area over the map. The power of the scan statistics that use circular-based search areas decline as clusters become more irregular in shape, for example, for cases clustered along a river valley or where disease transmission is linked to the road network. The spatial scan statistic has been extended to detect irregularly-shaped clusters in Patil and Taillie (2004) and Tango and Takahashi (2005). Extensions of these approaches to space-time are active areas of research. A space-time version of the Tango and Takahashi

(2005) method uses spatial adjacency of areal units added incrementally up to K nearest neighbour units which are connected through time to form 3-dimensional prism search areas (Takahashi et al. 2008). A similar approach is given by Costa et al. (2007). However, these methods are very computationally intensive.

Scan statistics are one of the most widely used statistical methods for outbreak detection in surveillance systems. Space-time scan statistics are able to detect and locate clusters of disease, and can condition expected counts for individual sub-regions on population data or on previous case data, making these methods suitable for implementation where data volume is large. The scope of scan statistics, like most statistical tests, is limited to monitoring case data, either case event point data or counts by sub-region, and therefore inclusion of covariates to further condition the expectation is limited. Scan statistics are best served to detect and locate discrete localized outbreaks. Secondary clusters can be identified by ranking candidate clusters by their likelihood ratio. Yet region-wide outbreaks cannot be detected with scan-statistics because of the assumed form of a cluster as a compact geographical region where cases are greater than expected. Novel space-time methods that search for raised incidence via graph-based connectivity may model spatial relationships of disease processes more accurately than circular search areas. However the computational burden and complexity of these approaches limits their use to expert analysts and researchers. At the root of the problem is a conceptual discrepancy between the definition of a disease outbreak (which disease surveillance systems are often interested in detecting) and a disease cluster (defined by spatial proximity) which is common to all statistical testing methods for space-time surveillance (Lawson 2005).

3.3.2 Model-based approaches

Model-based approaches to surveillance developed recently as the need emerged to include other variables into the specification of our expectation of disease incidence. For example, we often expect disease prevalence to vary with age, gender, and workplace of the population under surveillance. Statistical models allow for these influences to adjust the disease risk through space and time. A second impetus for the development of statistical models for disease surveillance is that a large part of epidemiology concerned with estimating relationships between environmental variables and disease risk provided a methodological basis from which to draw. Modelling for space-time disease surveillance is relatively recent, and this is a very active area of statistical surveillance research. Again we stratify statistical models into three broad classes: generalized linear mixed models, Bayesian models, and models of specific space-time processes.

3.3.2.1 Generalized Linear Mixed Models

Generalized linear mixed models (GLMM) offer a regression-based framework to model disease counts or rates using any of the exponential family of statistical distributions. This allows flexibility in the expected distribution of the response variable, as well as flexibility in the relationship between the response and the covariate variables (the link function). One application of this approach to prospective disease surveillance for detection of bioterrorist attacks is given by Kleinman et al. (2004). Here, the number of cases of lower respiratory infection syndromes in small geographic areas act as a proxy for possible anthrax inhalation. A GLMM approach is used to combine fixed effects for covariate variables (i.e., season, day of the week) with a random effect that accounts for varying baseline risks in different geographic areas. In Kleinman et al. (2004), the logit link function is used in a binomial logistic model to estimate the expected number of

cases y_{it} in area i for time t. This is a function of the probability of an individual being a case in area i at time t and the number of people n_{it} in area i at time t.

$$E(y_{it} \mid b_i) = n_{it} p_{it} \qquad (3)$$

This expectation is conditional on a location specific random effect b_i and is then converted to a Z-score and evaluated to determine if it is unusual (i.e., an emerging cluster). This approach was extended to a model using Poisson random effects in Kleinman (2005). The use of GLMM in prospective surveillance has also been suggested for use in West Nile virus surveillance due to the ease with which covariates can be included and flexibility in model specification (Johnson 2008).

The GLMM approach has attractive advantages as a flexible modelling tool. Particularly, relaxation of distributional assumptions, flexibility in link functions, and the ability to model spatial relationships (at multiple spatial scales) as random effects make GLMM useful for prospective space-time disease surveillance. The scale and scope of the surveillance context does not limit a model-based approach, and models may be even more useful when data abnormalities such as time lags occur (as estimates can be based on covariates alone). One feature of GLMM that are important for many disease surveillance contexts are the ease with which spatial hierarchies can be incorporated. Ecological relationships that are structured hierarchically that impact disease emergence (e.g., climate, vegetation, vector life-cycle development) can be represented and accounted for. Further, human drivers of disease emergence (e.g., land-use policies, travel patterns, demographics) are often organized hierarchically through administrative units. In social sciences GLMMs are often used (i.e., multi-level models) that incorporate these 'contextual effects' on an outcome variable. A further advantage of GLMMs is their

ability to incorporate spatial variation in the underlying population at risk by conditioning the expected value on the random effect component (b_i in eq. 3). Where fewer people are present, the expected value is adjusted toward the mean. This can somewhat account for the small-numbers problem of SMRs in epidemiology, reducing the likelihood of estimating extremely low expected values in rural areas.

3.3.2.2 Bayesian Models

Bayesian models have been used extensively in disease mapping studies (Best et al. 2005; Lawson 2009). Analysis of disease in a Bayesian framework centers around inference on unknown area-specific relative risks. Inference on this unknown risk distribution is based on the observed data *y* and a prior distribution. These are combined via a likelihood function to create a distribution for model parameters which can be sampled for prediction. Bayesian models have been applied for retrospective space-time surveillance (e.g., MacNab 2003) and are now being developed for prospective space-time disease surveillance.

The basic Bayesian model can incorporate space and time dependencies. In Abellan et al. (2008) a model is described where the counts of disease are taken to be binomial distributed, and the next level of the model is composed of a decomposition of the unknown risks into model parameters for general risk, spatial effects, temporal effects, and space-time interaction. Estimation requires specifying prior distributions for each of the model components and sampling the posterior distribution via monte carlo markov chain (MCMC) methods. Here, the authors describe space-time Bayesian models for explanation of overall patterns of disease, speculating on their use in disease surveillance contexts. Rodeiro and Lawson (2006a) offer a similar model based on a

Poisson distributed disease count. Specifically, the counts y_i are Poisson with mean a function of the expected number of cases e_{ij} in location i at time j and the area specific relative risk rr_{ij}.

$$\log(rr_{ij}) = u_i + v_i + t_j + \gamma_{ij} \quad (4)$$

Similar to Abellan et al. (2008), the $\log(rr_{ij})$ are decomposed into spatial effects u_i, uncorrelated heterogeneity v_i, temporal trend t_j, and space-time interaction γ_{ij}. Again, these components need prior distributions specified. For the spatial correlation term, a conditional autoregressive model (CAR) is suggested for modelling spatial autocorrelation. Residuals are then extracted from model predictions for incoming data and can be used to assess how well the data fits the existing model. As discussed in Rodeiro and Lawson (2006a), monitoring residuals in this way makes the detection of specific types of disease process change feasible by adjusting how residuals are evaluated. While adding to the complexity of the analysis, this may be of great use in a surveillance application.

Alternative proposals such as Bayesian cluster models with "a priori" cluster component for spatiotemporal disease counts was developed by Yan and Clayton (2006). More recently, Bayesian and empirical Bayes semi-parametric spatiotemporal models with temporal spline smoothing were developed for the analysis of univariate spatiotemporal small area disease and health outcome rates (MacNab 2007a; MacNab and Gustafson 2007, Ugarte et al. 2010) and multivariate spatiotemporal disease and health outcome rates (MacNab 2007b). Tzala and Best (2008) also proposed Bayesian hierarchical latent factor models for the modeling of multivariate spatiotemporal cancer

rates. These spatiotemporal models, with related Bayesian and empirical Bayes methods of inference, may also be considered for disease surveillance applications.

The statistical methodology for applying Bayesian models to surveillance in space-time is still being developed, and as such these approaches are suited primarily to researchers. Bayesian models are attractive because they allow expert and local knowledge of disease processes to be incorporated via the specification of prior distributions on model parameters. However, this can also be a drawback, as a subjective element is introduced to the model. It is generally recommended that sensitivity analysis be conducted on a variety of candidate priors for model parameters (e.g., MacNab and Gustafson 2007, MacNab 2007a). These technical aspects of model-fitting require advanced statistical training. A further complexity of Bayesian models is estimation. MCMC methods are required for generating the posterior distributions for these types of models and are computationally very demanding (although see Rodeiro and Lawson 2006b). This might negate the use of these approaches in surveillance contexts that require daily refitting of models (i.e., fine temporal resolution), however monthly or annual model refitting may be possible. As with GLMMs, Bayesian models lend themselves to modelling hierarchical spatial relationships, and this can be important for both ecological and human-mediated drivers of disease emergence.

3.3.2.3 Models of specific space-time processes

Some modelling approaches to surveillance have been designed to model specific types of spatial processes, generally represented as a realization from a statistical distribution. While all models require some distributional assumptions, those considered here purport to associate specific statistical processes with disease processes in the context of

surveillance. In Held et al. (2005), a model is based on a Poisson branching process whereby outcomes are dependent on both model parameters describing a particular property (e.g., periodicity) and past observed data. Spatial and space-time effects can also be included as an ordinary multivariate extension. A useful aspect of this formulation for disease surveillance is the separation of the disease process at time *t* into two parts: an endemic part *v* and an epidemic part with conditional rate λy_{t-1}

$$\mu_t = v + \lambda y_{t-1} \quad (5)$$

The endemic component can also be adjusted for seasonality, day of the week effects and other temporal trends. Extended to the multivariate case, the model becomes

$$\mu_t = n_{it} v + \lambda y_{i,t-1} \quad (6)$$

where the endemic rate adjusted by the number of people in area *i* at time *t*, and area-specific previous model estimates for the epidemic part. Spatial dependence can be incorporated by adding a spatial effects term that accounts for correlated estimates in $\lambda y_{i,t-1}$ via a weights matrix. However, this type of model yields separate parameters for each geographical unit.

A point process methodology for prospective disease surveillance is presented in Diggle et al. (2005). Point data representing cases are modelled with separate terms for spatial variation, temporal variation, and residual space-time variation. The method is local, in the sense that recent cases are used for prediction, producing continuously varying risk surfaces. However, there are also global model parameters which estimate the background variation in space and time estimated from historical data. Outbreaks are defined when variation in the residual space-time process exceeds a threshold value *c*.

Different values for the threshold parameter are evaluated and exceedence probabilities are mapped. Model parameters are fixed allowing the model to be run daily on new data. However, as noted in Diggle et al (2005), this may fail to capture unknown temporal trends, and periodic refitting may be required.

A different approach is given by Järpe (1999), which instead of decomposing the process into separate components, monitors a single parameter of spatial relationships in a surveillance setting. This is similar in spirit to Rogerson's work (Rogerson 1997) monitoring point patterns with spatial statistics, though here a specific underlying process is assumed: the Ising model. The Ising model represents a binary-state two dimensional lattice (sites coded 0 or 1). There are two parameters for the Ising model; one governs the overall intensity (probability of a site being a 1), and another the spatial interaction (probability of nearby sites being alike). In Järpe (1999), the intensity parameter is assumed equal and unchanging, and the surveillance is performed on the interaction parameter under different lattice sizes and types of change. The interaction parameter is essentially a global measure of spatial autocorrelation. This can then be monitored using temporal surveillance statistics such as CUSUM. Since the properties of the underlying model are known, Järpe is able to detect very small changes in spatial autocorrelation which could indicate the shift of a disease from endemic to epidemic. While significant spatial autocorrelation is often present at both endemic and epidemic states, changes in clustering can reveal threshold dynamics of the process in a surveillance setting. This is a common feature of forest insect epidemics (Peltonen et al. 2002). Further, the effect of the lattice size can easily be estimated, and as lattice size is increased, sensitivity to changes in the interaction parameter increases as well.

While most methods discussed thus far have been developed with the analysis of aggregated counts of disease in mind, analysis of sites on a lattice may have applicability in certain disease surveillance contexts. For example, square lattices are used for remotely sensed image processing, and surveillance of the presence or absence of a disease in these sampling units using an Ising model-based approach could incorporate remotely sensed environmental covariates (e.g., normalized differential wetness index) as is commonly done for zoonotic disease risk mapping and forecasting (Kitron et al. 1996; Rogers et al. 1996; Wilson 2002). However, it is unclear how covariates are included in the Ising model. This highlights an important point with model-based approaches to prospective surveillance: the main advantage of models is to incorporate extra information and to estimate smooth relative risks, yet as models grow in complexity they become more difficult to re-fit. This has implications for how suitable models are in different surveillance contexts. Where the temporal scale is large, expected counts can be based on observed data rather than using census or other data sources. This is particularly important where diseases follow seasonal trends. With limited temporal data available, estimated model parameters may be impacted by regular variation in disease occurrence. For surveillance systems monitoring many small areas (i.e., large spatial scale), the Held et al. (2005) model would be of limited value, as separate parameters need to be estimated for every sampling unit. Broad scale patterns over large areas might better captured by the point process approach of Diggle et al. (2005). Although here, case event data with fine spatial resolution is required.

For all modelling approaches, complex decisions are required such as what covariates to include, how often to re-fit the model, how to test incoming data for fit

against the existing model which require advanced statistical knowledge. This limits the applicability of modelling approaches to advanced analysts and researchers except for use in a black-box sense by analysts and public health practitioners. Surveillance models can be tailored to detect specific types of disease process changes, such as a region-wide increase, or small changes in spatial autocorrelation suggesting a shift from endemic to epidemic states. However, models also required additional tests to determine if incoming data differ from the expected (i.e., modelled) pattern of cases. Thus, in practice surveillance models are best utilized to estimate a realistic relative risk, and can then be combined with statistical tests such as CUSUM (Järpe 1999) and scan statistics (Kleinman et al. 2005).

3.3.3 Emerging research areas

Research into space-time disease surveillance methods has increased dramatically over the last two decades. Many new methods are designed for specific surveillance systems, or are in experimental/developmental stages and not used in practical surveillance. Here, we report on some newly developed approaches for public health surveillance to alert readers to the most recent developments in these emerging research areas.

While test and model-based approaches to surveillance build on classical statistical methods, many recent space-time disease surveillance methods have been developed specifically to take advantage of advanced computing power and data sources. While virtually all space-time disease surveillance makes use of computer technology, hybrid methods are developed specifically for massive datasets and computationally-based solutions. These approaches include networks (Reis et al. 2007; Wong and Moore

2005) simulation-based methods such as agent-based models (Eubank et al. 2004) and bootstrap models (Kim and O'Kelly 2008), and hidden markov models (Madigan 2005; Sun and Cai 2009; Watkins et al. 2009).

Other new methods are designed to address limitations of existing surveillance methods. One problem for most methods of surveillance, is the specification of the null hypothesis, or expected disease prevalence. While expected rates are generally conditional on population data, spatial heterogeneity in the background rates are rarely accounted for. That is, complete spatial randomness (CSR) is the underlying null model. Goovaerts and Jacquez (2004) have used geostatistical approaches, estimating spatial dependence of background rates via the semivariogram, to develop more realistic null models for disease cluster detection. The geostatistical framework has the advantage of estimating spatial dependence from the data, rather than defining it a priori via a spatial weights matrix as is common in disease mapping models.

Another problem common to most surveillance methods is that maps of disease represent either home address (case events) or small areas (tract counts). Unusual clusters on the map imply heightened risk is associated with those locations. However, movement of animals and people decouples the location of diagnosis from disease risk by modifying exposure histories. Methods that account for mobility may be an important area for future surveillance, especially in the context of real-time, prospective outbreak detection. The relationship between case, location, and exposure is further complicated by disease latency periods, which gives rise to space-time lags in diagnoses (Shaerstrom 1999). This may be most important in the context of retrospective cluster analysis and investigation of possible environmental risk factors. Statistical tests have been developed to account

for exposure history and mobility for case-control data (Jacquez and Meliker 2009) and case-only data (Jacquez et al. 2007). Kernel-based approaches to risk estimation that incorporate duration at each location have been utilized for amyotrophic lateral sclerosis (Sabel et al. 2003). The general approach is to model and analyze the space-time path of individuals in the sense of Hagerstrand (1967). As personal location data continues to become ubiquitous due to new technology such as GPS-enabled cell phones, surveillance methods that account for individual space-time histories may see more application in public health surveillance.

3.4 Summary

The development of space-time disease surveillance systems holds great potential for improving public health via early warning and monitoring of health. The selection of which method(s) to implement in a given context is dependent on a variety of factors (Table 2). This review has demonstrated that there is no best method for all systems. There are many aspects to consider when thinking about methods for space-time disease surveillance. Many of the methods described in this review are active areas of research and new methods are constantly being developed. As more data sources become available, this trend is expected to continue, and the methods described here provide a snapshot of options available to public health analysts and researchers. A brief outline of some of the factors reviewed and how they relate to surveillance methods is given below.

The spatial scale of the surveillance context is an important factor for selecting appropriate methods. Spatial effects (i.e., clustering) are likely only of interest when cases/counts collected over a relatively large, heterogeneous area are being analyzed. Over smaller more homogeneous areas, where spatial effects are negligible, temporal

surveillance is optimal. When space-time surveillance is warranted, choice of which surveillance approach to use may be impacted by how spatial effects can be incorporated. Where spatial scale is small, one would likely focus on either process models or statistical tests which use an underlying distribution for the null hypothesis (i.e., Poisson model). The temporal scale of surveillance is also important. Large temporal scales can use either testing or modelling methods, and most suit methods where baselines are estimated from previous cases, such as with the space-time permutation scan statistic. Short temporal scales are not appropriate for models when diseases have complex day of the week effects or seasonal variation in incidence. Scale will also affect the computational burden placed on the system. Many approaches reviewed here, particularly statistical tests such as scan statistics, use approximate randomization to generate a distribution of a test statistic under the null hypothesis. Methods that utilize randomization procedures, while powerful, impose constraints when applied with large spatial-temporal datasets.

Most methods are designed for a single disease and all methods are suitable for single host diseases, but finer detail in case distribution may be important for multiple host zoonotic diseases. Stratification into separate diseases by host type will result in a loss of information as associations between host types will be lost. As zoonotic diseases make up the majority of emerging infectious diseases (Greger 2007), multiple host surveillance methods are required. Multivariate tests such as multivariate CUSUM can be used to monitor multiple signals. Modelling approaches can also be used by creating a generalized risk index as the variable under surveillance. Multivariate extensions to

existing methods can be used to monitor associations between two diseases, for example, human and animal strains of the same pathogen.

The objective of surveillance is one of the main drivers of method selection. All statistical tests are commonly used for outbreak detection. In general, modelling approaches are better suited to monitoring space-time trends. For what has been termed situational awareness, multiple signals are usually monitored. This is often the case in large syndromic applications such as BioSense and ESSENCE. These contexts are best suited to a modelling approach, as often heterogeneity needs to be modelled with covariates.

Consideration of technical expertise is required for practical disease surveillance. Broadly speaking, greater statistical expertise is required for model-based methods than testing (understanding model assumptions, paramaterizing models, preparing covariate data, and interpreting output), while testing concepts are generally easier to grasp. However, for epidemiologists already familiar with generalized linear mixed models, some model approaches that incorporated space and time may be quickly attainable, such as that of Kleinmann et al. (2004). Yet for analysts from a health geography or spatial analysis background, testing methods might be more familiar. In any case, the use of space-time surveillance methods in public health will only increase in the future, and it is important that training and education keep pace with the changing methods available for surveillance data analysis.

3.5 Acknowledgements

This project was supported in part by the Teasdale-Corti Global Health Research Partnership Program, National Sciences and Engineering Research Council of Canada,

and GeoConnections Canada. The authors would like to thank Dr. Barry Boots for direction and suggestions during the starting phase of this research.

Table 3.1. Contextual factors for evaluation of methods for space-time disease surveillance

Factor	Description
Scale	The spatial and temporal extent of the system (e.g., local / regional / national / international)
Scope	The intended target of the system (e.g., single disease / multiple disease, single host / multiple host, known pathogens / unknown pathogens)
Function	The objective(s) of the systems (outbreak detection, outbreak characterization, outbreak control, case detection, situational awareness (Mandl et al. 2004; Buehler et al. 2004), bio-security and preparedness (Fearnley 2008))
Disease Characteristics	Is the pathogen infectious? Is this a chronic disease? How does it spread? What is known about the epidemiology of the pathogen?
Technical	The level of technological sophistication in the design of the system and its users (data type and quality, algorithm performance, computing infrastructure and/or reliability, user expertise)

Table 2 Summary of contextual factors and methods of space-time disease surveillance

Class	Type	Scale	Scope	Function	Characteristics	Technical
Test	CUSUM	Temporal statistical framework useful for long time periods of sequential surveillance.	Univariate CUSUM useful for single diseases while multivariate CUSUM useful when multiple diseases or syndromes are under surveillance.	Primarily for outbreak detection.	Multivariate CUSUM is not sensitive to outbreak type (one extreme vs. many subtle rises) whereas the univariate is.	Difficulty in specification and understanding of the threshold parameter.
Test	Interaction	Population shift bias increases with spatial and temporal scale.	Cannot analyze interactions and relationships in multiple host diseases	Can only detect presence of interaction. Limited utility for outbreak detection. Best used as screening method.	Interaction tests cannot capture interactions and flows between units under surveillance (spatial autocorrelation).	Require geocoded event data of cases of disease. Ease of understanding and interpretation. Subjectivity in specification of critical distances in space and time.
Test	Scan	Space-time scan statistics are able to detect and locate clusters. Using the permutation-based approach can make use of temporal history of data. Appropriate mostly where there is a large volume of data in space and time	Scan statistics are designed to monitor one data stream, and therefore in and of themselves are not suitable for multiple disease, nor can they include covariates. Can be combined with models as in Kleinman et al (2005).	Monitoring p-values of primary and secondary clusters can be useful for assessing trends over time, although primary function is for discrete localized outbreak detection.	Cylindrical search areas assume compact cluster form. Extensions using graph-based connectivity for search areas are computationally very demanding. Spatial relationships not defined by proximity may be more important for disease spatial processes	Can be used with point event data or count data. Ease of understanding and interpretation of results of analysis.
Model	GLMM	Increase in utility as the size of the surveillance database grows. Temporal trends can be incorporated as model parameters. Frequent	Models can be formulated for risks, incidence and counts of diseases. Very flexible in how dependent variable is structured.	Monitoring space-time trends in disease incidence, however all modelling approaches need to be coupled with a statistical test to	Can incorporate hierarchical effects of covariates easily including spatial effects.	The most accessible of modelling approaches but requires knowledge of statistical distributions. Limited mostly to researchers and statistical

		refitting of complex models can be difficult.		determine unexpected events (i.e., outbreaks).		analysts. Flexible choice of statistical distributions compared to OLS modelling.
Model	Bayesian	*Same as above*	*Same as above*	*Same as above*	*Same as above*	Priors need to be specified for model parameters. Advanced statistical knowledge required. Fitting complex space-time Bayesian models requires MCMC methods. Not suitable if need to be refit often.
Model	Processes	Can be used with data of any scale as testing is against a specified process.	Multiple hosts and pathogens can be accounted for though may be difficult to parameterize.	Generally high sensitivity to detecting different types of change such as periodic outbreaks or gradual shifts away from the process. Needs to be coupled with a statistical test.	Characteristics of disease (e.g., transmission, serial interval) can determine choice of process. Can also be used as exploratory tool.	Models in this class vary greatly. Technical factors will be specific to individual process models selected.

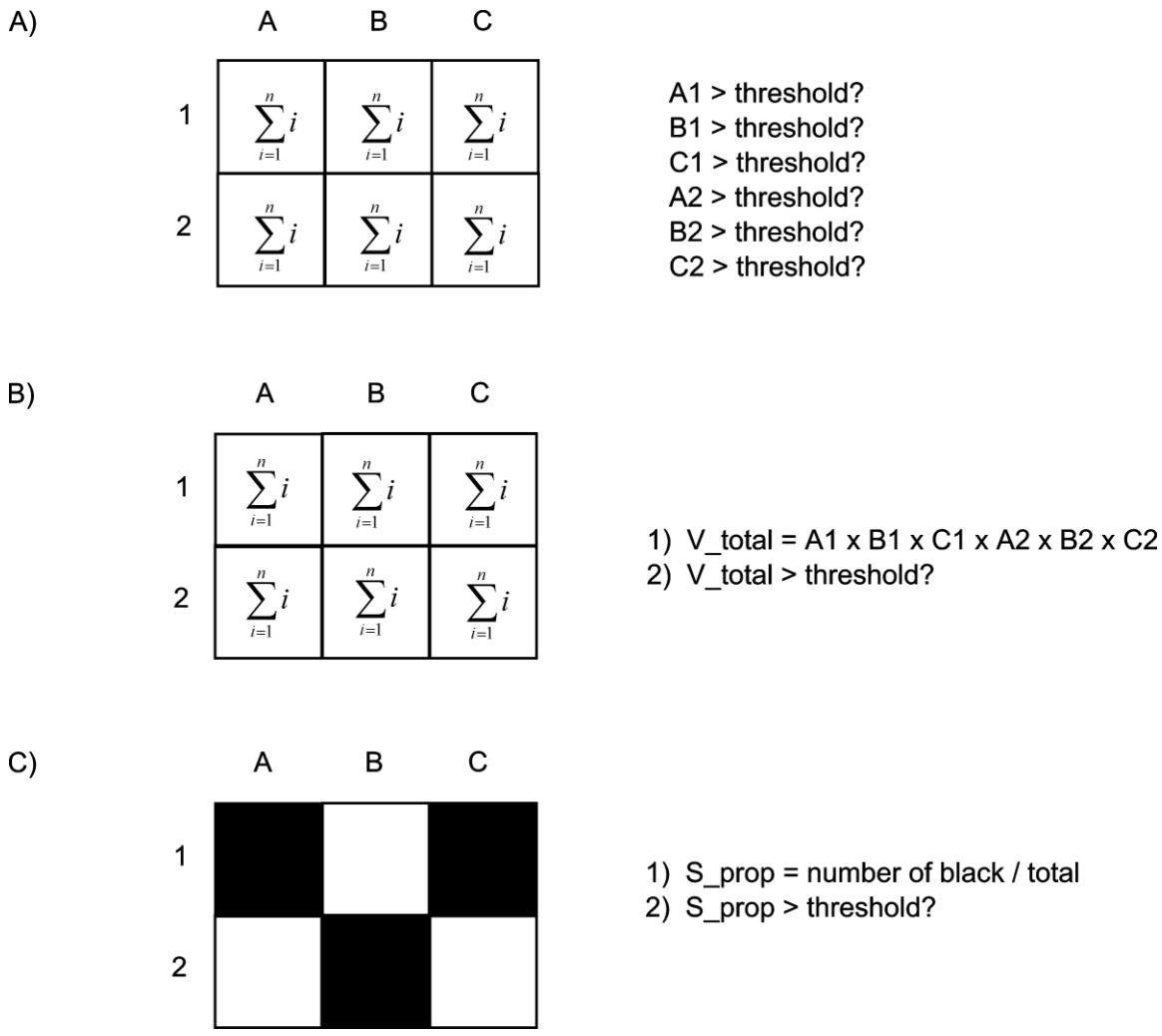


Figure 3.1. Methods for prospective surveillance. A) Parallel surveillance where a test statistic is computed separated for each region under surveillance and each assessed individually. B) Vector accumulation where test statistics in a parallel setting are combined to form one alarm statistic which is evaluated. C) Scalar accumulation where on statistic is computed over all regions under surveillance and evaluated.

Chapter 4: Review of software for space-time disease surveillance

4.1 Abstract

Disease surveillance makes use of information technology at almost every stage of the process, from data collection and collation, through to analysis and dissemination. Automated data collection systems enable near-real time analysis of incoming data. This context places a heavy burden on software used for space-time surveillance. In this paper, we review software programs capable of space-time disease surveillance analysis, and outline some of their salient features, shortcomings, and usability. Programs with space-time methods were selected for inclusion, limiting our review to ClusterSeer, SaTScan, GeoSurveillance and the Surveillance package for R. We structure the review around stages of analysis: preprocessing, analysis, technical issues, and output. Simulated data were used to review each of the software packages. SaTScan was found to be the best equipped package for use in an automated surveillance system. ClusterSeer is more suited to data exploration, and learning about the different methods of statistical surveillance.

4.2 Introduction

Disease surveillance is an ongoing process of information gathering, organizing, analyzing, interpreting, and communicating. It is the principal means by which public health information is generated and disseminated, informing policy, research, and response measures. For outbreaks of infectious disease, timely information on the spread of cases in space and time can facilitate action by public health officials (Ekpo et al. 2008). For chronic and endemic diseases, monitoring space-time trends in disease occurrence can highlight changing patterns in risk and help identify new risk factors (e.g.,

Kim et al. 2008). Analysis of spatial-temporal patterns in public health data is an increasingly common task for public health analysts as more surveillance data become available. Surveillance datasets are often massive in size and complexity, and the availability and quality of software capable of analyzing space-time disease surveillance data on an ongoing basis is integral to practical surveillance (Aylin et al. 1999; Richards et al. 1999; Rushton 2003). Geographic information systems (GIS) used for disease mapping can visualize the spatial variation in disease risk. However, statistical methods are often required to detect changes in the underlying disease process. GIS are also poorly equipped to handle temporal data (Langran 1992).

In Fall of 2008, a workshop on training priorities in the use of GIS in health research conducted in Victoria, British Columbia, polled 78 researchers, graduate students, faculty, and others working in health and GIS regarding barriers to the use of space-time disease surveillance (Population Data BC 2008). Training and software availability were cited as the primary barriers to the uptake of space-time disease surveillance. Currently, statistical methods for space-time disease surveillance are not included in most conventional GIS or statistical software. These methods are available in specialist cluster analysis software such as ClusterSeer (www.terraseer.com), or as extensions to general statistical analysis software packages (e.g., R, S-Plus). Our goal is to provide researchers and public health analysts with a review and demonstration of software packages for space-time disease surveillance. We aim to facilitate expanded use of these methods by providing a means to quickly determine the software options and to identify the ways in which programs differ. We limit our scope to methods that use both space and time, rather than purely temporal or spatial analysis.

This paper is organized as follows. First, we briefly review basic classes of methods for space-time disease surveillance in the background section. Readers familiar with these methods may wish to skip ahead. Second, in the methods section we outline how we selected software to review, the review methodology and datasets used to demonstrate software features. Third, we present the results of our review. Finally, we conclude with some guidelines for the use of these software packages for public health researchers and analysts.

4.3 Background

Statistical approaches to disease surveillance have been the subject of a number of texts and review papers (Sonesson and Bock 2003; Lawson and Kleinman 2005). A key factor in the selection of methods of analysis is the objective of surveillance, such as outbreak detection, trend monitoring, case detection, or situational awareness. Additional contextual factors are also important to consider such as scale and scope of the system, disease characteristics, and technical considerations (Robertson et al. 2010). Methods can be broadly categorized as either statistical tests or model-based approaches. Statistical tests are the dominant class of approaches used for outbreak detection. The aim of most methods is to test a subset of data, defined by spatial and temporal constraints (i.e., a window or kernel), against an expected rate of disease occurrence over the study area as a whole. Methods differ with respect to how the window that defines each subset is constructed, how statistical significance is determined, and how the baseline expectation varies over space and time.

The most widely used testing methods are cumulative sum (cusum) methods and scan statistics. Briefly, cusum approaches keep a running sum of deviations from the expected value, and once the cumulative deviation reaches some threshold, an alarm is triggered. For space-time applications, individual cumulative sums for each area under surveillance are monitored and can be adjusted for spatial relationships (Rogerson and Yamada 2004). Depending on the statistic being monitored in the cusum, different surveillance objectives can be addressed. For example, a measure of spatial pattern monitored in a cusum framework can be sensitive to slight changes in spatial pattern which may signal a shift in dynamics of an endemic disease (e.g., Rogerson 1997). Scan statistics are used mostly in outbreak detection contexts. Here, circular search windows of varying radii scan a map of disease and test if the number of cases within the search area is unexpectedly high. In the space-time scan statistic (Kulldorff 2001; Kulldorff et al. 2005), the search area is extended to a cylinder where the height of the cylinder is defined by time periods of varying lengths. The mostly likely cluster is assessed using monte carlo simulations.

Modeling approaches are used mostly for adjusting the expected number of cases (i.e., denominator) of disease. Disease incidence varies spatially with population and known risk factors. Disease mapping models aim to estimate the true *relative risk* across the study area by incorporating the spatial variation in these risk factors. The standardized mortality ratio (SMR) is the crudest measure of risk, computed as the observed number of deaths due to a disease divided by the expected in each area. The SMR is often of limited use in surveillance because it can fluctuate widely for rare diseases or in rural areas where populations are small. Further, abrupt (i.e., unrealistic) changes at the boundaries

of areal units are sometimes observed. Models allow both covariate effects to be estimated, and for sparsely populated areas to have their expected values adjusted towards the mean (i.e., borrow strength). When used in surveillance applications, models confer these same advantages. Disease surveillance models have been either space-time Bayesian models (e.g., Vidal-Rodeiro and Lawson 2006) or generalized linear mixed models (e.g., Kleinman et al. 2004). Modeling approaches are complementary to other methods as tests are still required to determine how well the most recently observed data fit with the model (Kleinman et al. 2005). Adjustments can also be such that models can be refit over time to adjust to long-term changes in disease occurrence or surveillance effort/efficacy (e.g., improved diagnostic tests), and parameters can be included to model spatial relationships and seasonal and day of the week effects, common features of some types of disease surveillance data.

In addition to testing and modeling methods, new computation-based tools are also being developed for surveillance. These approaches tend to be in either experimental and/or theoretical stages or algorithms designed for specific surveillance systems. Some hybrid approaches include networks (Reis et al. 2007), simulation-based methods (Kim and O'Kelly 2008), and space-time hidden markov models (Watkins et al. 2009). While many of these new approaches appear promising, most are not yet available in software.

4.4 Methods

4.4.1 Inclusion criteria

Software programs were included for review based on two criteria: the program had methods that handled both space and time, and methods were built-in to the software (i.e., not requiring programming). Software programs were found through internet searches,

review of the literature, and previous experience of the authors. This criteria constrained our review to four software packages (SaTScan 8.0, ClusterSeer 2.3, GeoSurveillance 1.1, Surveillance package 1.1-2 for R). Comprehensive disease surveillance systems (also sometimes called health information systems) software, that include data collection and processing routines, database components, and system-specific analysis and visualization modules were excluded (e.g., RODS - Tsui et al. 2003, AEGIS - Reis et al. 2007). These systems are large in scale and generally implemented at an enterprise level; they are not readily accessible to researchers/analysts. Research tools based purely on programming (e.g., WinBUGS; MatLab) were also excluded. Details of the software packages included in the review are outlined in Table 1.

4.4.2 Reviewing framework

Software programs were reviewed for broad steps of typical data analysis: preprocessing, analysis (methods and technical issues), and output (Table 2). Preprocessing is required to transform data into the appropriate structure for a particular software package. In many scenarios, health event data are collected at an address level, which needs to be compared to population estimates, available usually as polygon census data (Gotway and Young 2002). Our assessment of data preprocessing requirements reflected typical data by considering both point event case data, and polygonal administrative units. For each software package, we assessed the data formatting steps required to perform an analysis.

The second step is conducting the analysis and we briefly describe methods and analysis options for each software package. We highlight technical issues and potential problems or requirements such as stability, speed of computation, and required operating systems. All analyses were run on a Pentium 4 PC with 3.00 GHz processor and 2 GB of

RAM running the Windows XP operating system. The final step is outputting results and we overview output options available in each package. In addition, we qualitatively assess user facility based on our experience operating the software with test datasets. It should be noted that we do not discuss parameterization of different methods. This is a major issue in practical surveillance, suited to a review and comparison of surveillance methods themselves.

4.4.3 Datasets

Data were simulated to model a syndromic surveillance system monitoring calls to a health hotline in Vancouver, British Columbia, Canada. For simplicity, we refer to each simulated call as a case. Cases were simulated over one year from January 1st to December 31st. Cases were aggregated to census dissemination areas (DA) and were spatially allocated proportional to the population in each census DA. The total population in all DAs was 578,642, and total cases were 4303, giving an annual incidence of 743.64 cases per 100,000. This level of incidence is similar to what might be expected for the total volume of calls made to a telephone health hotline in a major Canadian city (Perry 2009).

Outbreaks were inserted into baseline data to indicate signals of a spike in calls which, in a sydnromic surveillance setting, indicate a signal of an unusual health event. Two outbreak scenarios were simulated in separate datasets. In outbreak one, a simulated outbreak started on March 4th and lasted until June 5th, with 148 cases occurring over 10 sq km, covering 33 geographically adjacent census DAs (light grey cluster, Figure 4.1). Outbreak cases were allocated proportional to census DA population. In outbreak two, 6 spatial clusters constituting a total 501 cases occurred over an area of 16 km^2, covering a

total of 104 census DAs (dark grey cluster, Figure 4.1). The number of cases in clusters ranged from 51 to 140, and cases occurred over the full year. Data were stored in Environmental Systems Research Institute (ESRI) shapefile format, a standard spatial data format which can represent data as points, polygons, or lines.

4.5 Review of Programs

4.5.1 Data preprocessing

The steps involved in preprocessing the test data for analysis in each software program are outlined in Table 3. SaTScan requires data to be input as three separate files to run the appropriate analysis for this data (retrospective space-time scan, Poisson model) where one file stores the spatial locations (geo file), another file stores the cases (case file), and a third stores the population of each area (population file). All SaTScan files are text-based, and an import tool is provided for importing common data formats (e.g., CSV, DBF). SaTScan also provides the functionality to aggregate the data temporally into years, months, or days. Thus, data can be input at the finest temporal resolution. This functionality turned out to be a key advantage over other programs as it limited the amount of data restructuring required when trying different analysis parameters.

ClusterSeer requires unique records for every space-time unit under surveillance. Running a daily space-time scan statistic for our simulated data would require a dataset with four columns (location, date, cases, population) and 478,515 records (365 days x 1311 census DAs). Additionally, all areas need a record for every time period. Generating the necessary table required use of specialized data restructuring functions in R statistical software (reshape package). Data were aggregated to counts of cases by week. (52 weeks

x 1311 census DAs) giving a table with 68,172 records. For weeks where DAs had no cases, zero counts had to be inserted.

Preparing data for analysis in GeoSurveillance required aggregation temporally and spatially. Counts of cases were required to be attributes of the polygon shapefile (or text file), and fields were required to be named in sequential order. This process was automated by custom programming in ArcGIS which performed spatial joins and added new fields to the attribute table. This was an extensive process to get the data in the proper format for analysis, and similar to ClusterSeer, GeoSurveillance does not allow flexibility in the level of temporal aggregation. ClusterSeer and GeoSurveillance can both read in polygon shapefiles and automatically calculate centroid coordinates.

For analysis with the Surveillance package in R, data were required to be in a matrix with temporal observations as rows and spatial units as columns, giving a 365x1311 matrix for daily analysis and 52x1311 for weekly analysis. All of the programs except SaTScan had inflexible data input requirements, specifically for temporal aggregation of cases. None of the software programs could input the two shapefiles (points and polygons) without any data preprocessing. This was surprising as previous experience and a review of SaTScan (Block 2007) suggested cumbersome input format as a major limitation of SaTScan.

4.5.2 Analysis methods

The programs reviewed here are of two types: specialized implementation of a specific class of surveillance algorithms (SaTScan, GeoSurveillance) and full suite surveillance / space-time analysis packages that implement multiple methods (ClusterSeer, R-surveillance). SaTScan offers a number of scan statistics such as spatial (Kulldorff and

Nagarwalla 1995), temporal (Nagarwalla 1996), and space-time versions (Kulldorff 2001, Kulldorff et al. 2005), as well as retrospective and prospective (clusters must be current) modes. Different data types can be accommodated by the many probability models including Poisson, Bernoulli, space-time permutation, multinomial, ordinal, exponential, and normal. The circular search area used in the classical scan statistic can also be altered to search using an ellipse, or along user-defined connections of spatial units. GeoSurveillance implements the cusum approach to surveillance (e.g., Rogerson 1997). The retrospective mode does global spatial analysis only (i.e., reports one cusum test statistic for the map), while the prospective mode does univariate parallel surveillance with the cusum statistic. The multivariate cusum is not yet implemented in GeoSurveillance.

ClusterSeer had the widest range of space-time methods implemented. Those particularly suited to disease surveillance included space-time scanning (Kulldorff 2001), a cusum approach similar to that in GeoSurveillance (Rogerson 1997), and tests for space-time interaction (Knox 1964; Mantel 1967; Jacquez 1995). This makes ClusterSeer a useful tool for exploring disease surveillance data. Once data is formatted for use in ClusterSeer, a variety of methods can be used to examine the data. The R-Surveillance package contains a number of algorithms such as the Farrington et al. (1996) method, Poisson cusum (Rossi et al. 1999), and the two-component negative binomial model in Held et al. (2006). The algorithms in the surveillance package are mostly model-based and non-spatial, though some space-time surveillance applications can be treated as a multivariate time series problem.

4.5.3 Technical issues

Technical issues encountered in running the software programs varied considerably. SaTScan was capable of running the space-time scan statistic in retrospective mode on daily case data. ClusterSeer was not run on daily data. Initially, memory requirements were a serious limitation of undertaking analysis in ClusterSeer with both test datasets; however an updated version (2.3.22.0) was obtained to complete the analysis on weekly data. The analysis took longer to run than on SaTScan with daily data, though results were very similar. GeoSurveillance ran the univariate cusum in parallel on each of the 1311 census DAs. The analysis ran well on weekly data, however the linked display between the maximum cusum and the map was very slow. The cusum methods were also used for our analysis in R-Surveillance. The time taken to run the analysis on the weekly data was similar to that of GeoSurveillance and results were also similar.

R-Surveillance is the only package that runs on windows, mac and linux operating systems. Currently, SaTScan has versions for windows and linux, and a mac version is in development. Both ClusterSeer and GeoSurveillance run only on the windows operating system. SaTScan completed analysis in the shortest time compared to all other programs.

4.5.4 Data output

Output options in SaTScan are limited to text file and database file output. Database files can be linked back to the input shapefile in a GIS for further examination of clusters, however no data exploration functionality is available in SaTScan itself. In GeoSurveillance results of an analysis can be written to text file which can be easily manipulated in other software. GeoSurveillance provides a basic map interface linked to a list of cusum scores. A cusum chart is also displayed showing the temporal pattern of cusum scores for the study area as a whole and individual units.

ClusterSeer has advanced data output facility such as mapping and graphing which can be exported as images. Results can also be exported with the data to new files for further examination inside statistical or GIS software. The Surveillance package has access to extensive visualization and exporting functions available in the R environment. The objects specific to the Surveillance package also have default methods for creating plots. This of course requires familiarity with the R programming language.

4.5.5 User Facility: Ease of learning, ease of use, help & documentation

Usability is an important part of software as public health organizations have limited resources available for technical training. Our review of user facility is presented in Table 4. ClusterSeer includes an extensive help menu explaining the parameters and required data for all of the methods. The help system also includes tutorials and example datasets that work through many of the methods. This is an important resource for learning methods of spatial and space-time analysis. The graphical user interface (GUI) of ClusterSeer makes learning and use straightforward. SaTScan is also a GUI-based system composed of three main screens: input, analysis, and output. The help menu in SaTScan is extensive with descriptions of the scan statistic methodology, explanations of parameters and data input and output options, sample datasets, and references for further reading. GeoSurveillance has two basic modes which are run from menus of a simple GUI. The program is easy to use after data has been formatted properly (as described above). Currently there is no help built into the system itself. The menus are described in a separate word document. A tutorial and sample datasets are also provided. Having these outside of the program itself makes navigating the documentation cumbersome. R-Surveillance is an R package and as such has help in the R package format, which can be

called directly from R. This includes descriptions of parameters and values for all of the implemented functions in the package. Basic examples are given, although detailed descriptions of the statistical methods is lacking. Users should be familiar with using R packages and the background statistical methodology before using the surveillance package.

4.6 Conclusions

With the advent of electronic medical records, syndromic data sources, and low-cost location sensors, disease data are increasingly encoded with both spatial and temporal information. These new data sources represent an opportunity for greater understanding of disease distributions, risk factors, and changes to population health over time and space. While analysis of surveillance data represents an expanding opportunity for public health practice and research, these new datasets, methods, and software also bring challenges. There are inherent problems in using traditional statistics for hypothesis testing, or applying simple GIS visualization, to these data sources. As is evidenced by the growing literature on statistical surveillance of disease data (Lawson and Kleinman 2005), methods need to be specifically suited to these data. In addition to statistical methods however, computer software is now essential for the analysis of surveillance data.

The four software programs reviewed in this paper provide functionality for different kinds of analysis and serve different purposes. Based on our review, SaTScan is the most developed and robust software package (i.e., fastest, least susceptible to crashing) for implementation in an automated cluster detection system. However, SaTScan only implements scan statistic methods, so those wishing to explore modeling-

based approaches may want to use the Surveillance package. Additionally, examining the results in detail requires other software for graphing and mapping. Reasons for taking a modeling approach include making refined estimates of expected rates based on modeled covariate effects, adjusting for spatial heterogeneity in disease rate, and smoothing relative risks. The Surveillance package implements models, but currently has very limited capability for true space-time surveillance. The large number of temporal methods make it a useful environment for exploring surveillance data, in addition to the advantages afforded by being able to integrate with other R packages. As a command-based system, it also is easy to automate and integrate with data processing scripts. The learning curve for R is quite steep, and those requiring a GUI-based system to explore surveillance data would be better served by ClusterSeer. The extensive documentation and many purely spatial and temporal methods, in addition to space-time methods, makes it a convenient tool for initial data exploration. There is also a range of output options in ClusterSeer. ClusterSeer may be more suitable for exploratory studies than as part of an ongoing, automated cluster detection system because there is limited capacity for automated surveillance. ClusterSeer project files can be set to run automatically, though because they are binary files they cannot be automatically configured to increment parameters (e.g., study period). Finally, though methods (and software) have been classified as testing or model-based approaches, it is important to note that these approaches are complimentary rather than opposing (Kleinman et al. 2005). For example, one approach is to develop a model of the expected risk of disease using the Surveilllance package, and use the estimated smoothed rates as the expected values in a SaTScan analysis.

All of the programs reviewed in this paper were applications installed on a local computer. While this is the architecture of most computer software applications, new developments in computing are taking advantage of the internet to perform ongoing, high-powered computing tasks (Armbrust et al. 2009). Online delivery of analytic services (such as cluster analysis) allows software to be centralized on one server, and accessible from anywhere with an internet connection. In the context of disease surveillance, this could facilitate standardization of analysis among different regional health authorities, increase transparency of analysis, and offer significant improvements in costs and performance. Initial steps towards web-based surveillance analysis are underway, with a web-based version of ClusterSeer (https://www.clusterseer.com) currently in development, RWeb (http://www.math.montana.edu/Rweb]) , a web-based interface to a server instance of R, as well as a newer project called rapache (Horner 2009), which integrates R into the popular Apache web server. These developments hold considerable promise for the development of future surveillance systems.

The threat of emerging diseases and the growing burden of chronic diseases requires integrated approaches to surveillance. Analysis of disease trends in space-time provides context which can be linked to possible risk factors in a research environment, flag unusual events in an automated surveillance system, and provide epidemiologists with current information during an outbreak. Well-studied and understood methods are required to ensure appropriate use and transparent and reproducible results. The literature on statistical surveillance is extensive and provides this basis, yet software implementations are far from standardized. As space-time surveillance statistical methods mature further, software is also surely to improve. The open-source environments, such

as R, may be the optimal venue for future development of surveillance software as they afford easy integration with many statistical and mapping packages, and being open-source, the underlying code can be viewed and modified easily. However data structure remains a major issue when handling space-time data, especially when data has to be moved between different software packages. Standardized space-time data classes in R or another open-source environment may be a fruitful area of development.

4.7 Acknowledgements

This project was supported in part by the Teasdale-Corti Global Health Research Partnership Program, National Sciences and Engineering Research Council of Canada, and GeoConnections Canada.

Table 4.1. List of software packages for review of space-time disease surveillance software.

Software Package	Source	Reference	Description
SaTScan 8.0	http://www.satscan.org	Kulldorff and Information Management Services 2009	Cluster detection software with several spatial, temporal and space-time scan statistics.
ClusterSeer 2.3	http://www.terraseer.com/	Jacquez et al. 2002	Cluster analysis software includes many methods for spatial, temporal, and space-time analysis.
GeoSurveillance 1.1	http://www.acsu.buffalo.edu/~rogerson/geosurv.htm	Yamada et al. 2009	Implementation of cumulative sum surveillance statistics.
Surveillance package 1.1-2	http://cran.r-project.org/web/packages/surveillance/index.html	Höhle 2007	Package for statistical surveillance includes test-based and model-based methods.

Table 4.2 Criteria and review approach for review of space-time disease surveillance software

Criteria	Review
Data preprocessing	Number of steps involved to process a point event (cases) shapefile and a polygon census shapefile (population)
Methods	Description of methods offered by each program
Technical issues	Speed of computation, system stability, automation, operating requirements
Analysis output	Output options (graphs, maps, reporting)
User facility	Qualitative assessment rated on scale of 1 – 5 on each of: • Ease of learning • Use • Set up • Documentation / Help

Table 4.3. Data preprocessing steps for each software package to perform a space-time analysis starting with daily data as point events in an ESRI point shapefile and a polygon shapefile of census dissemination area boundaries.

Software	Type of Analysis	Required Data Structure	Data Preprocessing Steps
SaTScan	Space-time cluster scan with Poisson model	• Case file with number of cases, date, and DA id • Population file with population, date, and DA id • Coordinates file with DA id, centroid X and Y coordinates	• Associate DA identifier with each point event • Calculate DA centroid coordinates
ClusterSeer	Space-time cluster scan with Poisson model	• One table with population • One table with counts of cases for each location and date during study period	• Associate DA identifier with each point event • Calculate week numbers • Aggregate cases by week for each DA (zero counts included)
GeoSurveillance	Univariate cusum on individual DAs	• DA shapefile with counts of number of cases for each time period named and ordered sequentially in the table	• Calculate week numbers • Split point events into unique shapefiles for each week • Count number of events in each DA by week (zero counts included) • Calculate weekly counts as new fields
R-Surveillance	Univariate cusum on individual DAs	• Matrix of counts of cases with spatial locations as columns and time periods as rows	• Calculate week numbers • Split point events into unique shapefiles for each week • Count number of events in each DA by week (zero counts included) • Calculate weekly counts as new fields • Read table into R as matrix and transpose

Table 4.4. Comparative review of software packages for space-time disease surveillance: User Facility.

Software	Learning	Use	Set Up	Help / Documentation	Comments
SaTScan	4	5	5	4	Requires knowledge of scan statistics. Basic analysis is straightforward though many advanced options available. Well referenced methodology in the user guide.
ClusterSeer	5	5	3	5	Excellent documentation and learning resources for the many different methods. Data format requirements can be cumbersome.
GeoSurveillance	3	3	3	3	Data structure is peculiar, though the basic user interface is straightforward. Documentation not integrated within the menu itself.
R – Surveillance	1	3	5	2	Command driven system requires knowledge of R language. Examples are easy to replicate. Very easy to install within R. Documentation is not extensive.

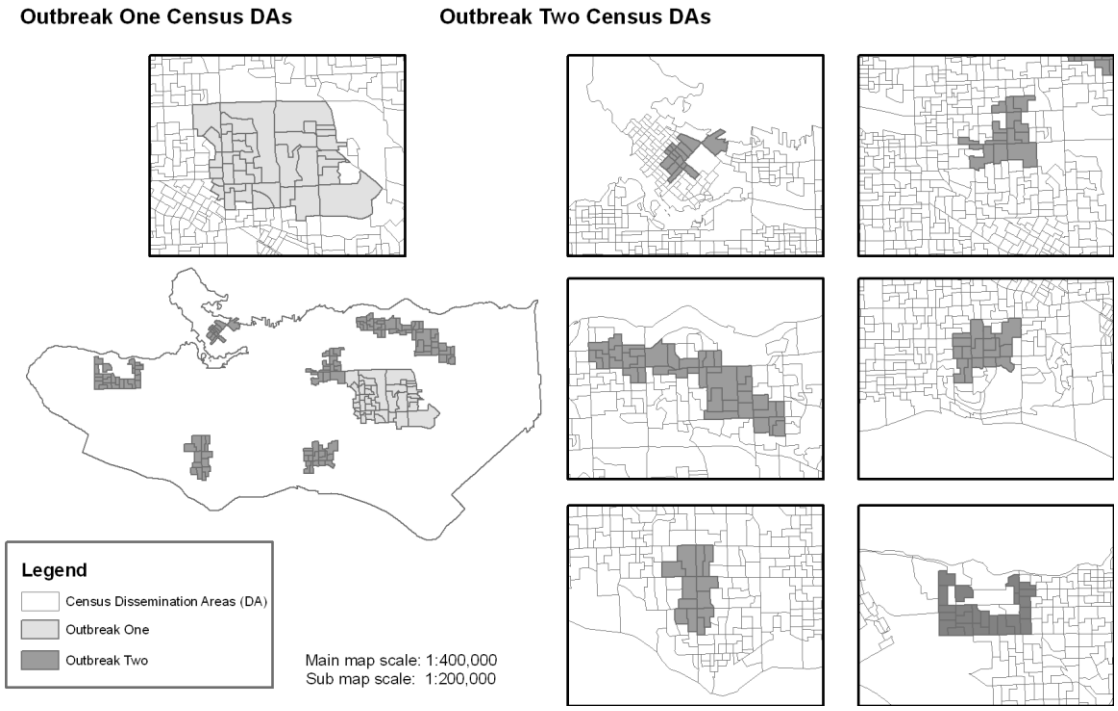


Figure 4.1. Outbreaks simulated to review software packages for space-time disease surveillance (Outbreak one – light grey; Outbreak two – dark grey). Outbreak one consisted of one large compact cluster. Outbreak two was composed of several clusters occurring at different times throughout the region.

Chapter 5: A hidden markov model for analysis of frontline veterinary data for emerging zoonotic disease surveillance

5.1 Abstract

Surveillance systems tracking health patterns in animals have potential for early warning of infectious disease in humans, yet there are many challenges that remain before this can be realized. Specifically, there remains the challenge of developing early warning signals for diseases that are not known or are not part of routine surveillance for named diseases. This paper reports on the development of a hidden markov model for analysis of frontline veterinary sentinel surveillance data from Sri Lanka. Field veterinarians collected data on syndromes and diagnoses using mobile phones. A model for submission patterns accounts for both sentinel-related and disease-related variability. Models for commonly reported cattle diagnoses were estimated separately. Region-specific weekly average prevalence was estimated for each diagnoses and partitioned into normal and abnormal periods. Visualization of state probabilities was used to indicate areas and times of unusual disease prevalence. The analysis suggests that hidden markov modelling is a useful approach for surveillance datasets from novel populations and/or having little historical baselines.

5.2 Introduction

Approximately 75 percent of emerging infectious diseases (EIDs) in people are estimated to have originated in animals (i.e., zoonoses) (Gregor 2007; Jones et al. 2008). Strategies to limit the impact of zoonotic EIDs can be broadly categorized as intervention at one or more of three levels: (i) controlling infections in people; (ii) blocking transmission of pathogens from animals to people; and/or (iii) preventing or controlling disease in

animals (Hayden et al. 2002). Despite significant effort and funds targeting the first strategy, the global public health community continues to be caught off guard by EIDs. It is now recognized that the third strategy, control of disease in animals, may hold considerable potential for prevention of zoonotic EIDs (Rabinowitz et al. 2008). To achieve this strategy, early detection of disease in animals is critical.

Surveillance for EIDs is confronted with the challenge of tracking something that has not yet happened. This has lead to the development of methods to track indicators of emergence or outbreaks such as risk factor surveillance and syndromic surveillance. Surveillance systems using novel (pre-diagnostic) data sources that track healthcare-seeking behaviour have become widespread in human health surveillance with an aim to detect both intentional (bioterrorist) and naturally-occurring infectious disease outbreaks. Data representing early stage disease related behaviours (e.g., staying home from work – absenteeism data) may therefore have predictive value and promote detection of disease at the earliest possible stage. However similar data is generally not available for animals. EID surveillance systems rely on pre-diagnostic, syndromic, or clinical diagnoses to gather early warning signals (). Syndromic surveillance for early outbreak detection often uses automated data collection and ongoing analysis for statistical signals to monitor patterns in health outcomes in near real-time to detect early signals of diseases outbreaks (Van Metre et al. 2009; Leblond et al. 2007). Analysis of conditions frequently seen by field veterinarians but rarely recorded or tracked can be thought of as similar to a syndromic surveillance approach, in that the data represent novel and unknown populations and may have early warning value for emerging diseases. The data presented in this study is from a system which recorded clinical diagnoses of field veterinarians

(Robertson et al. 2010). This system was developed as a prototypical complementary system to national disease reporting in Sri Lanka.

One of the drawbacks of pre-diagnostic, syndromic and clinical diagnostic is that they incur an increased chance of false alarms (Stoto et al. 2004). With pre-diagnostic data sources, the data do not represent actual cases of disease, but variables related to disease - such as over-the-counter pharmaceutical sales (Das et al. 2005), web site queries (Hulth et al 2009), or ambulance dispatch records (Mostashari et al. 2003). Such data sources exhibit non-disease related variations that need to be adjusted for in order to establish an accurate baseline level of risk. Similarly, clinical diagnoses data exhibit unknown variations that relate to how the data are collected. In many instances, making these adjustments is straightforward. For example, day of the week effects – that is, higher rates on certain days of the week - are features of many types of surveillance data. These higher rates could contribute to an outbreak signal when really the factors driving the increase are unrelated to disease, such as the greater propensity for people to visit the doctor on Mondays as compared to Fridays. With veterinary sentinel data, variability may be dependent on the sentinels themselves rather than the disease process. Therefore, with new and poorly understood surveillance data sources, developing a detailed understanding of baseline patterns (i.e., normal variation) is essential prior to conducting statistical analysis for cluster or outbreak detection.

Public health is increasingly looking towards surveillance of changing disease patterns in animals to enhance prediction and understanding of where and when EIDs in humans are likely to occur. Prediction of pre-emergence changes in pathogen dynamics in animals may hold the greatest potential of early detection in humans, and is therefore a

central goal of EID surveillance (Kahn 2006). A major challenge however, is the collection of appropriate data on animal health/behaviour (Vrbova et al. 2008). For livestock populations, veterinarians may serve as an important source of information. However, using veterinary clinical diagnoses instead of results from diagnostic laboratory tests, the traditional data source in animal health surveillance, carries similar inherent risks to novel data sources in human surveillance systems: false alarms and unknown baseline variations.

There have been rapid advances in the development of appropriate methods of analysis for surveillance data (Lawson and Kleinmann 2005; Wagner et al. 2006; Sonesson and Bock 2003). The detection of clusters in time (Naus 1965), space (Kuldorff and Nagarwalla 1995), and space-time (Kuldorff et al. 2001; 2005) are now routine analysis run in many surveillance systems (e.g., Heffernan et al. 2004). The majority of methods for cluster detection can be classified as hypothesis tests that evaluate a count or rate of some disease or syndrome within a subset defined by space/time, against some expected value estimated to be the normal state of the process. An alternate class of methods focus on estimation of the expected value using statistical models. A modelling approach can incorporate complex patterns; known demographic risk factors such as age and occupation, or environmental risks such as sources of pollution that affect disease outcomes. Models have been used widely in influenza surveillance to account for seasonal dynamics (Andersson et al. 2008), as well as long-term trends in retrospective analysis of chronic diseases (Xia and Carlin 1998).

Hidden markov models (HMM) have recently been developed for disease surveillance applications (Le Strat and Carat 1999; Rath et al. 2003; Madigan 2005;

Martinez-Beneito et al. 2008; Wall and Li 2009; Watkins et al. 2009). A Markov model can be used to examine the probability of transition from one state (e.g., normal) through a series of transitions to another state (e.g., abnormal) by considering the probability of the initial state and the transition probabilities from one state to another. In HMM, data are assumed to model a hidden or unobserved process with a fixed number of discrete states. An observed dataset related to the hidden process is analyzed at intervals to determine the state of the latent process. A first-order Markov assumption in the state accommodates the time-dependency in observed data. The state often relates to separate distributions for the observed data (in disease surveillance applications, often counts of cases of disease - Watkins et al. 2009). In health surveillance applications, states can describe the overall condition of the target population such as 'endemic' and 'epidemic', or 'normal' and 'flu season'. A transition probability matrix governs transitions between states. An advantage of HMMs for surveillance is that historical data is not required to train the model. Inferences about each of the states can be learned directly from available data, and in a Bayesian setting, the prior distributions. This is an attractive feature for new surveillance systems with short durations such as IDSAS as we do not know if any outbreaks actually occurred during the study period. The HMM approach to surveillance modelling offers a convenient framework for thinking about surveillance and novel data sources. If the aim of a surveillance system is to detect things that are "unusual", such that change preceding alteration in disease occurrence can be recognized and acted on quickly, the concept of process 'state' fits well within a decision-making framework. When changes are recognized quickly in the status of an animal population, this is of

value both inherently, to limit the impact of on the animal population, as well as a potential signal for human infection risk.

In the first application of HMMs to surveillance, Le Strat and Carat (1999) demonstrated a Poisson HMM for poliomyelitis that estimated weekly counts of cases at the national level as a mixture of two Poisson distributions. Recent examples of HMMs being used in disease surveillance include healthy and unhealthy states related to health services utilization from medical insurance data (Wall and Li 2009) and outbreak and non-outbreak states of influenza (Rath et al. 2003).

In this paper, we report on a study investigating baseline patterns in a newly established animal-based infectious disease surveillance system in Sri Lanka (Robertson et al. 2010). Data were collected for a period of a year describing clinical diagnoses of cattle, buffalo and poultry, in four regions of Sri Lanka. Field veterinary surgeons employed by the Department of Animal Production and Health submitted surveys via mobile phone to a central database. As these data describe syndromes and diagnoses not formerly tracked in Sri Lanka, there are no validation data available. We employ a modelling approach to examine different features of the data using hidden markov models (Le Strat and Carrat 1999). The objectives of the current study were to determine the sources of variation in animal-based EID surveillance in Sri Lanka, establish baseline rates for overall surveys, and explore spatial and temporal variability in commonly reported cattle diseases.

5.3 Methods

5.3.1 Data sources

The Infectious Disease Surveillance and Analysis System (IDSAS) was established in January 2009 as part of a collaboration between the authors and the Department of Animal Production and Health in Sri Lanka (Robertson et al. 2010). The system tracked syndromes and clinical diagnoses in cattle, buffalo, and poultry, in four districts of Sri Lanka. Forty government-employed field veterinary surgeons (FVS) from four administrative districts (Figure 5.1) participated as data collectors using mobile phone-based surveys coupled with global positioning systems (GPS). FVSs were instructed to submit surveys via email to a central surveillance database daily (with a suggested minimum rate of 2 surveys per week). The data used in the present study represent the period January 1st 2009 to December 31st 2009.

Each survey submitted by a FVS represented one visit to a farm or one examination in clinic of at least one of the three species. Surveys were classified by routine visits (yes/no) and presence or absence of an animal health issue. In the case of an animal health issue, cases were given a syndrome group and a clinical diagnosis. FVSs also had the option of classifying the cause of the health issue as unknown. There were a total of 17 syndrome groups for cattle and buffalo and 11 for poultry. Options for suspected diagnoses were based on the syndromic grouping selected. For example, under "lameness", possible diagnoses included Blackquarter, Footrot, Osteomyelitis, as well as 22 others. Each FVS was responsible for one geographic area called a range, so geographic locations could be associated with each survey. Farm level spatial data collected with GPS were not used in this analysis as we were primarily interested in determining broad-scale sources and patterns of variation in the IDSAS data.

Auxiliary data were collected to help account for non-disease variation in IDSAS data FVS-specific information such as sex and the number of years since graduation from veterinary school was collected when the FVS was enrolled in the project. There were also specific dates when re-training was conducted and indicator variables were used to represent these periods. The retraining sessions increased enthusiasm and participation levels of the FVSs as sharp increases in submissions were noted in exploratory analysis of the data (Robertson et al, 2010). These factors represent what we term a *sentinel process*; factors related to the FVS as disease sentinels, rather than disease.

We obtained monthly temperature and precipitation data as district averages from the Sri Lankan Department of Meteorology as disease patterns in animals are often seasonal and may therefore exhibit a relationship with local weather patterns or seasons.

5.3.2 Analysis of surveillance data

In this study, we model animal health conditions as seen by FVSs in Sri Lanka. We extend on the spatial Poisson HMM for disease surveillance given in Watkins et al. 2009, by simultaneously accounting for covariates impacting the observed data. The data collected by IDSAS can be conceptualized as deriving from two independent processes, the *sentinel process*, and the *disease process* (Figure 5.2). We were interested in accounting for variability related to the sentinel process, in order to learn more about variability related to disease during the study.

We formulated a two-state HMM where state one represents ‘normal’ conditions, and state two represents ‘abnormal’ conditions. The states were modelled as a mixture of two Poisson distributions. Or rather, the Poisson rate could take one of two values

depending on the current state of the HMM. The count y_{it} of submissions to the IDSAS system for week t by FVS i is modelled as follows:

$$y_{it} \sim Pois(\lambda_{c_{it}}) \quad (1)$$

where c_{it} is the state variable representing 'normal' ($c_{it} = 0$) or 'abnormal' ($c_{it} = 1$) animal health. The rate for the normal state was constrained to be less than that for the abnormal state, as we conceptualize 'abnormal' to indicate higher than normal submissions. We present the model for total submissions without covariates as HMM_1.

Covariates can be included in HMMs in two ways. They can link to the Poisson parameter for each count (i.e., covariates on state-dependent probabilities), where a stationary transition probability matrix is preserved. Alternatively, covariates can be incorporated into an HMM via the transition probability matrix itself (Zuchinni and MacDonald 2009 pg 126-127), resulting in an inhomogeneous HMM. For example, Wall and Li (2009) present a HMM for medical service utilization data where covariates relate to transitions between healthy and unhealthy states via a logistic regression. In the model here, the former approach is adopted, maintaining a stationary probability matrix. Covariates were included in the model by relating each Poisson intensity to μ_{cit} indicating the mean rate, and a vector of FVS (i.e. spatial) and time specific coefficients β and covariates X, via a log-link Poisson regression:

$$\log(q_{it}) = m_{c_{it}} + b\,\mathsf{C} \quad (2)$$

The mean rate, or intercept, is 'switched' between states based on the current state of the Markov chain, which is governed by stationary transition matrix p. The transition

probabilities are treated as unknown parameters that are estimated by the model. Following the parameterization in Watkins et al. (2009), a Dirichlet prior distribution for initial probabilities, and gamma priors on subsequent probabilities were employed. An outline of prior distributions for model parameters is given in Table 1.

Finally, spatial information was included in a final model. The four districts where IDSAS was operational were selected primarily to capture variation in environment, climate, and agricultural practices. For true outbreaks of disease or changes in pattern of disease, we might expect similar submissions among FVSs in the same district. To account for similarity of conditions within district versus other districts, submissions from FVSs in common districts were summed. The count y_{it} of submissions for FVS i at time t was added to counts for all FVS in the same district.

$$y_{it}^{*} = \sum_{\substack{j=1 \\ j \neq i}}^{40} y_{jt-1} D_{ij} \qquad (3)$$

where D_{ij} is an n x n matrix with 1s indicating FVSs in the same district and 0 otherwise. This information was included in a temporally lagged variable, representing the count of district wide submissions in the previous time period. We report results for the model with covariates included as HMM_2.

All models were run on the individual submission counts to generate an understanding of the factors affecting the IDSAS data. To investigate the patterns of individual diseases, the four most frequently reported suspected diagnoses in cattle were investigated. Cattle are one of the primary livestock species assessed and treated by FVSs in Sri Lanka and as such constituted the majority of submissions. For the disease-specific models, covariate effects for sentinel-level variables were taken from estimates

from the total submissions model, as we expect to these be constant factors effecting submissions equally. Disease-related variables (temperature, precipitation, and temporally lagged district-wide submissions) were estimated separately for each disease. Additionally, because natural disease prevalence varies by district, each district has separate mean rates for normal and abnormal states.

Models were implemented in a Bayesian setting via markov chain monte carlo (MCMC) sampling in WinBUGS (Lunn et al 2000). Bayesian modelling is a convenient choice for developing HMMs because sensitivity to distributional assumptions can be easily assessed, and a full probability distribution is obtained for model parameters in the posterior distribution. In all analysis, two MCMC chains were run for a 1000 iteration burn-in and 4000 iteration run. Convergence for parameters was assessed with both visual inspection of the posterior sampling history, and the Gelman-Rubin statistic, which is the ratio of within chain variability to between chain variability. When the Gelman-Rubin statistic is near one convergence is assumed to be reached.

Results for the state variable are reported for two thresholds. The posterior mean state for each FVS / week pair (a total 2080) yield values ranging from 1.0 for 'normal' to 2.0 for 'abnormal' and values in between. We set a lower threshold of 1.50 to define membership in state two, and an upper threshold of 2.00. In all modelling results reported, coefficients with 95% credible intervals covering zero are excluded.

5.3.4 Simulation study

A simulation study was developed to evaluate model performance. Data from two Poisson models were simulated onto a 10x10 spatial grid representing disease-reporting units in a hypothetical surveillance system (n = 100). Three covariates were also simulated for each area. The normal state (i.e., state 1) Poisson model was as follows

$$\lambda_{it} = \exp(1.8 + 1.3X_1 + 3X_2) \quad (4)$$

and the abnormal state model (i.e., state 2) was

$$\lambda_{it} = \exp(2.7 + 1.3X_1 + 3X_2) \quad (5)$$

Relationships for covariates X_1 and X_2 were the same between states but the intercept shifted from 1.8 during the normal state to 2.7 in the abnormal state. The purpose of the model is to detect shifts in state based on observations and simultaneously characterize the relationships between the mean and the covariate variables. We also evaluated whether the model could determine different covariate effects in different states, by changing the abnormal state model to include a third covariate:

$$\lambda_{it} = \exp(2.7 + 1.3X_1 + 3X_2 - 1.6X_3) \quad (6)$$

In the simulation study analysis, spatial information (neighborhood relationships) was not used, but could easily be incorporated through a conditional autoregressive random effect, pooling observations from neighbouring areas, or including region-specific dummy variables.

The normal state model was used to generate counts for 52 time periods (i.e., one year at weekly intervals) based on a normal distribution with a mean determined by Equation 4 and a standard deviation of 1. Different types of spatial patterns (outbreaks 1-5, see Figure 5.3) were created to establish areas where counts were replaced with counts estimated from the abnormal state model (Equation 5). Thus distinct spatial areas and time periods where counts and covariates in state two were created against a baseline of state one. In the second scenario, estimates for the abnormal state were obtained from Equation 6. Model performance was then evaluated as the percentage of correctly classified states.

5.4 Results

5.4.1 Simulation study

The HMM model correctly classified 99.7% of the observations in the shifted intercept scenario. Out of 5100 (51 time periods x 100 spatial units) observations (first week is not used because inference is based totally on initial values), 5088 were classified with the correct state. The 12 incorrectly classified states all occurred in outbreak five (see Figure 5.3), where all units were in the abnormal state, so all were errors of omission (i.e., incorrectly classified as normal). All parameters converged, and model diagnostics indicated a parsimonious model. The coefficient estimates were similar to the true values for both variables, though the mean for the normal state was slightly underestimated (Table 2). In contrast, in the scenario with shifted mean and the addition of a third covariate effect in the abnormal state model, the model failed to converge completely. Posterior estimates for the intercept and covariate $X1$ were similar and converged (not reported), however estimates for coefficients on $X2$ and $X3$ both failed to converge. The model was run for 20,000 iterations.

5.4.2 Animal health surveillance submission patterns

During the study period, there were a total of 5758 submissions to the IDSAS system reporting an animal health issue on the three included species. The HMM_1 without covariates yielded a total of 753 abnormal events during the study period based on a posterior mean threshold of greater than 1.5. When constrained to a higher degree of certainty (posterior mean threshold of 2.00), the number of abnormal events was 390 (Table 3). The mean submission rate for state one was 0.45 submissions per FVS, per week, and in abnormal periods the mean rate was 6.72. When covariates were added to the model (HMM_2), the number of abnormal events increased to 870 and 450 for the two

threshold levels, while mean rates adjusted to 0.34 and 6.65 submissions per FVS, per week for state one and two respectively. Covariate effects are reported in Table 3. Positive association with submission rates was limited to the variable indicating training periods, while covariates identifying male and less experienced FVS were negatively associated with submissions. Precipitation, temperature, and district reports had no effect in the total submissions model. The temporal pattern of abnormal events relative to all submission counts for each FVS are outlined in Figure 5.4. Using the upper threshold, the submission counts for state one ranged from zero to six, and from four to 103 for state two. The count densities plotted on a log scale are presented in Figure 5.5.

5.4.3 Commonly reported cattle diseases

In total, there were 3943 reported cattle cases during the study period. The most commonly-reported diagnoses in cattle were mastitis (543), ephemeral fever (234), babesiosis (212), and milk fever (210). Monthly cases for each of the districts is given in Figure 5.6, along with environmental variables maximum temperature and total monthly precipitation.

Model results for the four most common diagnoses are outlined in Table 4. As noted earlier, coefficients for sentinel-level variables were set as estimated in the total-submission model, and only covariate effects for temperature, precipitation, and district reports were estimated for disease-level models. Overall, the effects of the covariate variables in disease-level models were minimal, with rate ratios ranging from 0.93 to 1.10. Temperature was positively associated with reported diagnoses of all diseases. Precipitation was not associated with diagnoses of any of the four diagnoses. Temporally-

lagged district reports were negatively associated with mastitis, babesiosis, and milk fever, and positively associated with ephemeral fever.

The posterior mean states are presented in Figure 5.7 for each of the four main disease categories. A possible outbreak of ephemeral fever is evident in Anuradhapura towards the end of the study period. Other periods of high submissions for babesiosis, milk fever, and mastitis are found in the Nuwara Eliya district.

5.5 Discussion

Variation was modelled in data submitted to a mobile-phone infectious disease surveillance system in Sri Lanka. Results indicate that submission varied according sentinel level factors, and that HMMs are a convenient methodology to approach novel sources of surveillance data. The average submission rate for surveys varied by district, from 0.34 surveys per week during normal periods, to in 6.34 surveys per week during abnormal periods. The number of abnormally high submissions increased when covariates were added to the model. Baseline estimates for normal patterns of mastitis, babesiosis, and milk fever were highest in Nuwara Eliya, the main cattle-dairy region in Sri Lanka. The baseline estimate for the normal pattern of ephemeral fever was highest in Anuradhapura, a region which experiences seasonal droughts.

The number of new pathogens in animals and humans are increasing and known infections are changing in pattern as natural and social systems adapt to changes in climate. The role of animals in emergence of new diseases is widely recognized (Rabinowitz et al. 2008), and surveillance of EIDs via animal-based systems such as IDSAS holds potential for detection and response at an early stage, yet studying this in the absence of an actual EID is a major challenge. While detecting an EID was the goal

of the IDSAS system, enhanced understanding of the pathogen distribution as seen by veterinarians in the field represents an opportunity to both establish what is normal, and subsequently detect patterns that are unusual. This alone may be enough information to develop processes to inspire further action and promote early detection (Gubernot et al. 2008). Further, the improved timeliness of IDSAS data as compared to laboratory testing is another attractive feature of using clinical diagnoses data for EID surveillance.

As this analysis has demonstrated, there are complex variations driving surveillance data using novel sources such as field-based veterinary surveys. In Sri Lanka, sentinel process factors such as the sex and work experience of the submitter impact submission rates, as did periodic disruptions due to training and/or political events. The advantage of a modelling perspective to surveillance is that these sources of variation can be partitioned out in order to generate a finer understanding of the disease process. However, there is also value in learning about the sentinel process. This type of methodology could be used within ongoing surveillance systems to identify sentinel characteristics more common amongst high submitters, and therefore serve to inform the sentinel selection process and ongoing sentinel inclusion or exclusion. In addition, exploration of the factors driving temporal variation in submissions can help to guide sentinel retraining and electronic prompts reminding sentinels to submit data.

When examining the results of the model HMM_1 on total submissions, we note a high number of abnormal events. When variables are included in HMM_2, the overall effect of the important variables actually reduces expected mean submissions, which results in more 'unusual' events. The question becomes, what is the value of accounting for sentinel-level factors. Given that alerts generated by surveillance systems typically

overwhelm the number that can actually be investigated (Fearnley 2008), should adjustments be biased downwards? The analysis here suggests that adjustments are useful because they provide a more complete understanding of the processes generating the surveillance data. In the context of sentinels for disease surveillance, this might simply be helping to identify characteristics that predict a more engaged sentinel relative to others. Another issue is that variables such as sex and experience may have an overall effect, but cannot be attributed to individuals. While states are discrete, state probabilities can be visualized across space and time as in Figure 5.7, providing visual evidence of gradual changes after covariates have been taken into account.

The simulation study presented here provides evidence that the model performs well under the scenario where a shift in the mean occurs and covariate effects remain fixed. In this simulation scenario, both the mean and covariate effects were recovered well by the model. This analysis lends support to the results obtained from IDSAS data. We might therefore be able to conclude overall, the means detected for each state-district combination in the disease-level models represent baseline estimates of the weekly prevalence of these diseases as seen by FVSs based on clinical diagnoses and syndromic groupings. However, the low values for these estimates (Table 4) make interpretation somewhat cumbersome. The state one means for all four diseases range from 0.09 for babesiosis in Anuradhapura to 0.32 for milk fever in Nuwara Eliya, while state two means ranged from 0.63 for babesiosis in Matara, to 3.51 for babesiosis in Nuwara Eliya. It is important to quantify the differences in means between the districts for the different diseases as it provides a starting point from which to understand why these differences exist.

In developing this technique we chose to examine the four most frequently suspected diagnoses in cattle. However there are marked differences between babesiosis and mastitis in terms of epidemiology, etiology, and clinical presentation that are worth highlighting. Babesiosis is a tick-borne disease most commonly characterized by fever, inappetance, lethargy, weakness, red-tinged urine (hemoglobinuria), anemia and jaundice, though many cases are asymptomatic. Recovered cases become asymptomatic carriers, and duration of infection can be up to years. There is a large degree of variability in susceptibility between cattle breeds. Transmission of babesiosis is dependent on a bite from an infected Ixodes tick, and patterns in disease prevalence in cattle are dependent in part upon the prevalence of Babesia spp. in the vector and in the prevalence of the tick species itself (Bock 2004). In contrast, mastitis, defined simply as inflammation of the udder, can be caused by a variety of bacterial and fungal pathogens. It is often characterized by a drop in milk production, and when clinically evident may be accompanied by gross changes to milk or systemic illness. It can be caused by both contagious and environmental pathogens. Incidence and prevalence is impacted by a variety of individual animal characteristics, as well as environmental variables. Many cases are asymptomatic (Radostits, 1994). Given these differences, it is worth considering whether examination of their occurrence using the same method is appropriate, and whether covariates should be fixed across suspected diagnoses.

Visualizing the probability of state two in Figure 5.7 on a FVS/weekly basis provides some evidence for the stability and confidence in the model inferences. The outbreak of ephemeral fever in Anuradhapura is on face value, more unusual, than for example patterns of mastitis in Nuwara Eliya. This is because based on what we know

about ephemeral fever, transmitted by biting insects and often highly correlated with periods of rain, we expect, and are more concerned with ‘outbreaks’, than for mastitis, which is an endemic and pervasive condition. However, outbreak levels of mastitis may in fact represent clusters which also represent another, possibly unknown pathogen. The goal is to understand and establish the normal pattern for the population, so that unusual events can be quickly spotted and explored.

Models in general rest on assumptions of correct specification, when in practice, they are invariably missing important variables, and relationships often change over time in unforeseen ways. Visualizing patterns of the state variable over time provides a quick diagnostic tool to identify changes in pattern. Also, because we are working within a Bayesian setting and have a full distribution for model parameters, we can make similar plots for the model uncertainty using standard deviation. The modelling analysis here offers a robust framework for analysis of surveillance data with short temporal spans and multiple processes driving submissions, as is often the case with participant-generated data.

5.6 Acknowledgements

This project was funded in part by the Teasdale-Corti Global Health Partnership and the National Sciences and Engineering Research Council of Canada. IDSAS represents the culmination of a collaborative effort between stakeholders at the DAPH, the University of Peradeniya Faculty of Veterinary Science, and provincial levels of government in Sri Lanka.

Table 5.1. Description of prior distributions and hyper-parameters for model parameters.

Model	Parameter	Prior Distribution	Description
$HMM_{1,2}$	μ_1	*Normal*(0,0.01)*	Mean state 1
$HMM_{1,2}$	μ_2	*Normal*(0,0.01)*	Mean state 2
$HMM_{1,2}$	P_{Init}	*Dirichlet*(0.5,0.5)	Initial Probability
$HMM_{1,2}$	P	*Gamma(0.5,1)*	Probability transition matrix
$HMM_{1,2}$	Y	*Poisson(* λ *)*	Observed count data
HMM_2	X	*Normal*(0,0.001)*	Covariate coefficients

*Parameterized as mean and precision (1 / variance, as in WinBUGS). For disease-level models, a Normal(0,10) prior was used to accommodate very small expected counts.

Table 5.2. Model results from simulation study for five different outbreak scenarios occurring during a 52 week simulated surveillance system.

Parameter	True value	Posterior mean (95% credible interval)
μ_1	1.80	1.73 (1.71, 1.74)
μ_2	2.70	2.66 (2.63, 2.69)
X1	1.30	1.42 (1.29, 1.54)
X2	3.00	3.13 (3.03, 3.25)

Table 5.3. Submission pattern model parameter estimates reported as rate ratios.

Parameter	Posterior mean (95% credible interval)	Standard deviation
μ_1	0.34 (0.28-0.41)	0.10
μ_2	6.65 (6.05-7.29)	0.05
Training	1.19 (1.08-1.32)	0.05
Years	0.59 (0.55-0.63)	0.03
Male	0.90 (0.84-0.96)	0.03

Table 5.4. Model results for four commonly reported cattle diagnoses. Posterior mean estimates are per week, per field veterinary surgeon, reported as rate ratios. Maximum daily temperature and total precipitation are computed for each district and month. District reports are the number of cases within the district in the previous week.

	Anura-dhapura		Nuwara Eliya		Matara		Ratnapura				
Model	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	Temp-erature	Precip itation	District Reports
Mastitis	0.22	1.00	0.30	3.48	0.19	1.15	0.12	1.07	1.10	1.00	0.96
Ephemeral Fever	0.22	1.04	0.11	1.34	0.10	0.81	0.09	1.67	1.04	1.00	1.04
Babesiosis	0.09	0.79	0.24	3.51	0.08	0.63	0.13	1.05	1.10	1.00	0.93
Milk Fever	0.10	0.76	0.32	2.51	0.10	0.87	0.09	0.78	1.06	1.00	0.93

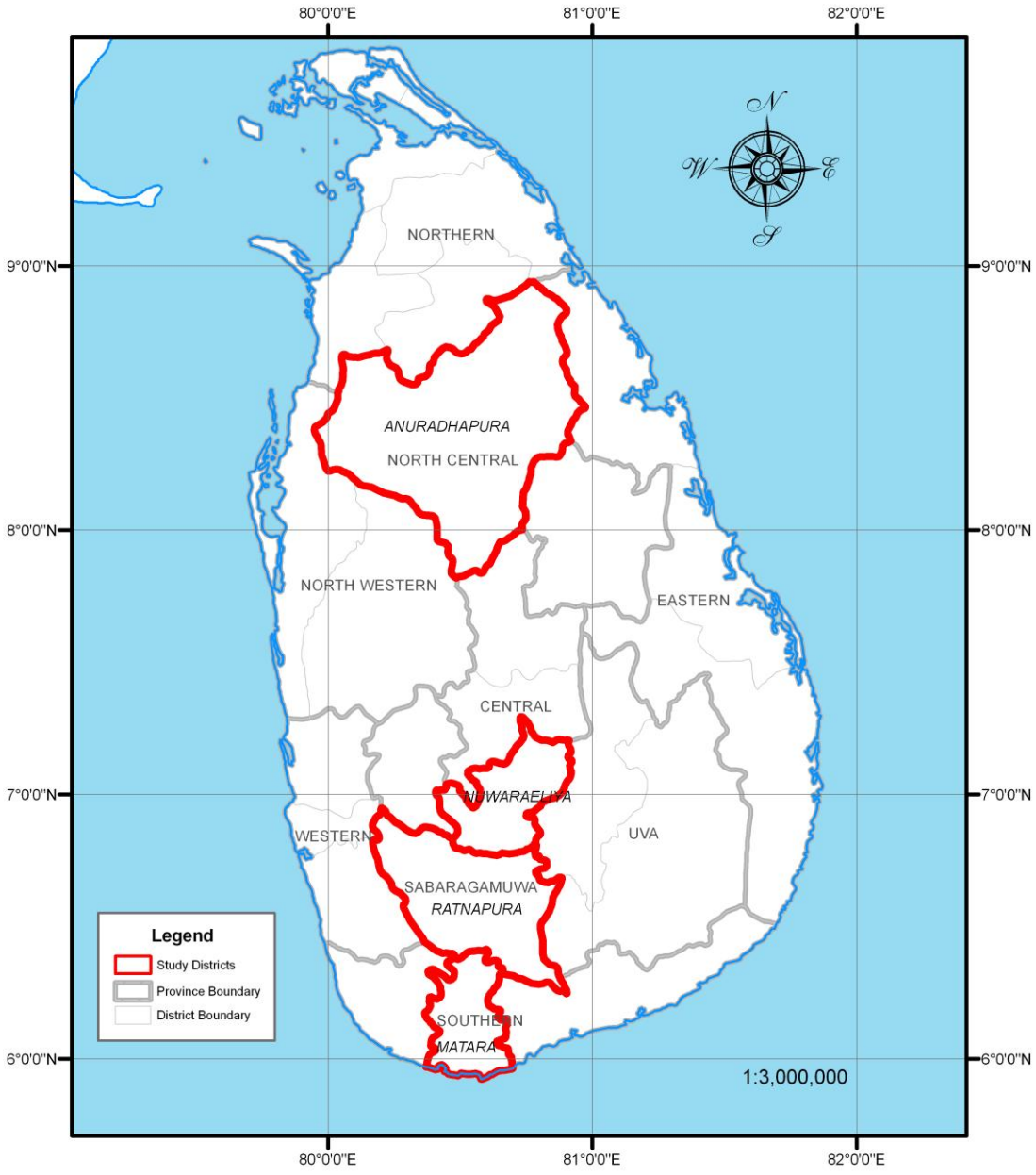


Figure 5.1 Map of Sri Lanka and study districts that were part of the Infectious Disease Surveillance and Analysis System.

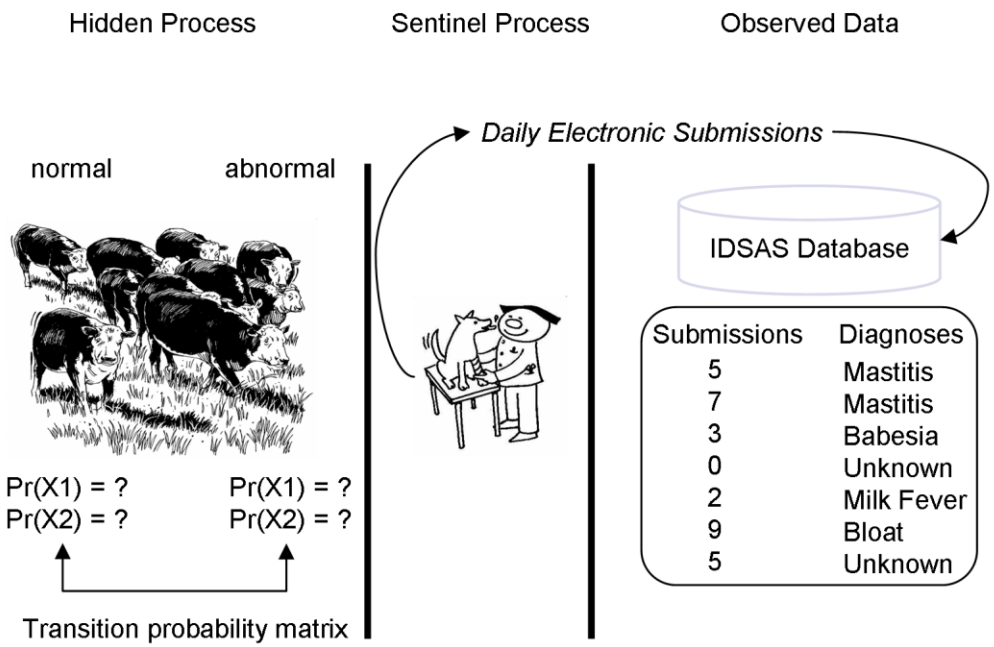


Figure 5.2. Conceptual model of data generating processes in the Infectious Disease Surveillance and Analysis System in the context of hidden markov models. The hidden states of interest are the normal or abnormal state of animal health as seen by field veterinary surgeons. Observed data may include weekly submission counts, or counts of specific reported diagnoses.

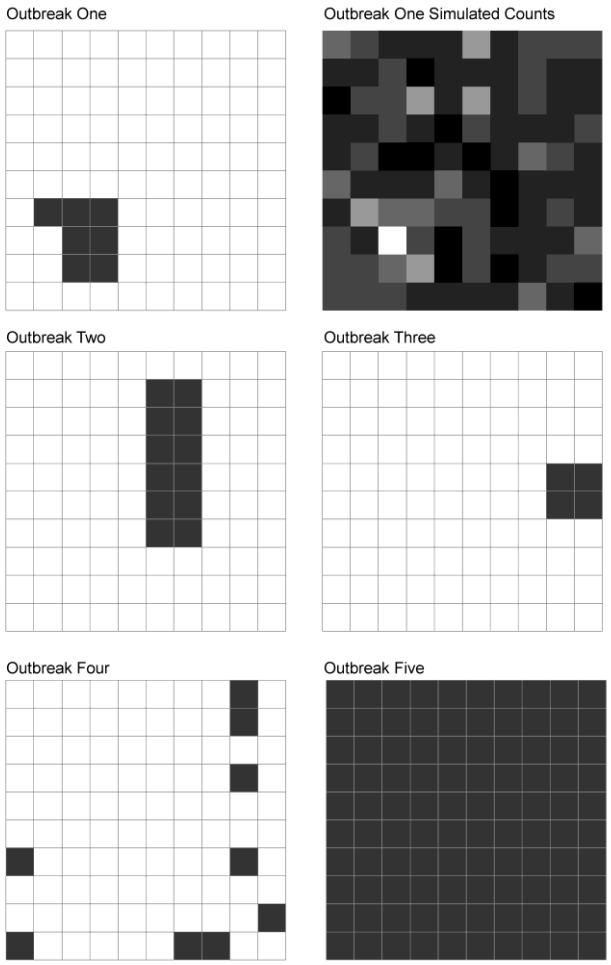


Figure 5.3 Simulated outbreak patterns in a hypothetical surveillance system: white cells generated under model for state one, and black cells generated under model for state two. The count data that was simulated using outbreak one is also shown: dark colours indicate low counts and lighter colours indicate high counts.

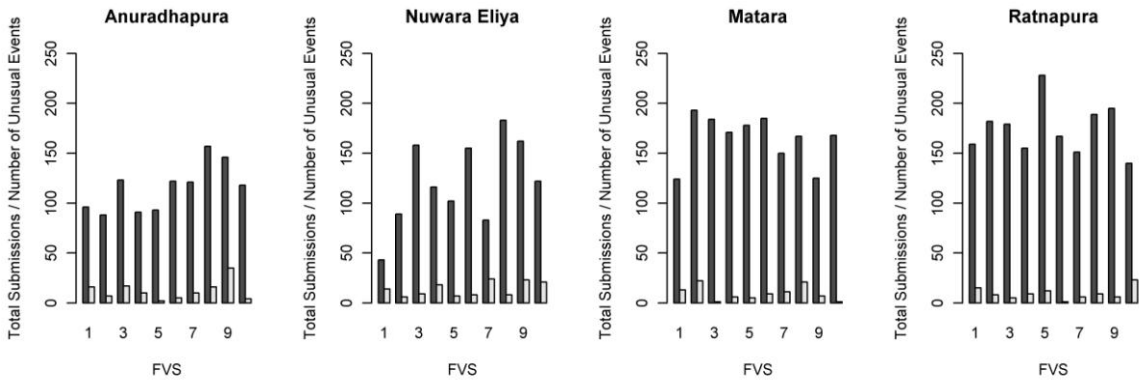


Figure 5.4 Total weekly submissions to the Infectious Disease Surveillance and Analysis System during the study period and the number of unusual states, by field veterinary surgeon and district. The number of weeks in state one is indicated in dark grey and the number of abnormal events in white.

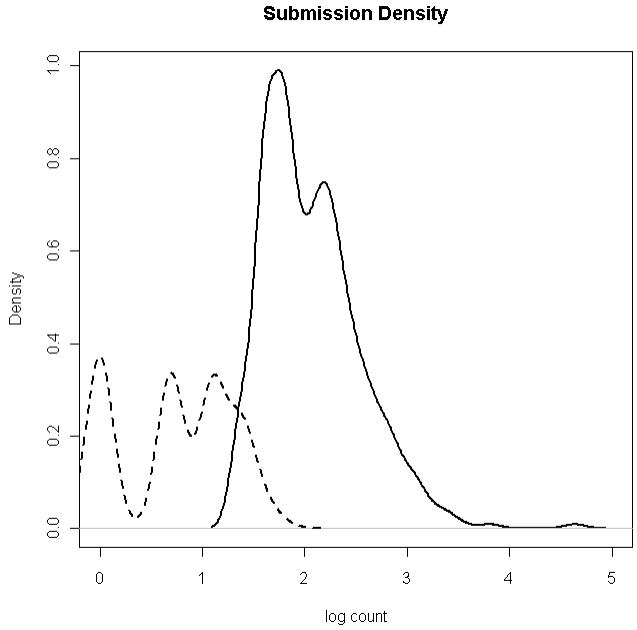


Figure 5.5 Density of the log count of submissions in state one (dashed) and state two (solid).

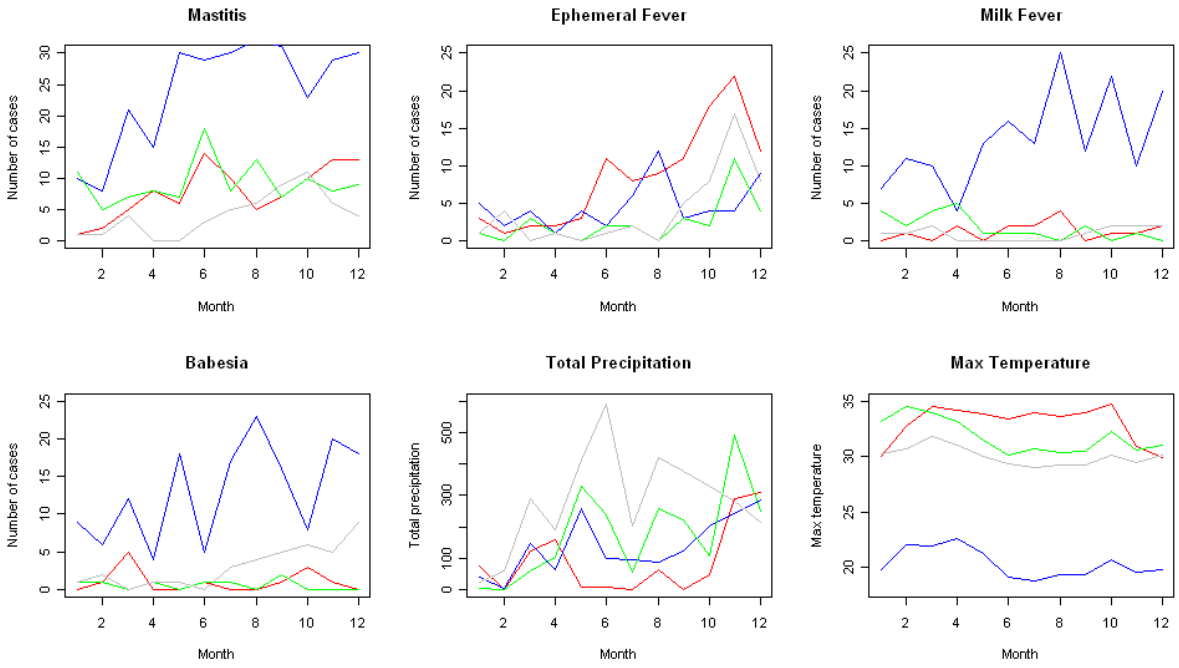


Figure 5.6 Monthly total cases for commonly reported diagnoses in each of the four districts: Anauradhapura (red), Nuwara Eliya (blue), Matara (green), and Ratnapura (grey). Monthly averages for district-wide total precipitation and maximum temperature.

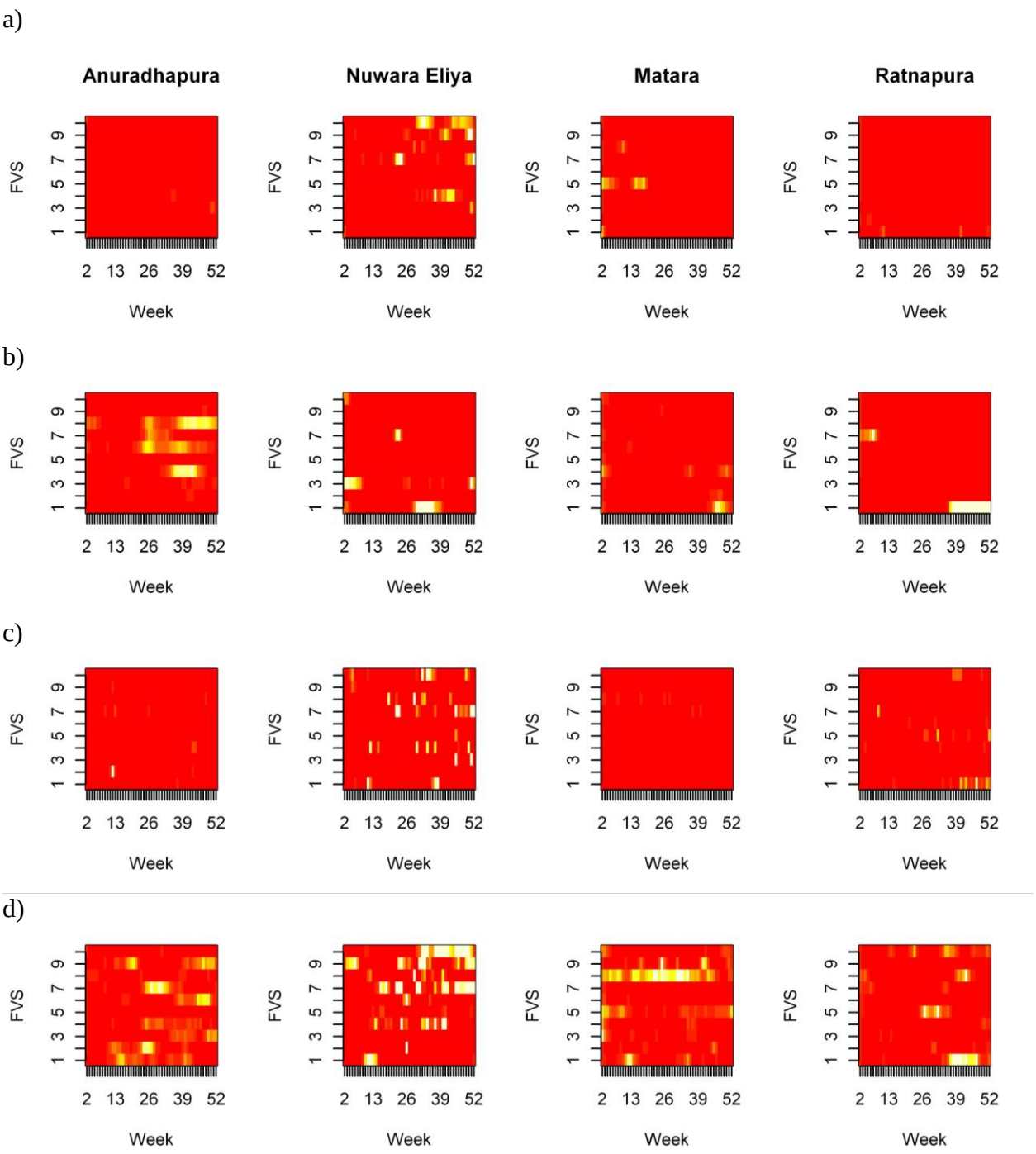


Figure 5.7 The model-adjusted posterior mean state for each field veterinarian surgeon by week, in each of the study districts for commonly reported cattle diagnoses. Red indicates state one and white indicates state two, and yellow intermediate values for a) Milk Fever, b) Ephemeral Fever, c) Babesiosis, and d) Mastits.

Chapter 6: Spatial epidemiology of suspected clinical leptospirosis in Sri Lanka

6.1 Abstract

Leptospirosis is one of the most widespread zoonosis in the world. A large outbreak of suspected human leptospirosis occurred in Sri Lanka during 2008. This study investigates spatial variables associated with suspected leptospirosis risk during endemic and outbreak periods. Space-time cluster scan statistics are combined with logistic regression modelling to test associations during outbreak periods. During the endemic period (2005-2007), leptospirosis risk was positively associated with shorter average distance to rivers and with higher percentage of agriculture made up of farms less than 0.20 hectares. Temporal correlation analysis of suspected leptospirosis cases and rainfall revealed a two-month lag in rainfall-case association during the baseline period. Outbreak locations in 2008 were characterized by lower distance to rivers and higher population density. Rice paddy density was not an important variable in either endemic or outbreak periods. The analysis suggests the possibility of household transmission in densely populated semi-urban villages as a defining characteristic of the outbreak. The role of rainfall in the outbreak remains to be investigated, though analysis here suggests a more complex relationship than simple correlation, possibly due to an intermediate relationship with rodent populations or other maintenance hosts.

6.2 Introduction

Leptospirosis is an emerging infectious disease of global importance affecting millions of people every year (Bharti et al. 2003; Pappas et al. 2008). As a water-borne zoonotic pathogen shed in the urine of infected animals, its prevalence in humans is highest where

people are exposed to contaminated surface waters (Levett 2001) . Animal hosts include dogs, cattle, pigs, sheep and wildlife hosts such as wild pigs (Sullivan 1974). The most common sources of infection in humans are domestic animals and rats (Levett 2001). Leptospirosis was previously considered mainly an occupational risk for people engaged in work requiring frequent contact with contaminated surface waters; common occupations associated with human infection included fish farming, rice and sugar agriculture, and workers in sewers and canals (Waitkins 1986; Higgins 2004; Sharma et al. 2006). Recently the prevalence of leptospirosis has been increasingly recognized in other risk groups, in part due to increased awareness (i.e., ascertainment bias), but also to changes in travel patterns, weather, poor sanitation, and urbanization (Vinetz et al. 2005). Adventure travelers (Sejvar et al. 2003), inner-city residents in developed (Vinetz et al. 1996) and the developing world (Barcellos and Sabroza 2000; 2001; Reis et al. 2008) have been identified as high-risk populations. However leptospirosis still predominately affects disadvantaged populations in the tropics, and due to substantial impacts and limited public awareness, it has been cited as a neglected tropical diseases (Hotez 2008). There is an increasing need for a better understanding of leptospirosis in relation to local risk factors and populations (Cachay and Vinetz 2005).

In Sri Lanka, a tropical developing country, leptospirosis is endemic. Since national reporting for this disease began in 1991, the number of annual cases has averaged around 7 per 100,000 people (Sri Lanka Epidemiology Unit 2008). A large outbreak of suspected leptospirosis starting in late 2007 and into 2008 occurred in Sri Lanka with over 7000 (35.7 cases per 100,000) suspected cases reported in 2008. Leptospirosis risk in Sri Lanka is typically seasonal, with a small spike occurring in the

spring and a large spike later in the fall/winter (Sri Lanka Epidemiology Unit 2008). This pattern roughly follows the seasonal variation in rainfall characterized by two monsoon seasons (*maha*, October to March and *yala*, April to September). It is generally thought that during paddy sowing and harvesting, farmers walking in flooded fields contract the disease when coming into contact with leptospires shed in the urine of infected rodents. There is also evidence that rodent populations expand in and around paddy fields during these periods (Department of Provincial Health Services 2008).

Leptospirosis is a spirochetal disease caused by human contact with pathogenic leptospires present in the environment. Specific leptospire serovars are typically adapted to particular animal reservoir hosts that shed spirochetes in urine (Levett 2001). Knowing which serovars are responsible for human infections can be extremely helpful in uncovering the local epidemiology of leptospirosis. Traditional understanding of the epidemiology of leptospirosis in Sri Lanka has held that the serovar commonly associated with human infection is *icterohaemorrhagiae*, residing in rodent populations. However there is contradictory evidence concerning the sources of the 2008 outbreak. In a preliminary study of 473 suspected cases in Kandy in the Central Province, only 15.6% tested positive for anti-leptospiral antibodies, 5.3% were equivocal, while 79.1% were negative (Agampodi et al. 2008). Identified serovars of 31 analyzed serum samples revealed a diverse array of serovars including *medensis*, *australis*, *ballum*, *canicola* as well as others. Similar analysis of 107 samples from the area found 24.3% of suspected cases tested positive for anti-leptospiral antibodies, with common serovars including *serjoe*, *icterohaemorrhagia*, *cynopteri*, and *tarassovi* among others (Koizumi et al. 2009).

Investigations of exposure history of 1957 confirmed cases in 2008 found that 60.9 % of patients reported exposure to paddy fields during interviews (Sri Lanka Epidemiology Unit 2008). Laboratory testing of 1404 suspected cases confirmed 37% as positive and 39% as equivocal. Nine samples had serovars isolated which revealed *pyrogenes*, *australis*, *weerasinghe*, *gem*, and *canicola*. While rodents are thought to be the main reservoirs for leptospirosis, the role of other animals in the transmission dynamics in Sri Lanka is largely unknown. Canicola for example, is often found in dogs, cattle and swine among other species (Galloway and Levett 2010). Testing in domestic animals is rarely done in Sri Lanka. In livestock, infection is often not severe or can be subclinical. Most commonly leptospirosis will mimic other production-related disease in buffalo and cattle which cause reductions in milk production, and in occasional cases, cause abortion (Sullivan 1974). The diversity of serovars circulating in Sri Lanka makes identifying the transmission and maintenance hosts, epidemiological risk factors, and appropriate control measures, difficult.

Where the understanding of disease risk is uncertain or incomplete, analysis of the geographic and temporal variation in cases can help reveal clues about the processes underlying observed disease patterns (e.g., Odiit et al. 2006; Moffett et al. 2007; Fevre et al. 2001). Spatial epidemiology is the study of the geographic variation of disease (Ostfeld et al. 2005). In the context of leptospirosis, this might include both the detection and analysis of clusters in space and time, and the analysis of spatial variables that help to identify locations where populations are at high risk. For example, outbreaks of leptospirosis in Brazil have been associated with spatial risk factors such as proximity to refuse, locations at risk of flooding, and presence of rats and chickens (Reis et al. 2008).

6.2.1 Variables possibly related to leptospirosis in Sri Lanka

Spatial risk factors in Sri Lanka may relate to characteristics of high-risk populations, the survival of the pathogen in the environment (i.e., surface waters), or factors related to population exposures to pathogens. In this analysis, we were primarily interested in outbreak-related spatial variables, that is, possible correlates related to the dramatic rise in suspected cases observed in 2008.

6.2.1.1 Rodents and environment

Rodent populations often expand where people and settlements are densely situated (Moore et al. 2003) so population density may be a proxy for leptospirosis risk in some areas. Household clustering of leptospirosis cases in Brazil suggested transmission dynamics are spatially structured (Maciel et al. 2008). A study from Thailand, which experienced a leptospirosis outbreak similar to Sri Lanka in 2000-2002 (Thaipadungpanit et al. 2007), found prevalence of leptospires in household rats to be twice as high as rice-field rats (Phulsuksombati et al. 2001).

Waterborne transmission of pathogenic leptospires in tropical settings often occurs in rice paddy fields (Victoriano et al. 2009). Locations where small-scale paddy fields are cultivated will be more highly populated, more heavily managed, and may therefore have more opportunities for exposure. Also, rivers can be sources of infection when people use them for swimming and bathing (Levett 2001).

6.2.1.2 Rainfall

In Sri Lanka, leptospirosis is highly seasonal, and seasons are characterized by variation in rainfall (Sri Lanka Epidemiology Unit 2008). Two major seasons have been identified related to rainfall variability: the northeast (maha) monsoon from October to March, and the Southwest (yala) monsoon from April to September. Within these, there are also

intermonsoons in March to April and October to November (Domroes and Ranatunge 1993). Generally the southwest monsoon sees rainfall on the order of 400-800 mm, and the northeast monsoon is less intense with rainfall ranging from 400 to 1200 mm (Zubair 2002). The spatial distribution of rainfall varies, with the southwest of the country receiving significant rainfall in all seasons, and the north and east parts of the country being hot and dry during the southwest monsoon season. Temperatures vary little in the lowlands, with mean monthly temperatures between 78°F and 85°F. While in the central highlands, mean monthly temperatures range from 55°F to 70°F due to more variable topography. Due to these climatic variations, Sri Lanka can be stratified into three agro-ecological regions based on annual rainfall: the wet zone in the southwest and central highlands area, the dry zone extending north and east of the highlands, and the intermediate zone transitioning between the two (Figure 6.1). The majority of leptospirosis cases occur in the wet zone.

There are three primary mechanisms linking rainfall and leptospirosis cases in humans. There is evidence that increased rainfall during the monsoon season, creates abundant food sources and optimal reproductive conditions for rats (Madsen and Shine 1999). Populations rise the following year, increasing opportunities for human exposure and disease transmission. However, links between rodent populations and disease prevalence in humans are not always clear (Davis and Calvet. 2005; Mills 1999). The second mechanism occurs over time scales of days or weeks, where rainfall-induced floods displace rats out of their burrows and into environments where they have more frequent contact with humans (e.g. households). This has been reported in both urban and rural environments, and is the most commonly cited link between rainfall and human

cases of leptospirosis (Sanders et al. 1999; Tassinari et al. 2008). The third mechanism is seasonal variation in rainfall which determines agricultural activities that put people at greater risk of exposure to contaminated surface waters. In Sri Lanka, this is thought to be the dominant process causing leptospirosis (Dassanayake et al. 2009).

The objectives of the current study were threefold. First, we investigated temporal correlation in suspected leptospirosis cases and rainfall pattern. Rainfall is often positively correlated to human leptospirosis in the tropics and large outbreaks can occur after rainfall events (Victoriano et al. 2009). Second, we to sought to identify clusters of cases during the 2008 outbreak of leptospirosis in Sri Lanka relative to the previous pattern of cases. Identifying locations and times of high risk and change in reported cases may elucidate factors driving the outbreak. Third, we investigated spatial factors useful in the context of disease surveillance activities. Geographic information systems combined with cluster analysis were employed to identify clusters and spatial associations. We developed a new index of cluster-adjusted risk to aid in the prediction of future outbreak patterns in Sri Lanka as well as the geographic targeting of disease control activities and further sero-epidemiological studies. The study concludes with a discussion of our findings in terms of spatial risk modelling for surveillance data, understanding of leptospirosis risk in Sri Lanka, and areas of future research.

6.3 Materials and Methods

6.3.1 Data & study area

Leptospirosis cases are reported to the Epidemiological Unit in the Sri Lankan Ministry of Health (MOH) on a weekly basis from public health inspectors in each of Sri Lanka's MOH administrative areas. Only a very small proportion of suspected cases are sent for

laboratory diagnostic testing. Rather, reported cases are based on clinical signs such as fever, headache, muscle pain, cough, haemoptysis for patients presenting at public hospitals. All analysis reported here is for suspected cases, indicating the earliest report of each case. Many cases may go unreported for mild forms of the disease where people do not seek treatment, as well as those who seek treatment at private hospitals (Department of Provincial Health Services 2008). While clinically diagnosed cases of leptospirosis could also be due to other febrile disease causing pathogens (e.g., Dengue fever, hanatavirus), a recent validation study found the clinical case definition to be 82% accurate when compared to results obtained by microscopy (Dassanayake et al. 2009)

Reported cases were obtained for all MOH areas for years 2005 – 2009. Over the course of the study period, the number of MOH areas increased from 277 in 2005 to 310 in 2010. All counts were standardized to 2005 MOH areas through aggregation using GIS. Population data were obtained for MOH areas from the most recent estimates made by the Epidemiology Unit from 2010 (http://www.epid.gov.lk). Where 2010 estimates weren't available, estimates from 2007 were used. In addition to providing a denominator in analysis of leptospirosis risk, population was used to calculate population density of each MOH area.

Data on agriculture and livestock were obtained for divisional secretariat divisions (DSDs), the main administrative planning units in Sri Lanka. DSDs generally correspond to the MOH area boundaries, although in some places multiple DSDs are covered by one MOH area and vice versa. DSD data were mapped onto MOH boundaries during GIS processing. The Census of Agricultural and Livestock was conducted in 2002 recorded a number of variables describing livestock and agriculture in each DSD in Sri Lanka. The

number of agricultural holdings greater than and less than 0.20 hectares (ha) were recorded, and these census data were used to calculate the percentage of 'small' agricultural holdings in each MOH area.

GIS maps were obtained from the Survey Department of Sri Lanka for areas of paddy agriculture mapped from aerial photographs, locations of urban centers, and rivers and streams. The total area of rice paddy fields in each MOH area was used to determine the paddy density. Finally, the average distance to urban centers and rivers/streams was computed for each MOH area, and new variables were created indicating if each MOH area had an average distance to rivers/streams and towns less than the median. The median average distance to streams/rivers was 400 meters and the median average distance to urban centers was 12 kilometers.

Data were obtained from the Meteorological Department of Sri Lanka on total monthly precipitation for four areas in Sri Lanka, for 2005-2009. Only one station had one missing value, which was estimated as the average for that month from the other years.

6.3.2 Temporal analysis of rainfall pattern and reported leptospirosis cases

To investigate the role of rainfall in the 2008 leptospirosis outbreak, four locations were used to compare total monthly precipitation and total monthly leptospirosis cases in the surrounding district (see Figure 6.1). Locations represent a large degree of variability in rainfall and leptospirosis cases. Comparison of cases and rainfall was done using the cross-correlation function which computes the correlation between series x and series y for multiple lags d. The cross correlation function (Diggle 1990) is

$$r = \frac{\sum_{i}^{n}(x_i - \bar{x})(y_{i-d} - \bar{y})}{\sqrt{\sum_{i}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

and was used to compute correlations between suspected cases of leptospirosis and total precipitation in Anuradhapura, Nuwara Eliya, Ratnapura, and Galle districts (Figure 6.1). A total of 12 lags for the cross-correlation analysis were used, indicating the correlation of cases and rainfall up to a maximum of 12 months previous in order to be able to identify both short-term and long-term correlations.

6.3.3 Baseline reported leptospirosis prevalence analysis

MOH areas where leptospirosis was present during the baseline period were selected to model associations between prevalence (i.e., number of cases divided by the population) and the covariates identified in Table 1. Leptospirosis prevalence in each MOH area for November was the dependent variable, as this is a month when high numbers of suspected cases are reported throughout all endemic areas. The November prevalence was log transformed to ensure normality and related to the dependent variables through a linear regression model (i.e., log-linear model).

6.3.4 Outbreak detection, modelling, and mapping

In order to detect clusters of cases in space and time during the 2008 outbreak year, a baseline of expected values was required. The first three years of data were used to determine the relative variation in the number of cases in each MOH area in each month. During these years, the number of annual cases ranged from 1552 in 2005 to 2198 in 2007, which are close to the historical norm (Sri Lanka Epidemiology Unit 2008). Establishing the space and time case distribution allows us to identify unusual patterns of

leptospirosis during years subsequent to 2007. A month-specific leptospirosis case ratio was calculated for each MOH during the endemic period using the monthly mean number of cases of all MOHs as the denominator and monthly mean number of cases over the 2005-2007 period for each area as the numerator. A value of 1 indicates the number of cases in a particular MOH area-month is equivalent to the average number of cases over all areas for that month, whereas values greater than 1 indicate higher numbers of cases in that month, and values less than 1 lower numbers. Figure 6.2 illustrates examples of the spatial distribution in relative case ratio for May and November. These estimates were then used to condition our analysis of the outbreak year (2008) in order to detect clusters in space and time.

The space-time scan statistic was used to identify unusual clusters of leptospirosis at the MOH-month level of resolution in 2008 (277 areas and 12 months). The null hypothesis in cluster analysis is that risk of disease is distributed proportional to the population at risk and any known relative risk variations. Scan statistics are one of the most widely used approaches to detecting clusters in disease data (Naus 1965; Kulldorff and Nagarwalla 1995; Kulldorff et al. 2001). Briefly, the space-time scan statistic uses cylindrical scanning volumes where radius is defined by space and height is defined by time (Kulldorff et al. 2001). Relative risk is calculated within each cylinder, and compared to the relative risk outside the cylinder, and cluster 'unusualness' is evaluated using a likelihood ratio. The cylinder that maximizes the difference between within-cylinder risk and outside-cylinder risk, adjusted for sample size, is the most likely space-time cluster. Cluster significance was evaluated using Monte Carlo randomization ($p < 0.05$). Secondary clusters were defined subject to the constraint that clusters do not

overlap geographically and the same time. The Poisson probability model was used for this analysis. SaTScan software was used for all scan statistic analysis (Kuldorff and Information Management Services 2010). In the analysis reported here, expected numbers of cases in each MOH were proportional to the population multiplied by the leptospirosis case ratio during the endemic period. We report the clusters identified by the space-time scan statistic as risk clusters.

A second analysis was performed to identify clusters of rapidly increasing risk. Linearly increasing trends in the rate of disease can be identified in the same way as clusters of high risk; however in this case, the temporal dimension is fixed (Kuldorff and Information Management Services 2010). The temporal trend in monthly rate of disease inside each cylinder is compared to the trend outside the cylinder, and the cylinder maximizing the difference in trend is the most likely cluster. Significance is determined using Monte Carlo randomization similar to the traditional scan statistic. Trend clusters of leptospirosis were used to indicate areas where disease patterns were changing rapidly.

Clusters identified in the scan analysis were then used to explore unusual risk factors at these locations. Because the scan statistic methodology uses circular scanning windows to identify compact geographic clusters, locations where risk or trend is negligible can be included in clusters if they are surrounded by high risk/trend areas. Also, areas of low relative risk can be seen as extremely high risk in the presence of only a few cases (i.e., small numbers problem). To account for these scenarios, MOH areas that were part of clusters but had fewer than 6 cases were re-coded as non-cluster areas in the risk-factor and subsequent analysis. This threshold constrains the risk factor analysis to ensure cluster locations are truly unusual and unlikely due to random variation.

Logistic regression was used to model cluster membership (1 – cluster, 0 – no cluster) with the MOH area spatial covariates outlined in Table 1. Significant variables in the models for risk and trend clusters were used to predict 2009 cluster locations. The objectives of the regression modelling was to both characterize risk factors and use this information in a way that could be used to aid in the prediction of future patterns. As such, model estimates were transformed from log-odds to probabilities and multiplied by the mean relative risk for each area. These new risk estimates, called cluster-adjusted-risks, were mapped and compared to clusters identified in the 2009 data.

Areas of high cluster-adjusted risk determined by combining the logistic risk model with the relative risk of reported leptospirosis obtained during the baseline period, represent locations where space-time clusters might be identified in 2009. The space-time scan statistic was run for the 2009 leptospirosis data and compared to the cluster-adjusted risk estimates.

6.4 Results

The 2008 outbreak of suspected leptospirosis cases in Sri Lanka is readily apparent from the graph of weekly cases in Figure 6.3 starting in around week 150. There were a total 7421 cases in 2008, yielding a national average of 34.9 cases per 100,000 people. The annual bi-modal distribution is also evident in the fluctuating peaks in the southwest (red – Figure 6.3) and northeast (blue – Figure 6.3) monsoon seasons. This seasonal pattern breaks down somewhat in 2007 with the delayed onset for the northeast peak into early 2008. High case numbers are sustained through 2008 and 2009. In both 2008 and 2009 the southwest peak in cases is much higher than the northeast peak. The spatial distribution of annual cases indicates that high-risk areas were located along a southwest

to northeast trajectory during the outbreak (Figure 6.4). The proportion of cases reported from the Wet Zone increased, with cases predominately occurring in the southwest-central corridor between Colombo and Matale.

Correlation analysis of the time-series of cases and total rainfall in four areas during the baseline period revealed significant correlations between reported cases of leptospirosis and total rainfall two months previous in three out of the four districts (Table 2). In 2008-2009, this correlation disappeared in Ratnapura and Galle, while strengthening in Anuradhapura. No significant correlations were found for Nuwara Eliya. Figure 6.5 shows the data used in the cross-correlation analysis. The number of cases in Nuwara Eliya during 2005-2007 was extremely low (Figure 6.5b), however began to increase in 2008. Ratnapura, where both rainfall and leptospirosis is more common, did not have any significant zero-lag correlations. One pattern that emerges from Figure 6.5 in Ratnapura, and to a lesser extent in Galle and Anuradhapura, is the timing of the first peak in rainfall. In 2007 and 2008, the first spike in rainfall occurs earlier in the year than in previous years, indicating heavy rainfall late in the northeast monsoon season.

Linear regression analysis revealed positive association between leptospirosis prevalence and the percentage of farms less than 0.20 ha in the MOH, and the average distance to rivers within an MOH less than 400 meters. No other variables had significant effects on leptospirosis prevalence (Table 3). The model was significant ($p < 0.01$) and model R^2 was 0.17, indicating 17% of the variation in risk was explained by the covariates. Residuals did not have any significant spatial autocorrelation and were approximately normal.

Space-time cluster analysis revealed one large significant risk cluster in 2008 making up almost the entire southern half of the island (Figure 6.6a). The timing of this cluster was from October to December, roughly corresponding to the northeast monsoon season. There were 800 cases, distributed over a population of 1.09 million. The MOHs that met the selection criteria were all on the western and north-western edge of the cluster. The trend cluster analysis (Figure 6.6b) revealed that the fastest increasing risks were in MOHs around Kandy, Matale, and Kurunegala. There were also significant trend clusters southeast of Colombo, though these MOHs had very low observed numbers of cases, indicating it is an area with very low relative risk in the baseline period. Further south near Matara, a significant trend cluster was detected. The characteristics of risk and trend clusters for 2008 are outlined in Table 4.

Results from logistic regression modelling are outlined in Table 5. Risk cluster locations had significant positive associations with population density and MOH areas where the average distance to river was less than 400 meters. Trend cluster locations were positively associated with population density, river distance, and negatively associated with percentage of farms less than 0.20 ha.

Cluster-adjusted risk model maps are presented in Figure 6.7. Risk clusters identified from space-time scan analysis of the 2009 data are also shown. The details of the 2009 clusters are outlined in Table 6. The major cluster occurring from Colombo southeast into Ratnapura, Galle, and Matara contained 1802 cases occurring from September to December 2009. Out of 52 MOH areas meeting the selection criteria in this cluster, 38 had cluster-adjusted risk greater than 1. The second major cluster, though far less severe, was found in the Matale area, occurring between January and May. This

cluster was made up of 138 cases in 8 MOH areas, of which 2 had cluster-adjusted risk greater than 1. The remaining clusters were very small, containing between 6 and 18 cases.

6.5 Discussion

There have been numerous hypotheses proposed regarding the cause of the 2008 leptospirosis outbreak in Sri Lanka. The first culprit is rainfall pattern, as it is typically the primary driver of leptospirosis incidence in the tropics (Victoriano et al. 2009). During the baseline period, two-month lags in correlation between rainfall and cases were detected in three out of four areas. Two mechanisms may explain this pattern. Rainfall occurs causing floods which displaces rodents to habitats in and around housing areas causing greater exposure risk to occupants. A small study in the Kandy area of south-central Sri Lanka which trapped 21 rats (*Rattus rattus* and *Mus musculus*) in and around houses found 13 tested positive for leptospira antibodies (Mukthar et al. 2010). Secondly, the observed rainfall – case two month correlation may reflect an occupational risk associated with farming occurring in paddy fields during rainy seasons in Sri Lanka. Disentangling the causal relationship between rainfall patterns and incidence and distribution of leptospirosis in Sri lanka requires further field study.

Correlations between rainfall and cases in Ratnapura and Galle in the baseline period disappear during 2008-2009, while the number of leptospirosis case reports increased dramatically. One reason this may have occurred is a large increase in the rat population. Rainfall and rat populations are often related over times scales of about a year, where rainfall alters habitat, resulting in changes in populations the following season (Madsen and Shine 1999, Taylor and Green 1976). However our analysis of long

term correlations (not shown), failed to find significant correlations. It may be that the relationship between rainfall and rat populations is more complicated than simple linear correlations (see Davis et al. 2005;). Temporal analysis of cross-correlations between suspected cases and rainfall did not provide any evidence of a causal link between rainfall pattern and leptospirosis in 2008 in terms of month-to-month correlation. Only Anuradhapura in the Dry zone had significant correlation with rainfall in 2008. Anurhadupra also experienced a large increase in reported cases in 2008 which was extremely unusual compared to the previous years as well as a large amount of late-monsoon rainfall in 2008 which preceded the rise in reported cases. We also observed an early peak in rainfall in locations in the Wet Zone, which could have altered the endemic transmission cycle of leptospirosis. Early rainfall in 2007 and 2008 actually represents a prolongation of the normal northeast monsoon season. This may have created abundant rodent habitat, prolonging reproduction and causing a spike in rat populations in 2008. The impact of this change would be greater in Anuradahpura, which does not receive year-round rainfall. However, in order establish this connection, a substantive analysis of rainfall pattern and leptospirosis cases all throughout high-risk areas is needed.

Secondary hypotheses regarding the 2008 outbreak have suggested that what are being reported as suspected leptospirosis cases are actually diseases caused by some other pathogen. The serological testing done on cases in Sri Lanka does suggest the possible presence of other pathogens, as positive laboratory confirmations have commonly ranged from only 15% to 24% of tested samples (Agampodi et al. 2009; Koizumi et al. 2009). In both India (Clement et al. 2006) and Brazil (Hinrichsen et al. 1993), hantavirus infections have been found in patients presenting with suspected leptospirosis. Differential

diagnoses of leptospirosis from most hantavirus cannot be made based on clinical signs (Clement et al. 1999). In India, a serological study found 12% of samples that were suspected to be leptospirosis but tested negative, reacted positively for hantavirus antibodies (Clement et al. 2006). In Brazil, similar analysis found 5% tested positively for hantavirus infection.

To what extent does our analysis indicate the possibility of other pathogens mimicking symptoms of leptospirosis? The space-time cluster analysis revealed unusually high risk in a large cluster covering most of the southern half of Sri Lanka. The drivers of this cluster, MOH areas with greater than 5 cases and relative risk greater than 1, were all found on the western and southwestern edges. These areas are typical locations of high risk for reported leptospirosis (see Figure 6.2). The timing of this cluster corresponded to the northeast monsoon season, a usual period of high risk suggesting an intensification of the normal pattern of leptospirosis in Sri Lanka. Intensifying endemic patterns is also illustrated by fact that the proportion of cases occurring in the Wet Zone increased during the study period and the annual spatial distribution of cases. The association with distance to rivers and number of cases in the Wet Zone, evident during both endemic and outbreak years, while suggestive of waterborne transmission risk associated with leptospirosis, does not rule out the presence of other pathogens (e.g., rodent habitat for hantaviruses, mosquito habitat for Dengue Fever). These observations suggest that areas of endemic febrile illness became more prevalent in endemic areas rather than expanded to new areas. Whether this febrile illness is leptospirosis cannot be determined based on this analysis. Expanded testing of the rodent population (e.g., Mukthar et al. 2010) in cluster locations might help shed more light on this.

The large number of areas included in the cluster but having very low numbers of cases is a due to two factors. First, the circular spatial search area (i.e., the base of the cylinder) used in the space-time scan statistic is not optimal for finding irregularly shaped clusters (Duczmal et al. 2006). Secondly, extremely low relative risks yield high-risk ratios when even a few cases are reported. A space-time model of relative risk, while outside the scope of this analysis, could provide more robust measures of risk (e.g., Bernardinelli et al. 2001).

The trend clusters identified in this analysis were concentrated north and west of Matale, and interior areas in the south situated northwest of Matara. These may indicate areas where dynamics of infection are rapidly changing, and may therefore be locations where surveys of patients or other forms of active surveillance might be targeted. In the annual risk maps in Figure 6.4, MOH areas north of Matale appear to have a change in risk starting in 2007. That these same areas are identified as statistically significant trend clusters in 2008 indicates these were areas that experienced the greatest change in risk.

Spatial variables associated with reported leptospirosis during the baseline period indicate higher risk in areas with a high percentage of small-scale agriculture and proximity to rivers and streams. These factors indicate a seasonal, endemic leptospirosis pattern, where human infections occur in areas near small-scale agriculture. Small-scale agriculture landscapes are heterogeneous, providing both habitat and food sources for rat populations.

During the outbreak year, covariates positively associated with risk cluster locations included proximity to rivers and population density. These relationships held for trend clusters as well, with the addition of the variable for percentage of farms less

than 0.20 ha, which had a negative relationship, indicating these are more rural areas where farms are larger. The 0.20 ha threshold might be too small to capture regular small-scale paddy fields in outlying areas, but in areas near large urban centers, this threshold does capture them. The effect of population density and rivers in transmission are similar for both risk and trend clusters. In both models, the variable for average distance to river being less than 400 meters has a strong effect (though large confidence intervals). Interestingly, rice paddy density was not a significant variable in any of models.

Higher population density does not necessarily imply urban environments in Sri Lanka. Most of the Wet Zone is densely populated, and a large subset of the population lives in villages and commutes to urban centers for work. A population density association to leptospirosis risk may indicate that transmission during the outbreak period was not occurring in paddy fields, but in densely populated semi-rural villages. Household level clustering of leptospirosis transmission in Brazil supports this hypothesis (Maciel et al. 2008). Closer to Sri Lanka in the Andaman Islands, leptospirosis has been associated with presence of cattle in the home, drinking from streams, and housing characteristics such as thatched roofs (Suganan et al. 2009), and in the Seychelles to a number of home-based factors such as walking barefoot, washing in streams, gardening, and presence of refuse around the home (Bovet et al. 1999). As noted earlier, a small study in a high-risk leptospirosis area in the Central Province of Sri Lanka found 62% (13/21) of rats trapped around homes tested positive for leptospire antibodies (Mukthar et al. 2010). Further study of household-level risk factors in Sri Lanka is needed.

The role of animal reservoirs other than rats remains an important area of inquiry in Sri Lanka. Leptospirosis is commonly known as ‘rat fever’ in Sri Lanka, and knowledge about rats as risks of infection is accordingly well known, but is extremely limited about infection from domestic animals or dogs (Agampodi et al. 2010). Additional laboratory testing of animals would aid in developing a thorough understanding of transmission processes in Sri Lanka.

When the outbreak-model probabilities were combined with relative risk estimates from 2005-2007 to produce cluster-adjusted-risk values for each MOH area, and compared to clusters detected in 2009 (Figure 6.7), the major cluster southeast of Colombo compared quite well to the cluster-adjusted risk estimates above one. However, large areas had cluster-adjusted risks greater than one that were not part of 2009 clusters. For example, areas north of Colombo, west of Kandy, and south of Kurunegala had positive cluster-adjusted risks. This area may be at risk of future outbreak levels of reported cases. Similarly, the areas north of Matale, which were detected as rapidly increasing risk in 2008, may be important areas to focus surveillance on. The alternate explanation however, is that there are unidentified factors keeping the reported number of cases below some threshold in these areas. This may be due to the dynamics of the pathogen itself, the role of other animal reservoirs, access to healthcare services, or education and surveillance initiatives in these MOH areas.

The major findings of this analysis are that there appears to be a semi-urban pattern to outbreak levels of leptospirosis in Sri Lanka. While, proximity to rivers was significant for both the endemic and outbreak periods, population density only became significant in the outbreak model. While the location and timing of cases does suggest

strong seasonal dynamics, monthly correlations to rainfall were not detected during 2008-2009. Further analysis of the role of rainfall during the outbreak is needed. Highly populated village areas in the Wet Zone are at greatest risk during outbreak periods, but not necessarily only those working in the fields. Paddy farm density was not an important variable in any of the models of risk. The cause of the outbreak remains unknown, but the analysis here suggests a possible shift in transmission dynamics, possibly from fields to households. One aspect of the rainfall analysis that may have affected rodent populations is the timing of rainfall. Unseasonably heavy rainfall late in the northeast monsoon may have affected rodent populations or other maintenance hosts and ultimately caused changes in transmission risks and/or improvement in environmental conditions for leptospire survival. However, the cause(s) contributing to the 2008 outbreak remain to be discovered.

6.6 Acknowledgements

This project was funded in part by the Teasdale-Corti Global Health Partnership and the National Sciences and Engineering Research Council of Canada. The authors would like to thank Dr.Paba Palihawadana of the Sri Lanka Epidemiolgical Unit for providing access to the data.

Table 6.1. Listing and rationale for covariates used in modelling reported leptospirosis risk and outbreak locations.

Variable	Rationale
Distance to town	Towns may have higher rodent populations due to refuse build up, and rice paddy farms in areas adjacent to urban areas may have high risk of leptospirosis.
Distance to river	Sources of freshwater are important for transmission of leptospirosis to humans. Areas with a high water table may facilitate leptospire survival in the environment.
Log(population density)	Rodent populations are often directly related to human populations. Highly populated areas on the outskirts of urban centers may serve as amplification areas for disease risk.
Rice paddy density	Rice paddy farmers are the traditionally most affected group in Sri Lanka and many Asian countries.
% of farms < 0.20 ha	The scale of rice paddy agriculture varies greatly in Sri Lanka. Areas with many small fields with villages and human settlements may be areas of increased risk.

Table 6.2. Cross-correlations between monthly cases of reported leptospirosis and total rainfall for baseline and outbreak periods.

	Months previous to reported cases of leptospirosis (lags)			
2005-2007	-3	-2	-1	0
NE	0.164	0.318	0.284	0.25
RAT	-0.105	0.404*	0.282	0.097
ANU	0.086	0.365*	0.131	-0.155
GALLE	0.386*	0.498*	0.192	0.062
2008-2009				
NE	0.225	0.034	-0.136	0.344
RAT	-0.027	0.131	0.381	0.146
ANU	0.279	0.554*	0.337	0.083
GALLE	-0.012	-0.116	0.297	0.146

* Significant correlation

Table 6.3. Linear regression model for reported leptospirosis prevalence, Sri Lanka, 2005-2007.

Variable	Coefficient	P-value
Distance to town < 12 km	-0.311	0.065
Distance to river < 400 m	0.3911	0.015
Log(population density)	0.033	0.742
Rice paddy density	0.888	0.137
% of farms < 0.20 ha	1.234	0.009

Table 6.4. Risk and trend space-time clusters detected in 2008 reported cases of leptospirosis, Sri Lanka.

Risk Clusters					
Cluster ID	Cluster Name	Duration	Relative Risk	Number of cases	# MOH Areas in Cluster*
1	Southern	Oct - Dec	3.76	800	49
Trend Clusters					
Cluster ID	Cluster Name	Cluster Trend	Relative Risk	Number of cases	# MOH Areas in Cluster*
1	Central	21.98%	1.37	763	29
2	Southern	36.44%	1.82	84	4

* total number of MOH areas meeting selection criteria in each cluster. Selection criteria was greater than 5 cases, and relative risk greater than 1.

Table 6.5 Spatial risk factors associated with risk and trend clusters identified in 2008 reported cases of leptospirosis, Sri Lanka.

A) Risk Cluster Model	N = 49 for clusters, N = 228 for non-clusters	
	Odds Ratio (95% Confidence Interval)	P
Distance to town < 12 km	0.65 (0.32-1.32)	0.23
Distance to river < 400 m	4.34 (2.04-9.25)	< 0.00
Log(population density)	1.57 (1.07-2.29)	0.02
Rice paddy density	0.43 (0.04-5.06)	0.51
% of farms < 0.20 ha	0.41 (0.07-2.48)	0.33

B) Trend Cluster Model	N = 33 for clusters, N = 244 for non-clusters	
	Odds Ratio (95% Confidence Interval)	P
Distance to town < 12 km	1.05 (0.48 - 2.32)	0.90
Distance to river < 400 m	2.84 (1.23 - 6.54)	0.01
log(population density)	1.58 (1.06 - 2.36)	0.02
Rice paddy density	1.43 (0.09 - 21.93)	0.80
% of farms < 0.20 ha	0.11 (0.01 - 0.81)	0.03

Table 6.6. Space-time risk clusters detected in 2009 reported cases of leptospirosis, Sri Lanka and cluster-adjusted risk model results from 2008.

Cluster ID	Cluster Name	Duration	Relative Risk	Number of cases	Cluster-adjusted risk > 1	# MOH Areas in Cluster*
1	Colombo-Matara	Sept - Dec	3.28	1802	38	52
2	Matale	Jan - May	21.99	138	2	8
3	Gampola	June	2790.57	18	1	1
4	Chilaw South	June - Oct	586.98	16	0	1
5	Chilaw North	July - Nov	1184.93	9	0	1
6	Pannala	July - Aug	688.47	7	0	1
7	Mirigama	Jan	1028.69	6	1	1

* total number of MOH areas meeting selection criteria in each cluster. Selection criteria was greater than 5 cases, and relative risk greater than 1.

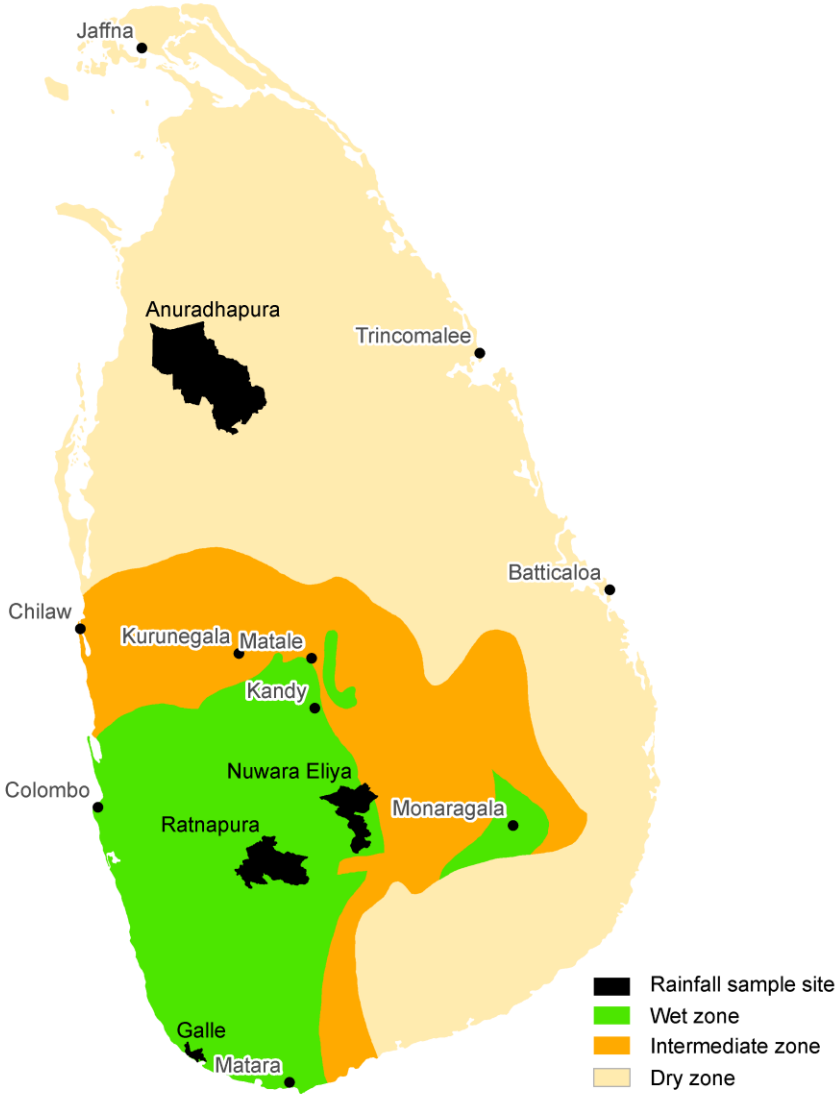


Figure 6.1. Map of Sri Lanka showing wet zone, dry zone, intermediate zone and locations where rainfall analysis was carried out.

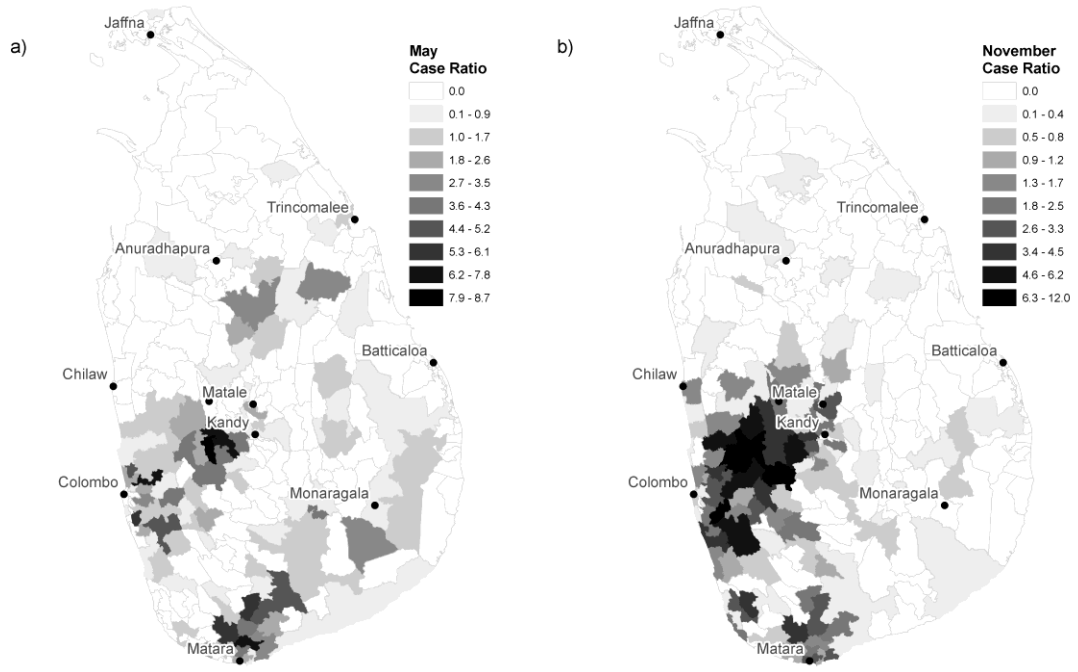


Figure 6.2. Leptospirosis reported case ratios estimated from 2005 – 2007 baseline period for a) May and b) November.

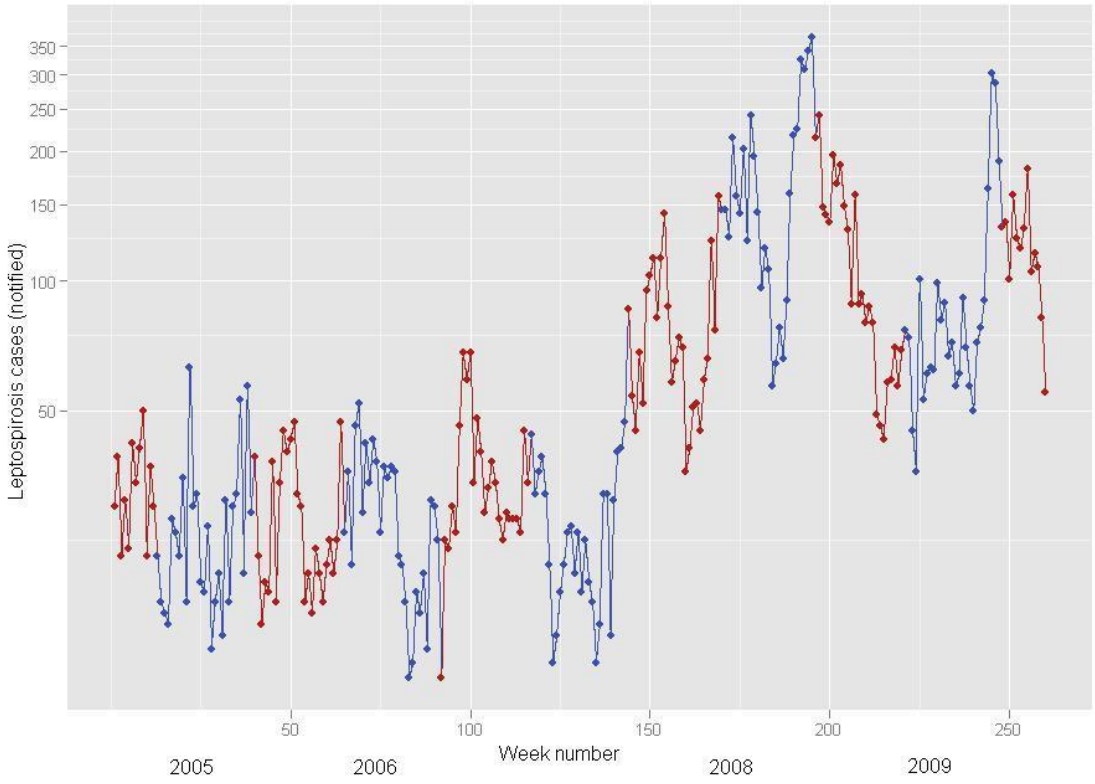


Figure 6.3. Weekly number of reported cases of leptospirosis plotted on logarithmic scale, Sri Lanka 2005-2009, northeast (*maha*) monsoon in red, southwest (*yala*) monsoon in blue.

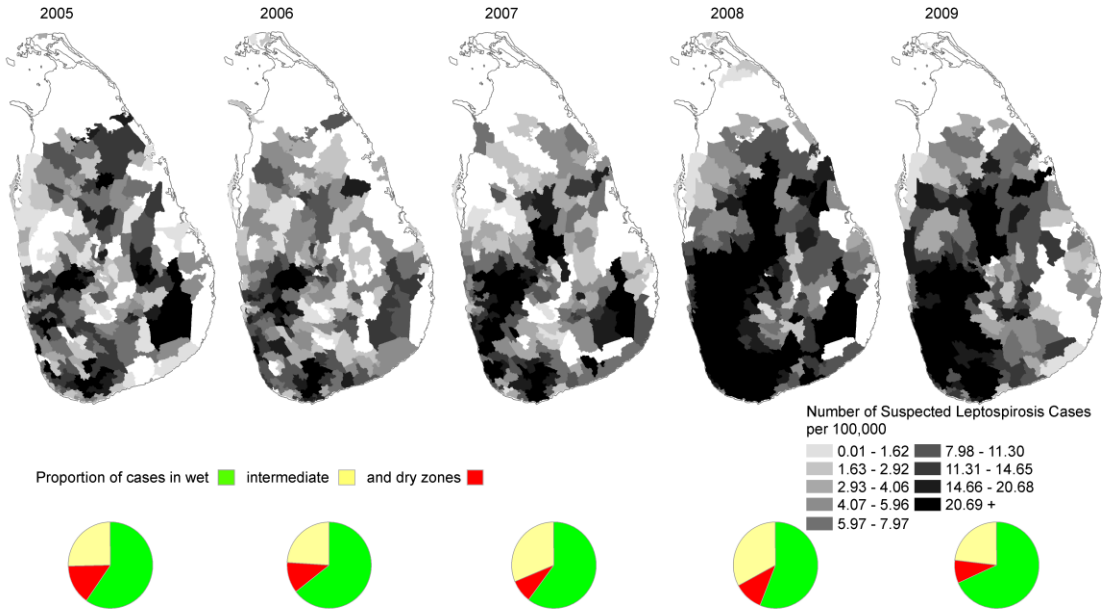
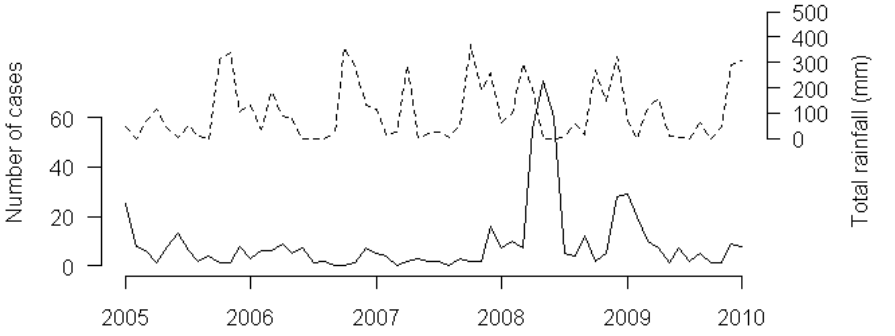
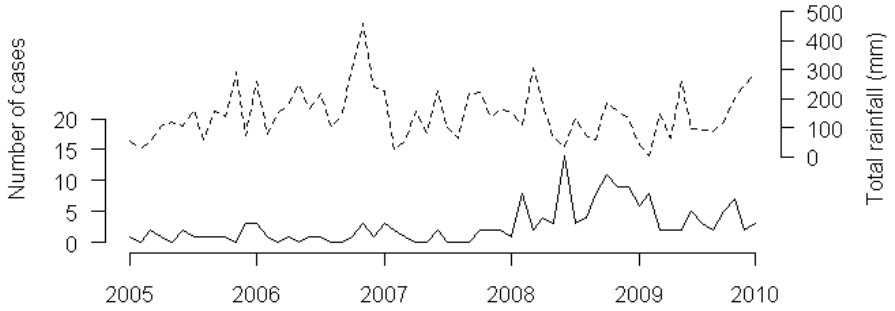


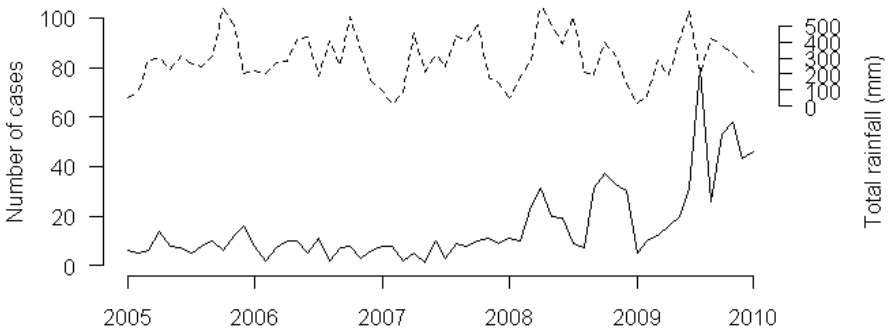
Figure 6.4. Annual number of reported cases of leptospirosis in Sri Lanka and the proportional distribution in ecological zones.



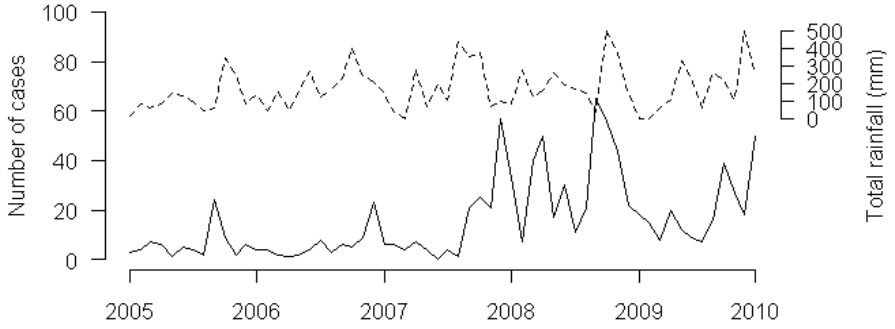
a)



b)



c)



d)

Figure 6.5. Total monthly rainfall and total number of reported leptospirosis cases for a) Anuradhapura, b) Nuwara Eliya, c) Ratnapura, and d) Galle.

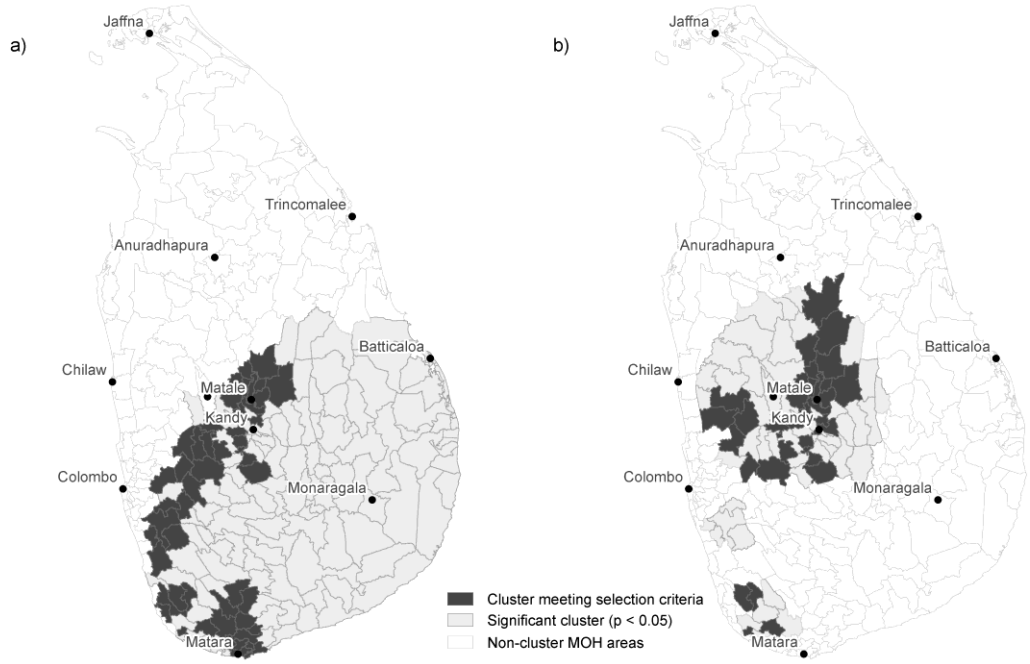


Figure 6.6. A) Risk and B) trend space-time clusters detected in 2008 reported cases of leptospirosis, Sri Lanka.

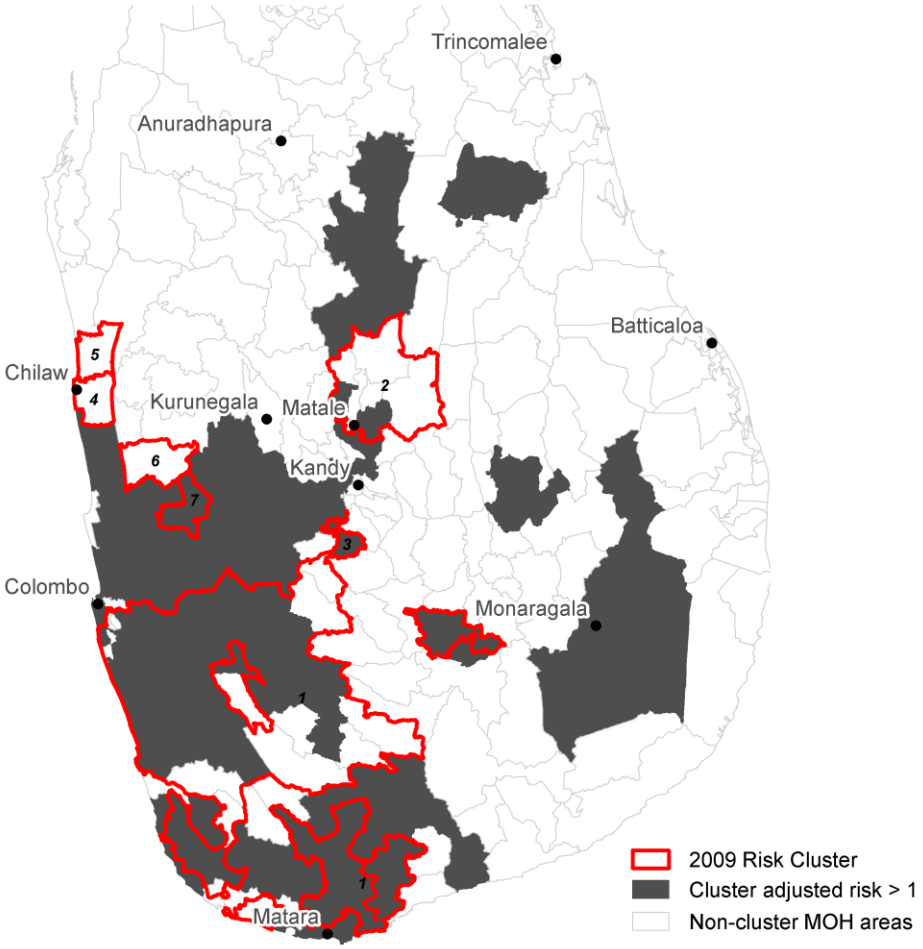


Figure 6.7. Cluster map showing areas with cluster adjusted risk > 1 (grey) and 2009 clusters of reported leptospirosis detected using the space-time scan statistic (red). Numbers refer to cluster number described in Table 6.

Chapter 7: Conclusions

7.1 Abstract

This dissertation brings together many aspects of emerging infectious disease surveillance. The research contained here extends theory of surveillance, with a focus on implementation in lower-resource settings (Chapter 2). The review of methods (Chapter 3) and software (Chapter 4) provides a novel compilation and practical guidance for public health researchers and epidemiologists with regard to method and software selection. A statistical model for frontline veterinary surveillance data presented in Chapter 5 offers a novel and promising approach for future surveillance systems. And finally, the application of space-time approaches to human leptospirosis in Sri Lanka contributes to the understanding of this disease and serves as an example of the power of spatial analysis of disease patterns (Chapter 6).

The problem of emerging diseases is inherently complex and multidisciplinary. This research has taken multiple approaches to developing new knowledge on EID surveillance. This chapter provides a summary of the major contributions of this research, discusses some issues and limitations, and suggests directions for future research.

7.2 Contributions of this research

In the course of this research, EID surveillance became part of the daily activity of 40 field veterinary surgeons in Sri Lanka. The impact that this has on the future of EID surveillance in Sri Lanka remains to be seen, but this project and related activities put emerging diseases on the minds of many people in Sri Lanka. The specific contributions

of the papers contained herein will enable developers of future systems to build on our successes in the infectious disease surveillance and analysis system (IDSAS) and learn from our missteps. The publication of our implementation experience and recommendations in *Emerging Infectious Diseases* will reach a wide audience interested in emerging diseases. This may have considerable impact on how EID surveillance systems are developed in the future. Further, we have demonstrated the feasibility of deploying an animal-based system in a lower-resource setting with open-source technologies. Increased communication and engagement among Sri Lankan organization involved in veterinary public health was another beneficial side effect of the system development process.

In Sri Lanka, mobile phone-based data collection is now being incorporated into ongoing disease surveillance initiatives and planning at the Department of Animal Production and Health. Nationally notifiable animal diseases which are reported to the World Organization for Animal Health (OIE) are going to be piloted using remote data collection and technologies based on the IDSAS system. Further, my own future work in Sri Lanka will aid in the development and evaluation of this system. The IDSAS system has proven to be an important first step in establishing timely disease surveillance of animal health in Sri Lanka. This alone will contribute to Sri Lanka's capability to detect and respond to an emerging disease in the future.

Methods of space-time disease surveillance made up a major component of this research. The review of methods in Chapter 3 provides a concise and practical assessment of a large and diverse set of statistical and computational approaches to analyzing surveillance data. Nowhere in the literature is such a review evident which will make

these approaches accessible to both researchers and epidemiologists and analysts working with surveillance data. This chapter has been published in the journal *Spatial and Spatial-temporal Epidemiology* which brings together research from geography, epidemiology, and statistics.

The review of software in Chapter 4 was published in the *International Journal of Health Geographics*, earning the designation 'Highly Accessed' which is reserved for especially highly accessed articles. The review provides a comparative assessment of four software programs available for analysis of surveillance data in space and time. Software implementation is an often overlooked aspect of methodology which is an essential part of all data analysis.

A new model for surveillance of emerging diseases is illustrated in Chapter 5. Hidden markov models (HMM) are shown to be an effective approach for surveillance data with short baselines. The model represents the first application of a Poisson HMM with covariates for disease surveillance. Publication of this paper will promote the use of HMM methods in future disease surveillance systems. The two-step modelling approach is also a novel way to deal with user-generated data effects, and particularly suits sentinel disease surveillance applications. The methods contained in Chapter 5 provide a robust example of space-time disease surveillance modelling with novel surveillance data.

The analysis of the spatial epidemiology of the human leptospirosis outbreak in Sri Lanka reported on in Chapter 6 revealed that the endemic pattern prior to 2007 changed dramatically in 2008. The covariate analysis suggested a possible shift in transmission dynamics, and the identified risk and trend clusters provide guidance for future research and surveillance initiatives.

The modelling of IDSAS data in Chapter 5 provides a regional breakdown of commonly reported cattle diagnoses in Sri Lanka. Baseline estimates on a weekly-veterinarian basis can be scaled up to the district level to determine district weekly and monthly averages. These can be used in future disease surveillance and analysis projects in Sri Lanka. The methodology is extendable to any disease as establishing appropriate baselines is a fundamental step for any new surveillance system.

7.3 Issues and limitations of this research

The major limitation of this research is that I am unable to say unequivocally that the IDSAS system can detect an EID. The central problem of evaluation with EID surveillance, since the goal is to develop systems, processes, and people to detect the new and unusual, is it is almost impossible to evaluate until an EID event occurs. During 2009 several events provided evidence that the reach of the system for increasing communication between field veterinarians and DAPH headquarters would improve the likelihood of EID detection. Related to this, the utility of tracking syndromes was never fully realized in the analysis here and remains an open question. Similarly, how IDSAS generated alerts would be acted-on in practice was never fully explored or realized. And finally, flowing the data collected in IDSAS back to participating veterinarians could have been improved greatly. This is definitely something that should be incorporated into future systems to promote and sustain engagement. Part of these limitations stemmed from the project's status as an external research project rather than a government initiative. Hopefully the research presented here demonstrates potential and, if current planning is an indicator, will serve as a stepping stone to enhanced surveillance capacity in Sri Lanka.

Another key limitation to all of the analysis is that representations of space are both coarse and static. While farm-level GPS data was collected during IDSAS, the data was not utilized in the analysis in Chapter 5 of baseline patterns. Initially, logistical problems delayed the delivery of GPS to field veterinarians. This and related factors led to a large proportion of IDSAS data not having GPS coordinates. GPS records that were obtained have been provided to the DAPH to be integrated with their farm mapping initiatives. Even when analyzed at the scale of veterinary ranges, the data and surveys have historical record. When unusual cases did arise such as Blackquarter in Anuradhapura, we had no way of determining if the sick cattle were recently moved from another area. Animal tracking would greatly improve the ability of the system to monitor and trace pathogens.

In the analysis of human leptospirosis in Sri Lanka, mobility of people could also have contributed to bias and error when making place-based associations. For example, people may work and live in different MOH areas than where they seek treatment. Further, the analysis provided some evidence of changes in processes in 2008; clusters were found to be positively associated with population density and distance to freshwater, however these covariates alone do not explain the outbreak. Further exploration of the effect of rainfall pattern, population dynamics of rodents, other animal hosts, and socioeconomic aspects of affected populations is required for a fuller understanding of the outbreak and endemic cycles of leptospirosis in Sri Lanka.

Misclassification of cases is a potential source of error in the leptospirosis analysis. Reference to the literature was inconclusive in this regard, with small serological studies yielding positive results for approximately 20% of suspected cases,

yet an evaluation of the clinical definition finding it to be over 80% accurate. Use of clinical diagnoses surveillance data to make ecological associations depends largely on the case definitions used. There remains the possibility of similar clinical diseases contributing to the cases reported as leptospirosis. Also, as the outbreak became apparent in Sri Lanka, physicians may have become more aware of leptospirosis, and reported cases increasing after 2008 may have been affected by this ascertainment bias. Isolating areas and times of high risk and changing risk can reveal patterns that suggest hypotheses about disease processes, such as transmission, but making inferences about risk factors based on these associations is only a first step towards establishing causal links. The risk factor analysis should be interpreted with caution and used to guide further, individual-level studies.

7.4 Directions for future research

This research opens up many exciting research questions. A central theme of this dissertation has been the integration of statistical and applied disease surveillance. Rarely are statistical methods developed in the same research project that developed the data collection system. However, as was shown here, understanding both aspects can have many residual benefits. A remaining challenge in EID surveillance is how surveillance outcomes can be linked to public health action. Development of appropriate theory and outcome-oriented surveillance projects in the future will undoubtedly be an active area of research. The role that methods of space-time disease surveillance can play in these processes will be a crucial part of these future developments.

Establishing baselines and monitoring changes requires statistical models that incorporate data from multiple contributing processes and can yield decision support

tools that are useful for applied surveillance contexts. The state probability visualization in Chapter 5 is just one example of such tool, and visualization of complex space-time models may be an important area for future research. Graphical methods for all parts of statistical modelling and analysis are becoming increasingly common (Gelman 2004), and this trend is likely to increase as more data sources become available.

In Sri Lanka, many areas for future research have been identified. Very few animal leptospirosis cases were reported to IDSAS during one year of surveillance, yet limited studies have demonstrated Leptospira serovars in cattle and buffalo (Gamage et al. 2009). Similarly little is known about the distribution of Leptospira in Sri Lanka's large street dog population, and what role these hosts may play in leptospirosis transmission dynamics. A study investigating animal Leptospira serovars may help elucidate the factors contributing to the 2008 outbreak and develop appropriate response measures in the future. In addition, a more complete understanding of rainfall-leptospirosis relationships is also needed. The analysis here focused on years 2005-2009, however, longer-term patterns need to be analyzed to truly ascertain seasonal dynamics. In future research, coupling large-scale geographic studies with person-level case control designs might help to address some of the issues of large-scale ecological studies of infectious disease.

While in some respects, IDSAS could be considered a space-time surveillance system for EIDs, there are many parts of the system which could have been improved. Auxillary datasets were analyzed post-hoc, but ideally these data sources would be integrated into the system seamlessly. Working within a resource-constrained setting limited some of the initial plans for automated, online data integration, however there is

potential here for the remotely sensed imagery to fill this void. Once models and relationships between disease prevalence and climate variables have been established, incorporating them into automated space-time disease surveillance and risk forecasting systems is the next step. Future research efforts should focus on developing automated data streams and generating decision support tools that are both acceptable and actionable in practical surveillance.

Chapter 8: References

Abellan JJ, Richardson S, Best N. 2008. Use of space–time models to investigate the stability of patterns of disease. Environmental Health Perspectives. 116:1111-1119.

Agampodi S, Peacock SJ, Thevanesam V. 2009. The potential emergence of leptospirosis in Sri Lanka. The Lancet Infectious Diseases. 9:524-526.

Agampodi SB, Thevanesam V, Wimalarathna H, Senarathna T, Wijedasa MH. 2008. A preliminary study on prevalent serovars of leptospirosis among patients admitted to teaching hospital, Kandy, Sri Lanka. Indian Journal of Medical Microbiology. 26:405-406.

Agampodi SB, Agampodi TC, Thalagala E, Fernando S, Perera S, Chandrarathna S. 2010. Do People Know Adequately About Leptospirosis? A knowledge assessment survey in post outbreak situation–Sri Lanka. International Journal of Preventive Medicine. 1(3).

Aldstadt J. 2007. An incremental Knox test for the determination of the serial interval between successive cases of an infectious disease. Stochastic Environmental Research and Risk Assessment. 21:487-500.

Andersson E, Bock B, Frisen M. 2008. Modeling influenza incidence for the purpose of on-line monitoring. Statistical Methods in Medical Research. 17:421-438.

Anselin L. 1995. Local indicators of spatial association-LISA. Geographical Analysis. 27:93-115.

Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, et al.. Above the clouds: A berkeley view of cloud computing. EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28: 2009.

Aylin P, Maheswaran R, Wakefield J, Cockings S, Jarup L, Arnold R, et al. 1999. A national facility for small area disease mapping and rapid initial assessment of apparent disease clusters around a point source: the UK Small Area Health Statistics Unit. Journal of Public Health. 21:289-298.

Banks AL. 1959. The study of the geography of disease. Geographical Journal 125:199-210.

Banos A, Lacasa J. 2007. Spatio-temporal exploration of SARS epidemic. Cybergeo. 408.

Barcellos C, Sabroza PC. 2000. Socio-environmental determinants of the leptospirosis outbreak of 1996 in western Rio de Janeiro: a geographical approach. International Journal of Environmental Health Research. 10:301-313.

Barcellos C, Sabroza PC. 2001. The place behind the case: leptospirosis risks and associated environmental conditions in a flood-related outbreak in Rio de Janeiro. Cadernos de Saúde Pública. 17:59-67.

Bernabe-Ortiz A, Curioso W, Gonzales M, Evangelista W, Castagnetto J, Carcamo C, et al. 2008. Handheld computers for self-administered sensitive data collection: A comparative study in Peru. BMC Medical Informatics and Decision Making. 8:11.

Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M. 1995. Bayesian analysis of space-time variation in disease risk. Statistics in Medicine. 14:2433-2443.

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

Best N, Richardson S, Thomson A. 2005. A comparison of Bayesian spatial models for disease mapping. Statistical Methods in Medical Research.14:35-59.

Bharti AR, Nally JE, Ricaldi JN, Matthias MA, Diaz MM, Lovett MA, et al. 2003. Leptospirosis: a zoonotic disease of global importance. The Lancet Infectious Diseases. 3:757-771.

Bithell JF. 1990. An application of density estimation to geographical epidemiology. Statistics in Medicine. 9:691-701.

Block R. 2007. Scanning for clusters in space and time. Social Science Computer Review. 25:272-278.

Bock R, Jackson L, De Vos A, Jorgensen W. 2004. Babesiosis of cattle. Parasitology. 129:S247-S269.

Bogh C, Lindsay SW, Clarke SE, Dean A, Jawara M, Pinder M, et al. 2007. High spatial resolution mapping of malaria transmission risk in The Gambia, West Africa, using landsat TM satellite imagery. American Journal Of Tropical Medicine And Hygiene. 76:875-881.

Bovet P, Yerson C, Merien F, Davis CE, Perolat P. 1999. Factors associated with clinical leptospirosis: a population-based case-control study in the Seychelles (Indian Ocean). International Journal of Epidemiology. 28:583-590.

Bradley CA, Rolka H, Walker D, Loonsk J. 2005. BioSense: implementation of a national early event detection and situational awareness system. MMWR Morbidty and Mortality Weekly Report. 54(Suppl):11-19.

Britch SC, Linthicum KJ, Rift Valley Fever Working Group. 2007. Developing a research agenda and a comprehensive national prevention and response plan for Rift Valley Fever in the United States. Emerging Infectious Diseases. 13:8.

Buchanan AV, Weiss KM, Fullerton SM. 2006. Dissecting complex disease: the quest for the Philosopher's Stone? International Journal of Epidemiology. 35:562-571.

Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore, AW. 2005. Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics. 38:99-113.

Buehler JW, Berkelman RL, Hartley DM, Peters CJ. 2003. Syndromic surveillance and bioterrorism-related epidemics. Emerging Infectious Diseases. 9:1197-1204.

Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. 2004. Framework for evaluating public health surveillance systems for early detection of outbreaks. MMWR Morbidity and Mortality Weekly Report. 53:1–11.

Cachay ER, Vinetz JM. 2005. A global research agenda for leptospirosis. Journal of postgraduate medicine. 51:174.

Canary Database: Animals as Sentinels of Human Environmental Health Hazards. [Internet] New Haven (CT): Yale University Occupational and Environmental Medicine. [cited 2009 Nov 23]. Available from: http://www.canarydatabase.org/.

Childs JE, Curns AT, Dey ME, Real LA, Feinstein L, Bjornstad ON. 2000. Predicting the local dynamics of epizootic rabies among raccoons in the United States. Proceedings of the National Academy of Sciences USA. 97:13666-13671.

Chretien J, Burkom HS, Sedyaningsih ER, Larasati RP, Lescano AG, Mundaca CC, et al. 2008. Syndromic surveillance: Adapting innovations to developing settings. PLoS Medicine.5:0367-0372.

Clement J, Maes P, Muthusethupathi C, Nainan G, vanRanst M. 2006. First evidence of fatal hantavirus nephropathy in India, mimicking leptospirosis. Nephrology Dialysis Transplantation. 21:826 -827.

Cliff AD, Haggett P. 1993. Statistical modelling of measles and influenza outbreaks. Statistical Methods in Medical Research. 2:43-73.

Costa MA., Kulldorff M, Assunção RM. 2007. A space time permutation scan statistic with irregular shape for disease outbreak detection. Advances in Disease Surveillance. 4:86.

Das D, Metzger K, Heffernan R, Balter S, Weiss D, Mostashari F. 2005. Monitoring over-the-counter medication sales for early detection of disease outbreaks—New York City. MMWR Morbidity and Mortality Weekly Report. 54:41–46.

Dassanayake D, Wimalaratna H, Agampodi S, Liyanapathirana V, Piyarathna T, Goonapienuwala B. 2009. Evaluation of surveillance case definition in the diagnosis of leptospirosis, using the Microscopic Agglutination Test: a validation study. BMC Infectious Diseases. 9:48.

Davis S, Calvet E. 2005. Fluctuating rodent populations and risk to humans from rodent-borne zoonoses. Vector-Borne & Zoonotic Diseases. 5:305-314.

Diero L, Rotich J, Bii J, Mamlin B, Einterz R, Kalamai I, et al. 2006. A computer-based medical record system and personal digital assistants to assess and follow patients with respiratory tract infections visiting a rural Kenyan health centre. BMC Medical Informatics and Decision Making. 6:21.

Diggle P, Rowlingson B, TingLi S. 2005. Point process methodology for on-line spatio-temporal disease surveillance. Environmetrics. 16:423-434.

Diggle P. 2003. Statistical analysis of spatial point patterns. Academic Press Inc, London, UK. 159 pp.

Diggle PJ, Chetwynd AG, Haggkvist R, Morris SE. 1995. Second-order analysis of space-time clustering. Statistical Methods in Medical Research. 4:124-136.

Doherr MG. 2000. Monitoring and surveillance for rare health-related events: a review from the veterinary perspective. Philosical Transactions of the Royal Society of London B Biological Science. 356:1097-1106.

Domroes M, Ranatunge E. 1993. Analysis of inter-station daily rainfall correlation during the Southwest Monsoon in the Wet Zone of Sri Lanka. Geografiska Annaler. Series A Physical Geography. 75:137-148.

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

Edge VL, Pollari F, Ng LK, Michel P, McEwen SA, Wilson JB, Jerrett M, Sockett PN, Martin SW. 2006. Syndromic surveillance of Norovirus using over-the-counter sales of medications related to gastrointestinal illness. Canadian Journal of Infectious Diseases and Medical Microbiology. 17:235-241.

Ekpo UF, Mafiana CF, Adeofun CO, Solarin AR, Idowu AB. 2008. Geographical information system and predictive risk maps of urinary schistosomiasis in Ogun State, Nigeria. BMC Infectious Diseases. 8:74.

Eubank S, Guclu H, Anil Kumar VS, Marathe MV, Srinivasan A, Toroczkai Z et al. 2004. Modelling disease outbreaks in realistic urban social networks. Nature. 429:180-184.

Farrington CP, Andrews N, Beale AD, Catchpole MA. 1996. A statistical algorithm for the early detection of outbreaks of infectious disease. Journal of the Royal Statistical Society Series A. 159:547-563.

Fearnley L. 2008. Signals come and go: syndromic surveillance and styles of biosecurity. Environment and Planning A. 40:1615-1632.

Fèvre EM, Coleman PG, Odiit M, Magona JW, Wellburn SC, Woolhouse MEJ. 2001.The origins of a new Trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern Uganda. The Lancet. 358:625-628.

Fricker Jr RD, Rolka H. 2006. Protecting against biological terrorism: statistical issues in electronic biosurveillance. Chance. 91:4-13.

Frisen M, Sonesson C. 2005. Optimal surveillance. In: Lawson AB, Kleinman K, editors. Spatial and syndromic surveillance for public health. John Wiley, West Sussex, UK. pp.31-52.

Gamage CD, Samaraweera S, Matibag GC, Obayashi Y, Tamashiro H. 2009. Leptospirosis surveillance in Sri Lanka, 2005-2008. 2009;13(1):21-21.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. Bayesian data analysis. Chapman & Hall/CRC, London, UK. 668 pp.

Getis A., Ord C. 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis. 24:189-206.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009. Detecting influenza epidemics using search engine query data. Nature. 457:1012-1014.

Goodchild, MF. 2007. Citizens as sensors: the world of volunteered geographic information. GeoJournal. 69:211-221.

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

neutral models: the case of lung cancer in Long Island, New York. International Journal of Health Geographics. 3:14.

Gotway C, Young L. 2002. Combining incompatible spatial data. Journal of the American Statistical Association. 97:632-648.

Government of Alberta, Agriculture and Rural Development. Alberta Veterinary Surveillance Network. 2010 Jan 7 [accessed 2010 May 1]. Available from: http://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/afs10440

Greger M. 2007. The human/animal interface: emergence and resurgence of zoonotic infectious diseases. Critical Reviews in Microbiology. 33:243-299.

Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, et al. 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science. 302:276-278.

Hägerstrand T. 1967. Innovation diffusion as a spatial process. University of Chicago Press, Chicago USA. 334 pp.

Haggett P. 1994. Geographical Aspects of the Emergence of Infectious Diseases. Geografiska Annaler. Series B, Human Geography. 76:91-104.

Haggett P. 1992. Sauer's' Origins and dispersals': its implications for the geography of disease. Transactions of the Institute of British Geographers. 17:387-398.

Halliday JEB, Meredith AL, Knobel DL, Shaw DJ, Bronsvoort BMDC, Cleaveland S. 2007. A framework for evaluating animals as sentinels for infectious disease surveillance. Journal of the Royal Society Interface. 4:973-984.

Haydon DT, Cleaveland S, Taylor LH, Laurenson MK. 2002. Identifying reservoirs of infection: a conceptual and practical challenge. Emerging Infectious Diseases. 8:1468-1473.

Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. 2004. Syndromic surveillance in public health practice, New York City. Emerging Infectious Diseases. 10:858-864.

Held L, Hofmann M, Hohle M, Schmid V. 2006. A two-component model for counts of infectious diseases. Biostatistics. 7:422-437.

Held L, Hohle M, Hofmann M. 2005. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Statistical Modelling. 5:187-199.

Higgins R. 2004. Emerging or re-emerging bacterial zoonotic diseases: bartonellosis, leptospirosis, Lyme borreliosis, plague. Revue Scientifique et Technique-Office International des Epizooties. 23:569-582.

Hindrichsen S, Medeiros A, Clement J, Leirs H, McKenna P, Matthys P, et al. 1993. Hantavirus infection in Brazilian patients from Recife with suspected leptospirosis. The Lancet. 341:50.

Höhle M. 2007. Surveillance: An R package for the monitoring of infectious diseases. Computational Statistics. 22:571-582.

Horner J. 2009. rapache: Web application development with R and Apache. 2009. Available from: [http://biostat.mc.vanderbilt.edu/rapache/]

Hotez PJ. 2008. Forgotten People and Forgotten Diseases. ASM Press, Washington D.C., USA. pp. 215.

Hulth A, Rydevik G, Linde A. 2009. Web queries as a source for syndromic surveillance. PLoS one. 4:2.

Jacquez G. 1996. A k nearest neighbour test for space-time interaction. Statistics in Medicine. 15:1935-1949.

Jacquez GM, Meliker JR. 2009. Case-control clustering for mobile populations. In: Fotheringham SA, Rogerson, PA, editors. The SAGE Handbook of Spatial Analysis. Sage, London, UK. pp. 355-374.

Jacquez GM, Greiling DA, Durbeck H, Estberg L, Do E, Long A, et al. 2002. ClusterSeer User Guide 2: Software for identifying disease clusters. Ann Arbor, MI: TerraSeer Press; 2002.

Jacquez GM, Meliker J, Kaufmann A. 2007. In search of induction and latency periods: Space-time interaction accounting for residential mobility, risk factors and covariates. International Journal of Health Geographics. 6:35.

Järpe E. 1999. Surveillance of the interaction parameter of the ising model. Communications in Statistics - Theory and Methods. 28:3009-3027.

Johnson G. 2008. Prospective spatial prediction of infectious disease: experience of New York State (USA) with West Nile Virus and proposed directions for improved surveillance. Environmental and Ecological Statistics. 15:293-311.

Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. 2008. Global trends in emerging infectious diseases. Nature. 451:990-993.

Kahn LH. 2006. Confronting zoonoses, linking human and veterinary medicine. Emerging Infectious Diseases. 12:556-61.

Kim A, Martinez A, Klausner J, Goldenson J, Kent C, Liska S, et al. 2008. Use of sentinel surveillance and geographic information systems to monitor trends in HIV prevalence, incidence, and related risk behavior among women undergoing syphilis screening in a jail setting. Journal of Urban Health. 86:79-92.

Kim Y, O'Kelly M. 2008. A bootstrap based space--time surveillance model with an application to crime occurrences. Journal of Geographical Systems. 10:141-165.

Kitron UL, Otieno H, Hungerford LL, Odulaja A, Brigham WU, Okello OO, et al. 1996. Spatial analysis of the distribution of tsetse flies in the Lambwe Valley, Kenya, using Landsat TM satellite imagery and GIS. Journal of Animal Ecology. 65:371-380.

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

Kleinman KP, Abrams AM, Kulldorff M, Platt R. 2005. A model-adjusted space–time scan statistic with an application to syndromic surveillance. Epidemiology and Infection. 133:409-419.

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

Koizumi N, Gamage CD, Muto M, Kularatne SAM, Budagoda SBDS, Rajapakse JRPV, et al. 2009. Serological and Genetic Analysis of Leptospirosis in Patients with Acute Febrile Illness in Kandy, Sri Lanka. Japan. Journal of Infectious Diseases. 62:474-475.

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

Kulldorff M, Hjalmars U. 1999. The Knox method and other tests for space-time interaction. Biometrics. 55:544-552.

Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. 2005. A space-time permutation scan statistic for disease outbreak detection. PLoS Medicine. 2:e59.

Kulldorff M, Information Management Services Inc. 2009. SaTScan™ v8.0: Software for the spatial and space-time scan statistics. Availabe from: [www.satscan.org].

Kulldorff M., Information Management Services, 2010. SaTScan, 2010.

Kulldorff M. 2001. Prospective time periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society Series A. 164:61-72.

Langran G. 1992. Time in Geographic Information Systems. Taylor Francis, New York, USA. pp. 189.

Lawson AB, Kleinman K, eds. 2005. Spatial and Syndromic Surveillance for Public Health. John Wiley, West Sussex, UK. pp. 269.

Lawson AB, Williams FLR. 1993. Applications of extraction mapping in environmental epidemiology. Statistics in Medicine. 12:1249-1258.

Lawson AB. 2008. Bayesian disease mapping: hierarchical modeling for spatial epidemiology. CRC Press, New York, USA. pp. 344

Lawson AB. 2005. Spatial and spatio-temporal disease analysis. In: Lawson AB, Kleinman K, editors. Spatial and syndromic surveillance for public health. John Wiley, West Sussex, London, UK. pp. 55-75.

Le Strat Y, Carrat F. 1999. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in Medicine. 18:3463-3478.

Leblond, A, Hendrikx P, Sabatier P. 2007. West nile virus outbreak detection using syndromic monitoring in horses. Vector-Borne and Zoonotic Diseases. 7:403-410.

Legendre P, Fortin MJ. 1989. Spatial pattern and ecological analysis. Vegetatio. 80:107-138.

Lescano AG, Larasati RP, Sedyaningsih ER, Bounlu K, Araujo-Castillo RV, Munayco-Escate CV, et al. 2008. Statistical analyses in disease surveillance systems. BMC Proceedings. 2(Suppl 3):S7.

Levett PN. 2001. Leptospirosis. Clinical Microbiology Reviews. 14:296-326.

Light RU. 1944. The progress of medical geography. Geographical Review. 34:636-641.

Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing.10:325-337.

Maciel EAP, de Carvalho AL, Nascimento SF, de Matos RB, Gouveia EL, Reis MG, et al. 2008. Household transmission of Leptospira infection in urban slum communities. PLoS Neglected Tropical Diseases. 2(1).

MacNab YC, Gustafson P. 2007. Regression B-spline smoothing in Bayesian disease mapping: with an application to patient safety surveillance. Statistics in Medicine. 26:4455-4474.

MacNab YC. 2003. A Bayesian hierarchical model for accident and injury surveillance. Accident Analysis and Prevention. 35:91-102.

MacNab YC. 2007a. Spline smoothing in Bayesian disease mapping. Environmetrics. 18:727-744.

MacNab YC. 2007b. Mapping disability-adjusted life years: A Bayesian hierarchical model framework for burden of disease and injury assessment. Statistics in Medicine. 26:4746-4769.

Madigan, D. 2005. Bayesian data mining for health surveillance. In: Lawson AB, Kleinman K, editors. Spatial and syndromic surveillance for public health. John Wiley, West Sussex, London, UK. pp. 203–221.

Madsen T Shine R. 1999. Rainfall and rats: climatically-driven dynamics of a tropical rodent population. Austral Ecology. 24:80-89.

Malone JB, Bergquist NR, Huh OK, Bavia ME, Bernardi M, El Bahy MM, et al. 2001. A global network for the control of snail-borne disease using satellite surveillance and geographic information systems. Acta Tropica. 79: 7-12.

Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F. 2004. Implementing syndromic surveillance: a practical guide informed by the early experience. Journal of the American Health Information Management Association. 11:141-150.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research. 27:209-220.

Martínez-Beneito, MA., Conesa D, López-Quílez A, López-Maside, A. 2008. Bayesian Markov switching models for the early detection of influenza epidemics. Statistics in Medicine. 27:4455-4468.

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

Meliker JR, Jacquez GM. 2007. Space–time clustering of case–control data with residential histories: insights into empirical induction periods, age-specific susceptibility, and calendar year-specific effects. Stochastic Environmental Research and Risk Assessment. 21:625-634.

Mills JN. 1999. The Role of Rodents in Emerging Human Disease: Examples from the Hantaviruses and Arenaviruses. In Singleton GR, Hinds LA, Leirs H, Zhang Z. eds. Ecologically based management of rodent pests. Australian Centre for International Agricultural Research. Canberra, Australia. pp.134-160.

Missinou M, Ololа C, Issifou S, Matsiegui P, Adegnika A, Borrmann S, et al. 2005. Short Report: Piloting Paperless Data Entry For Clinical Research In Africa. American Journal of Tropical Medicine and Hygiene. 72:301-303.

Mobile Active. Berhane Gebru: Disease surveillance with mobile phones in Uganda. 2008 Jul 30 [cited 2010 May 19]. Available from http://mobileactive.org/berhane-gebru-disease-surveillance-mobile-phones-uganda

Moffett A, Shackelford N, Sarkar S. 2007. Malaria in Africa: Vector Species' Niche Models and Relative Risk Maps. PLoS ONE. 2(9).

Moore K, Edge G, Kurc A. 2008. Visualization techniques and graphical user interfaces in syndromic surveillance systems. Summary from the Disease Surveillance Workshop, Sept. 11-12, 2007; Bangkok, Thailand. BMC Proceedings. 2:S6.

Mostashari F, Fine A, Das D, Adams J, Layton M. 2003. Use of ambulance dispatch data as an early warning system for communitywide influenzalike illness, New York City. Journal of Urban Health. 80:i43-i49.

Nagarwalla N. 1996. A scan statistic with a variable window. Statistics in Medicine. 15:845-850.

Naus JI. 1965. The distribution of the size of the maximum cluster of points on a line. Journal of the American Statistical Association. 60:532-538.

Ndiaye SM., Quick L, Ousmane S, Seydou N. 2003. The value of community participation in disease surveillance: a case study from Niger. Health Promotion International. 18:89-98.

Nityananda K Sulzer CR. 1969. A new leptospiral serotype in the Javanica serogroup from Ceylon. Tropical and Geographical Medicine. 21:207-209.

Odiit M, Bessell PR, Fevre EM, Robinson T, Kinoti J, Coleman PG, et al. 2006. Using remote sensing and geographic information systems to identify villages at high risk for rhodesiense sleeping sickness in Uganda. Transactions of the Royal Society of Tropical Medicine and Hygiene. 100:354-362.

Openshaw S, Charlton ME, Wymer C, Craft A. 1987. A mark 1 geographical analysis machine for the automated analysis of point data sets. International Journal of Geographic Information Systems. 1:335-358.

Ostfeld R, Glass G, Keesing F. 2005. Spatial epidemiology: an emerging (or re-emerging) discipline. Trends in Ecology & Evolution. 20:328-336.

Pappas G, Papadimitriou P, Siozopoulou V, Christou L, Akritidis N. 2008. The globalization of leptospirosis: worldwide incidence trends. International Journal of Infectious Diseases. 12:351-357.

Patil GP, Taillie C. 2004. Upper level set scan statistics for detecting arbitrarily shaped hotspots. Environmental and Ecological Statistics.11:189-197.

Peltonen M, Liebhold AM, Bjornstad O, Williams DW. 2002. Spatial synchrony in forest insect outbreak roles of regional stochasticity and dispersal. Ecology. 83:3120-3129.

Perry A. 2009. Forecasting hospital emergency department visits for respiratory illness using ontario's telehealth system: an application of real-time syndromic surveillance to forecasting health services demand. MSc Thesis Queen's University, Department of Community Health and Epidemiology.

Phulsuksombati D, Sangjun N, Khoprasert Y, Kingnate D, Tangkanakul W. 2001. Leptospires in rodent, northeastern region 1999-2000. Journal of Health Science. 10:516–525.

Pike MC, Smith PG. 1974. A case-control approach to examine diseases for evidence of contagion, including diseases with long latent periods. Biometrics. 30:263-279.

Population Data BC. 2008. Setting a health geomatics education and training agenda Victoria, BC, Canada.

Rabinowitz P, Scotch M, Conti L. 2009. Human and animal sentinels for shared health risks. Veterinary Italia. 45:23-34.

Rabinowitz PM, Gordon Z, Holmes R, Taylor B, Wilcox M, Chudnov D, et al. 2005. Animals as sentinels of human environmental health hazards: an evidence-based analysis. EcoHealth. 2:26-37.

Rabinowitz, MacGarr P, Odofin L, Dein FJ. 2008. From "us vs. them" to "shared risk": can animals help link environmental factors to human health? EcoHealth. 5:224-229.

Radostits OM, Leslie KE, Fetrow J. 1994. Herd health: Food animal production medicine. W.B. Saunders Company, Philadelphia, USA. pp. 631.

Rath T, Carreras M, Sebastiani P. 2003. Automated detection of influenza epidemics with Hidden Markov Models. In Advances in Intelligent Data Analysis, eds: Berthold MR, Lenz HJ, Bradley E, Kruse R, Borgelt C. Berlin, Germany, August 28-30, 2003 Proceedings. Springer:New York. pp. 521-532.

Reingold A. 2003. If syndromic surveillance is the answer, what is the question? Biosecurity and Bioterrorism: Biodefense strategy, practice, and science. 1:77-81.

Reis BY, Kirby C, Hadden LE, Olson K, McMurry AJ, Daniel JB, et al. 2007. AEGIS: A Robust and scalable real-time public health surveillance system. Journal of the American Medical Informatics Association. 14:581-588.

Reis BY, Kohane IS, Mandl KD. 2007. An epidemiological network model for disease outbreak detection. PLoS Medicine. 4:e210.

Reis RB, Ribeiro G, Felzemburgh R, Santana F, Mohr S, Melendez A, et al. 2008. Impact of Environment and Social Gradient on Leptospira Infection in Urban Slums. PLoS Neglected Tropical Diseases. 2:e228.

Richards TB, Croner CM, Rushton G, Brown CK, Fowler L. 1999. Geographic information systems and public health: mapping the future. Public Health Reports. 114:359-360.

Robertson C, Nelson TA, MacNab YC, Lawson AB. 2010. Review of methods for space-time disease surveillance. Spatial and Spatio-temporal Epidemiology. 1:105-116.

Robertson C, Sawford K, Daniel SLA, Nelson TA, Stephen C. 2010. Mobile surveillance system, Sri Lanka. Emerging Infectious Diseases. 15:1524.

Rogers DJ, Hay SL, Packer MJ. 1996. Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. Annals of Tropical Medicine and Parasitology. 90:225-242.

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

Rogerson P. 2005a. A set of associated statistical tests for spatial clustering. Environmental and Ecological Statistics. 12:275-288.

Rogerson P. 2005b. Monitoring spatial maxima. Journal of Geographical Systems. 7:101-114.

Rogerson PA., Yamada I. 2004. Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. Statistics in Medicine. 23:2195-2214.

Rossi G, Lampugnani L, Marchi M. 1999. An approximate CUSUM procedure for surveillance of health events. Statistics in Medicine.18:2111-2122.

Rushton G. 2003. Public health, GIS, and spatial analytic tools. Annual Review of Public Health. 24:43-56.

RWeb Project. 2009. Available from: [http://www.math.montana.edu/Rweb]

Sabel CE, Boyle PJ, Loytonen M, Gatrell AC, Jokelainen M, Flowerdew R, et al. 2003. Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death. American Journal of Epidemiology. 157: 898-905.

Sanders EJ, Rigau-Perez JG, Smits HL, Deseda HL, Vorndam VA, Aye T, et al. 1999. Increase of leptospirosis in dengue-negative patients after a hurricane in Puerto Rico in 1996. The American Journal of Tropical Medicine and Hygiene. 61:399-404.

Schaerstrom A. 1999. Apparent and actual disease landscapes. Some reflections on the geographical definition of health and disease. Geografiska Annaler: Series B. 81:235-242.

Sejvar J, Bancroft E, Winthrop K, Bettinger J, Bajani M, Bragg S, et al. 2003. Leptospirosis in" Eco-Challenge" athletes, Malaysian Borneo, 2000. Emerging Infectious Diseases. 9:702-707.

Sharma S, Vijayachari, P, Sugunan AP, Natarajaseenivasan K, Seghal AC. 2006. Seroprevalence of leptospirosis among high-risk population of Andaman Islands, India. The American Journal of Tropical Medicine and Hygiene. 74:278-283.

Shirima K, Mukasa O, Schellenberg J, Manzi F, John D, Mushi A, et al. 2007. The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania. Emerging Themes in Epidemiology. 4:5.

Siegmund D. 1985. Sequential analysis: tests and confidence intervals. Springer-Verlag, New York, USA. pp. 272.

Smith GJD, Vijaykrishna D, Bahl J, Lycett SL, Worobey M, Pybus OG, et al. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 459:1122-1125.

Sonesson C, Bock D. 2003. A review and discussion of prospective statistical surveillance in public health. Journal of the Royal Statistical Society Series A. 166:5-21.

Sosin DM, DeThomasis J. 2004. Evaluation challenges for syndromic surveillance—making incremental progress. MMWR Morbidity and Mortality Weekly Report. 53: Suppl:125-129.

Sri Lanka Department of Provincial Health Services. Annual Health Bulletin, Central Province 2007, Sri Lanka: 2008.

Sri Lanka Epidemiology Unit. An interim analysis of leptospirosis outbreak in Sri Lanka - 2008, Colombo, Sri Lanka: Ministry of Healthcare Nutrition: 2008.

Stoto G, Araujo-Castillo R, Neyra J, Fernandez M, Leturia C, Mundaca C, et al. 2008. Challenges in the implementation of an electronic surveillance system in a resource-limited setting: Alerta, in Peru. BMC Proceedings. 2(Suppl 3):S4.

Stoto MA, Schonlau M, Mariano LT. 2004. Syndromic surveillance: is it worth the effort? Chance. 17:19-24.

Sugunan AP, Vijayachari P, Sharma S, Roy, S, Manickam P, Natarajaseenivasan, K, et al. 2009. Risk factors associated with leptospirosis during an outbreak in Middle Andaman, India. The Indian Journal of Medical Research. 130:67-73.

Sullivan ND. 1974. Leptospirosis in animals and man. Australian Veterinary Journal. 50:216-223.

Sun W, Cai T. 2009. Large-scale multiple testing under dependence. Journal of the Royal Statistical Society Series B Statistical Methodology. 71:393-424.

Takahashi K, Kulldorff M, Tango T, Yih K. 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. International Journal of Health Geographics. 7:14.

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

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

Tassinari WS, Pellegrini DCP, Sa CBP, Reis RB, Ko AI, Carvalho MS. 2008. Detection and modelling of case clusters for urban leptospirosis. Tropical Medicine & International Health. 13:503-512.

Teutsch SM, Churchill RE. 2000. Principles and practice of public health surveillance. Oxford University Press, Oxford, UK. pp. 406.

Thaipadungpanit, J, Wuthiekanun V, Chierakul W, Smythe LD, Petkanchanapong W, Limpaiboon R, et al. 2007. A dominant clone of Leptospira interrogans associated with an outbreak of human leptospirosis in Thailand. PLoS Neglected Tropical Diseases. 1:e56.

Tsui FC, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM. 2003. Technical description of RODS: a real-time public health surveillance system. Journal of the American Medical Informatics Association. 10:399-408.

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

Tzala E, Best N. 2008. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. Statistical Methods in Medical Research. 17:97-118.

Ugarte MD, Goicoa T, Militino AF. 2010. Spatio-temporal modeling of mortality risks using penalized splines. Environmetrics. 21:270-289.

Van Metre, DC, Barkey, DQ, Salman, MD Morley, PS. 2009. Development of a syndromic surveillance system for detection of disease among livestock entering an auction market. Journal of the American Veterinary Medical Association. 234:658-664.

Vidal Rodeiro CL, Lawson AB. 2006a. Monitoring changes in spatio-temporal maps of disease. Biometrical Journal. 48:463-480.

Vidal Rodeiro CL, Lawson AB. 2006b. Online updating of space-time disease surveillance models via particle filters. Statistical Methods in Medical Research. 15:423-444.

Vinetz JM, Glass GE, Flexner CE, Mueller P, Kaslow DC. 1996. Sporadic urban leptospirosis. Annals of Internal Medicine. 125:794-798.

Vinetz JM, Wilcox BA, Aguirre A, Gollin LX, Katz AR, Fujioka RS, et al. 2005. Beyond Disciplinary Boundaries: Leptospirosis as a Model of Incorporating Transdisciplinary Approaches to Understand Infectious Disease Emergence. EcoHealth. 2:291-306.

Vital Wave Consulting. 2009. mHealth for Development: The opportunity of mobile technology for healthcare in the developing world. Washington, D.C. and Berkshire, UK: UN Foundation-Vodafone Foundation Partnership.

Vrbova L, Stephen C, Kasman N, Boehnke R, Doyle-Waters M, Chablitt-Clark A, et al. 2010. Systematic review of surveillance systems for emerging zoonoses. Transboundary and Emerging Diseases. 57:154-161.

Wagner MM, Moore AW, Aryel RM. 2006. Handbook of Biosurveillance. Elsevier, London, UK. pp. 605.

Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA, et al. 2001. The emerging science of very early detection of disease outbreaks. Journal of Public Health Management and Practice. 7:51-9.

Waitkins SA. 1986. Leptospirosis as an occupational disease. British Journal of Industrial Medicine. 43:721-725.

Wall MM, Li R. 2009. Multiple indicator hidden Markov model with an application to medical utilization data. Statistics in Medicine. 28:293-310.

Watkins R, Eagleson S, Veenendaal B, Wright G, Plant A. 2009. Disease surveillance using a hidden Markov model. BMC Medical Informatics and Decision Making. 9:39.

Wilson ML. 2002. Emerging and vector-borne diseases: Role of high spatial resolution and hyperspectral images in analyses and forecasts. Journal of Geographical Systems. 4:31-42.

Wiehe SE, Carroll AE, Liu GC, Haberkorn KL, Hoch SC, Wilson JS, Fortenberry JD. 2008. Using GPS-enabled cell phones to track the travel patterns of adolescents. International Journal of Health Geographics. 7:22

Wong WK, Moore AW. 2006. Classical time-series methods for biosurveillance. In: Wagner MM, Moore AW, Arye RM, editors. Handbook of Biosurveillance.London: Elsevier Academic Press; pp. 217-234.

Woodall WH, Ncube MM. 1985. Multivariate CUSUM quality-control procedures. Technometrics. 27:285-292.

World Health Organization. Global Early Warning System for Major Animal Diseases, including Zoonoses (GLEWS) 2007 WHO, Geneva. Available from: http://www.who.int/zoonoses/outbreaks/glews/en/.

Xia H, Carlin BP. 1998. Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. Statistics in Medicine. 17:2025-2043.

Yamada I, Rogerson P, Lee G. 2009. GeoSurveillance: a GIS-based system for the detection and monitoring of spatial clusters. Journal of Geographical Systems. 11:155-173.

Yan P, Clayton MK. 2006. A cluster model for space-time disease counts. Statistics in Medicine. 25:867-881.

Yan P, Zeng D, Chen H. 2006. A review of public health syndromic surveillance systems. Lecture Notes in Computer Science. 3975:249-260.

Zubair L. 2002. El Niño-southern oscillation influences on rice production in Sri Lanka. International Journal of Climatology. 22:249-260.

Zucchini W, MacDonald IL. 2009. Hidden Markov Models for Time Series: An Introduction Using R. Chapman & Hall/CRC, London, UK. pp. 275.