SPATIAL STATISTICS OF EPIDEMIC DATA: THE CASE OF CHOLERA EPIDEMIOLOGY IN GHANA

Frank Badu Osei

Examining committee:

Prof.dr. W. Albers	Twente University
Prof.dr. M.J. Kraak	Twente University
Dr. M.N.M. van Lieshout	Centrum Wiskunde & Informatica
Prof.dr. E. Pebesma	University of Münster

UNIVERSITY OF TWENTE.

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

ITC dissertation number 177
ITC, P.O. Box 6, 7500 AA Enschede, The Netherlands

ISBN 978-90-6164-299-2
Cover designed by Benno Masselink
Printed by ITC Printing Department
Copyright © 2010 by Frank Badu Osei

SPATIAL STATISTICS OF EPIDEMIC DATA: THE CASE OF CHOLERA EPIDEMIOLOGY IN GHANA

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the Rector Magnificus,
prof.dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Thursday 9 December 2010 at 15:00 hrs

by

Frank Badu Osei

born on 24 March 1980

in Kumasi, Ghana

This thesis is approved by
Prof.dr.ir. Alfred Stein, promoter
Dr. Alfred Allan Duker, assistant promoter

To my Father, William Osei Badu
&
Mother, Comfort Kusi

Table of Contents

1. Introduction 1
   - 1.1 Cholera and *V. cholerae* 2
     - 1.1.1 Biology and ecology of *V. cholerae* 2
     - 1.2.3 Transmission hypothesis 3
   - 1.3 Socioeconomic and environmental risk factors 4
   - 1.4 The burden of cholera 5
   - 1.4 Statistical methods for spatial epidemiology 6
     - 1.4.1 Cluster analysis 6
     - 1.4.2 Ecological analysis 8
   - 1.5 Spatial epidemiology of cholera 12
   - 1.6 Aims and objectives 13
     - 1.6.1 Specific objectives 13

     The specific objectives of this study are to 13
   - 1.7 The study area 14
     - 1.7.1 Cholera in Ghana 14
     - 1.7.2 Case definition of cholera in Ghana 14
   - 1.8 Research framework and methods 15
     - 1.8.1 Datasets and spatial data creation 15
     - 1.8.2 Spatial analyses and statistical modelling 16
   - 1.9 Outline of thesis 17

2. Spatial and demographic patterns of cholera 19
   - 2.1 Introduction 21
   - 2.2 The study area 22
   - 2.3 Methods 24
     - 2.3.1 Research methodology 24
     - 2.3.2 Spatial data preparation and cartographic display 24
     - 2.3.3 Spatial autocorrelation analyses 25
   - 2.4 Results and analyses 27
   - 2.5 Discussion 30
     - 2.5.1 Limitations of study 32
   - 2.6 Conclusion 32

3. Spatial and space-time clustering of cholera 33
   - 3.1 Introduction 35
   - 3.2 Methods 36
     - 3.2.1 Study area 36
     - 3.2.2 Data sources 37
     - 3.2.3 Cluster analysis 38
     - 3.2.4 Correlation between cholera and risk factors 40
   - 3.3 Results and analyses 41
     - 3.3.1 Purely spatial clusters 41
     - 3.3.2 Space-time clusters 42
     - 3.3.3 Correlation between cholera and risk factors 43
   - 3.4 Discussion 44

3.5 Conclusion....47

4. Spatial dependency of cholera on refuse dumps....49
4.1 Introduction....51
4.2 Materials and methods....53
4.2.1 The study area....53
4.2.2 Cholera case definition and data....54
4.2.3 Refuse dumps data....54
4.2.4 Spatial data input....55
4.2.5 Spatial data analysis and statistical modelling....55
4.2.6 Spatial clusters detection....61
4.2.7 Critical buffer distance....63
4.3 Results and analysis....63
4.3.1 Association between cholera and refuse dumps....63
4.3.2 Cholera incidence clusters....64
4.3.3 Critical buffer distance....66
4.4 Discussion....66
4.4.1 Association between cholera and refuse dump....66
4.4.2 Cholera incidence clusters....68
4.4.3 Critical buffer distance....68
4.5 Conclusion....69

5. Spatial dependency of cholera on potential cholera reservoirs....71
5.1 Introduction....73
5.2 Materials and methods....74
5.2.1 Study framework and methodology....74
5.2.2 The study area....75
5.2.3 Spatial data input....76
5.2.4 Delineation of potential cholera reservoirs....76
5.2.5 Spatial factors maps....77
5.2.6 Proximity analysis: regression modelling....78
5.2.7 Flexible spatial cluster analysis....81
5.3 Results and analysis....83
5.3.1 Dependency of cholera on $d_{(\mathrm{All})}$, $d_{(\mathrm{Up})}$ and $d_{(\mathrm{Dw})}$....83
5.3.2 Cluster analysis....85
5.4 Discussion....85
5.5 Conclusion....88

6. Multivariate Bayesian semi-parametric modelling of cholera in an urban environment....89
6.1 Introduction....91
6.2 Methods....92
6.2.1 Cholera data and risk estimation....92
6.2.2 Continuous, spatial and categorical covariates....93
6.2.3 Model specification....94
6.2.4 Prior distributions for covariates....95
6.2.5 Bayesian inference....98

6.2.6 Model implementation 99
6.3 Results and analysis 100
6.3.1 Sensitivity analyses and model selection 100
6.3.2 Fixed and nonlinear effects of covariates 102
6.3.3 Spatial effects 105
6.4 Discussion 105
6.5 Conclusion 107

7. Bayesian modelling of the space-time diffusion pattern of cholera epidemic 109
7.2 Methods and Data 112
7.2.1 Study area and Data 112
7.2.2 Defining the extent of contagiousness: variogram modelling 113
7.2.3 Defining transmission network routes of cholera diffusion 115
7.2.4 Time-ordered diffusion modelling 117
7.3 Results 120
7.4 Discussion 124

8. Research findings, conclusions and recommendations for further research: A synthesis 127
8.1 Overview 128
8.2 Major research findings 128
8.2.1 Spatial and temporal patterns of cholera 128
8.2.2 Demographic patterns of cholera 129
8.2.3 Dependency of cholera on refuse dumps 129
8.2.4 Dependency of cholera on surface water pollution 130
8.2.5 Dependency of cholera on slums 131
8.2.6 Dependency of cholera on spatial interaction 131
8.2.7 Diffusion dynamics of cholera 132
8.3 Research conclusions 133
8.4 Recommendations for future studies 135
8.4.1 Ecological fallacy 135
8.4.2 Modifiable areal unit problem 136
8.4.3 Edge effects 137
8.4.4 Remote sensing and cholera prediction 138
8.4.5 Spatial data quality issues 139

Bibliography 141

Curriculum vitae 161

List of symbols

Mathematical symbols

$\sim$	Distribution symbol
$\propto$	Proportionality symbol
Σ	Summation

Greek symbols

η	Linear predictor
σ^2	Variance parameter
Σ_σ	Variance-covariance matrix
ρ	Spatial autoregressive coefficient for spatial lag model
λ	Spatial autoregressive coefficient for spatial error model
χ^2	Chi square
$\chi^2(1)$	Chi square with one degree of freedom
$\gamma_{Chol_{(\mathrm{E(R)})}}$	Experimental variogram for the random variable $Chol_{(\mathrm{E(R)})}$
$\sigma^2_{Chol_{(\mathrm{E(R)})}}$	Variance of expected cholera risk
$\rho_{(\mathrm{pop})}$	Binary variable for population density; $\rho_{(\mathrm{pop})} = 0$ for moderate population density, and $\rho_{(\mathrm{pop})} = 1$ for high population density
$\varsigma_{(\mathrm{slum})}$	Binary variable for presence of slum settlers; $\varsigma_{(\mathrm{slum})} = 1$ denotes the presence of slum settlers, and $\varsigma_{(\mathrm{slum})} = 0$ otherwise
$\rho_{(\mathrm{dump})}$	Density of refuse dumps

Latin symbols

$E(y\|.)$	Conditional expectation of y
$E(y)$	Expectation of y
$\mathrm{Var}(.)$	Variance
I	Identity matrix
$p(.)$	Probability density
$p(\theta\|.)$	Conditional probability density of θ
$f(.)$	Nonlinear function
$N(a,b)$	Normal distribution with mean a and variance b

$MVN(a,b)$	Multivariate normal distribution with mean a and variance b
$N(0,\Sigma)$	Normal distribution with zero mean and variance-covariance matrix Σ
$\ln L(.)$	Log-likelihood function
$L(.)$	Likelihood function
LM_{ρ}	Langrage Multiplier statistic for spatial lag model
LM_{λ}	Langrage Multiplier statistic for spatial error model
tr	Trace operator for matrices
$f_{spat}(.)$	Function for spatial effects
$f_{str}(.)$	Function for structured spatial effects
$f_{unstr}(.)$	Function for unstructured spatial effects
I_{M}	Moran's Index
$C_{Chol_{(\mathrm{E(R)})}}$	Covariance function for the random variable $Chol_{(\mathrm{E(R)})}$

Data symbols

$Dist_{ij}$	Spatial weights matrix corresponding to the district pairs i and j
Com_{ij}	Spatial neighbourhood matrix corresponding to the community pairs i and j
$\overleftarrow{\overline{Com}}_{(\mathrm{t})i,j}$	Temporal neighbourhood matrix (Transmission matrix) with elements representing the probability of cholera transmission from community i to j with respect to time
$\overleftarrow{\overline{Com}}_{(\mathrm{t,s})i,j}$	Spatio-temporal transmission matrix with elements representing the probability of cholera transmission from community i to j with respect to space and time
$Chol_{(\mathrm{R})}$	Cholera incidence rate or risk
$\overline{Chol}_{(\mathrm{E(R)})}$	Mean of expected cholera risk
$Chol_{(\mathrm{E(R)})}$	Expected cholera risk
$\overline{Chol}_{(\mathrm{R})}$	Mean cholera incidence rate
$\tilde{Chol}_{(\mathrm{R})}$	Population weighted mean of cholera incidence rate
$Chol_{(\mathrm{RR})}$	Relative risk of cholera
$Chol_{(\mathrm{C})}$	Observed number cholera cases
$Chol_{(\mathrm{E(C)})}$	Expected number of cholera cases
$\overline{\overline{Chol}}_{(\mathrm{R})}$	Latent variable of cholera incidence rate

$Chol_{(C)}(W)$	Observed number of cholera cases within the window W
$Chol_{(E(C))}(W)$	Expected number of cholera cases within the window W
$Chol_{(C)t}$	Number of cholera cases at time t
$d_{(dump)}$	Distance to refuse dumps
$d_{(All)}$	Proximity to all cholera reservoirs
$d_{(Up)}$	Proximity to upstream cholera reservoirs
$d_{(Dw)}$	Proximity to downstream cholera reservoirs

Abbreviations

WHO	World Health Organization
CT	cholera toxin
GLM	Generalized linear model
GAMM	generalized additive mixed model
SAR	spatial autoregressive
CAR	conditional autoregressive
SMA	spatial moving average
GAM	Generalized additive model
MCMC	Markov Chain Monte Carlo
STAR	structured additive regression
OLS	Ordinary Least Squares
GIS	Geographic Information System
DCU	Disease Control Unit
KMHD	Kumasi Metropolitan Health Directorate
ESRI	Environmental System Research Institute
UTM	***Universal Transverse Mercator***
GTM	Ghana Transverse Mercator
EBS	Empirical Bayesian Smoothing
RR	rate ratios
GSS	Ghana Statistical Service
KVIP	Kumasi ventilated improved pit
GPS	Global Positioning System
WC	Water closet
DHD	District Health Directorates
MLC	most likely cluster
ML	maximum likelihood
AIC	Akaike Information Criterion values
DIC	deviance information criterion

Acknowledgements

"If I have been able to see further, it was only because I stood on the shoulders of giants"

Isaac Newton

This thesis began some five years ago; never did I know it would come to this far, let alone to a successful completion. I would not have made it this far without the help and assistance of many people.

First and foremost, my sincere thanks go to my supervisors, Professor Dr Alfred Stein (Promotor) and Dr Alfred Allan Duker (Assistant promoter) for their unconditional support and assistance. I started this thesis with Alfred Duker in Ghana. My sincere gratitude to him cannot be expressed in words. As a student without any research background, Alfred Duker accepted me as his PhD student and guided me step-by-step through research procedures and methods. He encouraged me about the need of publishing my research findings in peer reviewed journals. I owe a debt of gratitude to Prof Alfred Stein. Studying under Alfred Stein is a prestigious experience that will be unforgettable in the rest of my life. Alfred is a professor I admire so much. Alfred inspired, encouraged, and strengthened me into the field of spatial statistics. Once again, thank you for translating my abstract into Dutch. To Alfred and Alfred, I say thank you very much!

The contribution of Ellen-Wien Augustijn cannot be left unrecognized. I gained from Ellen significant knowledge in Geographic Information System and geo-processing. Ellen's contribution is partly evidenced in a paper we wrote together. I also acknowledge the contributions of Mohammed Ali, Michael Emch, Rene J Borroto, Havard Rue and Pierre Goovaerts. I never met these scientists personally, but they responded each time I wrote to them for help.

In special way, my sincere thanks go to Dr. Paul van Dijk, who together with Alfred Stein secured funding for me to continue my studies at ITC. I also knowledge the invaluable assistance of Loes Colenbrander who took care of all my travelling arrangements to the Netherlands, and made sure this document was correctly formatted and ready for printing. To Teresa Brefeld, I say thank you, especially for making sure coffee was available every Monday morning. To the staff of ITC, especially members of the EOS department, I say thank you very much.

I am grateful to the Kwame Nkrumah University of Science and Technology (KNUST) for providing funding for the initial stages of my study. I am also grateful to all the staff and workers of the Geomatic Engineering, KNUST. Ante Kate, thank you for taking care of all my letters and administrative matters whenever I called on you.

I am immensely grateful to all the friends I met in ITC and Enschede, especially the Ghanaian community. My warmest thanks go to Anthony, Amoa, Forson, Eric Adjei, and Gavu. My heartfelt thanks go to all my friends in Ghana, especially Akamah, Owusu, Okyere and Stephen.

My sincerest thanks go to all fellow PhD students, especially those I had the chance to interact with one way or the other. To Xia Li, I say thank you for the time you spent on me when I needed to utilize the space-time cube technique to explore my data. I acknowledge my debt to Gaurav Singh, Berhanu Kafale, Caroline Mathenge, Pablo Lopez, Mustafa, Abel Ramoelo, Tagel Gebrehiwot. I really enjoyed your friendship and the times and conversations we shared together during every coffee break.

My heartfelt thanks go to my family for their support and prayer. I am grateful to my parents whose support, sacrifice and encouragements have helped me to reach this far. I owe a great debt of appreciation to the family of Alfred Duker for always supporting me with prayer. I am grateful to all friends and love one whose names have not been written here.

Finally, I give thanks to the Almighty God through whose grace has sustained me to reach this level of life. It has been a long journey. Many obstacles stood the way; yet the Almighty God sort me through, turning every failure into successful and enjoyable moments. Through these struggles I have come to recognize the faithfulness of God.

Abstract

The growing number and increased frequency of major cholera outbreaks, especially in African countries, have heightened concerns about the disease, in particular about its spatial and temporal characteristics and their underlying risk factors. Cholera is transmitted mainly through contaminated water and food; however, demographic and geographic factors can predispose inhabitants to infection. Socioeconomic and environmental factors like environmental sanitation can influence the vulnerability of a population to cholera infection. In Ghana, the recurrent of cholera has raised possible endemic foci in urban communities which seem to report greater percentage of cases during outbreaks. Yet, little is known about the spatial and temporal characteristics of the disease. This thesis uses past cholera epidemic data and spatial statistical methodologies to better understand the effects of socioeconomic and environmental risk factors on the spatial epidemiology of the disease. Two separate study areas in Ghana are used, the Ashanti Region and its capital, Kumasi Metropolis.

Chapters 2 and 3 present a spatial cluster analysis to investigate and describe the spatial and spatio-temporal patterns of cholera. High cholera rates cluster around Kumasi Metropolis (the central part of the region), with significant Moran's I = 0.271 and $p < 0.001$. A Mantel-Haenszel *Chi square* test for trend analyses reflects a direct spatial relationship between cholera and urbanization ($\chi^2 = 2995.5$, $p < 0.0001$), overcrowding ($\chi^2 = 1757.2$, $p < 0.0001$), and an inverse relationship between cholera and adjacency/proximity with Kumasi Metropolis ($\chi^2 = 831.38$, $p < 0.0001$). Cholera prevalence is high if the majority of the people do not have access to good sanitation facilities, drink from rivers, wells and ponds, and if internal migration is high.

Chapter 4 uses proximity to and density of refuse dumps as proxies for environmental sanitation. Spatial lag and spatial error regression models are developed and implemented to determine the spatial dependence of cholera on open-space refuse dumps. In the regression models, cholera prevalence shows a direct relationship with density of refuse dumps, and an inverse relationship to their distance. The spatial autoregressive coefficients are significant for the spatial lag and spatial error models. This shows the dependence of cholera on spatial interaction between communities, and the existence of unobserved influential factors.

Chapter 5 shows a steepest downhill path analysis using a 3D elevation model and refuse dumps locations to delineate *potential cholera reservoirs*. Using proximity to the *potential cholera reservoirs* as explanatory variables, statistical models are developed and implemented to assess the effects of surface water pollution on cholera. High cholera is associated with this proximity, whereas the significance of spatial autoregressive coefficients in the statistical models reveals the dependence of cholera on the spatial interaction between communities and possible unobserved risk factors.

Chapter 6 combines the identified risk factors and other risk factors into a coherent multivariate statistical model. A Bayesian semi-parametric regression approach is used,

allowing us to carry out a joint analysis of nonlinear effects of continuous covariates, spatially structured variation, unstructured heterogeneity, as well as fixed covariates. The model reveals that the risk of cholera is high amongst communities with slum settlements (*posterior mean = 4.06, $p < 0.01$*) and densely populated communities (*posterior mean = 4.339, $p < 0.01$*). The relationship between cholera and dump density is almost linear with increasing posterior mean. The posterior mean of the proximity to dump sites deviates from linearity, with a decreasing risk up to approximately 500 m, and a slight increase for larger distances. Its relationship with proximity to potential cholera reservoirs is also almost linear, however, with the posterior means decreasing with increasing distance. There is evidence of distinct spatial variation, with significant increased cholera risk at the central part of Kumasi, and a significant reduced risk at the south-eastern part.

Chapter 7 presents a spatio-temporal statistical model to characterize the space-time diffusion patterns of cholera. An experimental variogram model explores and characterizes the spatial variability and the extent of contagiousness. Next, this model is integrated with spatial and temporal neighbourhood matrices to map the spatio-temporal transmission network routes of cholera transmission. The space-time diffusion dynamics of cholera is characterized by early disease transmission in populated communities and communities relatively close to primary cases. Likewise, the rate of cholera infection is high in those places.

In conclusion, this thesis shows that the distribution of cholera exhibits a distinct spatial and temporal variation. Such variation is influenced by demographic risk factors like urbanization, overcrowding, migration, sanitation and use of drinking water. Open-space refuse dumps and surface water pollution are important environmental risk factors for cholera transmission. Cholera outbreaks can start from multiple geographic locations that actually have no spatial connection. This thesis recommends that the influence of surface water pollution on cholera can be addressed by preventing faecal contamination of refuse dumps. Moreover, interventions targeting primary case locations and populated communities can effectively impede the spread of the disease.

Samenvatting

De toename van het aantal gevallen van cholera en de toegenomen frequentie, in het bijzonder in landen in Afrika hebben geleid tot een verhoogde zorg om deze ziekte, in het bijzonder betreffende de ruimtelijke en temporele karakteristieken en de onderliggende risicofactoren. Besmetting met cholera vindt vooral plaats via verontreinigd water en voedsel. Demografische en geografische factoren kunnen bewoners echter blootstellen aan besmetting. Socio-economische en milieu factoren zoals de kwaliteit van de leefomgeving kunnen de kwetsbaarheid beïnvloeden. In Ghana heeft de regelmatige terugkeer van cholera mogelijk endemische haarden gegenereerd in stedelijke gemeenschappen die een groter aantal ziektegevallen laten zien tijdens uitbraakperioden. Toch is er weinig tot niets bekend over de ruimtelijke en temporele karakteristieken van de ziekte. Dit proefschrift maakt gebruik van cholera gegevens uit het verleden en ruimtelijk statistische methoden om de effecten van socio-economische en milieurisicofactoren beter te begrijpen. Twee studiegebieden worden gebruikt: de Ashanti regio en haar hoofdstad, de Kumasi metropool.

De hoofdstukken 2 en 3 laten een ruimtelijke cluster analyse zien en beschrijven de ruimtelijke en ruimtelijk temporele patronen van cholera. Hoge cholera incidentie komt voor rond de Kumasi metropool in het centrale gebied van de regio, met een significante Moran's I waarde (0.271, $p < 0.001$). Een Mantel-Haenszel *Chi kwadraat* toets voor de analyse van een trend laat een directe relatie zien tussen cholera en verstedelijking ($\chi^2 = 2995.5$, $p < 0.0001$), overbevolking ($\chi^2 = 1757.2$, $p < 0.0001$) en een inverse relatie tussen cholera en afstand tot de Kumasi metropool ($\chi^2 = 831.38$, $p < 0.0001$). Cholera komt vooral als de meerderheid van de bevolking geen toegang heeft tot goede sanitaire voorzieningen, als deze drinkt uit rivieren, bronnen en poelen en als de interne migratie hoog is.

Hoofdstuk 4 gebuikt de nabijheid tot en de dichtheid van afvalhopen als maten voor milieukwaliteit. Ruimtelijke naijlingsmodellen en ruimtelijke fout modellen worden ontwikkeld en geïmplementeerd om de afhankelijkheid van cholera op afvalhopen in de open lucht vast te stellen. In deze regressiemodellen laat het voorkomen van cholera een directe relatie zien tot de dichtheid tot afvalhopen en een inverse relatie tot de afstand er naar toe. De ruimtelijke autoregressieve coëfficiënten zijn significant voor zowel de ruimtelijke naijlingsmodellen als de ruimtelijke fout modellen. Dit laat de afhankelijkheid van cholera zien op de ruimtelijke interactie tussen leefgemeenschappen en op de aanwezigheid van verborgen factoren.

Hoofdstuk 5 toont een steilste pad analyse waarin gebruik gemaakt wordt van een 3 dimensionale hoogte model en van de positie van afvalhopen om potentiële cholera reservoirs af te bakenen. Door gebruik te maken van de nabijheid van potentiële cholera reservoirs als verklarende variabelen worden statistische modellen ontwikkeld en geïmplementeerd om de effecten van oppervlakte water op cholera te kwantificeren. Hoge cholera cijfers zijn gekoppeld aan deze nabijheid, terwijl significantie van de

autoregressieve regressiecoëfficiënten in de statistische modellen de afhankelijkheid laat zien tussen cholera en leefgemeenschappen en verborgen risicofactoren.

Hoofdstuk 6 combineert de geïdentificeerde risicofactoren samen met andere risicofactoren in een coherent, multivariaat statistisch model. Een Bayesiaanse semi-parametrische regressie benadering wordt gehanteerd dat ons in staat stelt om een integrale analyse uit te voeren naar de niet-lineaire effecten van continue covariabelen, ruimtelijk gestructureerde variatie, ongestructureerde heterogeniteit en van deterministische covariabelen. Het model laat zien dat het risico op cholera hoog is binnen de leefgemeenschappen met sloppenwijken (*posterior mean = 4.06, $p < 0.01$*) en met een hoge bevolkingsdichtheid (*posterior mean = 4.339, $p < 0.01$*). De relatie tussen cholera en de dichtheid van afvalhopen is vrijwel lineair met een toenemende *posterior mean*. De *posterior mean* van de nabijheid van afvalhopen wijk af van lineariteit met een afnemend risico tot ca. 500 m en een lichte toename voor grotere afstanden. Er is sprake van uitgesproken ruimtelijke variatie met een significant risico op cholera in het centrale deel van Kumasi en een significante afname in het zuidoostelijke deel.

Hoofdstuk 7 presenteert een ruimtelijk temporeel statistisch model dat de spatio-temporele diffusie patronen van cholera karakteriseert. Een experimenteel variogram verkent en karakteriseert de ruimtelijke variatie en de mate van ruimtelijke samenhang. Vervolgens wordt het model geïntegreerd met ruimtelijke en temporele omgevingsmatrices om zo de ruimtelijk temporele verspreidingsnetwerken van het overbrengen van cholera te karteren. De ruimtelijk temporele diffusie dynamiek van cholera wordt gekenmerkt door een vroege infectie in dichtbevolkte gemeenschappen en gemeenschappen die dicht in de buurt van de eerste ziektegevallen liggen. Ook is de mate van cholera infectie hoog in deze plaatsen.

Als een conclusie laat dit proefschrift zien dat de verdeling van cholera sterke ruimtelijke en temporele variatie kent. Dergelijke variatie wordt beïnvloed door demografische risicofactoren zoals verstedelijking, overbevolking, migratie en gebruik van drinkwater. Afvalhopen in de open lucht en verontreiniging van oppervlakte water zijn belangrijke milieu risicofactoren. Uitbraken van cholera kunnen vanuit verschillende haarden starten die aanvankelijk geen ruimtelijke samenhang kennen. Dit proefschrift geeft als een aanbeveling dat de invloed van waterverontreiniging op cholera kan worden aangepakt door de afvalhopen niet te gebruiken voor menselijke uitwerpselen. Verder kunnen interventies die de verspreidingshaarden en dichtbevolkte gemeenschappen aanpakken de verspreiding van de ziekte effectief indammen.

1

Introduction

"The first step toward success is taken when you refuse to be a captive of the environment in which you first find yourself"

Caine Mark

1.1 Cholera and *V. cholerae*

Cholera is an acute intestinal infection caused by the water borne bacteria *Vibrio cholerae* O1 or O139 (*V. cholerae*). Infection is mainly through ingestion of contaminated water or food (Kelly, 2001). John Snow (1855) first associated cholera with contaminated drinking water in the 1850s, even before any bacterium was known to exist. Approximately 10^2-10^3 cells are required to cause severe diarrhea and dehydration (Sack et al., 1998; Hornich et al., 1971; Kaper et al., 1995). Ingested cholera *vibrios* from contaminated water or food must pass through the acid stomach before they are able to colonize the upper small intestine. After penetrating the mucus layer, *V. cholerae* colonizes the epithelial lining of the gut, secreting cholera toxin which affects the small intestine.

Clinically, the majority of cholera episodes are characterized by a sudden onset of massive diarrhea and vomiting. This is accompanied by the loss of profuse amounts of protein-free fluid along with electrolytes, bicarbonates and ions (Carpenter, 1971). The resulting dehydration produces tachycardia, hypotension, and vascular collapse, which can lead to sudden death. The diagnosis of cholera is commonly established by isolating the causative organism from the stools of infected individuals. The main mode of treatment is the replacement of electrolyte loss through the intake of a rehydration fluid, i.e. Oral Rehydration Salts (ORS) (Sack et al., 2004). Without prompt treatment, fatality rate can be as high as 50% (WHO, 1993; Sack et al., 2004). With adequate treatment, i.e. intravenous and oral rehydration therapy, supplemented with appropriate antibiotics, the fatality rate can drop to approximately 1.0% (Carpenter, 1996; Mahalanabis, 1992).

In its extreme manifestation, cholera is one of the most rapidly fatal infectious illnesses known. Within 3–4 hours of onset of symptoms, a previously healthy person may become severely dehydrated and if not treated may die within 24 hours (WHO, 2010). The disease is one of the most researched diseases in the world today; nevertheless, it is still an important public health problem despite more than a century of study, especially in developing tropical countries. The disease is currently listed as one of three internationally quarantainable diseases by the World Health Organization (WHO), along with plague and yellow fever (WHO, 2000). The growing number and frequency of major cholera outbreaks, especially in countries on the African continent, have heightened concerns of focusing epidemiological research on the underlying risk factors and the identification of high risk areas.

1.1.1 Biology and ecology of *V. cholerae*

The biology and ecology has been described by many authors (Yamai et al., 1977; Faruque et al., 1998; Ramamurthy et al., 1993; Felsenfeld, 1966; Singleton et al., 1982a, 1982b; Colwell et al., 1977; Carpenter, 1971; Barua et al., 1977; Glass et al., 1985). *V. cholerae* is an aerobic, motile, Gram-negative rod that is shaped like a comma (Hamer and Cash, 1999). Over 200 serogroups of *V. cholerae* have been documented (Yamai et al., 1997). The toxigenic *V. cholerae* serogroups which cause epidemic cholera are the O1 and O139 (Faruque et al., 1998). The two major biotypes of the *V. cholerae* O1 are

the classical and the El Tor (Hamer and Cash, 1999). Both of these biotypes can be further classified into two serotypes: Ogawa and Inaba.

Only the O1 serogroup was known to cause epidemics, until 1992 when a new variant serogroup, designated O139 Bengal, was identified after causing extensive outbreaks in India and Bangladesh (Ramamurthy et al., 1993). The *V. cholerae* O139 Bengal strain is a genetic derivative of the El Tor biotype in which the O1 biosynthetic genes are replaced by the O139 biosynthetic genes. The spread of the O139 serogroup, however, is restricted to Asia (WHO, 2010).

The general assumption by most workers, until the mid 1960, was that *V. cholerae* was an organism whose normal habitat was the human gut and/or intestine, and incapable of surviving for more than a few days outside the gut (Falsenfeld, 1966). Colwell et al. (1977, 1980) first isolated *V. cholerae* from plankton samples, and proposed that *V. cholerae* is ecologically autochthonous in estuarine and coastal waters. *V. cholerae* is now known to be a water-borne bacterium that is a natural inhabitant of estuarine and coastal waters, and survives and multiplies in association with zooplankton and phytoplankton (Colwell and Huq, 1994; Islam et al., 1989, 1990, 1994; Nair et al., 1988; Huq et al., 1983). Survival of *V. cholerae* in the aquatic environment, abundance and expression of virulence factors including cholera toxin (CT), and colonization factors such as the toxin-coregulated pilus (TCP), are strongly influenced by both biotic and abiotic factors. Abiotic factors such as sunlight, pH, temperature, salinity and organic nutrients enhance the growth and multiplication of biotic species such as phytoplankton and zooplanktons, whereas sequestration of CO_2 during photosynthesis of phytoplankton alter the dissolved O_2 and CO_2 contents of the surrounding, which in turn leads to elevated pH in the estuarine (Cockburn, 1960).

1.2.3 Transmission hypothesis

Two routes of cholera transmission have been described, *primary* and *secondary transmission*. Primary transmission occurs through exposure to an environmental reservoir of *V. cholerae* (Hartley et al., 2006) or contaminated water sources regardless of previously infected persons or faecal contamination, and is thus responsible for the beginning of initial outbreaks. Primary transmission is enabled by both micro-and macro-level environmental and climatic factors that affect the seasonal patterns of infection (Islam et al., 1994; Alam et al., 2006; Lipp et al., 2002; Sack et al., 2003; Colwell, 1996; Huq and Colwell, 1996; Islam et al., 1989, 1990a, 1990b, 1999). In Africa and South America where one yearly peak of cholera is often observed, the beginning of the epidemics has been associated with environmental conditions that favour the growth and survival of the bacterium (Codeco, 2001; Glass et al, 1991; Mintz and Guerrant, 2009; Swerdlow et al., 1992). Primary transmitted cases should be scattered in space, occurring almost simultaneously in different areas with no apparent interconnection, and should be located relatively close to water sources (Ruiz-Moreno et al., 2010). Hence, primary transmission appears to play a limited role in the epidemiological process since it does not fully explain the exponential growth of incidences during epidemics.

Secondary transmission occurs via the faecal-oral route, i.e. through exposure to faecally contaminated water sources. This is also known as faecal-oral transmission. Focal-oral transmission provides a mechanism for exhibiting a strong feedback between present and past levels of infection. The importance of faecal-oral transmission in cholera epidemics is also supported by recent time series models of cholera in Bangladesh (Koelle and Pascual, 2004; Koelle et al., 2005). In an epidemic situation, the initial reproduction rate of faecal-oral transmissions is increased by the degree of contamination of the water supply as well as the frequency of contacts with these waters (Codeço, 2001), which is in turn influenced by human dimensions such as local environmental factors, socioeconomic, demographic as well as sanitation conditions. Faecal-oral transmissions reflect a complicated transmission pattern, in that multiple factors may play a role in the spread of the disease. Although cholera control measures that target primary transmission is clearly important from the perspective of the disease persistence (Colwell et al., 2003), the dominant role of faecal-oral transmission (as observed in the studies by Ali et al., 2002a, 2002b; Mugoya et al., 2008; Borroto and Martinez-Piedra, 2000; Ackers et al., 1998; Sasaki et al., 2008; Sur et al., 2005) suggests that the containment of faecal-oral infections may be a viable and useful strategy to control epidemics.

1.3 Socioeconomic and environmental risk factors

Socioeconomic and environmental factors significantly enhance the vulnerability of a population to infection and contribute to the epidemic spread of cholera (Ali et al., 2002a, 2002b; Mugoya et al., 2008; Borroto and Martinez-Piedra, 2000; Ackers et al., 1998; Sasaki et al., 2008; Sur et al., 2005). Socioeconomic and environmental risk factors can also mandate the extent to which the disease will reach epidemic proportions (Miller, 1985; Emch et al., 2008), as well as modulate the size of the epidemic (Pascual et al., 2002, 2006; King et al., 2008; Koelle et al., 2005; Koelle and Pascual, 2004; Ruiz-Moreno et al., 2007; Hartley et al., 2005). Huq et al (2005) have specifically demonstrated links between cholera and environmental variables. Studies by Root (1997) and Siddique et al. (1992) show that increases in population density can strain existing sanitation systems, thus putting people at increased risk of contracting cholera. Ali et al. (2002a, 2002b) have identified proximity to surface water, high population density, and low educational status as the important risk factors of cholera in an endemic area of Bangladesh. Borroto and Martinez-Piedra (2000) have identified poverty, low urbanization, and proximity to coastal areas as the important geographic predictors of cholera in Mexico. In epidemic prone regions like Africa, cholera outbreaks have been linked to multiple environmental and socio-economic sources (Acosta et al., 2001; Shapiro et al., 1999; Hutin et al., 2003; Swerdlow et al., 1997; Tauxe et al., 1998; St Louis et al., 1990; Sinclair et al., 1982; Gunnlaugsson et al., 1998; Birmingham et al., 1997; Dubois et al., 2006; Siddique et al., 1995; Reller et al., 2001).

1.4 The burden of cholera

Throughout history, devastating outbreaks of cholera have resulted in millions of cases and hundreds of thousands of deaths. Cholera remains a major public health problem in developing countries where basic infrastructure to provide safe water and sanitation is lacking. The burden of cholera is characterized by both endemic disease and epidemics. Endemic cholera refers to cholera that recurs in time and place, whereas epidemic cholera denotes cholera that occurs unpredictably with respect to poor sanitation and lack of clean drinking water. Globally, the numbers of cholera cases and deaths have increased steadily since the beginning of the 21st century. From 2004 to 2008, a cumulative total of 838,315 cases were notified to WHO, compared with 676,651 cases between 2000 and 2004, representing a 24% increase in the number of cases (WHO, 2009). The burden of the disease is currently enormous on the African continent. Africa alone accounts for over 90% of worldwide cases and deaths. The seventh pandemic is the first to have established persistent residence on the African continent. The invasion and recurrence in Africa recorded over 2.4 million cases and 120,000 deaths from 1970 to 2005. Between 1999 and 2005, Africa alone reported a total of about 965,612 cases and 26,924 deaths to the WHO (WHO, 2000b, 2001, 2002, 20003, 2004, 2005, 2006); this is approximately 91% of the cases and 97% of the deaths worldwide. The burden of the disease on the African continent, however, is possibly worse than officially reported owing to underreporting, limitations in the surveillance and reporting system, as well as fear of unjustified restrictions on travel and trade (WHO, 2000a).

West Africa, and for that matter Ghana, had its share of epidemic cholera during the seventh pandemic which began in 1961 in Indonesia (Barua, 1972; Cvjetanovic and Barua, 1972; Goodgame and Greenough, 1975; Kustner et al., 1981). The disease has been a public health burden in Ghana since its introduction in the early 1970's (Pobee and Grant, 1970; Kwofie, 1976). Since the 1970's, cholera has persisted Ghana in an epidemic context, recurring approximately every 4-5 years (de Magny et al., 2007). Between 1999 and 2005, a total of 25,636 cases and 620 deaths were officially reported to the WHO by the Ghana Ministry of Health (WHO, 2000b, 2001, 2002, 2003, 2004, 2005, and 2006). The recurrence of cholera in Ghana has raised possible endemic foci in urban communities which seem to report greater percentage of cases during every outbreak period. Environmental and socioeconomic factors seem to have ensured the persistent recurrence of cholera in Ghana, inducing social and economic trauma on affected inhabitants.

Cholera diffuses rapidly in environments that lack basic infrastructure with regard to access to safe water and proper sanitation. The cholera *vibrios* can survive and multiply outside the human body and can spread rapidly in environments where living conditions are overcrowded and where there is no safe disposal of solid waste, liquid waste, and human faeces (Ali et al., 2002a, 2002b). Provision of good sanitary conditions, sewage treatment, and provision of clean water have long been known as important critical measures for prevention and eradication. These measures have eliminated cholera from industrialized and developed countries (Griffith et al., 2006). Chronic poverty and lack of awareness in developing countries make implementation of these measures almost unfeasible. Alternatively, in areas where large and prolonged outbreaks of cholera

occur, the WHO suggests the consideration of vaccination as an additional control measure. However, a thorough investigation of the current and historical epidemiological situation, and clear identification of high risk geographical areas to be targeted is a prerequisite since vaccinating an entire population is not warranted (WHO, 2010). A better understanding of the spatial distribution of incidences and associated risk factors, and identification of high risk geographical areas is a non-trivial task which is useful to develop and implement alternative and timely public health interventions to limit or prevent cholera.

1.4 Statistical methods for spatial epidemiology

Spatial epidemiology is the study of the spatial/geographical distribution of disease incidences and its relationship to potential risk factors. Knowledge of the spatial and temporal variations of diseases and characterization of its spatial structure is essential for the epidemiologist to better understand the population's interaction with its environment. The origin of spatial epidemiology dates back to 1855 with the classic epidemiologic studies of John Snow on cholera transmission. Snow's study of London's cholera epidemic provides one of the most famous examples of spatial epidemiology. Mapping the locations of cholera victims, Snow was able to trace the cause of the disease to a contaminated water source. Spatial analysis in the nineteenth and twentieth century mostly took the form of plotting the observed disease cases or rates (see, for example Snow, 1855). Advances in technology now allow not only disease mapping but also the application of spatial statistical methods, such as cluster analysis (Kulldorff et al., 1997; Rosenberg et al., 1999) and ecological analysis (Ali et al., 2001, 2002a, 2002b) in epidemiological research. Geographic Information System (GIS) methods and modern statistical methods allow an integrated approach to address both tasks; i.e. inference on the geographical distribution of the disease and its prediction at new locations.

1.4.1 Cluster analysis

Fundamental to the spatial epidemiologist is the investigation of possible disease clustering. Cluster analysis provides opportunities for the epidemiologist to understand the spatial distribution of diseases and the possible association between demographic and environmental exposures (Besag and Newell, 1991; Kulldorff and Nagarwalla, 1995; Kulldorff et al., 1997; Kulldorff, 2005, 2006). Searching for disease clustering involves an assessment of local or global accumulation of the disease incidences (Lawson et al., 2002; Tango, 2010). The focus of global cluster analysis is to determine the presence or absence of clustering in the whole study region. There are numerous methods for testing global clustering, including those proposed by Alt and Vach (1991), Besag and Newell (1991), Cuzick and Edwards (1990), Diggle and Chetwynd (1991), Grimson (1991), Moran (1950), Tango (1995, 1999, 2000), Walter (1992a, 1992b, 1993) and Whittemore et al. (1987). The most widely used measure of global clustering in epidemiology is the method proposed by Moran (1950). Moran's Index is a weighted correlation coefficient that is used to measure deviation from spatial randomness. Deviation from spatial randomness indicates specific spatial arrangements of

geographic location information such as clusters (Moran, 1950). Although Moran's Index was originally developed to analyze continuous data, its application to analyze count data of health events is enormous (Glavanakov et al., 2001; Perez et al., 2002; Perez et al., 2000; Bellec et al., 2006; Nodtvedt et al., 2007). Other health applications of Moran's Index include studies of Kitron and Kazmierczak (1997) of Lyme disease in the Wisconsin state, studies of Glick (1979) of cancer in Pennsylvania, the geographical distribution of human giardiasis in Ontario, Canada (Odoi et al., 2003), Lyme disease in the New York state (Glavanakov et al., 2001), and the geographical patterns of cholera in Mexico (Borroto and Martinez-Piedra, 2000).

Global cluster analysis can obscure local effects since the assumption of stationarity is rarely met. Local cluster analysis defines the characteristics of the clusters, such as size, location and intensity. Several formal methods and techniques for identifying local disease clusters have been developed for both point and areal level data (Kulldorff and Nagarwalla, 1995; Besag and Newell, 1991). Examples of local clustering methods include spatial correlograms (Isaaks and Srivastava, 1989; Liebhold et al., 1993; Rossi et al., 1992; Weisz et al., 1995; Upton and Fingleton, 1985), the Local Indicator of Spatial Association (Anselin, 1995), the local G_i^* statistics (Getis and Ord, 1992), Ripley's *K*-function (Ripley, 1976, 1981, 1988), Cluster Evaluation Permutation Procedure (CEPP) (Turnbull et al., 1990), the Knox test (1964, 1989), and Kulldorff's spatial scan statistic (Kulldorff, 1997). Other methods for space-time clustering include Mantel's test (1967), Ederer-Meyer-Mantel test (Ederer et al., 1964), Barton's test (Barton, 1965), Diggle et al. test (1995), Jacquez's *k* nearest neighbors test (1996), and Kulldorff's space-time scan statistic (Kulldorff et al., 1998).

The spatial scan statistic developed by Kulldorff (1997, 2005, 2006) offers several advantages over the others: (1) it corrects for multiple comparisons, (2) it adjusts for the heterogeneous population densities among the different areas in the study, (3) it detects and identifies the location of the clusters without prior specification of their suspected location or size thereby overcoming pre-selection bias, (4) and allows adjustment for covariates. Also Kulldorff's spatial scan statistic is both deterministic (i.e., it identifies the locations of clustering) and inferential (i.e., it allows for hypothesis testing and evaluation of significance). The spatial scan statistic has been used to detect and evaluate various disease clusters including leukaemia (Kulldorff and Nagarwalla, 1995; Hjalmars et al., 1996), cancer (Kulldorff et al., 1997; Michelozzi et al., 2002; Viel et al., 2000; Sheehan and DeChelo, 2005; Hjalmars et al., 1996; Turnbull et al., 1990, Kulldorff et al., 1998), giardiasis (Odoi et al., 2004), tuberculosis (Tiwari et al., 2006), diabetes (Green et al., 2003), Creutzfeldt-Jacob disease (Cousens, 2001), granulocytic ehrlichiosis (Chaput et al., 2002), and amyotrophic lateral sclerosis (Sabel et al., 2003).

The flexible spatial scan statistic is a recent cluster detection methodology developed by Takahashi and Tango (2005). This approach is based on the original idea of Kulldorff. Unlike Kulldorff's approach, however, which imposes a circular window to define the potential cluster areas (Kulldorff and Nagarwalla, 1995), Takahashi and Tango's (2005) flexible spatial scan statistic imposes an irregularly shaped window on each region connecting its adjacent regions. This approach is able to detect arbitrarily shaped

clusters, and this statistic is well suited for detecting and monitoring disease outbreaks in irregularly shaped areas.

1.4.2 Ecological analysis

A significant interest in spatial epidemiology lies in identifying associated risk factors which enhance the risk of infection, the so called *ecological analysis* (Lawson et al., 1999; Lawson, 2001) or *geographic correlations studies* (Elliott et al., 2000). The term ecological analysis is used loosely here to denote associating aggregated disease outcomes with related risk factors or covariates, where inference still remains at the aggregated level. The most prominent method is the classical linear regression model, where the response variable y is assumed to be independent normal or Gaussian distributed and covariates, say $x_1,...,x_p$ act linearly on the response. By assumption, the conditional expectation of y is:

$$\eta = E\left(y\middle|x_1,...,x_p\right) = \beta_0 + x_1\beta_1 + ... + x_p\beta_p\,, \qquad 1.1$$

where the regression coefficients $\beta_1,...,\beta_p$ determine the strength of the influences of the covariates, and the linear predictor η is the sum of the covariate effects. Here, each observation has an underlying mean of $\sum_i x_i\beta_i$ and normally distributed random error term ε. Generally, the random error term $\varepsilon = \left(\varepsilon_1,...\varepsilon_p\right)$ has zero mean and uncorrelated variance-covariance matrix Σ_σ, i.e. $\varepsilon_i \sim N\left(0,\Sigma_\sigma\right)$, where $\Sigma_\sigma = \mathrm{Var}\left(y\right) = \mathrm{Var}\left(\varepsilon\right) = \sigma^2 I$, and I is $p \times p$ identity matrix. The assumption of independent observations also implies that $E\left(\varepsilon_i\varepsilon_j\right) = E\left(\varepsilon_i\right)E\left(\varepsilon_j\right) = 0$

For disease counts of small areas with relatively small populations at risk and few observed cases, rates may not follow the assumptions of the linear model. In such cases, a direct connection between the expectation of y and the linear predictor η is not possible. Generalized linear models (GLMs) extend the classical linear model for Gaussian responses to more general situations such as binary or count data (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Fahrmeir and Tutz, 2001; McCulloch and Searle, 2001) to ensure the appropriate domain of $E\left(y\middle|x_1,...,x_p\right)$. By introducing a more general transformation or response function *h*, equation (1.1) can be rewritten as:

$$h\left(\eta\right) = E\left(y\middle|x_1,...,x_p\right) = h\left(\beta_0 + \beta_1 x_1 + ... + x_p\beta_p\right). \qquad 1.2$$

Both the classical linear model and GLMs provide the means to quantify and describe only *first-order* effects or *large-scale* variation in the mean of the disease outcome. These methods ignore *second-order* spatial effects or *small-scale* variations that arise from interactions between neighbors, i.e. spatial autocorrelation. Both methods assume

that any spatial pattern observed in the outcome y is entirely due to the spatial patterns in the covariates; therefore, no residual spatial variation is accounted for. If an important covariate is inadvertently omitted, however, estimates of β will be biased (e.g. Draper and Smith, 1998), and if this covariate varies spatially, residual spatial variation will often manifest itself as spatial autocorrelation in the residual process. Hence when these methods are used to analyze spatially correlated data, the standard error of the covariate parameters is underestimated and thus the statistical significance is overestimated (Cressie, 1993).

Spatial statistical methods, such as spatial regression, incorporate spatial autocorrelation according to the way spatial neighbors are defined. A spatial regression model may be parameterized as equation (1.1). A modification of the variance-covariance matrix Σ is then required to allow spatially correlated error terms. Common methods to incorporate spatially correlated error terms in the variance-covariance matrix Σ_σ is the simultaneous spatial autoregressive (SAR), conditional spatial autoregressive models (CAR), and spatial moving average models (SMA). Both the SAR and CAR correspond to autoregressive procedures in time series analysis (Ripley, 1981). These models are well explained in Cliff and Ord (1981), Haining (1990), Ripley (1981), and Cressie (1993).

Generalized additive models (GAM) provide a powerful class of models for modelling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. Modelling the nonlinear effects of continuous covariates may be based on smoothing splines (e.g. Hastie and Tibshirani, 1990), local polynomials (e.g. Fan and Gijbels, 1996), regression splines with adaptive knot selection (e.g. Friedman and Silverman 1989, Friedman 1991, Stone et al., 1997) and P-splines (Eilers and Marx, 1996; Marx and Eilers, 1998).

Bayesian estimation and inference in statistical modelling provides a number of advantages over the classical approaches. This includes a more natural interpretation of parameter intervals, and the ease with which the true parameter density may be obtained. Bayesian approach has recently been given intense focus due to the widespread adoption of Markov Chain Monte Carlo (MCMC) methods. In the past, Bayesian estimation and inference was often daunting due to the requirement of numerical integration. The MCMC estimation method decomposes complicated estimation problems into simpler problems that rely on conditional distributions for each parameter in the model (Gelfand and Smith, 1990). In classical approaches such as maximum likelihood estimation, inference is based on the likelihood of the data alone. In Bayesian approach, the likelihood of the observed data y given a d dimensional parameter set $\theta = (\theta_1, ..., \theta_d)$, denoted as $p(y|\theta)$, is used to modify the prior beliefs $p(\theta)$ with the updated knowledge summarized in a posterior density $p(\theta|y)$. Applying Bayes theorem, $p(\theta|y) = p(y|\theta)\,p(\theta)/p(y)$ is found, where the marginal likelihood $p(y)$ is obtained by integrating the likelihood over the prior densities,

i.e. $p(y) = \int p(y|\theta)\, p(\theta)\, d(\theta)$. Since $p(y)$ can be regarded as a normalizing constant, the posterior density can be simplified as $p(\theta|y) \propto p(y|\theta)\, p(\theta)$.

Much literature has been developed around methodological issues relating to the Bayesian approach (Manton et al., 1981; Tsutakawa, 1988; Baseg et al., 1991; Clayton and Kaldor, 1987; Clayton and Bernardinelli, 1992; Spiegelhalter et al., 2002; Lawson et al., 2003; Lawson, 2008). Bayesian approaches to GAM are currently either based on regression splines with adaptive knot selection (e.g. Smith and Kohn 1996, 1997; Denison et al., 1998; Biller, 2000; Biller and Fahrmeir, 2001; Di Matteo et al., 2001; Hansen and Kooperberg, 2002), or on smoothness priors (Hastie and Tibshirani 2000, Fahrmeir and Lang, 2001a, Fahrmeir and Lang, 2001b). Fahrmeir et al. (2004), Brezger (2004) and Kneib (2005) present a detailed description of Bayesian P-Splines and mixed model based inference in generalized structured additive regression (STAR) based on Bayesian P-Splines. Generalized STAR models are extensions of GAM models which allow one to incorporate small area spatial effects, nonlinear effects of risk factors, and the usual linear or fixed effects in a joint model. Typically, a generalized STAR model is parameterized as:

$$\eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + f_{spat}(s_i) + u_i'\gamma, \qquad 1.3$$

where $f_1, ..., f_p$ are nonlinear functions of the covariates $x_1, ..., x_p$. In such models, covariates of the parametric or fixed effects are subsumed in the term $u_i'\gamma$, where γ is an estimate of the fixed effect covariate u_i. The linear combination $u_i'\gamma$ corresponds to the usual parametric part of the predictor. The function $f_{spat}(s_i)$ accounts for spatial effects of the data.

Such STAR models are highly parameterized; therefore, inference is based on a fully Bayesian estimation of the posterior distribution of the model parameters. Since the posterior is analytically intractable, the parameter estimates are generated by drawing random samples from the posterior via MCMC simulation techniques. The unknown functions $f_1, ..., f_p$, $f_{spat}(s)$ and the fixed effects γ are considered as random variables and must be supplemented by appropriate prior assumptions. In the absence of any prior knowledge, diffuse prior $p(\gamma) \propto const$ may be assigned for the fixed effects. Alternatively, a weakly informative multivariate Gaussian distribution may be assigned. For modelling the unknown functions $f_1, ..., f_p$, there exists a variety of different approaches. Polynomials of degree *l* are often not flexible enough for small *l*, yet estimates become more flexible but also rather unstable for large *l*, especially at the boundaries (Brezger, 2004). Eilers and Marx (1996) suggest specific forms of polynomial regression splines which are parameterized in terms of B-spline basis functions together with a penalization of adjacent parameters, also known as P-splines. Following Eilers and Marx (1996), $f(x)$ can be approximated by a polynomial spline

of degree l with equally spaced knots $x_j^{\min} = \zeta_{j,0} < \zeta_{j,1} < \cdots < \zeta_{j,s-1} < \zeta_{j,s} = x_j^{\max}$ within the domain of x_j. The assumption that $f(x)$ can be approximated by a polynomial spline leads to a representation in terms of a linear combination of $d = s + l$ basic functions B_m, i.e. $f_j(x_j) = \sum_{m=1}^{d} \xi_{j,m} B_m(x_j)$. Thus, the estimation of $f(x)$ is reduced to the estimation of the vector of unknown regression coefficients $\xi = (\xi_1, ..., \xi_m)'$ from the data. Detailed description of Bayesian P-Splines in STAR models can be found in Brezger (2004).

The spatial effect is commonly introduced in a hierarchical fashion via prior distributions of location-specific random effects. Unlike the SAR, CAR, or SMA models, spatial dependencies are estimated for each spatial unit. A major significance of STAR modelling approach is that the spatial effect can be split into spatially structured (correlated) and a spatially unstructured (uncorrelated) effects. Thus, $f_{spat}(s) = f_{str}(s) + f_{unstr}(s)$ where the function $f_{str}(s)$ accounts for spatially correlated effects of the data, whereas the function $f_{unstr}(s)$ accounts for unobserved heterogeneity, occurring locally or at a large scale. The most common prior for modelling the structured spatial effects $f_{str}(s)$ is the Markov random field prior pioneered by Besag (1974, 1975):

$$p\left(f_{str}(s) \middle| f_{str}(s'), s' \neq s, \tau_{str}^2\right) \sim N\left(\frac{1}{N_s} \sum_{s' \sim s} f_{str}(s'), \frac{\tau_{str}^2}{N_s}\right). \qquad 1.4$$

Here $s \in \{1, ..., S\}$ represents the locations of connected geographical regions, N_s is the number of geographical neighbors and $s' \sim s$ denotes that geographical locations s' and s are neighbors. The uncorrelated $f_{unstr}(s)$ part may be estimated based on location-specific Gaussian random effects $p\left(f_{unstr}(s) \middle| \tau_{unstr}^2\right) \sim N\left(0, \tau_{unstr}^2\right)$. In a fully Bayesian estimation, hyper-priors for the variance parameters τ_j^2, $j = 1, ..., p$, τ_{str}^2 and τ_{unstr}^2 are also considered as unknown; therefore, appropriate hyper-parameters have to be assigned. Commonly, highly dispersed, but proper, inverse Gamma distributions $p\left(\tau_j^2\right) \sim IG\left(a_j, b_j\right)$ with known hyper-parameters a_j and b_j with density function $p\left(\tau_j^2\right) \propto \left(\tau_j^2\right)^{-a_j - 1} \exp\left(-b_j / \tau_j^2\right)$ are assigned in the second stage of the hierarchy.

Different forms of STAR models may be structured for both cross-sectional and longitudinal data. Well known models that can be structured include GAM, generalized additive mixed models (GAMM), spatial regression models, generalized geoadditive mixed models (GGAMM), dynamic models, varying coefficient models, and geographically weighted regression (Fotheringham et al., 2002) may be useful within a

unifying framework. Detailed description of these models and their applications can be found in Fahrmeir and Lang (2001a), Fahrmeir and Lang (2001b), Lang and Brezger (2004), Brezger and Lang (2003), Eilers and Marx (1996), Marx and Eilers (1998), Wahba (1978), and Hastie and Tibshirani (2000).

1.5 Spatial epidemiology of cholera

John Snow (1855) first mapped cholera cases together with the location of a water source in London, and showed that contaminated water was the major cause of the disease. Surprisingly, this was done 20 years before Koch and Pasteur established the beginnings of microbiology (Koch, 1884). After Snow's seminal work, most epidemiological studies of cholera have focused on the pathogenesis and biological characteristics of *V. cholerae* (Yamai et al., 1977; Faruque et al., 1998; Ramamurthy et al., 1993; Felsenfeld, 1966; Singleton et al., 1982a, 1982b; Colwell et al., 1977; Colwell, 1981; Carpenter, 1971; Mandel et al., 1995; Barua et al., 1977; Glass et al., 1985, Clement et al., 1989). However useful these studies are, they usually cannot establish accurate individual exposure levels for the critical risk factors of the disease (Haining, 1998). Moreover, such studies cannot indentify high risk geographical areas of the disease. Spatial epidemiological tools applied in cholera studies can facilitate the identification of high risk areas and the formulation of hypotheses about the causal factors responsible for such variations, as well as the optimal allocation of health facilities to improve health care provision.

Borroto and Martinez-Piedra (2000) have used GIS to map cumulative incidence rates of cholera in 32 Mexican states. Chevallier et al. (2004) used cartographic representation of raw cholera incidence rates to study the spatial distribution of cholera in Ecuador. Disease maps produced from raw rates, however, can lead to spurious spatial pattern due to heterogeneous population sizes across spatial units. Disease maps should therefore be based on smoothed estimates, clean of noise and adjusted for variation in population sizes (Lawson, 2001a, 2001b). With respect to evaluating global clustering, Siddiqui et al. (2006) applied Cuzick-Edward's k-Nearest Neighbors test (Cuzick and Edwards, 1990) to evaluate clustering of cholera cases in Pakistan. The well known global measure of spatial clustering or spatial autocorrelation, Moran's Index (Cliff and Ord, 1973), has been used to analyze the global clustering of cholera in Mexico (Borroto and Martinez-Piedra, 2000), in the Lusaka area of Zambia (Sasaki et al., 2008) and in Madras (Ruiz-Moreno et al., 2007).

Understanding the spatial relationship between cholera and risk factors has been a challenge for long. Most ecological studies of cholera make no, or limited, use of the spatial structure of the data. Thus, most studies utilize standard statistical methods that ignore methodological difficulties arising from the nature of the data, i.e. spatial correlation and/or spatial dependency of the data, especially when the population distribution and risk factors are particularly variable and spatially structured. Cholera is primarily driven by environmental and socioeconomic factors (Ali et al., 2001, 2002a, 2002b; Borroto and Martinez-Piedra, 2000), therefore, incidences in close geographical proximity are more likely to be influenced by similar environmental and socioeconomic factors and accordingly affected in a similar way. Similarly, socioeconomic and

environmental characteristics are particularly variable across space; the spatial distribution of cholera can therefore vary substantially between different spatial regions. Ali et al. (2001, 2002a, 2002b) have utilized logistic regression, simple and multiple regression models to study the spatial epidemiology of cholera in an endemic area of Bangladesh. In their study, spatial filtering methods (Talbot et al., 2000), spatial moving averages (Kafadar, 1996), and traditional kriging were only used to remove noise and transform cholera and environmental data into a spatially continuous form. Therefore, the effects of spatial proximity or spatial neighbourhood structure on cholera could not be quantified. Sasaki et al. (2008) investigated risk factors of cholera with a GIS and matched case-case control in a peri-urban area of Lusaka, Zambia. Although a spatial autocorrelation analysis using Moran's Index was found to be statistically significant, this was never incorporated into the logistic and multiple regression models to examine the risk factors of cholera. Mugoya et al. (2005) used logistic regression analysis to investigate the spread of cholera in Kenya. Ackers et al. (1998) have used Pearson correlation coefficient to determine the correlation between cholera incidence rates and socioeconomic and environmental risk factors in Latin America. Kuo and Fukui (2007) used a logarithmic regression model to model the diffusion of cholera in Japan. De Magny et al. (2008) used a GLM with a Poisson distribution and a log-link function to model environmental variables associated with cholera in Bangladesh.

1.6 Aims and objectives

This study focuses on the application of current spatial statistical methods, such as spatial regression and Bayesian STAR models, to study the spatial epidemiology of cholera in Ghana. The main aim is to use past cholera epidemic data and spatial statistical methodologies to study the spatial patterns of cholera, identify territories of high risk, and determine important demographic and environmental factors that increase the risk of infection. Studies on diarrhea related diseases in Ghana so far have focused solely on the biological factors and characteristics of the individuals affected. Knowledge of the spatial distribution and risk factors of cholera is woefully limited. Studying the spatial patterns of cholera and risk factor characteristics will prove useful for health officials and policy makers to develop appropriate and timely health interventions to limit the burden of cholera.

1.6.1 Specific objectives

The specific objectives of this study are to

- explore the spatial, temporal, and demographic patterns of cholera
- investigate the spatial dependency of cholera prevalence on local environmental risk factors
- investigate the space-time diffusion dynamics of cholera

1.7 The study area

1.7.1 Cholera in Ghana

Traditionally Ghana, and, for that matter, West Africa is presumed a virgin territory or a receptive area for cholera. The first bacteriological case report of cholera in Ghana was on 1st September, 1970. A Togolese national in transit at the Kotoka International Airport from Conakry, Guinea, collapsed and was found to have cholera (Pobee and Grant, 1970). However, an outbreak did not begin from then until it was later smuggled into the country (Ashitey, 1994). Later in that year, some Ghanaians went fishing in the waters of Togo, Liberia and Guinea. One of the fishermen died and although a sanitary cordon had been placed on Ghanaian borders, his family smuggled the corpse into the home town, and the usual burial rites were performed. It was after this that cholera began to spread along the shores of Ghana. The disease swept through many coastal villages in epidemic proportions. It kept on spreading and by July 1971, the Ashanti Region began to report cases, indicating that cholera was spreading across the country (Ashitey, 1994). During those days, reported outbreaks were investigated, treatment camps were set up, people were vaccinated against cholera, and the population was also educated on measures to prevent the spread of the disease. All these attempts to prevent cholera from taking root in Ghana failed, however. Since then cholera has remained entrenched, posing a major public health problem in Ghana.

From 1970 to 2005, the Ghana Ministry of Health officially reported a total of 111,406 cases to the WHO. Within this period, four major outbreaks of cholera have occurred. The first outbreak occurred between 1970 and 1972, recording 16,406 cases; the second outbreak between 1980 and 1985, recording 27,489 cases; the third outbreak between 1989 and 1992, recording 5,973 cases; the fourth outbreak between 1998 and 2003, recording 25,494 cases. Usually, heavy rains across the country seem to trigger and foster the epidemics. Although the periods of these outbreaks appear to be irregular, they are always preceded by the rainy season, and subside once the dry season begins.

1.7.2 Case definition of cholera in Ghana

In Ghana, a case definition of cholera is based on the WHO's definition which depends on whether or not the presence of cholera has been demonstrated in the area. According to the WHO (WHO, 1993) guidance on formulation of national policy on the control of cholera, in an area where the disease is not known to be present a case of cholera should be suspected, when a patient, 5 years of age or older develops severe dehydration or dies from acute watery diarrhea, or where an epidemic is occurring, a patient, 5 years of age or older develops acute watery diarrhea, with or without vomiting. First cases of cholera, however, are always confirmed by bacteriological tests (personal communication with the Kumasi Metropolitan Health Director). In this study, only cholera cases made known to the Disease Control Unit (DCU) through reporting facilities such as community volunteers, community clinics, and hospitals were used. In Ghana it is mandatory for all reporting facilities to report cholera cases weekly to the DCU. Although hospitals are scarcely found in many communities in the districts,

almost all communities in the districts have access to clinics, and community volunteers who monitor all communicable diseases. The communicable diseases surveillance network is purposely established from community level to district level to ensure effective surveillance of all communicable diseases (personal communication with head of DCU, Ashanti Region).

1.8 Research framework and methods

In spatial epidemiological studies, the closest link to an assumed biological model is achieved by using data from a variety of geographical points that describe the exact spatial locations of cases/events and exposure factors. As such, the average disease risk of an individual reflects its level of exposure to the risk factors. Due to several limitations in disease surveillance systems in Ghana, individual level data are rarely available. Hence, individual level studies are almost unfeasible in such situations. Case-control and cohort studies can give a relatively close approximation to the biologic model because they both provide point data that describes individual level characteristics. These studies are expensive and time consuming to carry out, and are not feasible in all situations. For example, in retrospective studies of infectious diseases where the recovery period is short, it is impossible to trace the affected individuals. Exploratory studies using aggregated data, such as ecological or geographic correlation studies, offer an alternative approach for analyzing aggregated epidemiological data to address specific hypotheses of disease causation. Although they too are prone to some biases and misclassifications, the so called *ecological bias* (Elliot and Wakefield, 2000), they are easier, quicker, and less expensive to conduct (Elliott and Wartenberg, 2004). Therefore, this study is undertaken within the framework of small area ecological studies where group level characteristics are the focus rather than individual level characteristics.

1.8.1 Datasets and spatial data creation

Previous to the year 2005, cholera surveillance and reporting was ineffective. This is evidenced in the aggregation of cholera cases from 1997 to 2000 across large geographical areas, i.e. district. With intensified surveillance and reporting systems during an outbreak in 2005 in the Kumasi Metropolis, cholera cases were recorded daily for each community. Accordingly, two main classes of cholera datasets are used for this study; (1) yearly cholera cases from 1997 to 2000 for each district in Ashanti Region, and (2) daily cholera cases of 2005 for each community in Kumasi Metropolis. Because these datasets cannot be synchronized, both Ashanti Region and Kumasi Metropolis are utilized as separate case study areas. There are 18 administrative districts in the Ashanti Region including Kumasi Metropolis. The Kumasi Metropolis is the capital of Ashanti Region, and is the only district which has gained a metropolitan status. Detailed description of the geography and demography of Ashanti Region and Kumasi Metropolis are provided in the respective chapters.

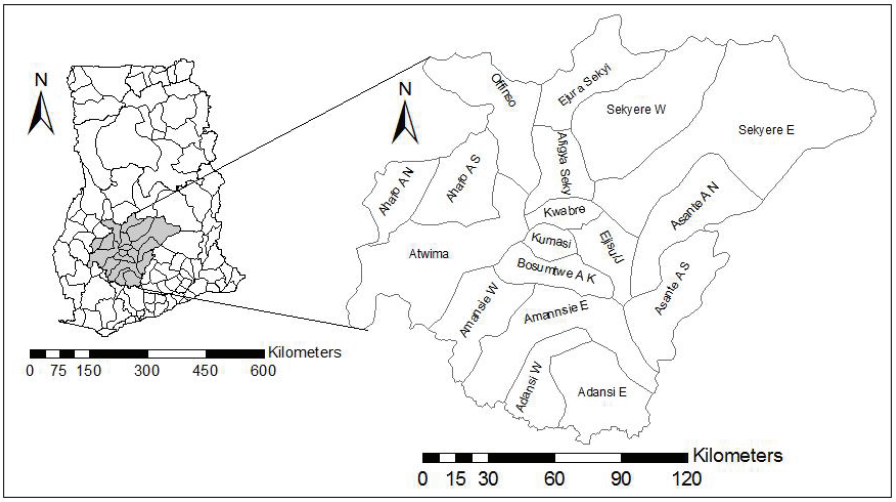

Figure 2.1: Map of Ghana and district map of Ashanti Region

From the topographic map of the Ashanti Region (Figure 1.1), the 18 administrative districts and the main boundary were digitized as polygon features. Reported cases of cholera over the period 1997-2001, obtained from the DCU were entered as attributes of the districts. Population and demographic data obtained from the 2000 Population and Housing Census of Ghana were also entered as attributes. From the topographic map, rivers and streams in Kumasi Metropolis were digitized as line segments, as were elevation contours. Values of elevations contours were input as spatial attributes. Locations of various communities in the Kumasi Metropolis were digitized as point features whereas locations of open-space refuse dumps were mapped during ground survey. Reported cases of cholera during the 2005 epidemic outbreak obtained from the Kumasi Metropolitan Health Directorate (KMHD) were entered as attributes of the communities.

1.8.2 Spatial analyses and statistical modelling

In this thesis, a spatial cluster analysis is conducted to investigate and describe the spatial and spatio-temporal patterns of cholera. A spatial autocorrelation indicator, Moran's Index, is used to describe the spatial patterns of cholera. To detect and map the specific locations of cholera clusters, the spatial scan statistic and flexible scan statistic methods are used. Using proximity to and density of refuse dumps as proxies for environmental sanitation, spatial statistical models are developed and implemented to determine the spatial dependency of cholera on open-space refuse dumps. Specifically, spatial autoregressive coefficients are incorporated in classical regression approaches to develop spatial lag and spatial error models that better explain the spatial dependency of cholera on open-space refuse dumps. Further, a GIS based steepest downhill path analysis is conducted using a 3D elevation model created from the elevation data to delineate water bodies which surface runoffs from refuse dumps will flow into, designated as *potential cholera reservoirs*. Thus, the study assumes that surface runoffs

from refuse dumps are the major sources of surface water pollution. Similarly, using proximity to the *potential cholera reservoirs* as explanatory variables, statistical models are developed and implemented to assess the effects of surface water pollution on cholera. Likewise, spatial autoregressive coefficients are incorporated in classical regression models to develop spatial lag and spatial error models that better explain the spatial dependency of cholera on potential cholera reservoirs. In order to bring all the hypothesized risk factors of cholera (i.e. proximity to and density of refuse dumps, and proximity to potential cholera reservoirs) and other known risk factors (i.e. slum settlements and population density), into coherence, Bayesian STAR models are developed and implemented to assess their mutual influences on cholera. Finally, Bayesian STAR approaches are used to develop a diffusion model to investigate and explore the space-time diffusion dynamics of cholera.

1.9 Outline of thesis

This thesis is made up of collection scientific papers, some of which have been published in peer reviewed journals and others under review. This structure allows spreading of the contents of the thesis to a wider audience. The structure and contents used in submitting the manuscripts have largely been retained. Therefore, overlaps and repetitions may occur between individual chapters.

Chapter 2 focuses on the application of a GIS based spatial analysis and statistical methods to study the spatial patterns of cholera, identify territories of high risk, and determine demographic risk factors that contribute to high rates of cholera.

Chapter 3 investigates the spatial and temporal clusters of cholera in the Ashanti Region using the *spatial scan statistics*. Correlation analyses of cholera rates with demographic factors are also explored to assess the extent to which these factors might explain high rate clusters of cholera.

chapter 4 the determines whether cholera prevalence is related to proximity and density of refuse dumps in Kumasi, detects and maps spatial clusters of cholera, and determine whether refuse dumps are a contributory factor to high rate cholera clusters and, determines a critical buffer distance within which refuse dumps should not be located away from communities.

Chapter 5 aims to determine (a) the impacts of surface water contamination on cholera infection and (b) detect and map arbitrary shaped clusters of cholera.

Chapter 6 develops a multivariate explanatory model that combines all the identified risk factors and other possible socioeconomic risk factors, both in a linear and a nonlinear way.

Chapter 7 analyzes the joint effects of primary and secondary transmission in the space-time diffusion dynamics of cholera. Specifically, Chapter 7 seeks to (1) define and map the transmission network routes of cholera diffusion from possible multiple primary

cases and (2) model the joint effects of population density and proximity to primary cases on the space-time dynamics of cholera diffusion.

Chapter 8 finally presents a summary of the results and conclusions of the preceding chapters, and offers recommendations for further research.

2

Spatial and demographic patterns of cholera

"The interpretation of our reality through patterns not our own, serves only to make us ever more unknown, ever less free, ever more solitary"

Gabriel Garcia Marquez

"Alexander received more bravery of mind by the pattern of Achilles, than by hearing the definition of fortitude."

Sir Philip Sidney

This chapter focuses on the application of a GIS based spatial analyses and statistical technology to study the spatial patterns of cholera distribution in Ashanti Region-Ghana. The main objectives are to describe the spatial patterns of cholera, identify territories of high risk, and determine demographic risk factors that contribute to high rates of cholera. We utilize Moran's Index to describe the spatial patterns of cholera. Moran's Index is the most widely used measure of global clustering in epidemiology. Since disease mapping is useful for initial exploration of relationships between exposure and the disease, empirical Bayesian smoothing techniques are utilized to map smooth rates of cholera. In a further analysis, a Mantel-Haenszel *Chi square* for trend analysis is utilized to determine the effects of some demographic risk factors on cholera prevalence. This chapter has originally been published as: Osei FB and Duker AA: Spatial and demographic patterns of Cholera in Ashanti Region-Ghana. *International Journal of Health Geographics* 2008, 7:44

Abstract

Cholera has claimed many lives throughout history and it continues to be a global threat, especially in countries in Africa. The disease is listed as one of three internationally quarantainable diseases by the World Health organization, along with plague and yellow fever. Between 1999 and 2005, Africa alone accounted for about 90% of over 1 million reported cholera cases worldwide. In Ghana, there have been over 27000 reported cases since 1999. In one of the affected regions in Ghana, Ashanti Region, massive outbreaks and high incidences of cholera have predominated in urban and overcrowded communities. A GIS based spatial analysis and statistical analysis, carried out to determine clustering of cholera, showed that high cholera rates are clustered around Kumasi Metropolis (the central part of the region) , with Moran's Index = 0.271 and $p < 0.01$. Furthermore, A Mantel-Haenszel *Chi square* for trend analysis reflected a direct spatial relationship between cholera and urbanization ($\chi^2 = 2995.5$, $p < 0.01$), overcrowding ($\chi^2 = 1757.2$, $p < 0.01$), and an inverse relationship between cholera and order of neighbourhood with Kumasi Metropolis ($\chi^2 = 831.38$, $p < 0.01$). The results suggest that high urbanization, high overcrowding, and neighbourhood with Kumasi Metropolis are the most important predictors of cholera in Ashanti Region.

2.1 Introduction

Cholera has claimed many lives throughout history and it continues to be a global threat (Mouriño-Pérez, 1998), especially in countries in Africa. Between 1999 and 2005, there were over 1million reported cholera cases and over 28,000 reported deaths worldwide. Africa alone accounted for about 90% of the cases and 96% of the deaths worldwide (WHO, 2000a, 2001, 2002, 2003, 2004, 2005, 2006). Cholera has gained both global and public health attention due to its mode of transmission and severity. For instance it has become one of the most researched communicable diseases. The disease is also listed as one of three internationally quarantainable diseases by the World Health Organization (WHO), along with plague and yellow fever (WHO, 2000b). In addition to human suffering and lives loss, cholera outbreak causes panic, disrupts the social and economic structure and can impede development in the affected communities (WHO, 2005).

Cholera reached West Africa and Ghana during the seventh pandemic (Barua D, 1972; Cvjetanovic and Barua, 1972; Goodgame and Greenough, 1975; Küstner et al., 1981). The disease has been endemic in Ghana since its introduction in the 1970's (Pobee and Grant, 1972). From 1999 to 2005, a total of 26,924 cases and 620 deaths were officially reported to the WHO (WHO, 2000a, 2001, 2002, 2003, 2004, 2005, 2006). Although the disease is transmitted mainly through contaminated water and food, several demographic and geographic factors can predispose an individual or groups of individuals to infection. For example, increase in population density can strain existing sanitation systems, thus putting people at increased risk of contracting cholera (Root, 1997; Siddique et al., 1992). Once the bacterial, *V. cholerae*, are present in water in sufficient dose, an outbreak can trigger and propagate depending on demographic factors such as population density (Ali et al., 2002a, 2002b), urbanization, and overcrowding (Borroto and Martinez-Piedra, 2000). In developing countries like Ghana, high incidence of cholera seems to predominate in the urban communities, and this is primarily due to high overcrowding and unsanitary living conditions in urban communities. While cholera is prevalent in low urban communities in certain geographical areas like Mexico (Borroto and Martinez-Piedra, 2000), the disease has predominated in urban and overcrowded communities in Ghana. Intermittent water supply coupled with indiscriminate sanitation practices in urban communities in Ghana puts inhabitants at risk of contracting cholera.

Studies on diarrhea related diseases in Ghana (for example Obiri-Danso et al., 2005) so far have focused solely on the biological factors and characteristics of the individuals affected. Although such studies are very useful, they omit the spatial and regional variations of the critical risk factors. Such studies also fail to define territories at high risk. Since health levels vary substantially between different regions, it is essential to characterize these regional variations and identify areas with an accumulation of health problems for epidemiological research, and to allow appropriate public health policy decisions (Cromley and Mclafferty, 2002; Rosenberg et al., 1999). Advances in Geographical Information Systems (GIS) technology provide new opportunities for environmental epidemiologist to study associations between demographic and environmental exposures and the spatial distribution of diseases (Vine et al., 1999). GIS

has been used in the surveillance and monitoring of vector-borne diseases (Beck et al., 1994; Glass et al., 1995), water-borne diseases (Clarke et al., 1991), in environmental health (Braddock et al., 1994; Barnes and Peck, 1994; Wartenberg et al., 1993), analysis of disease policy and planning (Roger and Williams, 1993). Several cholera studies (Ali et al., 2002a, 2002b; Borroto and Martinez-Piedra, 2000; Glass et al., 1982; Kwofie, 1976; Ackers et al., 1998; Fleming et al., 2007) have also employed GIS technologies. This study focuses on the application of a GIS based spatial analyses and statistical technology to study the spatial patterns of cholera, identify territories of high risk, and determine demographic risk factors that contribute to high rates of cholera. No study so far has looked at the spatial patterns of cholera in Ghana. There is therefore no information and/or knowledge about its spatial patterns and demographic correlates in Ghana. Studying the spatial and demographic patterns of cholera in Ghana will prove useful for health officials and policymakers to make appropriate planning and resource allocation.

2.2 The study area

2.2.1 History of Cholera in Ghana

On 1st September, 1970 a Togolese national in transit at the Kotoka International Airport from Conakry, Guinea, collapsed and was found to have cholera (Pobee and Grant, 1970). This was the announcement of the arrival of the seventh pandemic of cholera in Ghana. However, an outbreak did not begin from then until it was smuggled into the country through fishing (Ashitey, 1994). At that time, some Ghanaians went for fishing in the waters of Togo, Liberia and Guinea. One of the fishermen died and although a sanitary cordon had been placed on our borders, his family smuggled the corpse into his home town, and the usual burial rites were performed. It was after this that cholera began to spread along the shores of Ghana. The disease swept through many coastal villages in epidemic proportions. It kept on spreading and by July 1971, Ashanti Region began to report cases, indicating that cholera was spreading across the country (Ashitey, 1994). During those periods, reported outbreaks were investigated, treatment camps were set, people were vaccinated against cholera, and the population was also educated on measures to prevent the spread of the disease. However, all these attempts to prevent cholera from taking root in Ghana failed. Since then cholera has existed in both epidemic and endemic forms in Ghana.

2.2.2 Location and demography of the study area

The Ashanti Region is centrally located in the middle belt of Ghana. It lies between longitudes 0° 9'W and 2° 15'W, and latitudes 5° 30'N and 7° 27'N. The region shares boundaries with four of the ten political regions, Brong-Ahafo region in the north, Eastern region in the east, Central region in the south and Western region in the South-west (See Figure 2.1). Ashanti Region occupies a total land area of 24,389 km^2 representing 10.2% of the total land area of Ghana. Ashanti Region has a population density of 148.1 persons per km^2, compared with a national average of about 80 persons per km^2. The region consists of 18 administrative districts. Kumasi, which is the capital,

is the most populous district, and the only district that has gained a metropolitan status in the region. The 2000 census recorded the region's population as about 3.5 million people, representing 19.1 per cent of the country's population. The urban population (51.3%) in the region exceeds that of the rural population (48.7%). The region is currently the second most urbanized in the country after Greater Accra (87.7%), the national capital. The housing stock in the region is 329,478, of which about 37% are in urban areas and 63% in rural areas. This is in contrast to the 17.4% of houses in urban, and 82.6% in rural areas in 1970. The total stock also represents an increase of 86.8% over the stock in 1984. The relative increase in the proportion of urban housing is a reflection of the increase in urbanization, and perhaps overcrowding. Due to the high housing cost within the urban districts in the region, lots of slummy and/or squatter settlements are created. However, such areas have poor sanitation systems, and perceived to be niches where cholera outbreaks begin.

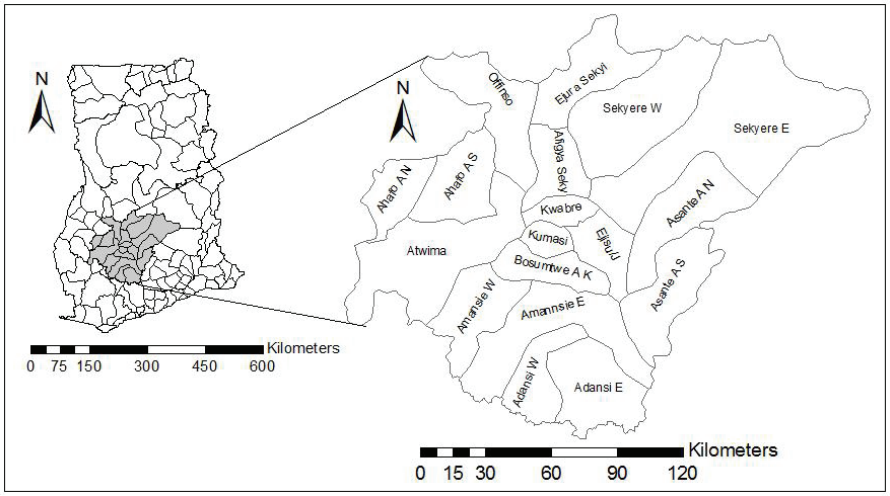

Figure 2.1: District map of Ashanti Region, Ghana

2.2.3 Case definition of cholera

In Ghana, a case definition of cholera is based on the WHO's definition which depends on whether or not the presence of cholera has been demonstrated in the area. According to the WHO (WHO, 1993) guidance on formulation of national policy on the control of cholera, in an area where the disease is not known to be present a case of cholera should be suspected, when a patient, 5 years of age or older develops severe dehydration or dies from acute watery diarrhea, or where an epidemic is occurring, a patient, 5 years of age or older develops acute watery diarrhea, with or without vomiting. The first case of cholera however was confirmed by bacteriological tests (personal communication with KMHD director). In this study, only cholera cases made known to the Disease Control Unit (DCU) through reporting facilities such as community volunteers, community clinics, and hospitals were used. In Ghana it is mandatory for all reporting facilities (i.e. hospitals, clinics, and community volunteers) to report weekly cholera cases to the

DCU. Although hospitals are scarcely found in many communities in the districts, almost all communities in the districts have access to clinics, and community volunteers who monitor all communicable diseases. The communicable diseases surveillance network is purposely established from community level to district level to ensure effective surveillance of all communicable diseases (personal communication with head of DCU, Ashanti Region).

2.3 Methods

2.3.1 Research methodology

An important part of health-needs assessment is the identification of high risk areas for a disease because understanding the characteristics of high risk areas may indicate what is needed to improve health care provision (Haining, 1996). Several disease clustering techniques have been developed to define territories of high risk (Myaux et al., 1997; Kulldorff and Nagerwalla, 1995; Besag and Newell, 1991). However, using clustering detection technique to define high risk territories is only an exploratory technique to locate clusters, but does not establish a relation between the disease and risk factors. In this study, Moran's Index for spatial autocorrelation was computed to ascertain evidence of cholera clustering. A global Bayesian smoothing technique was employed to smooth the crude rates of cholera, and then mapped to determine the spatial distribution of cholera. The districts in the region were classified into strata of districts based on demographic indicators. Population-based incidence rate ratios were then computed for each stratum to determine territories of high risk. The Extended Mantel-Haenszel *Chi Square* for trend analyses and associated *p-values* (one degree of freedom) (Schlesselman, 1982) were also computed to determine the trend between the demographic factors and *V. cholerae* infection.

2.3.2 Spatial data preparation and cartographic display

Topographic map of the study area at a scale of 1:2500 obtained from the planning unit of the Kumasi Metropolitan Assembly was digitized. Before digitizing, the map was georeferenced (by defining the X and Y coordinates of corner points of the map). The main boundary and the 18 districts within the study area were digitized as polygon features. Reported cases of cholera over the period 1997-2001, obtained from the DCU, Ashanti Region, were entered as attributes of the districts. The cumulative incidence rates of cholera were calculated for each of the 18 districts by including all cases over the period 1997-2001. The population database was obtained from the 2000 Population and Housing Census of Ghana (PHC, 2000). This census was conducted by the National Statistical Service of Ghana.

Disease mapping is useful for initial exploration of relationships between exposure and the disease. The raw cumulative incidence rates were smoothed using global Empirical Bayesian Smoothing (EBS) technique. This was to get rid of variance instability as result of heterogeneity in cholera cases and population data (small number problem).

The EBS technique consists of computing a weighted average between the raw rate for each district and the regional average, with weights proportional to the underlying population at risk (Anselin, 2005). In effect, districts with relatively small population will tend to have their rates adjusted considerably, whereas for districts with relatively large population the rates will barely change. The resulting smoothed rates were then mapped using GIS. The cut-off points for classification were based on the Jenk's (1977) optimal classification technique. This classification technique minimizes the total within group variation and is based solely on the statistical distribution of the variable to be classified.

2.3.3 Spatial autocorrelation analyses

In this study, spatial autocorrelation statistic was used to measure the correlation among neighbouring observations in a pattern and the levels of spatial clustering among neighbouring districts (Boots and Getis, 1998).

Global Moran's Index I_{M} statistic, which is similar to the Pearson correlation coefficient (Moran, 1950; Cliff and Ord, 1973), was calculated as:

$$I_{\mathrm{M}} = \frac{N_{(\mathrm{Dist})}}{S_o} \frac{\sum_i \sum_j \left(Chol_{(\mathrm{R})i} - \overline{Chol}_{(\mathrm{R})} \right) \cdot Dist_{ij} \cdot \left(Chol_{(\mathrm{R})j} - \overline{Chol}_{(\mathrm{R})} \right)}{\sum_i \left(Chol_{(\mathrm{R})i} - \overline{Chol}_{(\mathrm{R})} \right)^2}, \qquad 2.1$$

where $N_{(\mathrm{Dist})}$ is the number of districts, $Dist_{ij}$ is the element in the spatial weights matrix corresponding to the district pairs i, j; and $Chol_{(\mathrm{R})i}$ and $Chol_{(\mathrm{R})j}$ are cholera incidence rates for districts i and j with mean cholera incidence $\overline{Chol}_{(\mathrm{R})}$. Since the weights are not row-standardized, the scaling factor $N_{(\mathrm{Dist})} / S_o$ is applied, such that $S_o = \sum_i \sum_j Dist_{ij}$.

The first step in the analysis of spatial autocorrelation is to construct a spatial connectivity matrix or spatial weights matrix that contains information on the neighbourhood structure for each location. The (*i, j*)th element of the matrix $Dist$, denoted $Dist_{ij}$, quantifies the spatial dependence between districts i and j, and collectively, the $Dist_{ij}$ define the neighbourhood structure over the entire area. A first-order rook continuity weight matrix was constructed according to districts who share common boundaries. Thus,

$$Dist_{ij} = \begin{cases} 1 & \text{if districts } i \text{ and } j \text{ share a common boundary} \\ 0 & \text{otherwise} \end{cases}$$

A significance test against the null hypothesis of no spatial autocorrelation through a permutation procedure of 999 Monte Carlo replications was used to test for the significance of the statistic. Moran's Index was calculated using cholera data for the years 1998, 1999, 2001 and 1997-2001. Because very few cholera cases were reported in 1997 and 2000, autocorrelation was not calculated for these years.

2.3.4 Trend Analyses

Population based rate ratios were computed for strata of districts grouped by the following variables:

Proximity to Kumasi, the most urbanized city

Proximity to Kumasi Metropolis (the most urbanized district in Ashanti Region) was categorized into three strata based on the order of spatial neighbourhood and/or adjacency. First-order neighbors were defined as districts sharing common boundaries with the Kumasi Metropolis. Second-order neighbors were defined as districts sharing common boundaries with the first-order neighbors, whereas third-order neighbors were defined as district sharing common boundaries with the second-order neighbors. Population based rate ratios were computed for each stratum by taken the stratum of third-order neighbors as reference (baseline).

Urbanization level

The indicator for urbanization was population based. Each district within the region is made up of localities and/or communities. A locality with a population of 5,000 or more was classified as urban, and less than 5,000 as rural. This is the criteria given by the Ghana Statistical Service (PHC, 2000). The urbanization level for each district was then computed as the proportion of a district's population residing in localities and/or communities of $\geq$ 5,000 people in the year 2000. Three urbanization strata were determined, each representing a quartile of districts. Each quartile was composed of six districts. Population-based incidence rate ratios were calculated for each stratum by taking that of lower urbanization as reference.

Overcrowding

The indicator for overcrowding was based on four variables: (1) Population density; (2) population per house; (3) single room occupancy (i.e., percentage of households living in single rooms) and; (4) households per house. A household was defined as "a person or group of persons who live together in the same house or compound, sharing the same house-keeping arrangements and are catered for as one unit". Each variable was standardized to have a mean of zero and a standard deviation of one. The variables were combined to form a single index of risk, called overcrowding index (OI). The OI for each district was computed as the mean of the algebraic sum of the standardized values of the four variables. The assumption was that the variables carry equal weights. The

Jenk's (1977) method of classification was used to classify OI into three strata of districts. Population-based incidence rate ratios were calculated for each stratum by taking that of lower OI as reference (baseline).

Taking into account urbanization stratum and overcrowding level, neighbourhood population-based double stratification analyses were performed to explore whether cholera incidence rate was associated with the order of neighbourhood with Kumasi Metropolis.

The Extended Mantel-Haenszel *Chi Square* (χ^2) for trend analysis and 95% confidence intervals for rate ratios were computed to explore the relationship between cholera incidence rates and the variables under study (Schlesselman, 1982). Given a series of proportions representing increasing or decreasing exposure (risk factor) and numbers of affected and non affected people in each stratum (group), the Mantel-Haenszel's extension tests whether the rates in successive groups increase or decrease when compared to the baseline (reference). The results of such test include the rate ratios of successive exposure levels, and χ^2 and *p-value* (one degree of freedom). A *p-value* less than 0.05 may be taken as reasonable indication of trend in the rates of successive levels compared with the baseline (reference). Rate ratios and 95% confidence intervals, and the χ^2 for trend analysis were computed.

2.4 Results and analyses

The extent to which neighbouring values are correlated was measured using Global Moran's Index. A significance assessment through a permutation procedure was implemented to determine the significance of the computed Moran's Index. There is a positive and statistically significant spatial autocorrelation for cumulative incidence rate of cholera from 1997 to 2001 (I_M = 0.271, p = 0.0009). Moreover, a spatial autocorrelation statistic computed for each of the periods 1998, 1999 and 2001 were statistically significant ($p < 0.05$) for Moran's Index (See Table 2.1). This reflects clustering of high rates of cholera at the central part of the region (See Figure 2.2).

Table 2.1: Moran's Index for spatial autocorrelation computed for cumulative incidence of cholera (1997-2001), and the specific years of cholera outbreaks (1998, 1999, and 2001)

Year	I_M	*p*-value
1997-2001	0.271	0.0001
1998	0.331	0.004
1999	0.181	0.040
2001	0.240	0.010

Figure 2.2 also shows the Empirical Bayesian smoothed rates of cholera. A visual inspection reflects clustering of high rates of cholera at areas surrounding Kumasi

Metropolis and its adjoining neighbors (See Figure 2.2d). Moreover, clustering of high rates of cholera was persistent at Kumasi Metropolis and its adjoining neighbors in the years 1998, 1999, and 2001 (See Figures 2.2a, 2.2b, 2.2c).

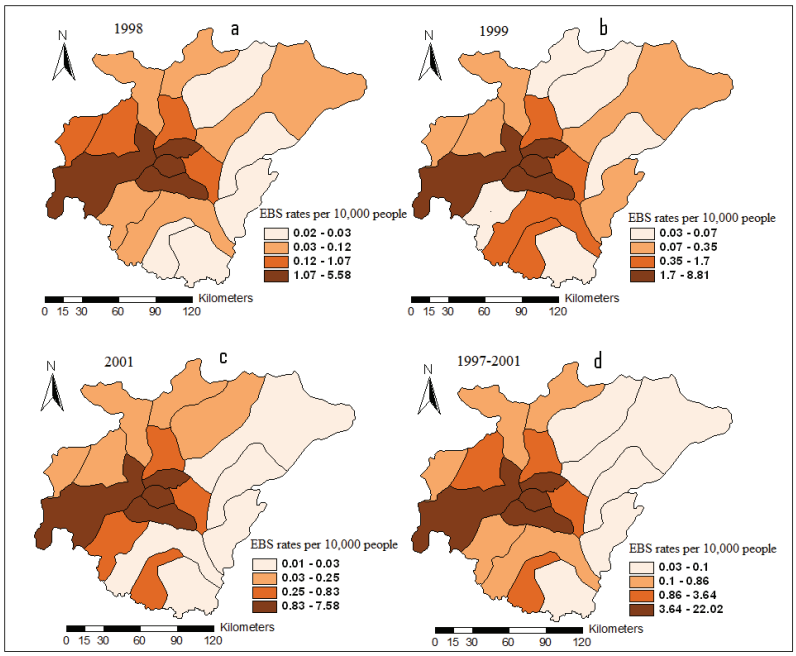

Figure 2.2: EBS Smoothed Rates of Cholera for 1998 (2a), 1999(2b), 2001(2c), and 1997–2001(2d)

The rate ratio within each stratum was computed, and the results shown in Tables 2.2-2.6. The cumulative incidence rate of cholera was 22 times higher in the first-order neighbourhood stratum than the third-order neighbourhood stratum. The cumulative incidence rate in the second-order neighbourhood (with Kumasi Metropolis) was not very high (1.4 times) compared to the third-order neighbourhood (See Table 2.2). Cholera incidence rate in the most urban stratum was about 20 times higher than the least urban stratum (Table 2.3), while the incidence rate in the most overcrowded stratum was about 30 times higher than the less overcrowded stratum (Table 2.4). A *Chi square* for trend analysis reflected a direct spatial relationship between cholera and urbanization (Table 2.3: $\chi^2 = 2995.5$, $p < 0.01$), overcrowding (Table 2.4: $\chi^2 = 1757.2$, $p < 0.01$), and an inverse relationship between cholera and order of neighbourhood (Table 2.2: $\chi^2 = 831.38$, $p < 0.01$).

The cumulative incidence rate of cholera was higher in the adjoining (first-order neighbourhood) stratum than in the non-adjoining stratum (second and third-order neighbourhoods) within each urbanization stratum (Table 2.5). Similar pattern was observed in the high and medium overcrowding strata (Table 2.6). No adjoining district was found within the low overcrowding stratum.

Table 2.2: Cholera incidence rate and population-based rate ratios by strata of districts classified according to order of neighbourhood and/or adjacency to Kumasi Metropolis, 1997-2001

Order of neighbourhood	Cases	Sub Population	Rate	Rate ratio (95%CI)
First-order	562	672482	8.36	22 (14.14-35.27)
Second-order	69	1307843	0.53	1.4 (0.84-2.36)
Third-order	21	558356	0.38	Reference

Chi-square for linear trend =831.375, p = 0.000001

Table 2.3: Cholera incidence rate and population-based rate ratios by strata of districts classified according to level of urbanization, 1997-2001

Urbanization (%)	Cases	Sub Population	Rate	Rate ratio(95%CI)
Low				
0.0-16.5	126	1065586	1.18	Reference
Medium				
19.2-35.6	357	848903	4.21	3.56 (2.89-4.38)
High				
38.9-100	2748	1170270	23.48	19.86 (16.55-23.84)

Chi-square for linear trend =2995.5, p = 0.000001

Table 2.4: Cholera incidence rate and population-based rate ratios by strata of districts classified according to level of overcrowding, 1997-2001.

Overcrowding index	Cases	Sub Population	Rate	Rate ratio(95%CI)
Low				
(-4.5<OI<-1.66)	51	1048225	0.49	Reference
Medium				
(-0.73<OI<-0.37)	310	764556	4.05	8.33 (6.14-11.34)
High				
(-0.25<OI<3.20)	2870	1896170	15.14	31.11 (23.40-41.46)

Chi-square for linear trend =1757.2, P-value=0.000001

Table 2.5: Cholera incidence rate and population-based rate ratios by strata of districts classified according to urbanization level and order of neighbourhood and/or adjacency to Kumasi Metropolis, 1997-2001.

Urbanization (%)	Neighbourhood	Cases	Population	Rate	Rate ratio(95% CI)
Low	Adjoining	84	146028	5.75	12.59 (8.57-18.55)
0.0-16.5	Non-adjoining	42	919558	0.46	Reference
Medium	Adjoining	330	361786	9.12	16.46 (10.96-24.88)
19.2-35.6	Non-adjoining	27	487117	0.55	Reference
High	Adjoining	148	164668	8.99	19.67 (12.22-31.95)
38.9-100	Non-adjoining	21	459524	0.46	Reference

Table 2.6: Cholera incidence rate and population-based rate ratios by strata of districts classified according to overcrowding level and order of neighbourhood or adjacency to Kumasi Metropolis, 1997-2001

Overcrowding	Neighbourhood	Cases	Population	Rate	Rate ratio (95% CI)
Low	Adjoining	0	0	0.00	0
(-4.5<OI<-1.66)	Non-adjoining	51	1048225	0.49	*
Medium	Adjoining	286	237610	12.04	26.43 (17.16-41.05)
(-0.73<OI<-0.37)	Non-adjoining	24	526946	0.46	Reference
High	Adjoining	276	434872	6.35	12.31 (7.17-21.51)
(-0.25<OI<3.20)	Non-adjoining	15	291028	0.52	Reference

**undefined*

2.5 Discussion

This study takes advantage of the advancements in GIS technologies such as spatial disease mapping and smoothing, exploratory spatial data analysis such as spatial autocorrelation, and spatial statistical techniques to identify demographic risk factors of cholera. The extent to which neighbouring values are correlated was measured using Global Moran's Index for spatial autocorrelation. All autocorrelation analyses suggest significant spatial clustering of cholera with positive Moran's Index (see Table 2.1). This non random distribution also suggests spatial clustering of high rates of cholera incidence at the central part of the region, and low rates at the peripheries (See Figure 2.2). This is also shown by the high rate ratio of 22 times in the first-order neighbourhood stratum (i.e. direct neighbors with Kumasi Metropolis, See Table 2.2). Visual inspections of the EBS maps also suggest possible sustained transmission of cholera at districts within the central part of the region (see Figures 2.2a, 2.2b, 2.2c, 2.2d). These patterns are plausible largely because of two main reasons. (1) *Demographic status*: Kumasi is the most urbanized and highly commercialized district in Ashanti Region, and therefore there is always a high daily influx of traders and civil workers from neighbouring districts to Kumasi Metropolis. Such a high daily influx strain existing sanitation systems, thereby putting people at increased risk of cholera transmission. Also, the rural poor most often migrate to city centres with the hope of better life. However, due to the high cost of housing, such migrants settle at slummy and/or squatter areas where environmental sanitation is poor. This largely explains the high *northern population* (inhabitants from the northern sector of Ghana; which is the most deprived sector) within Kumasi Metropolis (2) *Geographic location*: Kumasi Metropolis is the central nodal district of Ghana, and therefore, all road networks linking the northern sector and the southern sector of Ghana pass through Kumasi. There is the high probability of stoppage and transit by travellers, resulting in a high daily population increase and overcrowding at city centres.

This study has also shown that high urbanization and overcrowding are the most important predictors of cholera in Ashanti Region, Ghana (See Tables 2.3, 2.4, 2.5, 2.6). Although cholera is transmitted mainly through contaminated water or food, sanitary

conditions in the environment play an important role since the *V. cholerae* bacterium survives and multiplies outside the human body and can spread rapidly where living conditions are overcrowded and water sources unprotected and where there is no safe disposal of solid waste, liquid waste, and human faeces (WHO, 200b). These conditions are met in highly urbanized communities in Ashanti Region. The high rate of urbanization has led to the high level of overcrowding, which necessarily results in shorter disease transmission path. This is shown by the very high rate ratios within the high urban (RR = 19.86) and high OI (RR = 31.11) stratum (See Tables 2.3 and 2.4). In fact, the DCU has attributed outbreaks of cholera in urban communities to poor waste management and sanitation systems. In Ghana, urban population growth rate of about 4.3% has outstripped the overall national population growth rate of about 2.7%. The proportion of the population residing in urban areas rose from 32% in 1984 to 43.8% in 2000 (PHC, 2000). Such high urbanization rate strain existing resources meant for providing better sanitation systems and potable water. Inadequate sanitation systems coupled with intermittent supply of pipe borne water in urban communities puts the population at risk of cholera. Surface water pollution is particularly found to be worse where rivers pass through urban and overcrowded cities, and the commonest contamination is from human excreta and sewage. Due to the cosmopolitan (multi-ethnic) nature of the urban cities in Ghana (PHC, 2000), the traditional laws which were used to protect water bodies form faecal pollution are no longer adhered (Traditionally, it is a taboo to defecate or dispose waste in a water body). Therefore, defecating and dumping of waste in and at the banks of surface water bodies has become a common practice in most urban communities. However, urban inhabitants resort to such polluted water bodies for various household activities like cooking and washing during periods of water shortages.

Further, the rate of slums and/or squatter formation in urban communities is high due to the high rate of migration and population redistribution. Inhabitants living at slums and/or squatter settlements are generally poor, and face problems including access to potable water and sanitation. The urban poor (slums and squatter settlers), are worse off than their rural counterparts in terms of access and affordability to safe drinking water and sanitation. In many cases public utilities providers (e.g. Water distribution) legally fail to serve the urban poor living in slums due to factors regarding land tenure system, technical and service regulations, and city development plans. Most slums and/or squatter settlements are also located at low lying areas susceptible to flooding. Unfavourable topography, soil, and hydro-geological conditions make it difficult to achieve and maintain high sanitation standards among populations living in these territories (Barroto and Martnez-Piedra, 2000).

This study has shown the capabilities of spatial analysis and GIS in analyzing geographically referenced health data in Ghana. Moreover, the study has also proven that the demographic risk factors of cholera may not be the same in every geographical area or country. For example, Barroto and Martnez-Piedra (2000) identified low urbanization as one of the most important ecologic predictors of cholera in Mexico, a Latin American country. However, the results of this study show that high urbanization positively correlates with high cholera incidence.

Although some findings of this research reaffirms the already known hypothesis of cholera, we present the possibility of using GIS and spatial statistical tools for health research in this study area where GIS application in the health sector has not been extensively utilized.

2.5.1 Limitations of study

The results of the Extended Mantel-Haenszel *Chi Square* for trend analyses should be interpreted with caution. The number of cholera cases reported to the DCU may only be a fraction of cases that actually occurred, especially in lowly urbanized districts (or rural areas) of the country where level of education is extremely low. It has been suggested that educational level indirectly determines a person's healthcare seeking behaviour (Ali et al., 2002a, 2002b).

The spatial scale of the data may invariably affect the results of the spatial analysis. The areas identified as high risk of cholera are generally large areas defined by administrative boundaries. In such a large spatial scale, it is difficult to demonstrate the actual risk of cholera within a smaller group of people. A more detailed study at a smaller spatial scale is therefore required to assess the accurate individual or smaller groups of individuals' exposure levels. The spatial autocorrelation analysis should be interpreted with caution due the different shapes and sizes of the districts.

2.6 Conclusion

This study has demonstrated the use of spatial statistical analysis and GIS to map hotspots, and the spatial dependency of cholera distribution within a population. Through spatial statistical procedures, non-randomness in the distribution of cholera and the identification of unusual concentration of cholera incidence has been defined. This therefore prompts health planners in the country to take a critical look at these risk areas, and make appropriate health planning and resource allocation. In conclusion, the results of this research suggest that high urbanization, high overcrowding, and neighbourhood with Kumasi Metropolis are the most important predictors of cholera in Ashanti Region-Ghana. It is therefore necessary that health officials and policy makers reasonably improve their surveillance systems to prepare for the possibility of sustained transmission should an infection be introduced. Since this research is the first of its kind in Ghana, a more detailed research is required to consider factors like access to safe drinking water, and availability of waste disposal systems to thoroughly evaluate the risk of cholera in the region.

3

Spatial and space-time clustering of cholera

Time and space are modes by which we think and not conditions in which we live"

Albert Einstein

In chapter 2, a global cluster analysis showed that high cholera rates are clustered around Kumasi Metropolis (the central part of the region), with a significant positive Moran's. Since global cluster analysis ran the risk of obscuring local effects, we undertake local cluster analysis to define and map spatial and space-time clusters of cholera. Kulldorff's spatial scan statistics method is utilized since it is both deterministic (i.e., it identifies the locations of clustering) and inferential (i.e., it allows for hypothesis testing and evaluation of significance). Correlation analyses of cholera rates with demographic factors are also explored to assess the extent to which these factors might explain high rate clusters of cholera. This chapter is in preparation for publication as: Osei FB, Duker AA and Stein A: Investigating spatial and space-time clustering of cholera in Ashanti Region-Ghana, 1997-2001.

Abstract

After its inception in the early 1970's, cholera now exists in both endemic and epidemic forms in Ghana, recording cases almost every year. Between 1999 and 2005, a total of 25,636 cases and 620 deaths were officially reported to the World Health Organization (WHO). Since cholera is primarily driven by environmental factors, and since environmental processes are spatially continuous in nature, high disease rates are expected to cluster together. The objective of this study was to investigate the spatial and temporal clusters of cholera in Ashanti region using the *spatial scan statistic*. Correlation analyses of cholera rates with demographic factors are also explored to assess the extent to which these factors might explain high rates clusters of cholera.

The results show the presence of high rate clusters of cholera in areas surrounding Kumasi Metropolis, and a possible sustained transmission during the period under study. The correlation analyses also show that cholera prevalence is high when the majority of the people do not have access to good sanitation facilities; drink from rivers, wells and ponds; and when internal migration is high. The results show the presence of high rate spatial and space-time cholera clusters, suggesting possible sustained transmission of cholera in Kumasi Metropolis during the period under study.

3.1 Introduction

Basic problems in geographical surveillance for a spatially distributed disease data are the identification of areas of exceptionally high prevalence or clusters, test of their statistical significance, and identification of the reasons behind the elevated prevalence of the disease. Knowledge of the location of high risk areas of diseases and factors leading to such elevated risk is essential to better understand human interaction with its environment, especially when the disease transmission is enhanced by environmental or demographic factors. Cluster analysis provides opportunities for environmental epidemiologist to study associations between demographic and environmental exposures and the spatial distribution of diseases (Myaux et al., 1997; Kulldorff and Nagarwalla, 1995; Besag and Newell, 1991; Kulldorff, 1997, 2005, 2006).

Cholera has remained as an important cause of mortality and morbidity in the world, especially in developing tropical countries. Cholera is a global threat; it is listed as one of three internationally quarantainable diseases by the World Health Organization (WHO), along with plague and yellow fever (WHO, 2000b). The burden of cholera is enormous on the African continent. Between 1999 and 2005, there were over 1 million reported cholera cases and over 28,000 reported deaths worldwide. Africa alone accounted for about 90% of the cases and 96% of the deaths worldwide (WHO, 2000a, and 2001-2006). The disease has been a public health problem in Ghana since its introduction in the 1979's (Pobee and Gran, 1970). From 1999 to 2005, the Ghana Ministry of Health officially reported a total of 26,924 cases and 620 deaths to the WHO (WHO, 2000-2006).

Cholera is caused by specific strains of the water borne bacterial *Vibrio cholerae* O1 or O139 (*V. cholerae* here after), following ingestion of infective dose through contaminated water or food (Kelly, 2001). John Snow (1855) first associated cholera with contaminated drinking water in the 1850's even before any bacterial was known to exist (Koch, 1884). Once *V. cholerae* are present in drinking water in sufficient dose, an outbreak can trigger and propagate depending on several environmental and demographic risk factors (Ali et al., 2002a, 2000b; Borroto and Martinez-Piedra, 2000). The synergy of poverty, ignorance, poor hygiene, lack of good water supplies, poor housing and certain social setups creates conducive environment for the survival of *V. cholerae* outside its habitat. Cholera spreads rapidly through the faecal-oral route among communities that are poor, crowded, and characteristically without adequate disposal of human wastes (Greenough, 1995).

Since cholera is primarily driven by environmental factors (Huq et al., 2005), and since environmental processes are spatially continuous in nature (Webster et al., 1994), high incidence rates of the disease are expected to cluster together. A previous study carried out in Ashanti region used Moran's Index for spatial autocorrelation to explore the existence of clusters of cholera. Also in the above study, empirical Bayesians smoothed rates of cholera (i.e. visual inspection) revealed possible spatial and temporal clustering of cholera for the 5 year period, i.e. from 1997 to 2001 (Osei and Duker, 2008). However, the exact locations of these cluster, as well as the correlations with some demographic and socioeconomic factors were not systematically investigated. The

purpose of this study is to investigate spatial and space-time clusters of cholera in Kumasi. Correlation analysis of cholera rates with demographic factors, i.e. sanitation, drinking water and internal migration are also explored to assess the extent to which these factors might explain high rates clusters of cholera.

In this study we use the *spatial scan statistic* to detect geographical cluster of cholera. Several formal methods and techniques for identifying high risk areas by disease clustering techniques have been developed (Myaux et al., 1997; Kulldorff and Nagarwalla, 1995; Besag and Newell, 1991). However, the spatial scan statistic developed by Kulldorff (1997, 2005, 2006) offers several advantages over the others: (1) it corrects for multiple comparisons, (2) adjusts for the heterogeneous population densities among the different areas in the study, (3) detects and identifies the location of the clusters without prior specification of their suspected location or size thereby overcoming pre-selection bias, (4) and the method allows for adjustment for covariates. Also Kulldorff's spatial scan statistic is both deterministic (i.e., it identifies the locations of clustering) and inferential (i.e., it allows for hypothesis testing and evaluation of significance). The spatial scan statistic has been used to detect and evaluate various disease clusters including cancer (Michelozzi et al., 2002; Viel et al., 2000; Sheehan and DeChelo, 2005; Hjalmars et al., 1996; Turnbull, 1990, Kulldorff et al., 1998), giardiasis (Odoi et al., 2004) tuberculosis (Tiwari et al., 2006), diabetes (Green et al., 2003), Creutzfeldt-Jacob disease (Cousens, 2001), granulocytic elrlichiosis (Chput et al., 2002), and scelrosis (Sabel et al., 2003). The spatial scan statistic, as implemented in SaTScan software (Kulldorff, 1997; Kulldorff, 2005; Kulldorff, 2006) has the capabilities of detecting purely spatial clusters, temporal clusters, and space-time clusters. Using spatial scan statistic to detect purely spatial clusters for an extensive time period minimizes the power to detect recently emerging clusters (Tiwari et al., 2006). However, if data from the last few years are used, low to moderate excess risk that is present for a considerable length of time could be missed. Hence the purely spatial analysis used in addition with the space-time analysis helps the analyst to detect clusters which have persisted for a long period time as well as recently emerging clusters.

3.2 Methods

3.2.1 Study area

This study was conducted in Ashanti Region, one of the ten regions in Ghana. The region lies between longitudes 0° 9'W and 2° 15'W, and latitudes 5° 30'N and 7° 27'N. The Ashanti Region is dominated by Ashantis, who constitute 14.8% of all Ghanaians by birth. The Ashantis have a great history of culture of which the influence of the Ashanti Kingdom stretches beyond the borders of Ghana. The region occupies a total land area of 24,389 square kilometres representing 10.2% of the total land area of Ghana. The region has a population density of 148.1 persons per square kilometre, which is about two times higher than the overall population density in Ghana. There are 18 administrative districts in the Ashanti region including Kumasi Metropolis of which the capital is Kumasi, and is the only district which has gained a metropolitan status. The Kumasi Metropolis is the most populous district in the region. The 2000 census

recorded the region's population as 3,612,950, representing 19.1 per cent of the country's population. The urban population (51.3%) in the region exceeds that of the rural population (48.7%). In-house pit latrines and public toilets, which may be pit, Kumasi ventilated improved pit (KVIP) or bucket latrines, are the main toilet facilities used in the districts. Water closet (WC) is used by small proportions of households, ranging from 0.5 per cent in Ahafo Ano South to 27.8 per cent in the Kumasi Metropolis. The proportion of the population with access to potable (pipe-borne) water is relatively low in the districts, including the Kumasi Metropolis. A number of factors, particularly high fertility and internal migration, have accounted for the rapid population growth in the region. About two-thirds of the population in the region was born where they were enumerated; the remaining one third are in-migrants to the region. In 6 of the 18 districts, at least seven out of every ten persons were enumerated in the localities in which they were born, indicating that these districts have less in-migrant than other districts in the region (PHC, 2000).

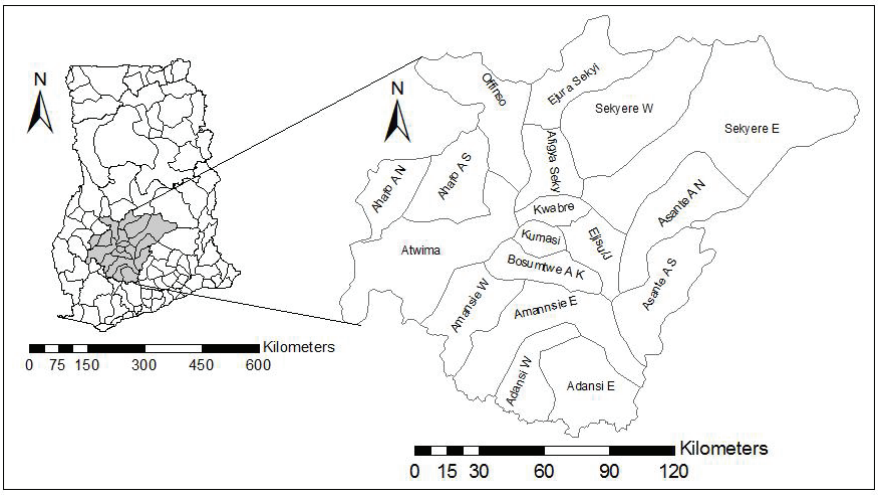

Figure 3.1: A map of Ghana showing Ashanti region, in gray colour. The figure also shows the spatial distribution of the various districts in Ashanti region

3.2.2 Data sources

The Ashanti region has a Disease Control Unit (DCU) to which all District Health Directorates (DHD) report suspected outbreaks of various infectious diseases at the end of each year. In this study, all cholera cases used were based on hospital data which were reported to the various DHD. For the detection of statistically significant clusters of cholera, the spatial scan statistics software, SaTScan, developed by Kulldorff, was used. This software requires three main data files to run:

Case file

Case file contains information about cholera cases for specified districts and times. Reported cases of cholera from 1997 to 2001 for each district within the region were retrieved from the DCU. Case definition of cholera was based on the WHO (1993) guidance on formulation of national policy on the control of cholera. According to this guidance, in an area where the disease is not known to be present a case of cholera should be suspected, when a patient, 5 years of age or older develops severe dehydration or dies from acute watery diarrhea, or where an epidemic is occurring, a patient, 5 years of age or older develops acute watery diarrhea, with or without vomiting.

Population file

The population file provides information about the background population at risk for each spatial district. The population database was obtained from the 2000 Population and Housing Census of Ghana conducted by the National Statistical Service (PHC, 2000).

Coordinate file

The coordinate file provides information about the spatial location of each district. In this study, the spatial scale of analysis was at the district level. The centroids of the districts were used as the coordinates of the districts.

3.2.3 Cluster analysis

The spatial scan statistic was used to detect the presence spatial and space-time clusters of cholera. The spatial scan statistic was developed by Kulldorff (1997, 2006) and it is been implemented in the SaTScan software. Spatial scan statistic has a disadvantage of being difficult to incorporate prior knowledge about the size and shape of an outbreak as well as its impact on disease rate (Neill et al., 2005). However, we used this as an advantage to get rid of pre-selection biases of clusters and their locations. Spatial scan statistic method is based on the principle that the number of cholera cases in a geographic area is Poisson-distributed according to a known underlying population at risk (Kulldorff, 2006). For the detection of purely spatial clusters, SaTScan imposes a circular window on the study region which is moved over the region and centred on the centroid of each district. The size of the circular window, which is also the cluster size, is expressed as a percentage of the total population at risk. This varies from 0 to a maximum (not exceeding 100), as specified by the user. The maximum window size should not exceed 50% of the total population because clusters of larger sizes would indicate areas of exceptionally low rates outside the circle rather than an area of exceptionally high rate within the circle. Possible clusters are tested within the window whenever it is centred on the centroid of each district. Whenever the window finds a new case, the software calculates a likelihood function to test for elevated risk within

the window in comparison with those outside the window. The likelihood function for any given window W is proportional to:

$$L(\mathrm{W}) = \sup_{W \in \mathbf{W}} \left(\frac{Chol_{(\mathrm{C})}(W)}{Chol_{(\mathrm{E(C)})}(W)} \right)^{Chol_{(\mathrm{C})}(W)} \left(\frac{Chol_{(\mathrm{C})}(\hat{W})}{Chol_{(\mathrm{E(C)})}(\hat{W})} \right)^{Chol_{(\mathrm{C})}(\hat{W})} \times I\left(\frac{Chol_{(\mathrm{C})}(W)}{Chol_{(\mathrm{E(C)})}(W)} > \frac{Chol_{(\mathrm{C})}(\hat{W})}{Chol_{(\mathrm{E(C)})}(\hat{W})} \right), \quad 3.1$$

where $\hat{W}$ indicates all the regions outside the window W, and $Chol_{(\mathrm{C})}(\,)$ and $Chol_{(\mathrm{E(C)})}(\,)$ denote the observed and expected number of cases within the specified window, respectively. The window W to be scanned by the spatial scan statistic is included in the set: $\mathbf{W} = \{W_{ik} \,|\, 1 \le i \le m, 1 \le k \le K_i\}$, where W_{ik}, $k = 1,...,K_i$, denote the window composed by the $(k-1)$ nearest neighbors to region i. The window W^* that attains the maximum likelihood is defined as the *most likely cluster* (MLC). The indicator function $I(\,)$ depends on the comparison between $Chol_{(\mathrm{E(C)})}$ and $Chol_{(\mathrm{C})}$. $I(\,)$ is 1 when $Chol_{(\mathrm{C})} > Chol_{(\mathrm{E(C)})}$, otherwise 0. The test of significance level of clusters is through the Monte Carlo hypothesis testing (Dwass, 1957). In this study, the maximum window size was set as 50% of the total population. The null hypothesis of no cluster was rejected when the simulated *p-value* was less than or equal to 0.05 for most likely clusters and 0.1 for secondary clusters since the latter have conservative *p-values* (Kulldorff, 2006).

A smaller window size (was defined as ≤ 25% of the total population) was also used to investigate the possibility of smaller clusters. This varied from ≤ 25% to ≤ 50% with successive increments of 5%. This was meant to check the sensitivity of spatial scan statistic to smaller window sizes when there are larger spatial units and small number of spatial units.

For the detection of space-time clusters, SaTScan imposes a cylindrical window with a circular geographic base and with height corresponding to the time of occurrences. In this way, the base of the cylinder is centred around one of several possible centroids located throughout the study region with the radius varying continuously in size, whereas the height of the cylinder reflects any possible time interval of less than or equal to half the total study period, as well as the whole study period. The window is then moved in space and time so that for each possible geographic location and size, it also visits each possible time interval (Kulldorf et al., 1998). The likelihood ratio test statistic is constructed in the same way as for the purely spatial scan statistic. However, the computational algorithm for calculating the likelihood for each window is in three rather than two dimensions (Kulldorff, 2001). Here, we used a spatial window that

could include up to 50% of population at risk and a maximum temporal window of 50%, without including purely spatial clusters. Moreover, most likely clusters for different time lengths (i.e. 1, 2, 3, or 4 year length) were scanned using a temporal cluster size of 90% of the study period and also included purely spatial clusters with temporal size of 100%. The maximum spatial cluster size was set at 50% of population at risk and included purely temporal clusters (spatial cluster size = 100%) as well.

3.2.4 Correlation between cholera and risk factors

All risk factors were obtained from the 2000 Population and Housing census of Ghana (PHC, 2000). Three main risk factors, i.e. sanitation, source of drinking water, and internal migration, were used to explore the extent at which these variables affect cholera prevalence within the study area. Four different types of sanitation facilities are used in the study area; WC, Pit latrine, KVIP, bucket or pan. A number of households in the districts have no access to toilet facilities. When a substantial number of households do not have toilet facilities, it is to be expected that inhabitants will defecate in the bush, drains, etc. Bucket or pan is the most unsafe sanitation method because the bucket is open and can attract filth breeding flies. Moreover, faeces have to be transferred to a different bucket when it is full; thus faeces can spread to nearby areas in the course of transfer. In this study, sanitation condition for a district is described as the percentage of the district's share of the region's population who do not have access to toilet facilities, and who use bucket or pan sanitation method. For this, larger values reflect poor or bad sanitation condition, while smaller values reflect good sanitation condition.

Since the natural reservoir of cholera is the aquatic environment, inhabitants who drink from wells, streams, rivers, ponds, dugouts and dams are assumed to be at a higher risk of cholera than those who drink from pipe borne water. Therefore, inhabitants who drink from wells, streams, rivers, ponds, dugouts and dams are classified as inhabitants who do not have access to potable water. The indicator for drinking water for each district was computed as the percentage of the district's share of the region's population who drink from wells, streams, rivers, ponds, dugouts and dams.

Internal migration is one of the important demographic characteristics that accounts for rapid population growth in a place. This variable was computed as a percentage of the district's share of the region's population in the year 2000 who were born outside the district during the time of enumeration.

Global Pearson's correlation coefficient was used to determine the correlation between cholera cumulative incidence rates from 1997 to 2001 and sanitation, drinking water, and internal migration. *p-values* were calculated to serve as a guide to access the significance of all correlation coefficients. Most health planning strategies in Ghana are based on the level of urbanization of a district. In other words, groups of districts with similar urbanization levels are planned together. With this in mind, all districts in the study region were stratified according to the level of urbanization; i.e. *low, medium and high*. Pearson's correlation analyses were repeated for each stratum of districts in order to assess the effects of the risk factors on cholera within each urbanization stratum.

3.3 Results and analyses

3.3.1 Purely spatial clusters

No cluster was detected for the years 1997 and 2000. Only most likely significant clusters were detected for the years 1998, 1999, 2001 (Table 3.1 and Figure 3.2). These clusters encompassed Kumasi, Bosumtwe AK and Kwabre in 1998 (relative risk $Chol_{(RR)} = 12.25$, $Chol_{(C)} = 733$, $Chol_{(E(C))} = 328.62$), Kumasi in 1999 ($Chol_{(RR)} = 7.42$, $Chol_{(C)} = 1033$, $Chol_{(E(C))} = 421.33$), Kumasi and Kwabre in 2001($Chol_{(RR)} = 15.60$, $Chol_{(C)} = 956$, $Chol_{(E(C))} = 383.32$), and Kumasi and Kwabre from 1998 to 2001($Chol_{(RR)} = 9.70$, $Chol_{(C)} = 2727$, $Chol_{(E(C))} = 1161.47$). No differences were observed between the results of the varying window sizes and the window size of ≤ 50% of the total population. Hence tables for these results are not shown.

Table 3.1: Most likely purely spatial clusters of cholera in Ashanti region, Ghana, detected by retrospective spatial analysis

Year	Cluster Area	$Chol_{(C)}$	$Chol_{(E(C))}$	$Chol_{(RR)}$	LLR	*p*-value
1998	Kumasi Kwabre Bosumtwe AK	733	328.62	12.253	434.73	0.001
1999	Kumasi	1033	421.23	7.42	592.29	0.001
2001	Kumasi Kwabre	956	383.32	15.597	673.86	0.001
1997-2001	Kumasi Kwabre	2727	1161.47	9.699	1618.26	0.001

This Table shows the results of the purely spatial cluster analysis using a spatial window that could include up to 50% of the population at risk in Ashanti region, Ghana, during 1998-2001: LLR (Log Likelihood Ratio).

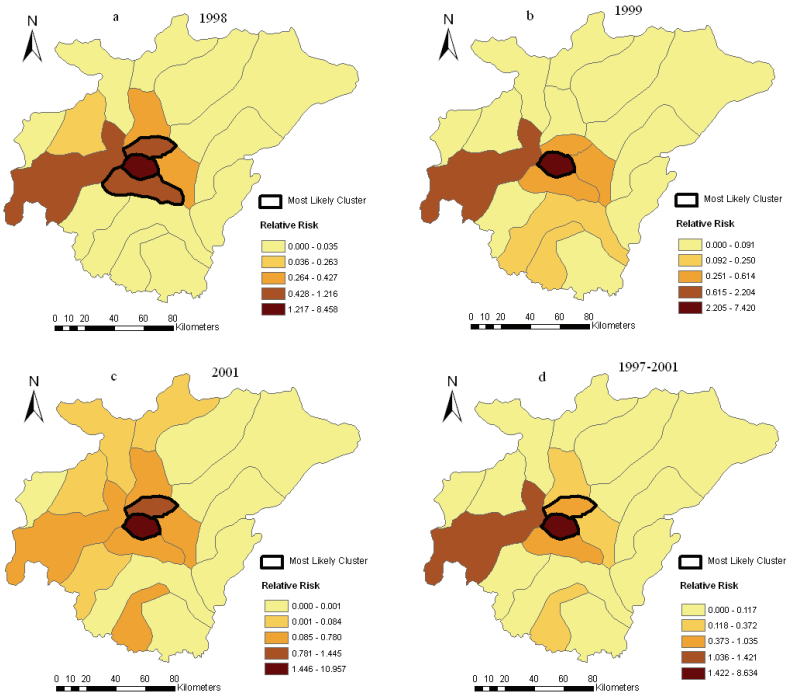

Figure 3.2: Locations of the detected clusters of cholera and spatial distribution of the relative risks for 1998(2a), 1999(2b), 2001 (2c), and 1998-2001 (2c).

3.3.2 Space-time clusters

While testing whether the purely spatial clusters were long term or temporary i.e. space-time analysis, a statistically significant (p = 0.001) most likely cluster was identified at Kumasi metropolis for the year 1998-1999. This cluster has $Chol_{(RR)} = 5.86$ with $Chol_{(C)} = 1668$ as against $Chol_{(E(C))} = 508.75$ (See Table 3.2). One statistically significant (P = 0.001) secondary cluster encompassing 3 districts (Ahafo Ano North, Ahafo Ano South, and Atwima) was detected for 1999. For this cluster, $Chol_{(RR)} = 1.91$ and $Chol_{(C)} = 179$ as against $Chol_{(E(C))} = 96.34$.

Table 3.2: Significant high rate spatial clusters of cholera in Ashanti region, Ghana, detected by retrospective space-time analysis

Cluster Area	Year	$Chol_{(C)}$	$Chol_{(E(C))}$	$Chol_{(RR)}$	LLR	*p-value*
Most Likely Cluster						
1. Kumasi Metro	1998-1999	1688	508.75	5.86	1149.02	0.001
Secondary Cluster						
2. Ahafo Ano North Ahafo Ano South Atwima	1999	179	96.34	1.908	29.34	0.001

This Table shows the results of the space-time cluster analysis using a spatial window that could include up to 50% of the population at risk and a maximum temporal window of 50% without including purely spatial clusters, in Ashanti region, Ghana, during 1998-2001: LLR (Log Likelihood Ratio).

The results of the space-time analysis when modified, i.e. when using a maximum temporal window of 90% (which included purely spatial clusters as well) and a spatial window that could include up to 50% of the population at risk (which included purely temporal clusters also) are shown in Table 3.3. Most likely statically significant (p = 0.001) cluster of high rates of cholera was again found to exist at the Kumasi Metropolis and Kwabre district for the year 1998-2001. This indicates that Kumasi Metropolis and Kwabre remained statistically significant throughout the year 1998-2001. One statistically significant (p = 0.001) secondary cluster encompassing Ahafo Ano South, Ahafo Ano North and Atwima for the year 1999 were also detected.

Table 3.3: Significant high rate spatial clusters of cholera in Ashanti region, Ghana, detected by retrospective space-time analysis

Cluster Area	Year	$Chol_{(C)}$	$Chol_{(E(C))}$	$Chol_{(RR)}$	LLR	*p-value*
Most Likely Cluster						
1. Kumasi Metro Kwabre	1998-2001	2727	1161.47	9.699	1618.26	0.001
Secondary Cluster						
2. Ahafo Ano North Ahafo Ano South Atwima	1999	179	96.34	1.91	29.33	0.001

This Table shows the results of the space-time cluster analysis when modified to find 1, 2, 3 or 4-year length clusters using a maximum temporal window of 90%, which included purely spatial clusters as well, and a spatial window of ≤50% of the population at risk, which included purely temporal clusters also, in Ashanti region, Ghana, during 1997-2001: LLR (Log Likelihood Ratio)

3.3.3 Correlation between cholera and risk factors

Pearson's correlation coefficients and their associated p-values were computed to determine the relationship between cholera cumulative incidence rate and the demographic risk factors (see Table 3.4). For the whole region, statistically significant relationship was observed for sanitation ($R^2 = 0.55$, $p = 0.001$), drinking water ($R^2 = 0.39$, $p = 0.001$), and internal migration ($R^2 = 0.73$, $p = 0.001$). However, when the analyses were repeated for each strata of urbanization, statistically significant correlations were observed for only the *high* urban strata (See Table 3.2). For instance there was a high, but non-significant correlation between cholera and drinking water

within the *medium-urban* strata ($R^2 = 0.62$, $p = 0.12$), and no significant correlation between cholera and drinking water within the *low-urban* strata ($R^2 = 0.001$, $p = 0.96$). However, there was a high and significant correlation between cholera and drinking water within the *high-urban* strata ($R^2 = 0.86$, $p = 0.007$).

Table 3.4: Pearson's correlation coefficients for the relationship between cholera rates and demographic risk factors

	Correlation and (*p-value*)		
	Sanitation	**Drinking water**	**Migration**
Global	[a]0.55 (0.001)	[a]0.39 (0.001)	[a]0.73 (0.001)
Low urban	[c]0.21 (0.36)	[c]0.04(0.66)	[c]0.001 (0.96)
Moderate urban	[c]0.48 (0.13)	[c]0.62 (0.12)	[c] 0.62 (0.11)
High urban	[b]0.86 (0.007)	[b]0.79 (0.018)	[b]0.89 (0.005)

This Table depicts both the Global Pearson's correlation analyses, and Pearson's correlation analyses for each urbanization strata of districts. The associated *p–values* are shown in brackets. *[a]significant correlations at 0.1% significance level; [b]significant correlations at 5% significance level. [c]not significant*.

3.4 Discussion

In this study, the purely spatial and space-time scan statistics methods implemented in SaTScan software have been used to analyze cholera cases from 1998 to 2001 in Kumasi, Ghana. These methods identifies whether unusual concentration of disease cases can be explained by chance or statistically significant. The findings of this study reveal several notable points. First, there is the existence of both purely spatial and space-time clusters, not explainable by chance (See Tables 3.1, 3.2, and 3.3). Also, the results of both the purely spatial and space-time analysis are somewhat similar. In particular, the purely spatial analysis reported an excess incidence of cholera in Kumasi during the years 1998, 1999, and 2001 (See Table 3.1 and Figure 3.2), and the space-time analysis also reported an excess incidence of cholera from 1999 to 2001 at the same area.

Second, the excess incidence of cholera mainly existed at Kumasi Metropolis throughout the period under study. Specifically, the purely spatial analysis reported excess incidence of cholera at Kumasi in 1998, 1999, and 2001. While testing whether the purely spatial clusters were long term or temporary, the space-time analysis also reported excess incidence of cholera at Kumasi Metropolis from the year 1999 to 2001. When the space-time analysis was modified to detect 1, 2, 3, 4, or 5 year length clusters, the space-time most likely cluster at Kumasi Metropolis became a purely spatial cluster (i.e. existed for 1997 to 2001, see Table 3.3). This indicates a sustained transmission of cholera at Kumasi Metropolis from 1997 through to 2001. Two main reasons may explain these patterns. (1) *Demographic status*: Kumasi is the most urbanized and highly commercialized district in Ashanti region, and therefore there is always a high daily influx of traders and civil workers from neighbouring districts to Kumasi Metropolis. Such a high daily influx strain existing sanitation systems, thereby putting people at increased risk of cholera transmission. The rural poor also often migrate to

city centres with the hope of a better life. However, due to the high cost of housing, such migrants settle at slummy and/or squatter areas where environmental sanitation is poor. This largely explains the high *northern population* (inhabitants from the northern sector of Ghana; which is the most deprived sector) within Kumasi Metropolis. (2) *Geographic location*: Kumasi Metropolis is the central nodal district of Ghana, and therefore, all road networks linking the northern sector and the southern sector of Ghana pass through Kumasi. There is the high probability of stoppage and transit by travellers, resulting in a high daily population increase and overcrowding at city centres.

Third, the findings of the space-time analysis clearly depict the statistical power of the scan statistics for detecting recently emerging clusters. The space-time analysis detected an important cluster during the year 1999 that would otherwise not be detected by a purely spatial analysis. This cluster encompassed areas surrounding Ahafo Ano North, Ahafo Ano South, and Atwima districts (See Tables 3.2 and 3.3).

Fourth, both the purely spatial and space-time cluster analysis detected no cluster during the years 1997 and 2000. This is somewhat consistent with both the overall global and national cholera trends. Although officially notified cases do not reflect the overall burden of the diseases, cholera cases reported to WHO in 1996 was 4.4 times higher than cases in 1997(a decrease of 77% from 1996 to 1997), and cases in 1998 was 9 times higher than cases in 1997(an increase of 803% from 1997 to 1998). Compared to 1999, the year 2000 saw 46% global reduction in the total number of cases, and about 65% reduction in the total number of cases reported in Ghana. After a massive outbreak in Ghana from 1998 to 1999, health officials and policy makers implemented several measures to curb the menace. Notable among these measures were effective waste collection and disposal (including solid waste, sewage and septage, industrial and clinical waste), cleansing of public areas, food hygiene, hygiene education and related programs. Consequently, the reduced number of cholera cases in the year 2000.

When the maximum window size was varied from $\leq$ 25% to $\leq$ 50% of the total population, the same results were obtained as with the window size of $\leq$ 50% of the total population. This clearly shows that for large geographical scales with fewer numbers of spatial units, spatial scan statistic will likely not be sensitive to varying window size. Chen et al. (2008) clearly demonstrated the sensitivity of the spatial scan statistic to the issues of varying window sizes (SaTScan scaling issues) through a geovisual analytic technique. Their study was partly a quest to determine a an optimal setting for SaTScan scaling parameters due to the confusing and even misleading results which are possible if the parameter choices are made arbitrarily. However, their data was across larger spatial geographical area with larger number of spatial units; giving SaTScan much flexibility on the varying window sizes. Contrary to our data used, there were only 18 spatial units; a number probably too small for spatial scan statistic. Therefore the interpretation of our findings should fall within the framework of the above limitation.

The findings of the correlations analysis suggest that cholera is high when majority of the people do not have access to good sanitation facilities; do not have access to potable water; and when internal migration is high. When the correlation analyses were repeated

for each strata of urbanization, statistically significant correlations were observed for only the *high-urban* strata. Considering drinking water for instance, there was no significant correlation within the *low-urban* strata and the *medium-urban* strata, but a high significant correlation was observed within the *high-urban* strata (See Table 3.4). This implies that drinking water, sanitation and internal migration affects only *high-urban* communities in the study area. This is consistent with the findings of the cluster analysis. Both the purely spatial and space-time analysis identified Kumasi Metropolis and Kwabre district as significant high rate clusters of cholera, which are also amongst the most urbanized and overcrowded areas in Ashanti Region.

Cholera primarily attack individuals with insufficient knowledge of and inappropriate attitudes towards hygienic practices, and who live in dwellings that lack access to safe drinking water supply and to adequate facilities for sanitation, sewerage disposal and treatment (Glass and Black, 1992; OPS, 1994). Majority of the region's population who do not have access to good sanitation systems, and drink from rivers, streams and ponds are people living in most urbanized and densely populated districts. For instance, Kumasi metropolis's share of the region's population who do not have access to potable water is close to 13%, a value 2.3 times higher than the mean percentage.

Faecal contamination of rivers is a major water quality issue in many fast growing cities like the Kumasi Metropolis where population growth far exceeds the rate of development of wastewater collection and treatment (Maybeck et al., 1989). The water bodies near densely populated areas may have high faecal concentrations due to defecation and sanitation practices of the people. Ali et al. (2002) has asserted that faecal contamination of surface water in densely populated area is higher than a sparsely populated area. Although Kumasi Metropolis and other urbanized districts are served with potable water, this water does not flow throughout the year. At certain times no water flow for a period of a week or two. Residents are therefore compelled to exploit nearby streams, rivers and ponds. If such water bodies are contaminated and is used for drinking or cooking, there is the likelihood of infection.

After several decades of research into cholera, the risk factors which contribute its transmission have not changed. The spatial and temporal patterns that the disease displays, however, are not the same from one outbreak area to another. Although several of the findings of this research are more confirmatory, it draws the attention of health officials and policymakers about the area where there has been sustained transmission of the disease over the years. The study also provides very useful information to health officials and policymaker about the spatial and temporal patterns of cholera in Ghana. For example, this study clearly shows that there has been a sustained transmission of cholera in Kumasi during the period under study. The findings of this study will also have important implications for public health officials since control strategies would vary depending on the most important risk factors in most important districts. With the important high rate cluster locations and risk factors identified, optimal efforts can be taken at appropriate districts to prevent and control cholera. There is no doubt that the faecal oral route of cholera transmission should be of primary concern because of its importance in the development of secondary cases and in subsequent spread of the disease. It should therefore be the concern of health officials and policymakers to

provide better sanitation systems to prevent faecal contamination of water bodies within *high-urban* districts. Moreover, potable (pipe-borne) water supply in urban and densely populated districts should be expanded an improved to prevent cholera outbreaks.

3.5 Conclusion

This study has shown the presence of both spatial and space-time hotspots of cholera in Ashanti region, suggesting that there has been sustained transmission of cholera within these hotspots. The study has also shown that drinking water, sanitation and internal migration are important risk factors of cholera in Ashanti region; however, these predictors do not have a significant impact in cholera transmission in low urban communities. This study has also demonstrated that using available health data, GIS and spatial scan statistics can provide public health officials in Ghana with new knowledge about the prevalence rate and hotspots of a disease. This new knowledge will help them to come out with optimal strategies to prevent and contain diseases outbreak. Since drinking water, sanitation and internal migration could not explain cholera prevalence in low urban areas, a more detailed research need to be carried out at individual levels to thoroughly understand the epidemiology of cholera in this study area.

4

Spatial dependency of cholera on refuse dumps

"The refuse which overflowed from the privies and a pigsty could be seen running into the well over the surface of the ground, and the water was very fetid; yet it was used by the people in all the houses except that which had escaped cholera.....It is not unlikely that insects, especially the common house-flies, aid in spreading the disease"

John Snow

In Chapters 2 and 3, spatial analyses of cholera have been conducted based on large scale spatial datasets. This, however, obscures the effect of local risk factors occurring at the community levels. Based on the findings of chapters 2 and 3, it is deduced that Kumasi Metropolis is the high risk area for cholera. In this chapter, we explore the effects of local environmental risk factors on cholera prevalence using Kumasi Metropolis as the case study area. The objectives are to (1) determine whether cholera prevalence is related to proximity and density of refuse dumps in Kumasi, (2) detect and map spatial clusters of cholera, and determine whether refuse dumps are a contributory factor to high rate cholera clusters and, (3) to determine a critical buffer distance within which refuse dumps should not be sited away from communities. This chapter has originally been published as: Osei FB and Duker AA: Spatial dependency of *V. cholera* prevalence on open-space refuse dumps in Kumasi, Ghana: a spatial statistical modelling. *International Journal of Health Geographics* 2008, 7:62

Abstract

Cholera has persisted in Ghana since its introduction in the early 1970s. From 1999 to 2005, the Ghana Ministry of Health officially reported a total of 26,924 cases and 620 deaths to the World Health Organization (WHO). Etiological studies suggest that the natural habitat of V. cholera is the aquatic environment. Its ability to survive within and outside the aquatic environment makes cholera a complex health problem to manage. Once the disease is introduced in a population, several environmental factors may lead to prolonged transmission and secondary cases. An important environmental factor that predisposes individuals to cholera infection is sanitation. In this study, we exploit the importance of two main spatial measures of sanitation in cholera transmission in an urban city, Kumasi. These are proximity and density of refuse dumps within a community. A spatial statistical modelling carried out to determine the spatial dependency of cholera prevalence on refuse dumps show that, there is a direct spatial relationship between cholera prevalence and density of refuse dumps, and an inverse spatial relationship between cholera prevalence and distance to refuse dumps. A spatial scan statistics also identified four significant spatial clusters of cholera; a primary cluster with greater than expected cholera prevalence, and three secondary clusters with lower than expected cholera prevalence. A GIS based buffer analysis also showed that the minimum distance within which refuse dumps should not be sited within community centres is 500 m. The results suggest that proximity and density of open-space refuse dumps play a contributory role in cholera infection in Kumasi

4.1 Introduction

The Ganges Delta region is believed to be the traditional home of cholera (Harmer and Cash, 1999). From this region, cholera has spread throughout the world, causing seven major pandemics since 1817 (Faruque et al., 1998). The seventh pandemic, which began in 1961 in Indonesia, reached West Africa in 1970 (Barua, 1972; Cvjetanovic and Barua, 1972; Goodgame and Greenough, 1975; Kustner et al., 1981, Glass et al., 1991). In Ghana the first bacteriological case report of cholera was on 1st September, 1970 (Pobee and Grant, 1970). Since then cholera has been endemic in Ghana, with occasional outbreaks. From 1999 to 2005, the Ghana Ministry of Health officially reported a total of 26,924 cases and 620 deaths to the WHO (WHO, 2000a, 2001, 2002, 2003, 2004, 2005, 2006).

Cholera is an acute intestinal infection caused by the bacterial *V. cholerae*. The main mode of infection is through contaminated food and drinking water. When ingested in the body, *V. cholerae* produces an exotoxin that either stimulates the mucosal cells to secrete large quantities of isotonic fluid, or increases the permeability of the vascular endothelium, thus allowing isotonic fluid to pass through in abnormal amount (Carpenter, 1970), resulting in watery diarrhoea. Without prompt treatment, *V. cholerae* can cause severe dehydration and death within hours of onset in a severely purging individual (Prestero, 2001). In an unprepared community, case-fatality rate or death can be as high as 50% of severe cases (Sack et al., 2004; WHO, 1993). In both concept and execution, *V cholerae* infection has extraordinarily simple and successful treatment (Mahalanabis et al., 1992). Oral rehydration salt (ORS) solutions are the most important component of treatment, although intravenous fluids are needed for patients with very severe dehydration (Crowcroft, 1994).

The general assumption by most workers, until quite recently (mid 60's), was that *V. cholerae* was an organism whose normal habitat was the human gut and/or intestine, and incapable of surviving for more than a few days outside the gut (Felsenfeld, 1966), however, recent studies suggest that the natural reservoir of cholera is the aquatic environment (Finkelstein, 1999; Colwell and Huq, 1994; Byrd et al., 1991; Feachem, 1981; Mosley and Khan, 1979). The ability of *V. cholerae* to survive outside the aquatic environment makes cholera difficult to eradicate. Etiological studies suggest that *V. cholerae* survives well in faecal specimens if kept moist (Sack et al., 2004). This makes cholera a complex health problem to manage. Once the disease is introduced in a population, several ecological and/or environmental factors may lead to prolonged transmission and secondary cases. Huq et al. (2005) has specifically demonstrated linkages between cholera and environmental variables. Ali et al. (2002a, 2002b) identified proximity to surface water, high population density, and low educational status as the important predictors of cholera in an endemic area of Bangladesh. Borroto and Martinez-Piedra (2000) identified poverty, low urbanization, and proximity to coastal areas as the important ecologic predictors of cholera in Mexico. Several scientific studies have also demonstrated the involvement of climatic factors in the recurrence of epidemic cholera (Gil et al., 2004; Pascual et al., 2000; De Magny et al., 2006). A very important ecologic and/or environmental factor that predisposes inhabitants to cholera infection is sanitation. Since cholera is hypothesized as a disease

of deficient sanitation (Ali et al*.,* 2002a, 2002b), an outbreak of cholera is therefore a stark reminder of deficiency in sanitations systems. However, sanitation as a spatial risk can have several measures depending on the geographical area. In Bangladesh, Ali et al. (2002b) used latrine types, classifying them as safe and unsafe, as a measure of sanitation. In studying the geographical patterns of cholera in Mexico, Borroto and Martinez-Piedra (2000) also used percentage dwellings without connection to sewerage or septic tanks to incorporate sanitation in a composite poverty index. In this study, we exploit two main spatial measures of sanitation in an urban city, Kumasi, in a developing country. These are proximity to refuse dumps, and density of refuse dumps within a community. In such settings, significant amount of human excreta reaches refuse dumps. During outbreak periods, surface runoff containing *V. cholerae* contaminate surface water if consumed perpetuates transmission of the organism. Also, refuse dumps can serve as breeding sites for some dangerous flies. Example, the common housefly, *Musca domestica,* is a eusyanthropic fly species i.e., linked to the human habitat, and a chief offender among the filth breeding flies worldwide (Greenberg, 1973). Studies show that the common housefly and flies in general can serve as mechanical vectors of many kinds of pathogens such as bacterial (Levine and Levine, 1990; Cohen et al*.,* 1991), protozoa (Fotedar et al*.,* 1992a), viruses (Ogata et al*.,* 1961), and helminth eggs (Sulaiman et al*.,* 1988; Dipeolu, 1982). Our hypothesis is that refuse dumps create environmental niches for *V. cholerae* infection, and therefore inhabitants who live in close proximity to open-space refuse dumps should have high cholera prevalence than those farther. Also areas with high density of refuse dumps should have higher cholera prevalence than areas with lower density. Although cholera is one of the most researched communicable diseases, no study so far has explored the spatial extent at which these ecologic and/or environmental factors influence its transmission.

Epidemiologists have long used maps to track the spread of disease, and in the past decade, geographic information system (GIS) technology has added powerful new tools that help reveal far more than simply the "where" and "when" of epidemics. Recent advances in GIS has allowed the application of not only disease mapping but also spatial analysis, such as spatial clustering and cluster detection in epidemiological research (Fukuda et al*.,* 2005; Rosenberg et al*.,* 1999; Kulldorff et al., 1997). A GIS is capable of analyzing and integrating large quantities of geographically distributed data as well as linking geographic data to non-geographic data to generate information useful in further scientific research and in decision making (Duker et al*.,* 2004). Spatial analyses (in the context of GIS), such as cluster analysis and geographic correlation studies are commonly used to characterize spatial patterns of diseases (Beck et al*.,* 1994; Glass et al*.,* 1995; Clarke et al*.,* 1991; Braddock et al*.,* 1994; Barnes and Peck, 1994; Wartenberg et al*.,* 1993; Odoi et al*.,* 2004). Such methodologies have also been utilised in several cholera studies (Ali et al*.,* 2002a; Ali et al*.,* 2002b; Borroto and Martinez-Piedra, 2000; Glass et al*.,* 1982; Kwofie, 1976; Fleming, 2007; Ackers et al*.,* 1998). In this study we use a GIS based statistical modelling to explore relationship between our spatial measure of sanitation (described above) and cholera prevalence, and spatial scan statistic to investigate geographical clusters of cholera. The objectives of this study are:

(1) determine whether cholera prevalence is related to proximity and density of refuse dumps in Kumasi,
(2) detect and map spatial clusters of cholera, and determine whether refuse dump is a contributory factor to high rate cholera clusters and,
(3) and determine a critical buffer distance within which refuse dumps should not be sited away from communities. The results of this study will provide invaluably information to health officials and policy makers to develop effective prevention and control programmes for cholera.

4.2 Materials and methods

4.2.1 The study area

This study was conducted in Kumasi, the capital of Ashanti Region (See Figure 4.1). Ashanti Region is centrally located in the middle belt of Ghana. It lies between longitudes 0.15°W and 2.25°W, and latitudes 5.50°N and 7.46°N. The region is divided into 18 districts. The Kumasi metropolis alone accounts for nearly one-third of the region's population (PHC, 2000). Kumasi is located in the south-central part of the country, about 250 km (by road) northwest of Accra, the capital city of Ghana, It lies at the intersection of latitude 6.04°N and longitude 1.28°W, covering an area of about 220 km^2. The Kumasi metropolis is the most populous district in the region. It has a population of about 1.2 million which accounts for just under a third (32.4%) of the region's population. Kumasi has attracted such a large population partly because it is the regional capital, and also the most commercialized town in the region. Other reasons include its centrality as a nodal town with major road arteries to other parts of the country. A greater proportion of households in the metropolis, 81.2%, use the public dump to dispose of solid waste. Only 2.2% of the metropolis population's wastes are collected (i.e. house to house collection), but only in few first class residential areas. The remaining population either burry their waste, burn or dump it elsewhere (PHC, 2000). Most of these refuse dumps also serve as transfer stations for transferring waste to a landfill site. Sanitation generally becomes a problem during the rainy season. Most access roads to refuse dumps are unpaved and become extremely deplorable during the rainy season, and therefore refuse collection vehicles are not able to ply such roads. As a result, refuse are left to pile up during the rainy season. The main source of drinking water in the metropolis is pipe borne water (about 82%). 11.5% drink from well, 1.5% drink from river, pond or lake, while 0.8% of the people obtain their drinking water from tanker supply. However, due to rampant water shortages, most inhabitants resort to nearby streams and rivers for various household activities such as cooking and washing.

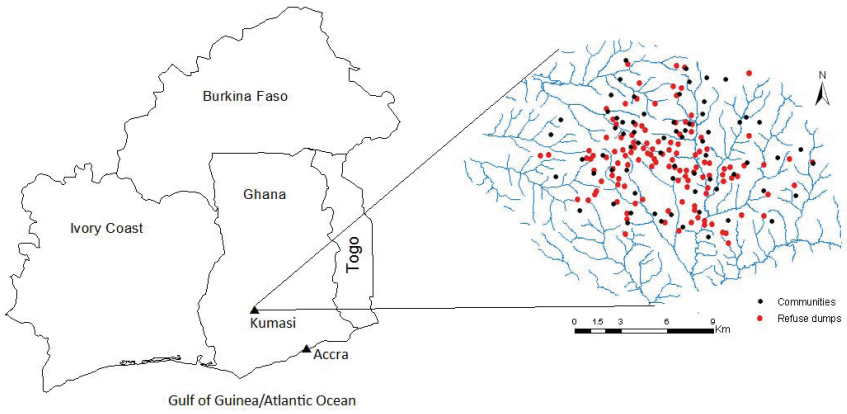

Figure 4.1: District Map of Ghana (left), and Kumasi (right)

4.2.2 Cholera case definition and data

In Ghana, a suspected case of cholera is based on the WHO's definition which depends on whether or not the presence of cholera has been demonstrated in the area. According to the WHO (1993) guidance on formulation of national policy on the control of cholera, in an area where the disease is not known to be present a case of cholera should be suspected, when a patient, 5 years of age or older develops severe dehydration or dies from acute watery diarrhea, or where an epidemic is occurring, a patient, 5 years of age or older develops acute watery diarrhea, with or without vomiting. However, the first suspected case of cholera has to be confirmed by bacteriological tests (personal communication with the director of the Kumasi Metropolitan Health Directorate (KMHD)).

During the year 2005 severe outbreaks of cholera occurred in most urban communities in Ghana. In Kumasi, this outbreak started from the last week of September, lasting for a period of 72 days, which was within the rainy season. The first suspected and confirmed case was recorded on 29th September, 2005. The outbreak source was traced to a slum settlement (Racecourse) which is an abandoned racecourse. All cholera data for this study were obtained from the Kumasi Metropolitan Disease Control Unit (DCU), since it is mandatory for all reporting facilities (i.e. hospitals, clinics, and community volunteers) to report weekly cholera cases to the DCU. According to the DCU, cholera surveillance and reporting before 2005 has been ineffective, and hence the existing data before 2005 has little or no spatial information. However, with intensified surveillance and reporting systems during the 2005 outbreak, cases were recorded at community level (spatial unit of reporting). Therefore, this study utilised only cholera cases reported during the 2005 outbreak.

4.2.3 Refuse dumps data

A Global Positioning System (GPS) was used to determine the geographic coordinates of all refuse dumps. The geographic coordinates (latitudes and longitudes) in the WGS

84 datum were then transformed into the Ghana Transverse Mercator (GTM) coordinate system using a simple transformation program written in Microsoft Visual Basic. The GTM coordinates were then imported into a GIS for mapping and further analysis. A total of 124 refuse dumps were mapped. Based on the hypothesis that cholera is a disease of deficient sanitation, the following predictions were made: inhabitants who live in close proximity to open-space refuse dumps should have higher cholera prevalence than those farther. Also areas with high density of open-space refuse dumps should have higher cholera prevalence than areas with lower density.

4.2.4 Spatial data input

Topographic map of the study area at a scale of 1:2500 obtained from the planning unit of the Kumasi Metropolitan Assembly was digitized using ArcGis version 9.0 developed by Environmental System Research Institute (ESRI). Before digitizing, the map was georeferenced (by defining the X and Y coordinates of corner points of the map) into the GTM coordinates system. The main boundary was digitized as a polygon feature while the locations of communities were digitized as point features. Reported cases of cholera in 2005 obtained from the KMHD were entered as attributes of the point features. Population estimates for 2005, obtained from the Ghana Statistical Service (GSS), were used in calculating the raw rates of cholera. Raw rates were calculated as the number of cholera cases in each community divided by the estimated population in 2005. In order to express the notion of risk more intuitively, the raw rates were rescaled by multiplying it by a factor, i.e. 10,000. This expresses the raw rate as per 10,000 people. The resulting layer was then overlaid on the refuse dumps layer for further analysis.

4.2.5 Spatial data analysis and statistical modelling

House-to-house collection of waste in Kumasi is limited to only few first class residential areas. The inhabitants of the rest of the population exploit open spaces and approved demarcated parcels as refuse dumps. Due to the rate of urbanisation and population growth, most of these refuse dumps have now approached the centres of communities and overcrowded areas. Inhabitants in close proximity to refuse dumps are assumed to have a higher risk of contracting cholera. Spatial analysis was therefore used to determine the spatial relationship between cholera prevalence per community and (a) proximity (distances) to refuse dumps, (b) density of refuse dumps.

Spatial analysis was carried out in two principal steps. Firstly, two spatial factor maps were generated: (a) *spatial distance surface*, showing distances of each point (cell or pixel) to the nearest refuse dump (Figure 4.2); (b) *kernel density surface*, showing the number of refuse dumps per unit area (Figure 4.3). Kernel density calculates the density of point features around each output raster cell. In this concept, smooth curved surfaces are fitted over each point. The surface value is highest at the location of the point and diminishes with increasing distance from the point, reaching zero at a search radius distance from the point. The density at each output raster cell is calculated by adding the values of all the kernel surfaces where they overlay the raster cell centre. The kernel

function is based on the quadratic kernel function described in Silverman (1986). In this study, a search radius of 1km was used. The spatial factor maps were subsequently crossed with the point map of communities to create two spatial covariates: (a) *Distance to refuse dumps* $d_{(\text{dump})}$: distances from each community to the nearest refuse dump; (b) *Density of refuse dumps* $\rho_{(\text{dump})}$: number of refuse dumps per unit area for each community. Summary statistics of the variables used for modelling are shown in Table 4.1.

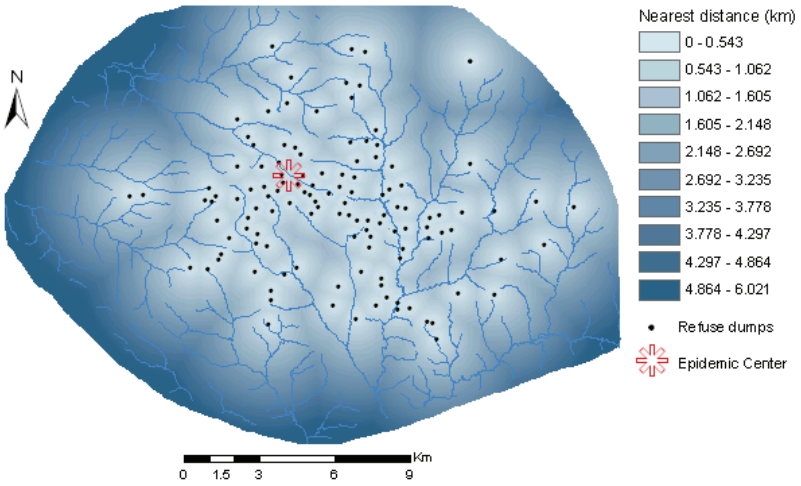

Figure 4.3: Distance surface (after neighbourhood statistics), showing distances from each pixel to the nearest potential cholera source (refuse dumps).

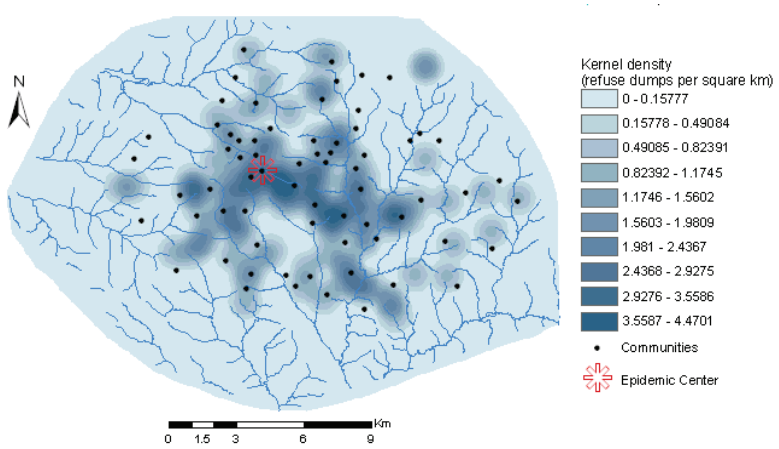

Figure 4.2: Kernel density surface (after neighbourhood statistics), showing the number of refuse dumps per unit area

Table 4.1: Summary statistics of variables used for the spatial modelling

Variable	Minimum	Mean	Maximum	Stan Dev
Incidence rate per 10,000 people	0.47	10.21	31.92	6.84
$d_{(\text{dump})}$ (m)	28.75	538.7	2375.5	446.22
$\rho_{(\text{dump})}$ (dumps per km^2)	0	0.29	4.14	0.59

Secondly, using the spatial covariates as explanatory variables, we developed a set of spatial models that attempted to relate cholera incidence rates to refuse dumps in Kumasi. Spatial regression methodologies in the context of spatial econometric framework (Anselin, 1988, 2002; Anselin and Bera, 1998) were. When standard linear regression, i.e. Ordinary Least Squares (OLS) models are estimated for cross-sectional data on neighbouring spatial units, the presence of spatial dependence may cause serious problems of model misspecification. The methodologies of spatial regression consist of examining and testing for the potential presence of such misspecifications and providing a more appropriate modelling that incorporates the spatial dependence (Anselin et al*.,* 1997; Varga, 1998).

In matrix notation, the general form of standard linear regression model is given by:

$$Chol_{(R)i} = d'_{(\text{dump})i} \cdot \beta_{(\text{dump})} + \varepsilon_{(\text{dump})i} \qquad 4.1$$

where $Chol_{(R)i}$ is cholera prevalence, $d_{(\text{dump})i}$ is an observation on the explanatory variable, with $i = 1.....,N$ (including a constant term, or 1), $\beta_{(\text{dump})}$ is the matching regression coefficient, and $\varepsilon_{(\text{dump})i}$ is a random error term. In this model, the error terms are assumed to have zero mean $E[\varepsilon_{(\text{dump})i}] = 0, \forall i$, and are identically and independently distributed (*i.i.d.*). Consequently, their variance is constant, $\text{Var}[\varepsilon_{(\text{dump})i}] = \sigma^2$, and they should be uncorrelated, i.e. $E[\varepsilon_{(\text{dump})i}\varepsilon_{(\text{dump})j}] = 0, \forall i, j$. These assumptions are usually violated due to the presence of spatial dependence (spatial autocorrelation) in the residuals of the standard regression model and consequently, model misspecification. Spatial dependence can be incorporated into the OLS model in two distinct ways: as an additional predictor in the form of a spatially lagged dependent variable (spatial lag model), or in the error structure (spatial error model).

In a spatial lag model, the spatial lag variable is introduced at the right hand side of equation (4.1) as:

$$Chol_{(R)i} = \rho \cdot Chol^*_{(R)i} + d'_{(\text{dump})i} \cdot \beta_{(\text{dump})} + \varepsilon_{(\text{dump})i}, \qquad 4.2$$

where ρ is an autoregressive coefficient of the lag variable $Chol^*_{(R)i}$. The spatial lag variable in the model can be expressed as: $Chol^*_{(R)i} = \sum_j Com_{ij} \cdot Chol_{(R)i}$, where Com_{ij} is

row-standardised spatial weight matrix corresponding to the community pair i, j; hence $\sum_j Com_{ij} = 1, \forall i$.

The spatial error variable Φ can also be introduced in the standard regression model as:

$$Chol_{(R)i} = d'_{(\text{dump})i} \cdot \beta_{(\text{dump})} + \Phi_{(\text{dump})i}, \qquad 4.3$$

with $\Phi_{(\text{dump})i} = \lambda \sum_j Com_{ij} \cdot \Phi_{(\text{dump})i} + \nu_{(\text{dump})i}$, where $\Phi_{(\text{dump})i}$ is the error vector, λ is the spatial autoregressive coefficient and $\nu_{(\text{dump})}$ is a random error term, assumed to be *i.i.d*. The errors $\Phi_{(\text{dump})i}$ are assumed to follow a spatial autoregressive process with autoregressive coefficient λ.

Model estimation

The parameters for the standard regression model, equation (4.1), can either be estimated using OLS estimation method or maximum likelihood (ML) method (that is, the parameters are estimated by maximizing the probability/likelihood of the sample data) since we are assuming that the data follow a multivariate Gaussian distribution, i.e. $Chol_{(R)} \sim MVN\left(d_{(\text{dump})} \cdot \beta_{(\text{dump})}, \sigma^2 I\right)$. The ML estimator maximizes the probability of obtaining the data over the set of possible values of the model parameters. The log-likelihood function for the OLS model is given as:

$$\ln L\left(\beta_{(\text{dump})}, \sigma^2 \left| Chol_{(\text{R})}, d_{(\text{dump})}\right.\right) = -\left(N/2\right)\ln\left(2\pi\right) - \left(N/2\right)\ln\sigma^2 - \left(1/2\sigma^2\right) \times \left(Chol_{(\text{R})} - d_{(\text{dump})} \cdot \beta_{(\text{dump})}\right)' \left(Chol_{(\text{R})} - d_{(\text{dump})} \cdot \beta_{(\text{dump})}\right) \qquad 4.4$$

Likewise, the spatial lag and error models are estimated by means of the maximum ML method. The ML estimation of spatial lag and spatial error regression models were first outlined by Ord (1975). The point of departure is an assumption of normality for the error terms $\varepsilon_{(\text{dump})}$ in the standard regression model. Unlike what holds for the standard regression model, the joint log-likelihood for a spatial regression does not equal the sum of the log likelihoods associated with the individual observations. Since $\varepsilon_{(\text{dump})} \sim MVN\left(0, \Sigma_\sigma\right)$, it follows that, with $\varepsilon_{(\text{dump})} = Chol_{(R)} - d_{(\text{dump})} \cdot \beta_{(\text{dump})}$ and a spatial autoregressive variance-covariance matrix $\sum = \sigma^2[\left(I - \lambda \cdot Com\right)' \left(I - \lambda \cdot Com\right)]^2$, the log-likelihood function for the spatial lag model is:

$$\begin{aligned}\ln L\left(\beta_{(\text{dump})},\sigma^2,\rho\middle|Chol_{(\text{R})},d_{(\text{dump})}\right) = &-(N/2)\ln(2\pi)-(N/2)\ln\sigma^2+\\ &\ln\left|I-\rho\cdot Com\right|-(1/2\sigma^2)\times\\ &\left(Chol_{(\text{R})}-\rho\cdot Com\cdot Chol_{(\text{R})}-d_{(\text{dump})}\cdot\beta_{(\text{dump})}\right)'\times\\ &\left(Chol_{(\text{R})}-\rho\cdot Com\cdot Chol_{(\text{R})}-d_{(\text{dump})}\cdot\beta_{(\text{dump})}\right)\end{aligned} \quad 4.5$$

where I is the N by N identity matrix.

The first conditions for the ML estimators yield nonlinear (in parameters) equations which are solved by numerical methods. For a ML estimate for ρ it is obtained from a numerical optimization of the concentrated log-likelihood function.

The maximum likelihood estimation for the spatial error model employs the error covariance term into log-likelihood function as follows:

$$\begin{aligned}\ln L\left(\beta_{(\text{dump})},\sigma^2,\lambda\middle|Chol_{(R)},d_{(\text{dump})}\right) = &-(N/2)\ln(2\pi)-(N/2)\ln\sigma^2+\\ &\ln\left|I-\lambda\cdot Com\right|-(1/2\sigma^2)\times\\ &\left(Chol_{(R)}-d_{(\text{dump})}\cdot\beta_{(\text{dump})}\right)'(I-\lambda\cdot Com)'\times\\ &\left(I-\lambda\cdot Chol_{(R)}\right)\left(Chol_{(R)}-d_{(\text{dump})}\cdot\beta_{(\text{dump})}\right)\end{aligned}. \quad 4.6$$

As in spatial lag model, the ML estimate can also be solved numerically and the estimates are obtained from the optimization of a concentrated log-likelihood function.

In a similar way, the standard regression, spatial lag, and spatial error models were fitted for $\rho_{(\text{dump})}$ using equations (4.1), (4.2) and (4.4).

Model specification and selection

A widely used diagnostic test for spatial error dependence is an extension of Moran's I to the regression context. In addition, Anselin (1988) provides the best guidance for model specification based on the joint use of the Langrage Multiplier (LM) tests for spatial lag and spatial error dependence. When Moran's I statistic for the error terms of the standard regression model is significant, LM test for spatial lag and spatial error dependence is used. When both tests have high values indicating significant spatial dependence in the data, the one with the highest value (lowest probability) will indicate the proper specification.

The test statistics for spatial error dependence is constructed from the standard regression residuals, and it is given by:

$$I = \varepsilon'_{(\text{dump})} \cdot Com \cdot \varepsilon_{(\text{dump})} \Big/ \varepsilon'_{(\text{dump})} \cdot \varepsilon_{(\text{dump})} \,, \tag{4.7}$$

where $\varepsilon_{(\text{dump})}$ is an N by 1 vector of regression residuals from the standard regression estimation. Inference is based on the normal distribution.

The LM-lag statistic has the following form:

$$LM_{\rho} = \frac{\left[\left(\varepsilon'_{(\text{dump})} \cdot Com \cdot Chol_{(R)}\right) \Big/ \left(\varepsilon'_{(\text{dump})} \cdot \varepsilon_{(\text{dump})} / N\right)\right]^2}{\left[\left(Com \cdot d_{(\text{dump})} \cdot \hat{\beta}_{(\text{dump})}\right)' M \left(Com \cdot d_{(\text{dump})} \cdot \hat{\beta}_{(\text{dump})}\right) \Big/ \hat{\sigma}^2 \right] + T}, \tag{4.8}$$

where $T = tr(Com' \cdot Com + Com \cdot Com)$, with tr as the matrix trace operator (the sum of the diagonal elements of a matrix), and $M = I - d_{(\text{dump})} \left(d'_{(\text{dump})} \cdot d_{(\text{dump})}\right)^{-1} d'_{(\text{dump})}$ is the projection matrix. The statistic is distributed as $\chi^2(1)$ with one degree of freedom.

The LM-error test for spatial error dependence was also suggested by Burridge (1980) and Anselin (1988), and has the following form:

$$LM_{\lambda} = \frac{\left[\varepsilon'_{(\text{dump})} \cdot Com \cdot \varepsilon_{(\text{dump})} \Big/ \left(\varepsilon'_{(\text{dump})} \cdot \varepsilon_{(\text{dump})} / N\right)\right]^2}{T} \,. \tag{4.9}$$

This statistic is also distributed as $\chi^2(1)$ with one degree of freedom. The best model that fits the data was based on the computed test statistics, selected using a step by step procedure shown in the flow chart below (Figure 4.4).

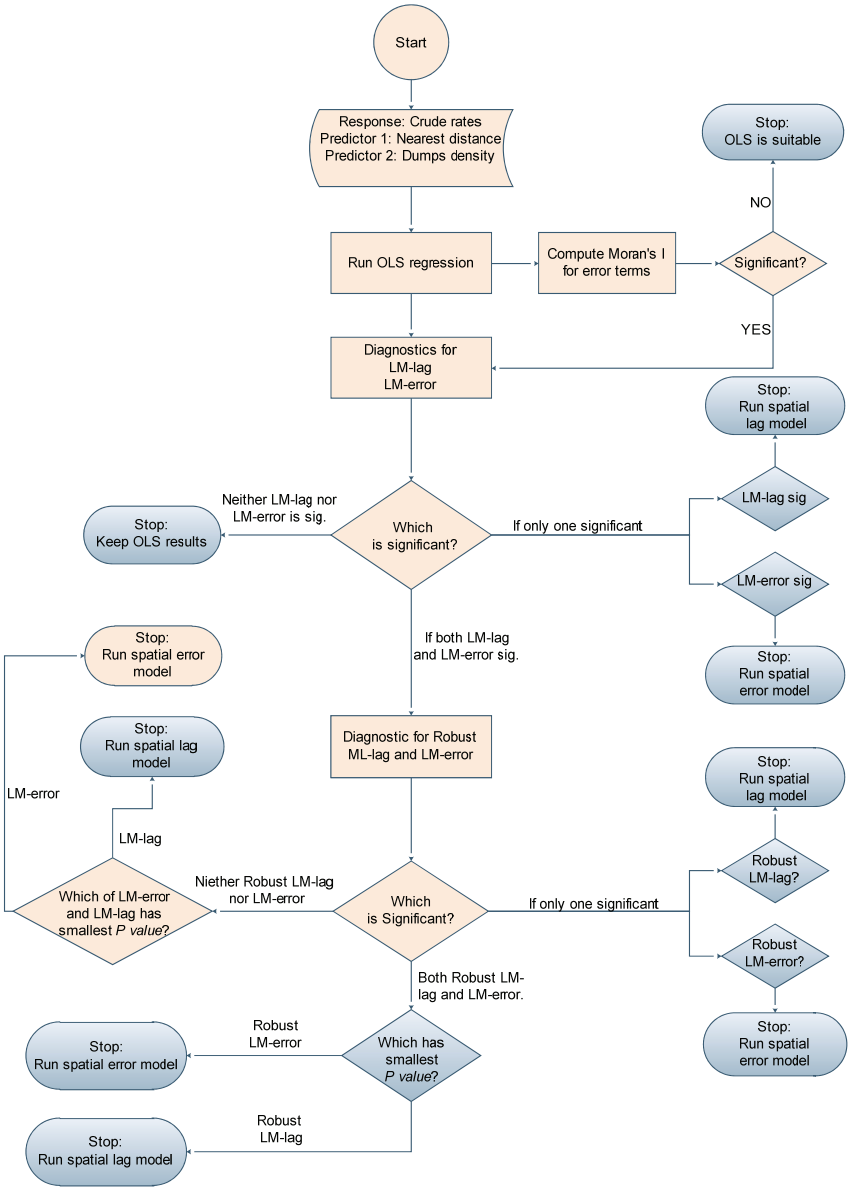

Figure 4.4: Flow chart showing step by step decisions for the spatial modelling. The shapes in pink show the decisions made to select the appropriate model that best fit our data.

4.2.6 Spatial clusters detection

One of the most important statistical tools for cluster detection, spatial scan statistic was used to detect most likely clusters for both high and low rates of cholera incidences. The

spatial scan statistic has a disadvantage of being difficult to incorporate prior knowledge about the size and shape of an outbreak and its impact on disease rate (Neill et al., 2005). We took advantage of this disadvantage to get rid of pre-selection biases of clusters and their locations. The spatial scan statistic is based on the likelihood ratio test, and analyses have been available for count data using either a Poisson or Bernoulli model (Kulldorff, 1997). The spatial scan statistic method is based on the principle that the number of cholera cases in a geographic area is Poisson-distributed according to a known underlying population at risk (Kulldorff, 2006). A spatial scan statistic software, SaTScan (Kulldorff, 2005), was used for all cluster analysis. A circular window was imposed on the study region which is moved over the region and centred on the centroid of each community. The size of the circular window or the cluster size was expressed as a percentage of the total population at risk. This varied from 0 to a maximum (not exceeding 100%), as specified by the user, expressed as a percentage of the population at risk. In this study, the retrospective spatial cluster analysis for both high and low rates was used, that is when SaTScan evaluates all temporal windows less than the specified maximum window size.

The maximum window size never exceeded 50% of the total population because clusters of larger sizes would indicate areas of exceptionally low rates outside the circle rather than areas of exceptionally high rates within the circle. Possible clusters were tested within the window whenever it was centred on the centroid of each community. Whenever the window finds a new case, the software calculates a likelihood function to test for elevated risk within the window in comparison with those outside the window. The likelihood function for any given window was proportional to:

$$L(\mathrm{W}) = \sup_{W \in \mathbf{W}} \left(\frac{Chol_{(C)}(W)}{Chol_{(\mathrm{E(C)})}(W)} \right)^{Chol_{(\mathrm{C})}(W)} \left(\frac{Chol_{(\mathrm{C})}(\hat{W})}{Chol_{(\mathrm{E(C)})}(\hat{W})} \right)^{Chol_{(\mathrm{C})}(\hat{W})} \times I\left(\frac{Chol_{(\mathrm{C})}(W)}{Chol_{(\mathrm{E(C)})}(W)} > \frac{Chol_{(\mathrm{C})}(\hat{W})}{Chol_{(\mathrm{E(C)})}(\hat{W})} \right), \quad 4.10$$

where $\hat{W}$ indicates all the regions outside the window W, and $Chol_{(\mathrm{C})}(\,)$ and $Chol_{(\mathrm{E(C)})}(\,)$ denote the observed and expected number of cases within the specified window, respectively. The indicator function $I(\,)$ is 1 when $Chol_{(\mathrm{C})} > Chol_{(\mathrm{E(C)})}$, otherwise 0. The test of significance level of clusters is through the Monte Carlo hypothesis testing (Dwass, 1957). This was used to test for the significance of the cluster that is least likely to have occurred by chance. The null hypothesis of no cluster was rejected when the simulated *p-value* was less than or equal to 0.05 for most likely clusters and 0.1 for secondary clusters since the latter have conservative *p-values* (Kulldorff, 2006).

To investigate whether proximity and density of open-space refuse dumps was associated with cholera within the cluster with higher than expected cholera prevalence,

(primary cluster); a similar modelling approach was used (See *Spatial data analysis and statistical modelling*).

4.2.7 Critical buffer distance

A major significance of this research is to give recommendation to public health officials and town planners about the critical (minimum) distance within which refuse dumps should not be sited away from inhabitants of communities. A GIS based buffer analysis and statistical analysis was used to estimate this critical distance. In this analysis, a buffer zone was defined as a specified distance around a selected map feature (Duker et al., 2004). Firstly, a series of *Boolean-distance* maps for experimental buffer distances from 200 m to 1000 m at regular buffer intervals of 100 m were created around refuse dumps. Boolean maps were created in such a way that distances less or equal to a buffer distance were considered high risk zones, whereas distances greater than the buffer distance were considered as low risk zone. Within each buffer zone, the mean cholera prevalence rate was computed. At each buffer zone, a test of the significance of the difference of mean cholera prevalence within the buffer and outside the buffer was calculated using the *t*-statistic:

$$t_{ij} = \frac{\overline{Chol}_{(R)i} - \overline{Chol}_{(R)j}}{\sqrt{s_p^2 (1/n_i + 1/n_j)}}, \qquad 4.11$$

where $\overline{Chol}_{(R)i}$, $\overline{Chol}_{(R)j}$ are sample means of cholera incidence rates, s_p^2 is the polled sample variance, n_i and n_j the sample sizes from population i and j. Using t_{ij} and degrees of freedom given by $n_i + n_j - 2$, a t distribution look-up table provides the probability p that the means are significantly different.

4.3 Results and analysis

4.3.1 Association between cholera and refuse dumps

The summary statistics of the study variables are shown in Table 4.1. For the period understudy, cholera incidence rates ranged from 0.47 to 31.92 per 10,000 people (*mean* = 10.21, *standard deviation* = 6.84). High incidence rates seem to have occurred at the central part of Kumasi (Figure 4.5). The results of the spatial regression models are shown in Table 4.2. Both spatial covariates i.e. nearest distance and dumps density were significantly correlated with cholera incidence. Both the spatial lag and error models are a significant improvement on the OLS model (see Table 4.2). Comparing ρ, λ and the Akaike Information Criterion values (AIC), the spatial error model best fits both covariates. As was expected, a direct spatial relationship between cholera prevalence and *dumps density*, and an inverse relationship with *nearest distance* was observed.

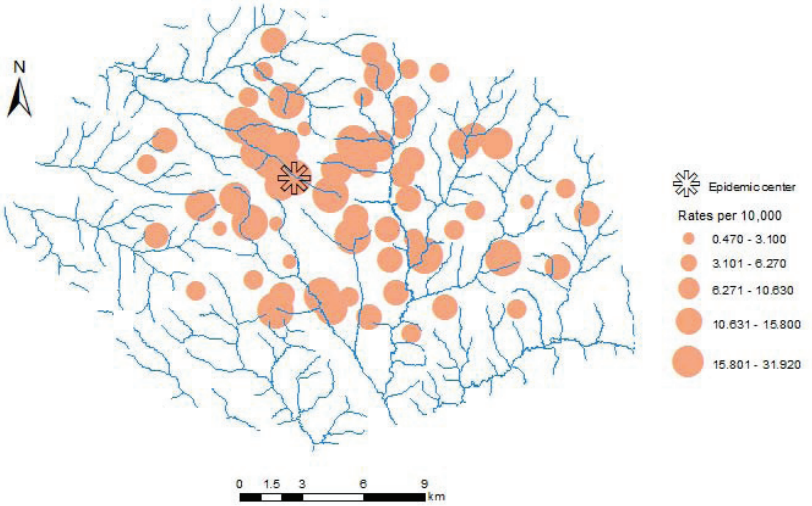

Figure 4.5: A proportional symbol map showing cholera prevalence for each community.

Table 4.2: Results of the spatial regression models; *P*-values are shown in brackets

	$d_{(dump)}$			$\rho_{(dump)}$		
Estimation	OLS	Spatial Lag	Spatial Error	OLS	Spatial Lag	Spatial Error
Constant	12.50‡	16.91‡	12.50‡	8.13‡	12.63‡	8.24‡
β	-0.0013†	-0.0013†	-0.0013†	1.86†	2.03†	1.81†
R^2	0.08†	0.16†	0.18†	0.07†	0.16†	0.17†
σ^2	44.57	39.50	38.64	45.00	39.32	38.69
ρ, λ	*	0.41†	0.47†	*	0.44†	0.48†
AIC	453.15	450.62	447.56	453.81	450.57	447.78
LM_ρ	*	3.581†	*	*	4.00†	*
LM_λ	*	*	4.049†	*	*	4.54†

* Not available; ‡$p < 0.001$; †$p < 0.05$

4.3.2 Cholera incidence clusters

High and low rate spatial clusters of cholera were detected within different window sizes using spatial scan statistic employed in the SaTScan software. Four statistically significant ($p < 0.001$) spatial clusters were detected when the maximum cluster size was ≤ 50% of the total population. The primary cluster (most likely cluster), which was the largest cluster, had greater than expected cholera prevalence rate. This cluster encompassed 23 communities in the study region, where about 27% of the people reside. The overall relative risk $Chol_{(RR)} = 1.790$, with $Chol_{(C)} = 376$ compared with $Chol_{(E(C))} = 255$. This cluster was in areas surrounding the central part of the study

region. Of the three secondary clusters with lower than expected prevalence rates, two encompassed a community each, whiles the third encompassed 17 communities where about 30% of the people reside. This cluster surrounded communities located at the south eastern part of the study region (Figure 4.6 and Table 4.3).

When spatial models were built within the cluster with higher than expected cholera prevalence, a result similar to the model within the whole study region was obtained. Cholera prevalence was positively associated with density of refuse dumps ($R^2 = 0.10$, $p < 0.1$), and negatively associated with proximity to refuse dumps ($R^2 = 0.24$, $p < 0.01$). However, neither the spatial lag nor spatial error variables were included in this model due to the absence of spatial dependency in the residuals. Hence OLS model best fitted the cluster with higher than expected prevalence.

Table 4.3: Results of cholera clusters using spatial scan statistics

Cluster	No. Communities	No. of cases	Exp. No. cases	Population	RR
1‡	23	376	254.52	268295	1.790
2‡	1	1	20.19	21281	0.049
3‡	17	214	279.97	295113	0.696
4‡	1	23	53.52	56417	0.416

‡$p < 0.01$

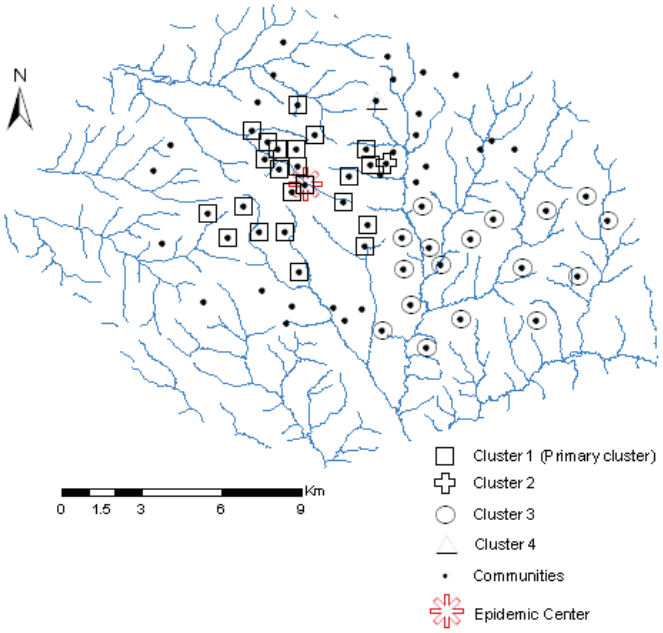

Figure 4.6: Results of spatial cluster analysis. This map shows the primary cluster with greater than expected cholera prevalence, and two secondary clusters with lower than expected cholera prevalence.

4.3.3 Critical buffer distance

The result of the *t*-tests for significant spatial determination of critical buffer distance is shown in Table 4.4 and Figure 4.6. For the buffers tested, *t-values* ranged from 8.86 to 0.88 whereas mean cholera prevalence ranged from 21.67 to 4.5 per 10,000. A quantitative assessment of distance discrimination of the experimental buffer zones around refuse dumps shows that the optimum spatial discrimination of cholera occurs at 500 m from refuse dumps (Figure 4.7). With this buffer 42 of the 68 communities (i.e., 62%) fall within the 500m distance to refuse dumps. For communities within this buffer, cholera prevalence ranged from 31.92 to 3.93 cases per 10,000 people, and mean prevalence of 11.77. For communities beyond this buffer, cholera prevalence ranged from 21.73 to 1.3 cases per 10,000 people, and mean prevalence of 7.71 (Table 4.4).

Table 4.4: Results of buffer distances and associated *p-values*

Distance (m)	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
Mean within	12.56	12.77	11.9	11.77	11.82	11.49	10.72	10.54	10.63
Mean outside	9.81	8.48	7.95	7.71	7.07	7.17	8.28	8.67	7.77
t-statistic	1.164	2.51	2.42	2.44	2.83	2.44	1.18	0.86	1.22
p-value	0.12	0.007	0.009	0.007	0.003	0.008	0.12	0.20	0.114

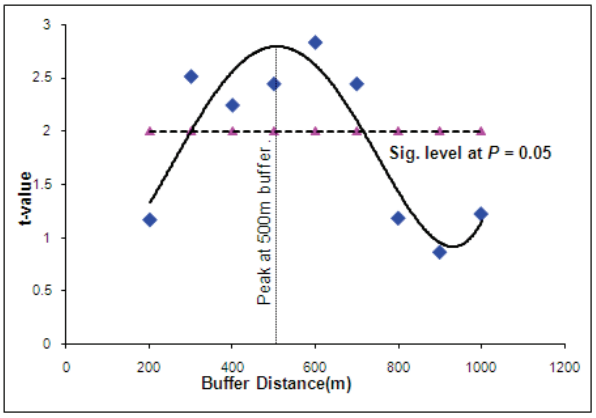

Figure 4.7: Quantitative assessment of critical distance discrimination obtained by applying experimental buffer zones around refuse dumps

4.4 Discussion

4.4.1 Association between cholera and refuse dump

The results of our spatial regression models suggest that proximity and density of refuse dumps are the important ecologic predictors of cholera (epidemic) in Kumasi. These findings are explicitly new, however, they support previous hypothesis of some reviews.

Cholera has been generally hypothesized as a disease of deficient sanitation. Since proximity and density of refuse dumps can serve as index of basic sanitation within an area, our findings support the general hypothesis of cholera. Two main reasons may explain the plausibility of our findings.

High rate of contact with filth breeding flies

Flies are attracted by the odour emanating from refuse dumps, especially the common housefly. This fly lives in close association with man feeding on all kinds of human food, garbage and excreta, and will travel no farther from its breeding site (refuse dumps) to the nearest resting place. The indiscriminate feeding habits (feeding on filth and human food) of this fly species combined with its structural morphology (presence of hair and sticky pads) make them ideally suited to carry and disseminate pathogenic micro organisms (Greenberg, 1973; Fotedar et al., 1992b; Kobayashi et al., 1999). Research has proven that the common housefly (*Musca domestica vicina*) and flies in general are mechanical vectors of many kinds of pathogens such as bacterial (Levine and Levine, 1990; Cohen at al., 1991), protozoa (Fotedar et al*.,* 1992a), viruses (Ogata et al*.,* 1961), and helminth eggs (Sulaiman et al*.,* 1988; Dipeolu, 1982). Fotedar (2001) undertook a study to ascertain the vector potential of the domestic housefly as a carrier of *V. cholerae* in Delhi, India, where an outbreak of cholera was encountered. Viable *V. cholerae* was isolated from six (60%) of the pooled fly samples, which confirmed that there were potentially contaminated mechanical vectors among the flies. Some published reports have also shown that fly control measures can be effective in reducing the incidence of diarrhoea (Watt and Lindsay, 1948; Cohen et al., 1991; Chavasse et al., 1999).

Flood water contamination

Significant amount of human excreta ultimately reaches solid waste systems through dump diapers, faeces of children, or even adults faeces are directly added to the solid waste in the homes. Some people also defecate along roadways, streets, and areas which are swept by public sweepers. These faecal matters also end up in the solid waste. Most often, the excreta of young children are also considered to be harmless, and hence end up in solid waste systems. Etiological studies have shown that *V. cholerae* survives well in faecal specimens if kept moist (Sack et al*.,* 2004). In the period of cholera outbreak, runoff from open space dumps during heavy rains may serve as the major pathway for the distribution of the bacteria, creating environmental niches for the bacteria infection. Excreta may be washed away by rain-water and can run into wells, streams and surface water bodies. The bacteria in the excreta may then contaminate these water bodies. The Kumasi metropolis suffers from frequent sporadic water shortages. During such periods, the people exploits nearby streams and surface water bodies for cooking, drinking, bathing and other activities.

4.4.2 Cholera incidence clusters

Four statistically significant spatial clusters of cholera were identified, one significant most likely cluster with greater than expected prevalence rates, and three secondary clusters with lower than expected prevalence rates. The most likely cluster which was the largest cluster (See Table 4.3 and Figure 4.5) encompassed 23 communities, where 27% of the people reside. For this cluster, 21 out of the 23 communities were within the critical buffer distance (i.e. within 500m of refuse dumps). This indicates that proximity to refuse dumps might be a significant contributory factor to the high rates of cholera. Also, this cluster was concentrated within the central part of the study area, around the epidemic centre (Racecourse, Bantama), i.e. where the epidemic began. The reasons for this could be several. Firstly, cholera is a contagious disease which can diffuse from its source of outbreak to other communities based on their proximity; hence high incidence rates are likely to occur at areas proximal to the outbreak source than areas farther. Secondly, the outbreak source, Racecourse is a waterlogged area and a forcibly created market centre with high commercial activities. This area generally has insufficient public toilets and garbage bins to accommodate the large daily influx of traders. Consequently, traders and buyers resort to unsanitary practices, urinate and defecate into open gutters and other open spaces and into polythene bags, which ultimately creates unsanitary conditions for livelihood. Since cholera outbreak is a stark reminder of deficiency in good sanitation practices, the patterns displayed may be partly explained by the above mentioned reasons. The cluster with lower than expected prevalence rates was concentrated at the south eastern part of Kumasi. Most of the communities within this cluster are residential and peri-urban areas where population density and commercial activities are relatively minimal. Also, house to house collection of refuse has been successfully achieved at the residential communities, hence dumping of refuse in open spaces is scarcely found

4.4.3 Critical buffer distance

A major significance of this research was to give recommendation to public health officials and town planners about the critical (minimum) distance within which dumps should not be sited away from inhabitants of communities. From our results, it is evident that the critical buffer distance within which refuse dumps should not be sited is 500 m (see Figure 4.6). This buffer distance included about 62% of communities, where about 68% of the people reside.

The city's expansion both spatially and in population has strained existing resources meant to achieve effective waste management systems (example *house to house* collection of waste). This has led to the creation of many open space dumps very close to community centres. Consequently, about 62% of communities have their refuse dumps within the critical buffer distance (*buffer distance with the lowest p-value*). Since cholera outbreak is an indication of poor sanitation, opens-space refuse dumps within community centres predispose inhabitants to cholera infection.

This present study provides useful information about the location of clusters of cholera and an ecological factor that might have led to increased cases during the period of outbreak. This new knowledge, the spatial dependency of high cholera prevalence on refuse dumps location will be useful to health officials and policy makers to make appropriate decisions. With the critical buffer distance identified in this study, it cautions policy makers not to locate any open-space refuse dump within 500 m radius from the centre of any community.

This study also has some potential limitations. Firstly, the data used is for only a single year outbreak. The best approach was to use data from several cholera outbreaks; however, cholera reporting at community level has not been available before the 2005 outbreak. Secondly, the assumption was that the population within a community has equal risk of exposure to refuse dumps. In reality, this is not so because within a particular community, individuals living close to refuse dumps have a higher risk of exposure than those living farther away. Thirdly, this study could not correlate periods of water shortages with the outbreak period of cholera.

4.5 Conclusion

The results of this study reveal the spatial dependency of cholera infection upon proximity and density of refuse dumps in Kumasi. This means that refuse dumps serve as niches for cholera infection. The results also show that the minimum distance within which refuse dumps should not be located is 500 m. We therefore hypothesize that proximity and density of refuse dumps may play a significant role in cholera transmission. House-to-house collection of refuse, which is at a limited service, should therefore be extended to all communities within the Kumasi metropolis.

5

Spatial dependency of cholera on potential cholera reservoirs

"Rivers always receive the refuse of those living on the banks, and they nearly always supply, at the same time, the drinking water of the community so situated"

John Snow

In the previous chapter, the results suggest that proximity and density of open space refuse dumps play a contributory role in cholera infection in Kumasi. Could the increased cholera prevalence near dump sites be explained by increased transmission through flood water contamination in the proximity of the dump sites? The chapter seeks to delineate water bodies possibly contaminated by runoff from refuse dumps, and assess their impact on cholera prevalence. This chapter has originally been published as: Osei FB, Duker AA, Augustijn E-W and Stein A: Spatial dependency of cholera prevalence on potential cholera reservoirs in an urban area, Kumasi-Ghana. *International Journal of Applied Earth Observation and Geoinformation* 2010, 12:5

Abstract

Cholera has been a public health burden in Ghana since the early 1970's. Between 1999 and 2005, a total of 25,636 cases and 620 deaths were officially reported to the WHO. In one of the worst affected urban cities, faecal contamination of surface water is extremely high, and the disease is reported to be prevalent among inhabitants living in close proximity to surface water bodies. Surface runoff from dump sites is a major source of faecal and bacterial contamination of rivers and streams in the study area. This study aims to determine (a) the impacts of surface water contamination on cholera infection and (b) detect and map arbitrary shaped clusters of cholera. A Geographic Information System (GIS) based spatial analysis is used to delineate potential reservoirs of the cholera *vibrios*; possibly contaminated by surface runoff from open space refuse dumps. Statistical modelling using OLS model reveals a significant negative association between (a) cholera prevalence and proximity to all the potential cholera reservoirs (R^2 = 0.18, $p < 0.001$) and (b) cholera prevalence and proximity to upstream potential cholera reservoirs (R^2 = 0.25, $p < 0.001$). The inclusion of spatial autoregressive coefficients in the OLS model reveals the dependency of the spatial distribution of cholera prevalence on the spatial neighbours of the communities. A flexible scan statistic identifies a most likely cluster with a higher relative risk (RR = 2.04, $p < 0.01$) compared with the cluster detected by circular scan statistic (RR = 1.60, $p < 0.01$). We conclude that surface water pollution through runoff from waste dump sites play a significant role in cholera infection.

5.1 Introduction

Cholera is listed as one of three internationally quarantainable diseases by the World Health Organization (WHO), along with plague and yellow fever (WHO, 2000a). Infection is mainly through the ingestion of sufficient dose of cholera *vibrios* through contaminated food and/or water (Sack et al., 1998; Hornick et al., 1971). The cholera *vibrios* constitute part of the normal aquatic flora in estuarine environments (Colwell et al., 1985; Colwell and Huq, 1994; Frauque et al., 2005); the most significant reservoir is brackish water (Colwell and Spira, 1992). Growth and survival of the cholera *vibrios* are influenced by abiotic parameters such as availability of organic nutrients (Singleton et al., 1982a, 1982b; Lipp et al., 2001). Surface runoff from point sources may cause increased organic nutrients concentration, and faecal contamination of rivers (Servais et al., 2007). Stagnation and slow flowing of rivers may lead to increased exposure to cholera *vibrios* (Ali et al., 2002). Faecal contamination of surface water during outbreak periods enhances the risk of exposure to the cholera *vibrios*.

Cholera has been a public health burden in Ghana since its introduction in the early 1970's (Pobee et al., 1970; Kwofie, 1976). Between 1999 and 2005, a total of 25,636 cases and 620 deaths were officially reported to the WHO by the Ghana Ministry of Health (WHO, 2000b, 2001, 2002, 2003, 2004, 2005, and 2006). The disease has received little attention from research. Available health research focuses solely on the biological aspects and characteristics of the individuals contracting the disease (see for example Obiri-Danso et al., 2003, 2005; Opintan et al., 2008), and neglect the spatial patterns of transmission. However useful these studies are, they cannot establish accurate exposure levels for the critical risk factors of the disease. Over the years, cholera in Ghana has predominated in the densely populated urban communities. These communities have limited sewage and sanitation facilities. Consequently, indiscriminate defecation practice, especially at waste dump sites, is common. Surface runoff from waste dumps may cause faecal contamination of surface water bodies, stagnation, and increased organic nutrients. In turn, these may put inhabitants at risk of waterborne diseases such as cholera.

Previous studies investigated cholera outbreaks between 1997 and 2001 in the Ashanti Region, Ghana. The findings revealed two important characteristics of the disease in Ashanti Region: (1) a non random distribution of cholera occurred in and around Kumasi Metropolis; (2) high cholera prevalence was associated with high urbanization and high overcrowding (Osei and Duker, 2008a). A second study, carried out in the Kumasi Metropolis, exploited the importance of two main spatial measures of sanitation in cholera infection, namely proximity to, and density of, refuse dumps within communities. The study found a statistically significant association between cholera prevalence and both dump density and proximity to dump sites (Osei and Duker, 2008b). The above study spelt out two main hypotheses to explain the plausibility of the findings; 1: *Contact with Filth Breeding Flies*, and 2: *Flood Water Contamination*. Since fly-control studies were not undertaken during the period of the outbreak, the study could not confirm fly transmission of cholera *vibrios* to humans, but only gave an indication of their possible involvement in transmission. No attempt, however, was made to ascertain the effect of runoff from waste dumps on cholera. It was questioned

whether the increased cholera prevalence near dump sites could be explained by increased transmission through flood water contamination in the proximity of the dump sites. Also, the relationship between cholera prevalence and proximity to potentially contaminated surface water bodies was as yet unclear. An objective answer to the second question is essentially needed to answer the first. The rational of this study has been to delineate water bodies possibly contaminated by runoff from waste dumps, and assess their impact on cholera prevalence. If runoff from waste dumps during heavy rains serve as the major pathway for faecal and bacterial contamination of rivers and streams, then it is likely that inhabitants living closer to water bodies where these runoffs flow into will have higher cholera prevalence than those who live farther.

Osei and Duker (2008b) utilized a spatial scan statistic methodology developed by Kulldorff (1997) to detect and map clusters of cholera. Although the significance of Kulldorff's scan statistics is widely acknowledged in the field of spatial epidemiology, the method imposes a circular window to define the potential cluster areas (Kulldorff and Nagarwalla, 1995), and thus has difficulty in correctly detecting actual noncircular clusters (Tango and Takahashi, 2005). Duczmal and Assunção (2004) introduced a simulated annealing spatial scan statistic that detects noncircular connected clusters with arbitrary shapes. Clusters thus detected, however, are much larger than the expected true cluster (Tango and Takahashi, 2005). Tango and Takahashi (2005) proposed a flexibly shaped spatial scan statistic that imposes an irregularly shaped window on each region connecting its adjacent regions. This approach is able to detect arbitrarily shaped clusters, and this statistic is well suited for detecting and monitoring disease outbreaks in irregularly shaped areas. Owing to socio-economic and environmental factors associated with cholera, clusters are likely to be irregular in shape following the patterns of these factors.

The main objectives of this study are to: (a) determine the spatial relation between possibly contaminated water bodies and cholera prevalence; (b) detect and map arbitrary shaped clusters of cholera. In this study, we concentrate on regression models that include interaction between spatial neighbors. We apply these models to determine the spatial dependency of cholera prevalence on potentially polluted surface water bodies. The flexible spatial scan statistic is utilized to detect and map arbitrary shaped clusters of cholera. These will help public health officials in Kumasi to track the underlying processes of cholera infection for guiding intervention strategies.

5.2 Materials and methods

5.2.1 Study framework and methodology

In spatial epidemiological studies, the closest link to an assumed biological model is achieved by using data from a variety of points that describe the exact spatial locations of cases/events and exposure factors. As such, the average disease risk of an individual reflects its level of exposure to the risk factors. Due to several limitations in disease surveillance systems, especially in Ghana, individual level data are rarely available. Hence, individual level studies are almost unfeasible in such situations. Case-control

and cohort studies can give a relatively close approximation to the biologic model because they both provide point data that describes individual level characteristics. These studies are expensive and time consuming to carry out, and are not feasible in all situations. For example, in retrospective studies of infectious diseases where the recovery period is short, it is impossible to trace the affected individuals. Exploratory studies using aggregated data, such as ecological or geographic correlation studies, offer an alternative approach for analyzing aggregated epidemiological data to address specific hypothesis of disease causation. Although they too are prone to some biases and misclassifications, the so called *ecological bias* (Elliot and Wakefield, 2000), they are easier, quicker, and less expensive to conduct (Elliott and Wartenberg, 2004). Therefore, this study is undertaken within the framework of small area ecological studies where group level characteristics are the focus rather than individual level.

Geographic Information Systems (GIS) are widely utilized by environmental epidemiologists and medical geographers to assess the association between disease distribution and exposure to environmental risk factors (Clarke et al., 1996; Vine et al., 1997; Moore and Tim, 1999; Glass, 2000; Rushton, 2003; Elliott and Wartenberg, 2004; Jarup, 2004; Nuckols et al., 2004; Rezaeian et al., 2007). Several studies, including works of Kwofie (1976), Borroto and Martinez-Piedra (2000), Ali et al. (2002a, 2002b), Glass et al., (1982) and Fleming et al. (2007) have successfully utilized the concepts and methodologies of GIS to understand the spatial variations of cholera. In this study, we use a GIS based spatial analysis as a tool to delineate potential cholera reservoirs, and assess their impacts on cholera prevalence through spatial regression modelling approaches.

5.2.2 The study area

The Kumasi Metropolis lies at the intersection of latitude 6.04^{o}N and longitude 1.28^{o}W, covering an area of about 220 km^2 (See Figure 5.1). It is the most populous Metropolis in Ashanti Region. As described in Osei and Duker (2008b), Kumasi has a population of about 1.2 million which accounts for just under a third (i.e. 32.4%) of Ashanti region's population. It is the regional capital, and the most commercialized city in the region. Kumasi is a nodal city linking major road arteries to the northern and southern parts of Ghana. There are two main seasons, the rainy season and the dry season. The rainfall pattern is bimodal with long rainy season peaks in May/June and a short season peaks in September/October. Approximately 82% of the inhabitants in Kumasi have access to potable, pipe-borne, water. Surface water from rivers and streams is, however, still largely used for cooking, bathing and washing utensils due to the rampant water shortages. The coverage rate of safe house-to-house collection of waste is still very low. A greater proportion of households in the metropolis, approximately 81.2%, dispose of solid waste at open space refuse dumps (PHC, 2000). High housing cost, coupled with high demand for accommodation has made most landlords convert sanitation rooms into sleeping rooms for renting. Therefore, most inhabitants have no access to household sanitation system. Less than 30% of the inhabitants in Kumasi have access to Water Closet (WC) toilet system (PHC, 2000). Most demarcated areas for public sanitation and waste disposal facilities have been sold out due to the high demand for land. The absence of both household and public sanitation systems, especially in highly populated

areas and squatter settlements, compels inhabitants to defecate at open space refuse dumps.

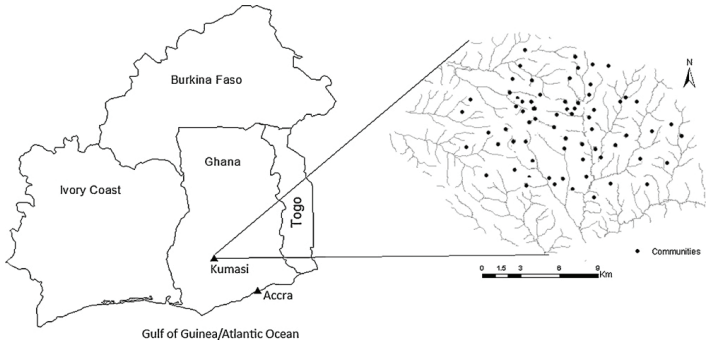

Figure 5.1: District map of Ghana (left), and Kumasi (right). Dots indicate the centroids of communities.

5.2.3 Spatial data input

In this study, the main sources of surface water pollution are the open space refuse dumps. A Global Positioning System (GPS) was used to locate the geographic positions of refuse dumps sites. The latitudes and longitudes in the WGS 84 datum were transformed into the Ghana Transverse Mercator (GTM) coordinate system. A topographic map of the study area at a scale of 1:2500 obtained from the planning unit of the Kumasi Metropolitan Assembly was digitized using ArcGIS version 9.2. Before digitizing, the map was georeferenced by defining the X and Y coordinates of corner points of the map into the GTM coordinates system. From the topographic map, rivers and streams were digitized as line segments, as were elevation contours. Values of elevations contours were input as spatial attributes. The main boundary was digitized as a polygon, whereas locations of communities representing centroids of densely populated areas within the community were digitized as point features. Reported cases of cholera during the 2005 epidemic outbreak obtained from the Kumasi Metro Health Directorate (KMHD) were entered as attributes of the communities. The detailed spatial information contained of those data is the community of residence of the affected individual. Cholera prevalence per community was calculated as the number of cases in a community divided by the population in 2005.

5.2.4 Delineation of potential cholera reservoirs

To evaluate runoff patterns from dump sites to the streams and rivers, a steepest downhill path analysis was conducted using a 3D surface model created from the elevation contours. A steepest path is a line that follows the steepest downhill direction on a surface; the path may terminate at the surface perimeter or in surface concavities. An overlay operation was subsequently performed to delineate all drainage segments which intersected with the steepest paths. This resulted in a layer containing all stream segments assumed to be potential cholera reservoirs. Strahler's (1952) method of stream order classification was used to delineate all 1st, 2nd and 3rd order stream segments which

intersected with the steepest paths. In this method, the smallest headwater tributaries are called 1st order streams. Where two 1st order streams meet, a 2nd order stream is created; where two 2nd order streams or 1st and 2nd order steams meet, a 3rd order stream is created; and so on. Although the water quality characteristics of these delineated stream segments have not been measured in this study, they are thought to be potential cholera reservoirs (Obiri-Danso et al., 2005). Hereafter, we refer to the potential cholera reservoirs as reservoirs.

5.2.5 Spatial factors maps

Three spatial factor maps were created: (a) spatial distance surface, showing distances of each point to the nearest reservoir that describes proximity to all delineated stream segments; (b) spatial distance surface, showing distances of each point to the nearest reservoir that describes proximity to 1st order stream segments, the so called upstream reservoirs; (c) spatial distance surface, showing distances of each point to the nearest reservoir that describes proximity to 2nd and 3rd order stream segments, the so called downstream reservoirs. To create the spatial factor maps, a spatial neighbourhood statistics was performed on the distance surface maps to calculate the mean pixel values within a neighbourhood of 1 km radius. The purpose of this was to obtain the mean distance within each community to the nearest reservoir. Since information about the exact location of cholera cases is not known, the mean distances best represent a measure of proximity to the reservoirs; assuming that inhabitants within each community have equal risk of exposure. Mean distances to reservoirs were used as a surrogate measure of exposure to cholera. The spatial factor maps were subsequently overlaid with the point map of communities to create three explanatory variables: (a) *proximity to all reservoirs* ($d_{(\text{All reser})}$); (b) *proximity to upstream reservoirs* ($d_{(\text{Up reseer})}$); and (c) *proximity to downstream reservoirs* ($d_{(\text{Dw reser})}$) (Figure 5.2). To simplify notations, $d_{(\text{All reser})}$, $d_{(\text{Up reseer})}$ and $d_{(\text{Dw reser})}$ will be denoted in the subsequent sections as $d_{(\text{All})}$, $d_{(\text{Up})}$ and $d_{(\text{Dw})}$, respectively.

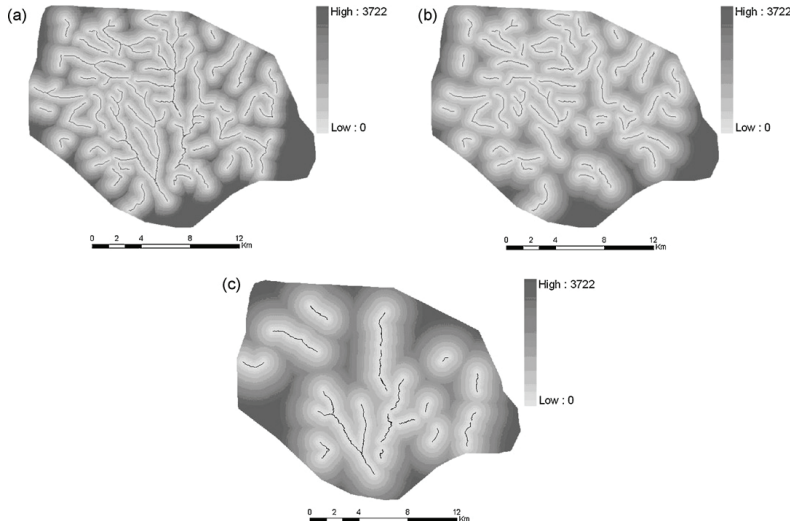

Figure 5.2: Distance surface map showing distance from each pixel to the nearest potential cholera reservoir: a (proximity to all reservoirs, $d_{(\mathrm{All})}$); b (proximity to upstream reservoirs, $d_{(\mathrm{Up})}$); c (proximity to upstream reservoirs, $d_{(\mathrm{Dw})}$)

5.2.6 Proximity analysis: regression modelling

Using the explanatory variables described above, we developed several regression models relating cholera prevalence to the reservoirs. The assumption of independence, identically and normally distributed error terms in Ordinary Least Squares (OLS) regression models are often violated due to the presence of spatial autocorrelation; consequently leading to model misspecifications. Therefore, spatial regression methodologies in the context of spatial econometric framework (Anselin and Bera, 1988; Anselin, 1988, 2002) were used. Spatial regression modelling tests the potential presence of such misspecifications, providing a more appropriate modelling that incorporates spatial autocorrelation. The modelling proceeds as follows. The OLS estimation model equals:

$$Chol_{(R)i} = d'_{(\mathrm{All})i} \cdot \beta_{(\mathrm{All})} + \varepsilon_{(\mathrm{All})i}, \qquad 5.1$$

with i indexing observations at $i = 1,...,N$ communities. The response variable $Chol_{(R)i}$ is an observation of cholera incidence rate for community i, $\beta_{(\mathrm{All})}$ is the matching regression coefficient for the distance variable $d_{(\mathrm{All})}$, and $\varepsilon_{(\mathrm{All})}$ is random error term. In this model, the error terms are assumed to have zero mean ($E[\varepsilon_{(\mathrm{All})i}] = 0, \forall i$), and are identically and independently distributed (*i.i.d.*). Consequently, their variance is

constant, $\mathrm{Var}[\varepsilon_{(\mathrm{All})i}] = \sigma^2$, and observations are supposed to be uncorrelated, i.e. $E[\varepsilon_{(\mathrm{All})i}\varepsilon_{(\mathrm{All})i}] = 0, \forall i, j$.

To include spatial autocorrelation, we identified two distinct ways: an additional predictor in the form of a spatially lagged dependent variable, i.e. spatial lag model; and an additional predictor in the error-covariance structure, i.e. spatial error model.

For the spatial lag model, the spatial lag variable is included at the right hand side of the OLS models as:

$$Chol_{(\mathrm{R})i} = \rho \cdot Chol^{*}_{(\mathrm{R})i} + d'_{(\mathrm{All})i} \cdot \beta_{(\mathrm{All})} + \varepsilon_{(\mathrm{All})i}, \qquad 5.2$$

where ρ is an autoregressive coefficient of the spatial lag variable $Chol^{*}_{(\mathrm{R})i}$. The motivation for the spatial lag specification is based on the possible effects of social interactions on cholera infection. The spatial lag variable in the model can be expressed as:

$$Chol^{*}_{(\mathrm{R})i} = \sum_j Com_{ij} \cdot Chol_{(\mathrm{R})j}, \qquad 5.3$$

where Com_{ij} is row-standardized spatial weight matrix corresponding to the community pair i, j; hence $\sum_j Com_{ij} = 1, \forall i$. The weights are supposed to decrease with increasing distance between pair of communities i and j. The weight matrix Com_{ij} is derived by equating the distance d_{ij} between communities i and j to d_i^k, being the distance from i and its kth nearest neighbour. The kth nearest neighbour (where k is the number of neighbors) is used to ensure equal number of neighbors for each community. Neighbors beyond this threshold are excluded. The diagonal elements Com_{ii} are set to zero to prevent an observation from predicting itself. Different weighing criteria (i.e. $Com_{ij} = 1$, $1/d_{ij}$ or $1/(d_{ij})^2$) may be used. In the present study, however, where the effect of spatial interaction is dependent on distance between pair of communities, the choice of $Com_{ij} = 1$ does not apply. Minor differences occurred between application of $1/d_{ij}$ and $1/(d_{ij})^2$, hence the most simple choice of $1/d_{ij}$ is used in the present study.

For the spatial error model, spatial autocorrelation is included in the covariance structure of the random error terms $\varepsilon_i^{(\mathrm{All})}$, following a spatial autoregressive process (SAR):

$$\varepsilon_{(\mathrm{All})i} = \lambda \sum_j Com_{ij} \cdot \varepsilon_{(\mathrm{All})j} + \nu_{(\mathrm{All})i}, \qquad 5.4$$

with λ as the spatial autoregressive parameter and $\nu_{(\mathrm{All})}$ as a random error term, assumed to be *i.i.d*. The motivation for the spatial error specification is that unmeasured effects spill over across units of observations and hence results in spatially correlated errors. Therefore, the spatial error specification serves as a surrogate measure of unmeasured risk factors of cholera.

The parameters of the spatial lag and error models are estimated by means of the maximum likelihood (ML) method, i.e. the parameters are estimated by maximizing the likelihood of the sample data. The log-likelihood function for the spatial lag equals:

$$\begin{aligned}\ln L\left(\beta_{(\mathrm{All})},\sigma^2,\rho\middle|Chol_{(\mathrm{R})},d_{(\mathrm{All})}\right) = &-(N/2)\ln(2\pi)-(N/2)\ln\sigma^2 + \\ &\ln\left|I-\rho\cdot Com\right|-\left(1/2\sigma^2\right)\times \\ &\left(Chol_{(\mathrm{R})}-\rho\cdot Com\cdot Chol_{(\mathrm{R})}-d_{(\mathrm{All})}\cdot\beta_{(\mathrm{All})}\right)'\times \\ &\left(Chol_{(\mathrm{R})}-\rho\cdot Com\cdot Chol_{(\mathrm{R})}-d_{(\mathrm{All})}\cdot\beta_{(\mathrm{All})}\right)\end{aligned}, \qquad 5.5$$

where *I* is a *N* by *N* identity matrix. The first order conditions for the ML estimators yield nonlinear equations which are solved by numerical methods. ML estimates for $\beta_{(\mathrm{All})}$, ρ and the variance σ^2 are obtained as solutions to the usual first order conditions, requiring numerical optimization. The log-Jacobian term $\ln\left|I-\rho\cdot Com\right|$ implies constraints on the parameter space for ρ, which must be such that $\left|I-\rho\cdot Com\right|>1$.

The maximum likelihood estimation for the spatial error model employs the error covariance term into log-likelihood function equals:

$$\begin{aligned}\ln L\left(\beta_{(\mathrm{All})},\sigma^2,\rho\middle|Chol_{(\mathrm{R})},d_{(\mathrm{All})}\right) = &-(N/2)\ln(2\pi)-(N/2)\ln\sigma^2 + \\ &\ln\left|I-\lambda\cdot Com\right|-\left(1/2\sigma^2\right)\times \\ &\left(Chol_{(R)}-d_{(\mathrm{All})}\cdot\beta_{(\mathrm{All})}\right)'\left(I-\lambda\cdot Com\right)'\times \\ &\left(I-\lambda\cdot Com\right)\left(Chol_{(R)}-d_{(\mathrm{All})}\cdot\beta_{(\mathrm{All})}\right)\end{aligned}. \qquad 5.6$$

As in the spatial lag model, the ML estimates are solved numerically and the estimates are obtained from the optimization of a concentrated log-likelihood function.

Anselin (1988) provides the best guidance for model specification based on the joint use of the Langrage Multiplier (LM) tests for spatial lag and spatial error dependence. The LM test for spatial lag and spatial error dependence is constructed from the OLS

residuals. This requires estimation of the model under the null hypothesis of no spatial dependence.

The LM test statistic for spatial lag dependence LM_ρ is expressed as:

$$LM_\rho = \frac{\left[\left(\varepsilon'_{(\mathrm{All})} \cdot Com \cdot Chol_{(R)}\right) \Big/ \left(\varepsilon'_{(\mathrm{All})} \cdot \varepsilon_{(\mathrm{All})} / N\right)\right]^2}{D}, \tag{5.7}$$

where $\varepsilon_{(\mathrm{All})}$ is a vector of OLS residuals, and the denominator term:

$$D = \left[\begin{array}{l} \left(Com \cdot d_{(\mathrm{All})} \cdot \hat{\beta}_{(\mathrm{All})}\right)' \left[I - d_{(\mathrm{All})}\left(d'_{(\mathrm{All})} \cdot d_{(\mathrm{All})}\right)^{-1} d'_{(\mathrm{All})}\right] \times \\ \left(Com \cdot d_{(\mathrm{All})} \cdot \hat{\beta}_{(\mathrm{All})}\right) \Big/ \hat{\sigma}^2 \end{array}\right] + T, \tag{5.8}$$

where $T = tr(Com' \cdot Com + Com \cdot Com)$, with tr as the matrix trace operator, and $I - d_{(\mathrm{All})}\left(d'_{(\mathrm{All})} \cdot d_{(\mathrm{All})}\right)^{-1} d'_{(\mathrm{All})}$ is the projection matrix. Estimates for $\hat{\beta}_{(\mathrm{All})}$ and $\hat{\sigma}^2$ are obtained from the OLS model. The statistic is asymptotically χ^2_1 distributed.

The LM test statistics for spatial error dependence LM_λ was also suggested by Burridge (1980) and Anselin (1988) as:

$$LM_\lambda = \frac{\left[\left(\varepsilon'_{(\mathrm{All})} \cdot Com \cdot \varepsilon_{(\mathrm{All})}\right) \Big/ \left(\varepsilon'_{(\mathrm{All})} \cdot \varepsilon_{(\mathrm{All})} / N\right)\right]^2}{T}. \tag{5.9}$$

This statistic is also asymptotically χ^2_1 distributed.

In a similar way, the OLS, spatial lag, and spatial error models were fitted for $d_{(\mathrm{Up})}$ and $d_{(\mathrm{Dw})}$ using equations (5.1), (5.2) and (5.4), respectively. Because the explanatory variables are correlated, no attempt was made to combine them in a single model.

5.2.7 Flexible spatial cluster analysis

The flexible scan statistic is based on the principle that the number of disease cases in each geographical area follows a Poisson distribution according to a known underlying population at risk. Flexible spatial scan statistic software, FleXScan (Takahashi et al., 2004), was used for all cluster analysis. Flexible scan statistic imposes an irregularly shaped window on each community by connecting its adjacent communities. For any

given community *i*, FleXScan creates a set of irregularly shaped windows with length *k* consisting of *k* connected communities including *i* and let *k* moves from 1 to a pre-set maximum *K*. To avoid detecting a cluster of an unlikely peculiar shape, the connected communities are restricted as the subsets of the set of communities *i* and (*K - 1*)-nearest neighbors to the community *i* where *K* is a pre-specified maximum length of cluster (*K* is also the window size, which is proportional to the population at risk). In effect a very large number of different but overlapping arbitrarily shaped windows are created. For community *i*, the flexible scan statistic considers *K* concentric circles plus all the sets of connected communities, including community *i*, whose centroids are located within the *K*th largest concentric circle. Let $W_{ik(j)}, j = 1,..., j_{ik}$ denote the *j*th window which is a set of *k* communities connected starting from the community *i*, where j_{ik} is the number of *j* satisfying $W_{ik(j)} \subseteq W_{ik}$ for $k = 1,..., K$. Then, all the windows to be scanned are included in the set:

$$\mathbf{W} = \left\{ W_{ik(j)} \middle| 1 \le i \le m, 1 \le k \le K, 1 \le j \le j_{ik} \right\}. \tag{5.10}$$

Therefore, the flexible scan statistic considers *K* concentric circles plus all the sets of connected regions, including community *i*, whose centroids are located within the *K*th largest concentric circle. Under the alternate hypothesis, there is at least one window *W* for which the underlying risk is higher inside the window when compared with outside. For each window, FleXScan computes the likelihood of the observed number of cases within and outside the window under the Poisson assumption. The test statistic is constructed with the likelihood ratio test as:

$$L(\mathrm{W}) = \sup_{W \in \mathbf{W}} \left(\frac{Chol_{(\mathrm{C})}(W)}{Chol_{(\mathrm{E(C)})}(W)} \right)^{Chol_{(\mathrm{C})}(W)} \left(\frac{Chol_{(\mathrm{C})}(\hat{W})}{Chol_{(\mathrm{E(C)})}(\hat{W})} \right)^{Chol_{(\mathrm{C})}(\hat{W})} \times I\left(\frac{Chol_{(\mathrm{C})}(W)}{Chol_{(\mathrm{E(C)})}(W)} > \frac{Chol_{(\mathrm{C})}(\hat{W})}{Chol_{(\mathrm{E(C)})}(\hat{W})} \right), \tag{5.11}$$

where $\hat{W}$ indicates all the regions outside the window *W*, and $Chol_{(\mathrm{C})}(\,)$ and $Chol_{(\mathrm{E(C)})}(\,)$ denote the observed and expected number of cases within the specified window, respectively. The indicator function $I(\,)$ is 1 when the reported number of cases within the window is more than the expected number of cases, and 0 otherwise. The window *W** that attains the maximum likelihood is defined as the most likely cluster (MLC). The test of significance level of clusters is through the Monte Carlo hypothesis testing. Probability values (*p-values*) are obtained by comparing the test statistic from the observed data set with the test statistics from 999 random data sets generated under the null hypothesis of no clustering. In this study, the null hypothesis of no cluster was rejected when the simulated *p-value* was less than or equal to 0.05.

5.3 Results and analysis

5.3.1 Dependency of cholera on $d_{(\mathrm{All})}$, $d_{(\mathrm{Up})}$ and $d_{(\mathrm{Dw})}$

A preliminary analysis shows no significant relationship between *Chol* and proximity to the surface water bodies ($R^2 = 0.02,\ p = 0.23$). As expected, the OLS model shows a significant negative relationship for both $d_{(\mathrm{All})}$ and $d_{(\mathrm{Up})}$ (Table 5.1). The distance variable $d_{(\mathrm{Up})}$ explains a greater percentage of the variations in cholera prevalence than $d_{(\mathrm{All})}$. This is shown by the relatively higher correlation coefficient between $Chol_{(\mathrm{R})}$ and $d_{(\mathrm{Up})}$ ($R^2 = 0.25,\ p < 0.001$) compared with $d_{(\mathrm{All})}$ ($R^2 = 0.18,\ p < 0.001$). No significant relationship is observed between $Chol_{(R)}$ and $d_{(\mathrm{Dw})}$ ($R^2 = 0.003, p = 0.70$), suggesting that the impact of proximity to downstream reservoirs on cholera prevalence is not significant. In what follows, only models fitted for $d_{(\mathrm{All})}$ and $d_{(\mathrm{Up})}$ are considered.

Table 5.1: Results of OLS, spatial lag and spatial error models

	$d_{(\mathrm{All})}$			$d_{(\mathrm{Up})}$			$d_{(\mathrm{Dw})}$		
Estimation	OLS	Lag	Error	OLS	Lag	Error	OLS	Lag	Error
Constant	15.81‡	21.15‡	16.83‡	16.73‡	21.40‡	16.98‡	9.69§	12.73§	9.73§
R^2	0.18‡	0.27‡	0.32‡	0.25‡	0.32‡	0.34‡	0.003§	0.05§	0.05§
β	-0.015‡	-0.017‡	-0.018‡	-0.015‡	-0.015‡	-0.015‡	0.0006§	0.0006§	0.0006§
ρ, λ	*	0.46†	0.62†	*	0.41†	0.51†	*	0.29§	0.30§
σ^2	38.53	34.18	31.78	36.47	31.89	30.84	46.72	44.57	44.52
AIC	445.28	441.21	435.25	439.51	435.51	431.25	458.38	458.03	455.98
LM_{ρ}	*	4.77†	*	*	4.20†	*	*	1.98§	*
LM_{λ}	*	*	5.81$^{\pm}$	*	*	4.67†	*	*	2.01§

* Not available; $^{\ddagger}p < 0.001$; $^{\pm}p < 0.01$; $^{\dagger}p < 0.05$; $^{\S}p > 0.1$

Table 5.2: Comparison of results with different number of neighbors (k) for the spatial lag and spatial error models; the explanatory variable used is $d_{(\mathrm{Up})}$

Spatial lag model for $d_{(\mathrm{Up})}$					
k	β	R^2	ρ	AIC	SC
4	-0.0152 (0.001)	0.300 (0.001)	0.310 (0.069)	437.37	444.030
5	-0.0154 (0.001)	0.319 (0.001)	0.413 (0.033)	435.51	442.171
6	-0.0151 (0.001)	0.319 (0.001)	0.426 (0.051)	436.18	442.840
7	-0.0151 (0.001)	0.319 (0.001)	0.426 (0.052)	436.18	442.839
8	-0.0151 (0.001)	0.317 (0.001)	0.500 (0.053)	436.47	443.128
Spatial error model for $d_{(\mathrm{Up})}$					
k	β	R^2	λ	AIC	SC
4	-0.0152 (0.001)	0.327 (0.001)	0.378 (0.041)	433.752	438.190
5	-0.0154 (0.001)	0.359 (0.001)	0.510 (0.017)	431.251	435.690
6	-0.0150 (0.001)	0.340 (0.001)	0.520 (0.031)	432.601	437.040
7	-0.0150 (0.001)	0.342 (0.001)	0.520 (0.032)	432.601	437.040
8	-0.0147 (0.001)	0.320 (0.001)	0.550 (0.058)	434.394	438.830

Since the effects of neighbourhood structure on the results of spatial modelling depends on the number of neighbors, we investigated the sensitivity of the results to different number of neighbors. Similar patterns of variations exist for the distance variables $d_{(\mathrm{All})}$, $d_{(\mathrm{Up})}$ and $d_{(\mathrm{Dw})}$. Hence only the results for $d_{(\mathrm{Up})}$ is presented (Table 5.2). Minor differences are observed between the estimated parameters for the different number of neighbors. The p-values of the autoregressive coefficients (λ in spatial error and ρ in spatial lag) however are smaller for $k = 5$ in both the spatial lag and spatial error models. Hence our choice of $k = 5$ is appropriate.

The LM tests for spatial lag and spatial error dependence for $d_{(\mathrm{All})}$ and $d_{(\mathrm{Up})}$ indicate a significant influence of spatial autocorrelation in the OLS models. Fitting spatial regression models show significant autoregressive coefficients λ and ρ for both $d_{(\mathrm{Up})}$ than $d_{(\mathrm{All})}$. This suggests that interaction between neighbors significantly affects the prevalence of cholera. Between the spatial lag and the error models, the latter may be preferred because of the relatively smaller variances (σ^2) and Akaike Information Criterion (AIC) values.

For 44 out of 68 communities $d_{(\mathrm{All})} = d_{(\mathrm{Up})}$, thus inducing a significant correlation between $d_{(\mathrm{All})}$ and $d_{(\mathrm{Up})}$ ($R^2 = 0.50, p < 0.001$). The OLS model fitted for the 44 communities where $d_{(\mathrm{All})} = d_{(\mathrm{Up})}$ shows a higher negative association between $Chol_{(\mathrm{R})}$ and $d_{(\mathrm{Up})}$ ($R^2 = 0.31, p < 0.001$). Likewise, similar OLS model fitted amongst the 24 communities where $d_{(\mathrm{All})} \neq d_{(\mathrm{Up})}$ shows a significant negative relationship between $Chol_{(\mathrm{R})}$ and $d_{(\mathrm{Up})}$ ($R^2 = 0.15, p < 0.05$). No significant relationship is observed between $Chol_{(\mathrm{R})}$ and $d_{(\mathrm{All})}$ ($R^2 = 0.015, p = 0.56$) for communities where $d_{(\mathrm{All})} \neq d_{(\mathrm{Up})}$.

5.3.2 Cluster analysis

A flexible scan statistic with a window size of 20% of the total population identified a significant ($p < 0.01$) MLC with greater than expected prevalence. This cluster contains 12 communities in the study region (Figure 5.3b). The relative risk $Chol_{(RR)} = 1.97$, with $Chol_{(C)} = 275$ compared with $Chol_{(E(C))} = 140$. One statistically significant ($p < 0.01$) secondary cluster with $Chol_{(RR)} = 2.04$, $Chol_{(C)} = 63$ and $Chol_{(E(C))} = 31$ also contains three communities. Some other secondary clusters were detected, none being significant. When repeating the analysis using a circular scan statistic of the same window size, a relatively larger significant ($p < 0.01$) MLC was detected. This cluster contains 18 communities, including the 12 communities identified by the flexible scan statistics, with $Chol_{(RR)} = 1.6$, $Chol_{(C)} = 306$ and $Chol_{(E(C))} = 194$. Only one significant secondary cluster was detected (Figure 5.3a). This is an indication that imposing a circular window overestimates the real shape and size of the true clusters.

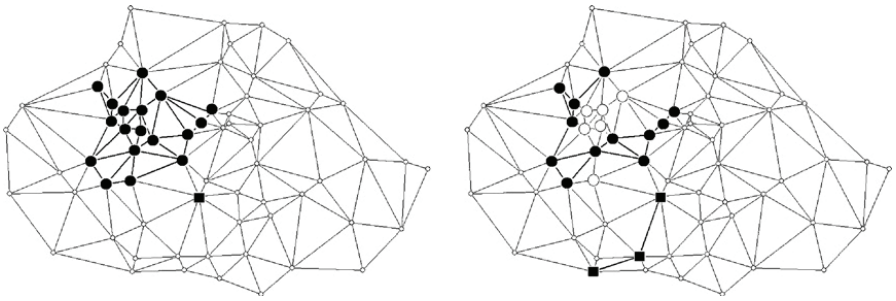

Figure 5.3: This map shows results of the clusters detected by circular scan statistics (Left: 3a) and flexible scan statistics (Right: 3b). Circular spots show the locations of the most likely clusters and rectangular spots for locations of secondary clusters. The circular rings show the locations of clusters detected by circular scan statistics but not detected by flexible scan statistics.

5.4 Discussion

The results of the study show a significant relationship between cholera prevalence and proximity to the potentially polluted water bodies. Significant amount of human excreta reaches waste dump sites due to the indiscriminate defecation practices of inhabitants. Surface runoffs from these dumps sites serve as a major pathway for faecal and bacterial contamination of rivers and streams. The runoffs also carry high organic loads, leading to stagnation and increased salinity of rivers and streams, thus creating suitable environmental niches for the cholera *vibrios*. For instance, a microbial water quality analysis of an urban river, also delineated as a potential cholera reservoir, has been shown to be polluted with faecal material with total califorms varying from 1.61×10^{9} to 4.06×10^{13} per 100 ml. Bacterial counts are also reported to be significantly higher during the rainy season compared with the dry season (Obiri-Danso et al., 2005). Use of

these water bodies for bathing, washing, cooking and other household activities is assumed to be greater for people who live closer to them (Ali et al., 2002a), and thus, will have higher cholera prevalence than those who live farther.

Proximity to upstream reservoirs has a stronger relationship with cholera prevalence compared with proximity to downstream reservoirs. This is shown by the relatively higher correlation between $Chol_{(R)}$ and $d_{(Up)}$ compared with $d_{(All)}$, and the non-significant relationship between $Chol_{(R)}$ and $d_{(Dw)}$. The relatively higher correlation amongst communities where $d_{(All)} = d_{(Up)}$, and the non-significant effects of $d_{(All)}$ amongst communities where $d_{(All)} \neq d_{(Up)}$ reaffirms this. This may be explained by the relatively higher likelihood of usage of upstream reservoirs compared with downstream reservoirs. The downstream reservoirs receive the pollution drained into the upstream reservoirs, of which the quality deteriorates as they flow downstream. This results in distasteful appearance (black colour and floating waste) and emanating bad odour; thus discouraging inhabitants from usage. The upstream reservoirs have clear colour which makes them appear clean and drinkable; therefore there is higher likelihood of usage for household activities during periods of water shortages.

Social interaction and unobserved confounders could induce spatially correlated effects in the spatial distribution of diseases; hence the inclusion of the spatial autoregressive coefficients in the spatial models is used as surrogate measures of: (a) the effects of social interaction on cholera infection, and (b) possible unobserved risk factors of cholera. The significance of the autoregressive coefficients ρ in the spatial lag models suggests the effects of interaction amongst communities on cholera prevalence. Communities which are closer in space tend to have similar cholera prevalence compared with communities farther apart. This suggests that similar environmental or demographic risk factors induce cholera transmission amongst the communities. The significance of the autoregressive coefficients λ in the spatial error models suggests the availability of important unobserved risk factors of cholera. These risk factors may be local or global depending on the spatial extent at which the spatial dependency persists. Further statistical analysis of the spatial auto-covariance structure of cholera prevalence is required to unveil the spatial nature of the unobserved risk factors. An alternative explanation for the significance of the autoregressive coefficients λ and ρ is the possible mismatch between the observed spatial unit and the true spatial scale of cholera transmission in the study area. Further analysis using household level data will provide invaluable information about the spatial transmission patterns of cholera.

The flexible scan statistic identified a significant MLC within the central part of Kumasi. This cluster encompassed relatively less number of communities but a higher RR than the MLC cluster detected by the spatial scan statistics. This suggests that imposing a circular window for cluster analysis may unduly include areas which are less likely to be hot-spots. For instance, six communities within the MLC of the circular scan statistic were not detected as hot-spots of cholera using the flexible scan statistic. These communities have a relatively lower mean prevalence of 6.27 as against 15.23 for circular scan statistic and 19.20 per 10,000 for flexible scan statistic. The MLC of the

flexible scan statistic seems to be close and enclose these six communities, yet they are not detected as hot-spots (Figure 5.3). This suggests that the characteristics that put inhabitants at increased risk of cholera may be more local than global. In fact about 94% of the potential cholera reservoirs lie within 1 km radius of communities, and thus influence individual communities rather than groups of communities. The arbitrary nature of the clusters detected by the flexible scan statistic may also follow patterns of socio-economic and environmental factors. For instance, the distribution and frequency of potable water supply to various communities may be implicated. Extensive data, however, are required to be able to make conclusive evidence.

The findings of this study reveal that cholera infection is enhanced by proximity to potentially polluted surface water bodies; and therefore, the increased prevalence near dump sites, as noted in Osei and Duker (2008b), can be explained by increased infection through flood water contamination as a result runoff from dump sites. The relationship between cholera and potentially polluted surface water bodies is stronger than that between cholera and dump sites (comparing the current study with Osei and Duker, 2008b). This, though arguable, is an indication that the effects of dump sites on cholera infection require surface water as an intermediate pathway. Therefore, any attempt to prevent defecation at dumps sites will reduce faecal contamination of rivers and streams. This will in turn reduce cholera infection during any outbreak. These findings support initial findings by other researchers. The classic epidemiological work of John Snow revealed the association between cholera and contaminated water even before any bacteria were known to exist (Snow, 1855). The epidemiological studies of Ali et al. (2002a, 2002b) have also reported on proximity to surface water bodies as an important cholera risk factor in an endemic area of Bangladesh. The current study, however, utilizes a GIS based spatial analyses to identify the steepest downhill paths along which runoff from dumps sites will flow. Methodologically, this study has improved on the existing techniques that environmental epidemiologists and medical geographers utilize to measure risk of exposure to an environmental determinant. The use of GIS and spatial analysis facilitates this type of methodological analysis which would be impossible in a non-spatial environment. The findings provide relatively detailed background information about the spatial characteristics of cholera, and its relationship with environmental risk factors in Kumasi. Such information will be useful to health officials and policy makers as background information for formulating measures to prevent future cholera outbreaks in Kumasi.

In this study, several assumptions are imposed by the available data, and therefore, the interpretation of the findings should be done within the framework of the limitations. First, the cholera case data used are count data aggregated at community level. This does not contain spatial information about the exact locations of the affected individuals. Consequently, the study assumes that inhabitants within each community have equal risk of exposure and, therefore, group levels of exposure represents individual levels of exposure. For that reason, the mean distance within a radius of 1 km^2 around the centroid of each community to potential cholera reservoirs is used as a surrogate measure of risk of exposure to cholera. This represents the exposure level of a group of individuals rather than that of individuals. Nevertheless, a number of epidemiological studies, including works of Ali et al. (2002a, 2002b) have used distance to surface water as a surrogate measure of exposure, and have successfully explained

the spatial epidemiology of cholera infection in Bangladesh. Others include works of Duker et al. (2004, 2006) in the epidemiology of Buruli Ulcer prevalence in Ghana. Notwithstanding the above limitations, this study has been able to address local issues that put inhabitants at risk of cholera. These findings prompt health official and policy makers to execute measures to prevent faecal contamination of surface water bodies in order to prevent future cholera outbreaks in Kumasi. In order to achieve this, the following strategies are suggested: (a) house-to-house collection of solid waste should be extended so as to reduce the dependency on open space dumps, (b) the Metropolitan Assembly should enforce the bylaws to prevent industrial pollution of surface water bodies; for example, the direct discharging of industrial waste from brewery companies, (c) the Metropolitan Assembly should increase the provision of good public sanitation facilities, such as flush toilets or water closets rather than the existing ventilated pit latrines. They should also draw and implement bylaws that will enforce landlords to provide toilet facilities for tenants in their houses, (d) above all, the Ghana Water Company Limited (GWCL) should improve on the water distribution system so as to ensure constant (24 hours a day) supply of treated piped water to all inhabitants in the metropolis.

5.5 Conclusion

This study uses statistical modelling to determine the dependency of cholera prevalence on contaminated surface water bodies. The findings reveal association between the spatial distribution cholera prevalence and proximity to contaminated surface water bodies. The flexible scan statistic reveals the existence of non circular clusters, with a relatively smaller size and higher RR than the clusters detected by the circular scan statistic. It is deduced that the spatial distribution of cholera prevalence is dependent on proximity to potentially contaminated surface water bodies and the spatial neighbors of communities. We conclude that proximity to potentially contaminated surface water bodies increases the risk of exposure to the cholera *vibrios*. The dependency of cholera prevalence on the spatial neighbors of communities indicates the existence of other confounding risk factors. Further studies, using purpose-collected household level data, will be very useful to elucidate all the critical risk factors of cholera in Kumasi.

6

Multivariate Bayesian semi-parametric modelling of cholera in an urban environment

". . . all models are wrong. The practical question is how wrong do they have to be to not be useful"

George Box and Norman Draper

In chapters 4 and 5, spatial statistical approaches was applied to model the effects of local environmental risk factors on cholera. Each chapter utilized a univariate spatial regression analysis to model the dependency of cholera on the risk factors. In order to bring the preceding chapters into coherence, this chapter develops a multivariate model using the multiple risk factors. The risk factors used are those identified in chapters 3 and 4, and other risk factors identified from literature. Since some of the risk factors are continuous and categorical, we apply a generalized structured additive regression modelling approach. Since such models are highly parameterized, a fully Bayesian estimation and inference based on Markov Chain Monte Carlo simulations is used. This chapter is under review in the journal of *Environmental and Ecological Statistics*, submitted as: Osei FB, Duker AA and Stein A: A multivariate Bayesian semi-parametric modelling for cholera risk in an urban environment

Abstract

This study develops a multivariate explanatory model for the risk of cholera infection in an urban area, Kumasi-Ghana. We apply a fully Bayesian semi-parametric regression modelling which allows joint analysis of nonlinear effects of continuous covariates, spatially structured variation, unstructured heterogeneity, as well as fixed covariates. Proximity to and density of dumps, and proximity to potential cholera reservoirs were modelled as smooth continuous functions; presence of slum settlers and population density were modelled as fixed effects, whereas spatial references to the communities were modelled as structured and unstructured spatial effects. We use a fully Bayesian estimation based on Markov Chain Monte Carlo (MCMC) simulations. The findings reveal that the risk of cholera infection is associated with slum settlements (*Posterior mean = 4.06, $p < 0.01$*) and high population density (*Posterior mean = 4.339, $p < 0.01$*). The relationship between cholera and dumps density is almost linear. The posterior mean of the proximity to dumps sites deviate from linearity, with a decreasing risk up to about 500 m, and a slightly increase for larger distances. The relationship between cholera and proximity to potential cholera reservoirs is almost linear, with the posterior means decreasing with increasing distance. We also observe distinct spatial variation in the risk of cholera infection, with evidence of significant increased cholera risk at the central part of Kumasi, and a significant reduced risk at the south-eastern part. These findings could serve as novel information to help health planners and policy makers in making effective decisions about cholera control measures.

6.1 Introduction

A significant interest in understanding the epidemiology of cholera lies in identifying associated risk factors which enhance the risk of infection, the so called *ecological studies* (Lawson et al., 1999; Gatrell and Bailly, 1996). Most ecological studies of cholera, however, make no, or limited use of the spatial structure of the data, as well as possible nonlinear effects of the risk factors. Thus, most studies utilize standard statistical methods such as the classical and generalized linear models that ignore methodological difficulties that arise from the nature of the data. Ali et al. (2001, 2002a, 2002b) have utilized logistic, simple and multiple regression models to study the spatial epidemiology of cholera in an endemic area of Bangladesh. Other ecological studies of cholera utilizing classical approaches include Ackers et al. (1998), Mugoya et al. (2005) and Sasaki et al. (2008). These methods when applied to spatially distributed data present severe problems with estimating small area spatial effects, and simultaneously adjusting for other risk factors, in particular when the effects of some risk factors are nonlinear. As noted by Cressie (1993), when standard statistical methods are used to analyze spatially correlated data, the standard error of the covariate parameters is underestimated and thus the statistical significance is overestimated.

Previous ecological studies have utilized spatial regression approaches to explore the dependency of cholera on local environmental risk factors (Osei and Duker, 2008b; Osei et al., 2010). Although the spatial regression approaches account for spatial dependency, it assumes a strictly linear relationship between the dependent variable and the predictor variable. Thus, the above studies did no account for the possible nonlinear effects of the risk factors. Also, no attempt has been made to combine all the identified risk factors into a single multivariate model to examine their joint effects on cholera. Generalized additive models (GAM) provide a powerful class of models for modelling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. Generalized structured additive regression (STAR) models are extensions of GAM models which allow one to incorporate small area spatial effects, nonlinear effects of risk factors, and the usual linear or fixed effects in a joint model.

This study applies a STAR modelling approach to develop a multivariate explanatory model for cholera. The study incorporates the effects of nonlinear risk factors and the usual fixed effects of some risk factors, while accounting for both structured and non structured spatial effects. A STAR model of this type has been termed *geoadditive* model (Kamman and Wand, 2003; Ruppert et al., 2003). The risk factors used are those identified from previous studies. These are density of refuse dumps, proximity to refuse dumps (Osei and Duker, 2008b), and proximity to potential cholera reservoirs (Osei et al., 2010). Other risk factors incorporated in the study are those identified from existing literature. These are livelihood at slummy environments and population density. Livelihood at slummy and squatter environments has been reported to increase the risk of cholera infection (Sur et al., 2005). High population density stresses existing sanitation systems, thus putting people at increased risk of cholera (Siddique et al., 1992; Root, 1997; Ali et al., 2002a, 2002b). Proximity to and density of dumps, and proximity to potential cholera reservoirs are modelled as smooth continuous functions; presence of slum settlers and population density are modelled as fixed effects, whereas

spatial references to the communities are modelled as structured and unstructured spatial effects. We use a fully Bayesian estimation based on Markov Chain Monte Carlo (MCMC) simulations, where probit link functions are represented by means of Gaussian distributed latent variables allowing simple Gibbs sampling updates. Making inferences based on a fully Bayesian approach is preferred because the functional of the posterior can be computed without relying on large Gaussian justifications, thereby quantifying the uncertainty in the parameters (Fahrmier et al., 2004). The deviance information criterion (DIC), as recently proposed by Spiegelhalter et al. (2002), is used for comparison and selection of competing models to assess model fit and complexity.

6.2 Methods

6.2.1 Cholera data and risk estimation

All cholera data for this study were obtained from the Kumasi (Figure 6.1) Metropolitan Disease Control Unit (DCU). Weekly reporting of cases to this unit is mandatory for all reporting facilities (i.e. hospitals, clinics, and community volunteers). The raw data consist of daily cases for every individual affected. They contain information such as name of patient, facility reporting, suburb and/or community of patient, date of onset and age of patient.

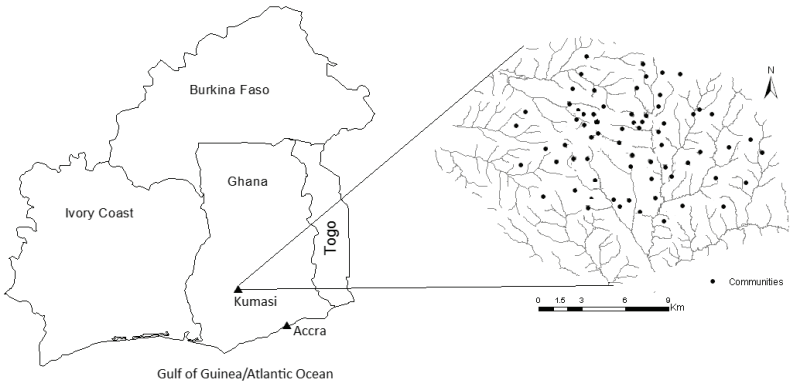

Figure 6.1: District map of Ghana (left), and Kumasi (right). Dots indicate the centroids of communities

Residential addresses were not recorded during the period of diagnoses; therefore we used the centroid of the communities as spatial references of the cases. We deleted cholera case records of communities outside the Kumasi metropolis, as well as the names of the patients from the data for security and privacy reasons. Data recording errors related to community names in the database were checked against the records in the spatial database. All case records for every community are given a unique identification code. These are cross-linked to the communities by unique identification codes to facilitate easy geo-referencing and analysis. The data are aggregated and restructured by community to describe the characteristics of the distribution of cholera incidence among the study subjects. The denominator (population data) for computing

community-specific cholera rates is obtained from the 2000 Population and Housing Census of Ghana (PHC, 2000).

We use the relative risk (also called excess risk) as the reference benchmark to estimate the risk of cholera infection. For each community i, $i = 1,...,N$ of population n_i, the observed number of cholera cases $Chol_{(C)}$ is assumed to be a realization of random variable that follows independent Poisson distribution with intensity $Chol_{(E(C))i} \cdot Chol_{(RR)i}$; thus: $Chol_{(C)i} \left| Chol_{(RR)i} \sim \text{Poisson}\left(Chol_{(E(C))i} \cdot Chol_{(RR)i}\right)\right.$, where $Chol_{(E(C))}$ is the expected number of cases and $Chol_{(RR)}$ is the relative risk. In standard principle, the maximum likelihood estimator for $Chol_{(RR)}$ is the ratio of observed number of cases to expected number of cases: $Chol_{(RR)i} = Chol_{(C)i} \big/ Chol_{(E(C))i}$. The expected number of cases is estimated as $Chol_{(E(C))} = Chol_{(\hat{R})} \cdot n_i$, where $\hat{Chol}_{(R)}$ a population weighted mean of cholera incidence rate, .i.e. an estimate of the overall risk of cholera within the study population. The population weighted mean $\hat{Chol}_{(R)}$ is obtained as a weighted average of the community-specific rates, each weighted by their share in the overall population, i.e. $\hat{Chol}_{(R)} = \sum_{i=1}^{N} \frac{Chol_{(C)i}}{n_i} \cdot \frac{n_i}{\sum_{i=1}^{N} n_i}$.

6.2.2 Continuous, spatial and categorical covariates

The continuous covariates used in this study are the *proximity to dumps* $d_{(dump)}$, the *density of dumps* $\rho_{(dump)}$, and the *proximity to potential cholera reservoirs* $d_{(reser)}$. These variables are extracted on per community basis via Geographic Information Systems (GIS) and spatial analysis. Details of the approaches for the calculation of these variables can be found in Osei and Duker (2008) and Osei et al. (2010). Here, $d_{(reser)}$ is used to represent *proximity to upstream potential cholera reservoirs* $d_{(Up)}$ is used as. The spatial locations of the communities are used to model the spatial effects. For ease of visualization and interpretation, the centroids of the communities are converted to Thiessen polygons whose boundaries define the area that is closest to each centroid relative to all other centroids.

In addition, two binary categorical covariates are used; presence of slum settlers in a community $\varsigma_{(slum)}$ and population density $\rho_{(pop)}$. For communities within which slum settlers dwell, $\varsigma_{(slum)} = 1$, otherwise $\varsigma_{(slum)} = 0$. We could not quantify the population density for each community since the boundaries of the various communities are not mapped. Therefore, we categorized the population density as moderately populated $\rho_{(pop)} = 0$ and densely populated $\rho_{(pop)} = 1$. The effects of the categorical covariates are

assumed fixed and constant and are estimated jointly with the continuous and spatial covariates.

6.2.3 Model specification

Since $Chol_{(RR)}$ is based on ratio estimators, i.e. $Chol_{(RR)} \propto 1/Chol_{(E(C))}$, large changes occurred in the estimates with relatively small changes in $Chol_{(E(C))}$. Consequently, extreme variations were observed in the distribution of $Chol_{(RR)}$. To circumvent this problem, we classify the N communities into *high* or *low* risk areas such that the risk of cholera in a community $Chol_{(R)} = 1$ if $Chol_{(RR)} > 1$, i.e. if the observed cases $Chol_{(C)}$ is greater than the expected cases $Chol_{(E(C))}$, and $Chol_{(R)} = 0$ otherwise.

We consider the triple $\left(Chol_{(R)i}, x_i, w_i\right), i = 1,...,N$ where $Chol_{(R)i}$ is a binary response of the risk of cholera infection in community *i*. The vector $x_i = \left(x_{i1},...,x_{ip}\right)'$ contains the p continuous covariates and $w_i = \left(w_{i1},...,w_{ir}\right)'$ is a vector of r categorical covariates. The response is distributed as a binary (Bernoulli) random variable which is based on a latent continuous normally distributed variable

$$\overline{\overline{Chol}}_{(R)i} = \eta_i + \varepsilon_i, \tag{6.1}$$

with $\varepsilon_i \sim N(0,1)$. The response variable $Chol_{(R)}$ and the latent variable $\overline{\overline{Chol}}_{(R)}$ are linked by: $Chol_{(R)} = 1$ if $\overline{\overline{Chol}}_{(R)} > 0$ and $Chol_{(R)} = 0$ if $\overline{\overline{Chol}}_{(R)} \le 0$. Such a formulation is computationally advantageous for MCMC simulations when Bayesian approaches for model estimation are being used. The binary response is then modelled with a binomial probit link function of the form:

$$p\left(Chol_{(R)i} = 1 | \eta_i\right) = \Phi\left(\eta_i\right), \tag{6.2}$$

where Φ is the probit link function modelled with the linear predictor:

$$\eta_i = x_i'\beta + w_i'\gamma. \tag{6.3}$$

Here, β is a p-dimensional vector of unknown regression coefficients for the continuous covariates x_i, and γ is a r-dimensional vector of unknown regression coefficients for the categorical covariates w_i. We use the probit link function because sampling the latent variable $\overline{\overline{Chol}}_{(R)}$ is relatively easy and fast since the full conditionals

are truncated Gaussian distributions. Thus, $\overline{\overline{Chol}}_{(\mathrm{R})i} \,|\cdot \sim N(\eta_i, 1)$ is truncated at the left by zero if $Chol_{(\mathrm{R})i} = 1$ and truncated by zero at the right if $Chol_{(\mathrm{R})i} = 0$. The advantage of defining a probit model through a latent variable is that the full conditionals for the regression parameters β and γ is again Gaussian. Hence efficient and fast sampling schemes for Gaussian responses can be used with slight modifications.

The above model imposes a strictly linear effect on the response, but it does not account for spatially effects, and the hierarchical structure of the data. Accounting for these underlining effects requires an additive model that simultaneously accounts for the spatial dependence in the variables. Such models are termed *geo-additive* models as described by Kamman and Wand (2003). The *geo-additive* model replaces the strictly linear predictor by a more flexible semi-parametric predictor as

$$\eta_i = f_1(x_{i,1}) + ... + f_p(x_{i,p}) + f_{unobs}(s_i) + w_i'\gamma \,. \qquad 6.4$$

Here, $f_1(x),...,f_p(x)$ are nonlinear smooth functions of the continuous covariates $x_{i,1},...,x_{i,p}$ and $f_{unobs}(s)$ is a function that accounts for the effect of unobserved influential factors or spatial covariates $s_i \in \{1,...,S\}$ representing the centroids of the communities. In this study, the function $f_{unobs}(s)$ is split into a spatially structured (correlated) and a spatially unstructured (uncorrelated) effects,

$$f_{unobs}(s_i) = f_{str}(s_i) + f_{unstr}(s_i) \,. \qquad 6.5$$

The function $f_{str}(s_i)$ accounts for spatial correlation in the data, whereas $f_{unstr}(s_i)$ accounts for unobserved heterogeneity, occurring locally or at a large scale. The final *geo-additive* model is then expressed as:

$$\eta_i = f_1(x_{i,1}) + ... + f_p(x_{i,p}) + f_{str}(s_i) + f_{unstr}(s_i) + w_i'\gamma \,. \qquad 6.6$$

This equation contains $p+2$ functions and *r* fixed parameters to be estimated.

6.2.4 Prior distributions for covariates

A fully Bayesian approach for modelling and inferences requires prior assumptions for all unknown functions $f_j(x), f_{unstr}(s), f_{str}(s)$ and for the fixed effect regression parameter γ. For γ, we assume an independent diffuse prior $p(\gamma) \propto const$ due to the absence of any prior knowledge. A possible alternative choice is a weakly informative multivariate Gaussian distribution.

For the unknown functions $f_j(x), j = 1,...,p$, we choose the Bayesian P(enalized)-splines (Eilers and Marx, 1996; Lang and Brezger, 1978, 2004). This approach assumes that an unknown smooth function f_j of a covariate x_j can be approximated by a polynomial spline of degree l defined on a set of equally spaced knots $x_j^{\min} = \zeta_{j,0} < \zeta_{j,1} < \cdots < \zeta_{j,s-1} < \zeta_{j,s} = x_j^{\max}$ within the domain of x_j. Such a spline can be written in terms of a linear combination of $d = s + l$ B-spline basis functions B_m, i.e.

$$f_j(x_j) = \sum_{m=1}^{d} \xi_{j,m} B_m(x_j). \qquad 6.7$$

The B-splines form a local basis since the basic functions B_m are only positive within an area spanned by $l+2$ knots. This property is essential for the construction of the smoothness penalty for P-splines. The estimation of $f_j(x_j)$ is thus reduced to the estimation of the vector of unknown regression coefficients $\xi_j = (\xi_{j,1},...,\xi_{j,m})'$ from the data. An essential factor in the estimation procedure is the choice of the number of knots. We chose a moderately large number of equally spaced knots (20), as suggested by Eilers and Marx (1996) to ensure enough flexibility to capture the variability of the data. In the Bayesian approach, penalized splines are introduced by replacing the difference penalties with their stochastic analogues, i.e., first or second order random walk priors for the regression coefficients. A first order random walk prior for equidistant knots is given by:

$$\xi_{j,m} = \xi_{j,m-1} + u_{j,m}, \qquad m = 2,...,d, \qquad 6.8$$

and a second order random walk for equidistant knots by:

$$\xi_{j,m} = 2\xi_{j,m-1} - \xi_{j,m-2} + u_{j,m}, \qquad m = 3,...,d, \qquad 6.9$$

where the $u_{j,m} \sim N(0, \tau_j^2)$ are Gaussian errors. Diffuse priors $\xi_{j,1} \propto const$, or $\xi_{j,1}$ and $\xi_{j,2} \propto const$, are chosen as initial values, respectively. The joint distribution of the regression parameters $\xi_{j,m}$ for a first order random walk is defined as:

$$\xi_{j,m} \left| \xi_{j,m-1} \sim N(\xi_{j,m-1}, \tau_j^2), \right. \qquad 6.10$$

and a second order random walk is defined as:

$$\xi_{j,m} \left| \xi_{j,m-1}, \xi_{j,m-2} \sim N(2\xi_{j,m-1} - \xi_{j,m-2}, \tau_j^2). \right. \qquad 6.11$$

The first order random walk induces a constant trend for the conditional expectation of $\xi_{j,m}$ given $\xi_{j,m-1}$ and a second order random walk results in linear trend depending on the two previous values $\xi_{j,m-1}$ and $\xi_{j,m-2}$. The joint distribution of the regression parameters $\xi_j = \left(\xi_{j,1},...,\xi_{j,m}\right)'$ is computed as a product of the conditional densities defined by the random walk priors. The general form of the prior for ξ_j is a multivariate Gaussian distribution with density:

$$p\left(\xi_j \middle| \tau_j^2\right) \propto \exp\left(-\frac{\xi_j' K_j \xi_j}{2\tau_j^2}\right), \qquad 6.12$$

where the precision matrix K_j acts as a penalty matrix that shrinks parameters towards zero, or penalizes too abrupt jumps between neighbouring parameters. Since the penalty matrix K is rank deficient, i.e. $k_j = \text{rank}\left(K_j\right) < \dim\left(\xi_j\right) = d_j$, it follows that the prior for $\xi_j \middle| \tau_j^2$ is partially improper with Gaussian prior $\xi_j \middle| \tau_j^2 \propto N\left(0; \tau_j^2 K_j^-\right)$, where K^- is a generalized inverse of K. The tradeoff between flexibility and smoothness is controlled by the variance parameter τ_j^2. The larger the variance, the rougher is the estimated functions, and vice versa.

Spatial components

For the structured spatial components $f_{str}\left(s\right), s = 1,...,S$, we choose the Markov random field prior based on the conditional distribution of the spatial neighbourhood relationship (Besag, 1974, 1975; Besag et al., 1991). This leads to a conditional spatially autoregressive specification

$$f_{str}\left(s\right) \middle| f_{str}\left(s'\right), s' \neq s, \tau_{str}^2 \sim N\left(\frac{1}{N_s}\sum_{s' \sim s} f_{str}\left(s'\right), \frac{\tau_{str}^2}{N_s}\right), \qquad 6.13$$

where N_s is the number of adjacent spatial units and $s' \sim s$ denotes that spatial unit s' is a neighbour of spatial unit s. Thus, the conditional mean of $f_{str}\left(s\right)$ is an unweighted average of the function evaluations of neighbouring spatial units. The $N \times S$ design matrix ψ is a 0/1 incidence matrix. The number of columns is equal to the number of spatial units since only one parameter is estimated for each spatial unit. Its value in the i*th* row and the s*th* column is 1 if the i*th* observation is located in the s*th* spatial unit, and 0 otherwise. In this study $N = S$, hence the matrix ψ is a square matrix. Since point data were used, we chose a neighbourhood structure based on the *k*th nearest neighbour method, where *k* is the number of neighbors. This approach results in an asymmetric neighbourhood matrix; therefore, false symmetry was imposed to ensure a

symmetrical neighbourhood structure. Like the continuous functions f_j, the tradeoff between flexibility and smoothness is controlled by the variance parameter τ^2_{str}.

For the unstructured spatial effects, we assume that the parameters $f_{unstr}(s)$ are *i.i.d.* Gaussian:

$$f_{unstr}(s)|\tau^2_{unstr} \sim N(0;\tau^2_{unstr}). \quad 6.14$$

Hyper-priors for the variance or smoothness parameters τ^2_j, $j = 1,...,p, str, unstr$, are considered as unknown. Therefore, highly dispersed, but proper, inverse Gamma distributions $p(\tau^2_j) \sim IG(a_j, b_j)$ with known hyper-parameters a_j and b_j are assigned in the second stage of the hierarchy. The corresponding probability density function is expressed as:

$$p(\tau^2_j) \propto (\tau^2_j)^{-a_j-1} \exp\left(-\frac{b_j}{\tau^2_j}\right). \quad 6.15$$

A standard choice for the hyper-parameters is to select small values for *a* and *b*, e.g. $a = b = 0.001$. Since the regression parameters depend on the choice of hyper-parameters used, other choices of hyper-parameters have been used as well to investigate the sensitivity of the models. Four alternatives of priors *IG* ($a = 0.5$, $b = 0.0005$), *IG* ($a = 1$, $b = 0.005$), *IG* ($a = 0.001$, $b = 0.001$) and *IG* ($a = 0.01$, $b = 0.01$) are used. The first and second choices are suggested by Kelsall and Wakefield (1999) and Besag and Kooperberg (1995) respectively. The third and fourth alternatives with equal scale and shape parameters, especially *IG* ($a = b = 0.001$), have often been used as a choice for the variances of random effects.

6.2.5 Bayesian inference

Bayesian inference stems from the posterior distribution, that is, the conditional distribution of the model parameters given the observed data $p(\theta|Chol)$, where θ denotes the vector of all model parameters, *Chol* the data vector, $p(.)$ represents the probability density function, and $L(.)$ is the likelihood function of the binomial probit model. In this study, we used a fully Bayesian inference based on analysis of posterior distribution of the model parameters by drawing random samples via MCMC simulation techniques. The probability density function of the posterior distribution is expressed as:

$$
\begin{aligned}
p(\theta|Chol) \propto & \prod_{i=1}^{n} L\left(Chol_{(\mathrm{R})i}, \eta_i\right) \\
& \times \prod_{j=1}^{p}\left[p\left(\xi_j \left|\tau_j^2\right.\right) p\left(\tau_j^2\right)\right] \\
& \times p\left(f_{str} \left|\tau_{str}^2\right.\right) p\left(f_{unstr} \left|\tau_{unstr}^2\right.\right) \\
& \times \prod_{j=1}^{r} p\left(\gamma_j\right) p\left(\sigma^2\right)
\end{aligned}
\quad . \qquad 6.16
$$

The full conditionals for the variance components $\tau_j^2, j = 1,...,p$, *str*, *unstr*, and σ^2 are inverse Gamma distributions. The full conditionals for the fixed parameters γ, the unknown parameter vector $\xi_1,...,\xi_p$, as well as $f_{str}(s)$, $f_{unstr}(s)$ are multivariate Gaussian. Gibbs sampler was employed for MCMC simulations, drawing successively from the full conditionals for $\xi_1,...,\xi_p$, $f_{str}(s)$, $f_{unstr}(s)$, $\tau_j^2, j = 1,...,p$, *str*, *unstr*, and σ^2. Cholesky decompositions for band matrices were used to efficiently draw random samples from the full conditional (Rue, 2001, 2005).

6.2.6 Model implementation

Three sets of explanatory models were developed to explain the risk of cholera infection. Model 1 is a strictly linear predictor that assumes a linear effect of the categorical and continuous covariates. Model 2 is an additive model which assumes nonlinear functions for the continuous covariates and linear effects of the categorical covariates. Model 3 is a geo-additive model, which is an extension of Model 2 that incorporates both structured and unstructured spatial effects.

Models:

$$Model\,1:\ \eta_i = \rho'_{(\mathrm{dump})}\beta_1 + d'_{(\mathrm{dump})}\beta_2 + d'_{(\mathrm{reser})}\beta_3 + \rho'_{(\mathrm{pop})}\gamma_1 + \varsigma'_{(\mathrm{slum})}\gamma_2$$

$$Model\,2:\ \eta_i = f_1\left(\rho_{(\mathrm{dump})}\right) + f_2\left(d_{(\mathrm{dump})}\right) + f_3\left(d_{(\mathrm{reser})}\right) + \rho'_{(\mathrm{pop})}\gamma_1 + \varsigma'_{(\mathrm{slum})}\gamma_2$$

$$Model\,3:\ \eta_i = f_1\left(\rho_{(\mathrm{dump})}\right) + f_2\left(d_{(\mathrm{dump})}\right) + f_3\left(d_{(\mathrm{reser})}\right) + f_{str}(s) + f_{unstr}(s) + \rho'_{(\mathrm{pop})}\gamma_1 + \varsigma'_{(\mathrm{slum})}\gamma_2$$

The models were implemented in a public domain program, BayesX ver 2.0 (Brezger et al., 2005; Belitz et al., 2009). We used a total number of 40,000 MCMC iterations and 10,000 number of burn in samples. Since, in general, these random numbers are correlated, only every 20th sampled parameter of the Markov chain were stored. This yielded 2,000 samples for parameter estimation.

We compared the strictly linear models with the additive models and the geo-additive models using the Deviance Information Criterion (*DIC*) values (Spiegelhalter, 2002). *DIC* is a Bayesian tool for model checking and comparison, where the model with the smallest *DIC* is preferred. The *DIC* is given by $DIC = \bar{D} + p_D$, where $\bar{D}$ is the posterior mean of the deviance, which is a measure of goodness of fit, and p_D is the effective number of parameters, which is a measure of model complexity and penalizes over-fitting.

6.3 Results and analysis

6.3.1 Sensitivity analyses and model selection

Results of the sensitivity analysis on the choice of hyper-parameters a and b are shown in Table 6.1. Since similar patterns of variations in *DIC* values were observed for all the three models, we report only on the sensitivity analysis for Model 3. From Table 6.1, it is noticed that the last two choices of hyper-parameters *IG* (a = 0.01, b = 0.01) and *IG* (a = 0.001, b = 0.001) yield better (but similar) inferences than the last two hyper-parameters *IG* (a = 0.5, b = 0.0005), *IG* (a = 1, b = 0.005). Since the fourth choice of hyper-parameters *IG* (a = 0.01, b = 0.01) overshadows the significance of the categorical covariate $\varsigma_{(\text{slum})}$, the choice of hyper-parameters *IG* (a = 0.001, b = 0.001) is appropriate for the analyses.

Model assessment and selection was based on the computed values for the goodness of fit (see Table 6.4). Models with a smaller *DIC* value are preferred. Again, models with differences in *DIC* of less than 3 cannot be distinguished, while those between 3 and 7 can be weakly differentiated (Besag and Kooperberg, 1995). Comparing goodness of fit of models, Model 3 is the preferred model. Although the extension of the basic model (Model1) to an additive model (Model 2) is an improvement; this improvement is indistinguishable (*DIC* = 43.25 in Model 1 versus *DIC* = 41.30 in Model 2, $\Delta DIC = 1.95$). The extension of Model 2 to include structured and unstructured spatial effects in Model3 significantly improved the model (*DIC* = 20.07 in Model 3 versus *DIC* = 41.30 in Model 2, $\Delta DIC = 21.23$). Therefore, our results are based on Model 3.

Table 6.1: Summary of the sensitivity analysis of the choice of hyper-parameters for Model 3

	a = 0.5 b = 0.0005	a = 1 b = 0.005	a = 0.001 b = 0.001	a = 0.01 b = 0.01
Model fit				
$\bar{D}$	24.77	29.3	10.64	9.49
pD	15.64	13.51	9.43	8.784
DIC	40.41	42.81	20.07	18.273
Fixed effects[†]				
Constant	-3.92 (-8.1, -1.41)	-2.83 (-5.44, -1.19)	-6.05 (-8.81, -3.14)	-6.46 (-9.76, -3.41)

$\varsigma_{(\text{slum})}, \gamma_2$	2.40 (0.19, 6.16)	1.6 (0.004, 3.29)	4.06 (0.53, 7.85)	3.98 (-0.08, 8.15)
$\rho_{(\text{pop})}, \gamma_1$	2.42 (0.39, 5.42)	2.05 (0.27, 3.86)	4.34 (1.12, 7.96)	4.71 (0.69, 8.79)
Spatial effects‡				
$f_{str}(s)$, τ^2_{str}	0.103 (0.0004, 0.051)	0.106 (0.002, 0.06)	39.46 (13.45, 69.92)	17.09 (0.05, 51.46)
$f_{unstr}(s)$, τ^2_{unstr}	4.24 (0.0005, 14.72)	1.90 (0.002, 7.53)	11.77 (0.005, 41.64)	7.92 (0.06, 18.54)
Smooth functions§				
$f_1\left(\rho_{(\text{dump})}\right), \tau_1^2$	0.02 (0.0004, 0.027)	0.02 (0.002, 0.04)	0.35 (0.002, 0.68)	0.55 (0.01, 1.22)
$f_2\left(d_{(\text{dump})}\right), \tau_2^2$	0.08 (0.0006, 0.155)	0.04 (0.003, 0.08)	0.71 (0.008,1.68)	0.72 (0.03, 1.74)
$f_3\left(d_{(\text{reser})}\right), \tau_3^2$	0.03 (0.0003, 0.026)	0.02 (0.002, 0.04)	0.28 (0.002, 0.61)	0.37 (0.01, 0.89)

†Estimates of posterior mean and 90% credible intervals for the fixed effects; ‡variance components and 90% credible intervals for the spatially structured and unstructured effects; §variance components and 90% credible intervals for the nonlinear smooth functions.

Table 6.2: Estimates of fixed effect parameters based on Model 1

Variable	Mean	Std. error	10%	90%
constant	0.139±	1.520	-1.753	2.088
$\rho_{(pop)}, \gamma_1$	1.309*	0.836	0.232	2.361
$\varsigma_{(slum)}, \gamma_2$	1.289*	0.803	0.295	2.270
$\rho_{(dump)}, \beta_1$	0.135±	0.302	-0.256	0.538
$d_{(dump)}, \beta_2$	-0.0009*	0.0005	-0.0016	-0.0002
$d_{(reser)}, \beta_3$	-0.0007±	0.0006	-0.0015	3.21E-05

* Significance at $p < 0.01$, ± not significant

Table 6.3: Estimates of fixed effect parameters based on Model 3

Variable	Mean	Std. error	10%	90%
constant	-6.050*	2.189	-8.819	-3.141
$\rho_{(pop)}, \gamma_1$	4.339*	2.747	1.123	7.964
$\varsigma_{(slum)}, \gamma_2$	4.060*	2.899	0.525	7.852

* Significance at $p < 0.01$

Table 6.4: Comparison of model fit using Deviance Information Criterion (*DIC*)

Model Fit	Model 1	Model 2	Model 3
$\bar{D}$	37.40	32.35	10.64
pD	5.85	8.95	9.43
DIC	43.25	41.30	20.07
ΔDIC	23.18	21.23	Reference

6.3.2 Fixed and nonlinear effects of covariates

The estimates of the fixed effects parameters in the linear model (Model 1) are shown in Table 6.2. Since Model 1 ignores important nonlinearities, the effects of $\rho_{(dump)}$ and $d_{(reser)}$ on $Chol_{(R)}$ appear to be not significant. The posterior means and the corresponding 90% credible intervals of the fixed effect parameters of Model 3 are shown in Table 6.3. The effect of the fixed parameters $\rho_{(pop)}$ equals 4.339 and $\varsigma_{(slum)}$ equals 4.06. These values are more than twice than those from Model 1, i.e. $\rho_{(pop)} = 1.309$ and $\varsigma_{(slum)} = 1.289$. This suggests that the risk of cholera infection is higher in slums and densely populated communities. The nonlinear effects of $\rho_{(dump)}$, $d_{(dump)}$, and $d_{(reser)}$ are shown in Figures 6.2, 6.3, and 6.4, respectively. The effect of $\rho_{(dump)}$ is almost linear, with increasing posterior means. For $d_{(dump)}$, the

posterior mean shows a major deviation from linearity, with decreasing risk up to about 500 m, and beyond that distance a slight increase. The effect of $d_{(\text{reser})}$ is also almost linear, with the posterior mean decreasing with increasing distance.

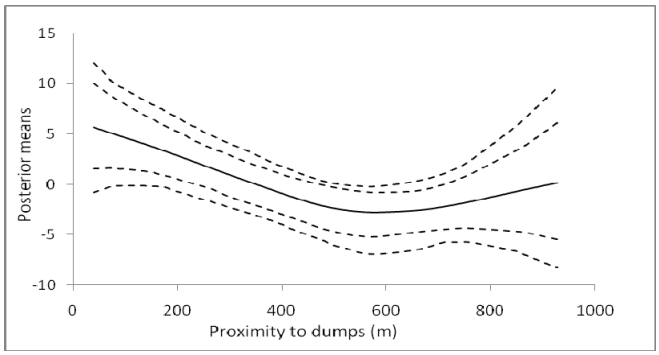

Figure 6.2: The estimated nonlinear effects of cholera risk on of proximity to refuse dumps in Kumasi. The posterior mean together with the 80% and 90% credible intervakls are shown

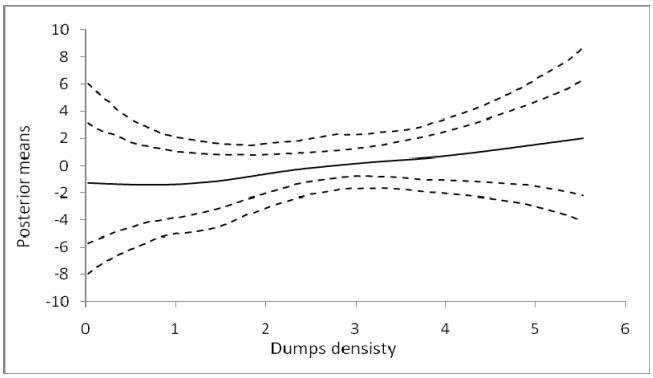

Figure 6.3: The estimated nonlinear effects of cholera risk on dumps density in Kumasi. The posterior mean together with the 80% and 90% credible intervakls are shown

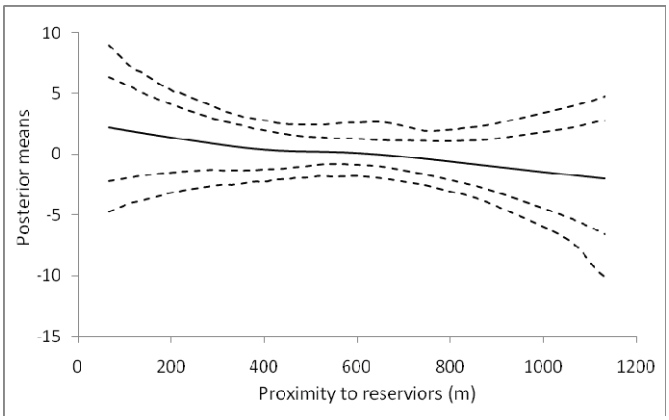

Figure 6.4: The estimated nonlinear effects of cholera risk on proximity to potential cholera reserviors in Kumasi. The posterior mean together with the 80% and 90% credible intervakls are shown

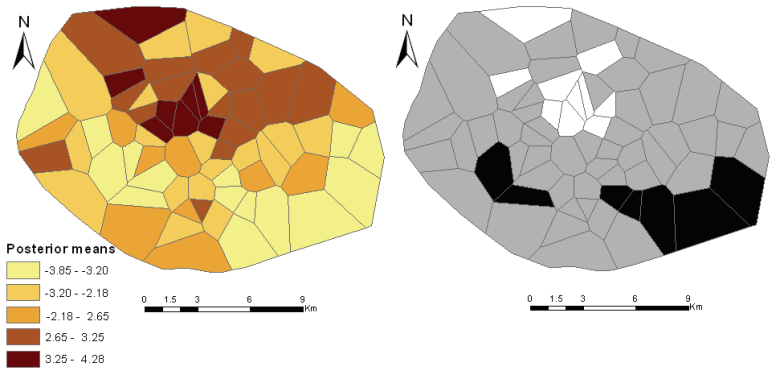

Figure 6.5: Spatial distribution of the posterior means of the total spatial effects on cholera risk (left), and posterior probabilities at nominal level of 80% (right). Black denotes areas with strictly negative credible intervals; white denotes areas with strictly positive credible intervals, while grey shows areas of no significant difference.

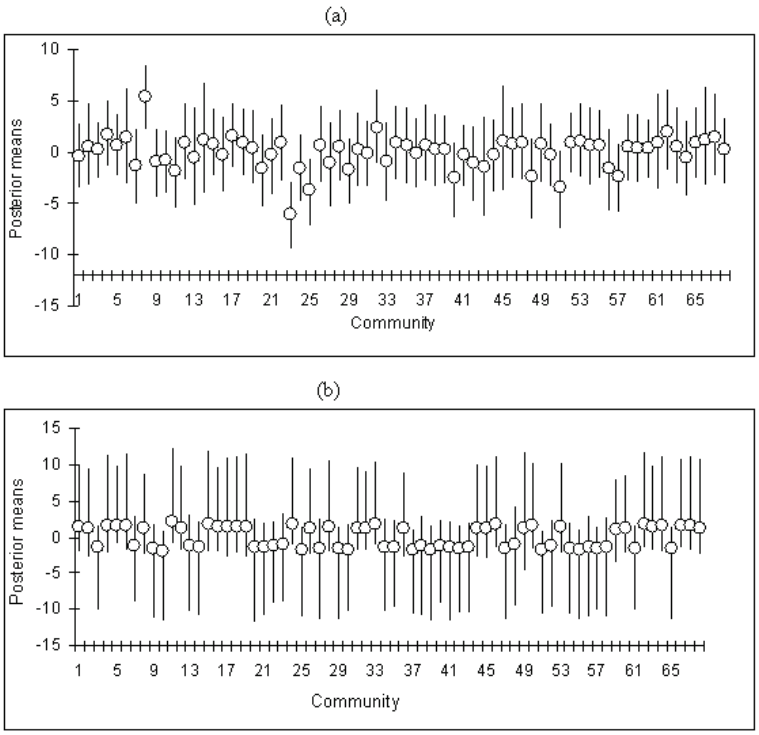

Figure 6.6: Caterpillar plots of the posterior means of the structured (a) and unstructured (b) spatial effects of the risk of cholera infection, with 90% error bars

6.3.3 Spatial effects

There is a considerable spatial variation in the risk of cholera infection. Figure 6.5 shows the estimated total spatial effects (left) and the corresponding 80% (credible interval) posterior probability map (right) of cholera risk. Areas shaded black show strictly negative credible intervals, while white areas depict strictly positive credible intervals, and grey indicate areas of non-significant spatial effects. The structured and unstructured spatial effects are given by the caterpillar plots in Figure 6.6. The spatially structured effects are dominant over the unstructured effects. This is shown by the wider variations in the caterpillar plots of Figure 6.6a compared with Figure 6.6b. The higher ratio of variance components $\phi = \tau_{str}^2 / \left(\tau_{str}^2 + \tau_{unstr}^2 \right) = 0.77$ (Table 6.1) confirms that the structured spatial effects are dominant over the unstructured spatial effects.

6.4 Discussion

This study utilizes *geo-additive* modelling approach to develop a multivariate explanatory model for the risk of cholera infection. We utilize the binomial probit regression model to elucidate the probability of cholera infection in relation to associated risk factors, some identified from previous studies (Osei and Duker, 2008; Osei et al., 2010). The *geo-additive* modelling approach is an extension of the GAM which allows the inclusion of both structured and unstructured spatial effects to account for possible unobserved factors and heterogeneity terms. To allow flexibility, the continuous covariates are modelled non-parametrically as nonlinear functions using P-splines with second-order random walk priors based, this based on contributions by Farhmeir and Lang (2001a, 2001b) and Fahrmeir et al. (2004); while the categorical covariates were modelled as fixed effects. The spatially structured and unstructured effects are modelled using Markov random filed priors and zero mean Gaussian heterogeneity priors respectively (Besag et al., 1991). In this modelling approach, fully Bayesian inferences based on MCMC simulations are preferred because the functionality of the posterior can be easily computed, thereby easily quantifying the uncertainty in the estimated parameters (Fahrmier et al., 2004).

The findings of the study show that the risk of cholera infection is higher within communities with slum settlers. The risk is also higher in densely populated communities (See Table 6.2). These relationships may exist because most communities with slummy settlers are densely populated. Although cholera is transmitted mainly through contaminated water or food, poor sanitary conditions in the environment enhance its transmission. The cholera *vibrios* can survive and multiply outside the human body and can spread rapidly where living conditions are overcrowded and where there is no safe disposal of solid waste, liquid waste, and human faeces (WHO, 2000). These conditions are mostly met in slummy and densely populated communities in Kumasi. Such high population density may necessarily result in shorter disease transmission paths, thus increasing the risk of cholera infection. Inhabitants living at slummy areas are generally poor, and face problems including access to potable water and sanitation. In many cases public utilities providers (e.g. water distribution companies) legally fail to serve these urban poor due to factors regarding land tenure

system, technical and service regulations, and city development plans. Most slum settlements are also located at low lying areas susceptible to flooding. Unfavourable topography, soil, and hydro-geological conditions make it difficult to achieve and maintain high sanitation standards among inhabitants living in these territories (Borroto and Martinez-Piedra, 2000).

The results of the nonlinear effects of $d_{(\text{dump})}$ and $\rho_{(\text{dump})}$ on $Chol_{(\text{R})}$ suggest that cholera risk is relatively high when inhabitants live in close proximity to waste dumps and where there are numerous refuse dumps. Due to the bad defecation practices of most inhabitants, the refuse dumps may contain high faecal matter. Surface drainage from such refuse dumps focally pollute water sources which when used perpetuates the transmission of cholera *vibrios*. If the runoff from waste dumps during heavy rains serve as the major pathway for faecal and bacterial contamination of rivers and streams, then it is likely that inhabitants living closer to water bodies where these runoffs flow into will have higher cholera prevalence than those who live farther. The observed decreasing cholera prevalence with increasing distance from potentially polluted surface water bodies (Figure 6.4), and the significant linear dependency of $d_{(\text{dump})}$ on $d_{(\text{reser})}$ (β = 0.67, R^2 = 0.34, $p < 0.001$) supports this hypothesis.

The relationship between $Chol_{(\text{R})}$ and $d_{(\text{dump})}$ is estimated to be strictly nonlinear, with an expected decreasing relationship to about 500 m, and a slight increasing relation afterwards. Since a decreasing relation is expected, we conclude that proximity to dumps may influence the risk of cholera infection only within a distance of about 500 m. This is consistent with the finding from previous studies when a quantitative assessment of critical distance discrimination on experimental buffer zones around refuse dumps showed that the optimum spatial discrimination of cholera occurs at 500 m way from refuse dumps (Osei and Duker, 2008b). Therefore, we hypothesis that refuse dumps located within 500 m away from inhabitants enhance the risk of cholera infection compared with those farther. The slight unexpected increase of $Chol_{(\text{R})}$ with $d_{(\text{dump})}$ after 500 m, however, is seemingly grounds for questioning the acceptance of this hypothesis.

The spatial effects included in the model are surrogate measures of unobserved risk factors of cholera. There is evidence of significant increased cholera risk at the central part, and a significant reduced risk at the periphery of the south-eastern part of Kumasi (Figure 6.2). These patterns may be explained by the fact that communities at the central part are highly populated with lots of slum settlers, while communities at the peripheries are moderately populated. These patterns clearly indicate possible unobserved risk factors of cholera, some of which may be individual or household level. Therefore, this gives leads for further epidemiological research using additional information within the study area.

In this study, highly dispersed (but proper) inverse Gamma priors $p\left(\tau_j^2\right) \sim IG\left(a_j, b_j\right)$ with $a_j > 0$ and $b_j > 0$ were assigned to all variance components. This ensures

propriety of the joint posterior despite the partial impropriety of the priors for the unknown parameters ξ_j. We find evidence that the results are variable to the choice of hyper-parameters (Table 6.1). Specifically, hyper-parameters with the unequal scale and shape parameters *IG* ($a \neq b$) and/or relatively large scale parameters *IG* ($a > b$) seems to produce similar results, whiles the hyper-parameters with equal scale and shape parameters *IG* ($a = b$) and/or relatively small values for a and b seems to produce similar results. The results, however, are more significant using the so called standard choices of hyper-parameters, thus relatively small values for a and b, e.g. $a = b = 0.001$ or 0.01. Therefore, our choice of hyper-parameters *IG* ($a = 0.001$, $b = 0.001$) is appropriate for the analyses.

In Model 3, CAR prior has been used to account for spatial random effects. The objective has partly been to estimate the posterior mean of the fixed effects $\rho_{(\text{pop})}$ and $\varsigma_{(\text{slum})}$ while accounting for spatial correlation. Although CAR prior is widely used for modelling spatial effects in complex hierarchical Bayesian models, its usage is at least debatable. Co linearity between the fixed effects and the CAR random effects can cause large changes in the posterior mean and variance of the fixed effects compared to a non-spatial regression model (Reich et al., 2006). In this study, changes in the posterior mean of the fixed effects $\rho_{(\text{pop})}$ and $\varsigma_{(\text{slum})}$ are observed after including CAR random effects. The posterior mean of the fixed effects changes from $\rho_{(\text{pop})} = 1.309$ and $\varsigma_{(\text{slum})} = 1.289$ in Model 1 (see Table 6.2) to $\rho_{(\text{pop})} = 4.339$ and $\varsigma_{(\text{slum})} = 4.060$ in Model 3 (see Table 6.3). These changes occurred despite the increase in model complexity or effective number of parameters from $pD = 5.85$ in Model 1 to $pD = 9.43$ in Model 3 (Table 6.4). The inflation of the posterior of $\rho_{(\text{pop})}$ and $\varsigma_{(\text{slum})}$ in Model 3 may be due to possible co linearity between the fixed effects and the CAR random effects rather than model suitability. However, causes of these changes have not been investigated in this study. Several diagnostics to investigate the change in the posterior of the fixed effects by adding spatial random effects have, however, been proposed in literature (Christensen et al., 1992; Christensen et al., 1993; Haslett, 1999). Reich et al. (2006) have investigated the effect of adding CAR random effects on the posterior of the fixed effects for disease mapping, and have proposed appropriate diagnostic methods to alleviate the co linearity between the fixed effect covariates and the CAR random effects.

6.5 Conclusion

This study applies a Bayesian semi-parametric modelling approach to develop explanatory models of cholera risk. Such flexible modelling approaches allow joint analysis of the nonlinear effects of continuous covariates, spatially structured variation, unstructured heterogeneity, and fixed covariates. Our model reveals that the risk of cholera infection is associated with slum settlements, high population density, proximity to and density of waste dumps, proximity to potentially polluted rivers and streams, as well as possible unobserved risk factors. The possible unobserved risk factors are shown

by the distinct spatial patterns exhibited by the spatial covariates; suggesting the need for further epidemiological research. These findings should serve as novel information to help health planners and policy makers in making effective decisions about cholera control measures.

7

Bayesian modelling of the space-time diffusion pattern of cholera epidemic

"Time, space, and causality are only metaphors of knowledge, with which we explain things to ourselves"

Friedrich Nietzsche

This chapter is the last of the series of chapters aimed at understanding the spatial distribution of cholera and its associated risk factors through spatial statistical methods. The preceding chapters focused on the effect of the risk factors on cholera prevalence. In this chapter, the focus is the effects of the risk factors on the diffusion dynamics of cholera. Here, we develop and present statistical models to investigate the transmission routes of cholera diffusion. Classical linear regression approaches have often been used to model the diffusion dynamics of infectious diseases. In this chapter, we apply a Bayesian structured additive regression modelling approach to model the diffusion dynamics of cholera epidemic in Kumasi. This chapter is accepted in the journal of *Statistica Neerlandica*. It has been submitted as: Osei FB, Duker AA and Stein A: Hierarchical Bayesian modelling of the space-time diffusion patterns of cholera epidemic in Kumasi, Ghana.

Abstract

Strategies for prevention and control of cholera *V. cholerae* depend on understanding the transmission dynamics and other geographic and demographic characteristics associated with the epidemic spread. Cholera epidemic is described by two transmission routes. The primary *environment-to-human* transmission route is by means of exposure to an aquatic reservoir of cholera. The secondary *human-to-human* route is by means of the faecal-oral contacts. This study analyzes the joint effects of these two transmission routes on the space-time diffusion dynamics of cholera epidemics. Statistical models are developed and presented to investigate the transmission routes of cholera diffusion, as well as possible primary cases. A Bayesian modelling approach is employed for a joint analysis of nonlinear effects of continuous covariates, spatially structured variation, and unstructured heterogeneity. Proximity to primary case locations and population density serve as continuous covariates. Reference to communities is modelled as a spatial effect. The study applied to the Kumasi area in Ghana shows that communities proximal to primary case locations are infected relatively early during the epidemics, with more remote communities infected at later dates. Similarly, more populous communities are infected relatively early and less populous communities at later dates. The rate of infections increases almost linearly with population density. A non systematic relation occurs between the rate of infection and proximity to primary case locations. It is discussed how these findings could serve as significant information to help health planners and policy makers in making effective decisions about cholera prevention and control measures.

7.1 Introduction

Mapping of disease transmission routes in human population and knowledge of its spatial and temporal transmission dynamics are essential for epidemiologist to better understand the population's interaction with its environment. Understanding the spatial distribution of diseases and transmission dynamics is facilitated by advancements in Geographic Information Systems (GIS) and spatial statistics. These provide opportunities for epidemiologist to analyze disease distribution in space and interactions with the environment. Most of these approaches, however, ignore methodological difficulties that arise from the nature of the data, especially when the population distribution and environment is particularly variable and spatially structured.

Classical linear regression approaches, where the response variable is assumed to be Gaussian distributed with the covariates acting linearly on the response, have been used to model the diffusion dynamics of infectious diseases (Kuo and Fukui, 2007; Trevelyan et al., 2005). Such diffusion models assume a strictly linear relationship between the dependent variable and the predictor variables, thereby ignoring the possible nonlinear and spatial effects of the predictor variables. Moreover, these diffusion models ignore the possibility and role of multiple index cases in the diffusion dynamics of the disease.

Cholera is a water-borne disease caused by *Vibrio cholera* (hereafter *V. cholera*). Comprehensive discussions about cholera are presented by Carpenter (1970), Colwell and Huq (1994), Finkelstein (1996, 1999), Prestero et al. (2001), Sack et al. (2004), Huq et al. (2005). The disease has been scrutinized since the beginning of epidemiology, yet it remains an important public health problem, especially in developing countries. Without treatment, case-fatality rate or death can be as high as 50% of severe cases (WHO, 1993; Sack et al., 2004). Cholera diffuses rapidly in environments that lack basic infrastructure with regard to access to safe water and proper sanitation. Provision of good sanitary conditions, sewage treatment, and provision of clean water have long been known as important critical measures for prevention and eradication. These measures have eliminated cholera from industrialized and developed countries. Chronic poverty in developing countries makes implementation of these measures almost unfeasible. A better understanding of the dynamics of cholera spread amongst communities could help to develop alternative and timely public health interventions to limit or prevent cholera epidemics.

Two routes of cholera transmission have been described. The primary route or *environment-to-human* transmission is the exposure of a human being to an aquatic reservoir of *V. cholera*. The secondary route or *human-to-human* transmission is through faecal-oral contacts induced by a previously infected person (Miller et al., 1985; Glass et al., 1991). Primary transmission is responsible for sparking initial outbreaks. Primary cases are therefore hypothesized to be scattered is space and time, occurring almost simultaneously in distant areas with no apparent connection. In contrast, once the outbreak has reached a threshold level, faecal-oral transmissions dominate and the disease becomes highly contagious. Consequently, geographic factors such as proximity to a primary case location and population density should spatially dominate the disease propagation. To examine these hypotheses, this study analyzes the joint effects of

primary and secondary transmission in the space-time diffusion dynamics of cholera. Specifically, the study seeks to (1) define and map the transmission routes of cholera diffusion from possible multiple primary cases and (2) model the joint effects of population density and proximity to primary cases on the space-time dynamics of cholera diffusion.

This paper is organized as follows. First, a variogram model is used to characterize the spatial auto-covariance structure of incidence rates in order to determine the threshold/extent of contagiousness of cholera. Thus, the variogram model is used to characterize the dominant scale at which cholera transmission occurs. Secondly, the threshold value and the times of cholera entrance in communities are applied to define the transmission routes and all probable primary cases. Third, a Bayesian model is built, where the time ordered sequence of cholera entrance in each community is modelled as nonlinear functions of proximity to respective primary cases and the urban level. In such a modelling approach, the unknown parameters are treated as random variables arranged in a hierarchy such that the distributions at each level are determined by the random variables in the previous levels. Next, we present the results and conclude the paper with discussion on the results.

7.2 Methods and Data

7.2.1 Study area and Data

The area studied is the Kumasi Metropolis, an urban and the most populous city in Ashanti Region, at approximately 250 km (by road) northwest of Accra. It is centred at the intersection of latitude 6.04^{o}N and longitude 1.28^{o}W, covering an area of about 220 km^2 (See Figure 7.1). Kumasi has a population of approximately 1.2 million which accounts for just under a third (i.e. 32.4%) of the region's population. After cholera introduction in Ghana in the 1970's, the country has experienced a series of epidemic outbreaks. Surveillance and reporting of the disease before 2005 has been ineffective, and hence the existing data before 2005 have little or no spatial and temporal information. With intensified surveillance and reporting systems during an outbreak in 2005, disease cases in Kumasi are being recorded daily at community level spatial units. Kumasi is therefore suitable for studying the dynamics of cholera in space and time.

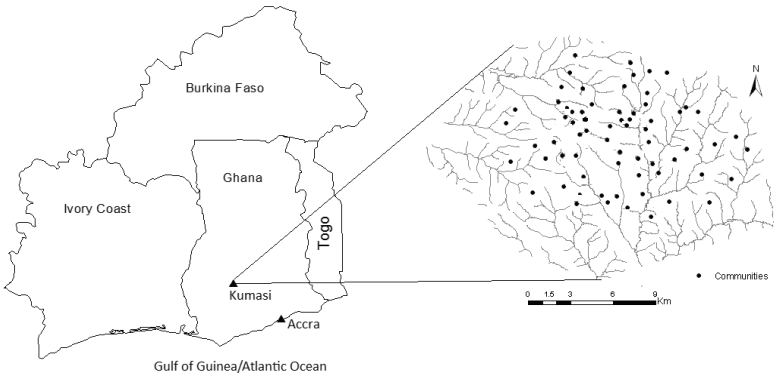

Figure 7.1: District map of Ghana (left), and Kumasi (right). Dots indicate the centroids of communities.

The topographic map of the metropolis and the $n = 68$ communities where cholera records are available was digitized. Cholera data for each community was extracted from case records obtained from the Kumasi Metropolitan Disease Control Unit (DCU). This data contains information about the index case records for each community, thus the time of cholera onset for each community. Residential addresses were not recorded during the period of diagnoses; therefore, the centroids of the communities were used as the spatial references of the case locations. Also, index case records for each community were extracted and assigned with unique identification codes. These were cross-linked to the communities by unique identification codes to facilitate easy geo-referencing and further analysis.

7.2.2 Defining the extent of contagiousness: variogram modelling

Let $S = (s_1, ..., s_n)$ represent the spatial locations of the n communities belonging to a domain D, denote the number of cholera cases as $Chol_{(C)}(s_\alpha)$, and the size of the population at risk as $n(s_\alpha)$. Then the observed incidence rate at s_α is expressed as $Chol_{(R)}(s_\alpha) = \frac{Chol_{(C)}(s_\alpha)}{n(s_\alpha)}$. In order to make a statistical inference, $Chol_{(C)}(s_\alpha)$ is interpreted as the realization of a random variable that follows a one parameter Poisson distribution with intensity $n(s_\alpha) \cdot Chol_{(E(R))}(s_\alpha)$, where $Chol_{(E(R))}(s_\alpha)$ is proportional to cholera incidences and measures the expectation of cholera cases per unit population.. Thus:

$$Chol_{(C)}(s_\alpha) \big| Chol_{(E(R))}(s_\alpha) \sim \text{Poisson}\left(n(s_\alpha) \cdot Chol_{(E(R))}(s_\alpha)\right), \alpha = 1, ..., n. \quad 7.1$$

Such modelling is based on the assumption that the spatial correlation among cholera cases is caused by spatial trends in either the population sizes or in the local individual

risks given the expected risk $Chol_{(\mathrm{E(R)})}(s_\alpha)$. Therefore, the count variables $Chol_{(\mathrm{C})}(s_\alpha)$ are assumed to be conditionally independent. The expected risk $Chol_{(\mathrm{E(R)})}(s_\alpha)$ is modelled as a positive random field honouring order two stationarity, with mean $\overline{Chol}_{(\mathrm{E(R)})}$, variance $\sigma^2_{Chol_{(\mathrm{E(R)})}}$, and covariance function:

$$C_{Chol_{(\mathrm{E(R)})}}(\mathrm{h}) = \mathrm{Cov}\left[Chol_{(\mathrm{E(R)})}(\mathrm{s}_\alpha), Chol_{(\mathrm{E(R)})}(\mathrm{s}_\beta)\right], \tag{7.2}$$

which depends only on the distance h between observation pairs s_α and s_β. From equation (7.1), it follows that:

$$\left.\begin{aligned}
&E\left[Chol_{(\mathrm{C})}(s_\alpha)\middle|Chol_{(\mathrm{E(R)})}(s_\alpha)\right] = n(s_\alpha)\cdot Chol_{(\mathrm{E(R)})}(s_\alpha)\\
&E\left[Chol_{(\mathrm{C})}(s_\alpha)\right] = \overline{Chol}_{(\mathrm{E(R)})}\cdot n(s_\alpha)\\
&\mathrm{Var}\left[Chol_{(\mathrm{C})}(s_\alpha)\middle|Chol_{(\mathrm{E(R)})}(s_\alpha)\right] = n(s_\alpha)\cdot Chol_{(\mathrm{E(R)})}(s_\alpha)\\
&\mathrm{Var}\left[Chol_{(\mathrm{C})}(s_\alpha)\right] = \left(n(s_\alpha)\right)^2\cdot\sigma^2_{Chol_{(\mathrm{E(R)})}} + \overline{Chol}_{(\mathrm{E(R)})}\cdot n(s_\alpha)
\end{aligned}\right\}. \tag{7.3}$$

Following Matheron's (1963, 1965) intrinsic hypothesis on expected mean differences and variances, the equivalent experimental variogram is:

$$\gamma_{Chol_{(\mathrm{E(R)})}}(\mathrm{h}) = \frac{1}{2N(\mathrm{h})}\sum_{i=1}^{N(\mathrm{h})}\left[\left(Chol_{(\mathrm{R})}(\mathrm{s}_\alpha) - Chol_{(\mathrm{R})}(\mathrm{s}_\beta)\right)^2\right]\mathrm{I}_{d_{\alpha\beta}\square\mathrm{h}},\ \forall\, Chol_{(\mathrm{R})} \geq 0, \tag{7.4}$$

where $\mathrm{I}_{d_{\alpha\beta}\square h}$ is the indicator function for observations pairs (s_α, s_β) separated by the distance h and $N(\mathrm{h})$ is the number of observation pairs separated by h. Strictly speaking, the above model is a semi-variogram but the prefix "semi" shall be omitted and this convention is followed in our subsequent references to it. The above variogram model, however, is not suited for the analysis of disease incidences since it does not account for the heterogeneous population distributions. Following the approach developed by Monestiez et al. (2005, 2006), the experimental variogram is estimated as:

$$\begin{aligned}
\hat{\gamma}_{Chol_{(\mathrm{E(R)})}}(\mathrm{h}) = &\frac{1}{2N(\mathrm{h})}\\
&\times\sum_{\alpha,\beta=1}^{N(\mathrm{h})}\left(\frac{n(\mathrm{s}_\alpha)\square n(\mathrm{s}_\beta)}{n(\mathrm{s}_\alpha)+n(\mathrm{s}_\beta)}\left[Chol_{(\mathrm{R})}(s_\alpha) - Chol_{(\mathrm{R})}(s_\beta)\right]^2 - \overline{Chol}_{(\mathrm{R})}\right)\mathrm{I}_{d_{\alpha\beta}\square\mathrm{h}}
\end{aligned}, \tag{7.5}$$

where $N(\mathrm{h}) = \sum_{\alpha,\beta} \frac{n(s_\alpha)\cdot n(s_\beta)}{n(s_\alpha)+n(s_\beta)} \mathrm{I}_{d_{\alpha\beta} \square \mathrm{h}}$ is a normalizing constant and $\overline{Chol}_{(\mathrm{R})}$ is an estimate of the mean of $Chol_{(\mathrm{E(R)})}$ expressed as population weighted mean of the rates. Thus:

$$\overline{Chol}_{(\mathrm{R})} = \frac{\sum_{\alpha=1}^{N} n(s_\alpha)\cdot Chol_{(\mathrm{R})}(s_\alpha)}{\sum_{\alpha=1}^{N} n(s_\alpha)}. \qquad 7.6$$

In equation (7.5) the different pairs $\left[Chol_{(\mathrm{R})}(s_\alpha) - Chol_{(\mathrm{R})}(s_\beta)\right]$ are weighted by the corresponding population sizes $\frac{n(s_\alpha)\cdot n(s_\beta)}{n(s_\alpha)+n(s_\beta)}$ to homogenize their variance terms by dividing them by a weight proportional to the standard deviation $\sqrt{\overline{Chol}_{(\mathrm{E(R)})} \cdot \frac{n(s_\alpha)+n(s_\beta)}{n(s_\alpha)\cdot n(s_\beta)}}$.

Monestiez et al. (2005, 2006) developed the above variogram to account for the spatially heterogeneous observation efforts and sparse animal sightings for mapping the relative abundance of species (*Balenoptera physalus*). In their approach, the heterogeneous distribution of the observation efforts was modelled as Poisson distribution. Simulation studies indicated that this approach outperforms simple population-weighted approaches and Bayesian smoothers (Goovaerts, 2005). Generalization of this approach for disease mapping can be seen in Goovaerts (2005). The approach, however, is similar to Oliver et al. (1998), except that the Poisson distribution in Monestiez et al. approach replaces the Binomial distribution. In this study, the approach developed by Monestiez et al is employed to model the spatial auto-covariance structure of cholera variability in Kumasi. Next, a permissible variogram model by means of least squares $\gamma_{Chol_{(\mathrm{E(R)})}}(\mathrm{h})$ is fitted to the experimental variogram. From the fitted model, the maximum distance at which no spatial autocorrelation occurs (i.e. the range) is noted as d^{Th}.

7.2.3 Defining transmission network routes of cholera diffusion

Let $T = (t_1, ..., t_n)$ be a vector of serially ordered observed times of cholera onset at each community and ds_{ij} be the distance between pairs of communities s_i and s_j. The elements in the vector T are ordered such that $t_i \le t_j \ \forall i < j$. Using the date of index case in each community, a pair wise n x n directional transmission matrix $\overline{Com}_{(\mathrm{t,s})} = \left(\overline{Com}_{(\mathrm{t,s})i,j}\right)$ is constructed based on neighbourhood with previously infected community. The elements in the matrix $\overline{Com}_{(\mathrm{t,s})} = \left(\overline{Com}_{(\mathrm{t,s})i,j}\right)$ represent the probability of transmission from spatial unit s_j to s_i with respect to time and distance.

First, a binary spatial neighbourhood matrix $Com = \left(Com_{i,j}\right)$ is defined, with elements representing the probability of transmission between pairs of communities with respect to distance from each other or threshold distance at which cholera is considered contagious, thus $Com_{i,j} \in [0,1]$. Formally:

$$Com_{i,j} = \begin{cases} 1 & if\ ds_{i,j} \le d^{Th} \quad \forall i \ne j \\ 0 & otherwise \end{cases}. \tag{7.7}$$

Since no knowledge of the extent of contagiousness of cholera exists in the study area, the threshold distance d^{Th} is used as the threshold distance at which cholera is considered contagious.

Next, a temporal neighbourhood (directional) matrix $\overleftarrow{Com}_{(t)} = \left(\overleftarrow{Com}_{(t)i,j}\right)$ is defined, with elements representing the probability of a transmission from s_j to s_i with respect to time. More precisely, the elements in the matrix represent the probability that a spatial unit s_i with time of onset $t_{i\ge 2}$ can be infected by another spatial unit s_j with time of onset t; such that:

$$\overleftarrow{Com}_{(t)i,j} = \begin{cases} 1 & iff\ t_i > t_j \ \forall t_{i\ge 2}\ \&\ i \ne j \\ 0 & otherwise \end{cases}. \tag{7.8}$$

Here the assumption is that $\sum_{j=1}^{n} \overleftarrow{Com}_{(t)i,j} = 0$ for $t_{i=1,}$.

The final transmission matrix $\overleftarrow{Com}_{(t,s)} = \left(\overleftarrow{Com}_{(t,s)i,j}\right)$, which is a spatio-temporal matrix, is defined based on an element-wise multiplication of the spatial neighbourhood matrix $Com = \left(Com_{i,j}\right)$ and the time dependent neighbourhood matrix $\overleftarrow{Com}_{(t)}$. Thus, $\overleftarrow{Com}_{(t,s)} = Com \square \overleftarrow{Com}_{(t)}$ Formally:

$$\overleftarrow{Com}_{(t,s)i,j} = \begin{cases} 1 & iff \begin{cases} \overleftarrow{Com}_{(t)i,j} = 1, \forall i \ne j \\ Com_{i,j} = 1, \forall i \ne j \end{cases} \\ 0 & otherwise \end{cases}. \tag{7.9}$$

The matrix $\overleftarrow{Com}_{(t,s)}$ is made up of several transmission trees with *1-1* (transmission from one community to one community), *M-1* (transmission from many communities to one community) and *1-M* (transmission from one community to many communities) relationships. Theoretically, however, only *1-M* and *1-1* transmissions are feasible;

therefore, nearest neighbour approach based on direct distance is chosen to extract all *1-1* and *1-M* transmissions. These are subsequently mapped with GIS to identify the various possible index cases and their locations (See Figure 7.2).

7.2.4 Time-ordered diffusion modelling

The present study hypothesizes that the time-ordered sequence of appearance of cholera patterns has a dynamic relationship with the urban level and proximity to the primary cases location of the diffusion system. As such, the urban population represents the hierarchical component in the spread process, whereas the geographic distance from a *respective index case* represents the contagious component in the diffusion. The term *respective index case* is used because there are multiple index cases and each infected community corresponds to a particular index case. Here, the study adopts a non linear nonparametric Bayesian modelling approach for the effect of population and proximity on the diffusion of cholera.

Consider the observations $(y_i, x_i), i = 1,...,n,$ with response $y_i = \ln(t_i)$, and t_i the time of cholera onset in communities $s_i \in \{1,...,S\}$. The time of cholera onset t_i is relative to a respective index case. The vector $x_i = (d(s_i), n(s_i))'$ contains two metrical covariates; the population size $n(s_i)$, and the direct distance from an infected community to a respective index case location $d(s_i)$.

The study assumes that the response variable follows Gaussian distribution, i.e. $y_i | \eta_i, \sigma^2 \sim N(\eta_i, \sigma^2 / c_i)$, with unknown mean η_i of a nonparametric *geo-additive* model of the form:

$$\eta_i = f_n(n(s_i)) + f_d(d(s_i)) + f_{str}(s_i) + f_{unstr}(s_i). \qquad 7.10$$

Here, $f_n(n(s))$ and $f_d(d(s))$ are nonlinear smooth functions of the metrical covariates $n(s), d(s)$, respectively and $f_{str}(s)$ and $f_{unstr}(s)$ are the structured and unstructured spatial effects of the spatial covariates $s_i \in \{1,...,S\}$.

In order to explore the dependence of the rate of cholera infection r on $d(s)$ and $n(s)$, we used the higher order moments of the frequency distribution of cholera infection against time (Cliff et al., 1986). Thus, the response variable $y = \ln(t)$ is replaced with $y = r = m_3^2 / m_2^3$, where m_i is the *i*th central moment about the mean (average) time to infection $\bar{t}$. For each community, $\bar{t}$ is defined as: $\bar{t} = \frac{1}{n} \sum_{t=1}^{t_n} t \cdot Chol_{(C)t}$, where

$Chol_{(C)t}$ is the number of cholera cases at time t, t_n is the number of days of cholera existence, and $n = \sum Chol_{(C)t}$ for all t. The ith central moment about $\bar{t}$ is expressed as: $m_i = \frac{1}{n}\sum_{t=1}^{t_n}(t-\bar{t})^i Chol_{(C)t}$.

The rate of infection r was evaluated for all communities, excluding communities for which cases where recorded in only one day of the epidemic period. The model in equation (7.8) was fitted for the set of communities for which r was available.

Prior assumptions

The unknown model parameters are estimated by a fully Bayesian approach. Prior assumptions for the smooth functions $f_n(n(s))$ and $f_d(d(s))$ are specified. First, a second order random walk prior is imposed on the function evolutions $f_n(n(s))$ and $f_d(d(s))$. Following Lang and Brezger (2004), we suppose that $x_{(1)} < ... < x_{(t)} < ... < x_{(m)}$ are m ordered distinct values with equally spaced observations x_i, $i = 1,...,m$ with $m \le n$ that are observed for the covariates x and define $\xi_t = f_j(x_{(t)})$. Then $f_j(x)$ can be written as $f_j(x) = \upsilon'\xi$, where υ is a 0/1 incidence vector taking the value of one if $x = x_{(t)}$ and zero otherwise, and $\xi = (\xi_1,...,\xi_m)'$ is a vector of regression coefficients. The first order random walk prior for non-equidistance observations of adjacent values is defined as:

$$\xi_t = \xi_{t-1} + u_t, \; t = 2,...,m \qquad 7.11$$

with Gaussian distributed error terms $u_{(t)} \sim N(0, \delta_t\tau^2)$, where the variance depends on $\delta_t = x_{(t)} - x_{(t-1)}$. Random walks of the second order are defined by:

$$\xi_t = \left(1 + \frac{\delta_t}{\delta_{t-1}}\right)\xi_{t-1} - \frac{\delta_t}{\delta_{t-1}}\xi_{t-2} + u_t. \qquad 7.12$$

Here, $u_t \sim N(0, w_t\tau^2)$, where the weights w_t define the variances of the random walks. In this study we chose the simplest approach, where $w_t = \delta_t$.

A first order random walk penalizes abrupt jumps $\xi_t - \xi_{t-1}$ between successive states while a second order random walk penalizes deviations from the linear trend $2\xi_{t-1} - \xi_{t-2}$. The joint distribution of the regression parameters $\xi_j = (\xi_1,...,\xi_m)'$ is computed as the

product of conditional densities defined by Eq. (7.11). Diffuse priors $\xi_1 \propto \text{const}$, or ξ_1 and $\xi_2 \propto \text{const}$, are chosen as initial values, respectively. These specifications act as smoothness priors that penalize too rough functions. The general form of the prior for ξ_j is a multivariate Gaussian distribution with density

$$p\left(\xi_j \left| \tau_j^2 \right.\right) \propto \exp\left(-\frac{\xi_j' K \xi_j}{2\tau_j^2}\right). \qquad 7.13$$

The penalty matrix of order k is of the form $K = D_k' D_k$ where D_k is a first or second order difference matrix for $k = 1$ or 2, respectively. Since the penalty matrix K is often not of full rank, it follows that $\xi_j \left| \tau_j^2 \right.$ is an improper Gaussian prior, $\xi_j \left| \tau_j^2 \right. \propto N\left(0; \tau_j^2 K^-\right)$, where K^- is a generalized inverse of K. The tradeoff between flexibility and smoothness is controlled by the variance parameter τ_j^2. Thus, a small (large) value of τ_j^2 correspond to an increase (decrease) of the penalty or shrinkage. Here, a weakly informative inverse Gamma prior $IG\left(a;b\right)$ with hyper-parameters for τ_j^2 is used.

For the structured spatial effects, the neighbourhood matrix $\omega = \left(\omega_{i,j}\right)$ is modelled as a Gaussian random filed prior (Besag et al., 1991; Rue and Held, 2005). This prior is defined by the conditional distribution of $Com = \left(Com_{i,j}\right)$. Spatial units near the edges of the study area are likely to have fewer neighbors than those in the centre of the study area. Estimates of spatial units near the edges are less reliable than estimates of spatial units in the centre of the study area as fewer neighbors may distort any estimates for spatial units near the edges, the so called *edge effects*. To reduce edge effects, the conditional mean of $f_{str}\left(s_i\right)$ is chosen to be a weighted average of the function evaluations of neighbouring spatial units, with weighting scheme based on the proportion of the number of observed neighbour. Thus:

$$f_{str}\left(s_i\right) \left| f_{str}\left(s_j\right), s_j \neq s_i, \tau^2 \sim N\left(\sum_{s_j \in \partial s_i} \frac{w_{ij}}{\sum_{s_j \in \partial s_i} w_{ij}} f_{str}\left(s_i\right), \frac{\tau^2}{\sum_{s_j \in \partial s_i} w_{ij}} \right), \right. \qquad 7.14$$

where $s_j \in \partial s_i$ denotes that spatial unit s_j is a neighbour of spatial unit s_i. Here, the weights $w_{ij} = \left|\partial s_i\right| / N_s$, where $\left|\partial s_i\right|$ is the number of neighbors of spatial unit s_i and N_s is the total number of spatial units. The design matrix ψ of the spatial effects is 0/1 incidence matrix where the number of columns is equal to the number of spatial units. The variance parameter τ^2 controls the amount of smoothing of the spatial covariates and the degree of similarity.

For the unstructured effects, the parameters $f_{unstr}(s)$ are assumed to be *i.i.d*. Gaussian. Thus:

$$f_{unstr}(s_i)|\tau^2_{unstr} \sim N(0;\tau^2_{unstr}). \quad 7.15$$

In a fully Bayesian approach, the variance parameters τ^2_j, j = *n*, *d*, *str*, *unstr* are also considered as unknown and estimated simultaneously with the corresponding unknown functions $f_n(n(s))$, $f_d(d(s))$, $f_{str}(s)$, $f_{unstr}(s)$. Highly dispersed inverse gamma distribution $IG(a;b)$, with hyper-parameters are assigned to them in a second stage of the hierarchy.

Posterior estimation

Fully Bayesian inference is based on the posterior distribution of the unknown parameters. In this approach, samples are drawn from the full conditionals of the unknown parameters given the data through MCMC simulations. Let β represent the vector of all unknown functions to be evaluated (i.e., $\beta = (f_n(n(s)); f_d(d(s)); f_{str}(s); f_{unstr}(s))$ and τ represent a vector of all variance components; the posterior distribution then equals

$$p(\beta,\tau|y) \propto p(y|\beta)\,p(\beta|\tau)\,p(\tau), \quad 7.16$$

where $p(y|\beta)$ is the likelihood function of the data given the parameters and $p(.)$ represents the probability density function. Full conditionals for the unknown functions $f_n(n(s))$, $f_d(d(s))$, $f_{str}(s)$, $f_{unstr}(s)$ are multivariate Gaussian and, as a consequence, a Gibbs sampler for MCMC simulation is employed. Cholesky decompositions for band matrices have been used to efficiently draw random samples from the full conditional (Rue and Held, 2005; Rue, 2001). The model has been implemented in public domain software for Bayesian analysis, BayesX ver 2.0 (Brezger et al., 2005; Belitz et al., 2009). We used a total number of 40,000 MCMC iterations and 10,000 number of burn-in samples. Since, in general, these random numbers are correlated, only every 20th sampled parameter of the Markov chain were stored.

7.3 Results

The population distribution in the study area is highly variable ranging from 587 to 56,417 people and standard deviation of approximately 13,506. Such spatially varying populations induced heteroscedasticity in the disease rates as well as non-stationarity in the variances. Consequently, the experimental variogram computed with the raw disease

rates is uneven and exhibits less continuous patterns, depicting little or no spatial correlation and/or structure among communities (Figure 7.2b). This, however, necessitated an alternative to the Matheron's variogram estimator to characterize the spatial variability of the disease rates. Monestiez et al. variogram model can reveal structures that might be blurred by the random variability of extreme population values.

Experimental variograms were computed for the 68 community-level incidence rates using the traditional variogram (Eq. 7.2) and the Monestiez et al variogram models (Eq. 7.3). The spatial variability is considered isotropic since no systematic differences are observed between the directional variograms; hence, only the omni-directional variograms are displayed in Figure 7.2. The traditional variogram model for cholera rates is exponential with a practical range of 1.36 km (Figure 7.2b); while the Monestiez et al variogram model is spherical with a practical range of 2.3 km (Figure 7.2a). The relatively larger range of autocorrelation for the Monestiez et al variogram model indicates a better spatial structure for cholera rates after heterogeneity in population distribution is accounted for. Therefore $d^{Th} = 2.3\,\text{km}$ is used for the subsequent analysis.

Figure 7.3 shows the transmission networks routes of cholera diffusion in Kumasi. The red spots show the location of primary cases and starting points of different diffusion systems. The primary case locations are shown to be scattered, occurring simultaneously at distant locations. 12 main primary cases are indentified, each corresponding to a different diffusion system. The largest diffusion system involves 19 communities and recorded approximately 37% of cholera cases during the outbreak period. From the transmission network routes, five isolated communities are observed. These are not included in any of the diffusion systems (Figure 7.3). The geographic locations of the primary cases are used for modelling the effect of $d(s)$ on $\ln(t)$ and r.

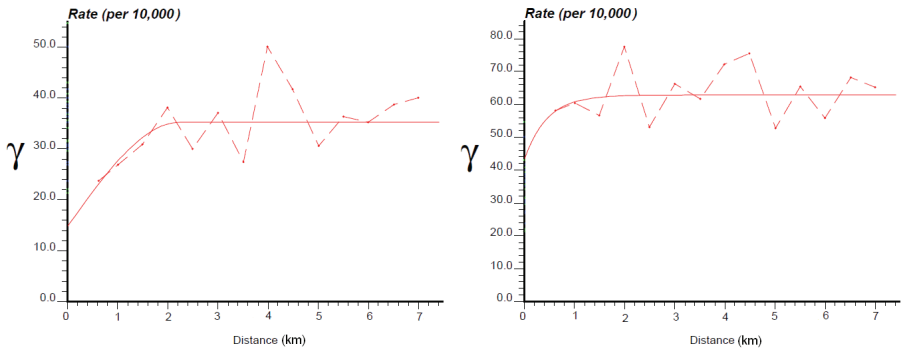

Figure 7.2: Experimental variograms computed for the 68 community-level cholera incidence rates. (left) Monestiez et al variogram model and (right) traditional variogram model.

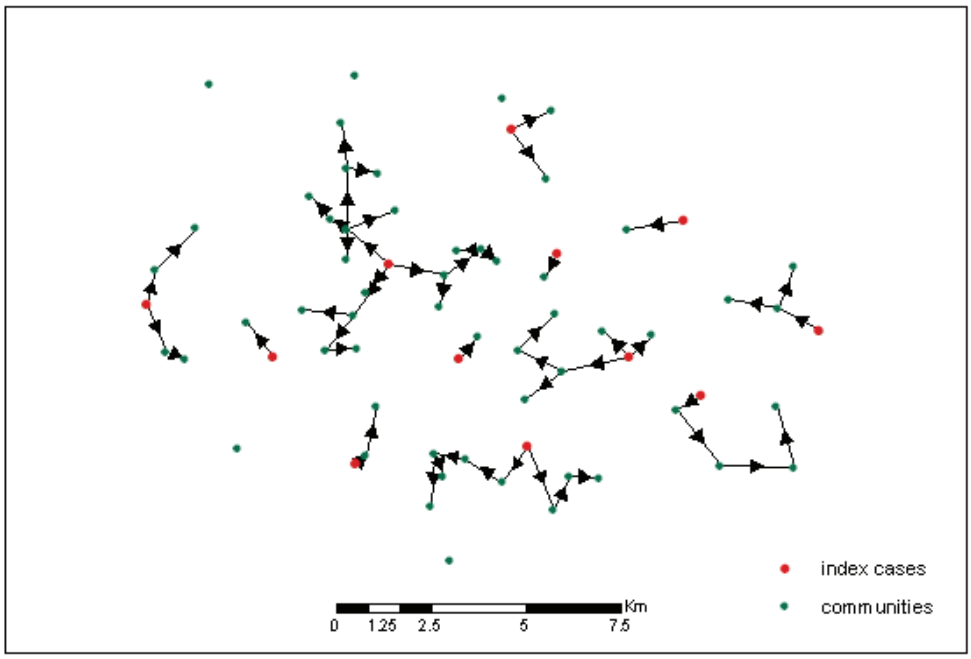

Figure 7.3: The diffusion network patterns showing *1-M* and *1-1* transmissions network routes. The arrows show the direction of the diffusion network, the red dots show the index/primary case locations, and the green dots show the location of various communities.

The nonlinear effects of the metrical covariates $n(s)$ and $d(s)$ on $\ln(t)$ and r are shown in Figure 7.4. The effect of $n(s)$ on $\ln(t)$ is nonlinear with decreasing posterior mean (Figure 7.4a). For $d(s)$, the posterior mean increases with increasing $\ln(t)$ (Figure 7.4b). The effect of $n(s)$ on r is almost linear with increasing posterior mean (Figure 7.4c). No systematic relationship is observed between r and $d(s)$ at $d(s) \leq 2.4\,\text{km}$ (Figure 7.4d). Thus, at $d(s) \leq 2.4\,\text{km}$, the relationship between r and $d(s)$ is fixed with neither decreasing nor increasing effect. At $d(s) > 2.4\,\text{km}$, however, a decreasing relationship is observed between r and $d(s)$.

Similar spatial patterns are observed for both $\ln(t)$ and r, hence only the patterns exhibited by r are shown. Figure 7.5 shows the estimated total spatial effects (left) and the corresponding 80% (credible interval) posterior probability map (right) of r. Areas shaded black show strictly negative credible intervals, whereas white areas depict strictly positive credible intervals; and grey indicate areas of non-significant spatial effects. There is a considerable spatial variation in the rate of cholera infection. Major spatial effects are observed at central part of the study area. The structured and unstructured spatial effects are given by the caterpillar plots in Figure 7.6. The wider

variations in the caterpillar plots of Figure 7.6a compared with Figure 7.6b show that the spatially structured effects are dominant over the unstructured effects.

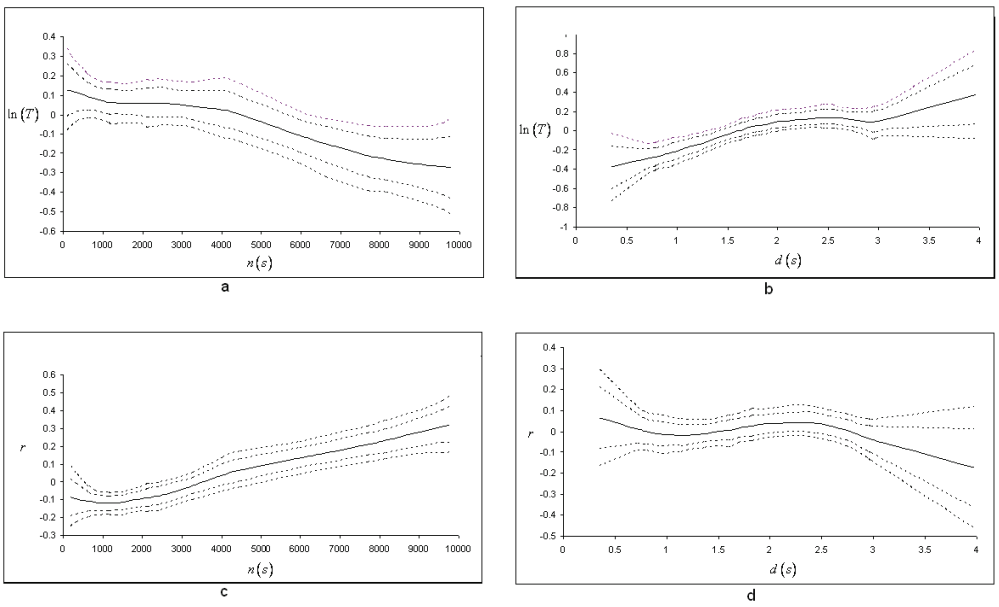

Figure 7.4: The estimated nonlinear effects of the metrical covariates (a) $n(s)$ on $\ln(t)$, (b) $d(s)$ on $\ln(t)$, (c) $n(s)$ on r, (d) $d(s)$ on r. The posterior mean together with the 80% and 90% credible intervals are also shown.

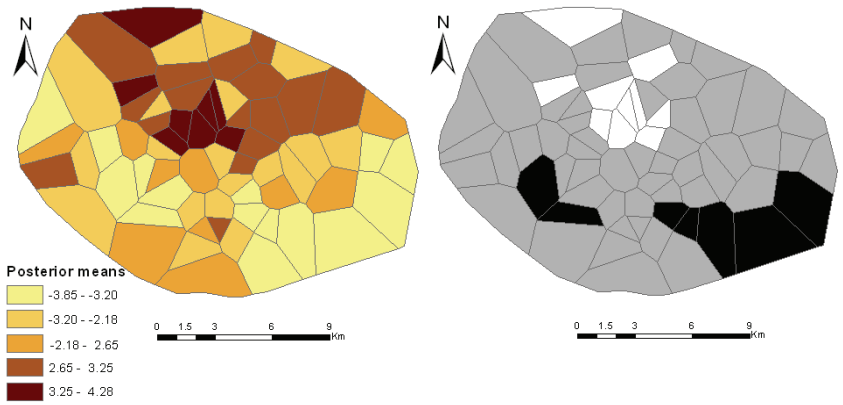

Figure 7.5: Spatial distribution of the posterior means of the total spatial effects of modelling the effects $n(s)$ and $d(s)$ on $\ln(t)$ (left), and posterior probabilities at nominal level of 80% (right). (left) Black denotes areas with strictly negative credible intervals; white denotes areas with strictly positive credible intervals, while grey shows areas of no significant difference.

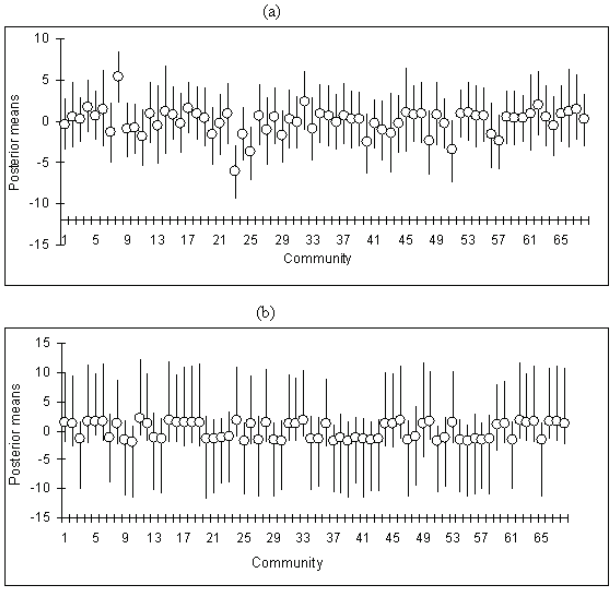

Figure 7.6: Caterpillar plots of the posterior means of the structured (a) and unstructured (b) spatial effects in Figure 7.5, with 90% error bars.

7.4 Discussion

This study utilizes statistical methods to explore the space-time diffusion dynamics of cholera incidences in Kumasi-Ghana. Variogram models are used to characterize the threshold of contagiousness of cholera. This threshold is subsequently used with the times of cholera entrance in each community to characterize all probable primary cases and diffusion systems during the outbreak. Finally, a Bayesian modelling approach is used to explore the space-time diffusion dynamics of cholera in Kumasi.

Several primary cases have been identified, each corresponding to a different diffusion system. This is an indication that the transmission of cholera during epidemic situations can start from several sources. This confirms the fact that primary transmission of cholera is responsible for sparking initial outbreaks. The primary case locations have been shown to be scattered, occurring almost simultaneously in distant areas with no apparent connection. This may be explained by the fact that *V. cholera* concentration in the environment is dominated by environmental drivers and the stochastic nature of these processes (Pascual et al., 2006).

The findings also imply that communities proximal to primary case locations are infected relatively early during the epidemics, with more distant communities infected at increasing later dates. Similarly, more populous communities are infected relatively early, with less populous communities infected at increasingly later dates. The plausibility of these implications could be explained by: (a) the existing hypothesis about the propagation of cholera and (b) the mode of cholera spread in a population and

the demographic structure of the study area. Firstly, cholera diffuses contagiously between surrounding communities following order of social interactions and/or geographic proximity (Pyle, 1969; Smallman-Raynor & Cliff, 1998a, 1998b; Trevelyan et al., 2005; Stock, 1976). Thus, it is likely for the disease to propagate from its origin to proximal communities earlier than communities which are farther away.

Secondly, cholera is a disease of deficient sanitation (Ackers et al., 1998) which is primarily driven by environmental (Huq et al., 2005) and demographic factors (Borroto and Martinez-Piedra, 2000; Osei and Duker, 2008). High population density puts pressure on existing sanitation systems, thus increasing the risk of early cholera infection in populous communities than less populous communities during an epidemic period.

When the response variable $y = \ln(t)$ is replaced by $y = r$, an expected observation is that r increases with $n(s)$, with an almost linear relationship. Thus, the effect of $n(s)$ on r is almost fixed. Such a relationship is plausible because in highly populous communities, many people live close together which results in shorter disease transmission paths; and therefore, higher rate of cholera infection. The passage of *V. cholerae* through the human host transiently increases the infectivity of *V. cholera*. Therefore, the existence of short-lived hyper-infective stage of *V. cholerae* could provide a mechanism for exhibiting a strong feedback between the past and present levels of infection (Merrell et al., 2002; Hartley et al., 2005), especially in a population where faecal contamination of water sources is high. The rate of exposure to short-lived hyper-infective *V. cholerae* could also be dominated by spatial interactions among infected communities, population density and/or urban level (Pyle, 1969; Smallman-Raynor and Cliff, 1998a, 1998b; Stock, 1976). Codeço (2001) reports that in an epidemic situation the initial reproduction rate of secondary cases is positively affected by the degree of contamination of water supply as well as the frequency of contacts with these waters, which is in turn influenced by demographic factors such as population density. The non systematic relation between r and $d(s)$ at approximately $d(s) \leq 2.4\,\text{km}$ may be explained by the contagious nature of cholera. Cholera is communicable which spreads contagiously amongst inhabitants from one community to another. Since proximal communities tend to exhibit similar socioeconomic and environmental characteristics, similar rate of cholera infection may be observed. Thus, for $d(s) \leq 2.4\,\text{km}$, no systematic relationship is observed between r and $d(s)$. The negative relationship, however, between r and $d(s)$ amongst distant communities, i.e. $d(s) > 2.4\,\text{km}$, is seemingly grounds for questioning the acceptance of this hypothesis.

This study cannot conclude that only the covariates $d(s)$ and $n(s)$ influence $\ln(t)$ and r due to the possibility of other unobserved influential covariates. Therefore, the inclusion of $f_{str}(s)$ and $f_{unstr}(s)$ in the model is meant to mimic the nature of unobserved influential covariates on $\ln(t)$ and r. From Figure 7.5, there is evidence of

significant increased rate of cholera infection at the central part, and a significant reduced rate of infection at the south-eastern part (the periphery) of Kumasi. The plausibility of these patterns may be explained by the fact that communities at the central part are highly populated with lots of slum settlers, while communities at the peripheries are moderately populated. As a consequence, shorter disease transmission path (higher rate of infection) is expected at the central part and longer disease transmission path is expected at the peripheries (lower rate of infection). These patterns also indicate the existence of possible unobserved covariates, some of which may be individual or household level, giving leads for further epidemiological research using purpose collected data.

Although several of these findings confirm existing hypothesis of cholera, this study resolves the methodological deficiencies of exploring the space-time diffusion patterns of infectious diseases. For instance in Trevelyan et al. (2005), a strictly linear model is imposed on the relationship between the period of observation of poliomyelitis and the joint effects population and distance from the epidemic centre. Also, in Kuo and Fukui (2007) the time-ordered cholera diffusion sequence is modelled as a linear logarithmic regression model, which is the functional relationship between the residents of infected counties and distances from epidemic origins. The strictly linear effects imposed in such models can obscure important non linear effects. Moreover, such models also underestimate important effects of spatial interactions amongst communities on the space-time diffusion patterns of infectious diseases.

7.5 Conclusion

This study applies statistical methods to explore the space-time diffusion patterns of cholera in Kumasi. We use Bayesian modelling approaches which allow joint analysis of the nonlinear effects of population hierarchy and geographic proximity on cholera infection. Our study reveals that the time-ordered sequence of appearance of cholera in a community has a dynamic relationship with the population hierarchy and proximity to primary case locations. Likewise, the rate of cholera infection increases with high population density. The geographic proximity to a primary case location, however, does not influence the rate of cholera infection. These findings provide significant information to help health planners and policy makers about the dynamics of cholera spread amongst communities.

8

Research findings, conclusions and recommendations for further research: A synthesis

"Life is like a circle. You walk and walk only to find yourself at the place you started from"

Henry Crow Dog,

8.1 Overview

Studies on cholera and other diarrhea related diseases in Ghana (example Obiri-Danso et al., 2005) so far have focused solely on the biological factors and characteristics of the individuals affected. Although such studies are useful, they omit the spatial and regional variation of critical risk factors that have s spatial component, thus failing to define territories at high risk. A better understanding of the spatial variation of incidences, and their relationships with important environmental risk will be useful to develop alternative and timely interventions to limit or prevent cholera. This study sought to study the spatial patterns of cholera, identify territories of high risk, and determine important environmental and socioeconomic risk factors that contribute to cholera transmission. The main objectives have been to utilize spatial statistical methodologies to (1) explore the spatial, temporal, and demographic patterns of cholera, (2) investigate the spatial dependency of cholera prevalence on environmental risk factors, and (3) investigate the space-time diffusion dynamics of cholera.

The focus of this final chapter is to bring together the most important results from the various chapters in order to gain a better understanding of the study. Detailed discussions of the findings have been given previously in the various chapters. In what follows, the major research findings, conclusions, and recommendations for further studies are presented. Since the separate datasets used could not be synchronized, possible confounding effects due to scale effects are not accounted for in this study. Detailed discussions of these effects and possible approaches of dealing with them are discussed.

8.2 Major research findings

8.2.1 Spatial and temporal patterns of cholera

Substantial distinct variation has been found in the spatial and temporal distribution of cholera. High cholera rates are clustered around Kumasi Metropolis (the central part of the region), with Moran's Index = 0.271 and $p < 0.001$(Chapter 2). Clustering of high rates was found to be persistent (1997-2001) at the central part of the region, while low rates persisted at the peripheries (Chapter 3). Significant cholera clusters were detected for the years 1998, 1999, 2001 (Table 3.1 and Figure 3.2). These clusters occur in areas surrounding Kumasi Metropolis and have persisted for the years 1998-1999. This indicates possible sustained transmission of cholera in districts within the central part of the region, especially in Kumasi Metropolis. Cholera is primarily driven by environmental factors (Huq et al., 2005), and since environmental processes are spatially continuous in nature (Webster et al., 1994), high incidence rates of the disease were expected to cluster together. Kumasi is the most urbanized and highly populated district in Ashanti Region. Such high population density can strain existing sanitation systems, thereby putting the inhabitants at increased risk of cholera transmission. Numerous slummy and/or squatter settlements exist in urban communities where environmental sanitation is poor.

In Chapters 2 and 3, Kumasi Metropolis was identified as the area where cholera has persisted with high transmission risk in the Ashanti Region. Therefore, it was worth zooming in and investigating clustering amongst communities (Chapters 4 and 5). Both the flexible and circular scan statistics were used. Cholera clusters were detected within the central part of Kumasi. The most likely cluster detected by the flexible scan statistics encompassed a relatively low number of communities but a higher relative risk than the most likely cluster detected by the spatial scan statistics (Chapter 5). This suggests that imposing a circular window for cluster analysis may unduly include areas which are less likely to be hot-spots. For instance, six communities within the most likely cluster of the circular scan statistic were not detected as hot-spots of cholera using the flexible scan statistic. The arbitrary nature of the clusters detected by the flexible scan statistic may also follow patterns of socio-economic and environmental factors. For instance, the distribution and frequency of potable water supply to various communities may be implicated.

8.2.2 Demographic patterns of cholera

High urbanization, high overcrowding, and adjacency to Kumasi Metropolis were found to be the most important risk factors of cholera in Ashanti Region (Chapter 2). In Chapter 3, high cholera rates were also found to be associated with poor sanitation, poor drinking water, and high migration. Surprisingly, the effects of drinking water, poor sanitation and high migration on cholera prevalence seem to be significant only in urban communities (Chapter 3, Table 3.4).

Although cholera is transmitted mainly through contaminated water or food, sanitary conditions in the environment play an important role. *V. cholerae* can spread rapidly in environments where living conditions are overcrowded and where there is no safe disposal of solid waste, liquid waste, and human faces. These conditions are met in urban communities in Ashanti Region. The high rate of urbanization leads to high level of overcrowding which necessarily results in shorter disease transmission. For instance in Ghana, surface water pollution is particularly found to be worse in urban and overcrowded communities. While these water bodies may contain the cholera *vibrios*, urban inhabitants resort to them for various household activities during periods of water shortages.

8.2.3 Dependency of cholera on refuse dumps

In Chapters 4 and 6, an objective was set to determine the dependency of cholera on environmental sanitation. This thesis utilized proximity to refuse dumps, and density of refuse dumps as surrogate measures of environmental sanitation. From Chapters 2 and 3, Kumasi Metropolis emerges as the area where sustained transmission of cholera has occurred; hence this district was chosen as the appropriate case study area to examine these effects. It was hypothesized that refuse dumps create environmental niches for cholera infection during the rainy season, and therefore inhabitants who live in close proximity to open-space refuse dumps should have higher cholera prevalence than those

farther. In addition, areas with high density of refuse dumps are expected to have higher cholera prevalence than areas with lower density.

A direct linear relationship was observed between cholera prevalence and dumps density (Chapters 4 and 6), and an inverse relationship with proximity to refuse dumps (chapters 4). However, proximity to refuse dumps influences the risk of cholera only in communities within a distance of approximately 500 m from refuse dumps (Chapter 6). This is consistent with the finding from Chapter 4 where a quantitative assessment of critical distance discrimination on experimental buffer zones around refuse dumps showed that the optimum spatial discrimination of cholera occurs at 500 m away from refuse dumps. The overall body of evidence suggests that cholera risk is relatively high when inhabitants live in close proximity to refuse dumps and where there are numerous refuse dumps. Two main reasons have been discussed (Chapter 4 and 6) to be plausible explanations to such findings i.e. (1) *High rate of contact with filth breeding flies*, and (2) *Flood water contamination*.

Flies are attracted by the odour emanating from refuse dumps, especially the common housefly. The indiscriminate feeding habits (feeding on filth and human food) of this fly combined with its structural morphology (presence of hair and sticky pads) make them ideally suited to carry and disseminate pathogenic micro organisms (Greenberg, 1973; Fotedar et al., 1992b; Kobayashi et al., 1999). This fly lives in close association with man feeding on all kinds of human food, garbage and excreta, and will travel not far from its breeding site (refuse dumps) to the nearest resting place. Therefore, inhabitants close to open-space refuse dumps tend to have a high rate of contact with these flies. Published reports have also shown that fly control measures can be effective in reducing the incidence of diarrhea (Watt and Lindsay, 1948; Cohen et al., 1991; Chavasse et al., 1999). Where high fly populations and poor hygiene conditions prevail, or where pathogens can grow within fly-contaminated food, the potential exists for transmitting pathogens with a high infectious dose (Nichols, 2005). Etiological studies have shown that *V. cholerae* survives well in faecal specimens if kept moist (Sack et al., 2004).

During cholera outbreaks, runoff from open space dumps during heavy rains may serve as the major pathway for the distribution of the bacteria, creating environmental niches for bacterial infection. Excreta may be washed away by rain-water and can run into nearby wells, streams and surface water bodies. The bacteria in the excreta may then contaminate these water bodies, and when used can perpetuate cholera infection.

8.2.4 Dependency of cholera on surface water pollution

In chapter 5, it was questioned whether the increased cholera prevalence near dump sites could be explained by increased transmission through flood water contamination in the proximity of the dump sites (as observed in Chapter 4). Also, the relationship between cholera prevalence and proximity to potentially contaminated surface water bodies was as yet unclear. To provide objective answers to the above questions, it was hypothesized that if runoffs from waste dumps during heavy rains serve as the major pathway for faecal and bacterial contamination of rivers and streams, then it is likely

that inhabitants living closer to water bodies where surface runoffs flow into will have higher cholera prevalence than those who live farther.

The hypothesized relationship between cholera prevalence and proximity to potential cholera reservoirs was largely confirmed (Chapters 5 and 6). Statistical models showed a significant relationship between cholera prevalence and proximity to the potentially polluted water. Thus, the increased cholera prevalence near dump sites, as observed in Chapter 4, can be explained by increased infection through flood water contamination as a result of runoff from refuse dumps. Comparing chapters 4 and 5, the relationship between cholera and potentially polluted surface water bodies seems stronger than that between cholera and refuse dumps. This, though arguable, is an indication that the effects of refuse dumps on cholera infection require surface water as an intermediate pathway. Therefore, any attempt to prevent defecation at dumps sites will reduce faecal contamination of rivers which will in effect reduce cholera infection. These findings support initial findings by other researchers. The classic epidemiological work of Snow (1855) revealed the association between cholera and contaminated water even before any bacteria were known to exist. The epidemiological studies of Ali et al. (2002a, 2002b) have also reported on proximity to surface water bodies as an important cholera risk factor in an endemic area of Bangladesh.

8.2.5 Dependency of cholera on slums

In chapter 6, risk of cholera infection was found to be high amongst communities with slum settlers. The risk is also higher in densely populated communities (See Table 6.2). These relationships may exist because most communities with slum settlers are densely populated. Although cholera is transmitted mainly through contaminated water or food, poor sanitary conditions in the environment enhance its transmission. The cholera *vibrios* can survive and multiply outside the human body and can spread rapidly where living conditions are overcrowded and where there is no safe disposal of solid waste, liquid waste, and human faeces (WHO, 2000). These conditions are mostly met in slummy and densely populated communities. Such high population density may result in shorter disease transmission paths, thus increasing the risk of cholera infection. Inhabitants living in slummy areas are generally poor, and face problems including access to potable water and sanitation. In many cases public utilities providers (e.g. water distribution) legally fail to serve these urban poor due to factors regarding land tenure system, technical and service regulations, and city development plans. Most slum settlements are also located at low lying areas susceptible to flooding. Unfavourable topography, soil, and hydro-geological conditions make it difficult to achieve and maintain high sanitation standards among inhabitants living in these territories (Borroto and Martinez-Piedra, 2000).

8.2.6 Dependency of cholera on spatial interaction

Spatial autoregressive coefficients were included in the statistical models to account for the possible effects of spatial interaction on cholera prevalence (Chapter 4 and 5). Significant autoregressive coefficients were observed for both the spatial lag and spatial

error models. This suggests that interactions between neighbors significantly influence the spread of cholera. The spatial autocovariance structure of cholera shows that this interaction may persist amongst communities within a neighbourhood of about 2.3 km (Chapter 7, Fig 7.3a). Social interaction and unobserved confounders could induce spatially correlated effects in the spatial distribution of diseases; hence the inclusion of the spatial autoregressive coefficients in the spatial models was used as surrogate measures of: (a) the effects of social interaction on cholera infection, and (b) possible unobserved risk factors of cholera. The significance of the autoregressive coefficients in the spatial lag models suggests the effects of interaction amongst communities on cholera spread. Communities which are closer in space tend to have similar cholera prevalence compared with communities farther apart. This suggests that similar environmental or demographic risk factors induce cholera transmission amongst the communities. The significance of the autoregressive coefficients in the spatial error models also suggests the existence of important unobserved risk factors of cholera.

The unobserved risk factors may be local or global depending on the spatial extent at which the spatial dependency persists. An alternative explanation for the significance of the autoregressive coefficients is the possible mismatch between the observed spatial unit and the true spatial scale of cholera transmission in the study area. Statistical maps of the spatial effects also provide evidence of distinct spatial variations (Chapters 6 and 7). A significant increase in cholera risk is observed at the central part, and a significant reduced risk at the south-eastern part (the periphery) of Kumasi (Figure 2). These patterns may be explained by the fact that communities at the central part are highly populated with lots of slum settlers, while communities at the peripheries are moderately populated.

8.2.7 Diffusion dynamics of cholera

Multiple primary cases were identified to spark the diffusion of cholera (Chapter 7, Fig 7.2). It was observed that the primary cases are scattered in space and time, occurring almost simultaneously in distant areas with no apparent connection. This is an indication that cholera outbreak could initiate and diffuse from multiple locations. This may be explained by the fact that *V. cholera* concentration in the environment is dominated by environmental drivers and the stochastic nature of environmental processes (Pascual et al., 2006).

In a diffusion model, it was realized that communities proximal to primary case locations are infected relatively early during epidemics, with more distant communities infected at later dates (Chapter 7, Fig). Similarly, more populous communities are infected relatively early, with less populous communities infected at increasingly later dates (Chapter 7, Fig 7).

Cholera is known to diffuse contagiously between surrounding communities following the order of social interactions and/or geographic proximity (Pyle, 1969; Smallman-Raynor & Cliff, 1998a, 1998b; Trevelyan et al, 2005; Stock, 1976). Thus, it is likely for the disease to propagate from its origin to proximal communities earlier than communities which are farther away. In densely populated communities, pressure on

sanitation systems can lead to an increase in the risk of early infection compared to less populous communities. The rate of infection is also observed to be high in densely populated communities (Chapter 7). Such a relationship is plausible because in highly populous communities, many people live close together which results in shorter disease transmission paths; and therefore, a higher rate of cholera infection. The passage of *V. cholerae* through the human host transiently increases the infectivity of *V. cholera*. Therefore, the existence of a short-lived hyper-infective *V. cholerae* could provide a mechanism for exhibiting a strong feedback between the past and present levels of infection (Merrell et al., 2002; Hartley, 2005), especially in a population where faecal contamination of water sources is high. The rate of exposure to short-lived hyper-infective *V. cholerae* could also be dominated by spatial interactions among infected communities, population density and urbanization (Pyle, 1969; Smallman-Raynor and Cliff, 1998a, 1998b; Stock, 1976). Codeço (2001) reports that in an epidemic situation the initial reproduction rate of secondary cases is positively affected by the degree of contamination of water supply as well as the frequency of contacts with these waters, which is in turn influenced by demographic factors such as population density.

8.3 Research conclusions

The main aim of this research was to use past cholera epidemic data and spatial statistical methodologies to study the spatial patterns of cholera, identify territories of high risk, and determine important environmental and socioeconomic risk factors that increase the risk of cholera infection. Based on the various research objectives and findings, the following conclusions have been drawn:

With regard to research objective 1 (Chapters 2 and 3), it is concluded that the observed non-random distribution and sustained transmission of cholera is influenced by demographic factors such as urbanization and overcrowding. The risk of cholera infection is also high when majority of the people do not have access to good sanitation facilities; drink from rivers, wells and ponds; and when migration is high.

With regard to research objective 2, the results from Chapters 4 and 6 reveal the spatial dependency of cholera infection upon proximity and density of refuse dumps in Kumasi. This means that refuse dumps serve as niches for cholera infection. The results also show that the minimum distance within which refuse dumps should not be located from a community is 500 m (Chapter 4). It is therefore hypothesize that proximity to and density of refuse dumps play a significant role in cholera transmission.

It is further deduced from Chapters 5 and 6 that the spatial distribution of cholera prevalence is dependent on proximity to potentially contaminated surface water bodies and the spatial neighbors of communities. Thus, proximity to potentially contaminated surface water bodies increases the risk of exposure to the cholera *vibrios*. The dependency of cholera prevalence on the spatial neighbors of communities indicates the existence of other confounding risk factors. Further studies, using purpose-collected household level data, will be very useful to elucidate all the critical risk factors of cholera.

Synthesizing Chapters 4, 5, and 6 together, the study concludes that runoffs from waste dumps during heavy rains serve as the major pathway for faecal and bacterial contamination of rivers and streams, thereby increasing the risk of cholera for inhabitants living closer to water bodies where these runoffs flow into (based on Chapters 4, 5, and 6). Therefore, any attempt to prevent faecal disposal at dumps sites will reduce faecal contamination of rivers which will in effect reduce cholera infection.

With regard to research objective 3, the study concludes that during cholera epidemics communities proximal to where index cases occur are infected relatively early compared with more distant communities (Chapter 7). Similarly, densely populated communities are infected relatively early in comparison with sparsely populated communities. The rate of infections is also higher amongst densely populated communities than sparsely populated areas. These findings provide significant information about the dynamics to help health planners and policy makers.

This study makes a number of novel contributions to understanding the epidemiology of cholera. In comparison with related cholera studies (Said, 2006: PhD thesis; Ruiz-Moreno, 2009: PhD thesis) and other scientific publication (For instance in Kwofie, 1976; Ackers et al., 1998; Borroto and Martinez-Piedra, 2000; Ali et al., 2002a, 2002b; Mugoya et al., 2005; Sasaki et al., 2008), the novelty of this thesis lies both in the significance of the statistical methods applied and the findings observed. First, the findings (Chapters 2 and 3) deviate from the already known popular idea that cholera is a rural disease. However, in Mexico, a Latin American country, cholera has been described as a rural disease (Barroto and Martnez-Piedra, 2000). This study has shown that, in developing countries like Ghana, cholera is an urban disease rather than a rural disease. Thus, the spatial and demographic patterns of cholera vary from one geographical area to another. Second, this study utilized proximity to and density of open-space refuse dumps to infer sanitation conditions in communities. This inference is novel compared with similar epidemiological studies of cholera. Most epidemiological studies infer mere proximity to water bodies as cholera risk (example in Ali et al., 2002a, 2002b). The current study utilizes GIS based spatial analyses to identify the steepest downhill paths along which runoff from point pollution sources (in this case refuse dumps sites) will flow. In which case a simple overlay operation could delineate all drainage channels these runoff will flow into. Methodologically, this study has improved on the existing techniques that environmental epidemiologists and medical geographers utilize to measure risk of exposure to an environmental determinant. The use of GIS and spatial analysis facilitates this type of methodological analysis which would be impossible in a non-spatial environment. Lastly, this study has shown the usefulness of spatial statistical methodologies in cholera research. For instance, the utilization of Bayesian Structured Additive Regression models to unveil the nonlinear nature of risk factors and diffusion dynamics of cholera is novel.

In conclusion, findings from this study prompt health officials and policy makers to execute measures to prevent faecal contamination of surface water bodies in order to prevent future cholera outbreaks in Kumasi. In order to achieve this, the following measures are suggested: (a) house-to-house collection of solid waste should be extended so as to reduce the dependency on open space dumps, (b) the Metropolitan Assembly

should enforce the bylaws to prevent industrial pollution of surface water bodies; for example, the direct discharging of industrial waste from brewery companies, (c) the Metropolitan Assembly should increase the provision of good public sanitation facilities, such as flush toilets or water closets rather than the existing ventilated pit latrines. They should also draw and implement bylaws that will enforce landlords to provide toilet facilities for tenants in their houses; (d) above all, the Ghana Water Company Limited (GWCL) should improve the water distribution system so as to ensure constant (24 hours a day) supply of treated piped water to all inhabitants in the metropolis.

8.4 Recommendations for future studies

In this study, inferences for the risk of cholera are based on groups of individuals both at the district and community levels. The inclusion of detailed house-hold level data in future studies will be useful to make inferences on relatively smaller groups of individuals. Case-control and cohort studies are better alternatives provided sufficient time and finances are available. This study could not prove fly transmission of *V. Cholerae* to humans, but only gives an indication of their possible involvement in transmission. Therefore, further epidemiological and fly control intervention studies are required to emphatically prove this hypothesis. However, an irrefutable acceptance or objection of this hypothesis can only be established if fly control studies are paralleled with cholera outbreak periods. Runoffs from open-space refuse dumps have been assumed as the major pathway for surface water pollution whereas other sources may exist. Further studies should look into the effect of other sources on cholera. Although the water quality characteristics of the delineated stream segments were not measured in this study, they were hypothetically thought to be potential cholera reservoirs. Microbiological and water quality analyses will be useful to provide a conclusive evidence of the related hypothesis. Finally, the link between periods of acute water shortages and cholera outbreaks should be established in future studies.

Notwithstanding the epidemiological significance of this study, several assumptions were imposed due to the available data. This led to several methodological limitations worth mentioning. Discussions of these effects and possible approaches of dealing with them are provided below.

8.4.1 Ecological fallacy

This study has been conducted within the framework of exploratory studies where aggregated health and exposure data are utilized. Generally, public health data aggregation is meant to protect patient privacy since the disclosure of patient locations and their associated conditions is a major breach of medical privacy protocols (Brownstein et al., 2006; Curtis et al., 2006). However, in Ghana, disease data aggregation is mainly due to lack of efficient reporting and surveillance systems. Socio-economic and demographic data collected at the individual level are often subjected to spatial aggregation prior to being made available to the public for research. Whether driven by privacy concerns or surveillance limitations, data aggregation is problematic

for data analysis since inferences regarding individuals become more difficult to understand (Anselin and Cho, 2002; Holt et al., 1996; Robinson, 1950). Spatial aggregation not only ignores the underlying variability of the data of interest, but also the spatial relationships between observations. In view of this, we reemphasize that inferences on the results of this study are based on group-level rather than individual-levels. This is crucial in order to avoid the so called *ecological fallacy* (Robinson, 1950), i.e. inferring individual-level relationships from group level data. It should be the aim of a further study to include possible individual or household exposure and health data to correctly estimate individual specific risk factors. For instance, case–control and cohort studies can give a relatively close approximation to the biologic model in investigating environmental health issues because both individual person characteristics and exposures are studied at the individual environment; thus the average disease risk of an individual will reflect individual characteristics.

8.4.2 Modifiable areal unit problem

It has long been known that the results of statistical analysis such as regression and correlation are dependent on the spatial framework within which data are collected, i.e. scale and aggregation (Gehlke and Biehl, 1934; McCarthy et al., 1956; Openshaw and Taylor, 1979; Openshaw, 1984). Thus, the spatial patterns of disease distribution may change if the same data are grouped into different sets of areal units, a phenomenon termed as modifiable areal unit problem (MAUP) (Openshaw, 1984). An extreme example was demonstrated by Openshaw (1984) showing variation of correlation coefficients. MAUP describes a geographic manifestation of ecological fallacy which arises from the uncertainty induced by the aggregation procedure, arising from the fact that areal units are not natural but arbitrary constructs that may not necessary have a relationship with the disease distribution (Openshaw, 1984). This phenomenon was first identified by Gehlke and Biehl (1934) and subsequently popularized by Openshaw and Taylor (1979, 1981). The effects of MAUP has been recognized in a variety of contexts including spatial interpolation (Cressie, 1995, 1996), regression analysis (Amrhein, 1995; Amrhein and Flowerdew, 1992; Clark and Avery, 1976; Fotheringham and Wong, 1991; Okabe and Tagashira, 1996), estimates of spatial autocorrelation (Chou, 1991), factor analysis (Hunt and Boots, 1996), image classification (Arbia et al., 1996), ecological modelling (Malanson and Armstrong, 1997) and regional economic forecasting (Miller, 1998). Two main separate effects of MAUP, i.e. *scale effect* and *zone effect*, usually occur simultaneously during the analysis of aggregated data. Scale effect causes variation in statistical results given different levels of aggregation, while *zone effect* describes variation in correlation statistics caused by the regrouping of data into different configurations but with the same scale.

Data limitations have enforced this study to be undertaken within a single-scale framework, thereby ignoring possible biases induced by MAUP. If data at different levels of spatial scales were available, possible bias of MAUP would be evaluated within a multi-scale analysis framework as exemplified in Odoi et al (2003). Re-aggregating the data to another set of areal units could assess the possible bias of MAUP (Atkinson and Molesworth, 2000). However, this was impossible due to the limited availability of higher resolution data and difficulties in assessing the ecological

fallacy associated. In accordance with the general rule of practice, the study analyzed aggregated data using the smallest areal units for which data were available to ameliorate the effects of aggregation. Accordingly, statistical inferences in this study are emphasized on the group-level rather than the individual-level.

Developing methods for dealing with MAUP still remains a crucial area of increasing research (Louie & Kolaczyk, 2006; Manley et al., 2006; Rushton and Lolonis, 1996; Swift et al., 2008). The effect of MAUP is widely recognized to be difficult to avoid and that no general solution has been agreed upon (Bailey and Gatrell, 1995; Goodchild, 2001). It may be considered as an unsolvable puzzle; nevertheless, it is imperative to recognize its effects, and become aware of its existence and impact. Quite a number of solutions have been proposed to evaluate and minimize the bias of MAUP (Openshaw and Charlton 1987; Besag and Newell 1991; Gatrell et al., 1996). Openshaw (1984) argues that the most appropriate response to the MAUP is to design purpose-specific zonal systems, also known as automated zoning procedure (Openshaw, 1977, 1984; Martin, 2003). Thus, rearrange the sets of areal units to match an optimal spatial variance. However, this approach is subject to criticism since it is focused towards achieving particular statistical results. A straightforward approach is to conduct spatial analysis at multiple scales (example Odoi et al., 2003). Adopting a multilevel approach using individual-level and aggregated-level data together can also assess the impact of MAUP bias (Greenland, 2001). It has been argued that the only real solution to avoid MAUP bias is to rely on individual-level data rather than aggregated data (Fotherinham et al., 2002; Zandbergen and Chakraborty, 2006). Yet, individual-level data are usually unavailable (Cromley and McLafferty, 2002; Meade and Earickson, 2000; Brownstein et al., 2006; Curtis et al., 2006). Alternatively, researchers can refrain from making inferences at the individual level when analyzing aggregated data (Waller and Gotway, 2004). Besides, data collection can be based on the features about which the researcher wants to make inferences. Fotherinham et al. (2002) suggest the use of local instead of a global parameter modeling, such as geographically weighted regression, to generate a surface of weighted local regression statistics. This approach sterns is based on the fact that global regression statistic over-simplify complex spatial relationships by smoothing over local spatial variations between datasets of predictor variables (Swift et al., 2008). Further discussions on current approaches to MAUP bias can be found in Swift et al. (2008), Matisziw et al. (2008), Hui (2009).

8.4.3 Edge effects

This study generally assumes a finite region for the study area, implying that a boundary is present and that any geographic distribution or spatial interaction occurring within the region may extend beyond its boundaries. Spatial units near the edges of the study area are likely to have fewer neighbors than those in the centre of the study area. Where interdependencies and interactions that occur among spatial units within and outside the boundary are ignored (as in this thesis), estimates of spatial units near the edges tend to be less reliable than estimates of spatial units in the centre of the study area as fewer neighbors may distort any estimates for spatial units near the edges, the so called *edge effects* (Griffith, 1983, 1985; Griffith and Amrheim, 1983). This effect arises when performing analyses that borrow strength from neighbouring spatial units. This has been

demonstrated to affect the analysis of small area health data (Lawson et al., 1999; Lawson, 2006; Vidal and Lawson, 2005).

Further work is needed to examine the implications of the edge effects in the estimation of the various results in this thesis. Several correction methods for edge effects have been proposed in the past, especially for spatially autoregressive models (Griffith, 1985; Griffith and Amrheim, 1983). The two main approaches for dealing with these effects are (1) the use of weighting/correction systems, which usually apply different weights to observations depending on their proximity to the study boundary, and (2) employing guard areas to provide external information to allow better boundary area estimation within the study window. Detailed discussion on other correction methods can be found in Dreassi and Biggeri (1998), Lawson (2006), Van Meter (2010).

8.4.4 Remote sensing and cholera prediction

Much still remain to be studied about the epidemiology of cholera. This study utilized spatial statistical methods to investigate the spatial and temporal patterns of cholera. Future studies incorporating remote sensing technologies will be a consequential effort to predict cholera outbreaks. Such integration will also be useful for the development of early warnings systems for cholera. Since *V. cholerae* is known to attach itself to the carapace and in the gut of copepods in large numbers, the copepod essentially serving as a vector the pathogen (Colwell, 1996; Nalin et al., 1979; Rawlings et al., 2007), climatic conditions favourable for multiplication of copepods and related chitinous zooplankton species can serve as proxies for the estimation of *V. cholerae* abundance in the environment. Although the bacterium cannot be sensed directly, remotely sensed data, such as sea surface temperature, sea surface height and chlorophyll concentration can be used to infer its presence, and therefore predicting cholera outbreaks. This discovery has essentially exposed the opportunities of utilizing remote sensing technologies for predicting cholera outbreaks. Only a handful of studies have been conducted to explore the potentials of remote sensing in predicting cholera outbreaks. Sea surface temperature, obtained from satellite data, has been found to be directly correlated with occurrence of cholera in Bangladesh (Colwell, 1996). Lobitz et al. (2000) have utilized remote sensing techniques to explore the relationship between sea surface temperature, sea surface height and cholera in Bangladesh. Significant correlation has also been found between remotely sensed data (i.e. precipitation, sea surface temperature, and chlorophyll concentration) and cholera cases in KwaZulu-Natal, South Africa (Mendelsohn and Dawson, 2007). Remote sensing can be an extremely useful tool in monitoring environmental conditions associated with cholera. Remote sensing techniques can provide a mechanism for monitoring cholera on a global scale and, most importantly, offers a model for testing the hypothesis that global climate phenomenon contributes to cholera outbreaks (Colwell and Huq, 2001).

8.4.5 Spatial data quality issues

The accuracy result of any spatial analysis depends on the quality of the datasets used. Certain issues relating to the concept of spatial data quality have, however, been ignored in this thesis. Consideration of such issues in future studies is vital, although it is doubtful whether it would lead to significant changes of results. Yet, it is worth discussing important aspects of data quality issues that should be considered in future studies. The subject of spatial data quality has long been a subject of importance for cartographers, surveyors, geographers and (see for example Blachut et al., 1979; Maling, 1989). An important part of spatial data quality concerns the description of error and uncertainty in spatial data (van Oort, 2005). Clarke and Clark (1995) defines data quality with respect to digital datasets as, "the part of the data statement that contains information that describes the source of observation or materials, data acquisition and compilation methods, conversations, transformations, analyses and derivation that the data has been subjected to, and the assumptions and criteria applied at any stage of its life. Spatial data are derived from the real world and understanding the processes involved determines its quality. Aalders (2002) partitioned this processes into two. These are: **Conceptualization**: The specification of what should be considered the real world and the abstraction of the selected objects; **Measurement**: The specification of the measuring methods and the measurement requirements for capturing the data. These are extensively explained in Laurini and Thompson (1992), Burrough and McDonnell (1998), Molenaar (1998), Fisher (1999), Raper (1999), Frank (2001), Uitermark (2001), Kresse and Fadaie (2004) and Leyk (2005). In Chapters 4 to 7 of this thesis, a community is conceptualized as a single point location and measured as the centroid of the community. Practically, such conceptualization and measurement is inappropriate; however, the non availability of polygon datasets for communities leaves no choice. This limits the ability to create and incorporate a boundary neighbourhood structure in statistical modeling; unless otherwise on creates artificial boundaries from the point location, which is also less applicable in spatial epidemiological studies. Also, exposure to risk factors are conceptualized as spatial proximity and density, and measured with Euclidean distance and kernel density, respectively. While these concepts and measurements are well understood in the field of spatial epidemiology, their usage is debatable.

The quality of spatial data also depends on the elements of spatial data quality. In this perspective, the International Standardization Organization (ISO) (2002) defines quality as "totality of characteristics of a product that bear on its ability to satisfy stated and implied needs" this definition remains vague and meaningless until an appropriate definition of these elements, i.e. *lineage, positional accuracy, attribute accuracy, logical consistency, completeness, semantic accuracy, usage, temporal quality, variation in quality, meta-quality and resolution*. Detail description of these elements can be found in Aalders (2002), Devillers et al. (2005), Aronoff (1989). These elements are self explanatory; hence the few relevant ones to this study are discussed. **Positional accuracy**: The presence of positional uncertainty can be seen in the various stages of the spatial data creation. Digitizing uncertainty can affect the accurate positions of communities, rivers and streams, and elevation contours. The effect of digitizing uncertainty in community location is considered insignificant since a whole community

is conceptualized as a single point. Digitizing uncertainty of the other spatial features, however, are worth exploring in future studies. **Attribute accuracy**: This refers to the accuracy of all attributes other than positional and temporal attributes of the datasets. Where hospital recorded disease datasets are used, it is possible the reported number of cases will not reflect the actual number of cases which occurred. In situations of mild clinical symptoms, patients might refuse to seek medical treatment, leading to a reduction in the actual number of cases. The incorporation of active disease reporting strategies in future studies will be useful; however, this will be expensive and time consuming. **Logical accuracy**: This describes the fidelity of relationships encoded in the data structure. The effects of assigning reported disease cases to wrong spatial locations and time periods should not be underestimated. Checks and balances to ensure logical accuracy should not be ignored in future studies. **Completeness**: Completeness is a measure of the absence of data and the presence of excess data. In epidemiological terms, completeness may be used as a measure of discrepancy in the reported number of cases. In the case of cholera, limitations of surveillance systems, such as inconsistency in case definition and a lack of a standard vocabulary normally lead to under reporting. Instances of over reporting, however, are less like to occur. **Temporal quality**: Ideally, explicit exploration of the temporal variability of a disease requires detailed information of the date of infection, preferably daily occurrences. Such data is rarely available since patients seek medical attention after clinical manifestation of the disease. Hence, hospital recorded data normally contains information about date of symptom rather than date of infection. Knowledge of the incubation period of the disease pathogen could be helpful to reconcile this anomaly, i.e. subtract the incubation period from the date of symptom to obtain the actual date of infection. This, however, will result in inconsistency if the incubation period depends on an interval's immunity to the pathogen.

Bibliography

Aalders HGJL (2002): The Registration of Quality in a GIS. In: Shi W, Fisher PF, Goodchild MF (Eds) Spatial Data Quality. Taylor and Francis, London, pp 186-299

Ackers M-L, Quick RE, Drasbek CJ, Hutwagner L, Tauxe RV (1998): Are there national risk factors for epidemic cholera? The correlation between socioeconomic and demographic indices and cholera incidence in Latin America. Int. J. Epidemiol. 27(2):330-334

Acosta JC, Galindo CM, Kimario J, Senkoro K, Urassa H, Casals C, Corachán M, Eseko N, Tanner M, Mshinda H, Lwilla F, Vila J, Alonso PL (2001): Cholera outbreak in southern Tanzania: risk factors and patterns of transmission. Emerg. Infect. Dis. 7(3):583-587

Alam A, LaRocque RC, Harris JB, Vanderspurt C, Ryan ET, Qadri F, Calderwood SB (2005): Hyperinfectivity of human-passaged Vibrio cholerae can be modeled by growth in the infant mouse. Infect. Immun. 73(1): 6674-6679

Ali M, Emch M, Donnay JP, Yunus M, Sack RB (2002a): Identifying environmental risk factors of endemic cholera: a raster GIS approach. Health & Place 8 (3): 201-210

Ali M, Emch M, Donnay JP, Yunus M, Sack RB (2002b): The spatial epidemiology of cholera in an endemic area of Bangladesh. Soc. Sci. & Med. 55(6):1015-1024

Ali M, Emch M, Yunus M, Sack RB (2001): Are the environmental niches of vibrio cholerae 0139 different from those of vibrio cholerae 01 El Tor? Int. J. Infect. Dis. 5(4):214-219

Alt KW, Vach W (1991): The reconstruction of "genetic kinship" in prehistoric burial complexes: Problems and statistics. In: Bock HH, Ihm P (Eds.) Classification, Data Analysis, and Knowledge Organization: Models and Methods with Applications. Springer, Berlin pp 299-310

Amrhein C G (1995): Searching for the elusive aggregation effect: Evidence from statistical simulations. Envron. & Plan. A 27(1) 105-119

Amrhein C G, Flowerdew R (1992): The effect of data aggregation on a Poisson regression-model of Canadian migration. Envron. & Plan. A 24(10):1381-1391

Anselin L (1988): Lagrange Multiplier test diagnostics for spatial dependence and spatial heterogeneity. Geogr. Anal. 20:1-17

Anselin L (1988): Spatial Econometrics: Methods and Models. Kluwer Press, Boston

Anselin L (1995): Local Indicators of Spatial Association: LISA. Geogr. Anal. 27:93-115

Anselin L (2002): Under the hood: Issues in the specification and interpretation of spatial regression models. Agric. Econ. 27:247-267

Anselin L, Acs Z, Varga A (1997): Local Geographic Spillovers between University Research and High Technology Innovations. J. Urban Econ. 42:422-448

Anselin L, Bera A (1988): Handbook of Applied Economic Statistics. Marcel Dekker Press, New York

Anselin L, Tam Cho, WK (2002): Spatial effects and ecological inference. Political Analysis 10(3):276-297

Arbia G, Benedetti R, Espa G (1996): Effects of MAUP on image classification. Geogr. Syst. 3:123-141

Aronoff S (1989): Geographic information systems: a management perspective. WDL, Ottawa, pp 294

Ashitey GA (1994): An epidemiology of disease control in Ghana 1901-1990. University Press, Accra-Ghana

Atkinson P, Molesworth A (2000): Geographical analysis of communicable disease data. In: Elliot P, Wakefield JC, Best NG, Briggs DJ (Eds) Spatial Epidemiology; Methods and Applications. Oxford University Press, New York, pp 253-266

Barnes S, Peck A (1994): Mapping the future of health care: GIS applications in health care analysis. Geogr. Inf. Syst. 4:31-3

Barton DE, David FN, Merrington M (1965): A criterion for testing contagion in time and space. Ann. Hum. Genet. 29:97-103

Barua D (1972): The global epidemiology of cholera in recent years. Proc. R. Soc. Med. 65:423-28

Barua D, Paguio AS (1977): ABO blood groups and cholera. Ann. Hum. Biol. 4:489-92

Beck LR, Rodrigues MH., Dister SW, Rodrigues AD, Rejmankova E, Ulloa A, Meza RA, Roberts DR, Paris JF, Spanner MA, Washino RK, Hacker C, Legters LJ (1994): Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. Am. J. Trop. Med. Hyg. 51(3):271-280

Belitz C, Brezger A, Kneib T, Lang S (2009): BayesX-Software for Bayesian inference in structured additive regression models. Version: 2.0. http://www.stat.uni-muenchen.de/~bayesx. Last accessed 3 November 2010

Belitz C, Brezger A, Kneib T, Lang S (2009): BayesX-Software for Bayesian inference in structured additive regression models. Version: 2.0. http://www.stat.uni-muenchen.de/~bayesx. Last accessed 3 November 2010

Bellec S, Hemon D, Rudant J, Goubin A, Clavel J (2006): Spatial and space-time clustering of childhood acute leukaemia in France from 1990 to 2000: a nationwide study. Br. J. Cancer 94(5):763-770

Besag J (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. B 36(2):192-225

Besag J (1975): Statistical analysis of non-lattice data. J. R. Stat. Soc. D 24(3):179-195

Besag J, Kooperberg C (1995): On conditional and intrinsic autoregressions. Biometrika 82(4):733-746

Besag J, Newell J (1991): The detection of clusters in rare disease. J. R. Stat. Soc. A 154(1): 143-155

Besag J, York Y, Mollié A (1991): Bayesian image-restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. 43(1):1-20

Biller C (2000): Adaptive Bayesian regression splines in semiparametric generalized linear models. J. Comput. Graph. Stat. 9(1):122-140

Biller C, Fahrmeir L (2001): Bayesian varying-coefficient models using adaptive regression splines. Stat. Model. 1:195-211

Birmingham ME, Lee LA, Ndayimirije N, Nkurikiye S, Hersh BS, Wells JG, Deming MS (1997): Epidemic cholera in Burundi: patterns of transmission in the Great Rift Valley lake region. Lancet 349(9057):981-985

Blachut TJ, Chrzanowski A, Saastamoinen JH (1979): Urban surveying and mapping. Springer, New York, pp 372

Blake PA (1993): Epidemiology of cholera in the Americas. Gastroenterol. Clin. North Am. 22(3):639-660

Bonilla-Castro E, Rodriguez P, Carrasquilla G (2000): La enfermedad de La pobreza: El colera en los tiempos modernos. Santafe de Bogota: Ediciones Uniandes. pp 292

Boots BN, Getis A (1998): Point pattern analysis. Sage Publications, Newbury Park, CA

Borroto RJ, Martinez-Piedra R (2000): Geographical patterns of cholera in Mexico, 1991-1996. Int. J. Epidemiol. 29(4):764-772

Braddock M, Lapidus G, Cromley E, Cromley R, Burke G, Branco L (1994): Using a geographic information system to understand child pedestrian injury. Am. J. Public Health 84(7):1158-61

Brezger A (2004): Bayesian P-Splines in Structured Additive Regression Models. PhD dissertation, Universität München

Brezger A, Kneib T, Lang S (2005): BayesX: Analyzing Bayesian structured additive regression models. J. Stat. Softw 14:11

Brezger A, Lang S (2003): Generalized additive regression based on Bayesian P-splines. SFB 386 Discussion paper 321, Department of Statistics, Universität München

Brownstein JS, Cassa CA, Mandl KD (2006): No place to hide-reverse identification of patients from published maps. N. Engl. J. Med. 355(16):1741-1742

Burridge P (1980): On the Cliff-Ord test for spatial autocorrelation. J. R. Stat. Soc. B 42: 107-108

Burrough PA, McDonnell RA (1998): Principles of Geographical Information Systems. Oxford University Press, Oxford UK, pp 333

Byrd J, Xu HS, Colwell RR (1991): Viable but nonculturable bacteria in drinking water. Appl. Environ. Microbiol. 57(3):875-878

Carpenter C (1971): Principles and Practice of Cholera Control. Ann. Intern. Med. 74(6):1021

Carpenter C, Barua D, Sack R (1966): Clinical studies in asiatic cholera. IV. Antibiotic therapy in cholera. Bull. Johns. Hopkins Hosp. 118:230-242

Chaput EK, Meek JI, Heimer R (2002): Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut. Emerg. Infect. Dis. 8(9): 943-948

Chavasse DC, Shier RP, Murphy OA, Huttly SRA, Cousens SN, Akhtar T (1999): Impact of fly control on childhood diarrhoea in Pakistan: community-randomised trial. Lancet 353(9146):22-25

Chen J, Roth RE, Naito AT, Lengerich EJ, MacEachren AM (2008): Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. Int. J. Health Geogr. 7:57

Chevallier E, Grand A, Azais J-M (2004): Spatial and temporal distribution of cholera in Ecuador between 1991 and 1996. Eur. J. Public Health 14(3):274-279

Chou Y-H (1991): Map resolution and spatial autocorrelation. Geogr. Anal. 23:228-246

Christensen R, Johnson W, Pearson LM (1992): Prediction diagnostics for spatial linear models. Biometrika 79:583-591

Christensen R, Pearson LM, Johnson W (1993): Case-deletion diagnostics for mixed models. Technometrics 34:133-169

Clark WAV, Avery K (1976): The effects of data aggregation in statistical analysis. Geogr. Anal. 8:428-438

Clarke DG, Clark DM (1995): Lineage. In: Guptil SC, Morrsion JL (Eds) Elements of spatial data quality. Elsevier Science Ltd, Oxford, UK

Clarke KC, Osleeb JR, Sherry JM, Meert JP, Larsson RW (1999): The use of remote sensing and geographic information systems in UNICEF's dracunculiasis (Guinea worm) eradication effort. Prev. Vet. Med. 11:229-35

Clarke KC, Sara LM, Barbara JT (1996): On Epidemiology and Geographic Information Systems: A Review and Discussion of Future Directions. Emerg. Infect. Dis. 2(2):85-92

Clayton D, Bernardinelli L (1992): Bayesian methods for mapping disease risk. In: Elliott P, Cuzick J, English D, Stern R (Eds) Geographical and Environmental Epidemiology: Methods for Small-Area Studies. Oxford University Press, Oxford, pp 205-220

Clayton D, Kaldor J (1987): Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43(3):671-681

Cliff AC, Ord J (1980): Spatial processes: models and applications. Pion Limited, London

Codeço CT (2001): Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir. BMC Infect. Dis. 1:1

Cohen D, Green M, Block C, Dlepon R, Ambar R, Wasserman SS, Levine MM (1991): Reduction of transmission of shigellosis by control of houseflies (*Musca domestica*). Lancet 337(8748):993-7

Colwell RR (1996): Global climate and infectious disease: the cholera paradigm. Science 274(5295):2025-2031

Colwell RR, A Huq (1994): Environmental reservoir of *Vibrio cholerae*. The causative agent of cholera. Ann. NY Acad. Sci. 740(1):44-54

Colwell RR, Brayton P, Harrington P, Tall B, Huq A, Levine M (1996): Viable but non-culturable *Vibrio cholerae* O1 revert to a cultivable state in the human intestine. World J. Microbio. Biotechnol. 12:28-31

Colwell RR, Brayton PR, Grimes DJ, Roszak DR, Huq SA, Palmer LM (1985): Viable but non-culturable *Vibrio holerae* and related environmental pathogens in the environment: implication for release of genetically engineered microorganisms. Nat. Biotechnolo. 3:817-820

Colwell RR, Huq A (1994): Vibrios in the environment: viable but nonculturable vibrio cholerae. In: Wachsmoth I, Blaker P, Olsvik O (Eds) Vibrio Cholerae and Cholera: Molecular to Global Perspectives. Am. Soc. Microbiol, Washington DC, pp 117-134

Colwell RR, Huq A (2001): Marine ecosystems and cholera. Hydrobiologia 460:141-145

Colwell RR, Huq A, Islam MS, Aziz KMA, Yunus M, Kahn NH, Mahmud A, Sack RB, Nair GB, Chakraborty J, Sack DA, Russek-Cohen E (2003): Reduction of cholera in Bangladeshi villages by simple filtration. Proc. Natl. Acad. Sci. 100(3):1051-1055

Colwell RR, Kaper J, Joseph SW (1977): *Vibrio cholerae*, *Vibrio parahaemolyticus*, and other vibrios: occurrence and distribution in Chesapeake Bay. Science 198(4315):394-396

Colwell RR, Patz JA (1998): Climate, infectious disease and health: an interdisciplinary perspective. Am. Acad. Microbiol. Washington DC

Colwell RR, Spira WM (1992): The ecology of Vibrio cholerae 01. In: Barua D, Greenough III WB (Eds) Cholera. Plenum Medical Book Company, New York, London pp 107-128

Cousens EK, Smith PG, Ward H, Everington D, Knight RSG (2001) Geographical distribution of variant Creutzfeldt-Jakob disease in great Britain, 1994-2000. Lancet 357(9261): 1002-1007

Cressie N (1993): Statistics for Spatial Data. John Wiley & Sons, New York
Cressie N (1995): Bayesian smoothing of rates in small geographic areas. J. Reg. Sci. 35:659-673
Cressie N (1996): Change of support and the modifiable areal unit problem. Geogr. Syst. 3:159-180
Cromley E, McLafferty S (2002): GIS and public health. The Guilford Press, New York
Crowcroft NS (1994): Cholera: current epidemiology. Commun. Dis. Rep. CDR Rev. 4(13):R157-164
Curtis A, Mills JW, Leitner M (2006): Keeping an eye on privacy issues with geospatial data. Nature 441(7090):150
Cuzick JC, Edwards R (1990): Spatial clustering for inhomogeneous populations. J. R. Stat. Soc. B 52:73-104
Cvjetanovic B, Barua D (1972): The seventh pandemic of cholera. Nature 239:137-38
de Magny GC, Cazelles B, Guegan JF (2007): Cholera threats to humans in Ghana as influenced by both global and regional climatic variability. EcoHealth 3(4):223-231
de Magny GC, Murtugudde R, Sapiano MRP, Nizam A, Brown CW, Busalacchi AJ, Yunus M, Nair GB, Gil AI, Lanata CF, Calkins J, Manna B, Rajendran K, Bhattacharya MK, Huq A, Sack RB, Colwell RR (2008): Environmental signatures associated with cholera epidemics. Proc. Natl. Aced. Sci. USA 2008 105(6):17676-17681
Denison DGT, Mallick BK, Smith AFM (1998): Automatic Bayesian curve fitting. J. R. Stat. Soc. B 60:333-350
Devillers R, Bedard Y, Jeansoulin R (2005): Multidimensional management of geospatial data quality information for its dynamic use within GIS. Photogramm. Eng. Remote Sens. 71(2):205-215
Diggle P, Chetwynd A (1991): Second order analysis of spatial clustering for inhomogeneous populations. Biometrics 47:1155-1163
Diggle P, Chetwynd A, Haggkvist R, Morris SE (1995): Second-order analysis of space-time clustering. Stat. Methods Med. Res. 4(2):124-36
DiMatteo I, Genovese CR, Kass RE (2001): Bayesian curve-fitting with free-knot splines. Biometrika 88(4):1055-1071
Dipeolu OO (1982): Laboratory investigations into the role of Musca vicina and Musca domestica in the transmission of parasitic helminth eggs and larvae. Int. J. Zoonoses 9: 57-61
Draper NR, Smith H (1998): Applied Regression Analysis. 3rd ed. John Wiley & Sons, New York
Dreassi E, Biggeri A (1998): Edge effect in disease mapping. J. Ital. Stat. Soc. 3:267-283
Dubois AE, Sinkala M, Kalluri P, Makasa-Chikoya M, Quick RE (2006): Epidemic cholera in urban Zambia: hand soap and dried fish as protective factors. Epidemiol. Infect. 134(6):1226-1230
Duczmal L, Assunção RA (2004): Simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Comput. Stat. & Data Anal. 45: 269-286
Duker AA, Caranza EJM, Hale M (2004): Spatial dependency of Buruli ulcer prevalence on arsenic-enriched domains in Amansie West District, Ghana: implications for arsenic mediation in Mycobacterium ulcerans infection. Int. J. Health Geogr. 3:19

Duker AA, Stein A, Hale M (2006): A statistical model for spatial patterns of Buruli ulcer in the Amansie West district, Ghana. Int. J. Appl. Earth Obs. Geoinf. 8:126-136

Dwass M (1957): Modified randomization tests for non-parametric hypothesis. Ann. Math. Stat. 28: 181-187

Ederer F, Myers MH, Mantel N (1964): A statistical problem in space and time: Do leukaemia cases come in clusters? Biometrics 20:626-638

Editorial (1971): Cholera in Spain. Br. Med. J. 3: 266

Eilers PHC, Marx BD (1996): Flexible smoothing with *B*-splines and penalties. Stat. Sci. 11(2):89-121

Elliott P, Wakefield JC (2000): Bias and confounding in spatial epidemiology. In: Elliott P, Wakefield J, Best NG, Briggs DJ (Eds) Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford, pp 68-84

Elliott P, Wakefield JC, Best NG, Briggs DJ (2000): Spatial Epidemiology: Methods and Applications. In: Elliott P, Wakefield J, Best NG, Briggs DJ (Eds) Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford, pp 1-29

Elliott P, Wartenberg D (2004): Spatial Epidemiology: Current Approaches and Future Challenges. Environ. Health Perspect. 112(9):998-1006

Emch M, Feldacker C, Yunus M, Streatfield PK, DinhThiem V, Canh DG, Ali M (2008): Local Environmental Predictors of Cholera in Bangladesh and Vietnam. Am. J. Trop. Med. Hyg. 78(5):823-832

Fahrmeir L, Kneib T, Lang S (2004): Penalized structured additive regression for space-time data: a Bayesian perspective. Stat. Sin. 14: 731-761

Fahrmeir L, Lang S (2001a): Bayesian semiparametric regression analysis of multicategorical time-space data. Ann. Inst. Stat. Math. 53(1): 11-30

Fahrmeir L, Lang S (2001b): Bayesian inference for generalized additive mixed models based on Markov random field priors. J. R. Stat. Soc. C 50(2):201-220

Fahrmeir L, Tutz G (2001): Multivariate Statistical Modeling Based on Generalized Linear Models. Springer, New York

Falsenfeld O (1965): Notes on food, beverages and fomites contaminated with *Vibrio cholerae*. Bull. World Health Organ. 33(5):725-734

Falsenfeld O (1966): A review of recent trends in cholera research and control. Bull. World Health Organ. 34(2): 161-195

Fan J, Gijbels I (1996): Local Polynomial Modeling and Its Applications, Chapman and Hall, London

Fang L, Yan L, Liang S, de Vlas SJ, Feng D, Han X, Zhao W, Xu B, Bian L, Yang H, Gong P, Richardus JH, Cao W (2006) Spatial analysis of hemorrhagic fever with renal syndrome in China. BMC Inf. Dis. 6:77

Faruque SM, Naser IB, Islam MJ, Faruque ASG, Gosh AN, Nair GB, Sack DA, Mekalanos JJ (2005): Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages. Proc. Natl. Acad. Sci. USA 102(5): 1702-1707

Faruque SM, Albert MJ and Mekalanos JJ (1998): Epidemiology, genetics and ecology of toxigenic *Vibrio cholerae*. Microbiol. Mol. Biol. Rev. 62(4):1301-1314

Feachem RG (1981): Environmental aspects of cholera epidemiology. II. Occurrences and survival of *Vibrio cholerae* in the environment. Trop. Dis. Bull. 78(10): 856-880

Finkelstein RA (1996): Cholera, Vibrio cholerae O1 and O139 and other pathogenic vibrios. In: Baron S (Ed) Medical Microbiology. 4th ed. Churchill Livingstone, New York

Fisher PF (1999): Models of uncertainty in spatial data. In: Longley PA, Goodchild MF, Maguire DJ, Rhind DW (Eds) Geographical Information Systems: Principles and technical issues. 2nd ed. John Wiley & Sons, New York, pp 191-205

Fleming G, Marwe M, McFerren G (2007): Fuzzy experts system and GIS for cholera health risk prediction in southern Africa. Environ. Modell. Softw. 22(4):442-448

Fotedar R (2001): Vector potential of houseflies (Musca domestica) in the transmission of Vibrio cholerae in India. Acta. Trop. 78(1):31-34

Fotedar R, Banerjee U, Samantary JC, Shriniwas (1992a): Vector potential of hospital houseflies with special reference to Klebsiella species. Epidemiol. Infect. 109(1):143-147

Fotedar R, Banerjee U, Singh S, Shriniwas, Verma AK (1992b): The housefly (*Musca domestica*) as a carrier of pathogenic microorganisms in a hospital environment. J. Hosp. Infect. 20(3):209-215

Fotheringham AS, Brunsdon C, Charlton ME (2002): Geographically weighted regression: The analysis of spatially varying relationships. John Wiley & Sons, Chichester

Fotheringham S, Wong D (1991): The modifiable areal unit problem in multivariate statistical analysis. Environ. & Plan. A 23(7):1025-1044

Frank AU (2001): Tiers of ontology and consistency constraints in geographical information systems. Int. J. Geogr. Inf. Sci. 15(7):667-678

Friedman JH (1991): Multivariate adaptive regression splines (with discussion). Ann. Stat. 19:1-141

Friedman JH, Silverman BL (1989): Flexible Parsimonious Smoothing and Additive Modeling (with discussion). Technometrics 31(1):3-39

Fukuda Y, Umezaki M, Nakamura K, Takano T (2005): Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan. Int. J. Health Geogr. 4:16

Gaffga NH, Tauxe RV, Mintz ED (2007): Cholera: A New Homeland in Africa? Am. J. Trop. Med. Hyg. 77(4): 705 -713

Gatrell AC, Bailey TC, Diggle PJ, Rowlingson BS (1996): Spatial point pattern analysis and its application in geographical epidemiology. Trans. Inst. Br. Geogr. 21(1):256-274

Gatrell AC, Bailly TC (1996): Interactive spatial data analysis in medical geography. Soc. Sci. Med. 42(6):843-855

Gehlke CE, Biehl K (1934): Certain effects of grouping upon the size of the correlation in census tract material. J. Am. Stat. Ass. 29(185):169-170

Gelfand AE, Smith AFM (1990): Sampling-based approaches to calculating marginal densities. J. Am. Stat. Ass. 85(410):398-409

Getis A, Ord JK (1992): The analysis of spatial association by use of distance statistics. Geogrr. Anal. 24:189-206

Gil AL, Louis VR, Rivera ING, Lipp E, Huq A, Lanata CF, Taylor DN, Russek-Cohen E, Choopun N, Sack RB, Colwell RR (2004): Occureence and dirtibution of Vibrio cholerae in the coastal environment of Peru. Environ. Microbiol. 6(7):699-706

Glass GE (2000): Update: Spatial Aspects of Epidemiology: The Interface with Medical Geography. Epidemiol. Rev. 22(1):136-139

Glass GE, Schwartz B, Morgan JM, Johnson DT, Noy PM, Israel E (1995): Environmental risk factors for Lyme disease identified with geographic information system. Am. J. Public Health 85(7):944-948

Glass R, Claeson M, Blake P, Waldman R, Pierce N (1991): Cholera in Africa: Lessons on transmission and control for Latin America. Lancet 338(8770):791-795

Glass RI, Becker S, Huq MI, Stoll BJ, Khan MU, Merson MH, Lee JV, Black RE (1982): Endemic cholera in rural Bangladesh, 1966-1980. Am. J. Epidemiol. 116(6): 959-970

Glass RI, Black R (1992): Epidemiology of cholera. In: Barua D, Greenough WB III (Eds) Topics in infectious diseases: cholera: Plenum Medical Company, New York, pp 129-154

Glass RI, Holmgren J, Haley CE, Khan MR, Svennerholm A, Stoll BJ, Hossain KMB, Black RE, Yunus M, Barua D (1985): Predisposition for cholera of individuals with O blood group: possible evolutionary significance. Am. J. Epidemiol. 121(6):791-796

Glavanakov S, White DJ, Caraco T, Lapenis A, Robinson GR, Szymanski BK, Maniatty WA (2001): Lyme disease in New York State: Spatial pattern at a regional scale. Am. J. Trop. Med. Hyg. 65(5):538-545

Goodchild M (2001): Models of scale and scales of modeling. In: Tate N, Atkinson P (Eds) Modeling scale in geographical information science. John Wiley & Sons, New York, pp 3-10

Goodgame RW, Greenough B (1975): Cholera in Africa: a message for the West. Ann. Intern. Med. 82(1):101-06

Goovaerts P (2005): Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. Int. J. Health. Geogr. 4:31

Green C, Hoppa RD, Young TK, Blanchard JF (2003): Geographical analysis of diabetes prevalence in an urban area. Soc. Sci. Med. 57(3):551-560

Greenberg B (1973): Flies & Disease. II. Biology and Disease Transmission. Princeton University Press, Princeton, pp 15

Griffith DA (1983): The boundary value problem in spatial statistical analysis. J. Reg. Sci. 23(3):377-387

Griffith DA (1985): An evaluation of correction techniques for boundary effects in spatial statistical analysis: Contemporary methods. Geogr. Anal. 17(1):81-88

Griffith DA, Amrhein CG (1983): An evaluation of correction techniques for boundary effects in spatial statistical analysis: Traditional methods. Geogr. Anal. 15:352-360

Griffith DC, Kelly-Hope LA, Miller MA (2006): Review of reported cholera outbreaks worldwide, 1995-2005. Am. J. Trop. Med. Hyg. 75(5):973-977

Grimson RC,Wang KC, Johnson PWC (1981): Searching for hierarchical clusters of disease: spatial patterns of sudden infant death syndrome. Soc. Sci. Med. D 15(2):287-293

Gunnlaugsson G, Einarsdottir J, Angulo FJ, Mentambanar SA, Passa A, Tauxe RV (1998): Funerals during the 1994 cholera epidemic in Guinea-Bissau, West Africa: the need for disinfection of bodies of persons dying of cholera. Epidemiol. Infect. 120(1):7-15

Haining R (1996): Designing of health needs GIS with spatial analysis capability. In: M Fischer, HJ Scholten, D Unwin (Eds) Spatial Analytical Perspective on GIS, GISDATA Series 4. Taylor & Francis, London

Haining RP (***1990***): ***Spatial*** Data Analysis in the Social and Environmental Sciences. Cambridge University Press, Cambridge, UK

Hansen MH, Kooperberg C (2002): Spline adaptation in extended linear models (with discussion and rejoinder by the authors). Stat. Sci. 17(1):2-51

Harmer DH, Cash RA (1999): Secretory Diarrheas: Cholera and Enterotoxigenic Escherichia Coli. In: Donald A, Cohen Jona(Eds) Infectious Diseases. Harcourt Publishers Ltd, Mosby

Hartley DM, Morris M, Smith DL (2005): Hyperinfectivity: A Critical Element in the Ability of *V. Cholerae* to Cause Epidemics? PLoS Med. 3(1): e7

Haslett J (1999): A simple derivation of deletion diagnostic results for the general linear model with correlated errors. J. R. Stat. Soc. B 61(3):603-609

Hastie T, Tibshirani R (1990): Generalized Additive Models. Chapman and Hall, London

Hastie T, Tibshirani R (2000): Bayesian back fitting (with comments and rejoinder by authors). Stat. Sci. 15(3):196-223

Hjalmars U, Kullforff M, Gustafsson G, Nagarwalla N (996): Childhood leukemia in Sweden: Using GIS and a spatial scan statistics for cluster detection. Stat. Med. 15(7-9):707-715

Holt D, Steel DG, Tranmer M, Wrigley N (1996): Aggregation and ecological effects in geographically based data. Geogr. Anal. 28(3):244-261

Hornick RB, Music SI, Wensel R, Cash R, Libonati JP, Snyder MJ, Woodward TE (1971): The Broad Street pump revisited; response of volunteers to ingested cholera *vibrios*. Bull. NY Acad. Med. 47(10):1181-1191

Hui C (2009): On the scaling patterns of species spatial distribution and association. J. Theor. Biol. 261(2):481-487

Hunt L, Boots B (1996): MAUP effects in the principal axis factoring techniques. Geogr. Syst. 3:101-121

Huq A, Colwell RR (1996): A microbiological paradox: viable but nonculturable bacteria with special reference to *vibrio cholerae.* J. Food Prot. 59(1):96-101

Huq A, Sack RB, Nizam A, Longini IM, Nair GB, Ali A, Morris JG, Khan MNH, Siddique AK, Yunus M, Albert MJ, Sack DA, Colwell RR (2005): Critical factors influencing the occurrence of *vibrio cholerae* in an environment of Bangladesh. Appl. Environ. Microbiol. 71(8):4645-4654

Huq A, Small EB, West PA, Huq MI, Rahman R, Colwell RR (1983): Ecologic relationships between *Vibrio cholerae* and Planktonic crustacean copepods. Appl. Environ. Microbiol. 45(1):275-283

Hutin Y, Luby S, Paquet C (2003): A large cholera outbreak in Kano City, Nigeria: importance of hand washing with soap and the danger of street-vended water. J. Water Health 1(1):45-52

Isaaks EH, Srivastava RM (1989): An Introduction to Applied Geostatistics. Oxford University Press, New York, pp 50-62

Islam MS (1990): Increased toxin production by *V. Cholerac* O1 during survival with a green algae, Rhizoclonium fintanam, in an artificial aquatic environment. Microbiol. Immun 34:557-563

Islam MS, Drasar BS, Bradley DJ (1990): Long-term persistence of toxigenic Vibrio O1 in the mucilaginous sheath of blue-green alga, Anabaena variabilis. J. Trop. Med. Hyg. 93(2):133-139

Islam MS, Drasar BS, Sack RB (1994): Probable Role of Blue-Green-Algae in Maintaining Endemicity and Seasonality of Cholera in Bangladesh-a Hypothesis. J. Diarrhoeal Dis. Res. 12(4):245-256

Islam MS, Draser BS, Bradley DJ (1989): Attachment of toxigenic *Vibrio cholerae* O1 to various freshwater plants and survival with filamentous green algae *Rhizoclonium fontanum*. J. Trop. Med. Hyg. 92(6):396-401

Islam MS, Rahim Z, Alam MJ, Begum S, Moniruzzaman SM, Umeda A, Amakao K, Albert MJ, Sack RB, Colwell RR (1999): Association of *Vibrio cholerae* O1 with the cyanobacterium, *Anabaena* sp., elucidated by polymerase chain reaction and transmission electron microscopy. Trans. R. Soc. Trop. Med. Hyg. 93(1):36-40

ISO (International Standardization Organization) (2002): ISO 19113:2002 Geographic Information-Quality principles, pp 29

Jarup L (2004): Health and Environment Information Systems for Exposure and Disease Mapping and Risk Assessment. Environ. Health Perspect. 112(9):995-997

Jenks G (1977): Optimal Data Classification for Choropleth Maps. Occasional paper No. 2, Department of Geography, University of Kansas

Kafadar K (1996): Smoothing geographical data, particularly rates of disease. Stat. Med. 15(23):2539-2560

Kamman EE, Wand MP (2003): Geoadditive Models. J. R. Stat. Soc. C 52(1):1-18

Kelly L (2001): The global dimension of cholera. Glob. Chang. Hum. Health 2(1):6-17

Kelsall J, Wakefield J (1999): Discussion of "Bayesian models for spatially correlated disease and exposure data". In: Best NG, Arnold RA, Thomas A, Conlon E, Waller LA, Bernado JM, Berger JO, Dawid AP, Smith AFM (Eds) Bayesian Statistics 6. Oxford University Press, Oxford, pp 151

King AA, Ionides EL, Pascual M, Bouma MJ (2008): Inapparent infections and cholera dynamics. Nature 454: 877-880

Kitron, U, Kazmierczak JJ (1997): Spatial analysis of the distribution of Lyme disease in Wisconson. Am. J. Epidemiol. 145(6):558-566

Kneib T (2005): Mixed model based inference in structured additive regression. PhD dissertation, Universität München

Knox EG (1964): The detection of space-time interaction. J. R. Stat. Soc. C 13(1):25-20

Knox EG (1989): Detection of clusters. In: Elliott P (Ed) Methodologies of Enquiry into Disease Clustering. Small Area Health Statistics Unit London, pp 17-22

Kobayashi M, Sasaki T, Saito N, Tamura K, Suzuki K, Watanabe H, Agui N (1999): Houseflies: not simple mechanical vectors of enterohemorrhagic Escherichia coli 0157:H7. Am. J. Trop. Med. Hyg. 61(4):625-629

Koch R (1884): An address on cholera and its bacillus. Br. Med. J. 2(1235):453-59

Koelle K, Pascual M (2004): Disentangling extrinsic from intrinsic factors in disease dynamics: A nonlinear time series approach with an application to cholera. The Am. Nat. 163(6):901-913

Koelle K, Rodo X, Pascual M, Yunus M, Mostafa G (2005): Refractory periods to climate forcing in cholera dynamics. Nature 436:696-700

Kresse W, Fadaie K (2004): ISO standards for geographic information. Springer, Berlin, pp 322

Kulldorff M (1997): A spatial scan statistic. Commun. Stat.-Theory Methods 269(6):1481-1496

Kulldorff M (2001): Prospective time-periodic geographical disease surveillance using a scan statistic. J. R. Stat. Soc. A 164:61-72

Kulldorff M (2005): Software for the spatial space time statistics, SaTScan v6.0. Information Management Service Inc. http://www.satscan.org/. Last accessed 4 November 2010

Kulldorff M (2006): SaTScan users guide for version 6.0. http://www.satscan.org/. Last accessed 4 November 2010

Kulldorff M, Athas W, Feuer E, Miller B, Key C (1998): Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. Am. J. Public Health 88(9):1377-1380

Kulldorff M, Feuer EJ, Miller BA, Freedman LS (1997): Breast Cancer clustering in the northeast United State, a geographic approach. Am. J. Epidemiol. 146(2):161-170

Kulldorff M, Nagarwalla N (1995): Spatial disease clusters: detection and inference. Stat. Med. 14(8):799-810

Kuo C-L, Fukui H (2007): Geographical structures and the cholera epidemic in modern Japan: Fukushima prefecture in 1882 and 1895. Int. J. Health Geogr. 6:25

Küstner HGV, Gibson IHN, Carmichael TR, Van Zyl L, Chouler CA, Hyde JP, du Plessis JN (1981): The spread of cholera in South Africa. S. Afr. Med. J. 60(3):87-90

Kwofie KM (1976): A spatio-temporal analysis of cholera diffusion in Western Africa. Econ. Geogr. 52(2):127-135

Lang S, Brezger A (2004): Bayesian P-splines. J. Comput. Graph. Stat. 13(1):183-212

Laurini R, Thompson D (1992): Fundamentals of spatial information systems. Academic Press, London, pp 680

Lawson AB (2001a): Statistical Methods in Spatial Epidemiology. John Wiley & Sons, Chichester

Lawson AB (2001b): Disease map reconstruction. Stat. Med. 20(14):2183-2204

Lawson AB (2006): Statistical Methods in Spatial Epidemiology. 2nd ed. John Wiley & Sons, New York

Lawson AB (2008): Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Chapman and Hall/CRC, Boca Raton

Lawson AB, Biggeri A, Bohning, Lesaffre E, Viel J-F, Bertollini R (1999b): Introduction to spatial models in ecological analysis Disease. In: Lawson AB, Biggeri A, Bohning, Lesaffre E, Viel J-F, Bertollini R (Eds) Disease Mapping and Risk Assessment for Public Health. John Wiley & Sons, Chichester, pp 181-191

Lawson AB, Biggeri A, Dreassi E (1999a): Edge effects in disease mapping. In Lawson A, Biggeri A, Bohning, Lesaffre E, Viel J-F, Bertollini R (Eds) Disease Mapping and Risk Assessment for Public Health. John Wiley & Sons, Chichester, pp 85-98

Lawson AB, Browne, WJ, Vidal Rodeiro CL (2003): Disease Mapping with WinBUGS and MLwiN. Wiley and Sons, Chichester

Lawson AB, Denison DGT (2002): Spatial cluster modeling. Chapman & Hall/CRC, London

Levine OS, Levine MM (1990): Houseflies (*Musca domestica*) as mechanical vectors of shigellosis. Rev. Infect. Dis. 13(4): 688- 696

Leyk S (2005): Computing the past-Utilizing Historical Data Sources for Map-Based Retrospective landscape Research., PhD dissertation, University of Zurich

Liebhold AM, Rossi RE, Kemp WP (1993): Geostatistics and geographic information systems in applied insect ecology. Annu. Rev. Entomol. 38:303-327

Lipp EK, Huq A, Colwell RR (2002): Effects of Global Climate on Infectious Disease: the Cholera Model. Clin. Microbiol. Rev. 15 (4):757-770

Lipp EK, Rodriguez-Palacios C, Rose JB (2001): Occurrence and distribution of the human pathogen Vibrio vulnificus in a subtropical Gulf of Mexico estuary. Hydrobiologia 460:165-173

Lobitz B, Beck L, Huq A, Wood B, Fuchs G, Faruque ASG, RR Colwell (2000) Climate and infectious disease: Use of remote sensing for detection of Vibrio cholerae by indirect measurement. Proc. Natl. Aced. Sci. USA 97(4):1438-1443

Louie MM, Kolaczyk ED (2006): A multiscale method for disease mapping in spatial epidemiology. Stat. Med. 25(8):1287-1308

Louis St ME, Porter JD, Helal A, Drame K, Hargrett-Bean N, Wells JG, Tauxe RV (1990): Epidemic cholera in West Africa: the role of food handling and high-risk foods. Am. J. Epidemiol. 131(4):719-728

Mahalanabis D, Molla A, Sack D (1992): Clinical Management of Cholera.In D. Barua and W. Greenough, III (Eds) Cholera. Plenum Medical Book Company, New York, pp 253-283

Malanson G, Armstrong MP (1997): Issues in spatial representation: Effects of number of cells and between-cell step size on models of environmental processes. Geogr. Environ. Modell. 1:47-64

Maling DH (1989): Measurements from maps. Pergamon Press, New York, pp 577

Manley D, Flowerdew R, Steel D (2006): Scales, levels and processes: Studying spatial patterns of British census variables. Comput. Environ.Urban Syst. 30(2):143-160

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

Manton K, Woodbury M, Stallard E (1981): A variance components approach to categorical data models with heterogenous mortality rates in North Carolina counties. Biometrics 37(2):259- 269

Martin D (2003): Extending the automated zoning procedure to reconcile incompatible zoning systems. Int. J. Geogr. Inf. Sci. 17(2):181-196

Marx B, Eilers PHC (1998): Direct generalized additive modeling with penalized likelihood. Comput. Stat. Data Anal. 28(2):193-209

Matheron G (1963): Principles of geostatistics. Econ. Geol 58(8):1246-1266

Matheron G (1965): Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature. Masson, Paris

Matisziw TC, Grubesic TH, Wei H (2008): Downscaling spatial structure for the analysis of epidemiological data. Comput. Environ. Urban Syst. 32(1):81-93

McCarthy HH, Hook JC, Knos DS (1956): The Measurement of Association in Industrial Geography. Department of Geography, State University of Iowa, Iowa City

McCullagh P, Nelder JA (1989): Generalized Linear Models. Chapman and Hall, London

McCulloch CE, Searle SR (2001): Generalized, Linear, and Mixed Models. John Wiley & Sons, New York

McNeill WH (1976): Plagues and People. Anchor Press Doubleday, New York

Meade M, Earickson R (2000): Medical geography. The Guilford Press, New York

Mendelsohn J, Dawson T (2007): Climate and cholera in KwaZulu-Natal, South Africa: The role of environmental factors and implications for epidemic preparedness. Int. J. Hyg. Environ. Health 211:156-162

Merrell DS, Butler SM, Qadri F, Dolganov NA, Alam A, Cohen MB, Calderwood SB, Schoolnik GK, Camilli A (2002): Host-induced epidemic spread of the cholera bacterium. Nature 417(6889):642-645

Michelozzi P, Capon A, Kirchmayer U, Forastiere F, Biggeri A, Barca A, Perucci CA (2002): Adult and Childhood leukemia near a high power radio station in Rome, Italy. Am. J. Epidemiol. 155(12):1096-1103

Miller CJ, Feachem RG, Drasar BS (1985): Cholera epidemiology in developed and developing countries: new thoughts on transmission, seasonality, and control. Lancet 1(8423):261-263

Miller JR (1998): Spatial aggregation and regional economic forecasting. Ann. Reg. Sci. 32(2):253-266

Molenaar M (1998): An introduction to the theory of spatial object modelling for GIS. Taylor & Francis, London, pp 246

Monestiez P, Dubroca L, Bonnin E, Durbec JP, Guinet C (2005): Comparison of model based geostatistical methods in ecology: application to fin whale spatial distribution in northwestern Mediterranean Sea. In: Leuangthong O, Deutsch CV. Dordrecht (Eds) Geostatistics Banff 2004 Volume 2. Kluwer Academic Publishers, The Netherlands pp 777-786

Monestiez P, Dubroca L, Bonnin E, Durbec JP, Guinet C (2006): Geostatistical modeling of spatial distribution of Balenoptera physalus in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. Ecol. Model. 193(3-4):615-628

Moore DA, Carpenter TE (1999): Spatial Analytical Methods and Geographic Information Systems: Use in Health Research and Epidemiology. Epidemiol. Rev. 21(2):143-161

Moran PAP (1950): Notes on continuous stochastic phenomena. Biometrika 37(1-2):17-23

Mosley WH, Khan MU (1979): Cholera epidemiology, some environmental aspects. Prog. Water Technol. 11:309-316

Mouriño-Pérez RR (1998): Oceanography and the seventh cholera pandemic. Epidemiology 9 (3):355-357

Mugoya I, Kariuki S, Galgalo T, Njuguna C, Omollo J, Njoroge J, Kalani R, Nzioka C, Tetteh C, Bedno S, Breiman RF, Feikin DR (2008): Rapid Spread of Vibrio cholerae O1 Throughout Kenya, 2005. Am. J. Trop. Med. Hyg. 78(3):527-533

Myaux J, Ali M, Chakraborty J, De Francisco A (1997): Flood control embankments contribute to the improvements of health status of children in rural Bangladesh. Bull. World Health Organ. 75(6):533-539

Nair GB, Oku Y, Takeda Y, Ghosh A, Ghosh RK, Chattopadhyay S, Pal SC, Kaper JB, Takeda T (1988): Toxin profiles of Vibrio cholerae non-O1 from environmental sources in Calcutta, India. Appl. Environ. Microbiol. 54(12):3180-3182

Nalin DR, Daya V, Reid A, Levine MM, Cisneros L (1979): Adsorption and growth of *Vibrio cholerae* on chitin. Infect. Immun. 25:768-770

Nelder JA, Wedderburn RWM (1972): Generalized linear models. J. R. Stat. Soc. A 135(3): 370-384

Nichols GL (2005): Fly Transmission of Campylobacter. Emerg. Infect. Dis. 11(3):361-364

Nodtvedt A, Guitian J, Egenvall A, Emanuelson U, Pfeiffer DU (2007): The spatial distribution of atopic dermatitis cases in a population of insured Swedish dogs. Prev. Vet. Med. 78(3-4):210-222

Nuckols JR, Ward MH, Jarup L (2004): Using Geographic Information Systems for Exposure Assessment in Environmental Epidemiology Studies. Envron. Health Perspect. 112 (9):1007-1015

Obiri-Danso K, Okore-Hanson A, Jones K (2003): The microbiological quality of drinking water sold on the streets in Kumasi, Ghana. Lett Appl Microb 37: 334-335.

Obiri-Danso K, Weobong CAA, Jones K (2005): Aspects of health related microbiology of the Subin, an urban river in Kumasi, Ghana. J. Water Health 3(1):69-76

Odoi A, Martin SW, Michel P, Holt J, Middleton D, Wilson J (2003): Geographical and temporal distribution of human giardiasis in Ontario, Canada. Int. J. Health Geogr. 2:5

Odoi A, Martin SW, Michel P, Middleton D, Holt J, Wilson J (2004): Investigation of clusters of giardiasis using GIS and spatial scan statistics. Int. J. Health Geogr. 3:11

Ogata K, Murata M, Furuno A, Uchida S, Sasa M (1961): Detection of the poliomyelitis viruses from flies in an epidemic area in Hokkaido, Japan (in Japanese with an English summary). Jpn. J. Sanit. Zool. 12:165-168

Okabe A Tagashira N (1996): Spatial aggregation bias in a regression model containing a distance variable. Geogr. Syst. 3:77-99

Oliver MA, Webster R, Lajuanie C, Muir KR, Parkes SE, Cameron AH, Stevens MCG, Mann JR: Binomial cokriging for estimating and mapping the risk of childhood cancer (1998): Math. Med. Biol. 15 (3):279-297

Openshaw S (1984): The modifiable areal unit problem. Geo Books, Norwich, United Kingdom

Openshaw S, Charlton M, Wymer C, Craft A (1987): A mark I geographical analysis machine for the automated analysis of point pattern data. Int. J. Geogr. Inf. Syst. 1(4):335-350

Openshaw S, Taylor PJ (1979): A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N (Ed) Statistical applications in the spatial sciences. Pion, London, pp 127-144

Openshaw S, Taylor PJ (1981): The modifiable areal unit problem. In RJ Bennett, N Wrigley (Eds) Quantitative geography. Routledge and Kegan Paul, London, pp 60-69

Opintan JA, Newman MJ, Nsiah-Poodoh OA, Okeke IN (2008): Vibrio cholerae O1 from Accra, Ghana carrying class 2 intergron and the SXT element. J. Antimicrob. Chemother. 62(5):929-933

OPS: Organization Pan-American de la Salud (1997): Situacion del Colera en Las Americans. Report number 16, Washington DC

Ord JK (1975): Estimation methods for models of spatial interaction. J. Am. Stat. Ass. 70(349):120-126

Osei FB and Duker AA (2008a): Spatial and demographic patterns of Cholera in Ashanti Region-Ghana. Int. J. Health Geogr. 7:44

Osei FB and Duker AA (2008b): Spatial dependency of *V. cholerae* prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modeling. Int. J. Health Geogr. 7:62

Osei FB, Duker AA, Augustijn E-W, Stein A (2010): Spatial dependency of cholera prevalence on potential cholera reservoirs in an urban area, Kumasi, Ghana. Int. J. Appl. Earth Obs. Geoinf. 12(5):331-339

Pascual M, Bouma MJ, Dobson AP (2002): Cholera and climate: revisiting the quantitative evidence. Microbes. Infect. 4(2):237-45

Pascual M, Koelle K, Dobson AP (2006): Hyperinfectivity in Cholera: A New Mechanism for an Old Epidemiological Model? PLoS Med. 3(6): e280

Pascual M, Rodo X, Ellner SP, Colwell R, Bouma MJ (2000): Cholera dynamics and the El Nino-southern oscillation. Science 289 (5485):1766-1769

Perez, A, Ward MP, Torres P, Ritacco V (2002): Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina. Prev. Ve.t Med. 56(1):63-74

PHC: Population and Housing Census (2000): Ghana statistical service

Pobee JOM, Grant F (1970): Case Report of Cholera. Ghana Med. J. 306-309

Pollitzer R (1959): Cholera. WHO, Geneva

Prestero T, Height M, Hwang R (2001): Rapid Cholera Treatment: Exploring Alternative IV Treatment Devices. In: Proceedings for 1st Development by Design Conference, Cambridge, Massachusetts

Pyle GF (1969): The diffusion of cholera in the United States in the Nineteenth Century. Geogr. Anal. 1:59-79

Ramamurthy T, Garg S, Sharma R, Bhattacharya SK, Nair GB, Shimada T, Takeda T, Karasawa T, Kurazano H, Pal A, Takeda Y (1993): Emergence of novel strain of *Vibrio cholerae* with epidemic potential in southern and eastern India. Lancet 341(8846):703-04

Raper JF (1999): Spatial representation: the scientist's perspective. In: Longley PA, Goodchild MF, Maguire DJ, Rhind DW (Eds) Geographical Information Systems: Principles and technical issues. 2nd ed. John Wiley & Sons, New York, pp 61-70

Rawlings TK, Ruiz GM, Colwell RR (2007): Association of *Vibrio cholerae* O1 El Tor and O139 Bengal with the copepods *Acartia tonsa* and *Eurytemora affinis.* Appl. Environ. Microbiol. 73:7926-7933

Reller ME, Mong YJM, Hoekstra RM, Quick RE (2001): Cholera prevention with traditional and novel water treatment methods: an outbreak investigation in Fort-Dauphin, Madagascar. Am. J. Public Health 91(10):1608-1610

Rezaeian M, Dunn G, Leger St S, Appleby L (2007): Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. J. Epidemiol. Community Health 61(2):98-102

Ripley B (1981): Spatial Statistics. John Wiley & Sons, Chichester

Ripley B (1988): Statistical Inference for Spatial Processes. Cambridge University Press, Cambridge

Ripley BD (1976): The second-order analysis of stationary point processes. J. Appl. Prob. 13(2):255-266

Robinson WS (1950): Ecological correlations and the behavior of individuals. Am. Sociol. Rev. 15(3):351-357

Roger DJ, Williams BG (1993): Monitoring trypanosomiasis in space and time. Parasitology 106:S77-S92

Root G (1997): Population density and spatial differentials in child mortality in Zimbabwe. Soc. Sci. Med. 44 (3):413-421

Rosenberg CE (1962): The cholera years, the United States in 1832, 1849, and 1866. University of Chicago Press, Chicago

Rosenberg MS, Sokal RR, Oden NL, DiGiovann D (1999): Spatial autocorrelation of cancer in Western Europe. Eur. J. Epidemiol. 15(1):15-22

Rossi RE, Mulla DJ, Journel AG, Franz EH (1992): Geostatistical tools for modeling and interpreting ecological spatial dependence. Ecol. Monogr. 62(2):277-314

Rue H (2001): Fast sampling of Gaussian Markov random fields with applications. J. R. Stat. Soc. B 63(2):325-338

Rue H, Held L (2005): Gaussian Markov Random Fields: Theory and Applications. Chapman and Hall, Boca Raton

Ruiz-Moreno D (2009): The role of primary and secondary transmission on the dynamics of cholera in endemic areas. PhD dissertation, The University of Michigan

Ruiz-Moreno D, Pascual M, Bouma M, Dobson A, Cash B (2007): Cholera Seasonality in Madras (1901-1940): Dual Role for Rainfall in Endemic and Epidemic Regions. EcoHealth 4(1):52-62

Ruppert D, Wand M, Carroll R (2003): Semiparametric Regression. Cambridge University Press, Cambridge

Rushton G (2003): Public health, GIS, and spatial analytic tools. Annu. Rev. Public Health 24:43-56

Rushton G, Lolonis P (1996): Exploratory spatial analysis of birth defect rates in an urban population. Stat. Med. 15(7-9):717-726

Sabel CE, Boyle PJ, Loytonen M, Gatrell AC, Jokelainen M (2003): Spatial clustering of amyotrophic lateral sclerosis in Finland at place of brith and place of death. Am. J. Epidemiol. 157(10): 898-905

Sack DA, Sack RB, Nair GB, Siddique AK (2004): Cholera. Lancet 363(9404):223-233

Sack DA, Tacket CO, Cohen MB, Sack RB, Losonsky GA, Shimko J, Nataro JP, Edelman R, Levine MM, Giannella RA, Schiff G, Lang D (1998): Validation of a volunteer model of cholera with frozen bacteria as the challenge. Infect. Immun. 66(5):1968-1972

Sack RB, Siddique AK, Longini IM Jr, Nizam A, Yunus M, Islam MS, Morris JG Jr, Ali A, Huq A, Nair GB, Qadri F, Faruque SM, Sack DA, Colwell RR (2003): A 4-year study of the epidemiology of *Vibrio cholerae* in four rural areas of Bangladesh. J. Infect. Dis. 187(1):96-101

Said MD (2006): Epidemic cholera in KwaZulu-Natal: the role of the natural and social environment. PhD dissertation, University of Pretoria

Sasaki S, Suzuki H, Igarashi K, Tambatamba B, Mulenga P (2008): Spatial Analysis of Risk Factor of Cholera Outbreak for 2003-2004 in a Peri-urban Area of Lusaka, Zambia. Am. J. Trop. Med. Hyg. 79(3):414-421

Schlesselman JJ (1982): Case-Control Studies: Design, Conduct, Analysis. Oxford University Press, New York

Servais Pierre, Tamara Garcia-Armisen, Isabelle George, and Gilles Billen (2007): Fecal bacteria in the rivers of the Seine drainage network (France): Sources, fate and modeling. Sci. Total Environ. 375(1-3):152-167

Shapiro RL, Muga RO, Adcock PM, Phillips-Howard PA, Hawley WA, Waiyaki P, Nahlen BL, Slutsker L (1999): Transmission of epidemic *Vibrio cholerae* O1 in rural western Kenya associated with drinking water from Lake Victoria: an environmental reservoir for cholera? Am. J. Trop. Med. Hyg. 60(2):271-276

Sheehan TJ, DeChelo LM (2005): A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997. Int. J. Health Georgr. 4:15

Siddique AK, Salam A, Islam MS, Akram K, Majumdar RN, Zaman K, Fronczak N, Laston S (1995): Why treatment centres failed to prevent cholera deaths among Rwandan refugees in Goma, Zaire. Lancet 345(8946):359-361

Siddique AK, Zaman K, Baqui AH, Akram KA, Mutsuddy P, Eusof A, Haider K, Islam S, Sack RB (1992): Cholera epidemics in Bangladesh:1985-1991. J. Diarrhoeal Dis. Res. 10(2):79-86

Silverman BW (1986): Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York, pp 76

Sinclair GS, Mphalele M, Duvenhage H, Nichol R, Whitehorn A, Kustner HG (1982): Determination of the mode of transmission of cholera in Lebowa. An epidemiological investigation. S. Afr. Med. J. 62(21):753-755

Singleton FL, Attwell RW, Jangi MS, and Colwell RR (1982a): Influence of salinity and organic nutrient concentration on survival and growth of Vibrio cholerae in aquatic microcosms. Appl. Environ. Microbiol. 43(5):1080-1085

Singleton FL, Attwell RW, Jangi MS, and Colwell RR (1982b): Effects of temperature and salinity on *Vibrio cholerae* growth. Appl. Environ. Microbiol. 44(5):1047-1058

Smallman-Raynor M, Cliff A (1998a): The Philippines insurrection and the 1902-4 cholera epidemic: Part 1- Epidemiological diffusion process in war. J. Hist. Geogr. 24(1):69-89

Smallman-Raynor M, Cliff A (1998b): The Philippines insurrection and the 1902-4 cholera epidemic: Part 2- Diffusion patterns in war and peace. J. Hist. Geogr. 24(2):188-210

Smith M, Kohn R (1996): Nonparametric regression using Bayesian variable selection. J. Econom. 75(2):317-343

Smith M, Kohn R (1997): A Bayesian Approach to Nonparametric Bivariate Regression. J. Am. Stat. Ass. 92(440):1522-1535

Snow J (1855): On the Mode of Communication of Cholera. 2nd ed. John Churchill, London

Spiegelhalter DJ, Best NG, and Carlin BP, van der Linde A (2002): Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. B 64(4):583-640

Stock RF (1976): Cholera in Africa: Diffusion of the Disease 1970-1975 with particular emphasis on West Africa. International African Institute, London

Stone CJ, Hansen MH, Kooperberg C, Truong YK (1997): Polynomial splines and their tensor products in extended linear modeling (with discussion). Ann. Stat. 25:1371-1470

Strahler AN (1952): Hypsometric (area-altitude) analysis of erosional topography. Geol. Soc. Am. Bull. 63(11):1117-1142

Sulaiman S, Sohadi AR, Yunus H, Iberahim R (1988): The role of some cyclorrhaphan flies as carriers of human helminthes in Malaysia. Med. Vet. Entomol. 2(1):1-6

Sur D, Deen J, Manna B, Niyogi S, Deb A, Kanungo S, Sarkar B, Kim D, Danovaro-Holliday M, Holliday K, Gupta V, Ali M, von Seidlein L, Clemens J, Bhattacharya S (2005): The burden of cholera in the slums of Kolkata, India: data from a prospective, community based study. Arch. Dis. Child 90(11): 1175-1181

Swerdlow DL, Greene KD, Tauxe RV, Wells JG, Bean NH, Ries AA, Blake PA, Mintz ED, Pollack M, Rodriguez M, Tejada E, Seminario L, Ocampo C, Vertiz B, Espejo

L, Saldana W (1992): Waterborne transmission of epidemic cholera in Trujillo, Peru: lessons for a continent at risk. Lancet 340 (8810):28-32

Swerdlow DL, Malenga G, Begkoyian G, Nyangulu D, Toole M, Waldman RJ, Puhr DN, Tauxe RV (1997): Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. Epidemiol. Infect. 118(3):207-214

Swift A, Liu L, Uber J (2008): Reducing MAUP bias of correlation statistics between water quality and GI illness. Comput. Environ. Urban Syst. 32(2):134-148

Tagashira N, Okabe A (2002): The modifiable areal unit problem in a regression model whose independent variable is a distance from a predetermined point. Geogr. Anal. 34(1):1-19

Takahashi K, Tango T (2005): A flexibly shaped spatial scan statistic for detecting clusters. Int. J. Health Geogr. 4: 4-11

Takahashi K, Yokoyama T, Tango T (2004): FleXScan: Software for the flexible spatial scan statistic. National Institute of Public Health, Japan

Talbot TO, Kulldroff M, Teven PF, Haley VB (2000): Evaluation of spatial filters to create smoothed maps of health data. Sat. Med. 18(17-18):2399-2408

Tanamal S (1959): Notes on Paracholera in Sulawesi (Celebes). Am. J. Trop. Med. Hyg. 8 (1): 72-78

Tango T (1995): A class of tests for detecting "general" and "focused" clustering of rare diseases. Stat. Med. 14(21-22):2323-2334

Tango T (1999): Comparison of general tests for disease clustering. In: Lawson AB *et al* (Eds) Disease Mapping and Risk Assessment for Public Health. John Wiley & Sons, New York, pp 111-117

Tango T (2000): A test for spatial disease clustering adjusted for multiple testing. Stat. Med. 19(2):191-204

Tango T (2010): Statistical Methods for Disease Clustering. Springer, New York

Tauxe RV, Holberg SD, Dodin A, Wells JV, Blake PA (1988): Epidemic cholera in Mali: high mortality and multiple routes of transmission in a famine area. Epidemiol. Infect. 100(2): 279-289

Tiwari N, Adhikari CS, Tewari A, Kandpal V (2006): Investigation of geo-spatial hotspost for the occureence of tuberculosis in Almora district, India, using GIS and spatial scan statistic. Int. J. Health Goegr. 5:33

Trevelyan B, Smallman-Raynor M, Cliff AD (2005): The spatial structure of epidemic emergence: geographical aspects of poliomyelitis in north-eastern USA, July-October 1916. J. R. Stat. Soc. A 168(4):701-722

Tsutakawa R (1988): A mixed model for analyzing geographic variability in mortality rates. J. Am. Stat. Ass. 83(401):37-42

Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC (1990): Monitoring for clusters of disease: application to leukemia incidence in upstate New York. Am. J. Epidemiol. 132:S136-S143

Uitermark HTJA (2001): Ontology-based geographic data set integration. PhD dissertation, University of Twente

Upton GJG, B Fingleton (1985): Spatial Data Analysis by Example. Volume 1: Point Pattern and Quantitative Data. John Wiley & Sons, Chichester

van Meter EM, Lawson AB, Colabianchi N, Nichols M, Hibbert J, Porter DE, Liese AD (2010): An evaluation of edge effects in nutritional accessibility and availability measures: a simulation study. Int. J. Health. Geogr. 9:40

van Oort P (2005): Spatial data quality: from description to application. PhD dissertation, Wageningen University

Varga Ab (1998): University Research and Regional Innovation: A Spatial Econometric Analysis of Academic Technology Transfers. Kluwer Academic Publishers, Boston-Massachusetts

Vidal RCL, Lawson AB (2005): An evaluation of the edge effects in disease map modeling. Comput. Stat. Data Anal. 49(1):45-62

Viel JF, Arveux P, Baverel J, and Cahn JY (2000): Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with dioxin emission levels. Am. J. Epidemiol. 152(1):13-19

Vine MF, Degnan D, Hanchette C (1997): Geographic Information Systems: Their use in epidemiologic research. Environ. Health Perspect. 105 (6):598-605

Wahba G (1978): Improper Prior, Spline smoothing and the problem of guarding against model errors in regression. J. R. Stat. Soc. B 44(3):364-372

***Waller LA, Gotway CA* (*2004*)**: Applied Spatial Statistics for Public Health Data. John Wiley & Sons, New Jersey

Walter SD (1992a): The analysis of regional patterns in health data: I. Distributional considerations. Am. J. Epidemiol. 136(6):730-741

Walter SD (1992b): The analysis of regional patterns in health data: II. The power to detect environmental effects. Am. J. Epidemiol. 136(6):742-759

Walter SD (1993): Assessing spatial patterns in disease rates. Stat. Med. 12(19-20):1885-1894

Wartenberg D, Greenberg M, and Lathrop R (1993): Identification and characterization of populations living near high-voltage transmission lines: a pilot study. Environ. Health Perspect. 101(7):626-32

Watt J, Lindsay DR (1948): Diarrheal disease control studies: effect of fly control on high morbidity area. Public Health Rep. 63(41):1319-1334

Webster R, Oliver MA, Munir KR, Man JR (1994): Kriging the local risk of rare disease from a register of diagnoses. Geogr. Anal. 26(2):168-185

Weil O, Berche P (1992): The cholera epidemic in Ecuador: towards an endemic in Latin America. Rev. Epidemiol. Sante. Publique. 40(3):145-55

Weisz R, Fleischer S, Smilowitz Z (1995): Site-specific integrated pest management for high value crops: sample units for map generation using the Colorado potato beetle (Coleoptera: Chrysomelidae) as a model system. J. Econ. Entomol. 88(5):1069-1080

Whittemore AS, Friend N, Brown BW, Holly EA (1987): A test to detect clusters of disease. Biometrika 74(3):631-635

WHO (1993): Guidelines for Cholera Control. World Health Organization, Geneva. http://helid.desastres.net/en/d/Jwho90e/. Last accessed 5 November 2010

WHO (2000a): Report on global surveillance of epidemic prone infectious diseases. World Health Organization, Geneva. http://whqlibdoc.who.int/hq/2000/WHO_CDS_CSR_ISR_2000.1.pdf. Last accessed 5 November 2010

WHO (2000b): Cholera, 1999. Weekly epidemiological record 75(31):249-256. http://www.who.int/wer. Last accessed 5 November 2010

WHO (2001): Cholera, 2000. Weekly epidemiological record 76(31):233-240. http://www.who.int/wer. Last accessed 5 November 2010

WHO (2002): Cholera, 2001. Weekly epidemiological record 77(31):257-268. http://www.who.int/wer. Last accessed 5 November 2010

WHO (2003): Cholera, 2002. Weekly epidemiological record 78(31):269-276. http://www.who.int/wer. Last accessed 5 November 2010

WHO (2004): Cholera, 2003. Weekly epidemiological record 79(31):281-288. http://www.who.int/wer. Last accessed 5 November 2010

WHO (2005): Cholera, 2004. Weekly epidemiological record 80(31):261-268. http://www.who.int/wer. Last accessed 5 November 2010

WHO (2006): Cholera, 2005. Weekly epidemiological record 81(31):297-308. http://www.who.int/wer. Last accessed 5 November 2010

WHO (2010): Cholera vaccines: WHO position paper. Weekly epidemiological record 85(13):117-128. http://www.who.int/wer. Last accessed 5 November 2010

Xu HS (1982): Survival and viability of non-culturable Escherichia coli and Vibrio cholerae in the estuarine and marine environment. Microb. Ecol. 8(4):313-323

Yamai S, Okitsu T, Shimada T, Katsube Y (1997): Distribution of serogroups of *Vibrio cholerae* non-O1 non-O139 with specific reference to their ability to produce cholera toxin, and addition of novel serogroups. Kansenshogaku Zasshi 71(10):1037-1045

Zandbergen P, Chakraborty J (2006): Improving environmental exposure analysis using cumulative distribution functions and individual geocoding. Int. J. Health Geogr. 5:23

Curriculum vitae

Frank Badu Osei was born on the 24th of March 1980 at Kumasi in the Ashanti Region, Ghana. In 2004, he obtained his Bachelor of Science (Bsc.) degree in Geodetic Engineering Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana. He worked in the Department of Geodetic Engineering as a Teaching Assistant from 2004 until 2005. Later in 2005, he obtained admission from KNUST to pursue a PhD degree. In May 2009, he obtained funding from the Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, The Netherlands, to continue his PhD.

List of publications

Osei FB and Duker AA (2008): Spatial and demographic patterns of Cholera in Ashanti Region-Ghana. International Journal of Health Geographics 7:44

Osei FB and Duker AA (2008): Spatial dependency of *V. cholerae* prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modelling. International Journal of Health Geographics 7:62

Osei FB, Duker AA, Augustijn E-W, Stein A (2010): Spatial dependency of cholera prevalence on potential cholera reservoirs in an urban area, Kumasi, Ghana. International Journal of Applied Earth Observation and Geoinformation 12(5):331-339

Osei FB, Duker AA, Stein A (2010): Hierarchical Bayesian modelling of the space-time diffusion patterns of cholera epidemic in Kumasi, Ghana. Statistica Neerlandica (Accepted)

Submitted manuscripts

Osei FB, Duker AA, Stein A: Multivariate Bayesian semi-parametric modelling of cholera in an urban environment. Environmental and Ecological Statistics (In review)

Osei FB, Duker AA, Stein A: Spatial and space-time clustering of cholera in Ashanti Region, Ghana. African Journal of Environmental Science and Technology (Submitted)