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

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 massage 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

List of Figures

List of Tables

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.