

**Spatial Modeling, Covariate Measurement Error
and Design Issues in Environmental Epidemiology**

by

Md Hamidul Huque

*Submitted to the School of Mathematical and Physical Sciences,
Faculty of Science in partial fulfillment of the requirements for the
degree of*

Doctor of Philosophy

at the

UNIVERSITY OF TECHNOLOGY SYDNEY

September, 2016

©Md Hamidul Huque, MHH XVI. All right reserved.

Permission is herewith granted to UTS to circulate and to have copied for non-commercial purposes, as its discretion, the above title upon request of individuals and institution.

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text. I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student:

Date: 12/09/2016

Acknowledgements

The successful completion of this research work is not only the result of my own effort, but also a series of contribution from many others ranging from my friends and family to the staff of the school of Mathematical and Physical Sciences. My deepest gratitude goes to Professor Louise Ryan for her ultimate guidance and supervision throughout of my doctoral study period. She has always been inspirational, supportive and talk through issues in details, sometimes long past a reasonable meeting length. I would also like to thank Richard Walton from NSW Cancer Institute for his guidance and inputs in cancer data analysis. I had certainly enjoyed numerous through provoking discussions at the NSW Cancer Institute. I would also like to thanks all other co-authors and collaborators for their constructive feedback. I thank Professor Raymond Carroll for introducing me the semiparametric measurement error models and for his collaboration in two of our projects and Craig Anderson for editorial service to this thesis in addition to work on various projects as a coauthor. I thank the Director of Industry Doctoral Training Center for organizing various courses and conferences that helped me to proper management of this research project. I would also like to thank professor Matt Wand for offering course on variational message passing. I am also grateful to other PhD students for their generous support and discussion on various issues. Most importantly, I would like to thank my family members. especially my wife Jannatul Ferdous for sacrificing her time and taking care of our lovely daughter Mawizat ul Huque. I am eternally indebted to her.

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Background	4
1.3	Objectives of the Study	6
1.4	Literature Review	10
1.5	Organization of the Thesis	27
1.6	List of Publications Arising From this Thesis	27
2	Individual Level Covariate Adjusted Conditional Auto-Regressive (indiCAR) Model for Disease Mapping.	29
2.1	Introduction	30
2.2	Methods	33
2.2.1	Data	33
2.2.2	The Model	34
2.3	Simulation Studies	40
2.4	Results and Discussion	41
2.4.1	Simulation Results	41
2.4.2	Application to the Neutropenia Data	43
2.5	Conclusions	49
3	Smooth Individual Level Covariates in Conditional Auto-Regressive (smooth-indiCAR) Model for Disease Mapping.	59
3.1	Introduction	60
3.2	Methodology	62
3.2.1	Data	62
3.2.2	Statistical Model	63
3.3	Simulation Study	72

3.4	Results	73
3.4.1	Simulation Results	73
3.4.2	Application to the Neutropenia Data	75
3.5	Discussion	79
4	On the Impact of Covariate Measurement Error on Spatial Regression Modeling.	87
4.1	Introduction	88
4.2	Model Formulation	91
4.3	Asymptotic Bias Analysis	92
4.3.1	Generalized Least Squares	93
4.4	Bias Correction	95
4.4.1	Method I: Method of Moments	96
4.4.2	Method II: Transformation Method	97
4.5	Simulation Study	97
4.6	Analysis of Ischemic Heart Disease Data	101
4.7	Discussion	104
5	Spatial Regression with Covariate Measurement Error: A Semiparametric Approach	109
5.1	Introduction	110
5.2	Model	112
5.2.1	Identifiability	114
5.2.2	Parameter Estimation	114
5.2.3	Asymptotic Theory	115
5.2.4	Estimating the Standard Error of $\hat{\beta}_1$	116
5.2.5	Smoothing Parameter Selection	117
5.3	Simulation Study	118
5.3.1	Data Generation	118
5.3.2	Generating Bi-variate Splines Basis Functions	120
5.3.3	Simulation Results	120
5.4	Analysis of Ischemic Heart Disease Data	123
5.5	Discussion	124

6	Exposure enriched case-control (EECC) design for the assessment of gene-environment interaction	141
6.1	Introduction	142
6.2	Methods	143
6.3	Simulation Study	146
6.3.1	Data Generation	146
6.3.2	Parameter Estimation	147
6.3.3	Power Calculation	147
6.4	Simulation Results	147
6.4.1	Estimation of Parameters	148
6.4.2	Estimation of Power	149
6.5	Application to the Arsenic Data from Bangladesh	157
6.6	Discussion	159
7	Conclusion	163

List of Figures

2.1	Estimated (a) Standardized Incidence Ratios (SIR) and (b) spatial random effects of neutropenia admissions in NSW, Australia using indiCAR.	48
3.1	Fitted non linear curves based on the first 50 simulations under scenario (i) for different values of spatial dependence parameter. The solid line indicates the true curve.	75
3.2	Fitted non linear curves based on the first 50 simulations under scenario (ii) for different values of spatial dependence parameter. The solid line indicates the true curve.	77
3.3	The estimated effect of age on neutropenia admission rates with associated 95 % Bayesian (light gray region) and frequentist (dark region) confidence intervals.	78
3.4	Distributions of estimated (a) Standardized Incidence Ratios (SIR) and (b) spatial random effects of neutropenia admissions in NSW, Australia.	78
4.1	Attenuation factor associated with varying degree of measurement error. . .	94
4.2	Distribution of estimated coefficient when estimated and true value of σ_u^2 used with Method I (a-b) and Method II (c-d) under different range parameters combinations with true and misspecified covariate structure.	101
4.3	Sensitivity analysis for IHD data. The assumed measurement error variance varied between 0 (naive) and 0.40	104
5.1	Contour plots of covariates (\mathbf{X} and \mathbf{W}) with different specification of measurement error variance	119
6.1	Comparison of estimated coefficients obtained using usual logistic regression ignoring sampling and EECC.	148

6.2	Comparison of the estimated parameters for (a) Logistic regression analysis ignoring sampling and (b) EECC design when true interaction parameters are 0 (upper panel) and -0.406 (lower panel), respectively.	150
6.3	Comparison of Estimated parameters for (a) Logistic regression analysis ignoring sampling and (b) EECC design when cases are randomly selected and controls are selected by oversampling of low exposed subjects.	151
6.4	Comparison of estimated power to detect gene-interaction effect obtained via a traditional case-control design and EECC design for a sample size of 2000.	152
6.5	Power as a function of control-case ratio and sample sizes: (a) EECC design (b) Traditional case control design.	152
6.6	Power as a function of ratio of high and low exposed sample. Different values of low exposed samples result in a different trajectory for power. . .	153

List of Tables

2.1	Simulation results for estimated regression coefficients following indiCAR and method proposed by Leroux, Lei, and Breslow (1999) where each area consists of a random number of subjects between 10 and 1000.	42
2.2	Simulation results for estimated regression coefficients following indiCAR and method proposed by Leroux et al. (1999) where each area consists of a random number of subjects between 10 and 50.	44
2.3	Descriptive analysis of neutropenia data.	45
2.4	Comparison of individual covariate adjusted conditional auto-regressive model (indiCAR) with the age-sex adjusted Leroux et al. (1999) method.	47
2.5	Application of indiCAR with age as a continuous predictor	54
3.1	Estimated regression coefficients and variance parameters for the proposed smooth-indiCAR method using simulation scenario (i).	73
3.2	Estimated regression coefficients and variance parameters for the proposed smooth-indiCAR method using simulation scenario (ii).	74
3.3	Comparison of estimated regression coefficients and variance parameters of smooth-indiCAR with indiCAR using neutropenia data.	76
4.1	Simulation results using different combinations of range parameters. Reported numbers are averaged over 1000 simulations with 100 observations per simulation with measurement error variance 0.2.	99
4.2	Analysis of Ischemic Heart Disease Data in NSW, Australia under different specification of measurement error	102
5.1	Simulation results using different combinations of range parameters and measurement error variance. Reported numbers are averaged over 1000 simulations with 500 observations per simulation.	122

5.2	Simulation results using different combinations of range parameters and sample sizes. Reported numbers are averaged over 1000 simulations with measurement error variance 0.5.	123
5.3	Analysis of Ischemic Heart Disease Data in NSW, Australia under different specification of measurement error.	124
5.4	Simulation results with varying number of knots (q_2) for covariate model in our proposed method. In each case, the number of knots for the residual model (q_1) were fixed at 125. Reported numbers are averaged over 1000 simulations. Data were simulated with sample size 500, regression coefficient 2, measurement error variance 0.25 and varying range parameters for spatial correlations.	135
6.1	Comparison of estimated regression coefficients and power for the asymmetry of the exposure distribution and varying cut offs with a sample size of 1600.	155
6.2	Comparison of estimated coefficients their standard errors, power of traditional case-control design vs modified case-control design for various sample size and exposure distribution. The various cut off ensures 10% of the total exposure data lies above these cut offs.	156
6.3	Comparison of estimated coefficients their standard errors, power of traditional case-control design vs modified case-control design for various sample size and exposure distribution. The various cut off ensures 10% of the total exposure data lies above these cut offs.	158

Abstract

In this thesis we develop methods to resolve a series of problems motivated by the analysis of administrative data to help explain geographical variation in disease rates. The Conditional auto-regressive (CAR) structure within a hierarchical generalized linear model offers a robust, flexible, and popular class of models for the exploration and analysis of geographical variation across small areas. However, lack of modeling strategies for individual level covariate data is a limitation of the existing methodology. We propose an individual level covariate adjusted conditional auto-regressive (indiCAR) model to incorporate both individual and area level covariates while adjusting for spatial correlation in disease rates. We also extend the indiCAR method to a semiparametric mixed model framework that allows adjustment for smooth covariate effects (smooth-indiCAR). We illustrate the applicability of both methods in a distributed computing framework that enhances its application in the Big Data domain with a large number of individual/group level covariates involved. We evaluate the performance of indiCAR and smooth-indiCAR through simulation studies. Our results indicate that both methods provide reliable estimates of all the regression and random effect parameters. The estimated regression coefficient based on the CAR modeling, however, appears to be sensitive to the assumed spatial correlation structure. We hypothesize that such sensitivity is especially likely to occur when the covariate of interest has been measured with error. We quantify the biases of covariate measurement error, showing that the amount of attenuation depends on the degree of spatial correlation in both the covariate of interest and the assumed random error from the regression model. These results explain why the estimates obtained from spatial regression modeling are often so sensitive to the assumed model error structure. We propose and develop both a parametric and a semiparametric approach to obtain bias corrected estimate. Statistical analysis of administrative data often helps in uncovering trends and patterns that need to be followed up via traditional epidemiologic investigations. Case control studies are often the first choice. However, appropriate selection of controls and lack of power to detect interaction effect are the main concerns of a case control design. We propose a variant of the classical case-control design, the exposure enriched case-control (EECC) design, where not only cases, but also high (or low) exposed individuals are over-sampled, depending on the skewness of the exposure distribution. We show that the judicious oversampling of exposure is possible and can boost the study power particularly when susceptibility genes are rare and environmental exposure is highly skewed.

