ANALYZING AND PREDICTING RISKS OF INFECTIOUS DISEASES BY GEOGRAPHIC INFORMATION SCIENCE

WANG BIN

Ph.D

The Hong Kong Polytechnic University

2015

The Hong Kong Polytechnic University

Department of Land Surveying & Geo-Informatics

Analyzing and Predicting Risks of Infectious Diseases by Geographic Information Science

WANG Bin

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

October, 2014

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

(Signed)

__________WANG Bin__________(Name of student)

Abstract

Epidemic waves of new emerging infectious diseases have awakened global concerns regarding their potential pandemic threats. Early prevention and control measures can prevent the spread or even control the outbreak of an infectious disease. Geographic information science (GIS) is used in this study as an integrated platform for epidemic surveillance. Combined with powerful spatial data management and statistical analysis methods, geospatial technologies offer a new perspective to model how a disease may spread together with its evolutionary path. This tool enables policy makers to explore the spatial interactions between disease emergence and its risk or protective factors, thereby allowing epidemiologists or public health officials to target areas with more effective means to control a disease spread.

This research aims to develop innovative models within the GIS framework for characterizing the spatial and temporal distribution patterns of an infectious disease, to assist the planning of preventive intervention measures. Firstly, it summarizes background, methods and research developments in emerging infectious diseases. Chapter 2 gives a description of relevant experimental data in GIS and Epidemiology based on H1N1 of Hong Kong in 2009, H7N9 of Mainland China in 2013, and the Ebola epidemic of West Africa in 2014. An in-depth discussion of elementary analysis methods - Standard Deviational Ellipse (SDE), is introduced in Chapter 3, using H1N1 infection of Hong Kong to highlight the spatiotemporal concentrations.

Mathematics has long been a powerful tool for understanding and assessing the disease spread. Understanding the how, when, and why an epidemic spreads across a geographic landscape is of critical importance, as effective preventive measures

can be put in place before a disaster occurs. Chapter 4 devotes to discussing how the temporal dynamics of infectious disease are modeled by basic SIR compartmental models, and how the meta-population model is used to characterize the spatiotemporal movement of a disease infection. The typical reaction diffusion equation models are also thoroughly explored, followed by a detailed description of computer implementation procedures using the Runge-Kutta method.

Spatiotemporal analysis can potentially contribute to characterizing the temporal evolution process and revealing possible spatial propagation patterns. As such, an innovative approach was proposed in Chapter 5 to examine the impact of spatiotemporal proximity upon the onset risk prediction of an emerging infectious disease. Experiments based on the avian influenza A H7N9 that occurred in eastern China from February to May 2013 demonstrated that such spatiotemporal proximity integrated approach was capable of providing approximately 70% correct prediction on average in predicting the H7N9 illness onset risk for the 5 days following the forecast date.

Furthermore, a sequential Bayesian inference combined with stochastic SEIR model has been employed to estimate the time-varying effective reproduction numbers, together with their 95% confidence intervals, for the Ebola virus epidemic in West Africa. Experimental results indicated that concerted efforts should be made to halt all transmission in Liberia for the dreadful reproduction number there. Based on the aforementioned theoretical models, a software prototype framework has resulted for further development and to enable the analysis of spatiotemporal spreading patterns and dynamic evolution trends of an infectious disease. Discussions regarding the limitations and potentialities of the models explored here are also carried out for guiding future research work to better prevent the infectious disease spread.

Acknowledgements

I am grateful to those who are involved in this research project. My enormous gratitude goes first and foremost to Professor Wenzhong (John) SHI, my chief supervisor, for his constant encouragement and guidance during the whole PhD study period. Not only has he opened the door of academic research for me, he has also set up an admirably example of a responsible, intelligent scholar and a positive, well-rounded, successful person for me to follow. Particularly, he has walked me throughout all the stages of the writing of this thesis. Without his consistent and illuminating instructions, this thesis could not have reached its present form. Besides, I highly appreciate his recommendation letter and great efforts in the application of my doctoral grant.

A particular debt of profound gratitude is owed to the financial support from the AXA Research Fund for me pursuing the PhD degree in PolyU. Special thanks go to Xun SHI from Dartmouth College, Ross Ka-Kit LEUNG from the Chinese University of Hong Kong and Lilian PUN from LSGI, PolyU for their generosity of providing me some data related to this study, as well as Dr. P.C. LAI from the Hong Kong University for her helpful recommended reading materials. Meanwhile, I also sincerely thank the two external examiners for their very constructive comments and suggestions which have greatly improved this work. Besides, I would also like to relish this opportunity to thank the administrative staff members in the Department of Land Surveying and Geo-informatics for their support of this study.

Perpetual gratitude is expressed to my previous mentors, Associate Professor Jianhua HU from China University of Mining & Technology of my undergraduate study and Professor Bo YU from Dalian University of Technology during me pursuing the Master degree, under the guidance of whom I had undergone solid mathematical training. I am also grateful of my friends Fu LIU just graduated from City University of Hong Kong, Teng ZHONG studying in Hong Kong University, Qunhu SUN in Shenzhen for your continuous concern and help. Furthermore, I want to thank all of our group members, especially Zelang MIAO, Xing YAN, Qunming WANG, Hao WU, Qianxiang XU, Yiliang WAN, Lipeng GAO, Anshu ZHANG, Cheng YANG, and so on. All joys and sorrows we have experienced together will always be in my memory. Thanks, everybody.

Last but not least, I am indebted to my parents for their continuous support and my wife, Ms Liying YU for her great patience and unconditional love of accompanying me through the past three years of doctoral study stage.

The dissertation is dedicated to my new born son, Huibo WANG. It’s regrettable I cannot accompany you when improving this work.

Table of Contents

Abstract ······ I

Acknowledgements ······ III

Table of Contents ······ V

List of Figures ······ VII

List of Tables ······ IX

List of Abbreviations ······ X

Chapter 1 Introduction ······ 1

1.1 Research Background ······ 1
1.2 Roles of GIS & Mathematics ······ 7
1.3 Research Scope ······ 9
1.4 Layout of this study ······ 10

Chapter 2 Summarization of limited Data ······ 13

2.1 H1N1 of Hong Kong in 2009 ······ 13
2.2 H7N9 of Eastern China in 2013 ······ 15
2.3 Ebola virus disease in West Africa, 2014 ······ 17

Chapter 3 Standard Deviational Ellipse and its Extension ······ 21

3.1 Introduction ······ 21
3.2 Standard Deviational Ellipse ······ 25
3.2.1 Exploring extreme standard deviations ······ 25
3.2.2 Optimal linear central tendency measure ······ 29
3.2.3 Spectral decomposition of covariance matrix ······ 29
3.3 Standard Deviational Hyper-Ellipsoid ······ 31
3.3.1 Construction of a Standard Deviational Hyper-Ellipsoid ······ 32
3.3.2 Confidence level analysis of SDHE ······ 34
3.4 Experiments & Applications ······ 43
3.4.1 Synthetic data experiments ······ 43
3.4.2 Spreading analysis of H1N1 infections ······ 46
3.5 Conclusions ······ 48

Chapter 4 Mathematical Approaches to Disease Spread ······ 51

4.1 SIR Compartmental Model ······ 51

4.2 Meta-Population Model 57
4.3 Reaction-Diffusion Equations 61
4.4 Runge-Kutta method based implementation 69
4.5 Applicability of Reaction-Diffusion Equations 70
4.5.1 Adaptive Diffusion Coefficient Matrix 70
4.5.2 Advantages & Disadvantages 72

Chapter 5 Investigating the Spatiotemporal Proximity Impact 73

5.1 Background 73
5.2 Spatiotemporal Proximity Integrated Approach 76
5.2.1 Retrospective inference of historical pathogens' existence 80
5.2.2 Spatial extrapolation of pathogens distribution 82
5.2.3 Infection risk prediction 84
5.3 Experiments & Discussions 85
5.3.1 Parameter Configuration 85
5.3.2 Historical pathogens distribution by retrospective inference 86
5.3.3 Forecast risk map 88
5.4 Conclusions 95

Chapter 6 Bayesian Inference of the Reproduction Number 99

6.1 Stochastic SEIR model 99
6.2 Reproduction number 100
6.3 Approximation of the early phase of exponential-growth 101
6.4 Sequential Bayesian Inference of R_t 105
6.5 Experiments and Discussion 109

Chapter 7 Conclusions 115

7.1 Concluding Summary 115
7.2 Main Contributions 118
7.2.1 Standard Deviational Ellipse 119
7.2.2 Reaction Diffusion Equations 119
7.2.3 Spatiotemporal Proximity Integrated Approach 120
7.2.4 Sequential Bayesian Inference of Reproduction Number 120
7.3 A Software Prototype 121
7.4 Discussion (Limitations and Future Work) 123

References 127

List of Figures

Figure 1.1 Illustration of theoretical relationships among Chapter 3 to Chapter 6 . 12
Figure 2.1 Locations of the confirmed H1N1 cases plotted on Hong Kong territory map with the Road Centerlines (RG1000) 14
Figure 2.2 Locations of the confirmed H1N1 cases plotted on Hong Kong territory map (289 TPUs) rendered of the population density distribution...... 15
Figure 2.3 Temporal evolution of notified cases of human infected with avian influenza A H7N9, 19 Feb to 21 May, 2013 16
Figure 2.4 Ebola virus epidemic in West Africa, up to Oct 12, 2014 17
Figure 2.5 Cumulative total numbers of Cases and Deaths over time in Guinea 18
Figure 2.6 Cumulative total numbers of Cases and Deaths over time in Sierra Leone 18
Figure 2.7 Cumulative total numbers of Cases and Deaths over time in Liberia..... 19
Figure 2.8 Cumulative total numbers of Cases and Deaths over time in Nigeria 19
Figure 3.1 An ellipse rotated with an angle θ in clockwise direction 26
Figure 3.2 One synthetic experiment of SDC and SDE constructed using 500 sampling points from a bivariate norm distribution 28
Figure 3.3 The confidence intervals correspond to 3-sigma rule of the normal distribution 37
Figure 3.4 The confidence regions corresponds to 1-3 multiples of SDEs............. 38
Figure 3.5 Visualization of 1-3 multiple SDEs for 2D scattered points.................... 45
Figure 3.6 Visualization of 1-3 multiple SDEs for 3D scattered points.................... 46
Figure 3.7 Exploratory analysis by 1-3 multiple SDEs for Hong Kong's H1N1...... 47
Figure 3.8 Exploratory analysis by 1-3 multiple SDHEs for Hong Kong's H1N1 .. 48
Figure 4.1 Phase portrait solutions of the basic SIR model 55
Figure 4.2 Example of SIR models for charactering the temporal evolutionary behavior 56
Figure 4.3 Snapshots of population proportion dynamic solved from the SIR-type PDE model 68
Figure 5.1 A schematic operational framework for the spatiotemporal proximity integrated approach... 79
Figure 5.2 Box plots of SRIOs with respect to different forecast time intervals..... 90
Figure 5.3 Box plot for the forecast time spans.. 90
Figure 5.4 3D bar graphs of the multiple "prediction-verification" chains generated by the spatiotemporal proximity integrated model for each H7N9 infected cases in chronological order. (a) maximum (best) forecasted illness onset risks, (b) median value of the forecasted illness onset risks and (c) the predictive time spans corresponding to the best forecasts... 93
Figure 5.5 One instance of "prediction-verification" chain link regarding the H7N9 illness onset risk.. 94
Figure 6.1 The R square index to the exponential growth phase of cumulative EVD case number for Guinea 103

Figure 6.2 The R square index to the exponential growth phase of cumulative EVD case number for Sierra Leone 103
Figure 6.3 The R square index to the exponential growth phase of cumulative EVD case number for Liberia 104
Figure 6.4 The R square index to the exponential growth phase of cumulative EVD case number for Nigeria 104
Figure 6.5 Time-varying Effective Reproduction Number for Guinea 110
Figure 6.6 Time-varying Effective Reproduction Number for Sierra Leone 110
Figure 6.7 Time-varying Effective Reproduction Number for Liberia 111
Figure 6.8 Time-varying Effective Reproduction Number for Nigeria 111
Figure 6.9 Choropleth map showing the effective reproduction number of Ebola affected countries 112
Figure 7.1 The framework of software prototype *ST* integrated with spatial-temporal epidemic models 122

List of Tables

Table 1.1 Roles of GIS & Mathematics 8

Table 3.1 Confidence levels of scaled SDHE vary with different magnification factors in spaces with the dimensionality not exceeding 10 40

Table 3.2 Magnification ratios of scaled SDHE corresponding to different specified confidence levels with space dimensionality not exceeding 10 43

Table 4.1 Other potential applications of Reaction Diffusion Equations 72

Table 5.1 A summary of primary variables and their descriptions 80

List of Abbreviations

BME	Bayesian Maximum Entropy
CHP	Center of Health Protection
EID	Emerging Infectious Diseases
EVD	Ebola Virus Disease
GIS	Geographic Information Science
HKSAR	Hong Kong Special Administrative Region of the People's Republic of China
KDE	Kernel Density Estimation
KTA	Kriging Trend Analysis
MCD	Minimum Covariance Determinant estimator
MCMC	Markov Chain Monte Carlo
PDE	Partial Differential Equation
PDF	Probabilistic Density Function
RDE	Reaction Diffusion Equations
SARS	Severe Acute Respiratory Syndrome
SBI	Sequential Bayesian Inference
SDC	Standard Deviation Curve
SDE	Standard Deviational Ellipse
SDHE	Standard Deviational Hyper-Ellipsoid
SEIR	Susceptible, Exposed, Infectious, Recovered / Removed
SIR	Susceptible, Infectious, Recovered / Removed
SRIO	Standardized Risk of Illness Onset indicator
STEM	Spatiotemporal Epidemiological Modeler
SVM	Support Vector Machine
TPU	Tertiary Planning Unit
WHO	World Health Organization
WKDE	Weighted Kernel Density Estimation method

Chapter 1 Introduction

The formidable task of developing models for endemic disease may be compared to building a house in a hurry. Practical workers insist on building a complete house, and are not too worried that it may need replacing later. Theoreticians insist on building reliable foundations and are not too worried if the house is never finished. Both viewpoints have their merits, and ideally we need to combine them together.

—— (Mollison and Kuulasmaa 1985)

1.1 Research Background

Over the past few years, epidemic waves of emerging infectious diseases (EID) appeared one after another. The more recent diseases included the severe acute respiratory syndrome (SARS) that happened in 2003 (Riley et al. 2003), the highly pathogenic H5N1 avian influenza with its peak of incidence in 2006 (Zhang et al. 2010), the swine influenza pandemic of H1N1 in 2009 (Fraser et al. 2009), the avian influenza A H7N9 that occurred in eastern China (Li et al. 2014) in the last two years, and the latest Ebola epidemic currently spreading in West Africa (Nishiura and Chowell 2014). The continuously emerging infectious diseases have always caused both very serious life risks and social-economic risks and been awakening the global concerns each time regarding their potential pandemic threats. Meanwhile, tens of thousands humans from all over the world unfortunately lost their life due to the spread of these infectious disease. There is an urgent need to develop theoretical methods and software to analyze and evaluate the risks due to the epidemic spread. Geographic information science (GIS) together spatial analysis can potentially be used for developing the solutions regarding this issue.

Studies on epidemic outbreaks began more than two thousand years ago (Bailey 1975). However, quantitative research on human diseases and deaths did not start until the 17th century (Graunt 1939). The earliest work on the geography of disease concentrated on mapping. A typical representative personage was John Snow, perceived to be one of the fathers of modern epidemiology, partially due to his pioneering work in 1854 about tracing the source of a cholera outbreak in Soho, London (Shiode 2012).

Much of the basic theory in Mathematical epidemiology was developed between 1900 and 1935. During that period, the well-known SIR model was formalized under a homogeneous mixing assumption, to characterize the temporal evolution of epidemic with the susceptible, infectious and immune compartments (Kermack and McKendrick 1927, Kermack and McKendrick 1932, Kermack and McKendrick 1933, McKendrick 1940). Another major contribution of Kermack and McKendrick was the Threshold Theorems, which claim that an introduction of a few infectious individuals into a population will not cause an epidemic outbreak, unless the proportion of susceptible population is above a certain critical value. Also supported by the Threshold Theorems is that not every susceptible person would necessarily be infected before an epidemic stops. This theory is in conjunction with the mass action principle, a cornerstone of modern theoretical epidemiology.

The deterministic version of SIR model was extended into a continuous-time and stochastic version by Kermack and McKendrick (1927). As infectious disease data often come in time-series format with unobservable latent disease states, Bayesian methods own exceptional advantages to make inferences on model parameters and predictions computed from these stochastic epidemic models (O'Neill 2002). Bayesian analysis implemented by the Markov Chain Monte Carlo method enables the estimation of a time-dependent transition rate and an effective reproductive number (Yin 2007). Informative prior distributions allow the inclusion of information from other analyses and expert opinion while posterior distributions can be used to inform additional data and structure inputs to simulation studies. The Bayesian approach has been applied by Spiegelhalter et al. (2004) in a spatial analysis of clinical trial data and Lawson et al. (2003) in disease mapping. Besides, Multilevel Bayesian models have been applied to investigate the spatial distribution of malaria in South Africa (Kleinschmidt et al. 2002), and the BSE(bovine spongiform encephalopathy) in Great Britain (Stevenson et al. 2005).

The concept of contact network epidemiology emerges as the idea of passing on a bacterial disease through contact between susceptible and infectious individuals becomes a consensus. This analytical framework intuitively captures the multifarious host interactions that underlie parasite transmission (Meyers 2007, Newman 2002). This approach firstly builds a realistic network model at an appropriate temporal and spatial scale for characterizing the contact patterns, through which the disease spreading can be predicted, based on intrinsic features of both the parasite and the network structure. Besides, percolation generating function methods originating from the discipline of statistical physics are often employed for mathematically analyzing the disease spread through networks (Grimmett 1999).

With rapidly increasing computing capacity, numerical simulations have been used to investigate more complex computational models, such as the "agent-based models" (ABMs) or micro-simulation models. An "artificial society" is created containing some population of "agents" in the computer program, typically the distinct individual of human beings. The agent-based model investigates the macro phenomena of the artificial society by simulating micro actions of individual, or "agents" (Epstein 2006). For example, Epstein (2004) built an agent-based model containing a population of 400 adults and 400 children living in two towns that share a school and a hospital. The 800 individuals formed 100 family households, each with two working adults and two school-age children. During the daytime, interactions between coworkers in work places and between students in the school were simulated, as were interactions within each household at night. After calibration, this program was used as a test-bed for comparing different smallpox outbreak control policies. The simulations showed that a combination of contact tracing by households and mass vaccination targeting at high-risk population provided the most effective control.

Due to the rapid rates of human mobility across the globe, the geographic spread of infectious diseases is of increasing concern. It is crucial to understand how, when, and why epidemics spread across the landscape so that effective planning, preparation, and control measures can be put in place before a disaster occurs. However, human mobility patterns are often complicated. Data sets that would aid in the understanding of these patterns are probably available but scattered in various proprietary locations. For example, Gonzalez et al. (2008) attempted to understand individual human mobility patterns by exploring the historical trajectory of 100,000 anonymized mobile phone users. Other potential sources of data on population travel patterns might include traffic counters along major roads or records from credit card use or hotel stays.

Objectively speaking, mathematical models play important roles in resisting the disease spreading. It can help to not only discover and elucidate important patterns from epidemiological data (Gonzalez et al. 2008), but also determine the potential risk factors for understanding and predicting disease spread (Gilbert et al. 2008). Besides, uncertainties hidden in epidemiological data can also be estimated by using mathematical models (Delmelle et al. 2014). However, it is of particular note that mathematical epidemiology differs from most sciences as it does not lend itself to experimental validation of models, which may probably be ethically unacceptable or technically unfeasible. It thus gives great importance to mathematical models as a versatile tool of comparing strategies for an anticipated epidemic or pandemic, and to cope with a disease outbreak in real time. Excellent models must either be able to be generalized to different infectious diseases and geographical conditions with ease or be readily adaptable to specific conditions. As conditions change in the progress of the disease, the model must be able to be updated promptly to take on new conditions. Of course, the outputs from the model must be rapidly available before a real time epidemic runs its course.

The Eclipse Foundation developed the Spatiotemporal Epidemiological Modeler (STEM), an auxiliary tool designed for public health officials to rapidly prototype and validate models for an emerging infectious disease (Edlund et al. 2010). STEM employs mathematical models of diseases (normally based on differential equations) to simulate the spatial-temporal behaviors of a disease (e.g., avian flu). These models can be used in understanding, and potentially preventing, the diseases spread process.

Humans with confirmed H7N9 virus infection may bring about rapid progressive pneumonia, acute respiratory distress syndrome (ARDS) and even death (Gao et al. 2013b). Up till July 2014, there were more than 450 laboratory confirmed influenza A (H7N9) cases including at least 150 deaths, reported by the Center for Health Protection of Hong Kong (http://www.chp.gov.hk/files/pdf/2014_avian_influenza_report_vol10_wk26.pdf). As the research continues, locally distributed density of live-poultry markets was perceived to be the most relevant environmental predictor for indicating the H7N9 infection risk (Li et al. 2014). Surveillance for influenza-like illness among susceptible individuals in close contact with laboratory-confirmed H7N9 cases indicated that there was no evidence of sustained onwards virus transmission between infected individuals, except at most a history of recent exposure to poultry. Most notably, an international scientific research team, with members from Belgium, United Kingdom, China, etc. adopted boosted regression tree models for correlating the risk of H7N9 market infection across Asia with locations of newly assembled thousands of live-poultry markets (Gilbert et al. 2014). However, their model only focuses on displaying of the static spatial distribution of H7N9 market infection risk, rather than indicating the future specific propagation tendency.

The currently ongoing epidemic of Ebola virus disease in West Africa has caused worldwide panic, mostly due to its significant mortality and rapid spreading rate. One research (Gomes et al. 2014) has even walk into the American congress for reference to the promotion of banning all the incoming flights from West Africa, soon after the first domestic cases of Ebola emerging. This research is known for estimating the importation probability of Ebola virus disease in countries by simulating the international outbreak spread using the worldwide daily airline passenger traffic data.

1.2 Roles of GIS & Mathematics

Sudden epidemic of an infectious disease urgently needs early prevention and control measures. To prevent the large-scale epidemic, using geographic information systems (GIS) to analysis the spatial-temporal clusters and diffusion patterns of diseases can assist in timely control measures for emergent infectious diseases or epidemics initiated from imported cases. The disease information, such as public health resources, spreading trends can be mapped together in relation to their surrounding environmental elements, also making GIS as a common platform for monitoring and management of epidemics.

Table 1.1 Roles of GIS & Mathematics

GIS	(1) Data management, processing, visualization and (2) Spatial analysis, (3) Monitoring of epidemic situation, (4) Interactive operation based on Web GIS.
Mathematics	(1) Discover transmission patterns from epidemiological data, (2) Determine the potential risk factors, (3) Understand and predict disease spread, (4) Estimate the hidden uncertainties from epidemiological data.

Objectively speaking, mathematical models play important roles in resisting the disease spread. It can help to not only discover important patterns from epidemiological data (Gonzalez et al. 2008), but also determine the potential risk factors for understanding and predicting disease spread (Gilbert et al. 2008). Besides, the uncertainties hidden in the epidemiological data can also be estimated by using mathematical models (Delmelle et al. 2014). Integrated with powerful mathematical models, geospatial technologies offer a new perspective to study the spatial interactions between disease emergence and protective factors. The Table 1.1 above lists out the roles of both GIS and mathematics in epidemic prevention.

1.3 Research Scope

This research mainly aims to explore the spatial and temporal characteristic patterns of epidemics by integrating mathematical models with the geographic information science technology, anticipating these innovative approaches can be conductive to understand and capture the behaviors of these frequent emerging infectious diseases.

To reach the above aim, specific research objectives are identified as follows,

(a) To dynamically manage and visualize epidemic data, identify the concentrated hot spots and interpret the evolutionary trends of the disease cases spatiotemporally using GIS solutions;

(b) Attempt to utilize the deterministic Partial Differential Equation (PDE) models to characterize the spatiotemporal dynamics of infectious diseases, thereby making it tractable for monitoring the whole epidemic evolution process;

(c) Propose innovative approach that can make (short-term) predictions upon the infection risk of emerging infectious disease, for the potential development of an early warning system;

(d) Reflect the real-time hazard level regarding the disease transmission intensity by estimating some time-varying indicators (such as the effective reproduction number), which can be hopefully merged into the Geographical maps so as to identify areas of higher risk for an outbreak;

(e) If possible, to develop a software prototype integrating with the functionalities of mapping hot spots, as well as flows and evolutionary trends of the disease, forecasting early warning and graphical risk maps based on suitable epidemic models and GIS technology.

1.4 Layout of this study

This thesis is divided into seven major chapters (plus this introduction), which deal with different characteristic patterns of epidemics and propose four typical models that can be conductive to understand and capture their behavior.

The structure of this thesis has been arranged as follows. The first chapter devotes to a comprehensive introduction for the background, methods and developments of the infectious disease related research. Chapter 2 is intended to give the description of the relevant experimental data in GIS and epidemiology formats for H1N1 of Hong Kong in 2009, H7N9 of Mainland China in 2013 and cumulative infections of Ebola virus in West Africa, 2014. After that, in-depth discussions of the elementary analysis methods - Standard Deviational Ellipse (SDE), are explored in Chapter 3. Mathematical approaches to disease spread are presented in Chapter 4 including the basic compartmental SIR model, traditional approach to modeling temporal dynamics of infectious disease, and the meta-population model for characterizing the spatiotemporal spreading dynamics of disease infection. Besides, the typical reaction-diffusion equation models are thoroughly explored, following with the detailed computer implementation procedures via the Runge-Kutta method. Chapter 5 innovatively proposes an approach for investigating the spatiotemporal proximity impact upon the prediction of illness onset risk of emerging infectious disease, with experiments upon the avian influenza A H7N9, February to May 2013 in eastern China. In addition, a sequential Bayesian inference combined with stochastic SEIR model has been employed in Chapter 6 for estimating the time-varying effective reproduction numbers, together with their 95% confidence intervals for the Ebola virus epidemic in West Africa. In the end, the final Chapter 7 provides the concluding remarks of this study and framework of one software prototype, to be developed for characterizing the spatiotemporal spreading dynamics and predicting future evolutionary trends for emerging infectious diseases. Discussions regarding the limitations and potentialities of the models involved here are also carried out for guiding the future research work to better prevent the infectious disease spread.

We acknowledge that there is no explicit relationship between each of these four models. However, the theory behind each chapter goes from the shallower to the deeper. All these models are contributing to characterize the transmission patterns of specific disease. Figure below presents concise illustrations of Chapter 3 to Chapter 6 to help with a rapid understanding of their theoretical relationships.

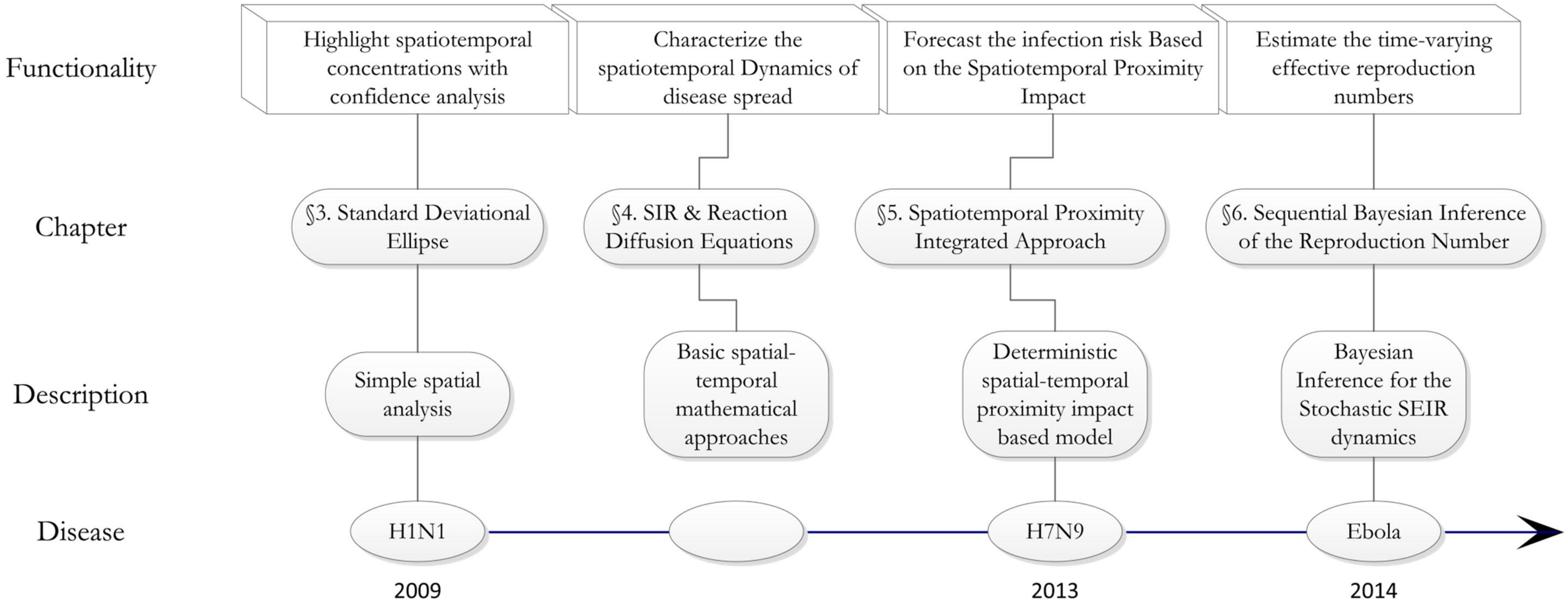

Functionality
Chapter
Description
Disease
Highlight spatiotemporal concentrations with confidence analysis
Characterize the spatiotemporal Dynamics of disease spread
Forecast the infection risk Based on the Spatiotemporal Proximity Impact
Estimate the time-varying effective reproduction numbers
§3. Standard Deviational Ellipse
§4. SIR & Reaction Diffusion Equations
§5. Spatiotemporal Proximity Integrated Approach
§6. Sequential Bayesian Inference of the Reproduction Number
Simple spatial analysis
Basic spatial-temporal mathematical approaches
Deterministic spatial-temporal proximity impact based model
Bayesian Inference for the Stochastic SEIR dynamics
H1N1
H7N9
Ebola
2009
2013
2014

Figure 1.1 Illustration of theoretical relationships among Chapter 3 to Chapter 6

Chapter 2 Summarization of limited Data

This chapter is devoted to the description of all relevant data acquired for this research. In general, three types of data are necessary for a spatial epidemiological study. They include the core case series data usually given as some spatial and temporal points; additional geographical layers, such as the population distribution layer describing the (dynamic) population density; and the mobility network layer of the host (like transportation routes) characterizing the potential spreading patterns of the disease. However, it is noteworthy that the contributing models of this research are mainly put forward under the circumstance of very with limited clinical database in hand.

2.1 H1N1 of Hong Kong in 2009

The core data with epidemiological date and residential address of human swine influenza cases (see Figures 2.1-2.2) from 1st May to 26th June, were gathered by NGAN (2010) and Zhang (2011) from lists of building(s) with confirmed case(s) of Human Swine Influenza on a daily basis released by Center of Health Protection (CHP), list of suspended school announced by Education Bureau, News Press in CHP and various local newspapers, websites. The way to acquire location of the publicized infected buildings is firstly extracting the Lat/Long in UTM upon Google Map, and further converting them into HK1980Grid.

Other Spatial and non-spatial additional data includes the Road Centerline data

(RG1000, see Figure 2.1) from the Geo-Reference Database and the shape files of 18 Administrative District Council (1996) in Hong Kong and 289 tertiary planning units (TPUs) boundaries. Non-spatial data includes the demographic data of Hong Kong population census distribution in 2011 (Figure 2.2).

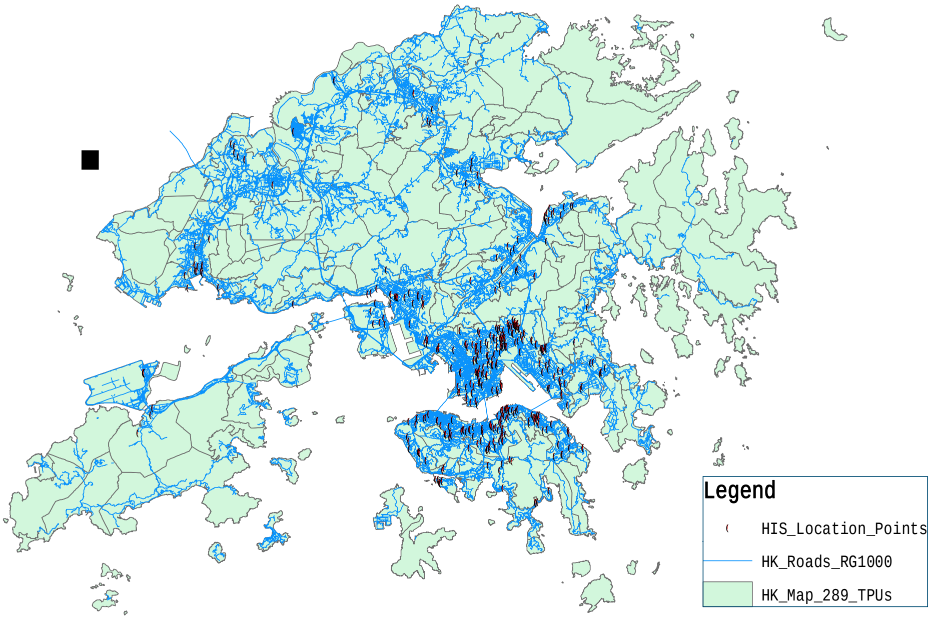


Figure 2.1 Locations of the confirmed H1N1 cases plotted on Hong Kong territory map with the Road Centerlines (RG1000)

Sincere thanks are owed to Mr. Raymond Tse, Census and Statistics Department for providing the guidance of building the own custom Census table from the 2011 Population Census Database, and Associate Professor, Lilian PUN CHENG, (LSGI, PolyU) for sharing me the collative Geo-Reference Database including the TPUs boundaries map data and Road Centerline data (RG1000). Derived from the Hong Kong Grid 1980 Coordinate System, all these geographic map layers are utilized for locating the position of H1N1 occurrences.

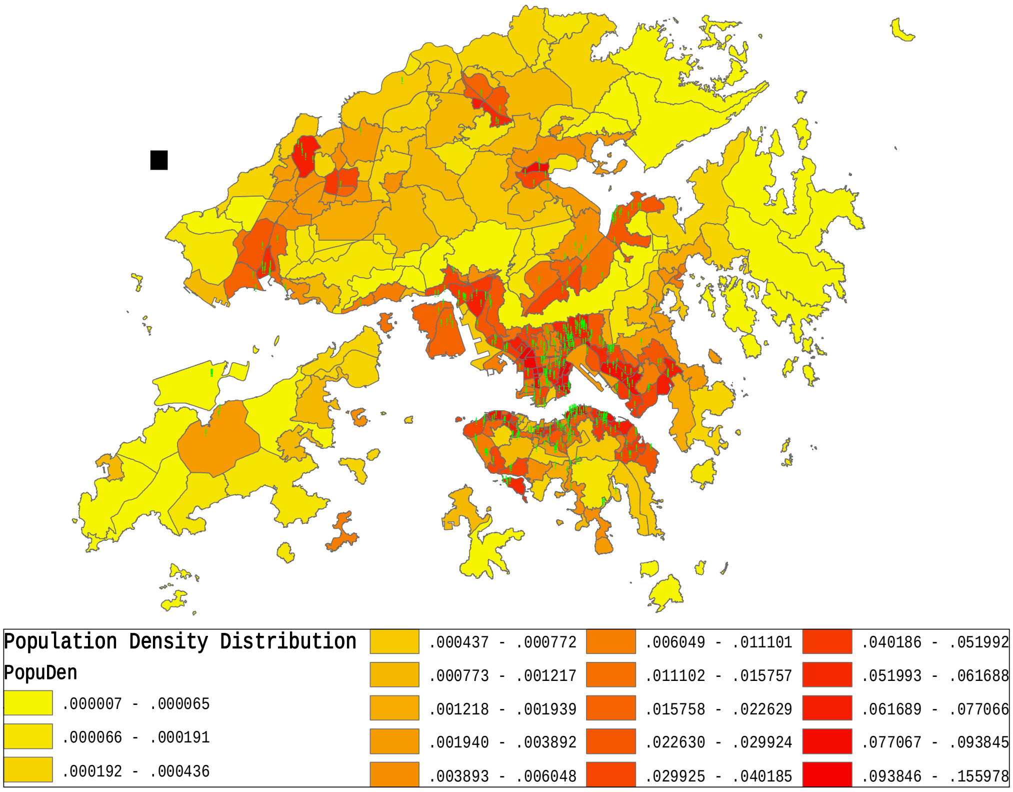


Figure 2.2 Locations of the confirmed H1N1 cases plotted on Hong Kong territory map (289 TPUs) rendered of the population density distribution

2.2 H7N9 of Eastern China in 2013

Database of H7N9 occurrence cases are gathered from two sources. One is a real-time H7N9 reporting system (http://goo.gl/maps/ZsVW8) developed by Yujun ZHAO & Dr. Jiankui HE from the South University of Science and Technology in China since the first outbreak, including a total of 87 cases as of 18 April, 2013. The other source is from an international public forum, FluTrackers.com (http://www.flutrackers.com/forum/showthread.php?t=202713), which investigates various emerging infectious diseases. It constantly publishes the confirmed or suspected cases of human avian influenza a (H7N9) with detailed hyperlinks for each case in chronological sequence. An additional 48 cases before June, 2013 are further filtered out from this forum. We have carefully checked each notification record by verifying them according to announcements from official websites of the national health and family planning commissions at the provincial level and local news reports for tracking the accurate infection site for each case, and then geocoded them into the WGS84 coordinate system. The collative data set finally combines with 135 records in total, covering the period between 19th February and 21th May, 2013. All cases are displayed with bar graph in Figure 2 from the aspect of temporal evolution. More detailed document with illness onset date and spatial site information (longitude and latitude) is also provided being the supplementary material (H7N9_Cases.xlsx).

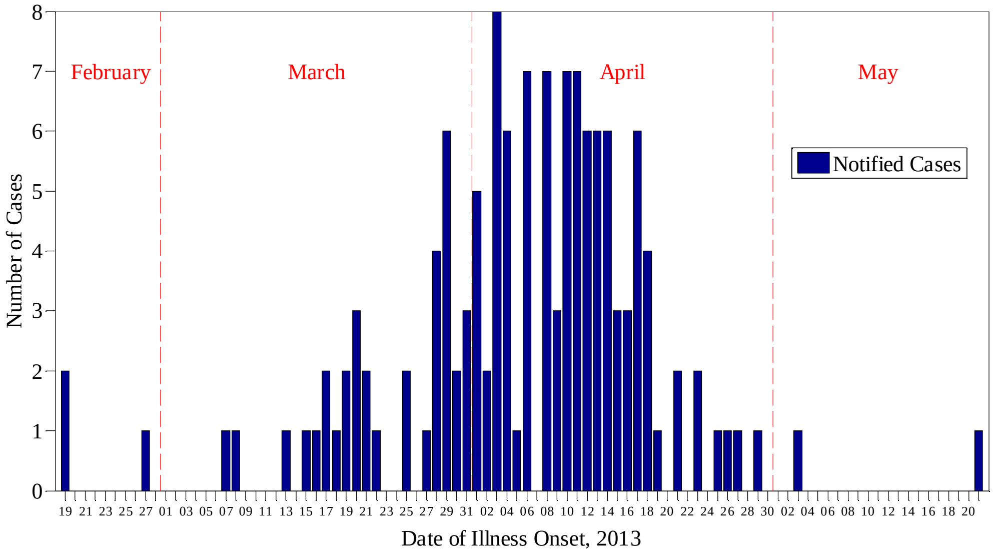


Figure 2.3 Temporal evolution of notified cases of human infected with avian influenza A H7N9, 19 Feb to 21 May, 2013

China's geographical administrative divisions are extracted from the GADM database (http://www.gadm.org/country) in shapefile format. Considering the entire epidemic area mainly concentrated in eastern China, all geographic maps, together with the 135 georeferenced occurrences, are both projected into the same Universal Transverse Mercator coordinate system of zone 50N for preferable display considerations of the later forecasted illness onset risk layer.

2.3 Ebola virus disease in West Africa, 2014

Data of the Ebola virus disease in West Africa was gathered from the Disease Outbreaks News of the WHO, which tracked new cases and deaths by date with generally biweekly reports. We carefully collating in Guinea, Sierra Leone, Liberia and Nigeria as officially affected by the EVD epidemics.

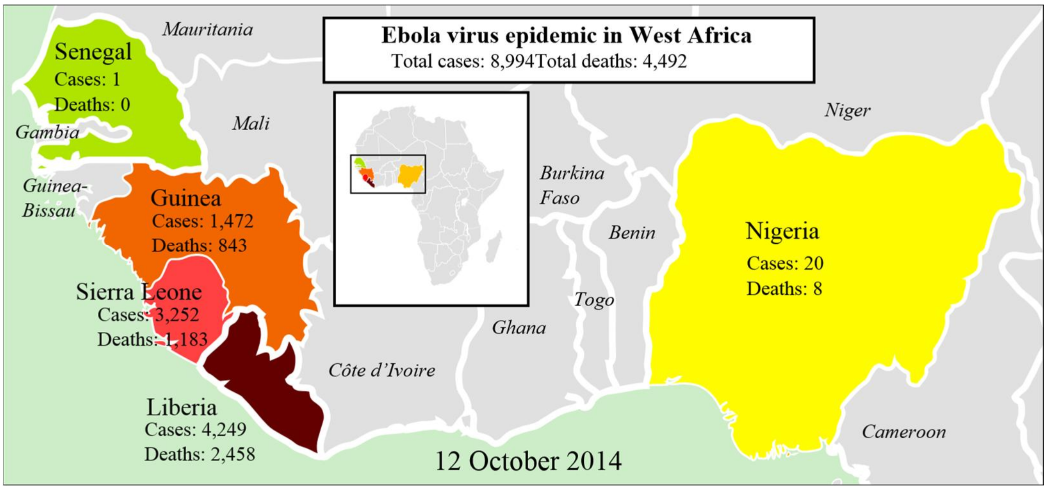


Figure 2.4 Ebola virus epidemic in West Africa, up to Oct 12, 2014[1]

As of 12 October, 2014, a total of 8997 confirmed, probable, and suspected cases

[1] https://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa

of Ebola virus disease (EVD) come together with 4493 deaths, reported from seven affected countries (Guinea, Liberia, Nigeria, Senegal, Sierra Leone, Spain, and the United States of America). The following Figures 2.5-2.8 illustrate the cumulative totals of infected cases and deaths by EVD over time in Guinea, Sierra Leone, Liberia and Nigeria.

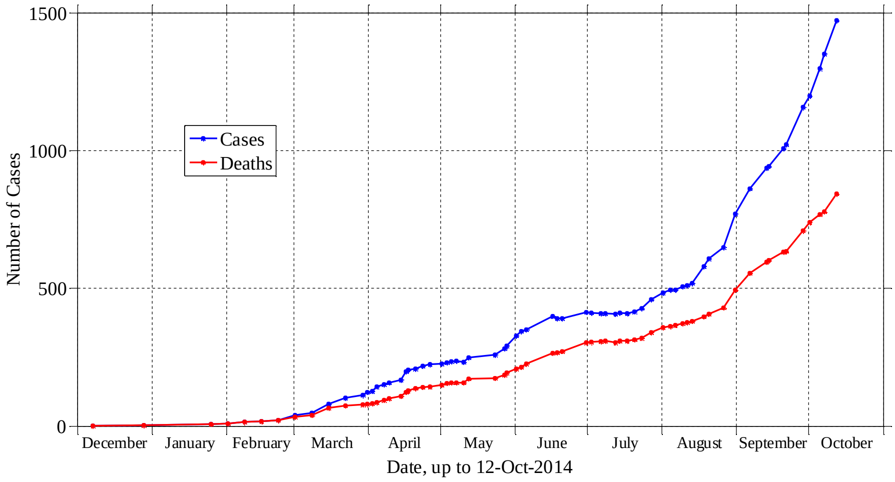


Figure 2.5 Cumulative total numbers of Cases and Deaths over time in Guinea

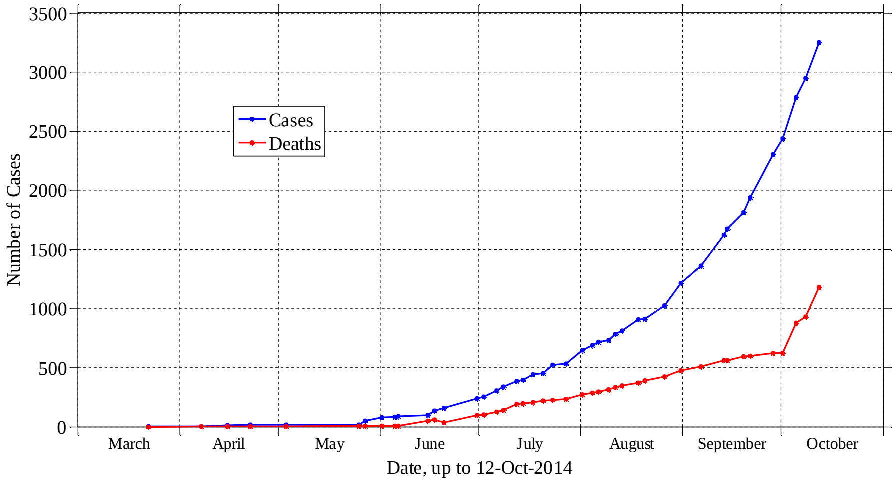


Figure 2.6 Cumulative total numbers of Cases and Deaths over time in Sierra Leone

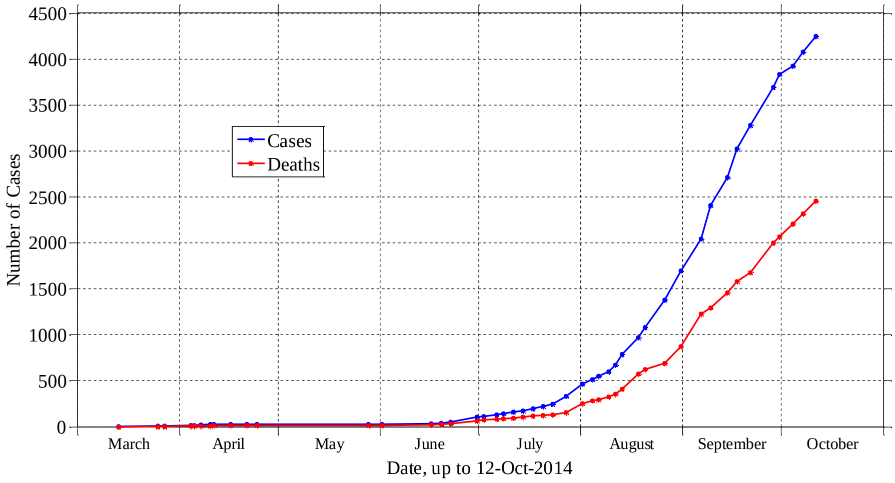


Figure 2.7 Cumulative total numbers of Cases and Deaths over time in Liberia

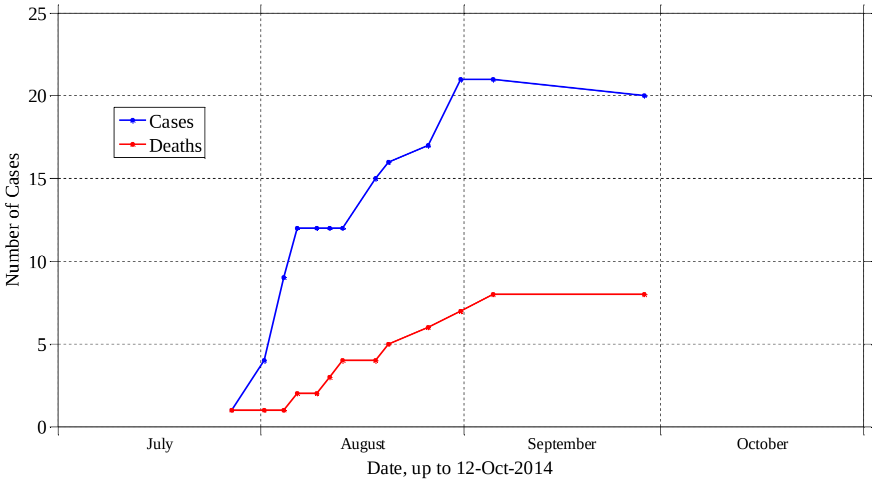


Figure 2.8 Cumulative total numbers of Cases and Deaths over time in Nigeria

From these figures, obviously it can be observed that the EVD started its journey from Guinea in December, 2013 and spread into Sierra Leone and Liberia during March, 2014. The cumulative infected case number in Guinea was shortly exceeded by the total cases in the other successor countries ever since July and August, 2014. At that time, the EVD case began reported Nigeria, where fortunately EVD seems to be under control. However, it was till unoptimistic of the epidemic situation in the preceding three countries, especially in Liberia, where the epidemic situation was even getting worse.

Chapter 3 Standard Deviational Ellipse and its Extension[1]

Standard deviational ellipse (SDE) has long been served as a versatile GIS tool for delineating the geographic distribution of concerned features. This chapter firstly summarizes two existing models of calculating SDE, and then proposes a novel approach to constructing the same SDE based on spectral decomposition of the sample covariance, by which the SDE concept is naturally generalized into higher dimensional Euclidean space, named standard deviational hyper-ellipsoid (SDHE). Then, rigorous recursion formulas are derived for calculating the confidence levels of scaled SDHE with arbitrary magnification ratios in any dimensional space. Besides, an inexact-newton method based iterative algorithm is also proposed for solving the corresponding magnification ratio of a scaled SDHE when the confidence probability and space dimensionality are pre-specified. These results provide an efficient manner to supersede the traditional table lookup of tabulated chi-square distribution. Finally, synthetic data were employed to generate the 1-3 multiple SDEs and SDHEs. And exploratory analysis by means of SDEs and SDHEs are also conducted for measuring the spread concentrations of H1N1 of Hong Kong in 2009.

3.1 Introduction

Standard deviation arises as one of classical statistical measures for depicting the

[1] This chapter are mainly based on such publication: WANG, B., SHI, W. and MIAO, Z. 2015. Confidence Analysis of Standard Deviational Ellipse and Its Extension into Higher Dimensional Euclidean Space. PLoS ONE, 10(3), e0118537.

dispersion of univariate features around its center. Its evolution in two dimensional space arrives at the standard deviational ellipse (SDE), which was firstly proposed by Lefever in 1926. Ever since then, SDE has long served as a versatile GIS tool for delineating bivariate distributed features. It is typically employed for sketching the distributional trend of geographical features by summarizing both of their dispersion and orientation. Although SDE's arrival had once aroused great attention, a certain amount of criticisms followed as well, mainly due to the fact that Lefever's defined curve is not an ellipse (Furfey 1927), but the standard deviation curve (SDC) as nominated by Gong (2002).

The utilization potentials of SDE have been found in many research fields and commercial industries. For instance, Smith and Cheeseman (1986) employed it for estimating spatial uncertainty between coordinate frames representing the relative locations of a mobile robot. Besides, SDE has also been adopted to quantitatively analyze the orientation anisotropy in contaminant barrier particles (Wang et al. 2008), and explore the geographical distribution of household activities or travel behaviors thereby promoting policy formulation in response to urban travel reduction strategies (Buliung and Kanaroglou 2006). Meanwhile, geographical profiling of the distributional trend for a series of crimes (Kent and Leitner 2007, Chainey et al. 2008) by SDE might detect a relationship to particular physical features such as some restaurants or apartments and even the lairs of the criminals. Mapping groundwater well samples for some kind of contaminant could identify how and to what extent the toxin is spreading, which consequently, may be conducive to deploy the responding mitigation strategies (Cloutier et al. 2008). Moreover, comparing the coverage area, shape, and overlap of ellipses for various racial or ethnic groups may provide insights regarding racial or ethnic segregation (Wong 1998). Furthermore, graphing ellipses for a disease outbreak such as malaria surveillance (Eryando et al. 2012) over time can potentially make the real-time prediction of its spatial path, since the central tendency and dispersion are two principal aspects attracting concerns from epidemiologists.

As a GIS tool for delineating spatial point data, SDE is mainly determined by three measures: average location, dispersion (or concentration) and orientation. In addition to the traditional mean center (gravity of the distribution) suggested by Lefever (1926), weighted mean or median could also be the alternative options, together with the weighted covariance of observations which evolve into some variants of the SDE (Yuill 1971). It is worth noting that SDE also lays the foundation for many other advanced models, such as the minimum covariance determinant estimator (MCD) (Rousseeuw and Driessen 1999, Hubert and Debruyne 2009) for outlier detection and elliptic spatial scan statistic (Kulldorff et al. 2006) employed in spatiotemporal disease surveillance. From the perspective of practical implementation, Alexandersson (2004) once wrote an *ellip* command for graphing the confidence ellipses in Stata 8, with the latest version being Stata 13.

Although SDE has extensive applications in various fields since 1926, it still has not been correctly clarified sometimes. For instance, from the latest resources in

ArcGIS Help 10.1 describing how standard deviational ellipse works, it is stated that one, two and three standard deviation(s) can encompass approximately 68%, 95% and 99% of all input feature centroids respectively, supposing the features concerned follow a spatially normal distribution. However, this content corresponds to the well-known 3-sigma rule with respect to univariate normal distribution, rather than the bivariate case. Worse still, there is even an attached illustration therein depicting several bivariate geographical features located within a planar map. Obviously, such confusing interpretation may mislead the GIS users to believe the univariate 3-sigma rule remains valid in two-dimensional Euclidean space, or even higher dimensions.

For fully clarifying the implications of SDE, section 3.2 below devotes to firstly summarizing two existing models of deriving the SDE's calculation formulas, and secondly proposing a novel approach for constructing the same SDE based on spectral decomposition of the sample covariance, by which SDE concept is further extended into higher dimensional Euclidean space, named standard deviational hyper-ellipsoid (SDHE). Most of all, rigorous recursive formulas are then derived for calculating the confidence levels of scaled SDHE with arbitrary magnification ratios in any dimensional space. Besides, an inexact-newton method based iterative algorithm is also proposed for solving the corresponding magnification ratio of a scaled SDHE when the confidence probability and space dimensionality are pre-specified. Finally, synthetic data is employed to generate the 1-3 multiple SDEs and SDHEs in two and three dimensional spaces, respectively. Meanwhile, exploratory analysis by means of SDEs and SDHEs are also conducted for measuring the spread concentrations of Hong Kong's H1N1 in 2009.

3.2 Standard Deviational Ellipse

First two subsections below devotes to a brief summarization of two classical approaches to generating the standard deviational ellipses in 2D. After that, a novel approach based on spectral decomposition of the covariance matrix is introduced which achieves the same calculation formula of SDE. This spectral decomposition based approach will be adopted for constructing the generalized standard deviational (hyper-)ellipsoids into higher dimensional Euclidean space in the next section 3.3.

3.2.1 Exploring extreme standard deviations

A standard deviational ellipse delineates the geographical distributing trend by summarizing both dispersion and orientation of the observed samples. There are already several approaches to obtaining the computational formula of SDE. The upcoming discussed method presented by Yuill (1971) was actually a melioration of Lefever's original model (Lefever 1926) despite suffering from certain criticisms (Furfey 1927).

Suppose a series of independent and identically distributed samples (x_i, y_i), $i = 1, \cdots, n$ are drawn from a Gaussian population. A standard deviational ellipse can be determined according to the following steps. Firstly, make the sample mean as the origin of new axes, thereby simultaneously centering all the observed samples,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,; \quad \begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \overline{x} \\ \overline{y} \end{pmatrix}. \tag{3.1}$$

Next, introduce a rotation matrix $G = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ with an angle θ in the clockwise direction as illustrated in Figure 3.1. All observed sample points are then transformed into a new planar coordinate system,

$$\begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = G\begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} \tilde{y}_i \sin\theta + \tilde{x}_i \cos\theta \\ \tilde{y}_i \cos\theta - \tilde{x}_i \sin\theta \end{pmatrix}. \tag{3.2}$$

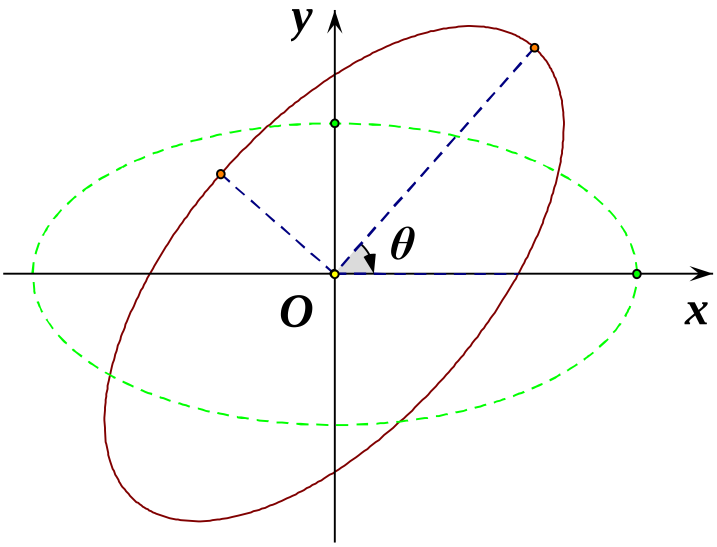


Figure 3.1 An ellipse rotated with an angle θ in clockwise direction

The maximum likelihood estimator (Li and Racine 2007) of the rotated samples' variance yields,

$$\begin{cases} \sigma_{x'}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i'\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{y}_i \sin\theta + \tilde{x}_i \cos\theta\right)^2 \\ \sigma_{y'}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i'\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{y}_i \cos\theta - \tilde{x}_i \sin\theta\right)^2 \end{cases} . \qquad (3.3)$$

Consequently, corresponding angles for producing the maximum and minimum standard deviations can be obtained by equating any derivative of the above variance estimators w.r.t. θ to be zero (Yuill 1971, Wang et al. 2008), that is

$$\frac{d\sigma_{x'}^2}{d\theta} = \frac{2}{n}\sum_{i=1}^{n}\left(\tilde{y}_i^2 \sin\theta\cos\theta + \tilde{x}_i\tilde{y}_i\left(\cos^2\theta - \sin^2\theta\right) - \tilde{x}_i^2 \sin\theta\cos\theta\right) = 0 .$$

According to Vieta's formulas, the general solution to the above quadratic equation is then given by

$$\tan\theta = \frac{\left(\sum_{i=1}^{n}\tilde{x}_i^2 - \sum_{i=1}^{n}\tilde{y}_i^2\right) \pm \sqrt{\left(\sum_{i=1}^{n}\tilde{x}_i^2 - \sum_{i=1}^{n}\tilde{y}_i^2\right)^2 + 4\left(\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i\right)^2}}{2\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i} . \qquad (3.4)$$

Each of these two angles corresponds to the maximum and minimum deviation in the new coordinate system, respectively. By merging equation (3.4) into equation (3.3), the major axis and minor axis of SDE can be determined for measuring the dispersion distribution of original observations.

It should be noticed that rotating $\sigma_{x'}^2$ equation (3.3) around the sample mean center defines an implicit locus curve (Lefever 1926). However, such a closed curve is not an ellipse (Furfey 1927), but actually the standard deviation curve (SDC) nominated by Gong (2002) with its expression as follows,

$$\left(\tilde{x}^2+\tilde{y}^2\right)^2=\sigma_x^2\tilde{x}^2+2\rho\sigma_x\sigma_y\tilde{x}\tilde{y}+\sigma_y^2\tilde{y}^2 . \qquad (3.5)$$

Here ρ is the correlation coefficient between x and y coordinates. For seeking a striking contrast between SDC and SDE, a numerical experiment is conducted, employing 500 synthetic points extracted from a bivariate normal variable with mean $\mu=(0,0)^{\mathrm{T}}$ and covariance matrix $C=\begin{pmatrix}0.9 & 0.4\\ 0.4 & 0.5\end{pmatrix}$. Based on these sampling points, contradistinctive profiles of 1-3 multiple SDC and SDE are illustrated in Figure 3.2. Conspicuously there are 4 tangency points for each corresponding pair, and SDC appears occupying an overall larger area then SDE.

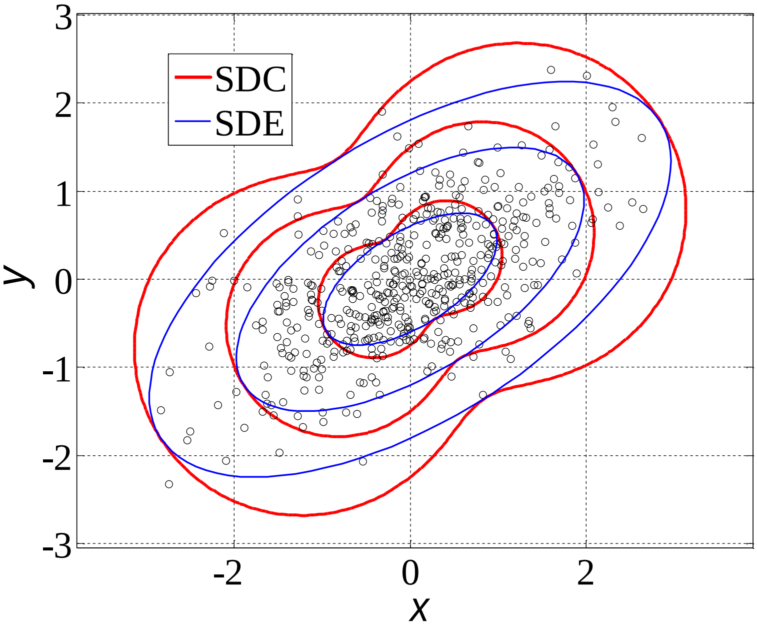


Figure 3.2 One synthetic experiment of SDC and SDE constructed using 500 sampling points from a bivariate norm distribution

3.2.2 Optimal linear central tendency measure

Another method described by Cromley (1992) aims to explore such an optimal linear central tendency measure, $ax + by + c = 0$, which passes through the distributed samples. This is equivalent to an optimization problem with the objective of minimizing the summation of total perpendicular distances from any observation point to this line subject to the constraint of $a^2 + b^2 = 1$, which guarantees the scale invariance, namely,

$$\begin{array}{ll} \min & \sum_{i=1}^{n} \left(ax_i + by_i + c \right)^2 \\ s.t. & a^2 + b^2 = 1 \end{array} . \tag{3.6}$$

The above constrained optimization problem can be solved by Lagrangian multiplier method, yielding the optimal linear central tendency which precisely coincides with the direction of the principal axis of SDE. Therefore, solution to the above optimization arrives at exactly the same calculation formulas of SDE as the aforementioned first approach.

3.2.3 Spectral decomposition of covariance matrix

Using symbols introduced in equation (3.1), this subsection devotes to presenting another approach for constructing SDE by means of spectral decomposition of the sample covariance matrix, which is formulated as follows,

$$C=\begin{pmatrix} \operatorname{var}(x) & \operatorname{cov}(x,y) \\ \operatorname{cov}(y,x) & \operatorname{var}(y) \end{pmatrix}=\frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n}\tilde{x}_i^2 & \sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i \\ \sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i & \sum_{i=1}^{n}\tilde{y}_i^2 \end{pmatrix}, \tag{3.7}$$

where $\operatorname{var}(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i^2$, $cov(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i$

and $\operatorname{var}(y)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i^2$.

It must be said there are two common textbook definitions of variance and covariance, as well as the standard deviation. One is the unbiased estimator while the other one is the maximum likelihood estimator proved by Li and Racine (2007). Their calculation formulas differ only in $n-1$ versus n in the divisor. To keep consistent with the previous equations involved, the latter estimator is employed hereafter.

After spectral decomposition of the sample covariance (3.7), SDE can be constructed by assigning square roots of eigenvalues as the lengths of its semi-major and semi-minor axes (Härdle and Simar 2012), to which being parallel by the corresponding eigenvectors. Solving of the characteristic polynomial equation of covariance matrix C, namely,

$$f(\lambda)=\det(\lambda I-C)=\det\begin{pmatrix} \lambda-\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i^2 & -\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i \\ -\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i & \lambda-\frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i^2 \end{pmatrix}=0, \tag{3.8}$$

yields the lengths of the SDE's semi-major and semi-minor axes, which are

$$\sigma_{1,2}=\left(\frac{\left(\sum_{i=1}^{n}\tilde{x}_i^2+\sum_{i=1}^{n}\tilde{y}_i^2\right)\pm\sqrt{\left(\sum_{i=1}^{n}\tilde{x}_i^2-\sum_{i=1}^{n}\tilde{y}_i^2\right)^2+4\left(\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i\right)^2}}{2n}\right)^{1\!/\!2}; \qquad (3.9)$$

Meanwhile, one group of base vectors from the characteristic vector space satisfying equation (3.8) can be obtained by

$$v_{1,2}=\left(\left(\sum_{i=1}^{n}\tilde{x}_i^2-\sum_{i=1}^{n}\tilde{y}_i^2\right)\pm\sqrt{\left(\sum_{i=1}^{n}\tilde{x}_i^2-\sum_{i=1}^{n}\tilde{y}_i^2\right)^2+4\left(\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i\right)^2},\ 2\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i\right)^{\mathrm{T}}. \qquad (3.10)$$

Thus, it takes no effort to verify that orientation angles intersected by the principle axes of SDE and the planar coordinate axes are exactly the same, namely, the optimal angle appeared in equation (3.4).

In conclusion, the above three approaches actually all calculate the same SDE according to formulas (3.1), (3.4) and (3.9), respectively, which lays the theoretical basis for SDE to be one functional component in the Spatial Statistics toolbox of ArcGIS 10.1.

3.3 Standard Deviational Hyper-Ellipsoid

In section 3.2, three approaches for constructing SDE have been summarized and compared upon the distributed samples in two-dimensional space. This section will generalize the SDE concept into higher dimensional Euclidean space, yielding the standard deviational hyper-ellipsoid (SDHE), be means of the spectral decomposition of covariance matrix. Meanwhile, rigorous mathematical derivations attempt to figure out the relationship between the confidence levels characterizing the probabilities of random scattered points falling inside a scaled SDHE and the corresponding magnification ratio under the assumption that samples follow the Gaussian distribution.

3.3.1 Construction of a Standard Deviational Hyper-Ellipsoid

Suppose $S \in R^n$ be an n-dimensional Gaussian random vector, that is $S \sim N\left(\mu, C\right)$ with its probability density function

$$f\left(s\right) = \frac{1}{\left(2\pi\right)^{n/2} \left|C\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(s-\mu\right)^{\mathrm{T}} C^{-1}\left(s-\mu\right)\right\}. \tag{3.11}$$

And $S_1, S_2, \cdots, S_m$ represent m independent and identically distributed samples extracted from population S. In general, the maximum likelihood estimators (Li and Racine 2007) for parameters μ and C employed in equation (3.11) can be given by

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} S_i \, , \quad \hat{C} = \frac{1}{m}\sum_{i=1}^{m}\left(S_i - \hat{\mu}\right)\left(S_i - \hat{\mu}\right)^{\mathrm{T}} . \tag{3.12}$$

Since covariance matrix C is real symmetric (positive semi-definite), there exists an orthogonal matrix Q (formed by eigenvectors of C) complying with the spectral decomposition,

$$C = QDQ^{\mathrm{T}} . \tag{3.13}$$

Without loss of generality, suppose all the main diagonal elements of $D = \operatorname{diag}(\sigma_i)$, $i = 1, 2, \cdots, n$ have been sorted in descending order, $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$. Due to the symmetry of covariance matrix C, its spectral decomposition is actually equivalent to its singular value decomposition which outputs a series of automatically sorted eigenvalues (singular values). As thus, mapping a unit sphere by square root of covariance matrix, $C^{1/2}$, yields a standard hyper-ellipsoid, with eigenvalues to be its principle semi-axes oriented by their corresponding eigenvectors (Trefethen and Bau III 1997).

Proceeding in this way, now comes to such an interesting question: how could this SDHE defined by equation (3.13) be represented graphically? This can be figured out by means of the Mahalanobis transformation (Härdle and Simar 2012) which is defined as

$$T = C^{-1/2}(S - \mu) = QD^{-1/2}Q^{\mathrm{T}}(S - \mu). \tag{3.14}$$

It can be verified that $T \sim N(0, I_n)$. In other words, Mahalanobis transformation eliminates correlation between the variables and standardizes each variable with variance. Apparently, random vector T 's SDHE happens to be a unit sphere ($\|T\|_2 = 1$) in view of its isotropic distribution along any direction. Therefore, SDHE of original random vector S can be constructed from the transformation of a unit sphere by firstly stretching with a ratio of $\sqrt{\sigma_i}$ along each axis successively, then rotating the ellipsoid by orthogonal matrix Q and a final translation of distribution center μ according to the following inverse

Mahalanobis transformation,

$$S = QD^{1/2}Q^{\mathrm{T}}T + \mu . \tag{3.15}$$

3.3.2 Confidence level analysis of SDHE

This section settles the relationship between confidence levels characterizing the probabilities of random scattered points falling inside the scaled ellipsoids and the corresponding magnification ratio of such an SDHE by means of the rigorous mathematical formulas derivations.

The following scalar quantity

$$r^2 = (S - \mu)^{\mathrm{T}} C^{-1} (S - \mu), \tag{3.16}$$

is known as the Mahalanobis distance of the vector S away from its mean μ. By merging equations (3.13) and (3.14) into equation (3.16), it can be easily perceived that the above defined quadratic function is exactly the magnified SDHE with a magnification ratio of r and follows the chi-square distribution with n degrees of freedom,

$$\Pr\left\{r^2 \leq \chi^2_{n,p}\right\} = p . \tag{3.17}$$

Table lookup of a tabulated chi-square distribution is always adopted as the traditional approach to acquire the exact confidence levels. Therefore, exploring to what extent the scattered samples obeying a Gaussian distribution is equivalent to examining whether they are falling inside such a scaled ellipsoid defined in terms of equation (3.16). Actually, calculation of the cumulative distribution function of chi-square distribution for a prescribed value x and the degrees-of-freedom n, namely, $F\left(x|n\right)=\int_0^x \frac{t^{\frac{n}{2}-1}e^{-\frac{t}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}dt$, is eventually transformed to calculate the gamma density function with parameters $n/2$ and 2 in computer implementation, since chi-square distribution can be perceived as one child of the gamma distribution family with two varying parameters. Knüsel (1986) has proposed a numerical algorithm with some supplement functions and a specified relative accuracy, which has been adopted in many modern statistical softwares, such as Matlab and R language. However, even using this algorithm, computation of the gamma density function is still extremely complex.

As mentioned above, SDE serves as a versatile spatial statistical tool for measuring the geographical distribution of features. Because of this, it has been embedded into many commercial software, like ArcGIS and Stata (Alexandersson 2004). As a result, the algorithm's practicability including the simplicity, speed and precision are of particular concern, which also originally stimulates us pursuing for an innovative approaches. In the subsequent portion, recursion formulas are derived for calculating the confidence levels and an iterative algorithm is proposed for solving the corresponding magnification ratio of the scaled ellipsoids after the prescribed scaling ratio or confidence level is given.

3.3.2.1 The confidence level defined by a scaled SDHE

Here an innovative recursion formula is presented by means of the multiple integral method for calculating the confidence level $P_n(r)$ of a scaled SDHE specified with a magnification factor r in n dimensional space so as to estimate the distribution of a random vector $S \sim N(\mu, C)$, which is equivalent to the confidence level value of $T \sim N(0, I_n)$, whose confidence region is exactly a sphere as explained in section 3.1; namely,

$$\Pr\left\{(S-\mu)^{\mathrm{T}} C^{-1} (S-\mu) \le r^2\right\} = \Pr\left\{T^{\mathrm{T}} T \le r^2\right\}.$$

Therefore, for 1D case,

$$P_1(r) = \Pr\left\{X_1{}^{\mathrm{T}} X_1 \le r^2\right\} = \int_{-r}^{r} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} ds$$

$$= \frac{2}{\sqrt{\pi}} \int_0^r e^{-\frac{x^2}{2}} d\left(\tfrac{x}{\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{r}{\sqrt{2}}} e^{-t^2} dt = \operatorname{erf}\left(\tfrac{r}{\sqrt{2}}\right); \qquad (3.18)$$

where the error function is defined as $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$, with another name being Gauss error function (Andrews 1992), which is a non-elementary function of sigmoid shape constantly occurring in probability, statistics and partial differential equations. As a matter of fact, equation (3.18) formulates the well-known 3-sigma rule of the most common normal distribution as illustrated in Figure3.3.

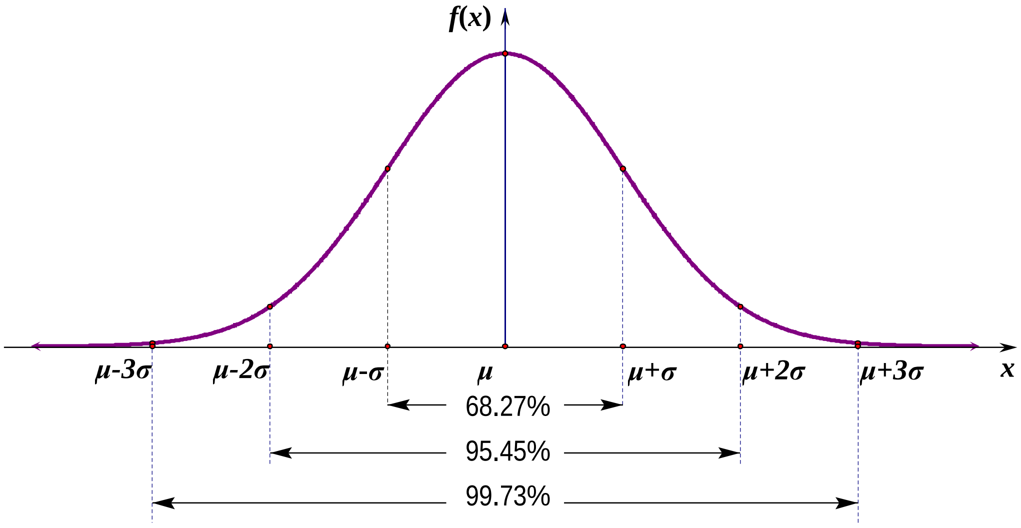


Figure 3.3 The confidence intervals correspond to 3-sigma rule of the normal distribution

For 2D case,

$$P_2(r)=\Pr\left\{X_2{}^{\mathrm{T}}X_2\le r^2\right\}=\iint\limits_{x_1^2+x_2^2\le r^2}\left(\frac{1}{\sqrt{2\pi}}\right)^2 e^{-\frac{x_1^2+x_2^2}{2}}dx_1dx_2$$

$$=\frac{1}{2\pi}\int_0^{2\pi}\int_0^r re^{-\frac{r^2}{2}}drd\theta=1-e^{-\frac{r^2}{2}}; \qquad (3.19)$$

Hereinto, polar coordinate transformation brings into the existence of the penultimate equal sign above. Next, the following Figure 3.4 demonstrates the confidence ellipses corresponding to 1-3 multiples of SDEs in red, blue and green, respectively.

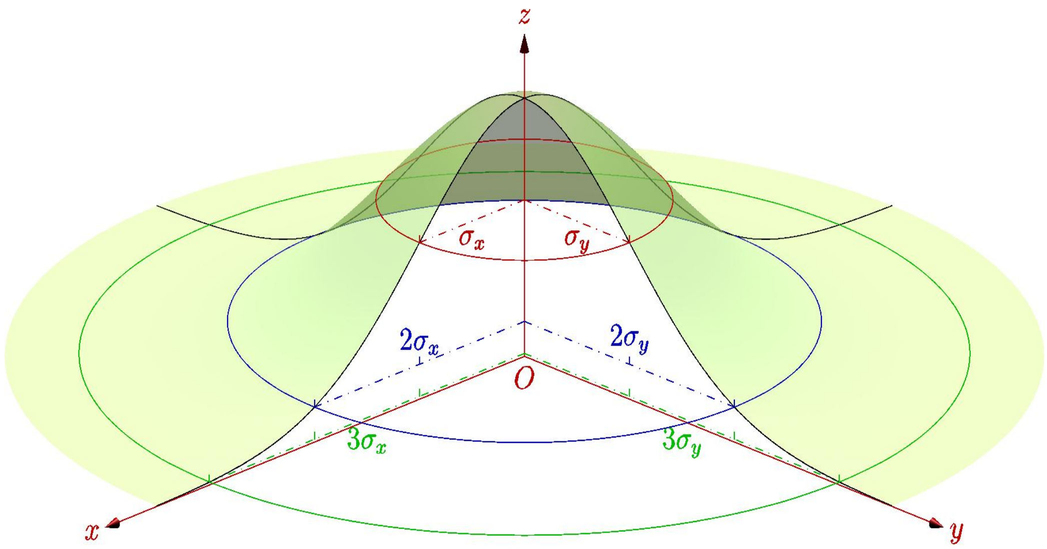


Figure 3.4 The confidence regions corresponds to 1-3 multiples of SDEs

It's worth noting that an inverse formula here exists,

$$r = \sqrt{-2\ln(1-p)}\,, \tag{3.20}$$

for determining the magnification factor r which corresponds to a prescribed confidence level.

Before proceeding to the general formulas applicable in n dimensional space, we introduce the cubature formula (Huber 1982) firstly, which calculates the volume of the n-sphere of radius r, with the quantity proportional to its nth power as follows,

$$V_n(r) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} r^n\,. \tag{3.21}$$

Accordingly, for a general dimensional number $n \geq 3$,

$$P_n(r) = P\left\{X_n{}^{\mathrm{T}} X_n \le r^2\right\} = \iint\cdots\int_{\sum_{i=1}^{n} x_i^2 \le r^2} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n} x_i^2}{2}} dx_1 dx_2 \cdots dx_n$$

$$= \iint\cdots\int_{\sum_{i=3}^{n} x_i^2 \le r^2} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-2} e^{-\frac{\sum_{i=3}^{n} x_i^2}{2}} \left(\iint_{x_1^2 + x_2^2 \le r^2 - \sum_{i=3}^{n} x_i^2} \left(\frac{1}{\sqrt{2\pi}}\right)^2 e^{-\frac{x_1^2 + x_2^2}{2}} dx_1 dx_2 \right) dx_3 \cdots dx_n$$

$$\doteq \iint\cdots\int_{\sum_{i=3}^{n} x_i^2 \le r^2} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-2} e^{-\frac{\sum_{i=3}^{n} x_i^2}{2}} \left(1 - e^{-\frac{r^2 - \sum_{i=3}^{n} x_i^2}{2}}\right) dx_3 \cdots dx_n$$

$$= \iint\cdots\int_{\sum_{i=3}^{n} x_i^2 \le r^2} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-2} e^{-\frac{\sum_{i=3}^{n} x_i^2}{2}} ds_3 \cdots ds_n - \iint\cdots\int_{\sum_{i=3}^{n} x_i^2 \le r^2} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-2} e^{-\frac{r^2}{2}} dx_3 \cdots dx_n$$

$$\doteq P_{n-2} - \left(\frac{1}{\sqrt{2\pi}}\right)^{n-2} e^{-\frac{r^2}{2}} \cdot V_{n-2}(r) = P_{n-2} - \left(\frac{1}{\sqrt{2\pi}}\right)^{n-2} e^{-\frac{r^2}{2}} \cdot \frac{\pi^{\frac{n-2}{2}}}{\Gamma\left(\frac{n}{2}\right)} r^{n-2}$$

$$= P_{n-2}(r) - \left(\frac{r}{\sqrt{2}}\right)^{n-2} \frac{e^{-\frac{r^2}{2}}}{\Gamma\left(\frac{n}{2}\right)}. \qquad (3.22)$$

Hereinto, Γ is the gamma function, with some useful properties: $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, $\Gamma(1) = 1$ and $\Gamma(x+1) = x\Gamma(x)$. It should be noted that the first $\doteq$ comes according to the results for 2D case in terms of equation (3.19) and the second $\doteq$ follows equation (3.21) representing a sphere's volume with radius r and dimensionality of $n-2$. Therefore, equation (3.22) totally characterizes the confidence probability for an arbitrary magnified SDHE with any specified magnification factor r in the form of a recursive formula applicable in any Euclidean space with dimensionality greater than 2. Similar findings regarding the confidence ellipse in terms of dimensionality n less than 3 have been provided in the appendix section of Smith and Cheeseman (1986)'s article. However, to our knowledge, there is no precedent of such analytical expression of confidence levels for an ellipsoid in higher dimensional Euclidean space.

Computation of confidence levels using equation (3.22) is rather simple and efficient. There is only some algebraic manipulations and calculation of the supplement error function $\mathrm{erf}(x)$ if n is assigned to be an odd number. For better quantitatively perceiving the confidence levels of these scaled ellipsoids, the following Table 3.1 lists probability values corresponding to the scaled SDHEs which are magnified with different integral multiples from 1 to 7 and the space dimensionality not exceeding 10.

Table 3.1 Confidence levels of scaled SDHE vary with different magnification factors in spaces with the dimensionality not exceeding 10

Dimensionality	Magnification factor						
	1	2	3	4	5	6	7
1	0.6827	0.9545	0.9973	0.9999	1.0000	1.0000	1.0000
2	0.3935	0.8647	0.9889	0.9997	1.0000	1.0000	1.0000
3	0.1987	0.7385	0.9707	0.9989	1.0000	1.0000	1.0000
4	0.0902	0.5940	0.9389	0.9970	0.9999	1.0000	1.0000
5	0.0374	0.4506	0.8909	0.9932	0.9999	1.0000	1.0000
6	0.0144	0.3233	0.8264	0.9862	0.9997	1.0000	1.0000
7	0.0052	0.2202	0.7473	0.9749	0.9992	1.0000	1.0000
8	0.0018	0.1429	0.6577	0.9576	0.9984	1.0000	1.0000
9	0.0006	0.0886	0.5627	0.9331	0.9970	1.0000	1.0000
10	0.0002	0.0527	0.4679	0.9004	0.9947	0.9999	1.0000

Observed from Table 3.1, 1-3 SDE(s) can encompass approximately 39.35%, 86.47% and 98.89% of all input feature centroids assuming these features follow a planar Gaussian distribution. It is evidently different from the content of our familiar 3-sigma rule. This finding can be conducive to clarify the confusing interpretation of confidence level regarding directional distribution in ArcGIS Help 10.1.

3.3.2.2 The corresponding magnification factor to a prescribed confidence level

Conversely, what size of a magnified SDHE can encompass the scattered features with a prescribed confidence probability? In other words, How to find the magnification factor r corresponding to a specified confidence level p in n dimensional space? This question can be answered by solving the following equation,

$$F(r) = P_n(r) - p, \tag{3.23}$$

with its derivative to be

$$F'(r) = P_n'(r) = \begin{cases} \sqrt{\frac{2}{\pi}} e^{-\frac{r^2}{2}} & n = 1 \\ r e^{-\frac{r^2}{2}} & n = 2 \\ P_{n-2}'(r) + \frac{r^{n-3} e^{-\frac{r^2}{2}}}{2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right)}\left(r^2 - n + 2\right) & n \geq 3 \end{cases}. \tag{3.24}$$

Thus, the approximate scaling ratio r can be solved according to the following iterative algorithm, which is put forward based on Newton method with Armijo rule (Kelley 2003).

Algorithm 3.1 $\textbf{nsolg}(r_0, n, p, \tau_a, \tau_r)$

```
Evaluate $F(r_0) = P_n(r_0) - p$; $\tau \leftarrow \tau_a + \tau_r |F(r)|$.
While $|F(r)| > \tau$ Do
    Calculate the Newton direction $d = -F'(r)^{-1} F(r)$ using (3.23)~(3.24), and
    set $\lambda = 1$.
    While $|F(r + \lambda d)| > (1 - \alpha\lambda)|F(r)|$ Do
        $\lambda \leftarrow \sigma\lambda$ where $\sigma \in [\frac{1}{10}, \frac{1}{2}]$ is the reduction factor of the line search
        computed by minimizing a quadratic polynomial $\varphi(\lambda) = |F(r + \lambda d)|^2$.
    End While
    $r \leftarrow r + \lambda d$
End While
```

Input arguments for this algorithm are the initial iterate r_0 with default value $\sqrt{n-1}$ which is an approximation of inflection point of the S-shape cumulative density function, space dimensionality n, confidence level p, relative and absolute termination tolerances $\tau_a = \tau_r = \sqrt{\varepsilon_{\text{machine}}}$ which need to be prescribed beforehand. Approximate solution with high accuracy can be soon obtained after a few iterations using this algorithm. Table 3.2 has tabulated the magnification ratios of scaled SDHEs for some commonly used confidence levels with space dimensionality not exceeding 10.

Table 3.2 Magnification ratios of scaled SDHE corresponding to different specified confidence levels with space dimensionality not exceeding 10

Dimensionality	Confidence Level (%)					
	80.0	85.0	90.0	95.0	99.0	99.9
1	1.2816	1.4395	1.6449	1.9600	2.5758	3.2905
2	1.7941	1.9479	2.1460	2.4477	3.0349	3.7169
3	2.1544	2.3059	2.5003	2.7955	3.3682	4.0331
4	2.4472	2.5971	2.7892	3.0802	3.6437	4.2973
5	2.6999	2.8487	3.0391	3.3272	3.8841	4.5293
6	2.9254	3.0735	3.2626	3.5485	4.1002	4.7390
7	3.1310	3.2784	3.4666	3.7506	4.2983	4.9317
8	3.3212	3.4680	3.6553	3.9379	4.4822	5.1112
9	3.4989	3.6453	3.8319	4.1133	4.6547	5.2799
10	3.6663	3.8123	3.9984	4.2787	4.8176	5.4395

Seen from Table 3.2, the corresponding magnification factors become larger and larger along with the increase of space dimensionality, indicating that only bigger magnified ellipsoids can maintain the same prescribed confidence level in higher dimensional space compared with the counterpart in lower dimensional space.

3.4 Experiments & Applications

3.4.1 Synthetic data experiments

In this section, two groups of synthetic data are employed to generate the 1-3 multiple SDEs and SDHEs in two and three dimensional spaces, respectively, to depict their aggregation extent and demonstrate the relationship between the scaled ellipse (or ellipsoid) size and their corresponding confidence levels.

3.4.1.1 2D case

Suppose that a series of scattered points $X_i \in R^2$ are randomly generated from a two dimensional Gaussian vector, that is $X_i \sim N(\mu, C)$. The following example employs 100 points with mean $\mu = (2, 3)^{\mathrm{T}}$, and covariance $C = \begin{pmatrix} 0.9 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}$. Overlaying upon these scattered samples, 1-3 multiple SDEs are then created in terms of equations. (3.7)~(3.10) encompassing their geographic distribution with corresponding confidence degrees listed in Table 3.1.

For a better visualization of SDEs in computer imaging, the observed samples can be overlaid by a warning coloration, for example a (gradually varied) red layer processed with a transparency function. Intuitionally it should be inversely proportional to the confidence probability density of the features. By incorporating equation (3.16) into (3.11), an desirable transparency function can be of the following form,

$$f = 1 - e^{-\frac{r^2}{2}}. \tag{3.25}$$

This function can also be considered as a projection of the Gaussian probability density function upon the sample space. In the end, Figure 3.5 presents a visualization of 1-3 multiple SDEs for these 2D scattered points.

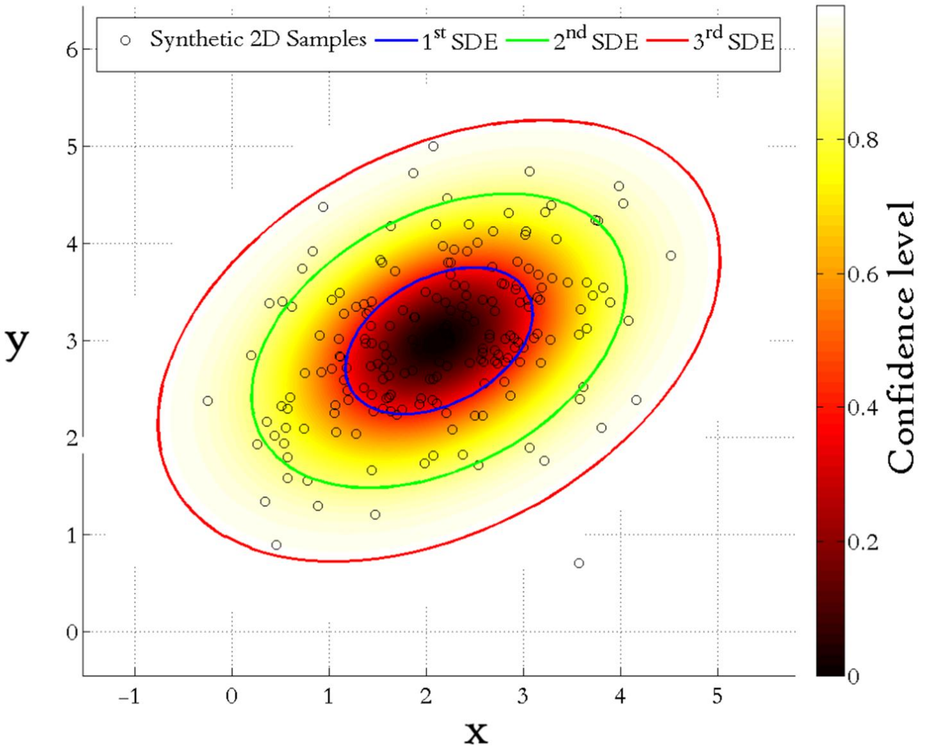


Figure 3.5 Visualization of 1-3 multiple SDEs for 2D scattered points

3.4.1.2 3D case

Once again, suppose that a series of scattered points $X_i \in R^3$ are randomly generated, following 3D Gaussian distribution, that is $X_i \sim N\left(\mu, C\right)$. The following example employs 600 points with mean $\mu = \left(1, 3, 2\right)^{\mathrm{T}}$, and covariance $C = \begin{pmatrix} 8 & -2 & 1 \\ -2 & 8 & 2 \\ 1 & 2 & 5 \end{pmatrix}$. Based on these data samples, Figure 3.6 exhibits 1-3 multiple SDEs constructed in terms of equations. (3.12)~(3.15), encompassing their geographic distribution with corresponding confidence degrees as listed in Table 3.1.

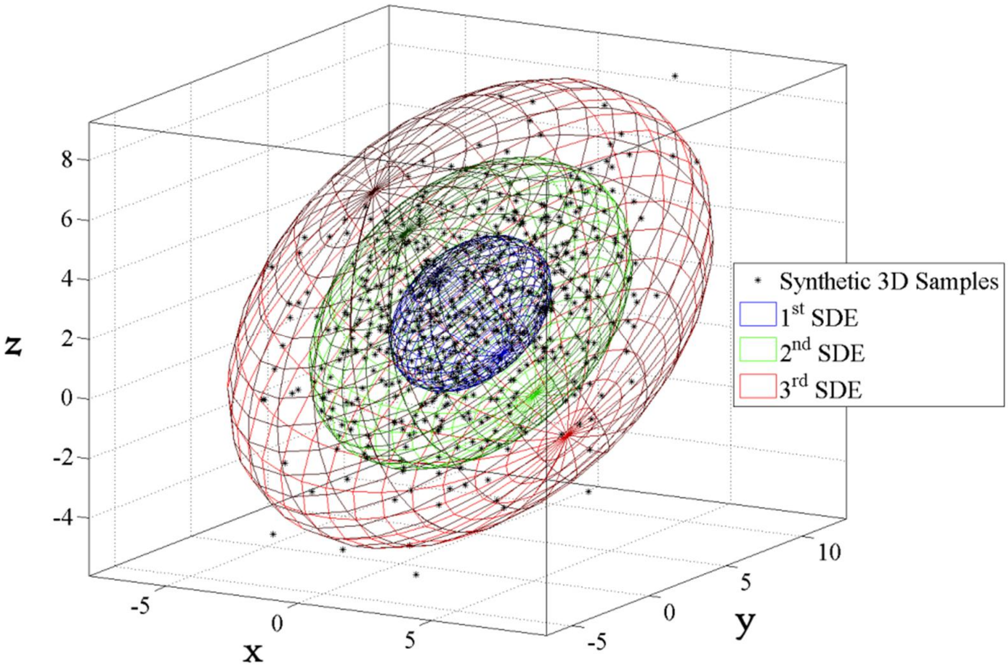


Figure 3.6 Visualization of 1-3 multiple SDEs for 3D scattered points

3.4.2 Spreading analysis of H1N1 infections

The spread of epidemic diseases causes both very serious life risks and social-economic risks. For example, the latest epidemic outbreak in Hong Kong was Swine Flu Virus A (H1N1) in 2009 causing hundreds of deaths and making many residents in fear of fatal infection.

Geographic information science (GIS) serves as a common platform for the convergence of disease surveillance activities. As one of its significant functional components, SDE, as well as SDHE, can be served to understand how a disease cluster together with its evolutionary trend, thereby assisting the epidemiologists or public health officials to raise more effective strategies so as to control the disease spread.

For the epidemic data, a total of 410 human swine influenza infected cases are gathered with epidemiological date and address from 1st May to 26th June on a daily basis released by Center of Health Protection (CHP), Hong Kong. Addresses of infected buildings were then geocoded into the WGS84 coordinate for the subsequent mapping. Exploratory analysis by 1-3 multiple SDEs was then conducted in order to keep the focus limited to only those areas with the most occurrences of infected cases (Figure 3.7). Although the resulting output map is simple, yet it conveys a strong message about where is the most severe region of H1N1 occurring.

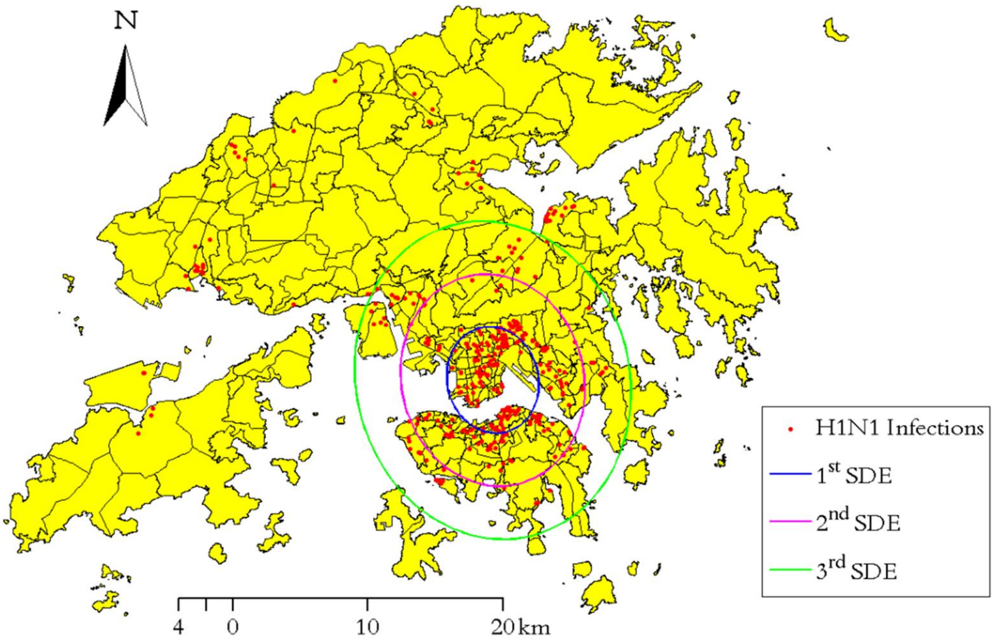


Figure 3.7 Exploratory analysis by 1-3 multiple SDEs for Hong Kong's H1N1

Further, 1-3 multiple SDHEs (in three-dimensional space) were employed for highlighting the spatiotemporal concentrations of H1N1 infections (Figure 3.8).

Apparently, most of the confirmed cases appeared densely during late June in time and converged on both sides of Victoria Harbor, including the Kowloon Peninsula and Hong Kong Island, in space.

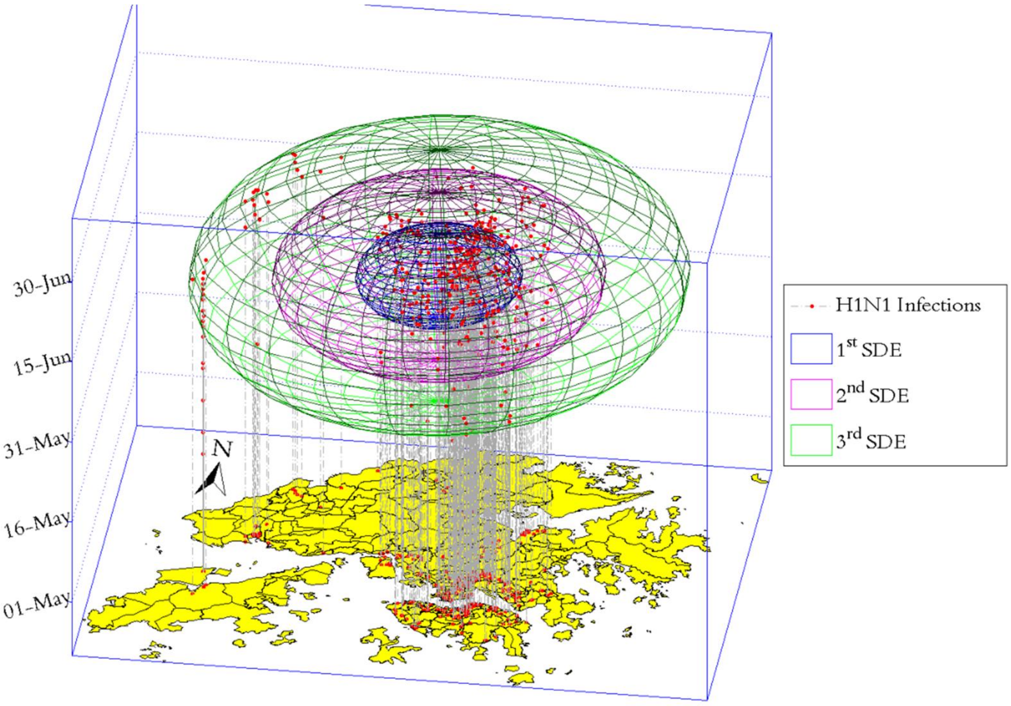


Figure 3.8 Exploratory analysis by 1-3 multiple SDHEs for Hong Kong's H1N1

3.5 Conclusions

In this chapter, confidence analysis of standard deviational ellipse (SDE) and its extension into higher dimensional Euclidean space has been comprehensively explored from the origin, formula derivations to algorithm implementation and applications. Firstly, two existing models are summarized and one novel approach is proposed based on the spectral decomposition of sample covariance for calculating the same SDE. After that, the SDE concept is naturally generalized into higher dimensional Euclidean space, namely the standard deviational hyper-ellipsoid

(SDHE). Then, rigorous recursive formulas were derived for calculating the confidence levels of scaled SDHE with arbitrary magnification ratios in any dimensional space. Such a formula can be employed for tabulating the confidence levels in relation to the magnification ratio and the space dimensionality more efficiently since the results obtained in low dimensional space can still be repeatedly utilized in subsequent higher dimensional spaces, whereas the traditional approach of calculating the chi-square distribution is mainly relying on the complex computation of gamma density function. Besides, an inexact-newton method based iterative algorithm is also proposed for solving the corresponding magnification ratio of a scaled SDHE when the confidence probability and space dimensionality are pre-specified, thereby making a commutatively computation of either the necessary scaled ratio or the confidence level of SDHE when one of these two parameters is given in any dimensional space. These results provide a more efficient manner to supersede the traditional table lookup of tabulated chi-square distribution.

Finally, synthetic data is employed to generate the 1-3 multiple SDEs and SDHEs. And exploratory analysis by means of SDEs and SDHEs are also conducted for measuring the spread concentrations of Hong Kong's H1N1 in 2009.

It is worth noting, standard deviational ellipses (or the SDHE) were derived under the assumption that observed samples follow the normal distribution. Therefore, the SDE tool must be employed with a certain degree of caution when measuring the geographic distribution of concerned features. Particularly, delineation of an area concerned by SDE may not be representative of the hotspot boundaries, but produce ambiguous outcomes when distribution of features is multimodal (Yuill 1971).

Fortunately, the aforementioned normal distribution assumption is no longer indispensable for the confidence ellipses owning to considerable progresses in the last three decades. Nonetheless, these shining ideas emerged during the SDE derivation process still sparkle for prompting innovative advanced models, among which the elliptically contoured distribution (Fang 2004) attracts wide attention, with its contours of constant density being ellipsoids, that is $(x-\mu)^{\mathrm{T}} C^{-1}(x-\mu)=$ *constant*. Amazingly, a scaled SDHE in terms of equations. (3.12)~(3.15) is actually depicted by this formulation, which also lays core foundation for many of the current popular method, such as the minimum covariance determinant estimator (MCD), multivariate kernel density estimation and support vector machine (SVM) with the Gaussian kernel.

Chapter 4 Mathematical Approaches to Disease Spread

Mathematics has long been perceived to be an important tool for characterizing the spreading patterns of infectious diseases. In the following, we firstly give an overview of compartmental models, the traditional approach to modeling dynamics of infectious disease, and then the meta-population model for characterizing the spatiotemporal spreading dynamics of disease infection. Besides, the typical reaction-diffusion models are thoroughly explored with its detailed computer implementation procedures using Runge-Kutta method. As illustrated, these methods lay theoretical foundations for addressing public health challenges and have the potentials of being coupled with powerful computational methods to simulate the real-time invasion process and predict the future evolutionary trend of infectious disease.

4.1 SIR Compartmental Model

In this section, we begin with the SIR model as a starting point for analyzing the temporal evolution behavior of an infectious disease into a well-mixed population. This basic epidemic model is based on dividing the host population into several compartments, each containing individuals of identical characteristics in terms of their status with respect to a disease. In the SIR model, there are three compartments,

- *Susceptible*: those who have no immunity to the infectious agent, therefore might become infected if exposed;

- *Infectious*: those currently infected and can potentially transmit the infection to susceptible individuals they contact;
- *Removed*: individuals who confer immunity to the disease after recovery, and consequently could not be involved in the transmission dynamics any more.

It is customary to denote the numbers of individuals in each of these compartments as S, I and R, respectively. After compartmentalization of the host population, an individual potentially transit his/her status from susceptible to infected when in contact with infected persons, and a transmission may also occur from an infected person to a recovered or immune person. Therefore, such a one-way only disease progress can be represented schematically as:

S ® I ® R.

Let $S(t)$ denote the number of individuals who are susceptible to the disease at time t (measured usually in days), $I(t)$ the number of infected individuals and $R(t)$ the number of individuals who are immune to the same infection strain hence they cannot be infected again during the outbreak. In a more general context, $R(t)$ may refer to either be immune, isolated or deceased individuals.

When a disease breaks out, individuals may be infected and ultimately recovered, and such a dynamic evolution process from one compartment to another can be formulated in terms of a set of differential equations that specify how the sizes of the compartments change over time.

$$\begin{aligned} \frac{dS}{dt} &= -\boldsymbol{b}SI/N, \\ \frac{dI}{dt} &= \boldsymbol{b}SI/N - \boldsymbol{g}I, \\ \frac{dR}{dt} &= \boldsymbol{g}I. \end{aligned} \tag{4.1}$$

Here parameter $\boldsymbol{b}$ is the transmission rate (per capita). The above first equation describes the disease transmission as a result of contacts between susceptible and infectious persons. Each infectious individual transmits the pathogen to $\boldsymbol{b}$ susceptible individual per unit time; nevertheless, new emerging infection arises only when the contact is with a susceptible person in terms of the probability S/N under an implicit assumption that the population is homogenous and randomly mixing. Besides, the incubation time has been ruled out meaning that a susceptible becomes infectious immediately once being infected. The other parameter $\boldsymbol{g}$ is the recovery rate, which corresponds to the inverse of the average of an exponentially distributed time to recovery, thereby making $1/\boldsymbol{g}$ to be the mean infectious period (average duration of the infection).

In addition, there is further assumption that no other entry or exit of population such as birth, natural death or migration from the compartments will be involved, i.e. $S + I + R = N$ for all time t. Correspondingly, if we define three groups as fractions (or densities) of the total population N in lower case, $s = S / N$, $i = I / N$ and $r = R / N$ thereby ensuring $s + i + r = 1$, the model (4.1) can have even more concise form

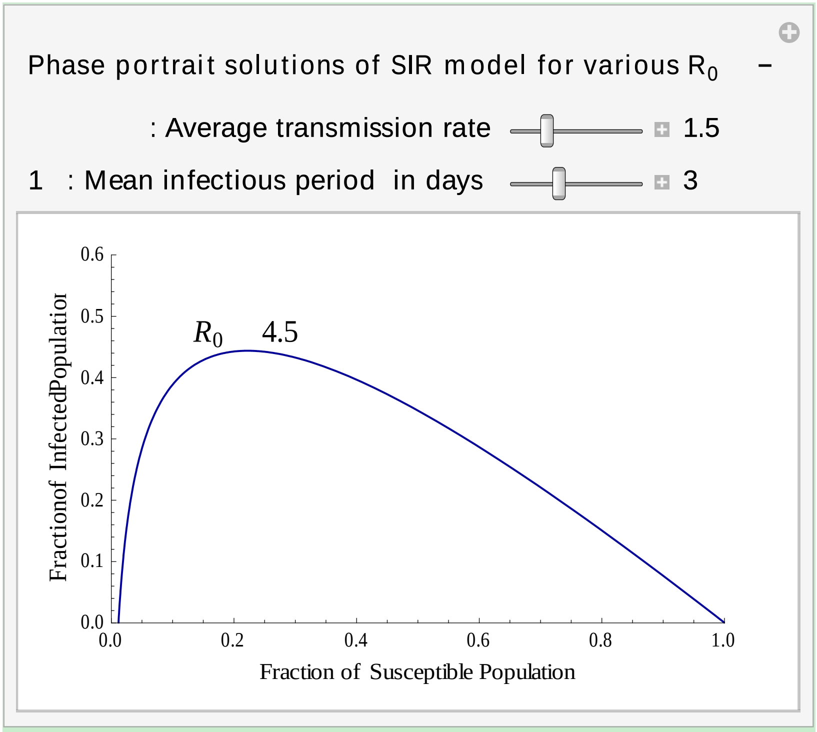
$$\frac{ds}{dt} = -\boldsymbol{b}si\,,\ \frac{di}{dt} = \boldsymbol{b}si - \boldsymbol{g}i\,,\ \frac{dr}{dt} = \boldsymbol{g}i. \qquad (4.2)$$

The formulation of above dynamic system can be completed with the specification of appropriate initial conditions s_0, i_0 and r_0. Often, epidemics are modeled with an introduction of a single infectious individual into a society where everyone else is susceptible, meaning that $s_0 = 1/N$, $i_0 = 1 - s_0$ and $r_0 = 0$.

By taking the ratio of the first two equations, we obtain $\frac{di}{ds} = -1 + \frac{\boldsymbol{g}}{\boldsymbol{b}s}$, which can be integrated immediately to yield

$$i = i_0 + s_0 - s + \boldsymbol{g}/\boldsymbol{b}\ln\left(s/s_0\right). \qquad (4.3)$$

This gives an exact expression for characterizing the dynamic density i as a function of s. Figure 4.1 shows the instantaneity interactive flexible plots of $i(s)$ according to various values of parameters $\boldsymbol{b}$ and $\boldsymbol{g}$ depicting the phase portrait solutions.



Phase portrait solutions of SIR model for various R_0

: Average transmission rate 1.5

1 : Mean infectious period in days 3

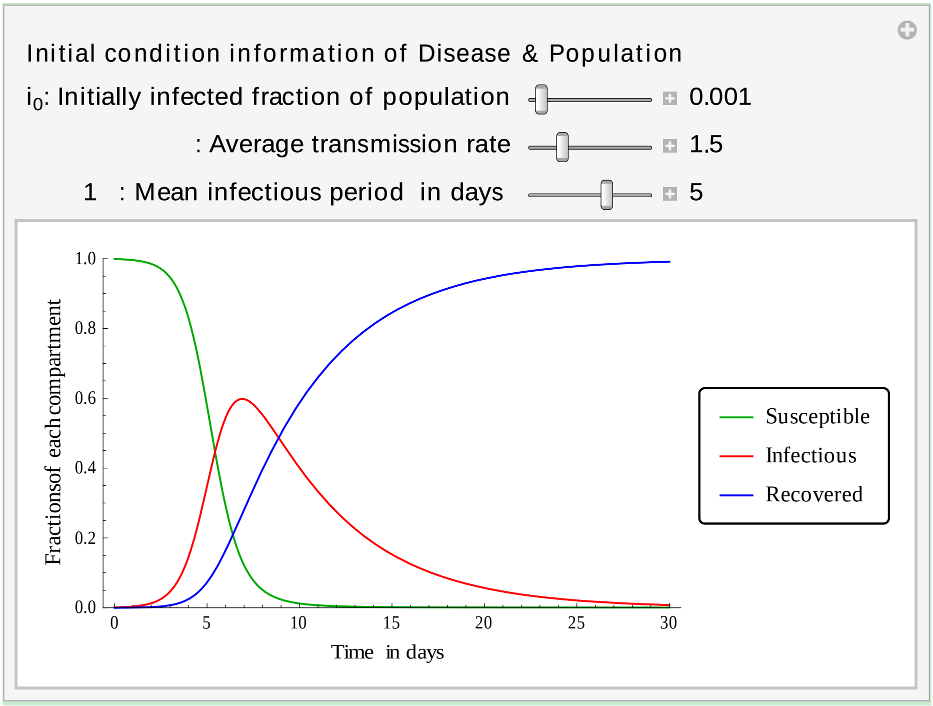
R_0 4.5
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Fractionof InfectedPopulation
Fraction of Susceptible Population

Figure 4.1 Phase portrait solutions of the basic SIR model

According to (4.2) , the implicit relationship between the infectious and susceptible compartmental population for a newly invading infectious disease with initial moment begins at the bottom right corner of the graph ($s_0 \simeq 1$ and $i_0 \simeq 0$). These dynamic curves are labeled by the basic reproduction ratio $\boldsymbol{R = b/g}$.

Although the exact solution (4.3) for phase portrait has been obtained, it is unfortunately not possible to deduce an analytical formula for describing each compartmental densities s_t , i_t and r_t trajectories over time, even for this extremely simple model. Solving the system of differential equations can be achieved numerically by using Euler method, or even the more precisely Runge-Kutta-Fehlberg (denoted RKF45) method (Lapidus and Pinder 2011).

If everyone is initially susceptible ($s(0) = 1$), then a newly introduced infected individual can be expected to infect other people at the rate bN during the expected infectious period $1/g$. Thus, this first infective individual can be expected to infect $R_0 = bN/g$ individuals. The number R_0 is called the basic reproduction number which is unquestionably one of the most important quantity to consider when analyzing any epidemic model for an infectious disease. Particularly, R_0 determines whether an epidemic can occur at all. It actually varies with the evolution of the disease over time, which indicates how the risk grade is of the current disease.



Initial condition information of Disease & Population

i_0: Initially infected fraction of population 0.001

: Average transmission rate 1.5

1 : Mean infectious period in days 5

Figure 4.2 Example of SIR models for charactering the temporal evolutionary behavior

Extended compartmental model with various modifications (including birth and death rates, migration, exposed compartment, vaccination campaign and further age-structured models) have proven extremely useful in analyzing epidemics and particularly for modeling the spread of moderately to highly infectious disease in a larger and well-mixed society.

4.2 Meta-Population Model

Intuitively, disease transmission can be predominantly perceived to be a localized process in most circumstances. For directly transmitted diseases, transmissions most likely occur between individuals with the most intense interaction. In addition, movement of individuals may facilitate the geographical spread of infectious diseases. The population distribution and the interaction patterns linking different groups are two important factors to consider that influence the infectious disease spread across space and time.

This chapter is mainly concerned with elucidating in more detail some typical population-based dynamic models. Before the late 1980s, rigorous analytical results for spatial epidemiological models remain rare (Keeling and Rohani 2008), however, the increasing ease of access to computational power has turned the simulation of such models into reality (Levin et al. 1997).

Most of the spatial models make provisions upon spatial scale of interaction and the scale at which hosts are aggregated. However, it operates difficult in practice to choose a “correct” scale, because creating many subpopulations of fine scale leads to computational prohibition, whereas subdividing the population at large scale may eliminate the spatial effects of primary interest. Therefore, answer to exploring the optimal, most informative scale (Keeling and Grenfell 1997, Pascual et al. 2001) should be based on sound epidemiological knowledge.

A meta-population model is a particular type of multi-group model in which a population is distributed into a collection of n spatially discrete groups that are linked to one another. Most often it is assumed that the individuals within a group are well-mixed and the groups of subpopulations are coupled to one another in some way. Coupling terms are used to represent the way that infection can be spread between groups. The spatial scale represented by the groups depends on the context of the study. For example, Lai et al. (2013) employed an environmental and social variations incorporated SEIR model upon the partitioned grids of HKSAR region consisting of 500 × 500 metric cells, and Brockmann and Helbing (2013b) explored the hidden geometry of complex, network-driven contagion phenomena using the global air-traffic lines. A meta-population model may also be regarded abstractly as a graph, with each population patch as a vertex of the graph and each travel route between two patches as an edge of the graph. Since travel may be in either direction, the graph is bidirectional. The population of a patch is said to have direct access to another patch if there is a route linking the two patches. If there is a sequence of more than one route to another patch, the population is said to have indirect access to the second patch. For simplicity, we will always assume that a meta-population is fully connected, that is, that there is a route linking every pair of patches.

Although meta-populations are one of the simplest spatial models, they are also one of the most applicable to modeling many human diseases like SARS or pandemic influenza A (H1N1). When used in studying the geographic spread of infectious diseases, these models do not incorporate explicit mobility among groups; rather, they attempt to mimic the effect of explicit mobility by defining an appropriate contact matrix that represents the strength of contact within and between groups (Brockmann and Helbing 2013a). In recent studies, the meta-population model is generally incorporated with other models involving description of the infectious diseases dynamics (Lai et al. 2013), such as the SIR model. Recall that the generalized SIR model with demography can be formulated as follows, with differences compared with the basic SIR is the introduction of the population births and deaths.

$$\begin{aligned} \frac{dS}{dt} &= \upsilon N - \beta SI/N - \mu S, \\ \frac{dI}{dt} &= \beta SI/N - \gamma I - \mu I, \\ \frac{dR}{dt} &= \gamma I - \mu R. \end{aligned} \tag{4.4}$$

The rate at which individuals (in any epidemiological class) suffer natural mortality is given by μ and the population's crude birth rate is denoted by υ. For brevity, these two rates are always supposed to be the same so as to ensure the total population size does not change through time ($\frac{dS}{dt}+\frac{dI}{dt}+\frac{dR}{dt}=0$).

A deterministic SIR cross-coupled meta-population model can be written to be the following differential equations:

$$\begin{aligned} \frac{dS_i}{dt} &= \upsilon_i N_i - \lambda_i S_i - \mu_i S_i, \\ \frac{dI_i}{dt} &= \lambda_i S_i - \gamma_i I - \mu_i I_i, \\ \frac{dR_i}{dt} &= \gamma_i I_i - \mu_i R_i, \end{aligned} \tag{4.5}$$

where the subscript i defines parameters and variables related to subpopulation i. Parameters υ_i and μ_i denote the birth rate and mortality rate, respectively. Besides, parameter γ_i is still the recovery rate. The infection force, λ_i, incorporates transmission from both the number of infected within subpopulation i and the coupling to other subpopulations. Demographic and epidemiological parameters may vary within subpopulations, reflecting differences in the local environments (Finkenstädt and Grenfell 1998, Broadfoot et al. 2001). In all honesty, meta-populations provide a useful framework for modeling disease dynamics for hosts which can be naturally partitioned into several spatial sub-units.

The relationship between infection force for population i and the number of infectious individuals in population j depends on the transmission mechanism of and between the two populations. In general terms, the infection force is written as a sum:

$$\lambda_i = \beta_i \sum_j \rho_{ij} \frac{I_j}{N_i}, \tag{4.6}$$

where parameter β_i is still the transmission rate (per capita), and the coefficients, ρ, measuring of the strength of interaction between populations. Specifically, ρ_{ij} measures the relative transmission rate to subpopulation i from subpopulation j.

An important aspect of this formulation concerns the precise scaling with population size in the expression of λ_i. The equation (4.6) contains N_i in the denominator, which reflects the implicit assumption that transmission takes place in population i, presumably resulting from the movement of an infectious individual from population j. Alternatively, the transmission due to a susceptible individual from population i picking up the infection during a temporary visit to population j would be incorporated by placing N_j in the denominator. Therefore, the force of infection within a subpopulation can be expressed as a weighted sum of the prevalence in all populations.

The meta-population models serve as spatial models which are usually high dimensional and contain many parameters. Simulations can easily be performed with parameters relevant for a particular disease with given demography and spatial structure. These models assume that each group population is sufficiently large so that a deterministic model is appropriate and there is homogeneous mixing with each subpopulation. Stochastic effects may be significant when group populations are small (Aparicio et al. 2002).

4.3 Reaction-Diffusion Equations

As a typical representative of meta-population model, square lattice-based model are often employed when there is no idea of partitioning the population into discrete subpopulations. However, one major disadvantage of the lattice-based models is the stationary lattice structure which may restrict the space discretization process. Resolution of an individual's position is generally limited by the scale of each grid cell. An alternative formulation is the reaction-diffusion equations which describe the dynamics of populations in continuous space using partial differential equations (PDE) (Murray 2003). Although these models generally provide theoretical predictions (Beardmore and Beardmore 2003, Reluga 2004), However, the insights from this type of model have also proved to be invaluable in understanding the spatial spread of infection (Kao 2003, Lai et al. 2013). Main theoretical advantage of the reaction-diffusion equations is the deterministic and tractable nature of the continuous-space models.

The standard PDE models are derived under the assumption that infectious individuals only transmit disease to susceptible ones at their current location, and that all individuals are freely diffusing at random through the landscape. A typical PDE model for a disease with SIR-type dynamics are,

$$\begin{aligned}\frac{\partial S}{\partial t} &= \upsilon N - \beta SI/N - \mu S + D_S \nabla^2 S,\\ \frac{\partial I}{\partial t} &= \beta SI/N - \gamma I - \mu I + D_I \nabla^2 I,\\ \frac{\partial R}{\partial t} &= \gamma I - \mu R + D_R \nabla^2 R,\end{aligned} \tag{4.7}$$

where S, I and R are functions of both space and time, and represent the local number of susceptible, infectious, and recovered individuals, and as always $N = S + I + R$ when omitting the demographic variations. Hence, if we're dealing with a two-dimensional landscape, $S(x, y, t)$ represents the number of the susceptible individuals at location (x, y) at time t.

Here the Laplacian operator ∇^2 is involved to characterize the local diffusion of individuals through space. Since ∇ is shorthand for the change rate of the quantity, thus ∇^2 is the change in the rate of change. In two dimensions, the diffusion term for the susceptible compartment becomes:

$$\nabla^2 S = \frac{\partial^2 S}{\partial x^2} + \frac{\partial^2 S}{\partial y^2} \, . \tag{4.8}$$

Inclusion of these spatial derivatives mimics the diffusion of individuals in the real-world situations. In general, susceptible, infectious, and recovered individuals may corresponds to different diffusing rates (D_S, D_Y, and D_Z), agreeing with the fact that sick individuals are unlikely to move.

For better understand the diffusion role of such PDE model, considering a group of susceptible individuals initially piled at the origin $(0,0)$. Ignoring demography, our equation becomes:

$$\frac{\partial S}{\partial t} = D_S \nabla^2 S \, .$$

The corresponding solution is,

$$S(x,y,t) \propto \frac{1}{4\pi D_S t} \exp\left(-\frac{\left(x^2+y^2\right)}{4D_S t} \right) . \tag{4.9}$$

This is actually the Gaussian distribution, with an ever-expanding bell-shape. The diffusion parameter D_S governs the varying speed of the Gaussian distribution. Therefore, the scene can be acquired of the Gaussian-like distributed individuals spreading radically away from the origin.

Reaction-diffusion models come out relying on spatial diffusion of hosts, with assumption of local interactive transmission of infection, to characterize the spatio-temporal dynamics of infectious disease. Very few PDE models can be analytically solved, with majority of these models being simulated via numerical methods. The common practice is firstly to translate the differential terms via discretizing the domain space into a lattice formulation.

Impose a square lattice structure onto the space of concern, with the lattice side length of a distance d. Therefore the lattice solution $S_{i,j}(t)$ is expected to approximate the PDE solution $S(i \times d, j \times d, t)$. We are familiar with the temporal derivatives (e.g. $\frac{d}{dt}$) in the ODE (ordinary differential equation) model, which can be numerically integrated forward in time using methods such as forward Euler or Runge-Kutta (Lapidus and Pinder 2011), However, the spatial derivatives involved in the PDF model normally need to be expressed in terms of the lattice structures for numerical implementation. In the following, the number of susceptible individuals, $S_{i,j}$, is taken as the instance for better interpretation of lattice formulation procedures.

Firstly, the second derivative along x dimension can be approximated as the change in the first derivative of S between $(i+\frac{1}{2}, j)$ and $(i-\frac{1}{2}, j)$ divided by the distance, d. That is,

$$\frac{\partial^2 S_{i,j}}{\partial x^2} \approx \frac{\frac{\partial S_{i+\frac{1}{2},j}}{\partial x} - \frac{\partial S_{i-\frac{1}{2},j}}{\partial x}}{d}.$$

Further, a similar handling for the first derivatives leading to the approximation,

$$\frac{\partial^2 S_{i,j}}{\partial x^2} \approx \frac{\left(\frac{S_{i+1,j}-S_{i,j}}{d}\right)-\left(\frac{S_{i,j}-S_{i-1,j}}{d}\right)}{d} \approx \frac{S_{i+1,j}-2S_{i,j}+S_{i-1,j}}{d^2}.$$

As thus, the full diffusion term becomes,

$$D_S \nabla^2 S_{i,j} = D_S \frac{\partial^2 S_{i,j}}{\partial x^2} + D_S \frac{\partial^2 S_{i,j}}{\partial y^2} \approx \frac{D_S}{d^2}\left(S_{i+1,j}+S_{i-1,j}+S_{i,j-1}+S_{i,j+1}-4S_{i,j}\right). \quad (4.10)$$

After substituting the second derivatives by spatial difference approximation in the PDE model, we arrive at our familiar ODE equation for each lattice point, which can be solved with ease in our familiar manner. Spatial diffusion therefore behaves like the movement of individuals between the four nearest-neighbor lattice sites. And D_S/d^2 is the coupling rate reflecting how fast of the individuals moving.

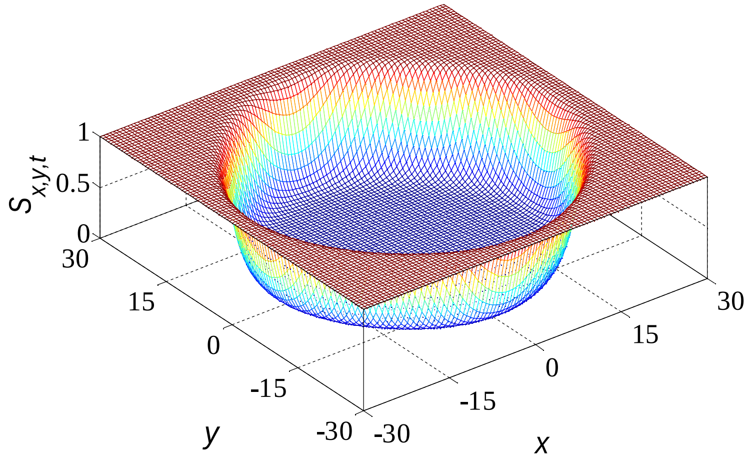
When considering the spatiotemporal dynamic diffusion process of the local density, but not the population size, of the susceptible, infectious, and recovered compartments, the foregoing SIR-type PDE model (4.7) becomes:

$$\begin{aligned}\frac{\partial s}{\partial t} &= \upsilon - \beta si - \mu s + D_s \nabla^2 s,\\ \frac{\partial i}{\partial t} &= \beta si - \gamma i - \mu i + D_i \nabla^2 i,\\ \frac{\partial r}{\partial t} &= \gamma i - \mu r + D_r \nabla^2 r,\end{aligned} \quad (4.11)$$

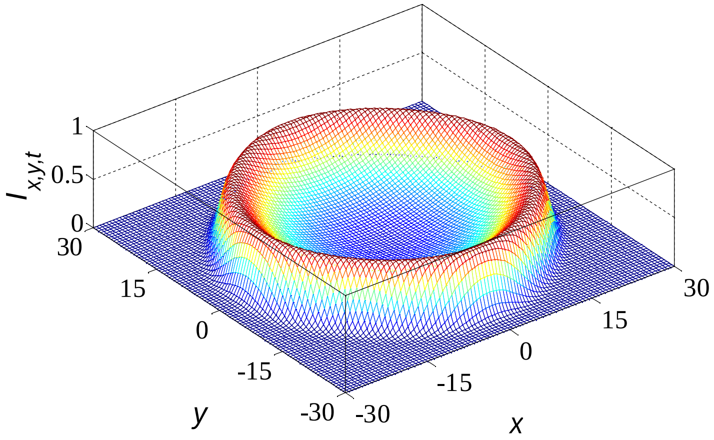
where the lowercase characters s, i and r represent the local density of the spatiotemporal distribution for susceptible, infectious, and recovered compartments. Figure 5.2 below illustrates a PDE model for characterizing the spatiotemporal spreading dynamics of an SIR-type infection. Starting at a point source, the infection spreads as an expanding epidemic wave, leaving secondary oscillations around the endemic equilibrium in its wake. The top left-hand three subfigures demonstrate a snapshot of such circular wave fronts for the susceptible, infectious and recovered compartments, respectively. However, the fourth graph located in lower right corner plots disease prevalence against the distance from initial source (blue solid line). This curve approximately coincides with the solution of the standard (non-spatial) SIR model (red dash-dot line), hinting a deeper relationship between each other. Involved mathematical deviations indicate that the PDE leads to a traveling wave with constant velocity (Keeling and Rohani 2008) once transient dynamics fading away.

Once an invading virus attaches a population, it may spread and expand its geographic range, at times an inexorable march begins. The invasion of new virus such as H1N1 influenza by organisms is a fundamental ecological process. Invading viruses have had tremendous epidemiological and socioeconomic, sometimes even potential catastrophic impact upon our whole human society. As thus, by integrating the related environmental information, here the discussed reaction diffusion models owe great potential to understand the real-time invasion process (Brockmann and Helbing 2013a) and predict the future evolution trend of infectious disease (Lai et al. 2013), as well as providing assistance for epidemiologist or government officers to formulate effective defensive measures timely.

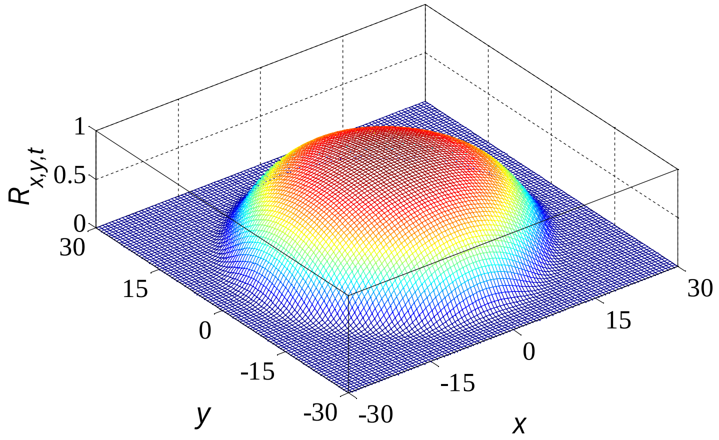
Proportion Dynamic of Susceptible Population



Proportion Dynamic of Infectious Population



Proportion Dynamic of Recovered Population



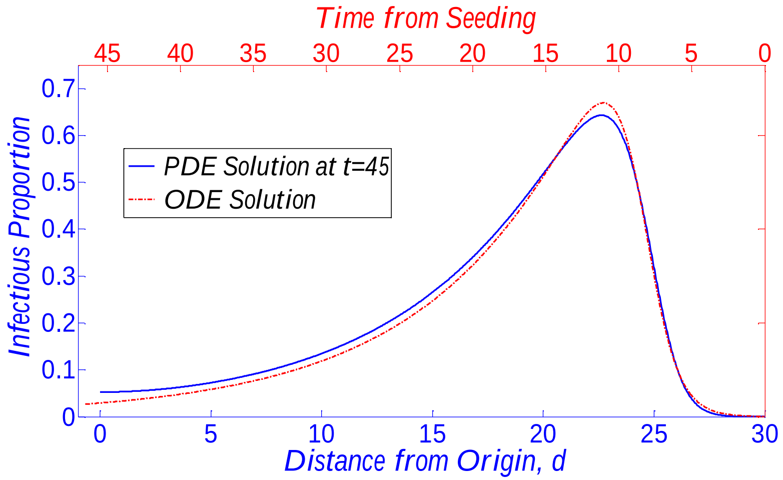


Figure 4.3 Snapshots of population proportion dynamic solved from the SIR-type PDE model

The top left-hand three subfigures present snapshot circular wave fronts at time $t = 45$ of the density of the susceptible, infectious and recovered compartments, respectively. This PDE model was simulated upon the region $x, y \in [-30, 30]$ which has been divided into a 101×101 lattice, with parameters configured as follows: $\upsilon = \mu = 10^{-3}$, $\beta = 1$, $\gamma = 0.1$, $D_s = D_i = D_r = 0.1$. Detailed implementation procedure are presented in section 4.4 by being integrated with difference approximation upon lattice formulation and Runge-Kutta methods of order four with step size $t = 0.1$. The lower right-hand figure compares the distribution of infection (at time $t = 45$) as a function of diffusion distance d from the initial source, with the results from a standard (non-spatial) SIR model equipped with the same basic parameters, whose solution domain ranging from 0 to 45. For the non-spatial model the x-axis represents the time from the beginning of this epidemic simulation, whereas for the PDE model the x-axis represents the distance from the initial point of infection. The values on the x-axis have been scaled by the wave speed so that the two curves coincide to the largest extent.

4.4 Runge-Kutta method based implementation

Population Number based SIR Model:	Population Proportion based SIR Model:
$\frac{\partial S}{\partial t} = \upsilon N - \beta SI/N - \mu S + D_S \nabla^2 S,$	$\frac{\partial s}{\partial t} = \upsilon - \beta si - \mu s + D_s \nabla^2 s,$
$\frac{\partial I}{\partial t} = \beta SI/N - \gamma I - \mu I + D_I \nabla^2 I,$	$\frac{\partial i}{\partial t} = \beta si - \gamma i - \mu i + D_i \nabla^2 i,$
$\frac{\partial R}{\partial t} = \gamma I - \mu R + D_R \nabla^2 R,$	$\frac{\partial r}{\partial t} = \gamma i - \mu r + D_r \nabla^2 r,$

Solution: Discretize the region into mesh grids and spatially approximate the second-order derivatives (the Laplacian Operator) involved by difference of subpopulation number/density from the neighbored lattices; and utilizing the Runge-Kutta method along the time dimension.

Parameter Configuration: $\upsilon = \mu = 10^{-3}$, $\beta = 1, \gamma = 0.1$, $D_s = D_i = D_r = 0.1$.

$$\begin{cases} \frac{\partial s}{\partial t} = \upsilon - \beta si - \mu s + D_s \nabla^2 s & \underline{\underline{def}} \quad f_s(t,s,i,r) \\ \frac{\partial i}{\partial t} = \beta si - \gamma i - \mu i + D_i \nabla^2 i & \underline{\underline{def}} \quad f_i(t,s,i,r) \\ \frac{\partial r}{\partial t} = \gamma i - \mu r + D_r \nabla^2 r & \underline{\underline{def}} \quad f_r(t,s,i,r) \end{cases}$$

$$s_{m+1} = s_m + \frac{h}{6}(K_{s1} + 2K_{s2} + 2K_{s3} + K_{s4})$$

$$\begin{cases} K_{s1} = f_s(t_m, s_m, i_m, r_m) \\ K_{s2} = f_s\left(t_m + \frac{\Delta t}{2}, s_m + \frac{\Delta t}{2}K_{s1}, i_m + \frac{\Delta t}{2}K_{i1}, r_m + \frac{\Delta t}{2}K_{r1}\right) \\ K_{s3} = f_s\left(t_m + \frac{\Delta t}{2}, s_m + \frac{\Delta t}{2}K_{s2}, i_m + \frac{\Delta t}{2}K_{i2}, r_m + \frac{\Delta t}{2}K_{r2}\right) \\ K_{s4} = f_s(t_m + \Delta t, s_m + \Delta t K_{s3}, i_m + \Delta t K_{i3}, r_m + \Delta t K_{r3}) \end{cases}$$

$$i_{m+1} = i_m + \frac{h}{6}(K_{i1} + 2K_{i2} + 2K_{i3} + K_{i4})$$

$$\begin{cases} K_{i1} = f_i(t_m, s_m, i_m, r_m) \\ K_{i2} = f_i\left(t_m + \frac{\Delta t}{2}, s_m + \frac{\Delta t}{2}K_{s1}, i_m + \frac{\Delta t}{2}K_{i1}, r_m + \frac{\Delta t}{2}K_{r1}\right) \\ K_{i3} = f_i\left(t_m + \frac{\Delta t}{2}, s_m + \frac{\Delta t}{2}K_{s2}, i_m + \frac{\Delta t}{2}K_{i2}, r_m + \frac{\Delta t}{2}K_{r2}\right) \\ K_{i4} = f_i(t_m + \Delta t, s_m + \Delta t K_{s3}, i_m + \Delta t K_{i3}, r_m + \Delta t K_{r3}) \end{cases}$$

$$r_{m+1} = r_m + \frac{h}{6}(K_{r1} + 2K_{r2} + 2K_{r3} + K_{r4})$$

$$\begin{cases} K_{r1} = f_r(t_m, s_m, i_m, r_m) \\ K_{r2} = f_r\left(t_m + \frac{\Delta t}{2}, s_m + \frac{\Delta t}{2}K_{s1}, i_m + \frac{\Delta t}{2}K_{i1}, r_m + \frac{\Delta t}{2}K_{r1}\right) \\ K_{r3} = f_r\left(t_m + \frac{\Delta t}{2}, s_m + \frac{\Delta t}{2}K_{s2}, i_m + \frac{\Delta t}{2}K_{i2}, r_m + \frac{\Delta t}{2}K_{r2}\right) \\ K_{r4} = f_r(t_m + \Delta t, s_m + \Delta t K_{s3}, i_m + \Delta t K_{i3}, r_m + \Delta t K_{r3}) \end{cases}$$

4.5 Applicability of Reaction-Diffusion Equations

Recall equation of (4.11), we consider spatio-temporal dynamic diffusion process of the local density of susceptible, infectious, and recovered compartments, a typical SIR-type PDE model is,

$$\begin{aligned} \frac{\partial s}{\partial t} &= \upsilon - \beta si - \mu s + D_s \nabla^2 s, \\ \frac{\partial i}{\partial t} &= \beta si - \gamma i - \mu i + D_i \nabla^2 i, \\ \frac{\partial r}{\partial t} &= \gamma i - \mu r + D_r \nabla^2 r, \end{aligned}$$

where the lowercase characters s, i and r represent the local density of the spatio-temporal distribution for susceptible, infectious, and recovered compartments.

4.5.1 Adaptive Diffusion Coefficient Matrix

A more general reaction-diffusion equation normally comprises a reaction term and a diffusion term, i.e. the typical form is as follows:

$$u_t = f(u) + D\nabla^2 u \,. \tag{4.12}$$

For an epidemic, $u = u(x,t)$ denotes the state variable describing density\concentration of the subpopulations at position $x \in \Omega$, at time t. ∇^2 denotes the Laplace operator. Thus, the first term, $f(u)$, describes the interactive disease infection process. And the second term characterizes the “physical diffusion” in space, including D as diffusion coefficient matrix.

Therefore, big challenge for applying the RD model is how to assign the diffusion coefficient so as to incorporate hosts mobility patterns into the disease spreading process, thereby forecasting the spatial spread trends of infectious disease. The diffusion coefficient can be determined by further integrating with the layers of hosts distribution, transmission routes, etc. so as to reflect the complicated disease spreading situation. Therefore, It is also possible, that the diffusion coefficient matrix D may depend on u, and/or explicitly on position x and time t. Sometimes, the geographic barrier constraints may also be added to limit the spreading range.

Approaches from meta-population model family usually consider both the temporal evolution and spatial propagation simultaneously. Analogous models to the RD model within the meta-population model family include the Lattice-based Model (Lai et al. 2013) and Contact Network Modeling (Brockmann and Helbing 2013a), which seem to be more popular from the application aspect. Specifically, the RD model is more suitable for the regionally distributed infectious disease and some other “diffusion” related applications with explicit transmission patterns which can be characterized by the relevant environmental factors.

Table 4.1 Other potential applications of Reaction Diffusion Equations

Other potential applications	Relevant factors
Hand-mouth-foot disease	Distribution of population, village, paths
Atmosphere pollution, PM 2.5	Wind speed & direction, Chemical plant location
Water pollution	River distribution, flow velocity
……	……

4.5.2 Advantages & Disadvantages

The Reaction Diffusion equation serves as a spatial model. Simulations can easily be performed with parameters relevant for a particular disease with given demography and spatial structure. Main theoretical advantage of the reaction-diffusion equations is the deterministic and tractable nature of the continuous-space models. However, this model assumes that each group population is sufficiently large and homogeneous mixing with each subpopulation. Stochastic effects may be significant when group populations are small.

Chapter 5 Investigating the Spatiotemporal Proximity Impact

Epidemic waves of emerging infectious disease come out successively one after another ever since the new millennium beginning. Spatiotemporal analysis may potentially contribute to characterizing the temporal evolutionary process and revealing the possible spatial propagation patterns, thereby being conducive to the optimal allocation of limited public health resources. Though different emerging infectious diseases may vary in transmission route and hazard level, yet they usually exhibit the aggregation tendency both in time and space. As such, this chapter intends to propose an innovative approach for investigating the spatiotemporal proximity impact upon the illness onset risk prediction of emerging infectious disease.

5.1 Background

Understanding the contact patterns between hosts and reservoir of infectious agents, as well as the corresponding transmission rates, are critical to developing effective responding measures so as to cope with the sudden outbreak of an infectious disease. However, investigation of the entire human's contact activities involved is often impracticable. Alternatively, spatiotemporal proximity, revealed by the First Law of Geography (Tobler 1970), can be employed as an metric for inferring the contact transmission modes (Williams et al. 2014). Numerous spatiotemporal analysis methods have emerged in the past few years. For instance, Zhang et al. (2010) proposed a novel spatiotemporal kernel technique for evaluating the global highly pathogenic avian influenza H5N1 outbreaks. Shi et al. (2014) analyzed the spatiotemporal pattern of hand-foot-mouth disease in China by using empirical orthogonal functions. Besides, Yu et al. (2014a) proposed an online spatiotemporal prediction approach by integrating the Susceptible Infected Recovered (SIR) model into the Bayesian maximum entropy (BME) framework for dengue fever epidemic in Kaohsiung (Taiwan). Furthermore, Jandarov et al. (2014) employed the Gaussian process approximation to emulate the gravity model for inferring the spatiotemporal dynamics of measles outbreaks in England and Wales.

Forecasting the spatiotemporal spreading dynamics of an infectious disease can be conducive to the decision-making regarding optimal allocation of limited public health resources and development of preventive intervention measures. Typical previous forecasting approaches (Nsoesie et al. 2014) include the time series models, compartmental models, agent-based models and the meta-population models, which are generally differing from the prediction mechanism and problem scale. However, all these models are always more or less accompanied by some inherent limitations. For instance, predictions provided by time series models (Held and Paul 2012, Kane et al. 2014) may be inconsistent with the epidemic (influenza) activity due to the seasonal variation. Homogeneous population assumption in the compartmental models (Biswas et al. 2014) likely fails to capture the contact patterns corresponding to demographic structure of the entire population. One major difficulty in applying agent-based models and meta-population models (Ajelli et al. 2010) is the challenge of empirically justifying modeling assumptions under which they operate, compounded by our lack of recognizing the human behavior via contact networks. Meanwhile, though some innovative spatiotemporal methods for the epidemics have been emerging in the recent related researches (Tsui et al. 2011, Angulo et al. 2013, Meyer et al. 2012), they are still inadequate for depicting any of the currently global pandemic threats, in general.

What is more, although there are indeed some well-thought models (Chowell et al. 2006, Balcan et al. 2010), whereby almost all of the relevant factors have been taken into account for simulating the disease spreading dynamics. However, during the early stage of a newly emerged infectious disease, external environmental factors (transmission route, climatic factor, etc.) normally still need to be further excavated (or even unknowable) whether they are relevant with the disease spreading or not. As thus, apart from the spatiotemporal location information of laboratory-confirmed cases, external environmental factors may not be directly employed, thereby cutting down the utilization potentiality of these complicated models.

On this occasion, here an innovative approach is proposed, totally based on the spatiotemporal location information of the laboratory-confirmed cases, without involving the external environmental factors. It is expected to be equipped into the disease detection and rapid response system (Yang et al. 2011) for forecasting the spatiotemporal illness onset risk especially during the early outbreak stage for an emerging infectious disease. The spatiotemporal proximity impact has been investigated upon the infection risk prediction of emerging infectious disease, illustrated with experiments upon avian influenza A H7N9, February to May 2013 in eastern China.

The rest of this chapter is structured as follows. Section 5.2 provides a detailed description of the spatiotemporal proximity integrated approach. It can be subdivided into three subsections, which are firstly retrospective inference of historical pathogens distribution and then spatial extrapolation of pathogens distribution by weighted kernel density estimation, finally forecasting the illness onset risk for the entire considered epidemic region. In Section 5.3, experiments upon avian influenza A H7N9 fully examines validities of the previous proposed spatiotemporal proximity integrated model. Meanwhile, subsequent discussions regarding these experimental results are also extensively carried out. Concluding remarks are provided in the final section 5.4.

5.2 Spatiotemporal Proximity Integrated Approach

This section devotes to proposing a spatiotemporal proximity integrated framework for the infection risk prediction of emerging infectious disease, illustrated upon avian influenza A H7N9 for an instance. Before further proceeding with in-depth analysis of this model, we intend to introduce some reasonable assumptions so as to simplify the relatively complicated disease infection and illness onset processes.

In consideration of the expression preciseness, there are altogether 5 assumptions introduced for simplifying the relatively complicated reality regarding the disease transmission process and practical operability during the data collection stage.

Assumptions:

① Infected cases were timely collected and released once their clinical symptoms developed (including the provisional suspected cases at notification time, but later laboratory-confirmed cases). Namely, the onset-to-report interval time was excluded.

② Susceptible individuals acquired H7N9 infection exclusively by exposing to the pathogens reservoir of poultry or a live poultry market, rather than via the human-to-human transmission route.

③ All infected cases reported were independent of each other from exposure to H7N9 pathogens until the clinical symptoms developing.

④ Somewhere infections happened or not is solely determined upon the local distributed concentration of H7N9 pathogens.

⑤ All infected individuals had a limited activity territory ever since their exposure to the H7N9 pathogens. In other words, their spatial locations can be regarded to be several stationary points within a relatively large scale epidemic region.

We believe these assumptions are reasonable and acceptable with appropriate justifications below. Assumption ① applies to the data collection stage, it can be acceptable as the automated system for outbreak early detection and rapid response has sprung up (Yang et al. 2011). Assumption ② concerns the H7N9 disease transmission route, actually it comes from the reference (Chowell et al. 2013), of which on Page 3 it states “most human cases are due to spillover events originating from exposure to an animal reservoir or the environment, and human-to-human transmission is limited”. As the main transmission route is from the pathogens reservoir of poultry, assumption ③ dealing with the case independence makes sense for quantitatively characterizing the likelihood of pathogens distribution. Besides, as there is no data of transmission routes involved, assumption ④ states that the spatial pathogen concentration directly determines the possibility of H7N9 infection, which accords with our common sense. Finally, the introduction of assumption ⑤ is due to the fact that occurrence data of new emerging infectious disease (such as H7N9) generally involves only the illness onset date and location information, rather than other unknowable external environmental factors (transmission route, climatic factor), which are still not verified or needing further excavated to be relevant with the disease spreading, especially during the early outbreak stage of the new emerging infectious disease. The final assumption put forward also paves the employment of Kernel Density Estimator for investigating the spatial impact.

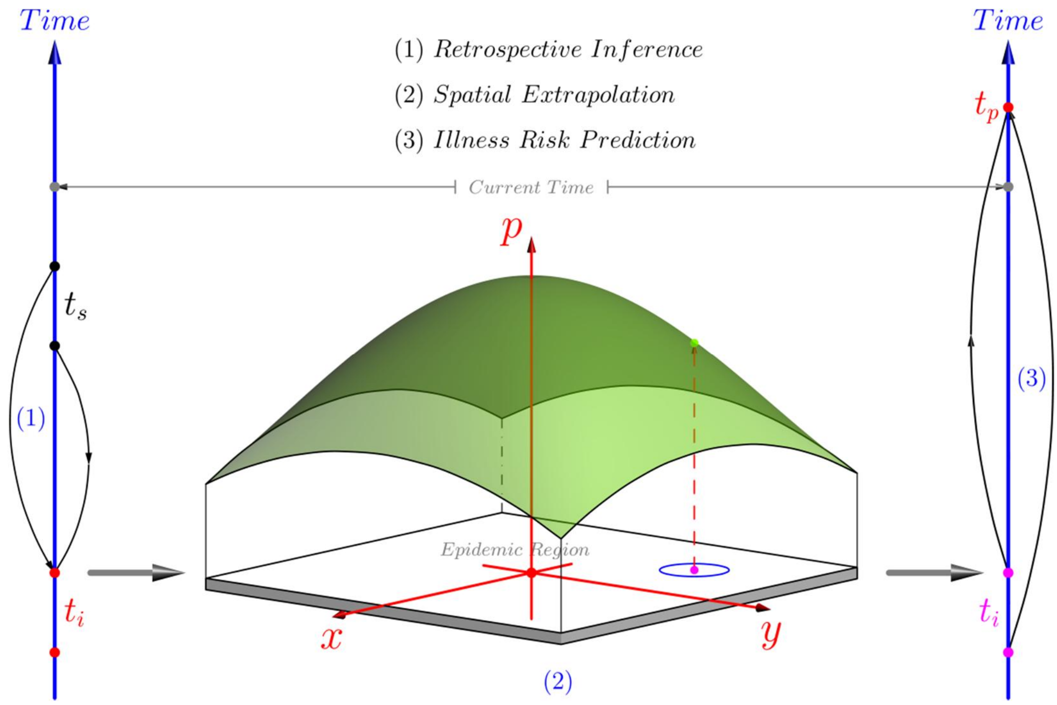

Time
(1) Retrospective Inference
(2) Spatial Extrapolation
(3) Illness Risk Prediction
Time
t_p
Current Time
p
t_s
(1)
(3)
Epidemic Region
t_i
x
y
t_i
(2)

Figure 5.1 A schematic operational framework for the spatiotemporal proximity integrated approach

This innovative approach, which operates totally based on the spatiotemporal site information of laboratory-confirmed cases, comes up with full consideration of the impact of spatiotemporal proximity upon the illness onset risk of an emerging infectious disease. There are altogether three steps: a) firstly retrospective inference of the historical existence likelihood of H7N9 pathogens from temporal dimension at each fixed notification site, b) then, spatial extrapolation of pathogens distribution by a weighted kernel density estimation on each fixed historical date, and c) finally, making prospective prediction of the illness onset risk for the entire epidemic region during some day in near future. The diagram illustrated in Figure 5.1 above provides a schematic operational framework for this spatiotemporal proximity integrated approach to the prediction risk of H7N9 infection. For better classifying the meaning of each involved symbols, Table 6.1 below summarizes the primary variables arising here, followed by the corresponding detailed descriptions.

Table 5.1 A summary of primary variables and their descriptions

Variable	Description
$P_{\mathrm{IP}}\left(t\right)$	The probabilistic density function for the incubation period distribution, with respect to the time interval t.
$P_{\mathrm{RV}}\left(s,t\right)$	The likelihood of retrospective Virus distribution of H7N9 at spatial site s, on date t.
$P_{\mathrm{PR}}\left(s,t\right)$	The predicted risk of H7N9 infection possibility at spatial site s, on date t.
$\mathrm{SIOR}_{\mathrm{PR}}\left(s,t\right)$	Standardized illness onset risk indicator for the predicted H7N9 infection risk, for fixed date t.

5.2.1 Retrospective inference of historical pathogens' existence

This section devotes to inferring the retrospective likelihood of historical existence of pathogens in terms of the probabilistic density function (PDF) of the incubation period, which is the time period elapsed from exposure to the pathogen until the emergence of clinical symptom. The probabilistic density function for characterizing the incubation period distribution can be generally expressed as $P_{\mathrm{IP}}\left(t\right)$, univariate function of the time interval from the moment of virus exposure to the symptom development date, that is independent of the spatial location of each notified H7N9 infection. Two commonly employed PDFs for characterizing the incubation period distribution are the Weibull distribution

(Cowling et al. 2013), $P\left(t\middle|\lambda,k\right)=k\lambda^{-k}t^{k-1}e^{-(t/\lambda)^k}$, including the exponential distribution to be one of its special case; and the lognormal distribution (Yu et al. 2014b), with its PDF formulated as $P\left(t\middle|\mu,\sigma\right)=\frac{1}{t\sigma\sqrt{2\pi}}e^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}$.

The core idea of retrospective inference process is given below as illustration using the avian influenza A H7N9 for easier clarification. Once an individual was exposed to the H7N9 virus, the possibility of subsequent illness onset risk can be depicted in terms of $P_{\mathrm{IP}}(t)$, where t is the elapsed time lag from the occasion of virus exposure to the moment of symptom onset. Similarly, the likelihood of historical exposure time (or exactly the existence of H7N9 virus) can also be represented as retrospective inference using $P_{\mathrm{IP}}(t)$, giving the temporal information of each infected case with clinical symptoms developed. This process coincides with the truth, as normally we are only aware of the illness onset date for most of the notified cases, but rarely getting the facts of their history exposure time or infections.

Now we concentrate on the formulation of the retrospective inference based on the foregoing discussions. Since most of the infections experienced a history exposure to the poultry or a live poultry market, rather than through human-to-human transmission, it is reasonable to assume that all susceptible individuals were infected after contact with H7N9 pathogens independently. That is to say, the clinical symptoms of infected cases would emerge separately at their own pace one after another. As a result, at a fixed spatial site where human infection with H7N9 virus was precisely reported, the potential likelihood of retrospective existence of H7N9 pathogens on date t_i implies the possibility of at least one susceptible individual infected on date t_i and later starting clinical symptoms on date t_s. Therefore, the existence probability of H7N9 virus on date t_i can be estimated by retrospective inference as follows,

$$P_{\mathrm{RV}}\left(s,t_i\right)=1-\prod_{t_s>t_i}\left(1-P_{\mathrm{IP}}\left(t_s-t_i\right)\right)^{n_{t_s}} , \tag{5.1}$$

where n_{t_s} denotes the number of notified cases on date t_s. The likelihood of P_{RV} can also be perceived to be the reservoir's concentration of H7N9 pathogens.

5.2.2 Spatial extrapolation of pathogens distribution

This section devotes to the spatial extrapolation of H7N9 pathogens distribution on any fixed historical date, by some spatial smoothing techniques based on the existence likelihood of dispersedly distributed pathogens estimated using retrospective inference in the previous section.

Considering the First Law of Geography introduced by Tobler (1970), suggesting that “everything is related to everything else, but that near things are more related than distant things”, thus the spatial proximity effects upon the historical distribution of H7N9 pathogens can be handled employing the weighted kernel density estimation (WKDE) method.

The WKDE method derived from KDE method, can be briefly formulated below.

For bivariate independent and identically distributed samples $s_1, s_2, \cdots, s_n$ drawn from some distribution with an unknown density function f, its kernel density estimator is

$$\hat{f}\left(s;\Sigma\right) = n^{-1}\sum_{i=1}^{n} K_{\Sigma}\left(s - s_i\right), \tag{5.2}$$

where $s = \left(s_1, s_2\right)^{\mathrm{T}}$ and $s_i = \left(s_{i1}, s_{i2}\right)^{\mathrm{T}}$, $i = 1, 2, \cdots, n$, the symbol Σ denotes the bandwidth matrix. Here the Gaussian kernel (De Smith et al. 2007) is adopted,

$$K_{\Sigma}\left(s\right) = \exp\left(-\tfrac{1}{2} s^{\mathrm{T}} \Sigma^{-1} s\right). \tag{5.3}$$

The bandwidth matrix can generally be determined by plug-in method or cross-validation method (Duong and Hazelton 2005). However, the Scott's rule of thumb (Scott 1979) provides a simple and alternative way of designating the bandwidth matrix to be an matrix proportional to the sample covariance matrix $\hat{\Sigma}$. Here we pick out the bandwidth matrix $\Sigma = n^{-\frac{1}{3}}\hat{\Sigma}$ (Ahamada et al. 2010). The WKDE can then be straightforwardly derived by appending additional weights ω_i, $i = 1, 2, \cdots, n$, yielding

$$\hat{f}\left(s;\Sigma\right) = \frac{1}{n}\sum_{i=1}^{n} \omega_i \exp\left(-\tfrac{1}{2}\left(s - s_i\right)^{\mathrm{T}} \Sigma^{-1}\left(s - s_i\right)\right). \tag{5.4}$$

This estimator can also be perceived as the weighted average of all Gaussian probability density surfaces centering at each sample position. Accordingly, for any fixed historical date t_i, the spatial extrapolation of existence likelihood of the historically distributed H7N9 pathogens can thus be formulated using the above-mentioned WKDE as follows,

$$\begin{aligned} P_{\mathrm{RV}}\left(s,t_i;\Sigma\right) &= n_{t_i}^{-1}\sum_{j=1}^{n_{t_i}} P_{\mathrm{RV}}\left(s_j,t_i\right)K_{\Sigma}\left(s-s_j\right) \\ &= n_{t_i}^{-1}\sum_{j=1}^{n_{t_i}} P_{\mathrm{RV}}\left(s_j,t_i\right)\exp\left(-\tfrac{1}{2}\left(s-s_j\right)^{\mathrm{T}}\Sigma^{-1}\left(s-s_j\right)\right), \end{aligned} \tag{5.5}$$

where $P_{\mathrm{RV}}\left(s_j,t_i\right)$ is the existence likelihood of historical H7N9 pathogens (or alternatively be regarded as the virus concentration) distribution by retrospective inference at spatial location s_j, on history date t_i, serving as the weighting coefficients of WKDE for extrapolating the potential distribution of H7N9 pathogens in space.

5.2.3 Infection risk prediction

The foregoing spatial extrapolation by WKDE produces the entire spatial distribution of H7N9 pathogen's reservoir on a historical date t_i. At each fixed spatial site s within the epidemic region, the predicted possibility of illness onset risk on some date t_p in near future, that is the likelihood of at least one infected individual developing clinical symptoms on t_p who acquired H7N9 infection from all the days t_i $\left(t_i < t_p\right)$ in the past, can similarly be formulated as,

$$P_{\mathrm{PR}}\left(s,t_p\right)=1-\prod_{t_i<t_p}\left(1-P_{\mathrm{RV}}\left(s,t_i\right)P_{\mathrm{IP}}\left(t_p-t_i\right)\right). \tag{5.6}$$

Among above equation, $P_{\mathrm{RV}}\left(s,t_i\right)$ indicates the existence likelihood of historical H7N9 pathogens by retrospective inference, and the quantity $P_{\mathrm{IP}}\left(t_p-t_i\right)$

measures the occurrence likelihood of time interval between the historical virus exposure date t_i and the future illness onset date t_p in terms of the PDF of incubation period.

5.3 Experiments & Discussions

This section intends to verify the innovative spatiotemporal proximity integrated model via experiments upon avian influenza A H7N9, February to May 2013, emerged in eastern China. Discussions and comments upon the experimental outputs are fully explored and provided subsequently.

5.3.1 Parameter Configuration

An epidemiology study conducted by Cowling et al. (2013), demonstrated that the Weibull models best fit the incubation period for H7N9 infection process, with an estimated mean incubation period of 3.1 days and the corresponding standard deviation of 1.4 days. Therefore, we employ the Weibull distribution to depict the incubation period of H7N9 occurrence cases,

$$f\left(t;\lambda,k\right)=\begin{cases} k\lambda^{-k}t^{k-1}e^{-(t/\lambda)^k} & t\geq 0, \\ 0 & t<0. \end{cases} \tag{5.7}$$

For a given Weibull random variable, T, its mean and variance can be formulated as the function of parameters λ and k, which can be calculated by some numerical iterative method (for example, the Newton-type method) as follows,

$$\begin{cases} E(T) = \lambda \Gamma\left(1 + \frac{1}{k}\right) \\ D(T) = \lambda^2 \left(\Gamma\left(1+\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2 \right) \end{cases} \Rightarrow \begin{cases} \lambda = 3.4981 \\ k = 2.3539 \end{cases}. \tag{5.8}$$

Besides, bandwidth matrix of the Gaussian kernel is designated following an empirical formula (Ahamada et al. 2010), in terms of the isotropic covariance of all these 135 spatially distributed infection sites,

$$\Sigma = n^{-\frac{1}{3}} \hat{\Sigma} = \begin{pmatrix} 0.7888 & -0.0846 \\ -0.0846 & 1.0985 \end{pmatrix} \times 10^{10} \,(\mathrm{m}^2). \tag{5.9}$$

5.3.2 Historical pathogens distribution by retrospective inference

Experiments based on avian influenza A H7N9 were implemented on the MATLAB platform. Given a series of confirmed cases, the historically spatiotemporal distribution of H7N9 pathogens' concentration can then be retrospectively inferred according to the first two procedures. Firstly, the existence likelihoods of historical pathogens are estimated at the sites of dispersedly distributed notifications via retrospective inference according to equation (5.1), which can be calculated complying with the recursive Algorithm 5.1 by programming.

Then, the spatially distributed existence likelihood of historical H7N9 pathogens can then be extrapolated by WKDE method according to equation (5.5) within the entire epidemic region.

Algorithm 5.1 Recursive iteration for estimating the pathogens' existence

```
$P_{\mathrm{RV}}\left(\cdot, t_i\right) = 0$;
For $t_s > t_i$,
    $P_{\mathrm{RV}}\left(\cdot, t_i\right) = 1 - \left(1 - P_{\mathrm{RV}}\left(\cdot, t_i\right)\right)\left(1 - P_{\mathrm{IP}}\left(t_s - t_i\right)\right)^{n_{t_s}}$.
End
```

Other than the formerly rendered risk map by Zhu and Peterson (2014) via a calibrated niche model to account for the primitive spatiotemporal information from the notified infections, these animated layers delineate the historical pathogens' distribution (existence likelihoods), which are normally one incubation period on average earlier before the illness onset date. Thus, it makes such retrospectively inferred risk maps be more conducive to deeply exploring the possible transmission patterns of H7N9 infection course, as well as inspecting the relationship between environmental risk factors (such as poultry markets and bird migration routes) and H7N9 incidence in the follow-up studies.

A GIF image (H7N9_History.GIF) has been made hereby serving as the supplementary material to demonstrate the existence likelihoods of historical H7N9 pathogens during covering the period from 3th February to 21th May, 2013, totally based on the collected H7N9 occurrence cases. It is revealed that the forepassed H7N9 pathogens were firstly prevalent in the Shanghai municipality in February, then spreading spatially around its adjacent geographic provinces including Jiangsu, Anhui, and Zhejiang, respectively, in March. Afterwards, the

H7N9 epidemic areas further expanded towards the central and southeast regions, including Henan, Jiangxi and Fujian provinces in late April, and diminishing thereafter. These pathogens also reemerged intermittently within the historical epidemic areas because periodic temperature changes due to seasonal variations and migratory birds inhabiting south in winter and flying northwards when spring comes. Such a pattern of spatiotemporal evolution of H7N9 pathogens coincides roughly with temperature variation and bird migration routes, as demonstrated by Zhang et al. (2014) that there were obvious correlation between H7N9 epidemic and environmental risk factors, such as the high-risk temperature range (9°C–19°C) occurring and bird migration coverage.

5.3.3 Forecast risk map

After the derivation of historical distribution of H7N9 pathogens via retrospective inference and spatial extrapolation, prospective forecasts of illness onset risk during some period in near future can then be made at each site within the entire epidemic region using equation (5.6). The onset possibility can be calculated with the following recursive Algorithm 5.2 by programming.

Algorithm 5.2 Recursive iteration for predicting H7N9 illness onset risk

```
$P_{\mathrm{PR}}\left(\cdot,t_p\right)=0$;
For $t_i<t_p$,
    $P_{\mathrm{PR}}\left(\cdot,t_p\right)=1-\left(1-P_{\mathrm{PR}}\left(\cdot,t_p\right)\right)\left(1-P_{\mathrm{RV}}\left(\cdot,t_i\right)P_{\mathrm{IP}}\left(t_p-t_i\right)\right)$.
End
```

There were altogether 45 unique dates in the original H7N9 database. Risk maps of illness onset can thus be generated according to the proposed approach by employing spatiotemporal information of the beginning few occurrence cases before each of the 45 different dates. All the other subsequent H7N9 occurrences later than that forecasting date can then be used to verify these forecast results of this spatiotemporal proximity integrated model.

It should be noted that exponential kernel function involved in WKDE, and the integral upon small area of both tail sides of incubation period' PDF may give rise to extremely tiny possibility values regarding the forecast risk of illness onset. In view of this, we standardize the predicted risk $P_{\mathrm{PR}}\left(s,t_p\right)$ in space, by introducing an additional SRIO (Standardized Risk of Illness Onset) indicator (falling behind 0 and 1) so as to highlight the severe areas of considered epidemic regions. That is,

$$\mathrm{SRIO}_{\mathrm{PR}}\left(s,t_p\right)=\frac{P_{\mathrm{PR}}\left(s,t_p\right)}{\max\limits_{s} P_{\mathrm{PR}}\left(s,t_p\right)}. \tag{5.10}$$

Figure 5.2 provides box plots of the predicted SRIOs at checkpoints of subsequent H7N9 occurrences with respect to different forecast time intervals from 1 to 14 days (the maximum prediction time span, beyond which the forecast results may be unreliable). From this figure, it can be seen that this innovative approach is capable of providing approximately 70% correct prediction on average in terms of the H7N9 illness onset risk during the future 5 days from the forecast date. However, correct predictions become less than 50% and getting even progressively smaller with increasing time span beyond 5 days. On the whole, a roughly negative correlation can be overserved between predicted illness onset risks and the corresponding time intervals, which exactly reflects the temporal proximity impact upon the predicted H7N9 illness risk gradually decreasing along with the increasingly broadening time spans.

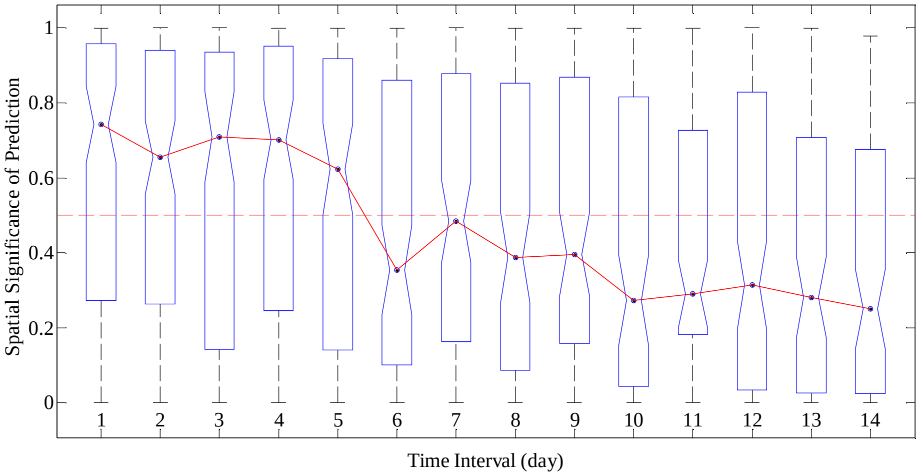


Figure 5.2 Box plots of SRIOs with respect to different forecast time intervals

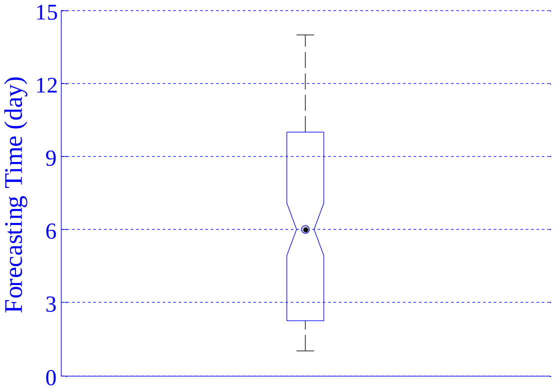
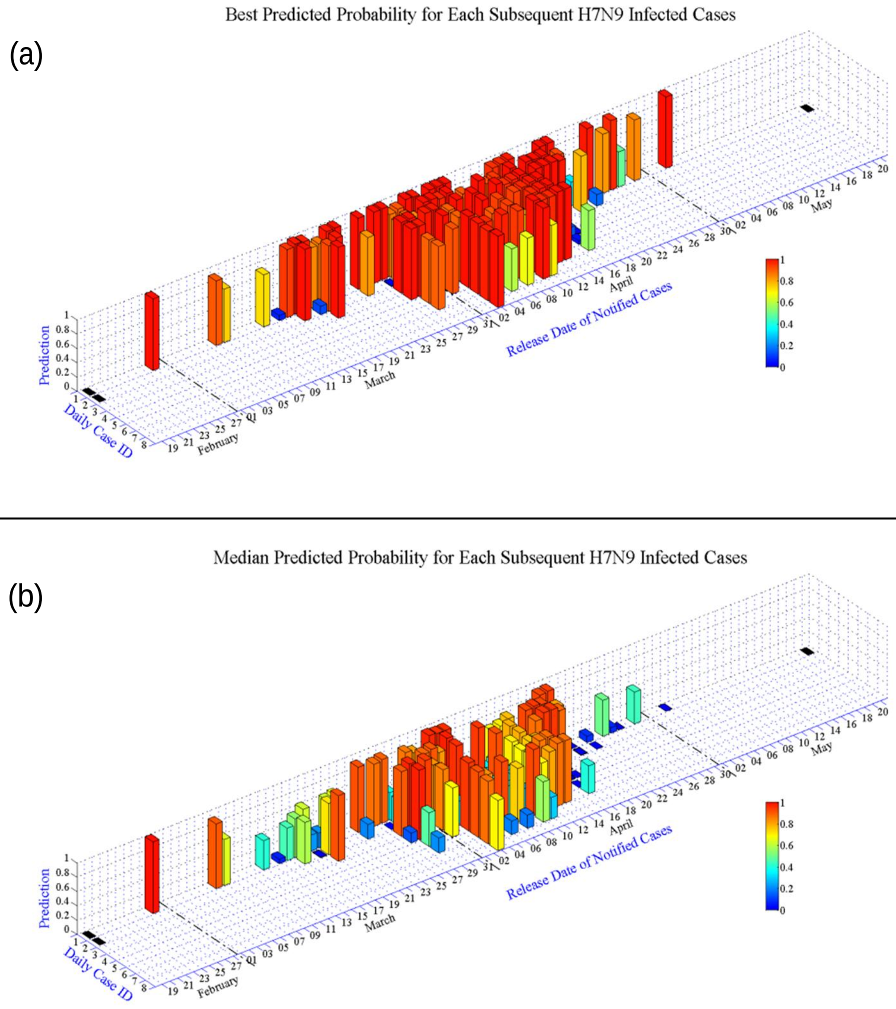


Figure 5.3 Box plot for the forecast time spans

The box plot for all the summarized forecasting time spans presented in Figure 5.3, shows that median of the predicted time intervals is about 6 days. As current data for avian influenza A (H7N9) infection indicate an incubation period ranging between 2 and 8 days, with an average of five days (Gao et al. 2013a) or even shorter, 3.1 days stated by Cowling et al. (2013), thus the maximum forecasting time span is thus best anticipated not more than twice the average incubation period, which is exactly reflected by these boxplots of forecasting time spans. As thus, by further consideration of Figure 5.2, the effective forecast time length can be asserted of 5 days for this spatiotemporal proximity integrated approach.

Besides, each of the subsequent occurrences can be used to verify multiple forecast results made by the first several cases notified earlier before the current predictive date. Figure 5.4 provides three-dimensional bar graphs reflecting these multiple “prediction-verification” chains in terms of the maximum and median furcating illness onset risk, together with the predictive time spans corresponding to the maximum (best) forecasting results in the day of each subsequent H7N9 occurrence. For each subfigure, the abscissa represents the issued date of notified cases, of which at most 8 cases (occurring on Apr, 3, 2013) simultaneously appear in a day from our collective data set. All notified cases coming up in one day were assigned a daily case ID number in turn according to the original recorded orders. Inapplicable cases (on the first-day or the forecast failure cases) were indicated as black blocks. With the best forecasting results (subgraph a in Figure 5.4), the spatiotemporal proximity integrated model seemingly provide mostly 90% or more accurate prediction for the future H7N9 illness onset risks. Median values (subgraph b in Figure 5.4) of the forecast risks performed barely satisfactorily, with only 50% or so verification accuracy. The bottom sugraph c in Figure 5.4 gives the predictive time spans corresponding to the best forecasts as illustrated in the top sugraph a. Actually, Figures 5.2-5.3 have already conveyed a general impression of the forecast time spans, while sugraph c presents more intuitive visualization of these time intervals from forecast to verification upon each H7N9 occurrence in chronological order.



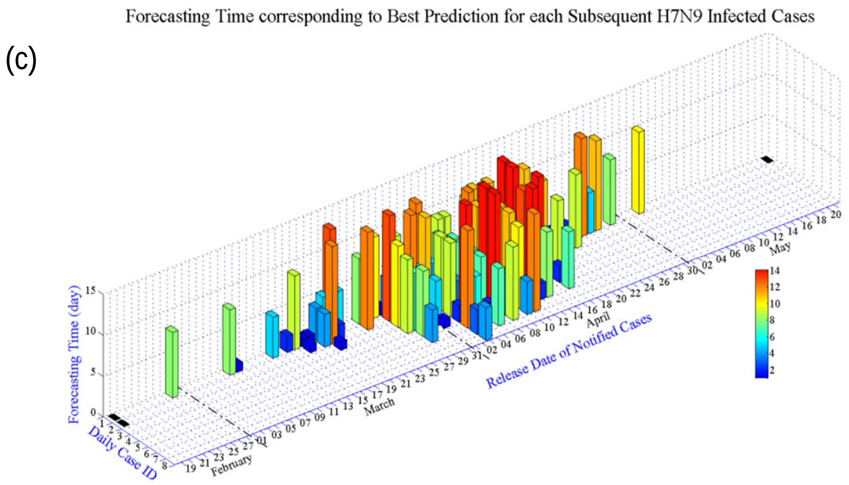


Figure 5.4 3D bar graphs of the multiple “prediction-verification” chains generated by the spatiotemporal proximity integrated model for each H7N9 infected cases in chronological order. (a) maximum (best) forecasted illness onset risks, (b) median value of the forecasted illness onset risks and (c) the predictive time spans corresponding to the best forecasts

As indicated above, H7N9 occurrences nearer the forecast dates exert more impact upon subsequent cases regarding the risk of illness onset. Thus, we further provide an animated image (H7N9_Predict.GIF) consisting of a continuous single-step “prediction-verification” chains for investigating the spatiotemporal proximity impact upon the forecasting performance along with a chronological order of the collected H7N9 occurrences. For each ring of the “prediction-verification” chains, the forecast was firstly made employing the information of all H7N9 occurrences before one H7N9 occurrence date. And then the verification was conducted immediately against subsequent H7N9 infections during one succeeding date. Moreover, every forecast risk map was also overlapped with its subsequent occurrences for visual comparison. As an example, Figure 5.5 below illustrates the forecast H7N9 illness risk map on 20 March, based on all earlier occurrences before 19 March, using the spatiotemporal proximity integrated approach. It is expected that these forecast risk maps can be served for depicting the tread of real-time epidemic evolution; especially in the under-development areas where statistics of human H7N9 infected cases are always overdue or incomplete; thereby guiding the development of more effective intervention measures and the optimized allocation of limited medical aid resources (vaccines) to potential targets.

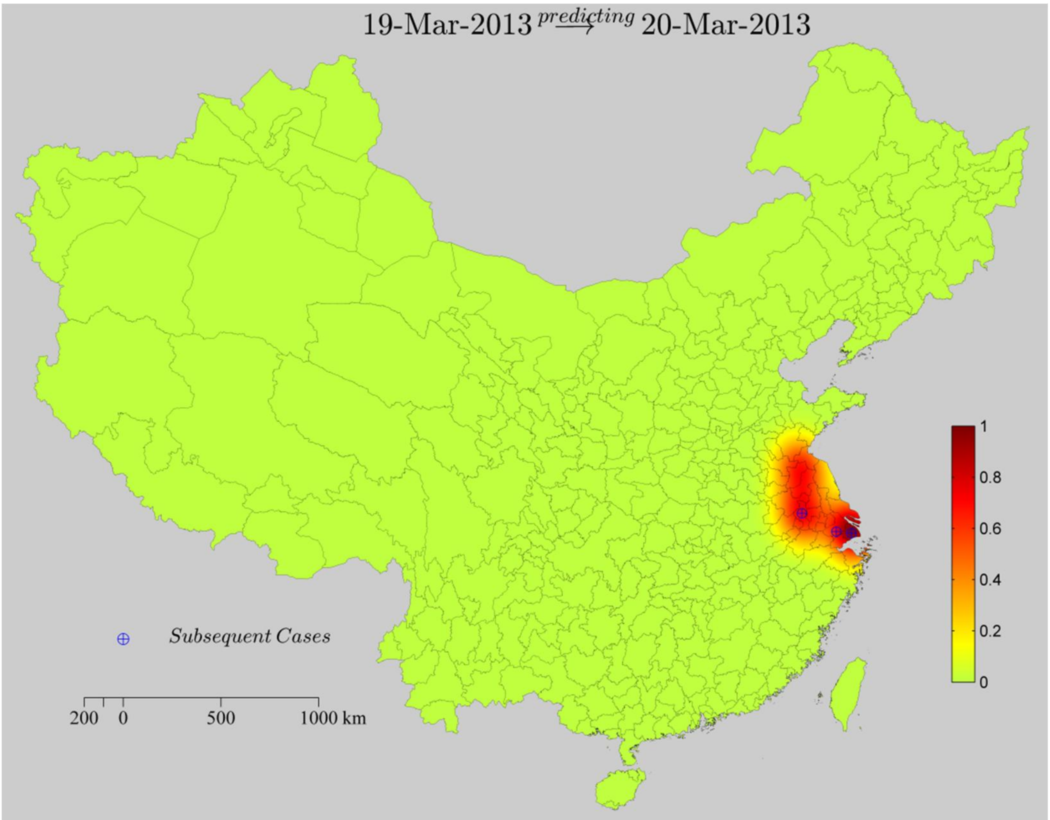


Figure 5.5 One instance of “prediction-verification” chain link regarding the H7N9 illness onset risk

5.4 Conclusions

This chapter presents a spatiotemporal proximity integrated approach (under appropriate assumptions) for predicting the infection risk of an infectious disease. This model is motivated by the principle that clinical symptoms appear certainly posterior to the historical virus exposure time, which can be described by PDF of the incubation period fitted using retrospective cases. There are altogether three procedures involved in the proposed model: a) estimating historical existence likelihoods of infectious pathogens via retrospective inference, b) undertaking spatial extrapolation of pathogens distribution by weighted kernel density estimations, and c) forecasting risk maps of illness onset. This approach also operates based entirely on the spatiotemporal information of laboratory-confirmed cases.

Validation experiments were implemented using a combined data set of avian influenza A H7N9 in eastern China comprising of 135 records, during the period between 19th February and 21th May, 2013. A preliminary outcome is the virus distribution map via the retrospective inference, which depicts the historical distribution of H7N9 infectious pathogens, thereby making it an axillary tool of value for further excavating possible dissemination pattern of the H7N9 contagions, and identifying potentially relevant environmental factors.

Besides, experiments based on avian influenza A H7N9 have demonstrated that the spatiotemporal proximity integrated approach is capable of providing an approximately 70% correct prediction on average in terms of the H7N9 illness onset risk during the near future 5 days ever since the forecasting date. Higher prediction accuracy of this straightforward model can be anticipated if further environmental factors were taken into account, such as road networks, distribution of poultry markets and human population density, which may more accurately account for the dissemination patterns of H7N9 dynamics. Nevertheless, these forecast risk maps of illness onset computed by the spatiotemporal proximity integrated model can function as early-warnings to identify areas where proactive surveillance efforts and preventive intervention measures should be targeted against further propagation of human H7N9 infections.

Tobler's First Law of Geography acknowledges the spatiotemporal proximity impacts are widely observed in both manmade and natural worlds. The innovative idea presented by the proposed approach can further serve proximity investigation of many other real phenomena, such as the praxeology analysis (criminology), atmospheric pollution and natural disasters (landslide, debris flow), etc. Although effectiveness of the approach has only been verified by employing the database of human H7N9 infections that occurred in eastern China, in 2003, it is felt that this approach can be readily applied to other infectious diseases that exhibit apparent spatiotemporal aggregation patterns; for instance, the foot and mouth disease (Wang et al. 2011, Wang et al. 2013). It is worth mentioning that spatiotemporal proximity are also observed in their propagating processes for some typical human-to-human transmitted infectious diseases, such as the SARS outbreak and

H1N1 pandemic. However, the conventional geographic distance may need to be substituted with a probabilistically motivated effective transmission distance (Brockmann and Helbing 2013a) so as to take the rapid host mobility and expanded scope of human's activities into considerations.

When an emerging infectious disease outbreak, occurrence cases generally afford only the illness onset date and spatial location information. Other external environmental factors (such as transmission route, climatic factor), are normally still unknowable, and needs to be further excavated whether relevant with the disease transmission or not along with the epidemic evolution. In this context, this proposed approach comes out operating entirely upon the spatiotemporal information of the laboratory-confirmed infections. It is expected to forecast the spatiotemporal illness onset risk, especially during the early stages of an emerging infectious disease. We are confident of that such model can be served to provide valuable scientific support for policy constitutors of public health to formulate more effective prevention and control measures.

Chapter 6 Bayesian Inference of the Reproduction Number

This chapter employs a Bayesian scheme for estimating the time-varying effective reproduction numbers so as to understand the real-time transmission potential of the Ebola epidemic situation in West Africa.

6.1 Stochastic SEIR model

To characterize the evolutionary dynamic process for the West African Ebola epidemic over time, we utilize a modified SEIR-type compartmental model (Anderson and May 1991) which classifies the time-varying individuals as susceptible, exposed, infectious, and removed, with the assumption of the population being homogeneously well-mixed. This nonlinear differential equation system can be formulated as follows,

$$\begin{cases} \dot{S}(t) = -\beta S(t) I(t) / N(t), \\ \dot{E}(t) = \beta S(t) I(t) / N(t) - \alpha E(t), \\ \dot{I}(t) = \alpha E(t) - \gamma I(t), \\ \dot{R}(t) = \gamma I(t), \\ \dot{C}(t) = \alpha E(t), \end{cases} \tag{6.1}$$

where the dot denotes time derivatives, and $C(t)$ is the cumulative case number counting all the infections. The parameters $\Theta = (\beta, \alpha, \gamma)$ are related to the transition rates from one disease stage to the next. Susceptible individuals enter the exposed compartment at the rate of $\beta I(t)/N(t)$ after contact with the virus, where β represents the pathogen transmission capacity from infectious to susceptible individuals per unit time, and $N(t)=S(t)+E(t)+I(t)+R(t)$ is the total population at time t. It is assumed that the susceptible occupies almost the entire population at the early stage of a disease outbreak. The "exposed but not yet infectious" individuals (E) enter into the infectious class at the rate of α per unit time, while γ is the diminishing rate (per unit time) of infectious individuals I due to recovery or death. In epidemiological terminology, parameters α and γ correspond to the inverse of the average of an exponentially distributed time to onset of infectiousness and to recovery since infection, respectively. Namely, $1/\alpha$ and $1/\gamma$ are the mean incubation and infectious period (duration of the infection). Besides, the demographic effects are ignored here in consideration of rapid spreading of the Ebola epidemic and the slow population growth during the interim.

6.2 Reproduction number

The basic reproduction number, which is one of key concepts in epidemiology, originates from the compartmental model (such as SIR, SEIR), where a population is assumed nearly all susceptible at the initial stage and well-mixed during the whole epidemic process. This concept is defined as the average number of secondary infections arising from a primary case during the course of its infectious period, with the calculation formula as

$$R_0 = \frac{\beta}{\gamma}. \tag{6.2}$$

It is a crucial quantity for identifying the hazard level of an infectious disease, and guiding the requisite intervention intensity to be adopted for preventing the further spread of an epidemic.

However, humans may have high mobility in reality, be isolated or confer lifelong immunity after some pathogens infected. Thus, the effective reproduction number, R_t, is timely put forward, more suitable for estimating the average number of secondary cases per infectious case. The subsequent section 6.4 provides the calculation formula for effective reproduction number. Unquestionably, the estimation of R_t indicator has proved to be of critical importance to gauge the risk level for an infectious disease; for instance, understanding the outbreak and potential danger from SARS (Riley et al. 2003) or the H1N1 (Chowell et al. 2007).

6.3 Approximation of the early phase of exponential-growth

The basic reproduction number can typically be estimated during the early stage of an epidemic, as there would not be interventions or evidence of depletion effects of the susceptible compartment at this stage. Thus, it is often assumed that the initial growth rate of the cumulative infected case number behaves exponentially. Based on the same considerations, Lipsitch et al. (2003) has estimated the basic reproduction number R_0 of the SARS outbreak in Singapore before control measures were instituted, by fitting the cumulative case number in a logarithmic scale with a straight line $b_0 + kt$. The optimal exponential phase scope can be determined by the R square or chi-square goodness of fit test (Favier et al. 2006). The Chi-Square Goodness of Fit Test is put forward to examine how "close" the observed values would be reflected by a fitted model. In general, the chi-square test statistic is of the form $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$, where χ^2 is the obtained Chi Square, O_i and E_i are observed and expected scores respectively. A large computed test statistic indicates the observed and expected values are not close and the employed model is a poor fit of the observations. Alternatively, the R square index characterizes the goodness of fit between the observations and the fitting function by comparing the variability of the estimation errors with the variability of the original values, with the formulation being

$$R^2 = 1 - \frac{SS_E}{SS_T}, \tag{6.3}$$

where $SS_E = \sum_i \left(O_i - E_i\right)^2$ is the sum of squared errors, and $SS_T = \sum_i \left(O_i - \bar{O}\right)^2$ is the total sum of squares. The following Figures 6.1-6.4 provide the optimal exponential phase scopes in terms of the R square index for each of the Ebola affected countries in West Africa.

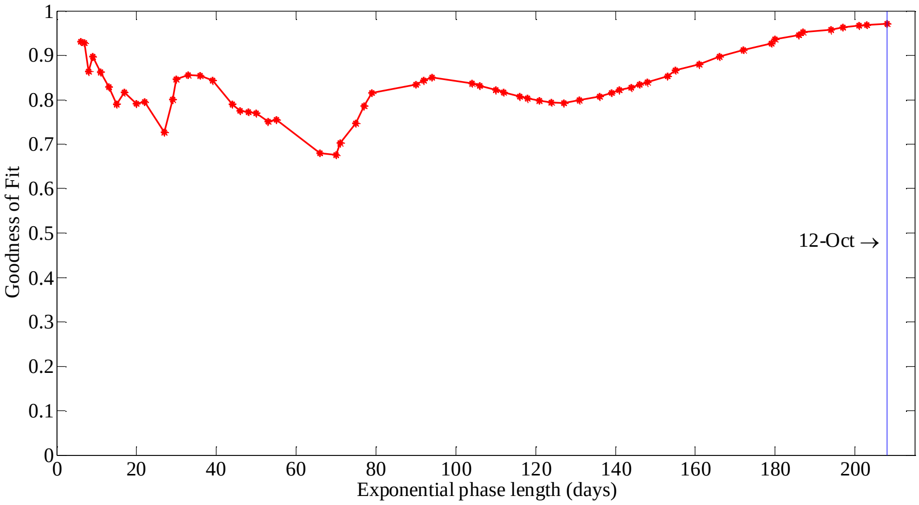


Figure 6.1 The R square index to the exponential growth phase of cumulative EVD case number for Guinea

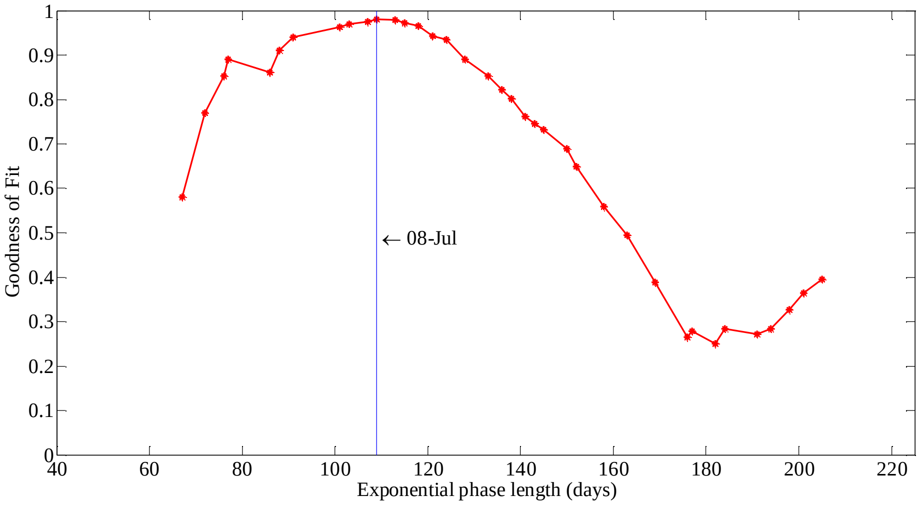


Figure 6.2 The R square index to the exponential growth phase of cumulative EVD case number for Sierra Leone

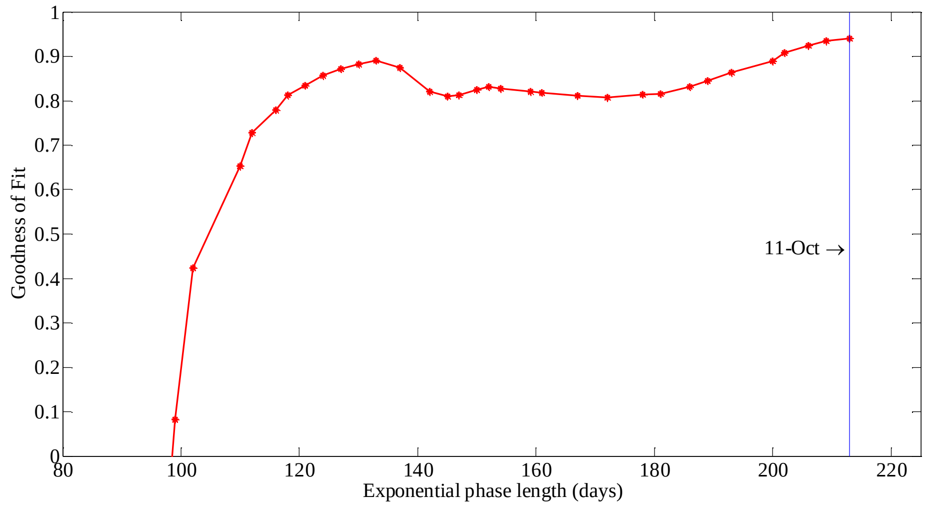


Figure 6.3 The R square index to the exponential growth phase of cumulative EVD case number for Liberia

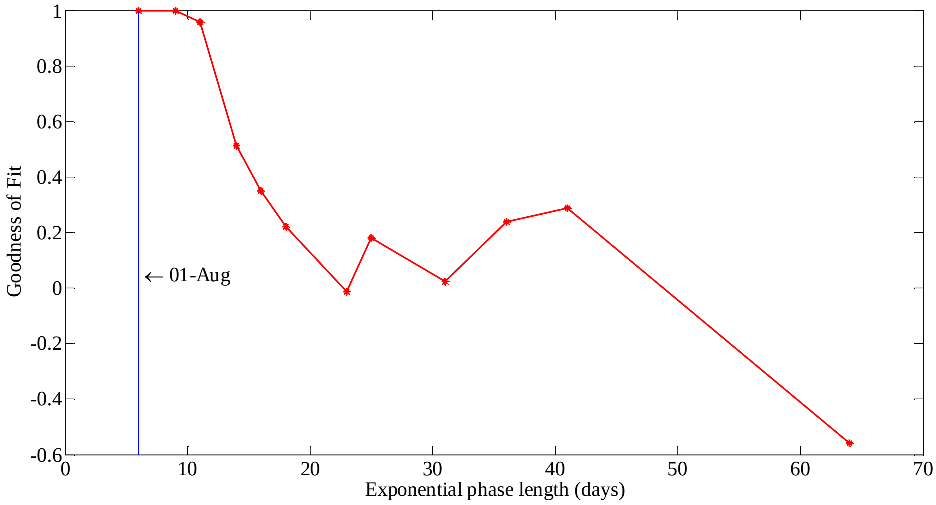


Figure 6.4 The R square index to the exponential growth phase of cumulative EVD case number for Nigeria

After curve fitting of the cumulative EVD case number, then the basic reproduction number R_0 can be approximately computed by substituting the estimated straight slope k with the spectral radius (dominant eigenvalue) of a linearized SEIR model (6.1) assuming no depletion of the susceptible compartment. That is achieved by setting $S(t) = N(t)$ in equation (6.1), yielding

$$\begin{pmatrix} \dot{E}(t) \\ \dot{I}(t) \end{pmatrix} = \begin{pmatrix} -\alpha & \beta \\ \alpha & -\gamma \end{pmatrix} \begin{pmatrix} E(t) \\ I(t) \end{pmatrix}, \tag{6.4}$$

with a Corresponding dominant eigenvalue $\lambda_+ = \frac{1}{2}\left(-(\alpha+\gamma) + \sqrt{(\alpha-\gamma)^2 + 4\alpha\beta}\right) = k$.

Recall of formula (6.2), we obtain the estimated basic reproduction number as

$$\tilde{R}_0 = \frac{1}{\alpha\gamma}(k+\alpha)(k+\gamma). \tag{6.5}$$

It should be noted that the above calculation formula has also be equivalently derived before involving extra quantities like the mean infectious period and serial interval as mentioned in other studies (Heffernan et al. 2005, Chowell et al. 2007).

6.4 Sequential Bayesian Inference of R_t

A communicable disease requires prompt adoption of emergency responses such as quarantine, or deployment of vaccine resources. Precise and prompt estimates of the reproduction number are of critical importance since this quantity can frequently be employed for not only indicating the critical level of the disease transmission over time, but also assessing the efficacy of the adopted countermeasures. This section describes a Sequential Bayesian inference combined with a stochastic SEIR model (Chowell et al. 2007) for estimating the real time effective reproduction number R_t, the actual number in average of secondary infected cases per primary case at time t (Nishiura et al. 2006) and is normally smaller than the basic reproduction number R_0. It may vary over time on account of changes in the demographic structure, as well as the virus evolution.

Epidemiological reports of Ebola that were updated intermittently on the WHO official website provided a tally of infected cases in chronological order. Thus, it is convenient to proceed with the estimation procedure through progressive increment in the cumulative number of notified cases, $C(t)$, as described in equation (6.1). The newly reported cases over the period $t \to t+\tau$ can be written as $\Delta C(t) = C(t) - C(t-\tau)$, where time interval τ between successive dates of epidemic information release may vary over time from 1 day to almost 1 week, or even longer. Suppose that the susceptible population portion remains approximately constant over each incremental period τ, namely, $\beta_t = \beta S(t)/N(t) = c_t\beta$ but may vary across two successive periods, then simultaneous differential equations associated with $E(t)$ and $I(t)$ can be reformulated by simplification from the linearization of the SEIR model (6.1) as follows,

$$\begin{pmatrix} \dot{E}(t) \\ \dot{I}(t) \end{pmatrix} = \begin{pmatrix} -\alpha & \beta_t \\ \alpha & -\gamma \end{pmatrix} \begin{pmatrix} E(t) \\ I(t) \end{pmatrix}. \tag{6.6}$$

Following the approach adopted in the literature (Lipsitch et al. 2003) and using formula (6.2)'s variant $R_t = \beta_t/\gamma$, we can figure out the dominant eigenvalue

$$\lambda_+ = \frac{1}{2}\left(-(\alpha+\gamma)+\sqrt{(\alpha-\gamma)^2+4\alpha\gamma R_t}\right),$$

thereby arriving at the approximate solution of $E(t)$ to equation (6.6), that is the relationship of $E(t+\tau)=b(R_t)E(t)$, or equivalently,

$$\Delta C(t+\tau)=b(R_t)\Delta C(t), \text{ where } b(R_t)=\exp(\lambda_+\tau). \qquad (6.7)$$

Parameters α and γ are constant for this model, thus the time-varying effective reproduction number R_t is the sole parameter to be estimated.

In general, core idea of the sequential Bayesian inference approach is to predict the distribution of future case number increment $\Delta C(t+\tau)$ based on currently new issued case number $\Delta C(t)$ after giving R_t (and other parameters such as α and γ). Fortunately, our familiar Poisson distribution can be utilized here to characterize such dynamic renewal process as it indicates probability of a given number of events occurring in a fixed time interval, with only one parameter λ to be estimated, which is exactly its expectation. Recall equation (6.7), future case number can thus be predicted as

$$P\left(\Delta C(t+\tau)\leftarrow\Delta C(t)\middle|R_t\right)=Pois(\lambda),\ \ \lambda=b(R_t)\Delta C(t). \qquad (6.8)$$

This probabilistic formulation implies that future increased case number $\Delta C(t+\tau)$ rests with currently new released case number $\Delta C(t)$, given R_t. Therefore, estimation of R_t with quantified uncertainty can be straightforwardly formulated as posterior distribution via the Bayesian rule, that is

$$P\left(R_t \middle| \Delta C(t+\tau) \leftarrow \Delta C(t)\right) = \frac{P\left(\Delta C(t+\tau) \leftarrow \Delta C(t) \middle| R_t\right) P(R_t)}{P\left(\Delta C(t+\tau) \leftarrow \Delta C(t)\right)}, \tag{6.9}$$

where $P(R_t)$ is a prior, reflecting any desirable guessing knowledge of R_t; and the denominator, independent of R_t, is a normalization factor, which is not necessarily in the subsequent simulated computation if utilizing the MCMC (Markov Chain Monte Carlo) iteration approach. Another advantage for employing the simulated MCMC approach is that the traditional analytical-likelihood-based inferences are generally computationally intractable.

Overall, the knowledge of two or more newly notified cases incorporated into this probabilistic contagion model, comes into the posterior estimation of the probability distribution of R_t, via the Bayesian theorem. A scheme of the estimation algorithm proceeds through continuous iterations, where the posterior probability estimation of R_t in the foregoing step is chosen to be the prior for new cases notified within the subsequent time interval, $t+\tau$. After successive iterations upon a series of incremental number of released cases, the maximum likelihood estimates (average or median) can be perceived as the optimum estimator for R_t, and the desired confidence intervals corresponding to different confidence levels. After a series of Bayesian recursive iterations, the effective reproduction number, R_t, can be elaborately estimated by excavating as much information as possible from the successive reported cases. However, the estimated R_t may have the declining trend due to depletion of the susceptible portion within the total population.

It is noteworthy that when applying the above SEIR model for a sporadically distributed population, the homogeneous mixing assumption may not be satisfied. Further, the evolutionary epidemic spreading may lead to a gradual depletion of susceptible individuals, thereby cutting down the estimated effective reproduction number. In the real-world situation, whether taking the susceptible depletion into consideration is also subject to the geographical scale of an epidemic region. Image that the EVD may infect all humans in a village but hardly affect the entire citizens of a big city. Therefore, if we take the whole country into account as an epidemic pool for the SEIR model, the susceptible portion can thus be regarded to be always constant during the whole Ebola spreading course.

6.5 Experiments and Discussion

For here the discussed sequential Bayesian inference model, an appropriate early time stage should be firstly picked out corresponding to the exponential epidemic growth with enough care. Because the approximate solution of $E(t)$ to the linearization equation of SEIR model (6.4) may directly influence the priori distribution of basic reproduction number. Normally, a prior of this sequential Bayesian iterative model can be assigned with a trimmed Gaussian $N\left(\mu_R = \tilde{R}_0,\ \sigma_R^2 = 1\right)$ restricted within the interval of $[0,\ 2\mu_R]$ for the estimation of R_t. Meanwhile, the latent and infectious periods are referred with $\alpha = 1/5.3$ and $\gamma = 1/5.61$ (Chowell et al. 2004).

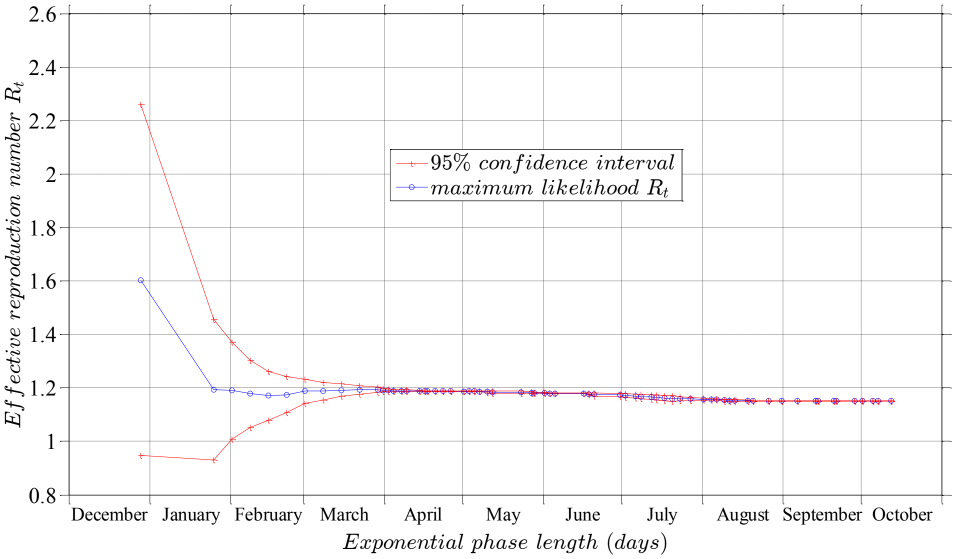


Figure 6.5 Time-varying Effective Reproduction Number for Guinea

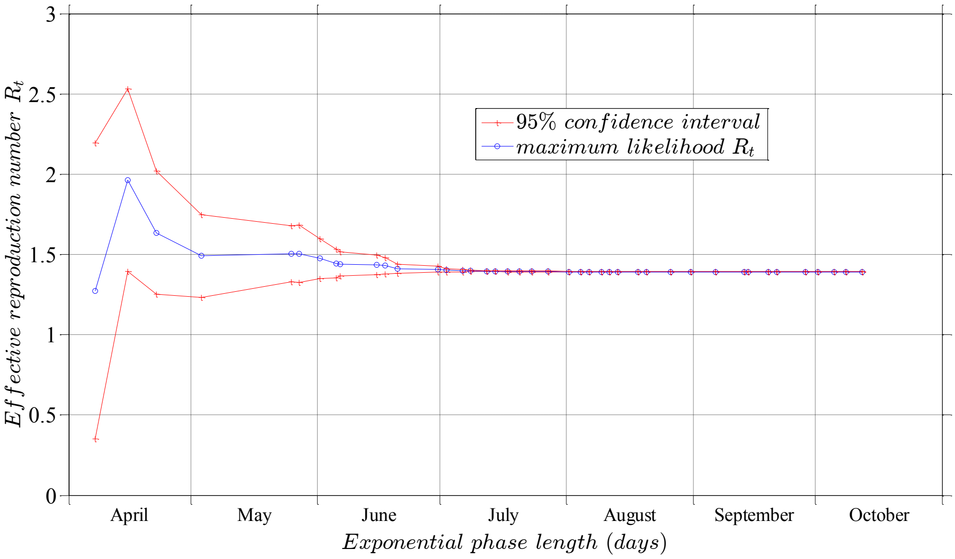


Figure 6.6 Time-varying Effective Reproduction Number for Sierra Leone

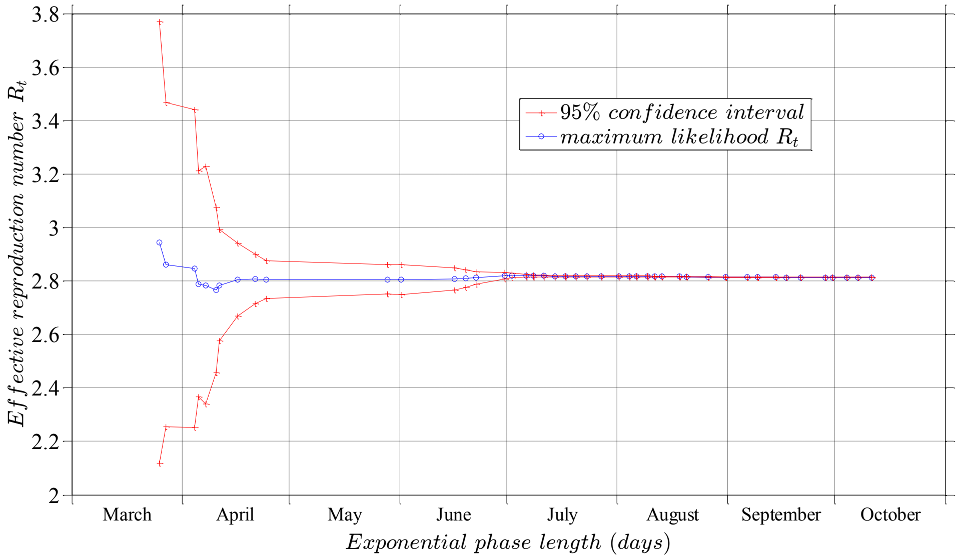


Figure 6.7 Time-varying Effective Reproduction Number for Liberia

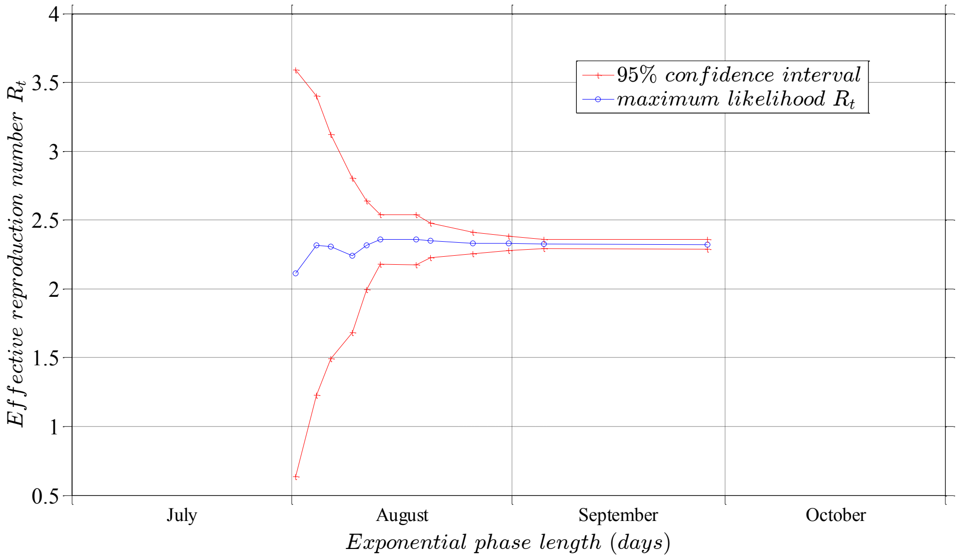


Figure 6.8 Time-varying Effective Reproduction Number for Nigeria

According to the above sequential Bayesian inference of stochastic SEIR model, we carry out the estimations of the time-varying effective reproduction number; as well as their 95% confidence intervals for the EVD infected countries in West Africa (Figure 6.5 to Figure 6.8). Solely based on the number of notified cumulative cases, the estimated effective reproduction number here provides an indicator of critical significance for perceiving the transmission intensity of EVD in each country. The final stabilized reproduction number for these four countries were, 1.15 for Guinea, 1.39 for Sierra Leone, dreadful 2.81 for Liberia, and 2.3 for Nigeria. It should be pointed out that the estimation for Nigeria may be not correct due to the insufficient EVD infected cases.

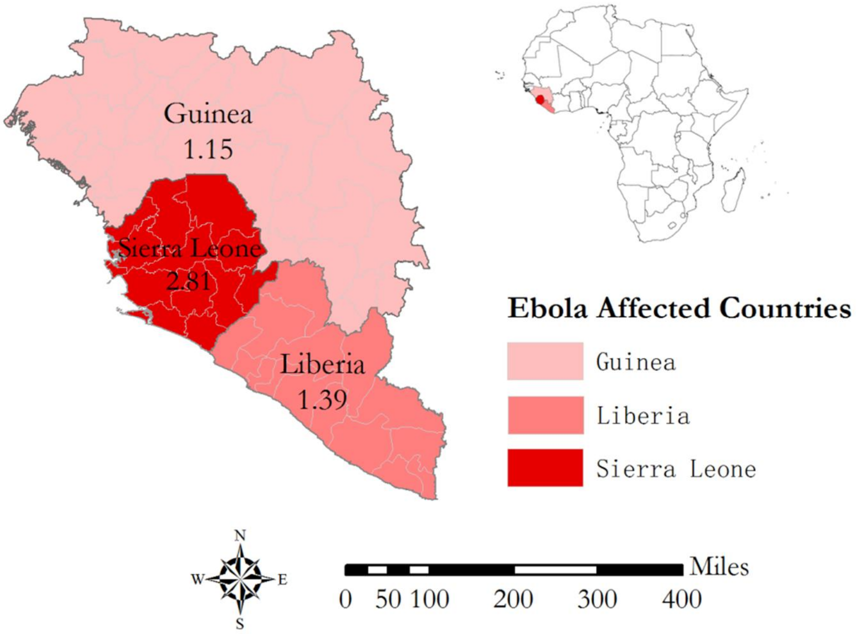


Figure 6.9 Choropleth map showing the effective reproduction number of Ebola affected countries

Geographical maps indicating the value of effective reproduction number can be used to identify areas of higher risk for an outbreak after an introduction. Figure 6.9 provides the choropleth map indicating the effective reproduction number of Ebola affected countries. Apparently, particular concerns and concerted efforts should be paid to halt all transmission in Liberia, where there was not only a high effective reproduction number, but also a prodigious base number of the infectious population. If no effective prevention measures were adopted, more and more EVD infected cases will be reported right from this country in the subsequent near future days.

In addition, such choropleth map can be further refined if the collected clinical database of EVD infection could be more location-specific. What's more, the smooth risk surface can also be expected (after being incorporated with the transport networks and population distribution) so as to precisely indicate where is the badly stricken area, thereby guiding the allocation of disaster relief resources.

Chapter 7 Conclusions

Epidemic waves of new emerging infectious diseases have awakened global concerns regarding their potential pandemic threats. In the meantime, rapid urbanization processes apparently accelerate the epidemic spread of emerging infectious diseases. Thus, spatiotemporal data analysis and mining can potentially contribute to characterizing the temporal evolution process and revealing possible spatial propagation patterns, thereby guiding the development of preventive intervention measures.

This research has investigated or developed several innovative models integrated with mathematical models within the GIS framework to understand and capture the behaviors of the frequent emerging infectious diseases, anticipating these models being conducive to the formulation of preventive intervention measures against infectious disease spread.

7.1 Concluding Summary

Firstly, the Standard deviational ellipse (SDE) has always been a versatile GIS tool for measuring the geographic distribution of concerned features. This research firstly summarizes two existing models of calculating SDE and then proposes a novel approach to construct the same SDE based on the spectral decomposition of the sample covariance, by which the SDE concept is further extended into higher dimensional Euclidean space, named standard deviational hyper-ellipsoid (SDHE).

A rigorous recursion formula is provided for calculating the confidence level of the scaled SDHE with an arbitrary magnification ratio in any dimensional space. Besides, an inexact-newton method based iterative algorithm is also proposed for solving the corresponding magnification ratio of a scaled SDHE when the confidence probability and space dimensionality are pre-specified. These results provide an efficient manner for substituting traditional table lookup of tabulated chi-square distribution. Finally, synthetic data is employed to generate the 1-3 multiple SDEs in two and three dimensions, and exploratory analysis by means of SDE is also conducted for measuring the spread concentrations of H1N1 of Hong Kong in 2009.

Mathematics is perceived as a powerful tool for understanding disease spread. Chapter 4 firstly devotes to modeling the temporal dynamics of an infectious disease by the basic SIR compartmental model, which potentially is capable of addressing some public health challenges and have recently been coupled with powerful computational methods to optimize the epidemic control strategies.

The geographic spread of infectious disease epidemics is of increasing concern, not only because of continuing threats of the liberal release of biological agents but also due to increasing rates of global travel, which provides an effective mechanism for disease spread. This situation makes it crucial to understand how, when, and why epidemics spread across the geographic landscape so that effective planning, preparation, and control measures can be in place before a disaster occurs.

In this connection, Chapter 4 further focuses upon one significant class of spatiotemporal methods come into being the form of partial differential equations, with reaction diffusion equation model as its typical representative, which possess the capacity of characterizing the dynamic evolution process for an infectious disease transmitting the virus from infectious individuals to the susceptible. Besides, it is also provided detailed computer implementation procedures using the Runge-Kutta method for simulating reaction diffusion equations.

Spatiotemporal analysis of new emerging infectious disease potentially contributes to characterizing the dynamic evolution process over time and revealing possible spatial propagation patterns. As such, an innovative approach was proposed in Chapter 5 for investigating the impact of spatiotemporal proximity upon the forecast infection risk of infectious disease outbreak. Experiments making use of the avian influenza A H7N9 in eastern China, from February to May 2013, demonstrated that the spatiotemporal proximity integrated approach is capable of providing approximately 70% correct prediction on average in terms of the H7N9 illness onset risk during the future 5 days ever since the forecasting date. Findings of this research can be served as an auxiliary means of great value for exploring the spatiotemporal propagation pattern of new emerging infectious disease, as well as making short-term predictions for deploying intensive effective prevention measures.

The epidemic situation of Ebola virus disease (EVD) in West Africa has been an ongoing concern in 2015. The effective reproduction number, average number of secondary infected cases, is an important indicator not only for reflecting the critical level of disease transmission over time, but also for assessing the efficacy of the adopted countermeasures. Thus, precise and prompt estimates of the reproduction number are of critical importance. A sequential Bayesian inference combined with stochastic SEIR model was used to estimate time-varying effective reproduction numbers, together with their 95% confidence intervals for each affected countries, so as to explore the transmission potential of the EVD. Experimental findings demonstrated that concerted efforts should be preferentially paid in Liberia to halt the dreadful transmission of EVD as indicated by a greater value of the effective reproduction number there.

7.2 Main Contributions

Main contributions of this thesis comprise four parts, being located in Chapter 3 to Chapter 6, which all deal with different characteristic patterns of epidemics. One chapter devotes to each model, contributing to characterize the transmission patterns of specific disease. The theory behind each chapter goes from the shallower to the deeper. However, there is no explicit relationship between each of these four models. Thus, we intend to conclude their contributions as per each model, respectively.

7.2.1 Standard Deviational Ellipse

(a) This research extends spatial analysis tool of Standard Deviational Ellipse (SDE) into higher dimensional Euclidean space, named standard deviational hyper-ellipsoid (SDHE).

(b) Meanwhile, it provides an efficient manner by a rigorous recursion formula and an inexact-newton method based iterative algorithm for conducting the confidence analysis.

(c) The proposed SDHE is obviously superior to the traditional SDE, since 1-3 multiple SDHEs (in three-dimensional space) can be employed for highlighting the spatiotemporal concentrations of Hong Kong's H1N1 infections.

7.2.2 Reaction Diffusion Equations

(a) The Reaction Diffusion equation serves as a spatial model. Simulations can easily be performed with parameters relevant for a particular disease with given demography and spatial structure.

(b) Diffusion coefficient of reaction diffusion equations can be determined if being integrated with the layers of hosts distribution, transmission routes, etc. so as to forecast the spatial spread trends of infectious disease.

(c) Main theoretical advantage of the reaction-diffusion equations is the deterministic and tractable nature of the continuous-space models.

7.2.3 Spatiotemporal Proximity Integrated Approach

(a) An innovative approach has been proposed for investigating the impact of spatiotemporal proximity upon the infection risk prediction of emerging infectious disease

(b) Experiments upon avian influenza A H7N9, February to May 2013 in eastern China, demonstrates that such spatiotemporal proximity integrated model can provide an approximately 70% correct prediction on average of the spatial significance level of the infection risk likelihood during the next five days ever since the notification release date.

(c) Findings of this research can be served as an auxiliary means for exploring the spatiotemporal propagation pattern of new emerging infectious disease, as well as making short-term predictions for the development of intensive effective prevention measures.

7.2.4 Sequential Bayesian Inference of Reproduction Number

(a) The sequential Bayesian inference approach only involves the data of cumulative number of case notifications, thereby making it versatile for most of the infectious disease.

(b) Time-varying effective reproduction numbers (including their 95% CIs) are conductive to reflecting the real-time hazard level regarding the disease transmission intensity.

(c) Geographical maps indicating the value of effective reproduction number can be used to identify areas of higher risk for an outbreak after an introduction.

7.3 A Software Prototype

Based on the aforementioned theoretical models, a software prototype framework is further presented below for analyzing the spatiotemporal spreading patterns and dynamic evolution trends of emerging infectious disease. From the practical viewpoint, its flexibility and extendibility allow rooms for improvement and wider applications in spatiotemporal analysis of the general infectious diseases.

Framework of a developing Software Prototype

Mission

Spatial-Temporal analysis of epidemic spread by mathematical models upon the platform of Geographic Information System

This icon is temporarily employed as the logo for this software prototype.

We named this software prototype as *ST*, which has been integrated with some functionality such as the elementary GIS operation of the geographic layers, spatial visualization and buffer analysis, as well as the Standard deviational ellipse (SDE), Kernel Density Estimation (KDE) and Kriging Trend Analysis (KTA), *etc*.

The Figure 7.1 illustrates the proposed framework of this software prototype, in which partial functionalities have been achieved (marked in blue color) while others indicated in red will be accomplished in the near future.

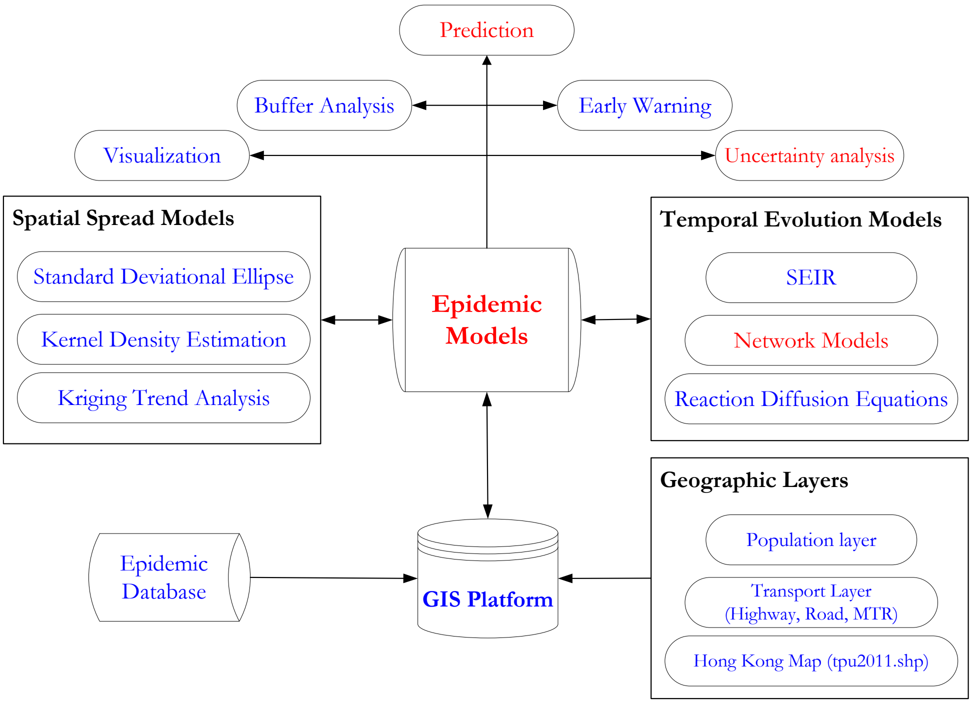


Figure 7.1 The framework of software prototype ***ST*** integrated with spatial-temporal epidemic models

This software prototype is put forward by integrating disease relevant data with typical mathematical models within the GIS platform. Meeting a need of solid expertise for effective decisions, this software prototype would probably be of interest for institutional decision makers and insurance industry.

7.4 Discussion (Limitations and Future Work)

This thesis mainly comprises four contributing models, which are anticipated to be conductive to understand and capture the behaviors of infectious disease. It is noteworthy that all these four models are proposed under the circumstance of very with limited clinical database in hand.

Derivations of both Standard Deviational Ellipses and its extension are under the assumption of observed samples following the normal distribution. Thus, a certain degree of caution is always necessary when employing the SDE tool for measuring the geographic distribution of concerned features. Particularly, delineation of an area concerned by SDE may not be representative of the hotspot boundaries, but produce ambiguous outcomes when distribution of features is multimodal.

Big challenge for applying the Reaction Diffusion Equations (RDE) is how to assign the diffusion coefficient so as to incorporate hosts mobility patterns into the disease spreading process, thereby forecasting the spatial spread trends of infectious disease. In theory, the diffusion coefficient can be determined by integrating with the layers of hosts distribution, transmission routes, etc. so as to reflect the complicated disease spreading situation. Therefore, It is also possible, that the diffusion coefficient D may depend on u, and/or explicitly on position x and time t. Sometimes, the geographic barrier constraints may also be added to limit the spreading range.

Still and all, research upon the RDE model is mainly staying from the theory level. Its enormous potential needs definitely to be attempted with more practical applications. Besides, the RDE model assumes that each group population is sufficiently large and homogeneous mixing with each subpopulation. Stochastic effects may be significant when group populations are small.

For the spatiotemporal proximity integrated approach, it is worth approving of the temporal proximity which creatively motivated by the principle that clinical symptom onset date is certainly posterior to the historical virus exposure time, as described by PDF of the incubation period fitted. However, the spatial proximity depicted by the KDE model seems a little bit arbitrary, not to mention the assumption of the spatial locations of infected individuals being static points within the epidemic region. As thus, higher prediction accuracy of this spatiotemporal proximity integrated approach can be anticipated if further environmental factors are taken into account, such as road networks, distribution of poultry markets and human population density, which may prominently account for the dissemination patterns of H7N9 dynamics.

Geographical maps with the value of effective reproduction number can be conductive for identifying areas of higher risk for an outbreak once infection introduced. Conservatively speaking, the Sequential Bayesian Inference (SBI) model can be further improved to output the spatial & temporal reproduction number (a GIS layer of the time-varying effective reproduction number) directly. Thus, choropleth map of R_t can be further refined (even the smooth risk surface) by utilizing more location-specific clinical database of EVD infections. Meanwhile, the calculated effective reproduction number by SBI can be employed to predict the future cumulative case number of EVD infections, thereby guiding the optimal allocation of limited public health resources.

References

AHAMADA, I., FLACHAIRE, E. and AHAMADA, I., 2010. *Non-parametric econometrics.* Oxford: Oxford University Press.

AJELLI, M., *et al.* 2010. Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. *Bmc Infectious Diseases,* 10(1), 190.

ALEXANDERSSON, A. 2004. Graphing confidence ellipses: An update of ellip for Stata 8. *Stata Journal,* 4(3), 242-256.

ANDERSON, R. M. and MAY, R. M., 1991. *Infectious diseases of humans.* Oxford university press Oxford.

ANDREWS, L., 1992. Special Functions of Mathematics for Engineers. McGraw-Hill, Inc.

ANGULO, J., *et al.* 2013. Spatiotemporal Infectious Disease Modeling: A BME-SIR Approach. *PLoS One,* 8(9), 12.

APARICIO, J. P., CAPURRO, A. F. and CASTILLO-CH VEZ, C. 2002. Markers of disease evolution: the case of tuberculosis. *Journal of Theoretical Biology,* 215(2), 227-237.

BAILEY, N. T., 1975. *The mathematical theory of infectious diseases and its applications.* Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

BALCAN, D., *et al.* 2010. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of Computational Science,* 1(3), 132-145.

BEARDMORE, I. and BEARDMORE, R. 2003. The global structure of a spatial model of infectious disease. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences,* 459(2034), 1427-1448.

BISWAS, M. H. A., PAIVA, L. T. and DE PINHO, M. 2014. A SEIR MODEL FOR CONTROL OF INFECTIOUS DISEASES WITH CONSTRAINTS. *Mathematical Biosciences and Engineering,* 11(4), 761-784.

BROADFOOT, J. D., ROSATTE, R. C. and O'LEARY, D. T. 2001. Raccoon and skunk population models for urban disease control planning in Ontario, Canada. *Ecological Applications,* 11(1), 295-303.

BROCKMANN, D. and HELBING, D. 2013a. The hidden geometry of complex, network-driven contagion phenomena. *science,* 342(6164), 1337-1342.

BROCKMANN, D. and HELBING, D. 2013b. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. *science,* 342(6164), 1337-1342.

BULIUNG, R. N. and KANAROGLOU, P. S. 2006. A GIS toolkit for exploring geographies of household activity/travel behavior. *Journal of Transport Geography,* 14(1), 35-51.

CHAINEY, S., TOMPSON, L. and UHLIG, S. 2008. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal,* 21(1), 4-28.

CHOWELL, G., *et al.* 2006. Transmission dynamics of the great influenza pandemic of 1918 in Geneva, Switzerland: assessing the effects of hypothetical interventions. *Journal of Theoretical Biology,* 241(2), 193-204.

CHOWELL, G., *et al.* 2004. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology,* 229(1), 119-126.

CHOWELL, G., NISHIURA, H. and BETTENCOURT, L. M. 2007. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of The Royal Society Interface,* 4(12), 155-166.

CHOWELL, G.*, et al.* 2013. Transmission potential of influenza A/H7N9, February to May 2013, China. *BMC medicine,* 11.

CLOUTIER, V.*, et al.* 2008. Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. *Journal of Hydrology,* 353(3), 294-313.

COWLING, B. J.*, et al.* 2013. Comparative epidemiology of human infections with avian influenza A H7N9 and H5N1 viruses in China: a population-based study of laboratory-confirmed cases. *Lancet,* 382(9887), 129-137.

CROMLEY, R. G., 1992. *Digital cartography.* Englewood Cliffs, N.J.: Prentice Hall.

DE SMITH, M. J., GOODCHILD, M. F. and LONGLEY, P., 2007. *Geospatial analysis: a comprehensive guide to principles, techniques and software tools.* Troubador Publishing Ltd.

DELMELLE, E.*, et al.* 2014. Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science,* 28(5), 1107-1127.

DUONG, T. and HAZELTON, M. L. 2005. Cross‐validation Bandwidth Matrices for Multivariate Kernel Density Estimation. *Scandinavian Journal of Statistics,* 32(3), 485-506.

EDLUND, S. B., DAVIS, M. A. and KAUFMAN, J. H., 2010. The spatiotemporal epidemiological modeler. *Proceedings of the 1st ACM International Health Informatics Symposium.* Arlington, Virginia, USA: ACM, 817-820.

EPSTEIN, J. M., 2004. *Toward a containment strategy for smallpox bioterror: an individual-based computational approach.* Brookings Institution Press.

EPSTEIN, J. M., 2006. *Generative social science: Studies in agent-based computational modeling.* Princeton University Press.

ERYANDO, T.*, et al.* 2012. Standard Deviational Ellipse (SDE) models for malaria surveillance, case study: Sukabumi district-Indonesia, in 2012. *Malaria Journal,* 11(Suppl 1), P130.

FANG, K. T. 2004. Elliptically contoured distributions. *Encyclopedia of Statistical Sciences*.

FAVIER, C.*, et al.* 2006. Early determination of the reproductive number for vector‐borne diseases: the case of dengue in Brazil. *Tropical Medicine & International Health,* 11(3), 332-340.

FINKENST DT, B. and GRENFELL, B. 1998. Empirical determinants of measles metapopulation dynamics in England and Wales. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 265(1392), 211-220.

FRASER, C.*, et al.* 2009. Pandemic potential of a strain of influenza A (H1N1): early findings. *science,* 324(5934), 1557-1561.

FURFEY, P. H. 1927. A Note on Lefever's "Standard Deviational Ellipse". *American Journal of Sociology,* 33(1), 94-98.

GAO, H.-N.*, et al.* 2013a. Clinical findings in 111 cases of influenza A (H7N9) virus infection. *New England Journal of Medicine,* 368(24), 2277-2285.

GAO, R.*, et al.* 2013b. Human infection with a novel avian-origin influenza A (H7N9) virus. *New England Journal of Medicine,* 368(20), 1888-1897.

GILBERT, M.*, et al.* 2014. Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. *Nature communications,* 5.

GILBERT, M., *et al.* 2008. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proceedings of the National Academy of Sciences,* 105(12), 4769-4774.

GOMES, M., *et al.* 2014. Assessing the international spreading risk associated with the 2014 West African Ebola outbreak. *PLOS Currents Outbreaks*.

GONG, J. X. 2002. Clarifying the standard deviational ellipse. *Geographical Analysis,* 34(2), 155-167.

GONZALEZ, M. C., HIDALGO, C. A. and BARABASI, A.-L. 2008. Understanding individual human mobility patterns. *Nature,* 453(7196), 779-782.

GRAUNT, J., 1939. *Natural and Political Observations made upon the Bills of Mortality.* The Johns Hopkins Press.

GRIMMETT, G., 1999. *What is Percolation?* : Springer.

H RDLE, W. K. and SIMAR, L., 2012. *Applied multivariate statistical analysis.* Springer.

HEFFERNAN, J., SMITH, R. and WAHL, L. 2005. Perspectives on the basic reproductive ratio. *Journal of The Royal Society Interface,* 2(4), 281-293.

HELD, L. and PAUL, M. 2012. Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal,* 54(6), 824-843.

HUBER, G. 1982. Gamma function derivation of n-sphere volumes. *The American Mathematical Monthly,* 89(5), 301-302.

HUBERT, M. and DEBRUYNE, M. 2009. Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics,* 2(1), 36-43.

JANDAROV, R., *et al.* 2014. Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *Journal of the Royal Statistical Society Series C-Applied Statistics,* 63(3), 423-444.

KANE, M. J., *et al.* 2014. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *Bmc Bioinformatics,* 15.

KAO, R. R. 2003. The impact of local heterogeneity on alternative control strategies for foot-and-mouth disease. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 270(1533), 2557-2564.

KEELING, M. and GRENFELL, B. 1997. Disease extinction and community size: modeling the persistence of measles. *science,* 275(5296), 65-67.

KEELING, M. J. and ROHANI, P., 2008. *Modeling infectious diseases in humans and animals.* Princeton University Press.

KELLEY, C. T., 2003. *Solving nonlinear equations with Newton's method.* Philadelphia: Society for Industrial and Applied Mathematics.

KENT, J. and LEITNER, M. 2007. Efficacy of standard deviational ellipses in the application of criminal geographic profiling. *Journal of Investigative Psychology and Offender Profiling,* 4(3), 147-165.

KERMACK, W. O. and MCKENDRICK, A. G. 1927. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character,* 115(772), 700-721.

KERMACK, W. O. and MCKENDRICK, A. G. 1932. Contributions to the mathematical theory of epidemics. II. The problem of endemicity. *Proceedings of the Royal society of London. Series A,* 138(834), 55-83.

KERMACK, W. O. and MCKENDRICK, A. G. 1933. Contributions to the mathematical theory of epidemics. III. Further studies of the problem of endemicity. *Proceedings of the Royal society of London. Series A,* 141(843), 94-122.

KLEINSCHMIDT, I., *et al.* 2002. Rise in malaria incidence rates in South Africa: a small-area spatial analysis of variation in time trends. *Am J Epidemiol,* 155(3), 257-264.

KN SEL, L. 1986. Computation of the Chi-Square and Poisson Distribution. *SIAM journal on scientific and statistical computing,* 7(3), 1022-1036.

KULLDORFF, M., *et al.* 2006. An elliptic spatial scan statistic. *Statistics in medicine,* 25, 3929 - 3943.

LAI, P. C., KWONG, K. H. and WONG, H. T. 2013. Spatio-temporal and stochastic modelling of severe acute respiratory syndrome. *Geospat Health,* 8(1), 183-192.

LAPIDUS, L. and PINDER, G. F., 2011. *Numerical solution of partial differential equations in science and engineering.* John Wiley & Sons.

LAWSON, A. B., BROWNE, W. J. and RODEIRO, C. L. V., 2003. *Disease mapping with WinBUGS and MLwiN.* Wiley.

LEFEVER, D. W. 1926. Measuring Geographic Concentration by Means of the Standard Deviational Ellipse. *American Journal of Sociology,* 32(1), 88-94.

LEVIN, S. A., *et al.* 1997. Mathematical and computational challenges in population biology and ecosystems science. *science,* 275(5298), 334-343.

LI, Q. and RACINE, J. S., 2007. *Nonparametric econometrics: Theory and practice.* Princeton University Press.

LI, Q., *et al.* 2014. Epidemiology of human infections with avian influenza A(H7N9) virus in China. *N Engl J Med,* 370(6), 520-532.

LIPSITCH, M., *et al.* 2003. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *science,* 300(5627), 1966-1970.

MCKENDRICK, A. 1940. The Dynamics of Crowd Infection. *Edinburgh Medical Journal,* 47, 117-136.

MEYER, S., ELIAS, J. and HOHLE, M. 2012. A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence. *Biometrics,* 68(2), 607-616.

MEYERS, L. 2007. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society,* 44(1), 63-86.

MOLLISON, D. and KUULASMAA, K. 1985. Spatial epidemic models: theory and simulations. *Population dynamics of rabies in wildlife,* 8, 291-309.

MURRAY, J. D., 2003. *Mathematical Biology: II: Spatial Models and Biomedical Applications.* New York, NY: Springer New York, New York, NY.

NEWMAN, M. E. 2002. Spread of epidemic disease on networks. *Physical Review E,* 66(1), 016128.

NGAN, H., 2010. *Analysis of Human Swine Influenza in Hong Kong Based on the Spatial and Temporal Approach.* BSc (Hons) in Geomatics. Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University.

NISHIURA, H. and CHOWELL, G. 2014. Early transmission dynamics of Ebola virus disease (EVD), West Africa, March to August 2014. *Eurosurveillance,* 19(36), 5-10.

NISHIURA, H., *et al.* 2006. Transmission potential of primary pneumonic plague: time inhomogeneous evaluation based on historical documents of the transmission network. *Journal of epidemiology and community health,* 60(7), 640-645.

NSOESIE, E. O., *et al.* 2014. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses,* 8(3), 309-316.

O'NEILL, P. D. 2002. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences,* 180(1–2), 103-114.

PASCUAL, M., MAZZEGA, P. and LEVIN, S. A. 2001. Oscillatory dynamics and spatial scale: the role of noise and unresolved pattern. *Ecology,* 82(8), 2357-2369.

RELUGA, T. 2004. A two-phase epidemic driven by diffusion. *Journal of Theoretical Biology,* 229(2), 249-261.

RILEY, S., *et al.* 2003. Transmission Dynamics of the Etiological Agent of SARS in Hong Kong: Impact of Public Health Interventions. *science,* 300(5627), 1961-1966.

ROUSSEEUW, P. J. and DRIESSEN, K. V. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics,* 41(3), 212-223.

SCOTT, D. W. 1979. On optimal and data-based histograms. *Biometrika,* 66(3), 605-610.

SHI, R. X., *et al.* 2014. Spatiotemporal pattern of hand-foot-mouth disease in China: an analysis of empirical orthogonal functions. *Public Health,* 128(4), 367-375.

SHIODE, S. 2012. Revisiting John Snow's map: network-based spatial demarcation of cholera area. *International Journal of Geographical Information Science,* 26(1), 133-150.

SMITH, R. C. and CHEESEMAN, P. 1986. On the Representation and Estimation of Spatial Uncertainty. *International Journal of Robotics Research,* 5(4), 56-68.

SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P., 2004. *Bayesian approaches to clinical trials and health-care evaluation.* Wiley.

STEVENSON, M. A., *et al.* 2005. Area-level risks for BSE in British cattle before and after the July 1988 meat and bone meal feed ban. *Preventive Veterinary Medicine,* 69(1-2), 129-144.

TOBLER, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 234-240.

TREFETHEN, L. N. and BAU III, D., 1997. *Numerical linear algebra.* Society for Industrial Mathematics.

TSUI, K. L., *et al.* 2011. Recent Research and Developments in Temporal and Spatiotemporal Surveillance for Public Health. *Ieee Transactions on Reliability,* 60(1), 49-58.

WANG, B., SHI, B. and INYANG, H. I. 2008. GIS-based quantitative analysis of orientation anisotropy of contaminant barrier particles using standard deviational ellipse. *Soil & Sediment Contamination,* 17(4), 437-447.

WANG, J. F., *et al.* 2011. Hand, foot and mouth disease: spatiotemporal transmission and climate. *Int J Health Geogr,* 10(25), 25.

WANG, J. F., *et al.* 2013. Spatial dynamic patterns of hand-foot-mouth disease in the People's Republic of China. *Geospat Health,* 7(2), 381-390.

WILLIAMS, D. M., QUINN, A. C. D. and PORTER, W. F. 2014. Informing Disease Models with Temporal and Spatial Contact Structure among GPS-Collared Individuals in Wild Populations. *PLoS One,* 9(1).

WONG, D. W. 1998. Measuring multiethnic spatial segregation. *Urban Geography,* 19(1), 77-87.

YANG, W., *et al.* 2011. A nationwide web-based automated system for outbreak early detection and rapid response in China. *Western Pac Surveill Response J,* 2(1), 10-15.

YIN, Y., 2007. *Bayesian analysis of infectious disease time series data and optimal constrained Bayesian updating.*

YU, H.-L., *et al.* 2014a. An online spatiotemporal prediction model for dengue fever epidemic in Kaohsiung (Taiwan). *Biometrical Journal,* 56(3), 428-440.

YU, H., *et al.* 2014b. Effect of closure of live poultry markets on poultry-to-person transmission of avian influenza A H7N9 virus: an ecological study. *The Lancet,* 383(9916), 541-548.

YUILL, R. S. 1971. The Standard Deviational Ellipse; An Updated Tool for Spatial Description. *Geografiska Annaler. Series B, Human Geography,* 53(1), 28-39.

ZHANG, S., 2011. *A Study of Spatial-temporal Distribution of Human Swine Influenza in Hong Kong Based on GIS and Spatial Analysis.* Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University.

ZHANG, Z., *et al.* 2010. Spatio-temporal data comparisons for global highly pathogenic avian influenza (HPAI) H5N1 outbreaks. *PLoS One,* 5(12), e15314.

ZHANG, Z., *et al.* 2014. Prediction of H7N9 epidemic in China. *Chinese medical journal,* 127(2), 254-260.

ZHU, G. and PETERSON, A. T. 2014. Potential geographic distribution of the novel avian-origin influenza A (H7N9) virus. *PLoS One,* 9(4), e93390.