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.

Chapter 1

Introduction

1.1 Introduction

The environment where we live in is a key determinant of health and wellbeing. According to the World Health Organization, 23% (95% CI: 13-34%) of all deaths and 22% (95% CI: 13-32%) of the total disease burden in the world is due to modifiable environmental factors (Prüss-Üstün, Wolf, Corvalán, Bos, & Neira, 2016). Cancer and ischemic heart disease are among the top diseases in the list that are influenced by environmental factors. The study of the distribution and determinant of health related adverse events in a specified population in relation to environmental exposures and to control the risks of such exposures are the main concerns of environmental epidemiology (Baker, Kjellstrom, Calderon, & Pastides, 1999).

One of the key features of environmental epidemiology is that the data often have both spatial and temporal dimensions, as risk factors for a disease often vary across time and space. Therefore, it is necessary to combine the knowledge from epidemiology, statistics and geographic information science in the assessment of risk factor variation (Beale, Abellan, Hodgson, & Jarup, 2008). Understanding such complex phenomena often requires simplification of the elements, idealization and modeling (Cox, 1990). Throughout this thesis we focus primarily on the spatial modeling aspect of the environmental epidemiology.

While environment has been classically interpreted to refer to aspects of our physical world (e.g. chemical exposures), environmental epidemiologists are increasingly aware that the social environment also plays an important role in health. There are a number of different pathways and mechanisms through which the social environment impacts on health. For example, exposure to chronic stress can directly affect the immune function and result in disease (Rich & Romero, 2005). Socio-economic status for areas that describes the circumstances in which people live, work and grow; has long been known to predict disease incidence and mortality, with the more disadvantaged areas typically experiencing higher risks (Pickett & Pearl, 2001). Because socio-economic status tends to vary geographically it confounds the spatial variation of the disease rates and if unaccounted for, may seriously bias the relationship and result in misleading conclusions about the possible effects of environmental risk factors on health (Jolley, Jarman, & Elliott, 1992).

The past decade has seen increasing interest in the use of routinely collected administrative data to provide new insights into the environment around us. The use of geographically referenced health databases, present an unique opportunity to investigate the environmental, social and behavioral factors underlying geographic variations in disease rates (Elliott & Wartenberg, 2004). Government agencies such as the NSW Cancer Institute are routinely collecting vast volumes of data as part of state-mandated cancer registries of disease incidence, treatment and outcome. Such data can be extremely helpful in understanding the geographical variation of cancer.

In the case of cancer, variations are complex because cancer is not merely a single disease but a collection of different types of diseases (U. S. Department of Health and Human Services. National Institutes of Health. National Cancer Institute, 2015). Cancer is a broad term used for diseases that arise from uncontrolled cell divisions in different tissues and organs in the body. Depending on tissue or organ type, each of these cancer diseases has its own characteristics of incidence, spread or survival rate (David, 1995). As a result, there are large variations in the risk factors for these different types of cancer. Several studies have revealed the relationship between genetic (Singh et al., 2002; Thompson & Easton, 2001) and environmental factors (Armstrong & Doll, 1975) with cancer.

In addition to the enormous variations in cancer incidence and mortality across geographical locations (Jemal, Center, DeSantis, & Ward, 2010; Parkin, Bray, Ferlay, & Pisani, 2005), some authors (Farrow, Hunt, & Samet, 1992; Nattinger, Gottlieb, Veum, Yahnke, & Goodwin, 1992) report that geographical variation also exist in cancer treatment and outcome. In Australia, Mitchell et al. (2006) reported presence of rural/urban differences in the presentation, management and survival outcomes of breast cancer. Jong et al. (2004) and Nattinger, Kneusel, Hoffmann, and Gilligan (2001) argue that geographical remoteness affects treatment choices made by both patients and clinicians, and consequently has a major impact in the quality of cancer treatment. MacNab (2003) studied geographical variation in incidence rates of intraventicular hemorrhage in neonates as an important health care performance indicator. One of the key mandates for cancer service providers is to ensure the same quality of care for each cancer patient, even though the cancer experience might be different from person to person. That is, the goal is to reduce the geographical variations in clinical outcomes for people diagnosed with cancer (Cancer Institute NSW, 2011).

Addressing the causes of geographical variation related to health outcomes, access, quality of care and productivity often necessitate policy reform (Hannan, 1999). Several authors (Gardner, 1973; Miller, 1994; Tierney & McDonald, 1996; Tierney, Overhage, & McDonald, 1997) suggest that routinely collected administrative data can provide valuable information for identifying possible causes of variations. For example, Gardner (1973) analyzed routinely collected data on mortality and environmental characteristics using a multiple regression method. Careful analysis of routinely collected health care data can identify potential causes of variations and improve the quality and cost effectiveness of treatment. However, the analysis of routinely collected data may be challenging due to complications such as non-Gaussian responses, hierarchical relationships, non-linear predictor effects, measurement error, missing potential covariates, missing values, spatial correlation and within-subject correlations. Therefore, sophisticated statistical models are needed that can handle the large volume of data involved, as well as doing appropriate adjustment for age, gender and other demographic factors, while accounting for spatial and temporal variations (Ash et al., 2012). Although there are strong desires to mine routinely collected administrative data sets, methods for doing so are relatively undeveloped.

My doctoral training has been through a new and unique program in Australia, the Industry Doctoral Training Centre (IDTC), which is loosely modeled on programs that exist in the UK. This training program aims to prepare graduates who are not only technically strong, but also who are experienced and proficient in collaboration and problem-based research. In my case, I had the opportunity to work with the NSW Cancer Institute, analyzing registry data with the aim of developing/identifying statistical methodology to assess geographical variation and hence enhance the quality of care for patients with cancer.

My project involves development and application of sophisticated regression models and methodologies for fitting and making inference with large complex cancer data sets, that involve hierarchical relationships reflecting characteristics not only of individual study subjects, but of their geographical location. The methodology was developed and tested using simulated data, and then applied to actual data. Specifically, my doctoral research has involved resolving a series of problems motivated by applications in environmental epidemiology. These include individual level covariates adjusted conditional auto-regressive modeling in disease mapping studies, quantifying and adjusting for covariate measurement error in spatial linear regression models and developing a new Exposure Enriched Case-Control (EECC) design for assessing gene environment interaction.

The next section of the thesis includes a brief background of the research problem along with main research questions, objectives and hypotheses, literature review and a section on the scope and organization of the thesis.

1.2 Background

Cancer is a major cause of morbidity and mortality in Australia, with an estimated 30% of all deaths attributed to cancer (Australian Institute of Health and Welfare (AIHW), 2014). However, cancer incidence and mortality is not uniformly distributed across population groups, regions and socio-economic status (Australian Institute of Health and Welfare (AIHW) & Australasian Association of Cancer Registries (AACR), 2012). One of the key goals of the NSW Cancer Plan 2011-15 was to reduce variation of cancer

outcomes in NSW by identifying areas of variation and developing strategies based on routinely collected administrative data (Cancer Institute NSW, 2011) obtained from the NSW Cancer Institute managed Centre for Health Record Linkage (CHeReL). Underpinning this goal is the need to examine the current methodology and develop appropriate methods for statistical analysis that enhance reporting.

In this thesis we focus on methods related to the geographical variation of febrile neutropenia across NSW. Febrile neutropenia is the most common infection-related syndrome and an early marker of sepsis in patients receiving myelosuppressive chemotherapy (Aapro et al., 2011; Baden et al., 2012; Crawford, Dale, & Lyman, 2004). Febrile neutropenia is a syndrome characterized by a low absolute neutrophil count and associated fever. More profound and prolonged neutropenia is associated with an increased risk of infection-related complications and death (Bodey, 2009). The importance of febrile neutropenia is underscored by its associated higher incidence of hospitalization and mortality irrespective of cancer type (Lyman et al., 2010). In Australia, chemotherapy and complicating infections accounted for more than 40% of hospital admissions for cancer-related illnesses during 2012-13 (Australian Institute of Health and Welfare (AIHW), 2014). Febrile neutropenia can result in chemotherapy dose reduction and delays which in turn compromise treatment outcomes, and is associated with increased length of hospital stay and health care costs (Crawford et al., 2004; Kuderer, Dale, Crawford, Cosler, & Lyman, 2006; Smith et al., 2006).

Geographical variation of neutropenia is of particular interest because of the uneven concentration of population across different geographical locations within NSW. As a result of this uneven population density, the access to health care services is not universally shared. Moreover, risks of many diseases and health outcomes are influenced by locally varying distributions of socioeconomic, behavioral and environmental risk factors (Elliott & Wartenberg, 2004). These spatially correlated risk factors can have important implications on the observed disease rates in small areas.

Therefore, sophisticated statistical methodologies are needed that can handle the large volume of data involved, account for spatial correlation in the disease rates and adjust for covariates on patients' socio-demographic and clinical characteristics. Additional complexity also arises from administrative data due to the fact that these data sets are

not collected for research purposes. Administrative data may have additional complexity due to the presence of multiple sources of errors, particularly measurement error, processing error and nonresponse error. In most instances administrative datasets are observational, therefore drawing causal inference from these datasets are quite challenging. On a practical level, the analysis of very large datasets will undoubtedly require powerful computing capacity, and data management requires more time than most research datasets. In this research project, we developed novel methodology that can handle large volumes of administrative data, appropriately adjust for age, gender and other socio-demographic factors and provide some guideline to appropriate analysis of administrative data by resolving some of the quality issues.

1.3 Objectives of the Study

The main aim of my thesis is to study the bottlenecks in applying statistical methods to routinely collected administrative data for assessing clinical variation and to develop new methodology that can overcome some of these difficulties. The overall question of interest of my research is:

> *"Can administrative data be effectively used to inform researchers and policy makers about factors that influence spatial variation of disease rates?"*

A major thrust of my thesis concerns the appropriate use of cancer registry data and other administrative data sets to explain geographical variation in disease rates, finding statistically significant factors that contribute to the disease variation, evaluate the impact of data quality in the assessment of those factors and to design future studies that can provide better quality data.

In the first part of my thesis, we develop spatial modeling techniques to explain the geographical variation of febrile neutropenia across NSW. Disease mapping via spatial regression modeling is a key tool in spatial epidemiology that provides a quantitative summary and visual illustration of the underlying geographical variation of the disease. It is also useful for generating new hypotheses on the disease aetiology through

identifying apparently high risk areas/disease clusters (Snow, 1855). However, producing reliable disease maps is complicated by the fact that raw incidence rates are often unstable due to small incidence counts, spatial correlation among rates and also due to the variation in individual patient characteristics (Besag, York, & Mollié, 1991; Clayton & Kaldor, 1987; Cressie, 1993).

Poisson mixed models with conditional auto-regressive random effects are commonly used for assessing the relationship between a rare disease outcome and risk factors in the presence of geographical variation (Lee, 2011). These models can adjust for region specific spatial random effects for correlated disease rates and region specific covariates. However, the currently available methods only allow adjustment based on the age and sex distribution of the underlying population through calculation of an offset in the model (Leroux et al., 1999). Therefore, the effect of age and sex on disease risk cannot be estimated from these models. Moreover, these models ignore a large number of potential individual level covariates related to the underlying disease process which are readily available in health registries. We therefore aim to develop a new methodology that can incorporate both individual and group level covariates while adjusting for spatial correlation in the disease rates.

Specifically, in the first part of the thesis, we will address the following research question:

> *Can we incorporate individual level covariates in currently available Poisson mixed models with conditional auto-regressive random effects?*

We hypothesized that reliable estimates of the regression coefficients for both individual level and group level covariates can be obtained from the individual level covariates adjusted Conditional Auto-Regressive (indiCAR) model. We further hypothesized that this model can be extended to study continuous covariate effects via splines. We also aim to formulate the indiCAR in a distributed computing framework so that individual and group level covariate effect can be estimated separately. This will help to reduce computational costs and overcomes memory space constraints, which is one of the major concerns in applying statistical methodology for disease mapping with large numbers of individual and group level covariates. Such formulations will also provide a convenient way to extend recent developments in Big Data for independent responses to spatially

correlated response.

On further exploration of the applicability of conditional auto-regressive (CAR) based model in assessing geographical variation, it appears that the estimated regression coefficient depends strongly on the assumed spatial correlation structure. Similar sensitivity to the assumed spatial correlation structure can also be seen in analysis of the well-known Scottish Lip Cancer data (Breslow & Clayton, 1993; Clayton, Bernardinelli, & Montomoli, 1993). In another spatial epidemiological study, Molitor et al. (2007) fit a model for the effect of NO_2 exposure on lung function. They considered a series of models, including one based on a conditional auto-regressive (CAR) model. They observed that models with spatial structure give smaller effect estimates as compared to models without spatial structure. These results suggest that estimated coefficients from a spatial regression model can be highly sensitive to whether and how spatial variation is accommodated. We hypothesize that such sensitivity is specially likely to occur when the covariate of interest has been measured with error.

Therefore, in the second part of the thesis, we aim to assess the impact of data quality, in particular the consequences of covariate measurement error in spatial linear regression. That is, our aim is to explore the nature of biases in the regression estimates when covariates of the regression model are measured with error. It is well-known that for classical linear regression, the presence of measurement error in the covariate attenuates the estimated regression coefficient towards no effect. However, the consequences of the measurement error in spatial linear model are not well studied. In particular, we will address the following research question:

> *How to calculate the attenuation factor in case of covariate measurement error in spatial linear regression?*

Relating to the above research question, we hypothesize that (a) reliable estimates of the true regression coefficient can be obtained using the covariance structure of the spatial linear mixed model and (b) semiparametric regression might provide better results in estimating unbiased estimates compared to the method based on modeling covariance structure.

Although the relationship between exposure and disease in a spatial modeling generates useful hypotheses on disease aetiology, variation in individual response to environmental exposures has been a major obstacle to understand the environmental exposure contribution to disease. Case-control studies with detailed individual level exposure information are often desired to support and test the hypothesis generated from spatial correlation studies (Elliott & Wartenberg, 2004; Lawson, 2013; Pacione, 2013). In an environmental case-control study, controls are selected from the neighborhood of the cases to match with the background characteristics of cases. However, selection of cases and controls with respect to same exposure source may result in over-matching: an inappropriate matching strategy that forces similar exposure distributions on the case and control groups, consequently the exposure effect cannot be determined. Moreover, the statistical power to detect interaction effect from a case control study may be limited when exposure distribution is highly skewed and the disease is rare. We hypothesize that judicious over-sampling of high/low exposed individuals can boost the study power considerably. Of course, a traditional logistic regression model is no longer valid in such cases and would result in biased estimation. We show that the addition of a simple covariate to the regression model removes this bias. We applied this concept in studying genetic and environmental interactions.

In brief, the main objectives of our study are as follows:

1. to study the geographical variation of febrile neutropenia across NSW by incorporating both individual and area level covariate information;
2. to develop statistical methods that can incorporate both (a) linear and (b) non-linear individual level covariate effects in disease mapping studies with area specific conditional auto-regressive random effects;
3. to explore the consequence of covariate measurement error in spatial linear regression and to develop new statistical methodologies that overcome some of the identified shortcomings;
4. to design a case control study that can provide better data and achieve higher power in detecting the joint influence of genetic and environmental factors on the risk of developing cancer.

1.4 Literature Review

Spatial modeling

The availability of administrative health databases indexed at a geographical resolution, presents a unique opportunity to investigate local variation in risk factors and disease outcomes. Analysis of such data, however, raises many interesting technical challenges such as how to best handle correlation among neighboring observations, uncertainty due to unequal population sizes, modeling large data sets and measurement error in covariates. Although there are well known statistical techniques to adjust for spatial correlation and uncertainty due to unequal population sizes, relatively little has been done in the context of spatial modeling when the covariate of interest is measured with error and fitting spatial models in large data sets that include both ecological (area level) and individual level covariates. In this thesis we will focus on these aspects.

Spatial modeling of geographical variation of disease in environmental epidemiology has been investigated for a number of reasons including (i) disease mapping to identify apparently high risks areas that could help in policy formulation and appropriate resource allocation, (ii) spatial correlation studies that examine geographic variation with respect to socio-economic and environmental exposures and further generate hypothesis regarding disease aetiology and (iii) spatial clustering to identify disease clusters by obtaining information on the background risk factors (Elliott & Wartenberg, 2004).

Disease mapping has a long history, dating back at least to the early or mid of the nineteenth century (Walter, 2000). Some of the earliest examples of disease mapping include the investigation of yellow fever in the United States just before the year 1800 (Walter, 2000) and cholera outbreak in London in the mid 1800's (Snow, 1855). Although at that time of cholera outbreak the aetiology of cholera was unknown, John Snow's dot map indicated that the outbreak was probably related to a contaminated water source. Since then, a broad range of methodological developments have emerged in relation to the mapping of spatial disease rates.

At the beginning of the Twentieth-Century, the Survey Gazetteers of British Isles produced crude mortality maps based on 1901 Census for England and Wales (Howe,

1964). However, the crude mortality rates used in the map produce erroneous conclusions because of unequal population sizes. In 1928, Percy Stock overcame some of the limitations of the crude mortality map by mapping Standardized Mortality Ratios (SMRs) of cancer stratified by different cancer sites, age and sex distribution of the population (Howe, 1964). The map based on SMRs describes the geographical variation of disease by identifying areas with apparently high risks. The apparent differences in risks are subject to subsequent modeling to identify the underlying disease aetiology.

Spatial modeling of disease rates, however, is complicated by correlations among rates from contiguous areas. In addition, map based on SMRs are unstable and imprecise for sparsely populated areas. Moreover, in the spatial modeling of disease rates, area serves as a surrogate for a combination of various environmental and genetic factors. Therefore, a map of residuals is used to examine the spatial variation of disease due to unmeasured confounders (Gardner, 1973).

Ord (1975) proposed using a first order auto-regressive residual error model in order to account for spatial correlation in the disease rates. Cook and Pocock (1983) considered an alternative approach based on an exponential isotopic correlation structure where the correlation between two observations declines exponentially if they are within a distance of d unit and 0, if the distance greater than d. This specification of spatial correlation is useful when diseases cluster around a point source.

Clayton and Kaldor (1987) proposed shrinking estimates of area specific SMRs towards a common mean, using an empirical Bayes estimation technique. Their approach represents a compromise where each estimated SMR sits somewhere between the overall mean and the region specific rate. The amount of smoothing is determined by the estimated mean from the data, its precision and a prior mean. To accommodate spatial correlation in the disease rates the authors assume that the logarithm of SMRs follow a multivariate Gaussian distribution with a Conditional Auto-regressive (CAR) structure. In this formulation, the conditional variance is assumed to be constant and the conditional expectation is assumed to be compromised between an overall mean and the sum (not the mean) of the neighbors. The constant variance assumes the number of neighbors is fixed which may not be the case in most of the disease mapping situations. Moreover, this method does not allow for adjustment of confounding variables.

Cressie and Chan (1989) proposed an alternative formulation of the CAR model described by Clayton and Kaldor (1987) where the conditional mean remains the same as the CAR model but the conditional variance changes across areas. However, such formulation requires the estimation of large number of parameters and is not extensively used in practice.

Marshall (1991) suggested a non-iterative empirical Bayes estimator for disease mapping based on method of moments. In this formulation, the spatial correlation is accounted for by using priors based on the neighborhood structure. However, such a formulation depends on how neighborhoods are chosen and therefore can be quite subjective.

Besag et al. (1991) extended the Clayton and Kaldor (1987) approach to a fully Bayesian formulation that can incorporate area level confounding variables as well as spatially structured and unstructured random effects. They proposed the use of MCMC algorithm for fitting such model. Their formulation leads to two classes of models: (i) Intrinsic Conditional Auto-regressive model (ICAR) and (ii) convolution model. In the ICAR formulation, the mean of the spatial structured random effect for any region given all other random effects is equal to the mean of the random effects of all neighboring regions, and the conditional variance is proportional to the number of such neighbors. One of the major shortcomings of the ICAR is that the conditional variance is inversely proportional to the number of neighbors even in the independent case.

The convolution approach combines the intrinsic model with a set of independent random effects, hence only the sum of these two is identifiable. The independent random effects are included to account for model over-dispersion. MCMC implementation is difficult for this approach as the prior for the intrinsic and independent components of the model are not identifiable.

In the discussion of the Besag et al. (1991) paper, Raftery and Banfield (1991) suggest use of Gaussian kriging in order to model spatial dependence. Kriging is the most popular method of spatial interpolation in geostatistics, and can describe spatial variation by measuring drift, spatially correlated random variation and noise (nugget effect). The advantage of kriging is that the correlation decreases with distance and it can be used for both areal and point sourced risk data.

Clayton and Bernardinelli (1992) later proposed spatial disease mapping in Bayesian Generalized Linear Mixed Model (GLMM) formulation where logarithmic SMRs can be decomposed into an overall mean plus area specific random effects. The GLMM formulation allows extension of the Clayton and Kaldor (1987) model to include ecological covariates in the model. To estimate parameters, the authors implement a full Bayesian analysis via MCMC. Chi-squared prior distributions for exchangeable and autocorrelated random effects are considered. They also compared the results of a full Bayesian estimates with those from the penalized quasi-likelihood implementation (Breslow & Clayton, 1993).

Bernardinelli and Montomoli (1992) also compared empirical and fully Bayesian approaches to disease mapping and conclude that the latter offer greater flexibility and convenience in the statistical analysis of geographical variation in disease rates.

Breslow and Clayton (1993) proposed a penalized quasi-likelihood estimation procedure as an alternative to the full Bayesian analysis of the generalized linear mixed model. In the application of the PQL, the authors illustrate disease mapping examples in a generalized linear mixed model formulation. They empirically compared estimated SMRs based on the Clayton and Kaldor (1987) approach without ecological covariates in the model, to the PQL with independent and intrinsic auto-regressive random effects. Their results revealed that estimated SMRs are sensitive to the assumed spatial correlation structure. The spatial dependence parameters considered are rather extreme (either no spatial correlation or intrinsic correlation, corresponding to spatial dependence parameter 0 and 1).

Clayton et al. (1993) fitted several spatial regression models with the aim to reduce spatial confounding bias due to omitted covariates. They showed that different assumptions on the spatial correlations lead to different estimates of regression coefficients. They re-analyzed the well-known Scottish Lip Cancer data (Breslow & Clayton, 1993) to study the relationship between exposure to sunlight and lip cancer. However, instead of measuring exposure to sunlight they used the percentage of the population employed in agriculture, fishing and forestry (AFF) as a surrogate measure. They observed that the addition of a clustering term to the model yields a significant decrease in the estimated regression coefficient of AFF. The authors hence argue that

the clustering term accounts for unmeasured confounding variables in the model.

Schlattmann and Böhning (1993) proposed an alternative approach using a discrete mixture of Poisson distributions under the assumption that the disease risk varies in sub-populations. The parameters of such a model are estimated by non-parametric maximum likelihood. The authors assumed that the rates are constant across a sub-population and thus did not account for spatial correlation in disease rates.

Devine, Louis, and Halloran (1994) argued that both empirical Bayes and fully Bayesian estimates experienced over-shrinkage towards their grand mean, particularly, in the case where the sample variances of the underlying rates are greater than the sample variances of corresponding estimates. To guard against such over shrinkage they proposed a constrained empirical Bayes estimation procedure.

Pickle, Mungiole, Jones, and White (1996) produced maps of observed age-adjusted rates and of predicted age-specific rates resulting from a linear mixed model. The model allowed mapping of age specific rates using a single knot-based cubic spline function of age. The knot captures a slope change in the age specific rates beyond the knot. The fitting of cubic splines functions, however, is subject to the selection of the appropriate number of knots and knot locations.

Bernadinelli, Pascutto, Best, and Gilks (1997) suggested a Bayesian hierarchical spatial model for disease mapping that adjusts for covariate measurement error. The authors empirically studied measurement error models using a metropolis Gibbs algorithm by specifying smoothing priors for both relative risks and for covariate with errors. The authors thus assumed that the log odds of the prevalence of the surrogate covariate follows a normal distribution with mean equal to the prevalence of the true covariates, and fixed the variance a priori. The choice of the variance is thus subjective.

Waller, Carlin, Xia, and Gelfand (1997) extended the hierarchical spatial models described by Besag et al. (1991) to accommodate temporal effects and spatio-temporal interaction in the context of disease mapping. They discussed an implementation of such a model using a Metropolis Gibbs algorithm. They also proposed a predictive criterion for model selection, validation and comparison. They reported a slow convergence rate

for the parameter of interest.

Xia and Carlin (1998) extended the spatio-temporal analysis of Waller et al. (1997) by including age adjusted disease rates in the model. They also studied covariate measurement error in the spirit of Bernadinelli et al. (1997). The authors fitted several alternative measurement error models using Metropolis Gibbs algorithm. However, their approach was empirical and they did not address any theoretical aspects of the impact of measurement error in regression modeling.

Diggle, Tawn, and Moyeed (1998) extended the Besag et al. (1991) approach of disease mapping through the use of continuous spatial dependence similar to the classical geostatistical approach. The authors proposed combining a generalized linear model for the outcome and a Gaussian process to represent underlying spatial correlation structure. The parameters of the resulting generalized linear statistical model are then estimated using MCMC. The authors assumed a linear relationship between covariate and disease rates.

Lin and Zhang (1999) proposed a general class of models for over dispersion and correlated data in a mixed model framework. The authors allowed the functional dependence of an outcome variable on covariates to be non-parametric and allowed for correlation between observations by using random effects. They proposed estimation and inference within a unified parametric mixed model framework by representing splines as fixed and random effects components.

Leroux et al. (1999) proposed an alternative to the Bayesian formulation of Besag et al. (1991) approach that allows for both structured and unstructured variation based on a single random effect rather than the sum of two random effects. Under this formulation the conditional mean (and variance) can be obtained as a weighted average of local mean (variance) based on an intrinsic auto-correlation model and mean (variance) based on an independent random effects model. They also described parameter estimation via penalized quasi-likelihood (PQL) (Breslow & Clayton, 1993). They showed that PQL parameter estimates provide nearly unbiased estimates of regression coefficients even for very small expected counts. Moreover, the PQL is computationally faster than the corresponding maximum likelihood (Empirical Bayes) or full Bayesian approaches.

Stern and Cressie (1999) use an alternative formulation of Cressie and Chan (1989) approach that leads to an invariant conditional variance, but replace the conditional mean by the weighted sum of the neighbors (rather mean). Hence the model is inappropriate in practice.

Langford, Leyland, Rasbash, and Goldstein (1999) extended the Besag et al. (1991) approach by decomposing spatially structured random effects as a weighted sum of independent random effects. They discussed various choices for weights including one with an exponential decay model. The parameters are then estimated by translating the spatial model as a multilevel model. However, such formulation requires the inclusion of large numbers of explanatory variables in the model and a large number of constraint vectors.

Leyland, Langford, Rasbash, and Goldstein (2000) latter extended Langford et al. (1999) to allow joint spatial analysis of two outcomes based on a multilevel model. Estimation of parameters associated with such formulation is obtained by iterative generalized least squares technique which relax the constraint structures.

Knorr-Held and Best (2001) described a joint modeling of two diseases in a Bayesian perspective. They introduced a shared component model which acts as a surrogate for unobserved spatially structured covariates affecting the risk of either both or only one of the two diseases. A joint modeling approach can be viewed as an improvement over the Bernadinelli et al. (1997) approach where one disease was used as a surrogate for the risk of other disease, hence introducing measurement error bias. However, underlying this approach is the key assumption that all shared covariates may be modeled by a single latent variable that adequately captured the underlying spatial structure. A lack of information about this assumption will make inference from such models highly prior dependent (Haneuse & Wakefield, 2008).

S. Wood (2003) suggested the use of isotropic thin plate regression smoothers for spatial modeling. He formulated an optimal approximation for the thin plate regression splines that can produce low rank smoothers for both single and multidimensional data. The low rank smoothers avoid the knot placement problems associated with penalized regression splines and provide low rank approximations to generalized smoothing spline

models. They also provide a sensible way of modeling interaction terms in generalized additive models and provide a means for incorporating smooth functions of more than one variable into non-linear models and improve the computational efficiency.

Kammann and Wand (2003) extended the model based geostatistics (Diggle et al., 1998) to an additive model framework. They proposed a semi-parametric formulation of spatial mixed models as a unification of kriging and additive models. Their approach accounts for linear or non-linear covariate effects under the additivity assumption and adjust for spatial correlation by expressing kriging as a linear mixed model. This geo-additive model is of low rank and can be implemented in the standard statistical packages which makes the approach very popular.

Ruppert, Wand, and Carroll (2003); Wand (2003) later describe the relationship between semiparametric regression models based on penalized splines and mixed models. They showed that methods based on maximum likelihood estimation of mixed models can be used for estimation and prediction based on semi-parametric models, using readily available mixed model software (Ngo & Wand, 2004). Wand (2003) argued that semi-parametric regression model can be extended in an efficient way to handle measurement error in predictors.

Burden et al. (2005) explore the relationship between Ischemic Heart Disease (IHD) and an area based measure of social disadvantage via spatio-temporal modeling. In the analysis the author stratified the data from around 300,000 IHD hospital separation records according to spatial location, day and month of hospitalization and age-gender combination resulting a dataset over 39 million records. This large data set is too large to be analyzed using a single spatio-temporal model. The authors explored a series of analyses based on some simplified assumptions. One such assumption is that the effect of time and area on log relative risk is mutually independent. This assumption is rather unrealistic as the spatial effect may change over time. Therefore, more general spatio-temporal models are needed that are capable of handling such Big Data. Moreover, in the application the social disadvantage variable is subject to measurement error and hence may have attenuated the estimated effect.

Ainsworth and Dean (2006) compared estimation of parameters based on the MCMC

and Penalized Quasi-likelihood (PQL) approaches to the spatial regression model described by Besag et al. (1991) through simulation studies. The authors reported that the PQL approach offers similar estimates of relative risks and confidence interval compared to the MCMC implementation. In addition, the PQL approach is computationally simple and requires less iteration to converge.

Lee (2011) empirically compared models proposed by Besag et al. (1991), Stern and Cressie (1999) and Leroux et al. (1999) using MCMC with different specifications of random effects that represent independence, moderate spatial dependence and strong spatial dependence in the disease rates across areas. The author noted that among the models compared, the Leroux et al. (1999) produced consistently better estimates across all the specification of the random effect considered.

The traditional ecological regression methodologies as discussed above use data that are measured at a variety of hierarchical levels but are accumulated to a common group level to facilitate ecological analysis. The heterogeneous exposure/confounders distribution within these groups, however, result in biased inference referred to as ecological bias or cross-level bias (Anselin, 2002; Greenland, 2001). Several authors argued that combination of ecological data with individual level data can provide a useful solution to the ecological bias (Greenland, 2001; Wakefield, 2004a). Although such data provide direct information about the relationship between exposure and response at individual level, little research has been done into statistical model that combine individual and group level data in the ecological modeling context.

Prentice and Sheppard (1995) studied an aggregated data study design in a non-spatial framework that combined age and sex specific disease rates from routinely collected administrative data and individual level covariate data from a random sample of individuals from each of the several population cohorts. They developed an estimating equations framework to obtain reliable estimates of individual level rate parameters. They considered outcome information at an aggregated level, ignoring individual level outcome data.

Best, Ickstadt, and Wolpert (2000) considered a Bayesian spatial modeling approach that combined both individual and group level information to reduce ecological bias

while accounting for spatial correlation in the rates. Underlying their approach is the key assumption that all the spatially varying data are related to a latent, spatially continuous stochastic process representing unexplained spatial variation in risk. They proposed modeling this latent spatial covariate via a kernel density. In this formulation, the baseline risk reflects unattributable risk which may confound the effect of the latent covariate for a misspecified true spatial variation.

Guthrie, Sheppard, and Wakefield (2002) extended the aggregate data model described by Prentice and Sheppard (1995) to allow residual spatial correlation in the disease rates. Specifically, the authors combined the aggregated data model and the Bayesian disease mapping model discussed by Besag et al. (1991). To represent structured variation the author used an exponential correlation structure and fitted using MCMC. They argued that only a small sample of covariate data can yield reliable estimates of the disease risk in aggregated data model.

Wakefield (2004a) discussed various sources of biases in ecological regression model. The author argued that modeling spatial variability in risks via random effects can control neither ecological bias nor covariate measurement error bias. Ecological bias arises due to within area variability of exposures/confounders. The author suggested the use of individual level covariate information in order to reduce ecological bias.

Wakefield (2004b) also studied combined aggregated disease data with individual cohort data. The author showed that the combination of aggregated and individual level data provides unbiased estimate with improve precision. However, such an approach is inefficient in the investigation of rare outcomes (Haneuse & Wakefield, 2008).

Jackson, Best, and Richardson (2006) suggest a joint modeling approach by combining the same covariate information from a random sample of individuals in each area with area level aggregation in an ecological regression framework. The authors considered a binomial modeling approach for group specific outcome counts and fitted using MCMC for a constant population size of 1000 within a area. The authors suggested that including individual level covariates improves inference. However, the generalization of this study is not clear in other applications as different covariates may be available for aggregated and individual level data. Moreover, in many applications individual

reporting may not corresponds to the aggregate level.

Martinez, Benach, Ginebra, G Benavides, and Yasui (2007) extended the aggregated data analysis described by Prentice and Sheppard (1995) in order to incorporate both individual and group level outcome data. They proposed an estimation-equation framework for estimation and inference. They showed that combining two sources of data through individual and group level modeling provides higher statistical power to detect exposure effect and improve estimation (Martínez et al., 2009). However, they did not consider spatial correlation in the disease rates.

Haneuse and Wakefield (2007) proposed a hybrid design via a hybrid likelihood that combines ecological data with a sample of individual level case control data to improve inference. The hybrid likelihood is derived by averaging the individual level likelihood over the uncertainty of the unobserved complete individual level data. They showed that estimation and inference of such hybrid designs can be carried out via either maximum likelihood or via MCMC in a Bayesian framework. However, this approach is computationally intensive and does not incorporate spatially structured random effect.

Wakefield and Sebastien (2008) proposed the use of a two-phase study design framework (Weinberg & Wacholder, 1990; White, 1982) in order to remove ecological bias through combining individual and group level ecological data. In this framework, outcomes are stratified according to the combination of geographical regions, confounders and exposure information at stage I. More detailed information on exposure and confounders are obtained at stage II. They showed this approach not only remove ecological biases but also provides efficient estimates of the regression parameters. However, this approach relies on supplementary exposure information at the individual level covariate rather than incorporating readily collected individual level covariates.

In this thesis we explore a novel individual level covariate adjusted conditional auto-regressive (CAR) model that can incorporate both individual and group level covariates while adjusting for spatial correlation in the disease rates. Our model allows adjustment for both linear and non-linear covariate effect in conditional auto-regressive (CAR) models. In both of these approaches we use the CAR structure proposed by Leroux et al. (1999) and estimation of the corresponding parameters are carried out

using PQL.

Covariate measurement error

Despite these excellent development in the theory of spatial modeling, relatively less attention has been paid to the context where the covariates in the spatial regression model are measured with error. The presence of measurement error in the covariate of interest arises in many epidemiological and social behavioral studies. For example, in the study of geographical variation in bladder cancer rates, lung cancer risk might be included in the model as a proxy for smoking exposure (Bernadinelli et al., 1997; Clayton et al., 1993; Xia & Carlin, 1998). In environmental epidemiology, individual air pollution exposures might be approximated by the distance from the polluted sites or by using the measures at a few monitoring sites (Carroll et al., 1997). These erroneous data can produce systematic bias or result in random errors or imprecision, which typically lead to bias toward no effect (Bernadinelli et al., 1997). Neither Bernadinelli et al. (1997) nor Xia and Carlin (1998) theoretically quantified the measurement error bias.

The measurement error problems have been widely studied in the context of independent data (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Fuller, 1987). Many approaches have been discussed in the literature for obtaining correct estimates of parameters in the presence of measurement error of independent data. These approaches include both parametric and non-parametric formulations with estimation based on maximum likelihood, quasi-likelihood, generalized least squares and conditional independence models (Carroll et al., 2006; Richardson & Gilks, 1993). However, relatively few papers have addressed the specific context of spatial modeling.

Li, Tang, and Lin (2009) derived asymptotic bias expressions for estimated regression coefficients in the context of a spatial linear mixed model. They showed that the regression estimates obtained from naive use of an error prone covariate attenuates the estimated regression coefficient, and that variance component estimates are also inflated. They proposed the use of a maximum likelihood approach based on the EM algorithm to adjust for measurement error. Their method performs well over various spatial correlation structures; namely exponential, Gaussian and conditional auto-regressive

structure (CAR). However, their simulation assumes that the measurement error variance is known and they did not assess the performance in the case of misspecification of the measurement error variance. Moreover, their result did not address the setting where the degree of spatial correlation associated with the covariates differs from the degree of spatial correlation in the error of the regression model. Their approach is also subject to a high computational burden. Gryparis, Paciorek, Zeka, Schwartz, and Coull (2009) and Szpiro, Sheppard, and Lumley (2011) noted that convergence can be very slow for large data sets and can lead to spurious result when there are outliers or in the case of model misspecification. Furthermore, Szpiro et al. (2011) argued that in the presence of spatial correlation, joint modeling of the kind proposed by Li et al. (2009) becomes challenging as it is very difficult to separate out the spatial correlation between exposure and outcome.

Paciorek (2010) addresses the impact of omitting an important covariate that is spatially correlated from a spatial linear regression model. He studied the effect of different residual spatial structures on the bias and precision of estimated regression coefficient in the context of omitted covariates. He argues that spatial models are particularly sensitive to estimation bias induced by unmeasured confounders that are spatially correlated.

In this thesis, 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. We proposed and developed several approaches based on both a parametric and a semi parametric method to obtain bias corrected estimates. The parametric approach has been published in *Environmetrics* and the non-parametric approach in *Biometrics*.

Design issues

Although spatial modeling with group level covariates are useful in generating hypotheses, the group level exposure may not reflect each individual's exposure experience in the group. Therefore, a detailed study of individual level exposure characteristics is often desired. Recent exploration of the human genome offers new opportunities to understand how genetic and environmental factors interplay to cause

disease. Case-control studies are often the first choice to explore the joint influence of genetic susceptibility and environmental risk factors on the risk of developing a rare disease. Such studies generally attain greater power to detect gene-environment interaction than comparably sized cohort studies, as cases are over-sampled from the underlying populations in a case control study (Clayton & McKeigue, 2001). Moreover, such design allows evaluation of the dose response relationship of the level of environmental exposure with the genotype of interest (Khoury, Adams Jr, & Flanders, 1988). However, the validity of the case-control study results largely depends on the appropriate selection of controls in the study (Miettinen, 1985). Case control studies generally use unrelated controls from the population and require large sample sizes in detecting gene-environment interactions (Foppa & Spiegelman, 1997; García-Closas & Lubin, 1999; Luan, Wong, Day, & Wareham, 2001). To address this, various alternative designs have been proposed in the literature.

White (1982) proposed a two stage design where exposure (or an appropriate surrogate) is first measured in a large number of case and control subjects (Stage I). At Stage II, detailed covariate information is obtained for a subset from each strata defined by case/control and exposure status. Breslow and Cain (1988) formalized and generalized White's approach to a general two-stage design with analysis proceeding via logistic regression applied to stage II data, but including an offset term that reflects the stage I sampling probabilities.

Weinberg and Wacholder (1990) suggest a slightly simpler approach to the analysis of two stage designs based on a so-called pseudo-likelihood approach that condition on being sampled in stage II. Their method also requires the inclusion of an offset reflecting sampling probabilities into the logistic regression. While these approaches all provide consistent estimate of main and interaction effects, they require knowledge of the screening variable specific disease rates.

Piegorsch, Weinberg, and Taylor (1994) proposed a case only design in the assessment of gene-environment interaction under the assumption that genetic susceptibility is independent of environmental exposure. The authors showed that under this assumption the case only study provides more efficient estimates of parameters and higher power to detect interaction effect than a case control study with a similar number of cases. The

inference, however, is highly sensitive to this assumption (Albert, Ratnasinghe, Tangrea, & Wacholder, 2001). Moreover, the case only design can only be used for testing interactions, not the main effect.

Khoury (1994) argued that both sibling controls and case-parental controls can adjust for genetic background and thus avoid bias from population stratification. Andrieu and Goldstein (1996) later proposed the use of relatives as control in a case control study to detect gene-environment interaction when the genetic factor is common. Although these family based designs provide higher power to detect interaction parameter, they are generally less powerful for testing main effects. Moreover, the use of relatives as controls may lead to over-matching on various genetic and environmental factors (Khoury & Flanders, 1996; Thomas, 2010).

Khoury and Flanders (1996) reviewed the family based designs along with case-only design and argued that the apparent benefit of these non-traditional case-control study may be inferior to a well conducted case-control study with unrelated control. Furthermore, they noted that neither the genotype nor the exposure effect can be estimated from a case only design.

Langholz and Borgan (1995) proposed counter-matching, an exposure stratified sampling method where a control is randomly sampled from those in the risk set that have exposure status opposite to that of the case. This allows for more variability in exposure in the sampled risk set compared to a random sample of controls and hence is more efficient than traditional case control design. They showed that the counter-matching partial likelihood is proportional to the full cohort partial likelihood in the case of modeling a single exposure variable. However, this approach require exposure or exposure-related surrogate information on the full cohort and detailed exposure variable is typically only collected on a subset.

Umbach and Weinberg (1997) showed that maximum likelihood estimates of the main and interaction effects corresponding to a logistic regression model can be obtained by fitting a constraint log linear model to the case control data. The constraint serves the basis for genetic and environmental independence. As a consequence, the interaction estimate depends on cases, and the genetic and environmental effect depends on control

only through their marginal total. This results improved precision for the main and interaction effect. Weinberg and Umbach (2000) later argued that the estimates of gene-environment interaction based on case only are more precise than those based on a classical logistic regression analysis. However, the author argued that a population based case-control study is needed to confirm the apparent gene-environment interaction in a case only study.

Andrieu, Goldstein, Thomas, and Langholz (2001) studied the counter-matching design to estimate the effect of gene-environment interaction as well as both the genetic and environmental main effects. Their approach requires exposure or surrogate information for both genetic and environmental factors in the whole cohort. They showed that the gain in efficiency of counter matching is highly dependent on the choice of highly specific and sensitive surrogates. The author noted similar efficiencies in estimated interaction effect can also be obtained from a two-phase design.

Chatterjee, Kalaylioglu, and Carroll (2005) proposed a conditional likelihood method for family based case-control design that utilize the gene-environment independence assumption. They showed that exploiting the gene-environment interaction effect in the model provides highly efficient estimates of the gene-environment interaction effect. Later Chatterjee and Carroll (2005) proposed a semiparametric approach for case control that uses data from both the cases and the controls. The authors showed that efficient estimates of the main and interaction effects can be obtained by employing a gene-environment independence assumption in the analysis.

Kraft, Yen, Stram, Morrison, and Gauderman (2007) proposed a joint likelihood ratio test of marginal genetic effect and gene-environment interaction effects for case control data. They showed that the joint test had greater power than a marginal test of gene effects and than a traditional test for gene environment interaction based on case-control data.

Mukherjee and Chatterjee (2008) proposed a shrinkage estimate of gene-environment interaction effects based on an empirical Bayes approach that provides weighted averages of the case-only and case-control estimators. This avoids a two step test for determining gene-environment independence based on control samples and test of interaction effects

based on case-only analysis. When gene-environment independence in the control population holds, the empirical Bayes estimators became more similar to the estimator from the case only design. Similarly, in the absence of gene-environment independence, the empirical Bayes estimator approximate the estimator from case-control study. This provides a trade off between bias and efficiency as it takes advantage of smaller standard error from a case-control study. The authors also argued that their approach is robust to the departure of gene-environment independence assumption.

Chen, Kang, VanderWeele, Zhang, and Mukherjee (2012) proposed a two stage design for detecting gene-environment interaction assuming a gene-environment independence. They showed that enriching exposure information in control selection for genotyping provides power advantages in binary exposure situations.

Stenzel, Ahn, Boonstra, Gruber, and Mukherjee (2015) empirically studied power properties of exposure enriched sampling for a binary exposure based on case-only, case-control and empirical Bayes approaches. They also considered the consequence of exposure misclassification on power and parameter estimates. They showed that exposure enriched sampling provides biased estimates of the parameters, but still exhibit greater power properties. However, the author did not outline any methodology to correct the biased estimates.

In this thesis, we further extend exposure enriched sampling with a continuous environmental exposure to estimate gene-environment interaction in case control studies. We show that the selection of individuals based on high (or low) value of the exposure results in biased estimation of the regression coefficients when standard logistic regression is used. We further show that valid statistical inference can be achieved simply by the addition of a single covariate that reflects this exposure-related category. We also discuss optimal design, showing that judicious over-sampling of high/low exposed individuals can boost study power considerably. Although existing two stage case control designs (Breslow & Cain, 1988) and their matched variant, counter matching (Langholz & Borgan, 1995), are known to have higher power than traditional case control design, they can only be used if surrogate information on gene, exposure or both is available. The efficiency obtained from these two designs though similar, counter matching designs are complex and require specific and sensitive surrogates for the risk

factor of interest. Our Exposure Enriched Case Control (EECC) design to detect gene-environment interaction uses similar underlying probability principle to pseudo-likelihood analysis based on a two stage design, hence will result in similar efficiency for an appropriate oversampling of high exposed individuals.

1.5 Organization of the Thesis

This thesis is organized in the following fashion. We start by developing a novel individual level covariate adjusted conditional auto-regressive (indiCAR) model that can incorporate both individual and group level covariates while adjusting for spatial correlation in the disease rates in Chapter 2. An extension of the indiCAR model to allow for non-linear smooth individual level covariate effect is presented in Chapter 3. In Chapter 4, we quantify the bias due to covariate measurement error in spatial linear regression and propose parametric approaches to calculate the attenuation factors and hence obtain reliable estimates of the regression coefficient. In Chapter 5, we formulate a semi-parametric model to adjust for covariate measurement error bias in spatial linear regression model. In Chapter 6, we discuss an Exposure Enriched Case Control (EECC) study that enhance power by oversampling exposure from the tail area of the exposure distribution. Finally, Chapter 7 gives an overview of the results of the hesis, discusses the contributions of the thesis, and presents areas for future work.

1.6 List of Publications Arising From this Thesis

All of the methodology chapters (Chapter 2 - 6) in this thesis are submitted to journals for possible publications. Among these chapters, Chapter 2, 4, 5 and 6 has already been published in *International Journal of Heath Geographics*, *Environmetrics*, *Biometrics* and *Genetic Epidemiology*, respectively. The remaining Chapter 3 is in revision in the Statistics in Medicine. The full list of all the submissions is provided below:

Chapter 2: Huque, Md Hamidul; Anderson, Craig; Walton, Richard and Ryan, Louise (2016). Individual level covariate adjusted Conditional Auto-Regressive (indiCAR)

model for disease mapping. International Journal of Health Geographics, 15, 25. DOI: 10.1186/s12942-016-0055-7.

Chapter 3: Huque, Md Hamidul; Anderson, Craig; Walton, Richard and Ryan, Louise. Smooth individual level covariates in Conditional Auto-Regressive (smooth-indiCAR) model for disease mapping studies. Statistics in Medicine (in revision).

Chapter 4: Huque, Md Hamidul; Bondell, Howard and Ryan, Louise (2014). On the impact of covariate measurement error on spatial regression modeling. *Environmetrics*, 25 (8), 560-570.

Chapter 5: Huque, Md Hamidul; Bondell, Howard; Carroll, Raymond and Ryan, Louise (2016). Spatial regression with covariate measurement error: A semiparametric approach. *Biometrics*, DOI: 10.1111/biom.12474.

Chapter 6: Huque, Md Hamidul; Carroll, Raymond and Ryan, Louise (2016). Exposure Enriched Case-Control (EECC) design for the assessment of gene-environment interaction. Genetic Epidemiology. DOI: 10.1002/gepi.21986.

Chapter 2

Individual Level Covariate Adjusted Conditional Auto-Regressive (indiCAR) Model for Disease Mapping.

Summary

Mapping disease rates over a region provides a visual illustration of underlying geographical variation of the disease and can be useful to generate new hypotheses on the disease aetiology. However, methods to fit the popular and widely used conditional autoregressive (CAR) models for disease mapping are not feasible in many applications due to memory constraints, particularly when the sample size is large. We propose a new algorithm to fit a CAR model that can accommodate both individual and group level covariates while adjusting for spatial correlation in the disease rates, termed indiCAR. Our method scales well and works in very large datasets where other methods fail.

The content of this chapter is published as: Huque, MH; Anderson C; Walton R; and Ryan LM. (2016). Individual level covariate adjusted conditional autoregressive (indiCAR) model for disease mapping. *International Journal of Health Geographics*, 15:25, DOI: 10.1186/s12942-016-0055-7.

Results: We evaluate the performance of our indiCAR method through simulation studies. Our simulation results indicate that the indiCAR provides reliable estimates of all the regression and random effect parameters. We also apply indiCAR to the analysis of data on neutropenia admissions in New South Wales (NSW), Australia. Our analyses reveal that lower rates of neutropenia admissions are significantly associated with individual level predictors including higher age, male gender, residence in an outer regional area and a group level predictor of social disadvantage, the Socio-Economic Index For Areas (SEIFA). A large value for the spatial dependence parameter is estimated after adjusting for individual and area level covariates. This suggests the presence of important variation in the management of cancer patients across NSW.

Conclusions: Incorporating individual covariate data in disease mapping studies improves the estimation of fixed and random effect parameters by utilizing information from multiple sources. Health registries routinely collect individual and area level information and thus could benefit by using indiCAR for mapping disease rates. Moreover, the natural applicability of indiCAR in a distributed computing framework enhances its application in the Big Data domain with a large number of individual/group level covariates.

2.1 Introduction

The risks of many diseases and health outcomes may vary across geographical locations because of locally varying distributions of socioeconomic, behavioural and environmental risk factors (Elliott & Wartenberg, 2004). These spatially correlated risk factors can have important implications for the observed disease rates in small areas. Mapping disease rates over a region offers a visual illustration of geographical variation. These maps are particularly useful for generating new hypotheses through identifying apparently high risk areas or disease clusters (Snow, 1855). However, producing such maps is complicated by the fact that raw incidence rates are often unstable due to small incidence counts, spatial correlation among rates and also due to the variation in individual patient characteristics (Besag et al., 1991; Clayton & Kaldor, 1987; Cressie, 1993).

Poisson mixed models with conditional autoregressive random effects are commonly used

for assessing the relationship between a rare disease outcome and risk factors in the presence of geographical variation (Lee, 2011). These models can adjust for region specific spatial random effects for correlated disease rates and both individual- and region specific covariates. However, the fitting of such models is subject to high computational burden, particularly when the sample size is large and when the number of individual and group level covariates are large. To alleviate such problems, investigators often adjust for the age and sex distribution of the underlying population through calculation of an offset in the model (Leroux et al., 1999). Therefore, the effect of age and sex on disease risk can not be estimated from these models. Moreover, such an approach ignores a large number of potential individual level covariates that may be related to the underlying disease process and readily available in health registries.

Health registries routinely collect geo-coded information relating to the patient's residence at diagnosis, their socio-demographic status and their clinical characteristics. In addition, information on locally varying socioeconomic, behavioral and environmental risk factors for each area under study can also be obtained from other data sources. For example, in Australia, New South Wales (NSW) cancer registries collect cancer treatment and outcome information for each patient diagnosed with cancer, along with their socio-demographic characteristics. Additionally, a Socio-Economic Index For Areas (SEIFA) and an area specific index for remoteness (ARIA) of each patient's residence can be obtained from the Census Bureau. Combining these individual and area level characteristics in mapping studies can help researchers and policy makers to understand the relative contribution of both individual and group level covariates to the observed cancer rates. In addition, combining such data can also reduce ecological bias, which occurs when the group level exposure-disease relationship does not reflect the individual level relationship. A reduction in this bias leads to improved inference about both group and individual level covariates(Haneuse & Bartell, 2011; Jackson et al., 2006). In this paper we propose a novel approach that enables the study of individual level risk factors in mapping studies.

The aim of our current research is to make use of routinely collected administrative cancer treatment and outcome data to explore the possible geographical variation in the rate of neutropenia admissions corresponding to all cancer types across New South Wales (NSW). Neutropenia is a blood disorder with an abnormally low number of neutrophil

granulocytes (a type of white blood cell in the blood), often associated with fever. It is a life threatening complication of cancer chemotherapy and a major cause of morbidity and associated healthcare resource costs. Furthermore, neutropenia results in compromised efficacy due to delays and dose reductions in chemotherapy (Cameron, 2009).

New South Wales is the most populated state in Australia with a population of approximately 7.6 million people. Geographical variations in neutropenia admissions are of particular interest because of the uneven geographical concentration of the population within the state. As a result of this uneven population density, the level of access to health care services is not uniform across the whole region (Australian Bureau of Statistics, 2015). Moreover, neutropenia incidence might also depend on patient age and cancer type, as treatment modalities often vary across different types of cancer and age groups. Therefore, appropriate analysis of geographical variation of neutropenia admissions requires adjustment for both the patient's demographic characteristics and covariates reflecting the patient's geographic location of residence. In our current application, we explore whether there is any spatial variation in the rates of neutropenia admissions after adjusting for patients' individual and clinical characteristics.

In our proposed method, hereafter known as indiCAR, we incorporate individual level covariate information in a two step iterative procedure following an initialization step. At the initialization step, individual level outcome data were fitted against individual level covariates with a Poisson Generalized Linear Model (GLM), ignoring random effects and group level covariates. Then, at the first step, the individual level outcome data were aggregated at the area level and fitted via a Poisson Generalized Linear Mixed Model (GLMM) against area level covariates including a conditional autoregressive spatial random effect, and an offset calculated based on individual covariate contributions. At the second step, the individual level outcome data is fitted via a Poisson GLM with individual level covariates and a second offset calculated based on the contribution of area specific covariates and random effects obtained from the previous step. Steps 1 and 2 are repeated until convergence.

We evaluate the performance of our indiCAR method through simulation studies and also compare indiCAR to the traditional method of age-sex standardisation. (Leroux et al., 1999). Our simulation results show that the proposed indiCAR approach is able to

correctly estimate coefficients associated with both individual and group-level covariates. We illustrate our proposed indiCAR method using data on neutropenia admissions from the NSW Cancer Institute and conclude with some practical guidelines.

2.2 Methods

2.2.1 Data

New South Wales (NSW) cancer registries were used to identify patients diagnosed with cancer, associated treatment procedures and co-morbidities. Specifically, we used data from the NSW Central Cancer Registry (CCR) linked to NSW Admitted Patient Data Collection (APDC). Detailed descriptions of the data items can be obtained from the Centre for Health Record Linkage (CHeReL http://www.cherel.org.au/master- linkage -key). Data were checked for consistency across data sources and linked by assigning a unique Project Person Number (PPN) to each patient. Our study population comprises all cancer patients that were diagnosed with cancer and were hospitalized during the period between 2001 and 2009.

Demographic variables including age at diagnosis, gender, residence at diagnosis, postal area of residence, and the Accessibility/Remoteness Index of Australia (ARIA) were obtained from the CCR database. The ARIA variable was recorded at individual level rather than postal area level because the ARIA index varies within postal areas. The Socio-Economic Index For Areas (SEIFA; an index of social disadvantage) and the geo-coded shape files for mapping corresponding to 2006 census postal areas were obtained from the Australian Bureau of Statistics (ABS). Individual level clinical characteristics such as type of cancer were also obtained from the CCR. The diagnosis of neutropenia admissions and co-morbidity were obtained using data from the Admitted Patients Data Collection (APDC). The ICD-10-AM (International Statistical Classification of Disease and Related Health problem, 10th revision, Australian modification) code D70 (Agranulocytosis) was used to identify admissions with possible neutropenia.

2.2.2 The Model

Suppose the total area under study is divided into M contiguous regions and the number of neutropenia admissions for the i^{th} $(i = 1, 2, ..., n_j)$ individual in the j^{th} $(j = 1, 2, ..., M)$ region is denoted by $\{y_{ij}\}$. Let $\boldsymbol{Y}$ be a vector with elements $\{y_{ij}\}$ that represents the number of neutropenia admissions for all individuals in the study regions of interest. Similarly, let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p)$ and $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_q)$ represent individual and area level covariate matrices with dimensions $n \times p$ and $M \times q$, respectively, where n is the total sample size i.e., $n = \sum_{j=1}^{M} n_j$. We define a replication matrix, $\boldsymbol{Z}$ of dimension $n \times M$ to map group level covariates and random effects to the individual level as

$$\boldsymbol{Z} = \begin{bmatrix} \mathbf{1}_{n_1 \times 1} & \mathbf{0}_{n_1 \times 1} & \cdots & \mathbf{0}_{n_1 \times 1} \\ \mathbf{0}_{n_2 \times 1} & \mathbf{1}_{n_2 \times 1} & \cdots & \mathbf{0}_{n_2 \times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_M \times 1} & \mathbf{0}_{n_M \times 1} & \cdots & \mathbf{1}_{n_M \times 1} \end{bmatrix}.$$

Under the above specifications, conditional on the area specific random effect vector, $\boldsymbol{b}$, the number of neutropenia admissions for each cancer patient is assumed to be a Poisson random variable with mean $\boldsymbol{\mu}$, given by

$$\ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}. \tag{2.1}$$

Here, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vectors of regression coefficients associated with the individual level and group level covariates, respectively. Of course, it is possible to express model (2.1) by replicating group level covariate data to the individual level and including them within the design matrix, $\mathbf{X}$. However, such a formulation often results in high computational burden and a large amount of storage memory allocation. Instead, formulation (2.1) hints on an algorithm that does not necessarily require the replication of area level covariate to the individual level data for model fitting, thus helps to fit individual and group level data separately in a distributed computing framework as will be shown at the end of the current section.

Many different choices for modelling the random effect, $\boldsymbol{b}$ are available in the mapping literature (see Lee 2011, for a recent review). Among these, the method of Leroux et al.

(1999) is appealing because it allows varying weights between spatially structured and unstructured variation (Leroux et al., 1999). Within this framework, the random effect vector, $\boldsymbol{b}$ has a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix, $\boldsymbol{D}$ delivered through its Moore-Penrose generalized inverse, $\boldsymbol{D}^{-} = \sigma^{-2}\{(1-\boldsymbol{\lambda})\boldsymbol{I} + \lambda\boldsymbol{R}\}$, where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{R}$ is the intrinsic auto regression matrix reflecting the neighbourhood structure. Typically, neighbours are those areas which share a common boundary, but distance based neighbourhood structures can also be used (Earnest et al., 2007).The typical element of $\boldsymbol{R}$ is given by

$$\boldsymbol{R}_{jj'} = \begin{cases} m_j, & j = j' \\ -I\{j \sim j'\} & j \neq j', \end{cases}$$

where, m_j is the number of neighbours of region j, and $I\{j \sim j'\}$ is an indicator function that takes value 1 if regions j and j' are neighbours and 0 otherwise. The parameters characterising the random effect distribution, $\boldsymbol{\theta} = (\sigma^2 > 0, \boldsymbol{\lambda} \in [0,1])$ quantify overdispersion and spatial dependence respectively. A larger value of $\lambda \in [0,1]$ indicates a higher degree of spatial correlation among proximal areal units. This specification results in two extreme cases: i) completely independent random effects when $\boldsymbol{\lambda} = 0$ and ii) the intrinsic autoregressive model when $\boldsymbol{\lambda} = 1$ (Besag et al., 1991). In cases where $0 < \boldsymbol{\lambda} < 1$, a weighted combination of these extreme cases is assumed. With the expression of the Moore-Penrose generalized inverse of the covariance matrix as $\sigma^{-2}\{(1-\boldsymbol{\lambda})\boldsymbol{I} + \lambda\boldsymbol{R}\}$ therefore avoids inverting the covariance matrix $\boldsymbol{D}$. Alternatively, one can restrict $\boldsymbol{\lambda}$ to the range $(0,1)$, thus ensuring that $\boldsymbol{D}$ is invertible.

Since the random effects, $\boldsymbol{b}$ are unobserved, inference about $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ can be made by integrating out the distribution of the random effects, $\boldsymbol{b}$. The corresponding integrated quasi-likelihood function is equal to (see equation (2) of Breslow and Clayton (1993))

$$|\boldsymbol{D}|^{-\frac{1}{2}} \int \exp\left[-\frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{n_j} d_{ij}(Y_{ij}, \mu_{ij}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}^{-}\boldsymbol{b}\right] d\boldsymbol{b},$$

where $d(Y, \boldsymbol{\mu})$ refers to the deviance residual associated with observation Y.

The maximum likelihood estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are simply those values which maximize the above quasi-likelihood. However, no simple closed form expression exists

for the integral. Instead, Breslow and Clayton (1993) proposed the penalized quasi-likelihood (PQL) approach for parameter estimation and inference. The PQL uses Laplace's method for integral approximation and jointly maximizes the following quasi-likelihood function to obtain estimates for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{b}(\boldsymbol{\theta})$ (see equation (6) of Breslow and Clayton 1993)

$$-\frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{n_j} d_{ij}(Y_{ij},\mu_{ij}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}^{-}\boldsymbol{b}. \tag{2.2}$$

Under the above specification the approximate log-likelihood can be expressed as

$$\begin{aligned} const + \boldsymbol{Y}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}) - \\ \mathbf{1}^{\mathrm{T}}\exp(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}^{-}\boldsymbol{b}. \end{aligned} \tag{2.3}$$

Differentiating (2.3) with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{b}$ using vector matrix calculus (Wand, 2002), we obtain the following score equations

$$\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\}^{\mathrm{T}}\mathbf{X} = 0, \tag{2.4}$$

$$\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\}^{\mathrm{T}}\boldsymbol{Z}\mathbf{U} = 0, \tag{2.5}$$

and

$$\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\}^{\mathrm{T}}\boldsymbol{Z} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}^{-}. \tag{2.6}$$

Iterative Re-weighted Least Squares (IRLS) can be applied to solve the above equations for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{b}$. However, high computational costs and memory space constraints often make it difficult to apply these iterative procedures to data sets with a very large number of cases. An alternative computational strategy is the use of the Gauss-Seidel algorithm. In this method, at each iteration, one of the parameters is estimated while keeping other parameters fixed at current values. The advantage of such an approach is that substantial simplifications can be obtained at each step. Using this approach, we first initialize $\boldsymbol{\beta}$ and then obtain updated estimates for $\boldsymbol{\gamma}$ and $\boldsymbol{b}$ in the following two step

procedure:

Step 0: Set the coefficients corresponding to area level covariates, $\boldsymbol{\gamma}$ and random effects, $\boldsymbol{b}$ to $\mathbf{0}$ in equation (2.4). Then we have

$$\left\{\boldsymbol{Y} - \exp(\mathbf{X}\widehat{\boldsymbol{\beta}})\right\}^{\mathrm{T}} \mathbf{X} = 0.$$

This equation is the estimating equation for a Poisson generalized linear model (Wand, 2002) and thus can be fitted using the existing glm package in the ***R*** statistical computing environment (R Core Team, 2013). This gives initial estimates of the regression coefficient $\boldsymbol{\beta}$ associated with individual level covariates.

Step 1. Substitute the current estimated individual level coefficients, $\widehat{\boldsymbol{\beta}}$ in equation (2.5) and (2.6) and with some simple algebra, we have

$$\{\boldsymbol{Y}_c - \exp(\mathrm{O}_1 + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{b})\}^{\mathrm{T}} \mathbf{U} = 0$$

and,

$$\{\boldsymbol{Y}_c - \exp(\mathrm{O}_1 + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{b})\}^{\mathrm{T}} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}^{-},$$

where $\boldsymbol{Y}_c^{\mathrm{T}} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Z}$ is a vector of aggregated disease counts of length M at the group level and $\mathrm{O}_1 = \log\{\boldsymbol{Z}^{\mathrm{T}} \exp(\mathbf{X}\widehat{\boldsymbol{\beta}})\}$ is a vector of offset with length M.

The above two equations are well known PQL estimating equations for the Poisson mixed model (Breslow & Clayton, 1993). Since, the outcome $\boldsymbol{Y}_c$, offset O_1, covariate $\mathbf{U}$ and random effects $\boldsymbol{b}$ are all measured at the group level, estimates of parameters for the group level coefficient $\widehat{\boldsymbol{\gamma}}$ and random effects $\boldsymbol{b}$ can be estimated using the PQL method (Breslow & Clayton, 1993; Leroux et al., 1999) with only group level data. The detailed procedure is described in Appendix 2A.

Step 2. Now substitute the estimated area-specific regression coefficient, $\widehat{\boldsymbol{\gamma}}$ and random effect parameter, $\widehat{\boldsymbol{b}}$ estimated at step 1 in (2.4). Then we have

$$\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + \mathrm{O}_2)\}^{\mathrm{T}} \mathbf{X} = 0,$$

where $O_2 = \boldsymbol{Z}(\mathbf{U}\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{b}})$ is an offset vector of length n. Under the above specification, the individual level coefficients estimate $\widehat{\boldsymbol{\beta}}$ can then be updated using ordinary Poisson regression with individual level data.

Steps 1 and 2 are then repeated until the algorithm converges. Estimates obtained by this iterative procedure will be the same, aside from rounding error as the solution obtained by a standard IRLS algorithm.

Estimation of standard error

The approximate standard error estimates for $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$ in step 1 and 2 assume fixed $\boldsymbol{\beta}$ and fixed $\boldsymbol{\gamma}$, respectively. Therefore, we re-calculated the standard error of these regression coefficients by adjusting the variability of the estimated $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$. This can be done via the Iterative Re-weighted Least Squares (IRLS) estimation of score equations (2.4, 2.5 and 2.6). The IRLS estimation requires us to define a working dependent variable and a weight matrix that are updated at each iteration and solved via Fisher scoring (Breslow & Clayton, 1993).

Let the GLM adjusted dependent variable, $\boldsymbol{Y}_{pseudo}$ be

$$\boldsymbol{Y}_{pseudo} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b} + \mathbf{W}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \tag{2.7}$$

where $\mathbf{W}$ is a $n \times n$ diagonal matrix with diagonal elements $\boldsymbol{\mu}$. Harville (1977) and Robinson (1991) showed that the Fisher scoring corresponding to the score equations (2.4, 2.5 and 2.6) and GLM dependent variable as in (2.7), is identical to the normal equation of the best linear unbiased predictors (BLUPs) of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ corresponding to the following linear mixed model

$$\boldsymbol{Y}_{pseudo} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}_{pseudo}, \tag{2.8}$$

where the pseudo-error $\boldsymbol{\epsilon}_{pseudo} \sim N(0, \mathbf{W}^{-1})$. Following Robinson (1991), the estimated regression coefficients for the fixed effects, $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and BLUP estimate for the random

effect $\boldsymbol{b}$ can be obtained as

$$\begin{aligned}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) &= (\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{C})^{-1}(\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}_{pseudo}) \\ \widehat{\boldsymbol{b}} &= \boldsymbol{D}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\{\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\mathbf{U}\widehat{\boldsymbol{\gamma}}\} \end{aligned} \tag{2.9}$$

where $\mathbf{C} = [X|ZU]$ and $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^{\mathrm{T}} + \mathbf{W}^{-1}$, the variance of pseudo-response $\boldsymbol{Y}_{pseudo}$. Thus, the variance-covariance matrix for the fixed effect $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ can be estimated by

$$\boldsymbol{Q} = (\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{C})^{-1}. \tag{2.10}$$

Note that equation (2.9) suggests that estimates of the regression coefficients and variance components can be obtained using the Leroux et al. (1999) model with appropriate specification of the design matrix ($\boldsymbol{Z}$) associated with spatial random effect (2.1). Indeed, a back-fitting approach such as indiCAR will be effective in situations where memory constraints may prohibit fitting a single model consisting of all individual and group level covariates. A useful feature of our indiCAR method is that we can calculate the above standard error in a distributed computing framework. This is because $\boldsymbol{V}^{-1}$ can be expressed as $\mathbf{W} - \mathbf{W}\boldsymbol{Z}\boldsymbol{D}(I + \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D})^{-1}\boldsymbol{Z}^{\mathrm{T}}\mathbf{W}$ (Henderson & Searle, 1981). Therefore, the above variance-covariance matrix can be written as

$$\boldsymbol{Q} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1},$$

where,

$$\begin{aligned} a_{11} &= \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D}(\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D})^{-1} \times \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\mathbf{X} \\ a_{12} &= \mathbf{X}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\mathbf{U} - \mathbf{X}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D}(\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D})^{-1} \times \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\mathbf{U} \\ a_{21} &= a_{12}^{\mathrm{T}} \\ a_{22} &= \mathbf{U}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\mathbf{U} - \mathbf{U}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D} \times (\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D})^{-1}\boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\mathbf{U}. \end{aligned}$$

Among the various components of the above variance-covariance matrix, $\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}$ are the only terms involving individual level data, and the rest of the terms involve a lower dimension corresponding to the group level data. These components are therefore straightforward to calculate. Hence, upon convergence, calculation of the variance-covariance matrix is also carried out in a distributed computing framework for

individual and group-level data separately.

The covariance matrix for $\widehat{\boldsymbol{b}}$ was obtained from the Fisher information matrix from Step 2 in the usual way, assuming that parameters for the individual and area specific covariates are fixed. Of course there is additional variability due to the fact that the individual and area specific covariates parameters are estimated. However, following Breslow and Clayton (1993) we ignore this additional variability when making inference about the parameters which characterize the random effect distribution, $\widehat{\boldsymbol{\theta}}$. The detailed procedure is given in Appendix 2A.

In the next section we describe a simulation study to evaluate the performance of our method.

2.3 Simulation Studies

To evaluate our proposed method we design a simulation study involving 400 regions in a 20×20 square lattice grid with varying sample sizes. Specifically, we consider cases with i) 10 to 1000 and ii) 10 to 50 subjects in each area. We declare two regions to be neighbours if they share a common border. The random effects are generated following a multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{D} = [\sigma^{-2}\{(1-\boldsymbol{\lambda})\boldsymbol{I} + \boldsymbol{\lambda}\boldsymbol{R}\}]^{-1}$. The value of σ is set to 0.4 and five different values of spatial dependence parameters, $\boldsymbol{\lambda} = \{0, 0.25, 0.50, 0.75, 0.99\}$ are considered in order to represent different strengths of spatial correlation. We then generate three individual level covariates (one binary, one categorical and one continuous) and one group level covariate. The binary covariate represents the distribution of sex in the area and is generated following a Bernoulli random variable with probability ranging from 0.45 to 0.55 across groups. The categorical variable with six categories is generated to represent the age distribution of the neutropenia admissions data with prespecified probabilities (similar to the neutropenia admissions data). The continuous individual level variable is generated as Uniform (0.2, 1). The group level covariate is generated from a standard normal distribution. The outcome variable is then generated using model (2.1). The full list of the parameters used to generate data is given in Table 1. The binary and the categorical individual level variables help us to compare our simulation results for the

indiCAR with the age-sex adjusted Leroux et al (1999) approach. Conditional on the area specific random effect $\boldsymbol{b}$, the Leroux et al (1999) model is given by $\ln \boldsymbol{\mu} = \ln \boldsymbol{E} + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{b}$, where $\boldsymbol{E}$ is the expected count, which may be based on the age distribution in the region and a set of standard rates.

2.4 Results and Discussion

In this section we discuss our results obtained from the simulation study and an application to the neutropenia admission data. We compare the results obtained by indiCAR with the existing Leroux et al. (1999) method. When applying indiCAR to the simulated data, we adjust for all individual and areal covariates. However, in the existing Leroux et al. (1999) method we were only able to incorporate the binary and categorical variable by calculating offsets based on direct standardization of these covariates.

2.4.1 Simulation Results

Table 2.1 displays the average of estimated regression coefficients along with their estimated standard errors for indiCAR and Leroux et al. (1999) methods based on 1000 simulation runs assuming a random number of subjects between 10 and 1000 in an area and varying the spatial dependence parameter λ between 0, 0.25, 0.50, 0.75 and 0.99. We estimated two different standard errors of estimated regression coefficients: namely, (i) empirical standard errors i.e., taking the standard deviation of the 1000 simulated regression coefficient estimates, (ii) average of model based standard errors. The first column of Table 2.1 specifies the spatial dependence parameter used in that particular simulation. The next eight columns list the estimated regression coefficients for the individual level covariates using indiCAR method. The tenth, eleventh and twelfth columns list the estimated group level regression coefficients, the estimated over-dispersion parameters and estimated spatial dependence parameters for the spatial random effect using indiCAR method. The last three columns list the estimated regression coefficients for the group specific covariate and estimated over-dispersion and spatial dependence parameter using the Leroux et al. (1999) method. The Leroux et al. (1999) method adjusts only for the binary and categorical variables.

Table 2.1: 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 1000.

	indiCAR											Leroux et al (1999) approach		
True value	β_0 -0.20	β_1 -2.50	β_2 0.70	β_{32} -2.00	β_{33} -1.50	β_{34} 0.20	β_{35} 0.50	β_{36} 0.80	γ 0.20	σ 0.40	λ	γ 0.20	σ 0.40	λ
λ	Estimated coefficient													
0.00	-0.175	-2.500	0.700	-2.005	-1.501	0.200	0.500	0.799	0.198	0.393	0.017	0.197	0.439	0.064
0.25	-0.172	-2.500	0.700	-2.002	-1.498	0.201	0.499	0.799	0.198	0.394	0.251	0.197	0.421	0.308
0.50	-0.163	-2.500	0.699	-1.997	-1.498	0.202	0.502	0.802	0.200	0.395	0.503	0.199	0.412	0.523
0.75	-0.147	-2.500	0.700	-2.005	-1.501	0.198	0.499	0.798	0.199	0.393	0.730	0.198	0.406	0.719
0.99	-0.117	-2.501	0.700	-1.998	-1.500	0.202	0.502	0.803	0.198	0.398	0.956	0.198	0.413	0.947
	Empirical standard error													
0.00	0.034	0.015	0.012	0.061	0.037	0.026	0.027	0.026	0.020	0.028	0.025	0.021	0.035	0.051
0.25	0.038	0.016	0.012	0.062	0.038	0.026	0.026	0.027	0.018	0.029	0.098	0.018	0.029	0.099
0.50	0.044	0.016	0.012	0.061	0.040	0.028	0.027	0.028	0.016	0.028	0.123	0.016	0.027	0.111
0.75	0.053	0.016	0.012	0.061	0.040	0.028	0.027	0.027	0.014	0.025	0.114	0.015	0.024	0.105
0.99	0.206	0.017	0.012	0.061	0.039	0.027	0.026	0.026	0.013	0.019	0.043	0.013	0.020	0.049
	Average of the simulated standard error													
0.00	0.034	0.016	0.012	0.061	0.039	0.027	0.027	0.027	0.021	0.028	0.045	0.022	0.031	0.055
0.25	0.036	0.016	0.012	0.061	0.039	0.028	0.027	0.027	0.017	0.029	0.091	0.017	0.030	0.098
0.50	0.041	0.016	0.012	0.062	0.039	0.028	0.027	0.027	0.015	0.026	0.113	0.015	0.026	0.114
0.75	0.050	0.016	0.012	0.062	0.039	0.028	0.027	0.027	0.014	0.021	0.101	0.014	0.022	0.102
0.99	0.130	0.016	0.012	0.061	0.039	0.028	0.027	0.027	0.013	0.019	0.035	0.013	0.020	0.039

As expected, the indiCAR method provides consistent estimates of the individual level and region specific regression coefficients and the parameters in the spatial random effect. Though the Leroux et al. (1999) method provides similar consistent estimates of the true region-specific regression parameters, however, the parameters in the random effect are slightly biased.

To evaluate the performance of the proposed method under small sample settings, we also conducted simulations with only 10 to 50 subjects per region. The results are given in Table 2.2. As indicated in the table, the proposed method performs very well in this setting providing consistent estimates of all the parameters. In contrast, the Leroux et al. (1999) method provides slightly less efficient estimates of spatial dependence parameters.

2.4.2 Application to the Neutropenia Data

We applied our methodology to the data on neutropenia admission in New South Wales, Australia. One of the key objectives of this analysis is to assess the geographical variation of neutropenia admission rates and its association with area level measures of socio-economic status. Data also includes patient age, gender, year of diagnosis, ARIA, cancer types at diagnosis, number of major comorbidities excluding cancer during hospital discharge and geographic location reported via postcode of residence.

Table 2.3 shows the descriptive statistics of cancer patients diagnosed and treated between years 2001 and 2009 in New South Wales, Australia. The proportion of neutropenia admissions decreases gradually with increasing age (9.2 % for 20-30 years of age to 1.7 % for 80+ years of age). Overall, the rates are similar ($\approx$ 5 %) across the years 2001 to 2008 but are considerably lower (3.0 %) in the year 2009. This might be due to the fact that the data are date limited to those patients diagnosed with cancer and treated in 2009. As cancer treatment often requires long duration and subsequent neutropenia admissions may have happened beyond the study period. The proportion of neutropenia is highest (4.9 %) in the major cities followed by inner regional Australia (3.9 %). Among the various types of cancer, the highest proportion of neutropenia admissions are observed for hematological malignant cancer patients (25.0 %) followed by lung (6.2 %) and breast cancer (5.3 %). The proportion of neutropenia admissions

Table 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.

	indiCAR											Leroux et al (1999) method		
True value	β_0 -0.20	β_1 -2.50	β_2 0.70	β_{32} -2.00	β_{33} -1.50	β_{34} 0.20	β_{35} 0.50	β_{36} 0.80	γ 0.20	σ 0.40	λ	γ 0.20	σ 0.40	λ
λ	Estimated coefficient													
0.00	-0.159	-2.496	0.698	-2.020	-1.508	0.206	0.506	0.805	0.195	0.383	0.045	0.186	0.449	0.058
0.25	-0.178	-2.507	0.703	-2.017	-1.493	0.212	0.510	0.810	0.199	0.379	0.236	0.193	0.427	0.228
0.50	-0.176	-2.500	0.699	-2.021	-1.502	0.202	0.500	0.802	0.198	0.376	0.449	0.194	0.426	0.378
0.75	-0.178	-2.500	0.702	-2.024	-1.506	0.202	0.503	0.803	0.198	0.380	0.671	0.197	0.440	0.576
0.99	-0.148	-2.502	0.698	-2.033	-1.509	0.199	0.500	0.800	0.199	0.405	0.922	0.198	0.514	0.849
	Empirical standard error													
0.00	0.117	0.067	0.052	0.256	0.160	0.115	0.112	0.116	0.032	0.062	0.077	0.034	0.066	0.079
0.25	0.093	0.052	0.039	0.196	0.128	0.090	0.083	0.085	0.023	0.048	0.155	0.025	0.052	0.150
0.50	0.090	0.049	0.041	0.209	0.124	0.085	0.084	0.085	0.023	0.044	0.191	0.025	0.050	0.180
0.75	0.099	0.053	0.037	0.198	0.131	0.086	0.085	0.088	0.024	0.041	0.169	0.026	0.045	0.182
0.99	0.238	0.066	0.048	0.265	0.165	0.120	0.118	0.117	0.027	0.051	0.098	0.030	0.057	0.146
	Average of the simulated standard error													
0.00	0.116	0.066	0.049	0.254	0.161	0.114	0.111	0.113	0.031	0.063	0.124	0.032	0.061	0.105
0.25	0.092	0.051	0.038	0.198	0.125	0.088	0.086	0.087	0.025	0.049	0.164	0.027	0.050	0.144
0.50	0.093	0.052	0.038	0.198	0.125	0.088	0.086	0.087	0.024	0.044	0.196	0.025	0.046	0.171
0.75	0.097	0.052	0.038	0.198	0.125	0.088	0.086	0.087	0.023	0.041	0.170	0.025	0.043	0.167
0.99	0.164	0.067	0.049	0.259	0.163	0.115	0.112	0.114	0.028	0.056	0.067	0.029	0.054	0.095

Table 2.3: Descriptive analysis of neutropenia data.

Variables	Neutropenia N (%)	Total N=279,623
Age group		
20-30 years	408 (9.2)	4418
30-39 years	851 (7.7)	10,988
40-49 years	1649 (6.2)	26,395
50-59 years	2942 (5.6)	52,281
60-69 years	3465 (4.8)	71,446
70-79 years	2577 (3.7)	69,236
80+ years	769 (1.7)	44,859
Sex		
Female	6,363 (5.0)	127,519
Male	6,298 (4.1)	152,104
Year of Diagnosis		
2001	1,343 (4.9)	27,356
2002	1,411 (5.0)	28,451
2003	1,503 (5.1)	29,560
2004	1,478 (4.8)	30,970
2005	1,596 (5.1)	31,533
2006	1,452 (4.6)	31,865
2007	1,453 (4.5)	32,603
2008	1,405 (4.2)	33,343
2009	1,020 (3.0)	33,942
ARIA		
Major Cities	9,199 (4.9)	189,322
Inner Regional Australia	2,638 (3.9)	67,086
Outer Regional Australia	774 (3.6)	21,664
Remote or Very remote Australia	50 (3.2)	1,551
Cancer Type		
Breast Cancer	2,059 (5.3)	38,620
Lung cancer	1,401 (6.2)	22,744
Colon & rectum cancer	1,011 (3.0)	34,018
Haematological Malignancy	5,134 (25.0)	20,518
Other cancer	3,056 (1.9)	163,723
No. of major comorbidities		
0	6,072 (3.7)	163,645
1	2,228 (4.9)	45,817
2	2,315 (6.7)	34,670
3	976 (5.7)	17,264
4+	1,070 (5.9)	18,227
SEIFA		
Most disadvantaged	1,388 (4.6)	30,302
2	1,750 (4.1)	42,558
3	3546 (4.5)	78,006
4	2800 (4.6)	60,880
Least disadvantaged	3177 (4.7)	67,877

are very similar across various SEIFA index categories.

Table 2.4 reports the multivariable analysis of neutropenia admission data using indiCAR and the Leroux et al. (1999) method based on age-sex adjustments. We calculate age-sex adjusted standardized incidence ratios (SIR) by dividing the observed number of neutropenia admissions by the age-sex adjusted expected number of neutropenia admissions (Breslow & Day, 1987). Our results reveal that higher age, male gender, patients residing in outer regional and remote area and with higher socioeconomic status all have a significantly lower rate of neutropenia. The estimated over-dispersion (σ) and spatial dependence parameters (λ) with indiCAR are 0.204 and 0.992, respectively. These figures with Leroux et al. (1999) are estimated as 0.210 and 0.989. Both the methods estimate a very strong spatial dependence (≈ 1) in the neutropenia admission across NSW.

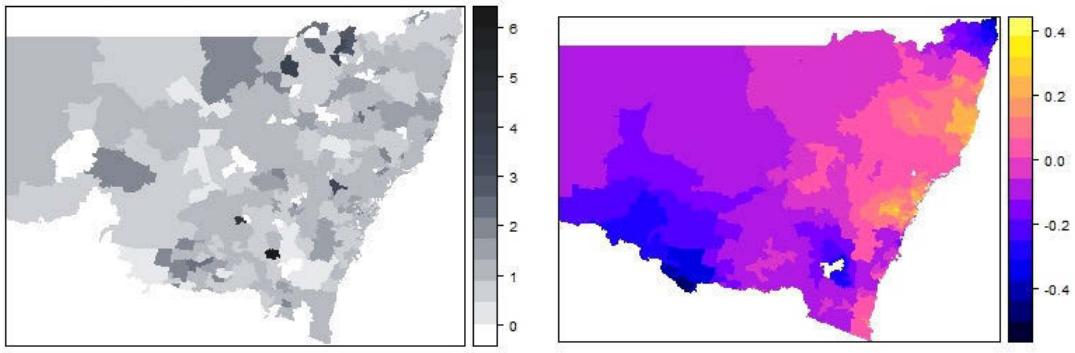
Although advanced age has been identified as a significant predictor for neutropenia admissions in previous studies (Klastersky et al., 2000), we observed a lower risk of neutropenia admissions associated with increasing age. This might be due to the fact that the current risk based prophylactic administration of Colony Stimulating Factor (CSF) guidelines account for patients advanced age (Aapro et al., 2011).

The dependence between average neutropenia rate on ARIA and SEIFA are in the opposite direction, which is counter intuitive as remote areas in NSW are mostly associated with disadvantage SEIFA categories. However, the observed contrast in estimated regression coefficients might be due to the differences in the health care practices. Patients in the remote areas where the patients are geographically distant to the treating medical oncologist are best managed by their primary care physicians, hence, may be treated with lower doses of chemotherapy (Fox & Boyce, 2014). On the contrary, patients in the major cities might get intensive chemotherapy to treat them early, and are better managed due to availability of resources. Previous studies also indicate that remoteness have greatest effect in quality of cancer treatment (Jong et al., 2004) and it affect treatment choices made by both patients and clinicians (Nattinger et al., 2001).

Figure 2.1(a), shows the Standardized Incidence Ratios (SIR) of neutropenia admission in NSW. Six postal areas in NSW had an estimated SIR>3 as shown in the map. Five

Table 2.4: Comparison of individual covariate adjusted conditional auto-regressive model (indiCAR) with the age-sex adjusted Leroux et al. (1999) method.

	indiCAR		**Leroux et. al.**	
Regression coefficients	Estimates	Std. Error	Estimates	Std. Error
Intercept	-2.781	0.110	—	—
Age group				
20-30 years	0.124	0.056	—	—
30-39 years	0.208	0.042	—	—
40-49 years	Ref			
50-59 years	-0.119	0.031	—	—
60-69 years	-0.287	0.031	—	—
70-79 years	-0.712	0.033	—	—
80+ years	-1.586	0.045	—	—
Sex				
Female	Ref			
Male	-0.082	0.020	—	—
Year of Diagnosis				
2001	Ref			
2002	0.018	0.038	—	—
2003	0.083	0.038	—	—
2004	0.021	0.038	—	—
2005	0.096	0.037	—	—
2006	0.036	0.038	—	—
2007	0.026	0.038	—	—
2008	-0.001	0.038	—	—
2009	-0.315	0.042	—	—
ARIA				
Major Cities	Ref			
Inner Regional Australia	-0.023	0.047	—	—
Outer Regional Australia	-0.147	0.068	—	—
Remote/ Very remote Australia	-0.231	0.163	—	—
Cancer Type				
Breast Cancer	Ref	—	—	—
Lung cancer	0.253	0.038	—	—
Colon & rectum cancer	-0.434	0.040	—	—
Haematological Malignancy	1.572	0.029	—	—
Other cancer	-0.942	0.031	—	—
No. of major comorbidities				
0	Ref	—	—	
1	0.413	0.026	—	—
2	0.670	0.026	—	—
3	0.609	0.036	–	—
4+	0.605	0.035	–	—
SEIFA				
Most disadvantaged	Ref			
2	-0.083	0.044	-0.075	0.042
3	-0.071	0.041	-0.068	0.038
4	-0.125	0.047	-0.121	0.044
Least disadvantaged	-0.131	0.056	-0.129	0.052
Variance parameter				
σ	0.204	0.023	0.210	0.022
λ	0.992	0.012	0.989	0.015



(a) Standardized Incidence Ratio.

(b) Estimated spatial random effects.

Figure 2.1: Estimated (a) Standardized Incidence Ratios (SIR) and (b) spatial random effects of neutropenia admissions in NSW, Australia using indiCAR.

of these postcodes are in the remote or very remote area. There is also very high spatial dependence of the neutropenia rates across NSW (Figure 2.1(b)). The white region in the map of NSW is the Australian Capital Territory (ACT). Two other Australian states, Queensland (QLD) and Victoria (VIC) are located to the North-East and South-West of NSW, respectively. The strong spatial correlation after adjusting for individual and group specific covariates indicates that geographical variation of neutropenia might be due to differences in health care practices or access to care across NSW. Further investigation needed possibly at the hospital level for a comprehensive explanation of these findings. The lower spatial random effect might indicate a low number of cancer patients recruited in our study due to a border effect (i.e., getting admitted for neutropenia in other states: ACT, Victoria or Queensland).

Variation across clinical practices of neutropenia treatment has been identified in Australia in a previous survey (Lingaratnam et al., 2011). The authors showed that the treatment approach for management of neutropenia varies across oncologists, hematologists and clinicians as well as different sectors of cancer care. Therefore, it might be interesting to explore whether the observed variation is due to variation across different hospitals (eg., metropolitan hospital vs. Non-metropolitan hospitals) in NSW or across various healthcare providers. However, data items for such analysis are not collected in the registry and are beyond the scope of our present paper.

Our study was based on data linked from a state-based cancer registry and administrative data from the Admitted Patient Data Collection (APDC). An advantage of such linked data is that it provides us with a large, population based sample. Registry

based analysis is more comprehensive than that based on single centre studies, and provides more complete information than may be obtained from clinical trials where patient selection and loss to follow-up may impact validity and generalizability of study findings. However, it is important to keep in mind that the resulting data quality may be inferior to that obtained from prospective studies.

Despite various limitations, indiCAR is an useful addition to the existing methodology to explore clinical variation across geographical locations. One of the major advantages of indiCAR is the ability to analyze age as a continuous variable rather than grouping them using arbitrary cut-offs. The results of such analysis are given in Appendix Table 2.5, though they are very similar to those using age groups. However, in many applications age grouping might induce residual confounding and result in spurious relationship between age and outcome variable (Rothman, Greenland, & Lash, 2008). In our simulation study, we evaluate indiCAR for a continuous area level covariate; however, to ease our interpretation we considered SEIFA as a categorical variable. The continuous SEIFA index scores are an ordinal measure, so care should be taken when comparing scores. For example, an area with a index of 1000 is not twice as advantaged as an area with a index of 500. Moreover the scores are relative measures of Social economic disadvantages, so while this type of measure is useful for considering inequality between households, it cannot provide information on absolute levels of poverty within a community (McKenzie, 2003). Therefore, for ease in interpretation it is recommended that the index rankings and quantiles (e.g. decile) should be used for analysis, rather than using the index scores. The results are quite similar and indicate a significant negative relationship between high SEIFA score and neutropenia admission (result not shown in table).

2.5 Conclusions

In this paper we propose a novel methodology to incorporate individual level covariate information in disease mapping studies. Our method provides reliable estimates of individual and area level covariate effects. The natural applicability of indiCAR in a distributed computing framework make it potential for possible implementation in the Big Data situation. The gain in speed and computation for large data set with spatial

correlation can be obtained using indiCAR in conjunction with recently developed statistical methodology for uncorrelated Big data (Enea, 2012; Lumley, 2011). Cancer registries routinely collect individual level cancer information and thus could benefit by using indiCAR to incorporate individual level information in the analysis and mapping of disease rates.

In this chapter we assumed a linear dependence of age on neutropenia admission rates, however, exploratory analysis suggests a non-linear relationship between age and neutropenia admission rates. We therefore extend indiCAR in a semiparametric mixed model formulation where effect of a continuous individual level covariate is accommodated via a semiparametric splines formulation. The next chapter presents such an extension in detail.

Appendix 2A

Implementation of PQL in Step 2

The PQL estimation procedure is a iterative procedure where at each step requires defining following working dependent variable and a weight matrix that are updated at each iteration and solve via Fisher scoring (Breslow & Clayton, 1993; Leroux et al., 1999). The detailed procedure has been illustrated elsewhere (Breslow & Clayton, 1993; Leroux et al., 1999).

The GLM adjusted dependent variable ($\boldsymbol{Y}_{c-pseudo}$) at group level is calculated as

$$\boldsymbol{Y}_{c-pseudo} = \hat{\eta}_c + (\boldsymbol{Y}_c - \hat{\mu}_c)\frac{d\hat{\eta}_c}{d\hat{\mu}_c}, \tag{2.11}$$

where, $\eta_c = g(\mu_c) = \mathrm{O}_1 + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{b}$ and $\mathrm{O}_1 = \log\{\boldsymbol{Z}^{\mathrm{T}} \exp(\mathbf{X}\widehat{\boldsymbol{\beta}})\}$ is a vector of offset with dimension M. The poisson link ($g(\mu_c) = \log\mu_c$) and variance function $V(\mu_c) = \mu_c$ are used. The covariance matrix of $\boldsymbol{Y}_{c-pseudo}$ is then approximated by

$$\widehat{V}_c = \widehat{W}_c^{-1} + \widehat{D}, \tag{2.12}$$

where $\hat{D}$ is the covariance matrix of the random effects, $\boldsymbol{b}$, evaluated at the current estimate for the variance parameters, and $\hat{W}_c$ is the $M \times M$ diagonal matrix with diagonal terms $w = \hat{\mu}_c$. Updated estimates of the fixed effect vector $\boldsymbol{\gamma}$ and random effect vector $\boldsymbol{b}$ are then can be obtained from the solution of the following mixed model equations:

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{U}^{\mathrm{T}}\widehat{V}_c^{-1}\mathbf{U})^{-1}\mathbf{U}\widehat{V}_c^{-1}(\boldsymbol{Y}_{c-pseudo} - \mathrm{O}_1), \tag{2.13}$$

and

$$\widehat{\boldsymbol{b}} = \hat{D}\widehat{V}_c^{-1}(\boldsymbol{Y}_{c-pseudo} - \mathrm{O}_1 - \mathbf{U}\widehat{\boldsymbol{\gamma}}). \tag{2.14}$$

The updated estimates of the variance parameters, λ and σ are obtain by a

Newton-Rapson iterative procedure as follows:

$$\begin{pmatrix} \widehat{\sigma} \\ \widehat{\lambda} \end{pmatrix}^{updated} = \begin{pmatrix} \widehat{\sigma} \\ \widehat{\lambda} \end{pmatrix}^{old} + \boldsymbol{I}^{-1}\mathbf{S}. \tag{2.15}$$

where $\mathbf{S}$ is the score vector and $\boldsymbol{I}$ is the expected information matrix based on REML likelihood for $\boldsymbol{Y}_{c-pseudo}$. The expression for the element of score vector and information matrix, letting $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\sigma, \lambda)$ is given by

$$\begin{aligned} S_i &= \frac{1}{2}(\boldsymbol{Y}_{c-pseudo} - \mathbf{U}\widehat{\boldsymbol{\gamma}} - \mathrm{O}_1)^{\mathrm{T}} P \frac{\delta V_c}{\delta \theta_i} P \\ &\quad (\boldsymbol{Y}_{c-pseudo} - \mathbf{U}\widehat{\boldsymbol{\gamma}} - \mathrm{O}_1) - \frac{1}{2} tr\left(P \frac{\delta V_c}{\delta \theta_i}\right) \end{aligned}$$

and

$$I_{jk} = -\frac{1}{2} tr\left(P \frac{\delta V_c}{\delta \theta_j} P \frac{\delta V_c}{\delta \theta_k}\right),$$

where, $P = V_c^{-1} - V_c^{-1}\mathbf{U}(\mathbf{U}^{\mathrm{T}} V_c^{-1}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}} V_c^{-1}$. The derivative of V_c with respect to σ and λ are given below:

$$\begin{aligned} \frac{\delta V_c}{\delta \sigma} &= 2\sigma \boldsymbol{R}_\lambda^{-1} \\ \frac{\delta V_c}{\delta \lambda} &= \sigma^2 \boldsymbol{R}_\lambda^{-1}(\boldsymbol{R} - \boldsymbol{I})\boldsymbol{R}_\lambda^{-1}, \end{aligned}$$

where $\boldsymbol{R}_\lambda = (1-\lambda)\boldsymbol{I} + \lambda \boldsymbol{R}$ and $\boldsymbol{R}$ is the intrinsic auto-regression matrix determined by neighborhood structure .

Repeated iteration of equations (2.11)-(2.15) are considered to obtain a consistent estimates of the region specific fixed effect and random effect parameters. Convergence is acheived when the change in parameter estimates are less than a prespecified tolerance level (less than $1e-3$, in the simulation study reported). Approximate standard errors for λ and σ are obtained from the above information matrix in the usual way.

Additional Tables

Table 2.5: Application of indiCAR with age as a continuous predictor

Regression coefficients	Estimates	Std. Error
Intercept	-1.493	0.047
Age	-0.027	0.001
Sex		
Female	Ref	
Male	-0.043	0.020
Year of Diagnosis		
2001	Ref	
2002	0.021	0.038
2003	0.083	0.038
2004	0.019	0.038
2005	0.095	0.037
2006	0.038	0.038
2007	-0.022	0.038
2008	-0.004	0.038
2009	-0.315	0.042
ARIA		
Major Cities	Ref	
Inner Regional Australia	-0.006	0.022
Outer Regional Australia	-0.118	0.037
Remote or Very remote Australia	-0.192	0.142
Cancer Type		
Breast Cancer	Ref	
Lung cancer	0.240	0.037
Colon & rectum cancer	-0.463	0.040
Haematological Malignancy	1.497	0.029
Other cancer	-0.986	0.031
No. of major comorbidities		
0	Ref	
1	0.413	0.026
2	0.682	0.026
3	0.589	0.036
4+	0.594	0.035
SEIFA		
Most disadvantaged	Ref	
2	-0.089	0.043
3	-0.078	0.038
4	-0.134	0.045
Least disadvantaged	-0.144	0.053
Variance parameter		
σ	0.209	0.022
λ	0.992	0.012

R Code

Function for data generation

```
library(MASS)
DataSimulation<-function(nsamp=20000,nGroup=400,sigma=0.4,lambda=0.75,
        beta0=-0.2, beta1=-2.5, beta2=0.7, beta32=-2.0, beta33=-1.5,
        beta34=0.2, beta35=0.5, beta36=0.8, gamma=0.2)
{
#         nGroup: Total number of group
#         nSamp: sample size
#          lambda:
#          sigma: variance parameter
#          beta0-beta3: individual level coefficients
#          gamma: group level coefficient
#   ID: Area ID e.g., postcode
repeat{
  indivGroup<-as.vector(rmultinom(n=1, size=nsamp, prob=runif(400,0.05)))
  if (all(indivGroup)>=1){break}
 }
totalSample<-sum(indivGroup)
#generate covariate values
pr =runif(totalSample,0.45,0.55)
x2<-rbinom(totalSample,1,pr)
x1<-0.2*x2+runif(totalSample,0.2,2)
x3<-sample(1:6,totalSample, replace=TRUE,
                        prob=c(0.06,0.09,0.19,0.25,0.25,0.16))
#generate covariate values
z2<-rnorm(nGroup)
x4<-rep(z2,indivGroup)
ID<-rep(1:nGroup,indivGroup)
#Generating spatial random effect in a lattice grid
#### Set up a square lattice region
x.easting <- 1:20
x.northing <- 1:20
Grid <- expand.grid(x.easting, x.northing)
n <- nrow(Grid)
#### set up distance and neighbourhood  matrices W
distance <-array(0, c(n,n))
W <-array(0, c(n,n))
for(i in 1:n)
 {
 for(j in 1:n)
        {
        temp <- (Grid[i,1] - Grid[j,1])^2 + (Grid[i,2] - Grid[j,2])^2
        distance[i,j] <- sqrt(temp)
        if(temp==1)  W[i,j] <- 1
        }
 }
R <- -W
diag(R) <- as.numeric(apply(W, 1, sum))
Sigma.inv<-1/sigma^2*(lambda*R + (1-lambda)*diag(rep(1,n)))
Sigma<-solve(Sigma.inv)
phi <- mvrnorm(n=1, mu=rep(0,n), Sigma=Sigma)
phi_long<-rep(phi,indivGroup)
mu <- exp(beta0 +beta1 *x1 +beta2*x2+beta32*I(x3==2)+beta33*I(x3==3) +
 beta34*I(x3==4)+beta35*I(x3==5)+beta36*I(x3==6)+ gamma*x4 + phi_long)
#generate Y-values
y <- rpois(totalSample, lambda=mu)
#data set
data <- data.frame(ID,y=y, x1,x2,x3,x4)
output<-list(data=data,R=R)
output
}
```

Necessary functions for implementing the proposed method

```
# Load required library
library(lme4)
library(plyr)
library(reshape)
library(parallel)
# Load data set and neighbourhood matrix
DataSim<-DataSimulation()
sampleData1<-DataSim$data
R <- DataSim$R
#Fit generalized linear model with individual level covariates
fit_ind<-glm(y~x1+x2+as.factor(x3),data=sampleData1,family="poisson")
# Extract model matrix and coefficient
frame_ind <- model.frame(fit_ind)
Y <- model.response(frame_ind)
X_ind <-model.matrix(object = attr(frame_ind,"terms"), data = frame_ind)
n_ind<-nrow(X_ind)
beta <- fit_ind$coefficient
#Calculate the fitted value
sampleData1$Predict<-exp(X_ind%*%beta)
#head(sampleData1)
#Grouplevel data: Location data
locationData<-sampleData1[,c("ID","x4")]
GroupCovData<-aggregate(.~ ID, data = locationData, mean)
#head(GroupCovData)
#Aggregate data Over postcode
AreaData<-sampleData1[,c("y","ID","Predict")]
aggdata <-aggregate(. ~ ID, data = AreaData, sum)
#head(aggdata)
nGroup<-nrow(aggdata)
aggdataU<-aggdata[!duplicated(aggdata$ID),]
GroupData<-merge(GroupCovData,aggdataU,by="ID")
names(GroupData)<-c("ID","x4","totalY","totalePredict")
#head(GroupData)
#Fitting PQL model
gamma.iter<-NULL
theta.iter<-NULL
beta.iter<-NULL
#Set initial values
gamma.hat<-0
b.rand<-rep(0,nGroup)
sigma.hat <-0.5
lambda.hat <-0.5
theta.hat<-c(sigma.hat,lambda.hat)
#Generate covariance matrix
R <- DataSim$R
betaCombined<-c(beta,gamma.hat)
X_group <-as.matrix(GroupData$x4)
Y<-as.matrix(GroupData$totalY)
OffSet<-as.matrix(log(GroupData$totalePredict))
repeat{
        repeat{
# Estimate covariance matrices
R.lambda <- lambda.hat*R + (1-lambda.hat)*diag(rep(1,nGroup))
R.lambda.inv<-solve(R.lambda)
D.hat.inv<-1/(sigma.hat^2)*R.lambda
D.hat<-solve(D.hat.inv)
# Calculate the PQL elements
                repeat{
eta.est<-OffSet+X_group%*%gamma.hat+b.rand
mu.est<-exp(eta.est)
zz.est<-eta.est+(Y-mu.est)/mu.est
W.hat<-diag(as.vector(mu.est))
W.hat.inv<-diag(as.vector(1/mu.est))
#Estimate covariance matrix of Y
V.hat<-W.hat.inv+D.hat
V.hat.inv<-solve(V.hat)
#Estimate the fixed and random effect
gammaUpdate<-solve(t(X_group)%*%V.hat.inv%*%X_group)%*%(t(X_group)%*%
                V.hat.inv%*%(zz.est-OffSet))
```

```
b.rand.update<-D.hat%*%V.hat.inv%*%(zz.est-OffSet-X_group%*%gammaUpdate)
diff<-abs(gammaUpdate-gamma.hat)
diff.rand<-abs(b.rand.update-b.rand)
gamma.iter<-rbind(gamma.iter,gammaUpdate)
if (all(diff< 1e-5)){break}
if (all(diff.rand< 1e-3)){break}
gamma.hat<-gammaUpdate
b.rand<-b.rand.update
                        }
# Extract score and observed information matrix
P<-V.hat.inv-(V.hat.inv%*%X_group%*%solve(t(X_group)%*%V.hat.inv
                        %*%X_group)%*%t(X_group)%*%V.hat.inv)
dV.sigma<-2*sigma.hat*R.lambda.inv
dV.lambda<--1*sigma.hat^2*R.lambda.inv%*%(R-diag(rep(1,nGroup)))
                                %*%R.lambda.inv
#Score vector
score.sigma<-0.5*(t(zz.est-OffSet-X_group%*%gammaUpdate)%*%P)%*%
                dV.sigma%*%(P%*%(zz.est-OffSet-X_group%*%gammaUpdate))-
                0.5*sum(diag(P%*%dV.sigma))
score.lambda<-0.5*(t(zz.est-OffSet-X_group%*%gammaUpdate)%*%P)%*%
                dV.lambda%*%(P%*%(zz.est-OffSet-X_group%*%gammaUpdate))-
                0.5*sum(diag(P%*%dV.lambda))
score<-c(score.sigma,score.lambda)
exp.inf11<-0.5*sum(diag(P%*%dV.sigma%*%P%*%dV.sigma))
exp.inf12<-0.5*sum(diag(P%*%dV.sigma%*%P%*%dV.lambda))
exp.inf21<-0.5*sum(diag(P%*%dV.lambda%*%P%*%dV.sigma))
exp.inf22<-0.5*sum(diag(P%*%dV.lambda%*%P%*%dV.lambda))
exp.infor<-matrix(c(exp.inf11,exp.inf12,exp.inf21,exp.inf22),ncol=2)
thetaUpdate<-theta.hat+solve(exp.infor)%*%score
if (thetaUpdate[2]>1) {thetaUpdate[2]<-0.99999}
if (thetaUpdate[2]<0) {thetaUpdate[2]<-0.00001}
if (thetaUpdate[1]<0) {thetaUpdate[1]<-0.1}
sigma.hat<-thetaUpdate[1]
lambda.hat<-thetaUpdate[2]
diff.theta<-abs(thetaUpdate-theta.hat)
theta.iter<-rbind(theta.iter,as.vector(thetaUpdate))
if (all(diff.theta< 1e-5)){break}
theta.hat<-thetaUpdate
        }
# Repeat individual level data fitting
GroupData$Predict_group<-X_group%*%gamma.hat+b.rand
combinedData<-merge(sampleData1,GroupData[,c("ID","Predict_group")]
                        ,by=c("ID"))
formula_ind=y~offset(Predict_group)+x1+x2+as.factor(x3)
fit_ind<-glm(formula_ind,data=combinedData,family="poisson")
#Calculate the fitted value
betaUpdate<-fit_ind$coefficient
#Calculate the fitted value
combinedData$Predict<-exp(X_ind%*%betaUpdate)
AreaData<-combinedData[,c("y","ID","Predict")]
aggdata <-aggregate(. ~ ID, data = AreaData, sum)
aggdataU<-aggdata[!duplicated(aggdata$ID),]
GroupData<-merge(GroupCovData,aggdataU,by="ID")
names(GroupData)<-c("ID","x4","totalY","totalePredict")
Y<-as.matrix(GroupData$totalY)
OffSet<-as.matrix(log(GroupData$totalePredict))
betaCombined.Update<-c(betaUpdate,gamma.hat)
diff.est<-abs(betaCombined.Update-betaCombined)
beta.iter<-rbind(beta.iter,betaCombined.Update)
if (all(diff.est< 1e-5)) {break}
betaCombined<-betaCombined.Update
# End of indiCAR
}
#Estimate corrected standard error
M<-nGroup
group.size<-as.vector(table(combinedData$ID))
group.cov.long<-data.matrix(X_group[rep(1:nrow(X_group),
                                times = group.size),])
cov.combined<-cbind(X_ind,group.cov.long)
fitted.combined<-cov.combined%*%betaCombined+rep(b.rand,group.size)
mu.combined<-as.vector(exp(fitted.combined))
Xcov.m<-X_ind*mu.combined
```

```
XTWX<-t(Xcov.m)%*%X_ind
groupID<-rep(c(1:M),group.size)
XTWZ<-t(rowsum(Xcov.m, groupID))
XTWZD<-XTWZ%*%D.hat
ZTWZ.vec<-aggregate(mu.combined,by=list(groupID),sum)
ZTWZ<-diag(ZTWZ.vec$x)
ZTWZD<-ZTWZ%*%D.hat
I.ZTWZD<-(diag(M)+ZTWZD)
I.ZTWZD.inv<-solve(I.ZTWZD)
a11<-XTWX-XTWZD%*%I.ZTWZD.inv%*%t(XTWZ)
XTWZU<-XTWZ%*%X_group
ZTWZU<-ZTWZ%*%X_group
a12<-XTWZU-XTWZD%*%I.ZTWZD.inv%*%ZTWZU
a21<-t(a12)
a22<-t(X_group)%*%ZTWZ%*%X_group-t(X_group)%*%ZTWZ%*%D.hat%*%
        I.ZTWZD.inv%*%ZTWZ%*%X_group
Q.inv<-as.matrix(rbind(cbind(a11,a12),cbind(a21,a22)))
Q<-solve(Q.inv)
se.coef<-sqrt(diag(Q))
se.theta<-sqrt(diag(solve(exp.infor)))
coef<-c(betaCombined,as.vector(theta.hat))
names(coef)<-c("(Intercept)","x1","x2","as.factor(x3)2","as.factor(x3)3",
"as.factor(x3)4","as.factor(x3)5","as.factor(x3)6", "U","sigma","lambda")
se.coef<-c(se.coef,se.theta)
names(se.coef)<-c("(Intercept)","x1","x2","as.factor(x3)2",
        "as.factor(x3)3","as.factor(x3)4","as.factor(x3)5",
        "as.factor(x3)6","U","sigma","lambda" )
```

Chapter 3

Smooth Individual Level Covariates in Conditional Auto-Regressive (smooth-indiCAR) Model for Disease Mapping.

Summary

Conditional Auto-Regressive (CAR) models have been extensively used in disease mapping studies. However available implementations of such models only incorporate area level covariates to help explain spatial variation in disease rates. In many epidemiological study settings, individual level covariates also have considerable impact on the outcome of interest. Therefore, spatial models for disease mapping should ideally account for covariates measured at both individual and area levels.

In the current study, we propose a novel conditional auto-regressive model that can incorporate both individual and group level covariates while adjusting for spatial

correlation in the disease rates. In this formulation the effect of a continuous individual level covariate is accommodated via semi-parametric splines. We describe a two-step estimation procedure to obtain reliable estimates of individual and group level covariate effects. We evaluate the performance of our smooth-indiCAR method through simulation studies. Our results indicate that the smooth-indiCAR method provides reliable estimates of all regression and random effect parameters. We also apply smooth-indiCAR to the analysis of data on neutropenia admissions in New South Wales (NSW), Australia.

Incorporating individual covariate data in disease mapping studies improves the estimates of fixed and random effect parameters by utilizing information from multiple sources. Health registries routinely collect individual and area level information and thus could benefit by using smooth-indiCAR to incorporate individual level information in the analysis and mapping of disease rates. Moreover, the natural applicability of smooth-indiCAR in a distributed computing framework enhances its application in the Big Data domain with a large number of individual/group level covariates.

3.1 Introduction

Rapid growth of Geographic Information Systems (GIS), together with advances in high performance computing environments, presents a unique opportunity to examine the relationship between risk factors and outcomes that vary across geographical locations. Careful analysis of spatial data can lead to useful explanation of the exposure and disease relationship through natural experimentation (Rothman et al., 2008; Snow, 1855). It also helps in understanding spatial variation of disease, disease clustering, distribution of socio-demographic structure, environmental exposure distribution and its impact on health outcomes (Elliott & Wartenberg, 2004).

Analysis of spatially indexed data is complicated by correlations among neighboring observations (Besag et al., 1991; Clayton & Kaldor, 1987; Cressie, 1993). Regression analysis ignoring this spatial correlation leads to incorrect inference on the estimated regression coefficients by narrowing of associated confidence intervals (Waller & Gotway, 2004). Mixed effects models provide a convenient way of adjusting for spatial correlation by incorporating spatially defined Conditionally Auto-Regressive (CAR) random effects

in the model (Breslow & Clayton, 1993; Leroux et al., 1999). The use of this model allows one to map disease rates by borrowing information about each small area from its surrounding areas, thus stabilizing estimation based on the sparse data for small areas. Use of the CAR-based structure within a hierarchical generalized linear model offers a robust, flexible, and enormously popular class of models for the exploration and analysis of small area rates for disease mapping. However, a lack of modeling strategies for individual level covariates is a limitation of existing software which may lead to ecological bias (Wakefield, 2007).

In the previous chapter we propose an individual level covariate adjusted CAR (indiCAR) model that can incorporate both individual and area level covariates. In such a formulation individual and group level data were fitted in separate steps of an iterative process. Although this approach is very useful in modeling large number of individual and group level covariate effects, it relies on an assumption of log-linear dependence between the outcome and covariates. In many epidemiological study settings, such log-linear dependence may not be applicable, rather other types of non-linearity may be in operation. Spline based techniques are appealing to model the effect of covariates in a flexible non-linear fashion (Wakefield, 2007). In the case study that motivates this chapter, researchers from the NSW Cancer Institute explored the use of the indiCAR method to model the geographical variation of neutropenia infection across New South Wales, Australia. However, exploratory analysis reveals that the age of the patient exhibits a non-linear association with the observed neutropenia rate. The non-linear age effect has been also noted in many previous studies (MacNab, 2004; Rosenbaum & Rubin, 1984).

Therefore, in our current study we extend the indiCAR method to a semi-parametric mixed model context. Following indiCAR, we incorporate individual level smooth covariate information in a two step iterative procedure following an initialization step. In this method, the individual level and the group level covariate effects are fitted in separate iterations with existing software by appropriate calculation of an offset at each step. We illustrate that the estimation and inference based on smooth-indiCAR can be carried out in a distributed computing framework, thus achieving a helpful reduction in computational cost and memory requirements.

We evaluate the performance of the smooth-indiCAR method through simulation studies. Our results show that the smooth-indiCAR is able to correctly estimate coefficients associated with both individual and group-level covariates. We further illustrate this method through the analysis of data on neutropenia admissions from the New South Wales (NSW) Cancer Institute and conclude with some practical guidelines.

The structure of this chapter is as follows: Section 3.2 describes the data and model formulation, estimation and inference procedures. Section 3.3 presents the data generation process for our simulations. In section 3.4 we present results from the simulation study and an application of the proposed method to data on neutropenia. We conclude with general discussion in section 3.5.

3.2 Methodology

3.2.1 Data

NSW cancer registries were used to identify patients diagnosed with cancer, associated treatment procedures and co-morbidities. Specifically, we used NSW Central Cancer Registry (CCR) linked to NSW Admitted Patient Data Collection (APDC). Detailed descriptions of the data items can be obtained from the Centre for Health Record Linkage (CHeReL http://www.cherel.org.au/master- linkage -key). Data were checked for consistency across data sources and linked by assigning a unique Project Person Number (PPN) to each patient. Our study population comprises all cancer patients that were diagnosed with cancer and were hospitalized during the period between 2001 and 2009.

Demographic variables including age at diagnosis, gender, residence at diagnosis, postal area of residence, and Accessibility/Remoteness Index of Australia (ARIA) based on patient residence were obtained from the CCR database. The ARIA variable was recorded at individual level rather than postal area level because the ARIA index varies within postal area. The Socio Economic Index For Areas (SEIFA; an index of social disadvantage) and the geo-coded shape files for mapping corresponding to 2006 census postal areas were obtained from Australian Bureau of Statistics (ABS). Individual level

clinical characteristics such as type of cancer were also obtained from CCR. The diagnosis of neutropenia admission and co-morbidity were obtained using data from the Admitted Patients Data Collection (APDC). The ICD-10-AM (International Statistical Classification of Disease and Related Health problem, 10th revision, Australian modification) code D70 (Agranulocytosis) was used to identify admissions with possible neutropenia.

3.2.2 Statistical Model

Suppose the total area under study is divided into M contiguous regions and the number of outcomes for the i^{th} $(i = 1, 2, ..., n_j)$ individual in the j^{th} $(j = 1, 2, ..., M)$ area is denoted by $\{y_{ij}\}$. Let $\boldsymbol{Y}$ be a vector with elements $\{y_{ij}\}$ that represents the number of events for all individual in the study regions of interest. Similarly, let $\mathbf{X} = (X_1, X_2, ..., X_p)$ and $\mathbf{U} = (U_1, U_2, ..., U_q)$ represent individual and area level covariate matrices with dimensions $n \times p$ and $M \times q$, respectively, where n is the total sample size i.e., $n = \sum_{j=1}^{M} n_j$. Further suppose that in addition to the log-linear relationship of $\mathbf{X}$ and $\mathbf{U}$ with $\boldsymbol{Y}$, an additional individual level covariate, T exhibits a non-linear relationship with $\boldsymbol{Y}$. Under the above specifications, conditional on the area specific random effect vector, $\boldsymbol{b}$, the number of events for each cancer patient is assumed to be Poisson distributed with mean $\boldsymbol{\mu}$ where

$$\ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + f(T) + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}. \tag{3.1}$$

Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients associated with the individual level covariates, $f(T)$ is an unknown smooth function, $\boldsymbol{\gamma}$ is a $q \times 1$ vector of area-specific regression coefficients and $\boldsymbol{Z} = \text{blockdiag}(Z_1, Z_2, ..., Z_M)$ is a replication matrix that replicates group level covariate and random effect to the individual level, where Z_j is a vector of length n_j with all elements equal to 1. We further assume that the unknown smooth function $f(T)$ can be represented by a linear combination of spline basis functions, i.e., $f(T) = B^{\mathrm{T}}(T)\eta$. Here $B(T)$ is a vector of spline basis functions and η is a vector of corresponding basis coefficients.

Note that the proposed model (3.1) represents various study designs, such as clustered, hierarchical and spatial designs depending on the specification of the random effect $\boldsymbol{b}$.

For example, the random effect may represent specification for a) random slope and intercept for the multilevel (hierarchical) models (Gelman, 2007), b) random intercept and stochastic process as of longitudinal studies (Zhang, Lin, Raz, & Sowers, 1998) and c) modeling spatial correlation in disease mapping (Leroux et al., 1999). Leroux et al. (1999) includes group level smooth predictor effect and group specific random effects but does not incorporate individual level covariates effect. Throughout this paper we will focus on modeling random effects so that they reflect spatial correlation. Our postulated model (3.1) is an extension of Lin and Zhang (1999) that incorporates individual level predictors and area specific conditional auto-regressive random effects in the context of disease mapping studies.

To fit model (3.1), many different choices of random effects, $\boldsymbol{b}$ are available in the mapping literature (see Lee (2011), for a recent review). Among these, the method of Leroux et al. (1999) is appealing because it allows for a weighted combination of spatially structured and unstructured area-level variation. Within this framework, the random effect vector, $\boldsymbol{b}$ has a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix, $\boldsymbol{D}$ with Moore-Penrose generalized inverse, $\boldsymbol{D}^{-} = \sigma^{-2}\{(1-\lambda)\boldsymbol{I} + \lambda\boldsymbol{R}\}$, where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{R}$ is the intrinsic auto-regression matrix reflecting neighborhood structure. Typically, neighbors are those areas which share a common boundary. The typical element of $\boldsymbol{R}$ is given by

$$\boldsymbol{R}_{jj'} = \begin{cases} n_j, & j = j' \\ -I\{j \sim j'\} & j \neq j', \end{cases}$$

where, n_j is the number of neighbors of region j, and $I\{j \sim j'\}$ is an indicator function that takes value 1, if regions j and j' are neighbors, 0, otherwise. Alternatively, a distance based neighborhood structure could be used (Earnest et al., 2007). The parameters characterizing the random effect distribution, $\boldsymbol{\theta} = (\sigma^2 > 0, \lambda \in [0,1])$ quantify over-dispersion and spatial dependence. A larger value of $\lambda \in [0,1]$ indicates a higher degree of spatial dependence. This specification results in two extreme cases: i) completely independent random effects when $\lambda = 0$ and ii) the intrinsic auto-regressive model when $\lambda = 1$ (Besag et al., 1991). In general, a combination of the two is assumed.

Now consider the case without any non-linear predictor in the model (3.1), inference about $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and θ can be made by integrating out or averaging over the distribution of

the unobserved random effects, $\boldsymbol{b}$. The corresponding integrated quasi-likelihood function is equal to (see equation (2) of Breslow and Clayton Breslow and Clayton (1993))

$$|D|^{-\frac{1}{2}} \int \exp\left[-\frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{n_j} d_{ij}(Y_{ij},\mu_{ij}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}} D^{-}\boldsymbol{b}\right] d\boldsymbol{b},$$

where $d(Y,\boldsymbol{\mu})$ is the deviance residual.

The maximum quasi-likelihood estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$ are those values of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ that maximize the above quasi-likelihood. However, no simple closed form solution exists. Instead, Breslow and Clayton (1993) proposed the penalized quasi-likelihood (PQL) approach for parameter estimation and inference. The PQL uses the Laplace method for integral approximation and jointly maximize the above quasi-likelihood function to obtain estimates for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{b}(\boldsymbol{\theta})$.

In the presence of a non-linear predictor, however, statistical inference about $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ must account for the non-parametric function, $f(T)$ which requires estimation of the basis coefficient, η and smoothing parameter δ. Lin and Zhang (1999) showed that approximate estimates of the regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and η can be obtained by maximizing the following Double Penalized Quasi-Likelihood equation with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{b}$ and η:

$$-\frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{n_j} d_{ij}(Y_{ij},\mu_{ij}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}} D^{-}\boldsymbol{b} - \frac{1}{2}\eta^{\mathrm{T}}\mathbf{S}\eta, \tag{3.2}$$

where $\mathbf{S} = \delta\boldsymbol{K}$ with smoothing parameter δ and penalty matrix $\boldsymbol{K}$. Here, $\boldsymbol{K}$ is a $(q+N)\times(q+N)$ matrix where q is the number of knots and N is the dimension of the unpenalized function. Given the knot locations $\{x^{*}_{(k)} : k = 1,2,...,q\}$, the penalty matrix have zeroes everywhere except in its lower right $q \times q$ block with $\boldsymbol{K}_{(ik)} = \|x^{*}_{j(i)} - x^{*}_{j(k)}\|^{3}$, for $k \leq q$. The penalty matrices map the spline basis functions to the data whereas the penalty parameters control the amount of smoothing (Ruppert et al., 2003; S. Wood, 2006). For now, assume that the smoothing parameter δ is known.

Under the above specification the approximate log likelihood can be expressed as

$$const + \boldsymbol{Y}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}) - \mathbf{1}^{\mathrm{T}}\exp(\mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^{\mathrm{T}}D^{-}\boldsymbol{b} - \frac{1}{2}\delta\eta^{\mathrm{T}}\mathbf{S}\eta. \tag{3.3}$$

Differentiating (3.3) with respect to $\boldsymbol{\beta}$, η, $\boldsymbol{\gamma}$ and $\boldsymbol{b}$ using vector matrix calculus (Wand, 2002), we obtain the following score equations

$$\left\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\right\}^{\mathrm{T}}\mathbf{X} = 0, \tag{3.4}$$

$$\left\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\right\}^{\mathrm{T}}B(T) = \eta^{\mathrm{T}}\mathbf{S}, \tag{3.5}$$

$$\left\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\right\}^{\mathrm{T}}\boldsymbol{Z}\mathbf{U} = 0, \tag{3.6}$$

and

$$\left\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b})\right\}^{\mathrm{T}}\boldsymbol{Z} = \boldsymbol{b}^{\mathrm{T}}D^{-}. \tag{3.7}$$

Penalized Iteratively Re-weighted Least Squares (P-IRLS) can be applied to solve the above equations for $\boldsymbol{\beta}$, η, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ (S. Wood, 2006). However, high computational costs and memory space constraints often make it difficult to apply these iterative procedures to data sets with large number of group level covariates and large sample size. An alternative computational strategy is the use of the Gauss-Seidel algorithm to obtain the same estimate as of P-IRLS of the associated parameters (Guha, Ryan, & Morara, 2009). In this approach, at each iteration one of the parameters is estimated while keeping others fixed at current values. Within this framework, we first initialize $\boldsymbol{\beta}$ and η and then obtain updated estimates for $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ in the following two step procedure:

Step 0: Set the coefficients of area level covariates, $\boldsymbol{\gamma}$ and random effects, $\boldsymbol{b}$ to zero in equation (3.4) and (3.5). Then we have

$$\left\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta)\right\}^{\mathrm{T}}\mathbf{X} = 0,$$

and,

$$\left\{\boldsymbol{Y} - \exp(\mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta)\right\}^{\mathrm{T}} B(T) = \eta^{\mathrm{T}}\mathbf{S}.$$

If the value of penalty parameter δ is known, the solution of the above equations can be computed by a penalized version of the iterative re-weighted least square method used for GLM estimation (Green, 1987; S. Wood, 2006). The smoothing parameter can be estimated using the Generalized Cross Validation score (GCV) or the generalized Akaike's Information Criterion (S. Wood, 2006). Computationally, this can be done using existing *gam* function in the *mgcv* package (S. Wood, 2006) in ***R*** (R Core Team, 2013). Thus we can obtain an estimate of the regression coefficients $\boldsymbol{\beta}$ and η associated with individual level covariates and the penalty parameter δ. This step provides initial estimates of the regression coefficients $\boldsymbol{\beta}$ and η.

Step 1. Now substitute the current estimated individual level coefficients, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}$ into equations (3.6) and (3.7). With some simple algebra, we have

$$\{\boldsymbol{Y}_c - \exp(\mathrm{O}_1 + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{b})\}^{\mathrm{T}}\mathbf{U} = 0$$

and,

$$\{\boldsymbol{Y}_c - \exp(\mathrm{O}_1 + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{b})\}^{\mathrm{T}} = \boldsymbol{b}^{\mathrm{T}}D^{-},$$

where, $\boldsymbol{Y}_c^{\mathrm{T}} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Z}$, is a vector of aggregated outcome counts of length M at the group level and $\mathrm{O}_1 = \log\{\boldsymbol{Z}^{\mathrm{T}}\exp(\mathbf{X}\widehat{\boldsymbol{\beta}} + B^{\mathrm{T}}(T)\widehat{\eta})\}$ is a vector of offsets.

The above two equations are well known PQL estimating equations associated with the Poisson mixed model (Breslow & Clayton, 1993). Since, the outcome $\boldsymbol{Y}_c$, offset O_1, covariate $\mathbf{U}$ and random effects $\boldsymbol{b}$ are all available at the group level, estimates of parameters for the group level coefficient $\widehat{\boldsymbol{\gamma}}$ and random effects $\boldsymbol{\theta}$ can be estimated using the existing PQL method (Breslow & Clayton, 1993; Leroux et al., 1999) with only group level data.

Step 2. Substitute the estimated area-specific regression coefficient, $\widehat{\boldsymbol{\gamma}}$ and random effect

parameter, $\widehat{\boldsymbol{\theta}}$ estimated at Step 1 into (3.4) & (3.5). With some simple algebra, we have

$$\left\{\boldsymbol{Y} - \exp(\mathrm{O}_2 + \mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta)\right\}^{\mathrm{T}} \mathbf{X} = 0,$$

and

$$\left\{\boldsymbol{Y} - \exp(\mathrm{O}_2 + \mathbf{X}\boldsymbol{\beta} + B^{\mathrm{T}}(T)\eta)\right\}^{\mathrm{T}} B(T) = \eta^{\mathrm{T}}\mathbf{S},$$

where $\mathrm{O}_2 = \boldsymbol{Z}(\mathbf{U}\widehat{\gamma} + \widehat{\boldsymbol{b}})$ is a offset vector of dimension $n \times 1$. Under the above specification, the individual level coefficients estimate $\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}$ and smoothing parameter can then be updated using *gam* function with individual level data.

Step 1 and Step 2 are then repeated until the algorithm converges. The estimated coefficients and parameters in random effect obtained by this iterative procedure will be similar to the estimates of the true regression coefficients and parameters in the random effect that would be obtained by solving equations (3.3) - (3.7) directly.

Approximate Standard Error

Model fitting at Step 1 and Step 2 assumes fixed $(\boldsymbol{\beta}, \eta)$ and fixed $\boldsymbol{\gamma}$, respectively. Therefore the corresponding standard errors of $(\boldsymbol{\beta}, \eta)$ from *gam* and $(\boldsymbol{\gamma}, \boldsymbol{\theta})$ from the PQL method based on Step 2 and Step 1 will not be exactly correct. We re-calculate the standard error of these regression coefficients by adjusting the estimated standard errors of $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\eta}$. This can be done via the Iterative Re-weighted Least Squares (IRLS) estimation based on score equations (3.4, 3.5, 3.6 and 3.7) for a known smoothing parameter, δ (Lin & Zhang, 1999). The IRLS estimation requires us to define a working dependent variable and a weight matrix that are updated at each iteration and solved via Fisher scoring (Breslow & Clayton, 1993; S. Wood, 2006).

Now assume that an unpenalized linear combination of basis functions is adequate to represent the nonlinear function $f(T)$. In this case the linear combination of the basis function contributes to the log-likelihood equation (3.3) via the fixed effect components only. Therefore, the corresponding GLM adjusted dependent variable, $\boldsymbol{Y}_{pseudo}$ can be

obtained as

$$\boldsymbol{Y}_{pseudo} = \mathbf{X}\boldsymbol{\beta} + B(T)\eta + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b} + \mathbf{W}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \tag{3.8}$$

where $\mathbf{W}$ is a $n \times n$ diagonal matrix with diagonal term $\boldsymbol{\mu}$. Following Harville (1977) and Robinson (1991), it can be shown that the Fisher scoring corresponding to the score equations (3.4, 3.5, 3.6 and 3.7) and GLM dependent variable as in (3.8), is identical to the normal equation of the best linear unbiased predictors (BLUPs) of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, η and $\boldsymbol{\theta}$ corresponding to the following linear mixed model

$$\boldsymbol{Y}_{pseudo} = \mathbf{X}\boldsymbol{\beta} + B(T)\eta + \boldsymbol{Z}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}_{pseudo},$$

where the pseudo-error $\boldsymbol{\epsilon}_{pseudo} \sim N(0, \mathbf{W}^{-1})$. The estimated regression coefficients for the fixed effect, $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \eta)$ and BLUP estimate for the random effect $\boldsymbol{b}$ can be obtained as (Robinson, 1991)

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\eta}) = (\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{C})^{-1}(\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}_{pseudo})$$

and

$$\widehat{\boldsymbol{b}} = \boldsymbol{D}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\{\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\mathbf{U}\widehat{\boldsymbol{\gamma}} - B(T)\widehat{\eta}\}, \tag{3.9}$$

where $\mathbf{C} = [X|ZU|B(T)]$ is a design matrix consisting of the individual level covariate matrix, group level covariates and basis functions, and $\boldsymbol{V} = \boldsymbol{Z}D\boldsymbol{Z}^{\mathrm{T}} + \mathbf{W}^{-1}$ is the variance of pseudo response $\boldsymbol{Y}_{pseudo}$.

Now consider the fact that the nonlinear function $f(T)$ is represented using splines regression bases, with associated roughness penalties in the log-likelihood equation (3.3). Following S. Wood (2006) and Marra and Wood (2012), it can be shown that the maximum penalized likelihood estimate, ($\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$ and $\widehat{\eta}$) can be obtained as

$$\begin{aligned}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\eta}) &= (\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{C} + \mathbf{S}_1)^{-1}(\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}_{pseudo}) \\ \widehat{\boldsymbol{b}} &= \boldsymbol{D}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\{\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\mathbf{U}\widehat{\boldsymbol{\gamma}} - B(T)\widehat{\eta}\},\end{aligned} \tag{3.10}$$

where $\mathbf{S}_1$ is the smooth matrix consisting of 0s except in the block corresponding to the basis coefficients η, where it is replaced by smoothing matrix $\mathbf{S}$. Thus, the frequentist

variance-covariance matrix for the fixed effect ($\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$ and $\widehat{\eta}$) can be estimated by

$$\boldsymbol{Q}_{Freq} = (\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{C} + \mathbf{S}_1)^{-1}\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{C}(\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{C} + \mathbf{S}_1)^{-1}. \tag{3.11}$$

Following the normality of $\boldsymbol{Y}_{psudo}$, S. Wood (2006) showed that $\hat{\eta} \sim N(E(\hat{\eta}), \boldsymbol{Q}_{freq})$. However, $E(\hat{\eta}) = \eta$ if $\eta = 0$only, therefore in general $E(\hat{\eta}) \neq \eta$. Hence the above estimated standard error for non-parametric function is only useful when testing model terms equal to zero. S. Wood (2006) further suggests the use of an alternative Bayesian approach to calculate uncertainty, which results in a Bayesian posterior covariance matrix for the parameters as

$$\boldsymbol{Q}_{Bayes} = (\mathbf{C}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{C} + \mathbf{S}_1)^{-1}. \tag{3.12}$$

Note that the frequentist and the Bayesian estimates of standard error differ in the inference about basis coefficients, but virtually are identical for linear individual and group level covariate effects. Further note that reliable estimates of the regression coefficients and variance components can also be obtained using the model of Lin and Zhang (1999) with appropriate specification of the design matrix ($\boldsymbol{Z}$) associated with the spatial random effect model (3.1). However, their formulation requires representation of smooth terms as a linear combination of fixed and random effect covariates. Back-fitting approaches such as the smooth-indiCAR method calculate the tuning parameters at Step 1 and will be effective in situations where memory constraints prohibit the fitting of a single model consisting of a large number of individual and group level covariates. Smooth-indiCAR not only provides a convenient way of fitting large number of individual and group level covariates in a distributed computing framework, it also allows us to calculate the standard error in a distributed computing framework. This is because $\boldsymbol{V}^{-1}$ can be expressed as $\mathbf{W} - \mathbf{W}\boldsymbol{Z}\boldsymbol{D}(I + \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D})^{-1}\boldsymbol{Z}^{\mathrm{T}}\mathbf{W}$ (Henderson & Searle, 1981). Therefore, the above Bayesian variance-covariance matrix can be written as

$$\boldsymbol{Q}_{Bayes} = \left(\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \mathbf{S}_1 \right)^{-1},$$

where,

$$
\begin{aligned}
a_{11} &= \bar{\mathbf{X}}^{\mathrm{T}}\mathbf{W}\bar{\mathbf{X}} - \bar{\mathbf{X}}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D}(\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D})^{-1} \times \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\bar{\mathbf{X}} \\
a_{12} &= \bar{\mathbf{X}}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\mathbf{U} - \bar{\mathbf{X}}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D}(\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D})^{-1} \times \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\mathbf{U} \\
a_{21} &= a_{12}^{\mathrm{T}} \\
a_{22} &= \mathbf{U}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\mathbf{U} - \mathbf{U}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D} \times (\boldsymbol{I} + \boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\boldsymbol{D})^{-1}\boldsymbol{Z}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}\mathbf{U},
\end{aligned}
$$

where $\bar{\mathbf{X}} = [X|B(T)]$, is the design matrix combining individual level covariates and basis functions. Thus, among the various components of the above variance-covariance matrix, $\bar{\mathbf{X}}^{\mathrm{T}}\mathbf{W}\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}^{\mathrm{T}}\mathbf{W}\boldsymbol{Z}$ are the only terms involving individual level data, and the rest of the terms involve a lower dimension corresponding to the group level data. Hence, upon convergence, calculation of the variance covariance matrix can also be carried out in a distributed computing framework for individual and group-level data separately.

However, the above standard errors for the non-linear function, $f(T)$ relies heavily on large sample assumption and treating the smoothing parameter as a known quantity (S. Wood, 2006). In reality, the smoothing parameter is estimated from the data, hence the confidence intervals for non-linear function based on the standard errors as calculated above may not be appropriate. Marra and Wood (2012) proposed and developed an alternative confidence interval based on the above frequentist and Bayesian variance covariance matrices. In this formulation the Bayesian and frequentist confidence intervals for the non-linear function, $f(T)$ can be obtained as $\widehat{f}(T) \pm z_{\alpha/2}\sqrt{[\boldsymbol{V}_f]_{ii}}$, where $\widehat{f}(T) = \mathbf{B}^{\mathrm{T}}(T)\widehat{\eta}$, $\boldsymbol{V}_f = \mathbf{B}(T)\boldsymbol{V}_\eta\mathbf{B}^{\mathrm{T}}(T)$, and $z_{\alpha/2}$ is the critical value of the standard normal distribution with level of significance, α . Here, $\boldsymbol{V}_\eta$ is the variance covariance matrix of $\widehat{\eta}$ that can be obtained from the corresponding block of $\boldsymbol{Q}_{Freq}$ in (3.11) and of $\boldsymbol{Q}_{Bayes}$ in (3.12) in order to obtain frequentist and Bayesian confidence interval, respectively.

The covariance matrix for $\widehat{\boldsymbol{\theta}}$ was obtained from the Fisher information matrix from Step 1 in the usual way, assuming parameters for the individual and area specific covariates are fixed. Of course there is additional variability due to the fact that the individual and area specific covariate parameters are estimated. However, following Breslow and Clayton (1993) we ignore the additional variability due to estimation of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$ for inference about estimated random effect, $\widehat{\boldsymbol{\theta}}$.

In the next section we describe a simulation study to evaluate the performance of smooth-indiCAR method.

3.3 Simulation Study

To evaluate our proposed smooth-indiCAR, we design a simulation study involving 400 regions in a 20×20 square lattice grid with varying sample sizes. To evaluate the smooth-indiCAR in both a large and a small sample settings, we divided a total of 20000 individuals (scenario i) and 5000 individuals (scenario ii) randomly among 400 areas. In this allocation, we ensured at least one individual for each region. We define two regions as neighbors if they share a common border. The random effects are then generated following a multivariate normal distribution with mean 0 and covariance matrix $D = [\sigma^{-2}\{(1-\lambda)\boldsymbol{I} + \lambda R\}]^{-1}$. The value of σ is set to 0.4 and five different values of spatial dependence parameters, $\lambda = \{0, 0.25, 0.50, 0.75, 0.99\}$ are considered in order to represent different strengths of spatial correlation. We then generate three individual level covariates (one binary, one categorical and one continuous) and one group level covariate. The binary covariate represents the distribution of sex in the area and is generated following Bernoulli random variable with probability ranging from 0.45 to 0.55 across groups. The categorical variable with five categories is generated with pre-specified probabilities. The continuous individual level covariate, T is generated using a univariate bump function as $f(t) = \frac{1}{1+t} - 2e^{-25(t-0.7)^2}$, to represent an age effect. The group level covariate is generated as a standard normal random variable. The outcome variable is then generated using model (3.1). The overall intercept of the model is set to zero in this simulation. The full list of the parameters used to generate the simulated data is given in the header row of Table 3.1.

Table 3.1: Estimated regression coefficients and variance parameters for the proposed smooth-indiCAR method using simulation scenario (i).

True value	β_2 -2.00	β_{32} -1.50	β_{33} 0.15	β_{34} 0.50	β_{35} 0.20	γ 0.20	σ 0.40	λ
λ	Estimated coefficient							
0.00	-2.000	-1.499	0.150	0.501	0.201	0.198	0.391	0.021
0.25	-2.001	-1.499	0.150	0.500	0.200	0.197	0.392	0.249
0.50	-1.999	-1.500	0.149	0.499	0.199	0.198	0.392	0.496
0.75	-2.000	-1.499	0.150	0.500	0.202	0.200	0.391	0.723
0.99	-2.000	-1.502	0.147	0.497	0.197	0.198	0.392	0.932
	Empirical standard error							
0.00	0.020	0.053	0.030	0.028	0.029	0.023	0.036	0.035
0.25	0.020	0.053	0.030	0.029	0.030	0.018	0.041	0.133
0.50	0.020	0.055	0.031	0.028	0.029	0.016	0.042	0.189
0.75	0.020	0.052	0.032	0.029	0.030	0.015	0.034	0.164
0.99	0.020	0.053	0.032	0.031	0.030	0.014	0.024	0.066
	Average of the simulated standard error							
0.00	0.020	0.052	0.030	0.028	0.029	0.021	0.030	0.049
0.25	0.020	0.053	0.030	0.029	0.029	0.018	0.031	0.095
0.50	0.020	0.053	0.030	0.029	0.029	0.016	0.027	0.113
0.75	0.020	0.053	0.030	0.029	0.029	0.015	0.023	0.099
0.99	0.020	0.054	0.030	0.029	0.029	0.014	0.021	0.046

3.4 Results

3.4.1 Simulation Results

Table 3.1 displays the average of the estimated regression coefficients of linear individual level covariates, group level covariate and parameters in the spatial random effects corresponding to model (3.1) along with their estimated standard errors based on 500 simulation runs of scenario (i). We calculate two different standard errors for the estimated regression coefficients: namely, (a) empirical standard errors i.e., taking the standard deviation of the 500 simulated regression coefficient estimates, (b) average of model based standard errors. The first column of Table 3.1 specifies the spatial dependence parameter used in that particular simulation. The second column represents the estimated coefficients corresponding to the binary variable, the next fours columns list the estimated regression coefficients for the categorical individual level covariates. The last three columns list the estimated regression coefficients for the group specific covariate, estimated over-dispersion and spatial dependence parameter.

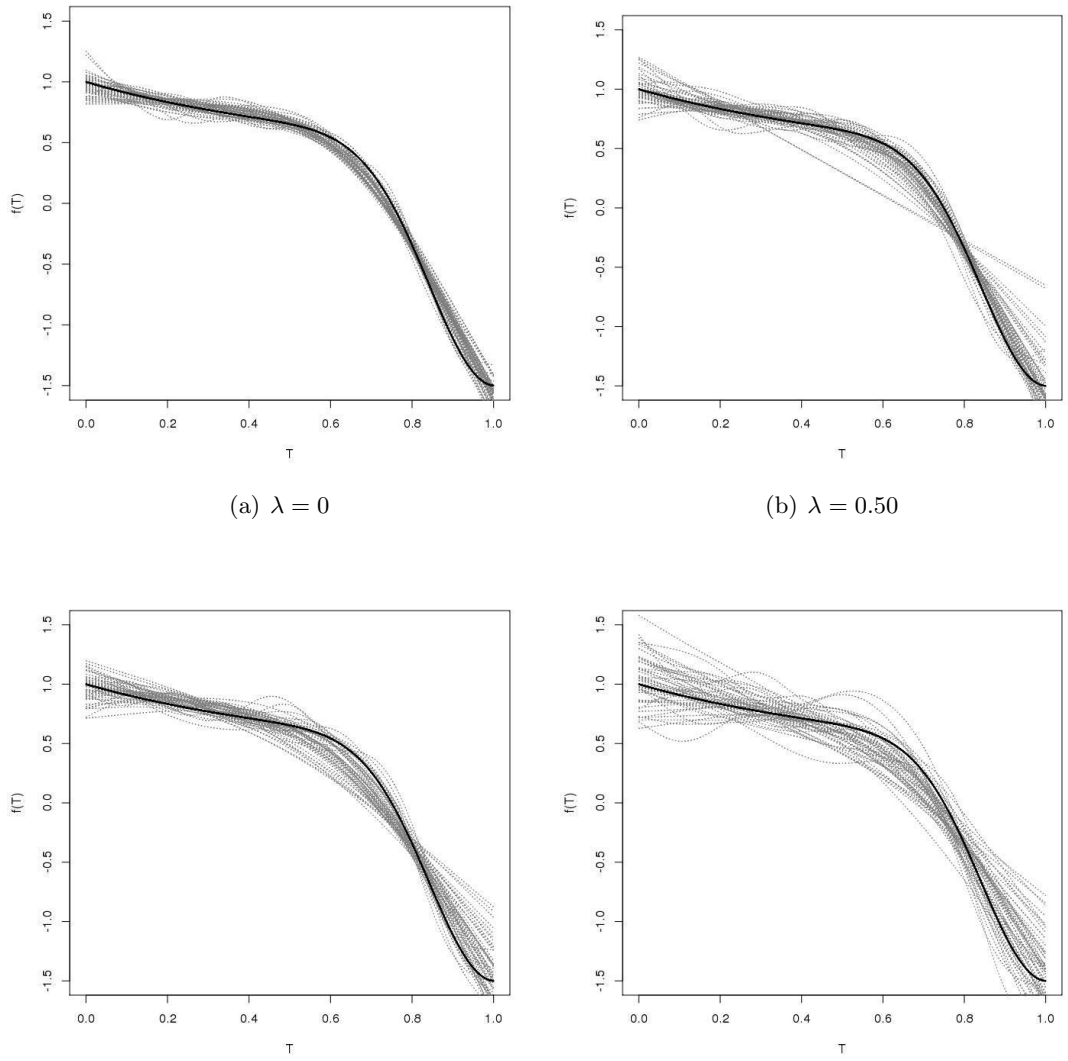
Table 3.2: Estimated regression coefficients and variance parameters for the proposed smooth-indiCAR method using simulation scenario (ii).

True value	β_2 -2.00	β_{32} -1.50	β_{33} 0.15	β_{34} 0.50	β_{35} 0.20	γ 0.20	σ 0.40	λ
λ	Estimated coefficient							
0.00	-1.999	-1.499	0.151	0.500	0.199	0.197	0.379	0.021
0.25	-1.999	-1.494	0.154	0.504	0.203	0.197	0.371	0.207
0.50	-2.000	-1.495	0.156	0.506	0.208	0.199	0.372	0.403
0.75	-2.004	-1.511	0.153	0.503	0.203	0.199	0.371	0.619
0.99	-2.004	-1.499	0.154	0.503	0.205	0.200	0.391	0.901
	Empirical standard error							
0.00	0.040	0.107	0.059	0.058	0.059	0.026	0.045	0.039
0.25	0.040	0.104	0.063	0.060	0.059	0.025	0.048	0.141
0.50	0.039	0.110	0.063	0.058	0.060	0.021	0.043	0.167
0.75	0.040	0.110	0.062	0.058	0.061	0.021	0.039	0.178
0.99	0.038	0.107	0.062	0.059	0.060	0.018	0.035	0.108
	Average of the simulated standard error							
0.00	0.040	0.106	0.061	0.058	0.059	0.025	0.040	0.070
0.25	0.040	0.107	0.062	0.059	0.060	0.022	0.041	0.127
0.50	0.040	0.108	0.062	0.059	0.060	0.021	0.037	0.161
0.75	0.040	0.108	0.062	0.059	0.060	0.020	0.033	0.154
0.99	0.040	0.108	0.062	0.059	0.060	0.019	0.032	0.067

As expected, the smooth-indiCAR method provides reliable estimates of the individual level and the region specific regression coefficients and the spatial random effect parameters. Moreover, the estimated standard error matches well with the empirical standard error for the individual level and the region specific regression coefficients. However, our proposed method underestimated the empirical standard error for the spatial random effect parameters.

To evaluate the performance of the proposed method under small sample settings, we also conducted another simulation study with 5000 subjects distributed randomly in 400 regions (scenario (ii)). The results are given in Table 3.2. As indicated in the table, the proposed method also performs well in the case when the number of individuals in a group is low.

The estimated non-linear functions also approximate the true non-linear function well for both scenarios (i) and (ii) as shown in Figure 3.1 and Figure 3.2, respectively. The solid line in these figures represents the true non-linear curve and the dotted lines represent the estimated non-linear functions from the first 50 simulations. The variability of the fitted curve increases with the degree of spatial dependence parameter.



(a) $\lambda = 0$ (b) $\lambda = 0.50$

(c) $\lambda = 0.75$ (d) $\lambda = 0.99$

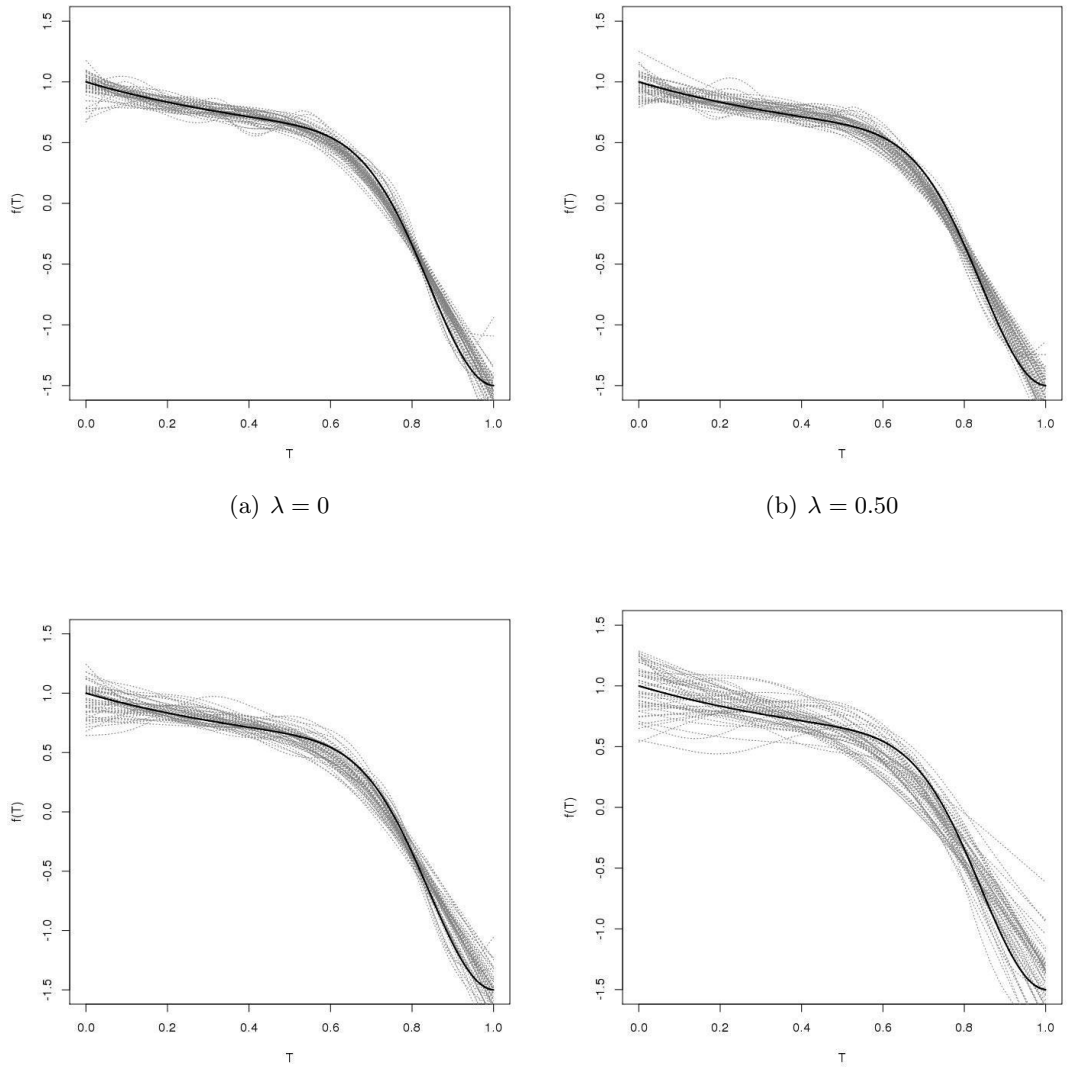
Figure 3.1: Fitted non linear curves based on the first 50 simulations under scenario (i) for different values of spatial dependence parameter. The solid line indicates the true curve.

3.4.2 Application to the Neutropenia Data

We applied our proposed smooth-indiCAR to the data on neutropenia admission. One of the key objectives of this analysis is to assess the geographical variation of neutropenia admission rates and its association with area level measures of socio-economic status. Data also includes patient age, gender, year of diagnosis, ARIA, cancer types at diagnosis, number of major comorbidities other than cancer and geographic location reported via postcode of residence.

Table 3.3: Comparison of estimated regression coefficients and variance parameters of smooth-indiCAR with indiCAR using neutropenia data.

	smooth-indiCAR		**indiCAR**	
Regression coefficients	Estimates	Std. Error	Estimates	Std. Error
Intercept	-2.316	0.108	-1.493	0.047
Sex				
Female	Ref		Ref	
Male	-0.091	0.020	-0.043	0.020
Year of Diagnosis				
2001	Ref		Ref	
2002	0.016	0.038	0.021	0.038
2003	0.083	0.038	0.083	0.038
2004	0.023	0.038	0.019	0.038
2005	0.097	0.037	0.095	0.037
2006	0.038	0.038	0.038	0.038
2007	0.029	0.038	-0.022	0.038
2008	0.000	0.038	-0.004	0.038
2009	-0.313	0.042	-0.315	0.042
ARIA				
Major Cities	Ref		Ref	
Inner Regional Australia	-0.024	0.047	-0.006	0.022
Outer Regional Australia	-0.150	0.068	-0.118	0.037
Remote/ Very remote Australia	-0.244	0.163	-0.192	0.142
Cancer Type				
Breast Cancer	Ref		Ref	
Lung cancer	0.259	0.038	0.240	0.037
Colon & rectum cancer	-0.428	0.040	-0.463	0.040
Haematological Malignancy	1.579	0.029	1.497	0.029
Other cancer	-0.934	0.032	-0.986	0.031
No. of major comorbidities				
0	Ref		Ref	
1	0.424	0.026	0.413	0.026
2	0.680	0.026	0.682	0.026
3	0.625	0.036	0.589	0.036
4+	0.623	0.035	0.594	0.035
SEIFA				
Most disadvantaged	Ref		Ref	
2	-0.082	0.044	-0.089	0.043
3	-0.070	0.041	-0.078	0.038
4	-0.125	0.047	-0.134	0.045
Least disadvantaged	-0.129	0.056	-0.144	0.053
Variance parameter				
σ	0.203	0.023	0.209	0.022
λ	0.992	0.012	0.992	0.012
AIC	87402.88		87580.13	



(a) $\lambda = 0$ (b) $\lambda = 0.50$

(c) $\lambda = 0.75$ (d) $\lambda = 0.99$

Figure 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.

Table 3.3 reports the multivariable analysis of neutropenia admission using the smooth-indiCAR and indiCAR methods as discussed in the previous chapter. The only difference between these two methods is that the former includes the age effect as a non-linear predictor and the latter includes the age effect as linear. In general, the results are quite similar although the magnitude of the regression coefficients differs. As shown in Figure 3.3, the risk of neutropenia in cancer patients is non-linear and decreases rapidly beyond the age of 65. This might explain the difference in the estimates of the regression parameters between indiCAR and smooth-indiCAR. Our results with a non-linear age effect are very similar in terms of coefficients of other

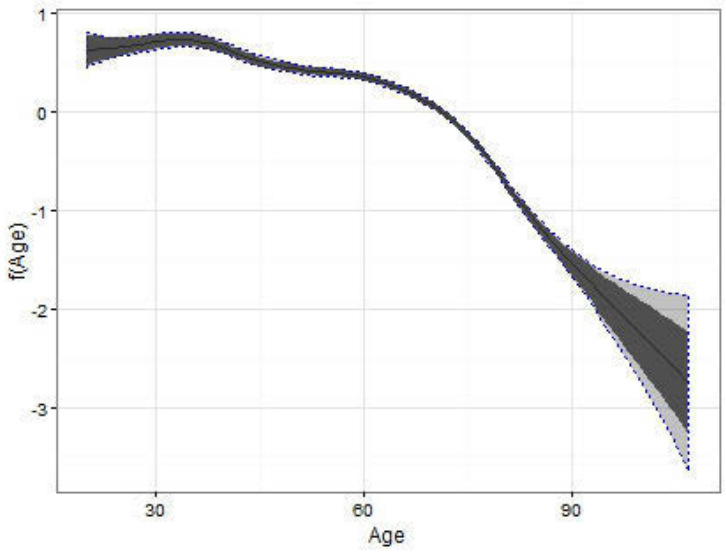
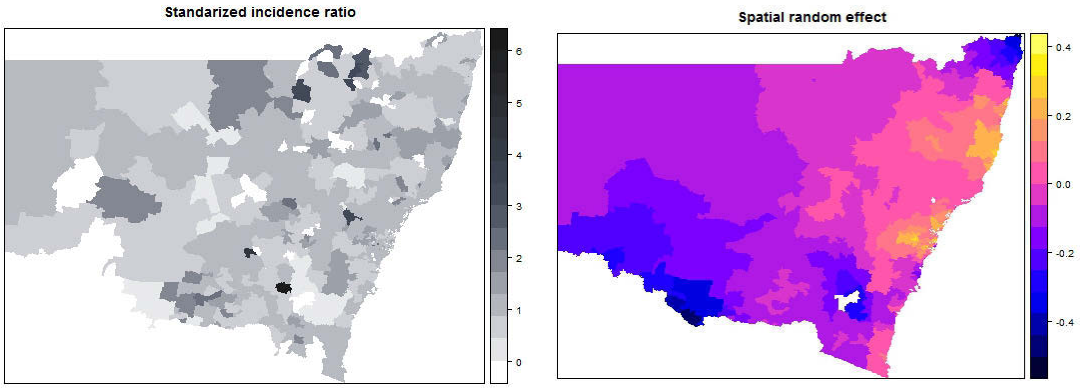


Figure 3.3: The estimated effect of age on neutropenia admission rates with associated 95 % Bayesian (light gray region) and frequentist (dark region) confidence intervals.

covariates in the model to the results of the indiCAR method with age as a categorical predictor (see Table 2.4 in the previous chapter). Male patients compared with female patients, patients living in outer remote area compared with patients in major cities and patients with a higher socio-economic index have lower risk of neutropenia infections.



(a) Standarized Incidence Ratio.

(b) Estimated spatial random effects.

Figure 3.4: Distributions of estimated (a) Standardized Incidence Ratios (SIR) and (b) spatial random effects of neutropenia admissions in NSW, Australia.

We obtained similar distributions of estimated Standardized Incidence Ratios (SIR) and spatial random effects of neutropenia admissions in NSW for the smooth-indiCAR method and indiCAR method. Figure 3.4, shows the estimated Standardized Incidence Ratios (SIR) and estimated spatial random effects using indiCAR. The white region in the map of NSW is the Australian Capital Territory (ACT), a standalone state

surrounded by NSW. Two other Australian states, Queensland (QLD) and Victoria (VIC) are located to the North-East and South-West of NSW, respectively. Some edge effect is inevitable due to patients living near these border regions receiving treatment interstate. The strong spatial correlation after adjusting for individual and group specific covariates indicates that geographical variation of neutropenia might be due to differences in health care practices or access to care across NSW.

3.5 Discussion

In this chapter, we have developed a framework to semi-parametrically adjust for a continuous individual level covariate effect in spatial disease mapping. Our results suggest that smooth-indiCAR provides reliable estimates of the true regression parameters. Due to the natural applicability of the smooth-indiCAR method in a distributed computing framework, this method has potential for Big Data implementations. One of the key problems in Big Data analysis is to divide the data so that this division retains the inherent correlation structure of the data. Our proposed methodology provides a convenient method for such division by separating data according to the natural characteristics of the data, based on individual and group level covariates. The individual level covariate data can then be analyzed with the recent development of generalized additive models for large data sets (S. Wood, Goude, & Shaw, 2015). Thus our proposed smooth-indiCAR provides a convenient way to extend recent developments in Big Data for independent responses to the spatially correlated responses. This could also speed up the process and reduce computational costs.

For simplicity we assume single smooth function, however (3.1)can be extended to include more than one smooth functions. In our proposed smooth-indiCAR we have considered the smooth term as a nuisance parameter and that interest lies in inference on other things, adjusted for smooth covariate. In this thesis, we have presented formula for the confidence intervals for the smooth component. The proposed confidence interval is known to exhibit good coverage probabilities (S. N. Wood (2012)). However, testing the smooth term for equality to zero is beyond the scope of the current thesis. Future research should be carried out whether an indiCAR and a smooth-indiCAR model is more appropriate. Although our simulation results suggests that smooth-indiCAR model

provides consistent estimates of the standard error of the main effect, we however observed underestimation of the standard error for the variance component. Future research should carried out to obtain appropriate standard error for variance component.

Health registries routinely collect geo-coded information for patient's residence at diagnosis and their individual level socio-demographic and clinical characteristics and thus could benefit by using our proposed method to incorporate individual level information in the analysis and mapping of disease rates. We illustrated the proposed approaches using the analysis of neutropenia data. There are a number of areas where future study would be useful.

We observed a lower risk of neutropenia admissions associated with increasing age, although advanced age has been identified as a significant predictor for neutropenia admissions in previous studies (Klastersky et al., 2000). This might be due to the fact that the current risk based prophylactic administration of Colony Stimulating Factor (CSF) guidelines account for patients advanced age (Aapro et al., 2011). The lower than expected counts near the borders of NSW are almost certainly a result of some patients being admitted for neutropenia out of state. The higher than expected areas are located in two areas of higher population, this pattern invites further investigation.

The dependence between area-specific neutropenia rates on ARIA and SEIFA are in the opposite direction. This is counter intuitive as remote areas in NSW are mostly associated with disadvantaged SEIFA categories. However, the observed contrast in estimated regression coefficients might be due to the differences in the health care practices. Patients in the remote areas where the patients are geographically distant to the treating medical oncologist are best managed by their primary care physicians, and therefore may be treated with lower doses of chemotherapy (Fox & Boyce, 2014). On the contrary, patients in the major cities might get intensive chemotherapy to treat them early, and are better managed due to availability of resources. Previous studies also indicate that remoteness has the greatest effect in quality of cancer treatment (Jong et al., 2004) and it affects treatment choices made by both patients and clinicians (Nattinger et al., 2001).

Variation across clinical practices of neutropenia has been identified in Australia in a

previous survey (Lingaratnam et al., 2011). The authors showed that the treatment approach for management of neutropenia varies across oncologists, hematologists and clinicians as well as different sectors of cancer care. Therefore, it might be interesting to explore whether the observed variation is due to variation across different hospitals (eg., metropolitan hospital vs. Non-metropolitan hospitals) in NSW or across various healthcare providers. However, data items for such analysis are not collected in the registry and are beyond the scope of our present paper.

Our study was based on data linked from a state-based cancer registry and administrative data from the Admitted Patient Data Collection (APDC). An advantage of such linked data is that it provides us with a large, population based sample. Registry based analysis is more comprehensive than that based on single centre studies, and provides more complete information than may be obtained from clinical trials where patient selection and loss to follow-up may impact validity and generalizeability of study findings. However, it is important to keep in mind that the resulting data quality may be inferior to that obtained from prospective clinical studies.

Despite various limitations, smooth-indiCAR is an useful addition to the existing methodology to explore clinical variation across geographical locations where covariates might have non-linear effects. One of the major advantages of our proposed method is the ability to obtain both individual and group level covariate effects when employing spatial regression models for disease mapping.

Although both indiCAR and smooth-indiCAR modeled individual and group level data very well, it appears that the estimated regression coefficient of a random effect model depends strongly on the assumed spatial correlation structure. We hypothesized that such sensitivity of the model parameters on spatial correlation structure is especially likely to occur when the covariate of interest has been measured with error. We examine the effect of covariate measurement error in the spatial linear regression setting. The results are presented in the next chapter.

R Code

Function for data generation

```
library(MASS)
DataSimulation<-function(nsamp=20000,nGroup=400,sigma=0.4,lambda=0.75,
 beta0=0, beta1=1, beta2=-2.0, beta32=-1.5,beta33=0.15,beta34=0.5,
 beta35=0.2, gamma=0.2)
{
#        nGroup: Total number of group
#        minIndivGroup: Minimum number of individual per group
#        maxIndivGroup: Maximum number of individdual per group
#         lambda:
#         sigma: variance parameter
#         beta0-beta3: individual level coefficients
#         gamma: group level coefficient
repeat{
  indivGroup<-as.vector(rmultinom(n=1, size=nsamp,prob=runif(400,0.05)))
  if (all(indivGroup)>=1){break}
  }
totalSample<-sum(indivGroup)
#generate covariate values
pr =runif(totalSample,0.45,0.55)
x2<-rbinom(totalSample,1,pr)
x1<-sort(runif(totalSample))
f <-function(x) 1/(1+x) - 2*exp(-20*(x-1)^2)
f1<-f(x1)
x3<-sample(1:5,totalSample, replace=TRUE,
      prob=c(0.06,0.09,0.19,0.25,0.25))
#generate covariate values
z2<-rnorm(nGroup)
x4<-rep(z2,indivGroup)
ID<-rep(1:nGroup,indivGroup)
#Generating spatial random effect in a lattice grid
#### Set up a square lattice region
x.easting <- 1:20
x.northing <- 1:20
Grid <- expand.grid(x.easting, x.northing)
n <- nrow(Grid)
#### set up distance and neighbourhood  matrices W
distance <-array(0, c(n,n))
W <-array(0, c(n,n))
for(i in 1:n)
 {
 for(j in 1:n)
  {
  temp <- (Grid[i,1] - Grid[j,1])^2 + (Grid[i,2] - Grid[j,2])^2
  distance[i,j] <- sqrt(temp)
  if(temp==1)  W[i,j] <- 1
  }
}
R <- -W
diag(R) <- as.numeric(apply(W, 1, sum))
Sigma.inv<-1/sigma^2*(lambda*R + (1-lambda)*diag(rep(1,n)))
Sigma<-solve(Sigma.inv)
phi <- mvrnorm(n=1, mu=rep(0,n), Sigma=Sigma)
phi_long<-rep(phi,indivGroup)
mu <- exp(beta0 + beta1*f1 +beta2*x2+beta32*I(x3==2)+beta33*I(x3==3)+
 beta34*I(x3==4)+beta35*I(x3==5)+gamma*x4+phi_long)
#generate Y-values
y <- rpois(totalSample, lambda=mu)
#data set
data <- data.frame(ID,y=y, x1,x2,x3,x4,f1)
output<-list(data=data,R=R)
output
}
```

Necessary functions for implementing the proposed method

```
# Load required library
library(mgcv)
DataSim<-DataSimulation()
sampleData1<-DataSim$data
R <- DataSim$R
#Use of generalized linear model with individual level covariates
# Fit model
fit_ind<-gam(y~s(x1,k=10)+x2+as.factor(x3), data=sampleData1,
                        family="poisson")
#coef(fit_ind)
# Extract model matrix and beta
#X_ind <-predict(fit_ind,type="lpmatrix")
#n_ind<-nrow(X_ind)
beta <- coef(fit_ind)
#Calculate the fitted value
sampleData1$Predict<-exp(predict(fit_ind))
#sampleData1$Predict1<-exp(X_ind%*%beta)
head(sampleData1)
#Grouplevel data: Location data
locationData<-sampleData1[,c("ID","x4")]
GroupCovData<-aggregate(.~ ID, data = locationData, mean)
#Aggregate data Over postcode
AreaData<-sampleData1[,c("y","ID","Predict")]
aggdata <-aggregate(. ~ ID, data = AreaData, sum)
nGroup<-nrow(aggdata)
aggdataU<-aggdata[!duplicated(aggdata$ID),]
GroupData<-merge(GroupCovData,aggdataU,by="ID")
names(GroupData)<-c("ID","x4","totalY","totalePredict")
#Fitting PQL model
gamma.iter<-NULL
theta.iter<-NULL
beta.iter<-NULL
#Set initial values
gamma.hat<-0
b.rand<-rep(0,nGroup)
sigma.hat <-0.5
lambda.hat <-0.5
theta.hat<-c(sigma.hat,lambda.hat)
#Generate covariance matrix
betaCombined<-c(beta,gamma.hat)
X_group <-as.matrix(GroupData$x4)
Y<-as.matrix(GroupData$totalY)
OffSet<-as.matrix(log(GroupData$totalePredict))
repeat{
   repeat{
# Estimate covariance matrices
R.lambda <- lambda.hat*R + (1-lambda.hat)*diag(rep(1,nGroup))
R.lambda.inv<-solve(R.lambda)
D.hat.inv<-1/(sigma.hat^2)*R.lambda
D.hat<-solve(D.hat.inv)
# Calculate the PQL elements
      repeat{
eta.est<-OffSet+X_group%*%gamma.hat+b.rand
mu.est<-exp(eta.est)
zz.est<-eta.est+(Y-mu.est)/mu.est
W.hat<-diag(as.vector(mu.est))
W.hat.inv<-diag(as.vector(1/mu.est))
#Estimate covariance matrix of Y
V.hat<-W.hat.inv+D.hat
V.hat.inv<-solve(V.hat)
#Estimate the fixed and random effect
gammaUpdate<-solve(t(X_group)%*%V.hat.inv%*%X_group)%*%(t(X_group)%*%
                                  V.hat.inv%*%(zz.est-OffSet))
b.rand.update<-D.hat%*%V.hat.inv%*%(zz.est-OffSet-X_group%*%gammaUpdate)
diff<-abs(gammaUpdate-gamma.hat)
diff.rand<-abs(b.rand.update-b.rand)
gamma.iter<-rbind(gamma.iter,gammaUpdate)
if (all(diff< 1e-5)){break}
if (all(diff.rand< 1e-3)){break}
```

```
gamma.hat<-gammaUpdate
b.rand<-b.rand.update
                    }
# Extract score and observed information matrix
P<-V.hat.inv-(V.hat.inv%*%X_group%*%solve(t(X_group)%*%V.hat.inv%*%
                                  X_group)%*%t(X_group)%*%V.hat.inv)
dV.sigma<-2*sigma.hat*R.lambda.inv
dV.lambda<--1*sigma.hat^2*R.lambda.inv%*%(R-diag(rep(1,nGroup)))
                                  %*%R.lambda.inv
#Score vector
score.sigma<-0.5*(t(zz.est-OffSet-X_group%*%gammaUpdate)%*%P)%*%
                dV.sigma%*%(P%*%(zz.est-OffSet-X_group%*%gammaUpdate))-
                        0.5*sum(diag(P%*%dV.sigma))
score.lambda<-0.5*(t(zz.est-OffSet-X_group%*%gammaUpdate)%*%P)%*%
                dV.lambda%*%(P%*%(zz.est-OffSet-X_group%*%gammaUpdate))-
                        0.5*sum(diag(P%*%dV.lambda))
score<-c(score.sigma,score.lambda)
exp.inf11<-0.5*sum(diag(P%*%dV.sigma%*%P%*%dV.sigma))
exp.inf12<-0.5*sum(diag(P%*%dV.sigma%*%P%*%dV.lambda))
exp.inf21<-0.5*sum(diag(P%*%dV.lambda%*%P%*%dV.sigma))
exp.inf22<-0.5*sum(diag(P%*%dV.lambda%*%P%*%dV.lambda))
exp.infor<-matrix(c(exp.inf11,exp.inf12,exp.inf21,exp.inf22),ncol=2)
thetaUpdate<-theta.hat+solve(exp.infor)%*%score
if (thetaUpdate[2]>1) {thetaUpdate[2]<-0.99999}
if (thetaUpdate[2]<0) {thetaUpdate[2]<-0.00001}
if (thetaUpdate[1]<0) {thetaUpdate[1]<-0.1}
sigma.hat<-thetaUpdate[1]
lambda.hat<-thetaUpdate[2]
diff.theta<-abs(thetaUpdate-theta.hat)
theta.iter<-rbind(theta.iter,as.vector(thetaUpdate))
if (all(diff.theta< 1e-5)){break}
theta.hat<-thetaUpdate
        }
# Repeat individual level data fitting
GroupData$Predict_group<-X_group%*%gamma.hat+b.rand
combinedData<-merge(sampleData1,GroupData[,c("ID","Predict_group")],
                                        by=c("ID"))
fit_ind<-gam(y~s(x1,k=10)+x2+as.factor(x3),offset=Predict_group,
                        data=combinedData,family="poisson")
#Calculate the fitted value
betaUpdate<-coef(fit_ind)
#Calculate the fitted value
combinedData$Predict<-exp(predict(fit_ind))
AreaData<-combinedData[,c("y","ID","Predict")]
aggdata <-aggregate(. ~ ID, data = AreaData, sum)
aggdataU<-aggdata[!duplicated(aggdata$ID),]
GroupData<-merge(GroupCovData,aggdataU,by="ID")
names(GroupData)<-c("ID","x4","totalY","totalePredict")
Y<-as.matrix(GroupData$totalY)
OffSet<-as.matrix(log(GroupData$totalePredict))
betaCombined.Update<-c(betaUpdate,gamma.hat)
diff.est<-abs(betaCombined.Update-betaCombined)
beta.iter<-rbind(beta.iter,betaCombined.Update)
if (all(diff.est< 1e-5)) {break}
betaCombined<-betaCombined.Update
# End of smooth-indiCAR
        }
# Extract the smoothing parameters and vector
G<-gam(y~s(x1,k=10)+x2+as.factor(x3),offset=Predict_group,
        fit=FALSE, data=combinedData,family="poisson")
X_bar<-G$X
lambda<-as.vector(fit_ind$sp)
K<-G$S[[1]]
S<-lambda*K
M<-nrow(GroupData)
group.size<-as.vector(table(combinedData$ID))
group.cov.long<-data.matrix(X_group[rep(1:nrow(X_group),
                times = group.size), ])
cov.combined<-cbind(X_bar,group.cov.long)
fitted.combined<-cov.combined%*%betaCombined+rep(b.rand,group.size)
mu.combined<-as.vector(exp(fitted.combined))
```

```
#Doing in another way
Xcov.m<-X_bar*mu.combined
XTWX<-t(Xcov.m)%*%X_bar
groupID<-rep(c(1:M),group.size)
XTWZ<-t(rowsum(Xcov.m, groupID))
XTWZD<-XTWZ%*%D.hat
ZTWZ.vec<-aggregate(mu.combined,by=list(groupID),sum)
ZTWZ<-diag(ZTWZ.vec$x)
ZTWZD<-ZTWZ%*%D.hat
I.ZTWZD<-(diag(M)+ZTWZD)
I.ZTWZD.inv<-solve(I.ZTWZD)
a11<-XTWX-XTWZD%*%I.ZTWZD.inv%*%t(XTWZ)
XTWZU<-XTWZ%*%X_group
ZTWZU<-ZTWZ%*%X_group
a12<-XTWZU-XTWZD%*%I.ZTWZD.inv%*%ZTWZU
a21<-t(a12)
a22<-t(X_group)%*%ZTWZ%*%X_group-t(X_group)%*%ZTWZ%*%
                D.hat%*%I.ZTWZD.inv%*%ZTWZ%*%X_group
Q.inv<-as.matrix(rbind(cbind(a11,a12),cbind(a21,a22)))
# Bayesian estimates of regression coefficients
S1<-matrix(0,16,16)
S1[8:16,8:16]<-S
Q.bayes.inv<-Q.inv+S1
Q<-solve(Q.bayes.inv)
se<-sqrt(diag(Q))
cbind(betaCombined,se)
se.theta<-sqrt(diag(solve(exp.infor)))
se.theta
#Frequentist estimates of regression coefficients
Q.freq<-Q%*%Q.inv%*%Q
se.freq<-sqrt(diag(Q.freq))
cbind(betaCombined,se,se.freq)
```

Chapter 4

On the Impact of Covariate Measurement Error on Spatial Regression Modeling.

Summary

Spatial regression models have grown in popularity in response to rapid advances in GIS (Geographic Information Systems) technology. In health research, for example, it is common for epidemiologists to incorporate geographically indexed data into their studies. However, it turns out that there are some pitfalls. In contrast to many regression analysis settings where parameter estimation is fairly robust to covariance specification, we describe some empirical findings which suggest that spatial regression analysis can be acutely sensitive to specification of the spatial correlation structure. While some authors have studied the impact of omitted covariates as an explanation for this phenomenon, we found that the presence of covariate measurement error can lead to significant sensitivity of parameter estimation to the choice of spatial correlation structure. We elucidate the effect of measurement error on parameter estimates and

The content of this chapter is published as: Huque, MH; Bondell, HD and Ryan. LM. (2014). On the impact of covariate measurement error on spatial regression modelling. *Environmetrics 25 (8)*,560-570. This research was also presented at the Australian Statistical Conference in conjunction with the Institute of Mathematical Statistics Annual Meeting, Sydney, 2014.

discuss and evaluate several different ways to produce unbiased estimates.

keywords: Measurement Error, Spatial regression.

4.1 Introduction

Advances in statistical methodology, together with geographically referenced health databases, present an unique opportunity to investigate the environmental, social and behavioural factors underlying geographic variations (Elliott & Wartenberg, 2004). In health research, for example, social epidemiologists seek to assess the impact of socio-demographic characteristics of a community on the health of individuals living in that community (Elliot, Wakefield, Best, & Briggs, 2000). Analysis of geo-coded data is complicated by correlations among observations located near each other. Regression analysis ignoring these spatial correlations leads to incorrect inference on the estimated regression coefficients by narrowing confidence intervals. Mixed effect models provide a convenient way of modeling spatial correlations by incorporating random effects with spatial correlation structure (Waller & Gotway, 2004). In this paper, we focus on how such models perform when covariates of interest are measured with error.

In the case study that motivates this paper, Australian researchers explored the relationship between the SEIFA index (an area-based measure of socio-economic status produced by the Australian Bureau of Statistics) and acute hospitalization for Ischemic Heart Disease (IHD) for approximately 600 postcodes in NSW, Australia (Burden et al., 2005; Guha et al., 2009). Regression models suggest a strong association between SEIFA and IHD, even after adjusting for factors such as age, gender, population density and other factors that might influence the outcome. However, exploratory analysis reveals that the estimated coefficient of the SEIFA index from such models depends strongly on the assumed spatial correlation structure. Briefly, the estimated SEIFA coefficients are all significantly negative, confirming that IHD rates decrease as social advantage increases. However, the magnitude of the effect varies by more than a factor of 2, depending on whether or not a spatial correlation adjustment is made. Similar sensitivity to assumed spatial correlation structure can be seen in analysis of the well-known Scottish Lip Cancer data (Breslow & Clayton, 1993; Clayton et al., 1993). In

another spatial epidemiological study Molitor et al. (2007) fit a model for the effect of NO_2 exposure on lung function. They considered a series of models included one based on a conditional auto-regressive (CAR) model. They observed that models with spatial structure give smaller effect estimates as compared to models without spatial structure. These results suggest that estimated coefficients from a spatial regression model can be highly sensitive to whether and how spatial variation is accommodated. In this paper, we show that such sensitivity is especially likely to occur when the covariate of interest has been measured with error.

Presence of measurement error in the covariate of interest arises in many epidemiological and socio behavioural studies. For example, in the study of geographical variation in bladder cancer rates, lung cancer risk might be included in the model as a proxy for smoking exposure (Clayton et al., 1993). In environmental epidemiology, individual air pollution exposures might be approximated by the distance from the polluted sites or by using the measures at a few monitoring sites (Carroll et al., 1997). Further examples include geographical studies relating cancer incidence and mortality to dietary intakes (Prentice & Sheppard, 1990).

Many papers have appeared in the literature over the years on covariate measurement error in the context of independent data (Carroll et al., 2006; Fuller, 1987; Wansbeek & Meijer, 2000). In case of linear regression with independent data, it is well known that presence of exposure measurement error causes estimated regression coefficients attenuate toward the null. However, relatively few have addressed the effect of exposure measurement error in the context of correlated data with spatial structure. In epidemiological studies of association between air pollutants and health outcome, typically data are available from few monitoring sites. Therefore, the measured exposure used in the analysis might be different from the underlying true exposure.

Xia and Carlin (1998) presented a spatio-temporal analysis of spatially correlated data with errors in the covariates, in the context of disease mapping. The authors empirically studied several alternative measurement error models using a metropolis Gibbs algorithm. Li et al. (2009) derived asymptotic bias expressions for estimated regression coefficients in the context of a spatial linear mixed model. They showed that the regression estimates obtained from naive use of an error prone covariates attenuates the

estimated regression coefficient and variance component estimates are inflated. They proposed the use of a maximum likelihood approach based on the EM algorithm to adjust for measurement error under the assumed error structure. However, their simulation assumes that the measurement error variance is known and they did not assess the performance of their method in the case of misspecification. Their approach is also subject to a high computational burden and may lead to spurious result in the presence of outliers or model misspecification (Gryparis et al., 2009; Szpiro et al., 2011). Furthermore, Szpiro et al. (2011) argued that in the presence of spatial correlation, joint modeling becomes challenging as it is very difficult to separate out the spatial correlation between exposure and outcome.

In this paper, we explore the sensitivity of estimated regression coefficients in spatial regression models, showing that it arises in settings where the covariate of interest has been measured with error. We show that ignoring measurement error attenuates estimated regression coefficients and observe that estimates can be very sensitive to the choice of assumed correlation structure in the model formulation. We derive expressions for the bias when measurement error is ignored and present some technical derivations that characterize the bias as a function of the degree of measurement error as well as the degree of spatial correlation in the covariate of interest and in the residuals. We show that the bias due to attenuation depends on the spatial correlation structure. When there is no or the same degree of spatial correlation in both covariate or the measurement error the bias in spatial linear model reduces to the familiar attenuation factors under OLS modeling of independent data, namely $\rho = \sigma_X^2/(\sigma_X^2 + \sigma_U^2)$, where σ_X^2 is the variance of the true covariate and σ_U^2 is the variance of the measurement error.

Based on these expressions, we propose two different strategies for obtaining consistent estimates: (i) adjusting the estimates using an estimated attenuation factor; and (ii) using an appropriate transformation of the error prone covariate. We then evaluate the performance of these two approaches via simulations. These approaches do not require complex programming and can be implemented via readily available mixed model software. Moreover, we suggest ways to estimate measurement error variance from the data rather than assuming measurement error variance as a known quantity. Our simulation results show that bias correction methods using the estimate of the measurement error work reasonably well in obtaining consistent estimates. However,

estimation of the measurement error variance requires additional data or assumptions related to the underlying measurement error process. In the case of spatial epidemiology, validation data are typically rare. Therefore we suggest employing a sensitivity analysis when dealing with measurement error problems in practice. We illustrate the methods using data on Ischemic Heart Disease (IHD) and conclude with some practical guidelines.

4.2 Model Formulation

Suppose that X_i represents the true covariate of interest for spatial location i, $i = 1, ..., n$, and suppose that it is related to an outcome Y_i according to a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{4.1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \sim N(0, \boldsymbol{\Sigma}_\epsilon)$ and $\boldsymbol{\Sigma}_\epsilon$ is a covariance matrix, for now kept arbitrary. Let W_i be the observed covariate for spatial location i, related to the true covariate according to a classical measurement error model:

$$W_i = X_i + U_i,$$

where $\mathbf{U} = (U_1, ..., U_n)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}_U)$. When $\mathbf{X} = (X_1, ..., X_n)^T$ is also normally distributed (say with mean $\boldsymbol{\mu}_X$ and covariance $\boldsymbol{\Sigma}_X$), straightforward algebra establishes that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and $\mathbf{W} = (W_1, ..., W_n)^T$ have a multivariate normal distribution,

$$\begin{pmatrix} \boldsymbol{Y} \\ \mathbf{W} \end{pmatrix} \sim MVN \left(\begin{pmatrix} (\beta_0 + \beta_1 \boldsymbol{\mu}_X)\mathbf{1} \\ \boldsymbol{\mu}_X \mathbf{1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_\epsilon + \beta_1^2 \boldsymbol{\Sigma}_X & \beta_1 \boldsymbol{\Sigma}_X \\ \beta_1 \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U \end{pmatrix} \right)$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones. Standard theory for the multivariate normal establishes that $\boldsymbol{Y}|\mathbf{W}$ is normally distributed with conditional mean

$$E(\boldsymbol{Y}|\mathbf{W}) = \beta_0 \mathbf{1} + \beta_1 (\boldsymbol{I} - \boldsymbol{\Lambda}) \boldsymbol{\mu}_x + \beta_1 \boldsymbol{\Lambda} \mathbf{W} \tag{4.2}$$

and conditional variance

$$Var(\boldsymbol{Y}|\mathbf{W}) = \boldsymbol{\Sigma}_\epsilon + \beta_1^2 (\boldsymbol{I} - \boldsymbol{\Lambda}) \boldsymbol{\Sigma}_X,$$

where

$$\mathbf{\Lambda} = \mathbf{\Sigma}_X(\mathbf{\Sigma}_X + \mathbf{\Sigma}_U)^{-1}. \tag{4.3}$$

For ease of discussion, assume that the variable $\mathbf{X}$ has been centered so that $\boldsymbol{\mu}_X = \mathbf{0}$. In direct analogy with standard measurement error settings, these results suggest that regression coefficients obtained by regressing the outcome ($\boldsymbol{Y}$) on the observed, but error prone covariate ($\mathbf{W}$) will lead to bias as well as inaccurate variance modeling. We proceed now to explore the nature of this bias under varying assumptions about the correlation structure for $\boldsymbol{Y}$, $\mathbf{X}$ and the measurement error term, $\mathbf{U}$.

4.3 Asymptotic Bias Analysis

Suppose we fit model (4.1), naively replacing $\mathbf{X}$ with the error prone version of the covariate $\mathbf{W}$ and assuming independence of the error terms in the model on $\boldsymbol{Y}$. The ordinary least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{ols} = (\mathbf{W}_*^T\mathbf{W}_*)^{-1}\mathbf{W}_*^T\boldsymbol{Y}, \tag{4.4}$$

where $\mathbf{W}_*$ is the $n \times 2$ matrix with elements of the first column all equal to 1 and second column corresponding to the $n \times 1$ vector $\mathbf{W}$. Under the true model and assuming $\boldsymbol{\mu}_X = 0$, it is straight forward to show that the limiting value of this estimate is

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}^{ols} &= \begin{pmatrix} 1 & 0 \\ 0 & trace\,(\mathbf{\Sigma}_X + \mathbf{\Sigma}_U) \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & trace\,(\mathbf{\Sigma}_X) \end{pmatrix} \boldsymbol{\beta} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & \rho^{ols} \end{pmatrix} \boldsymbol{\beta},
\end{aligned}$$

where $\rho^{ols} = trace\,(\mathbf{\Sigma}_X)\,/trace\,(\mathbf{\Sigma}_X + \mathbf{\Sigma}_U)$, see the Appendix 4A.

Using basic properties of the trace function, this simple formula leads to a number of interesting observations. For example, suppose that both $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_U$ have constant diagonal elements σ_X^2 and σ_U^2, respectively, then the bias factor can be written as $\rho^{ols} = \sigma_X^2/\left(\sigma_X^2 + \sigma_U^2\right)$. This is the standard measurement error result (see Carroll et al.

2006), namely that the estimated regression coefficient is biased towards the null by an attenuation factor that reflects the proportion of the variability in the observed covariate $\mathbf{W}$, explained by the true covariate $\mathbf{X}$. Note that there is no bias in the estimated intercept in this case since we have assumed that $\mathbf{X}$ has mean zero. It is interesting to note that the result holds regardless of the correlation structures on the error term, $\boldsymbol{\Sigma}_\epsilon$.

In the next section, we consider the bias associated with fitting a generalized least squares model in the presence of covariate measurement error. We will see that in this case, the degree of bias also depends on the assumed error structure.

4.3.1 Generalized Least Squares

Suppose we obtain a generalized least squares (GLS) estimator of $\boldsymbol{\beta}$, under that assumption that the error term $\boldsymbol{\epsilon}$ has covariance matrix $\boldsymbol{\Sigma}_a$, with the subscript "a" denoting "assumed". For fixed $\boldsymbol{\Sigma}_a$, the estimator is:

$$\hat{\boldsymbol{\beta}}^{gls} = (\mathbf{W}_*^T \boldsymbol{\Sigma}_a^{-1} \mathbf{W}_*)^{-1} \mathbf{W}_*^T \boldsymbol{\Sigma}_a^{-1} \boldsymbol{Y}. \tag{4.5}$$

In the limit under the true model and following similar arguments as in the OLS case, this estimate converges in probability to

$$\begin{aligned} \tilde{\boldsymbol{\beta}}^{gls} &= \begin{pmatrix} n & 0 \\ 0 & trace\left[\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U)\right] \end{pmatrix}^{-1} \begin{pmatrix} n & 0 \\ 0 & trace\left[\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_X\right] \end{pmatrix} \boldsymbol{\beta} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \rho^{gls} \end{pmatrix} \boldsymbol{\beta}, \end{aligned}$$

where $\rho^{gls} = trace\left[\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_X\right] / trace\left[\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U)\right]$.

As in the OLS case, this simple formula also yields a number of interesting observations

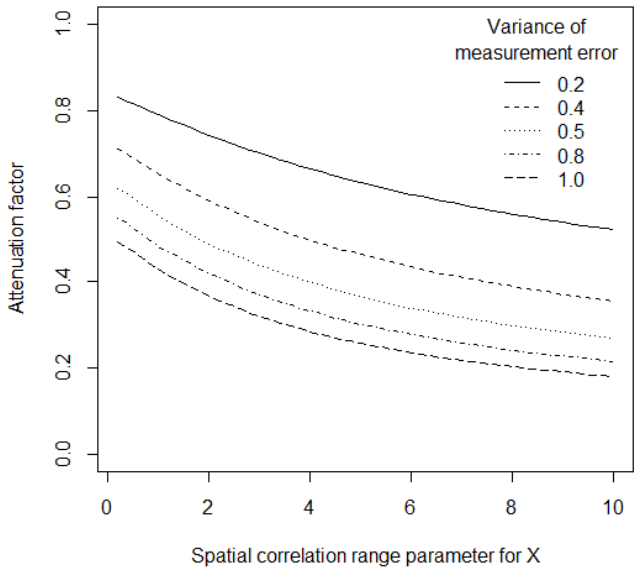


Figure 4.1: Attenuation factor associated with varying degree of measurement error.

with important practical implications. First of all, because we can write

$$\rho^{gls} = \left[1 + trace\left(\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_U\right)/trace\left(\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_X\right)\right]^{-1}, \tag{4.6}$$

it follows that there will always be an attenuation of the estimated regression coefficient towards the null.

The following figure shows the attenuation factor associated with fitting generalized least squares, ρ^{gls}, under the assumption that $\mathbf{X}$ and $\boldsymbol{\epsilon}$ each have unit variance and an exponential spatial covariance structure with the correlation between two observations a distance h units apart is given by $Cor(h) = exp(-h/\tau)$, where τ denote the range.

Each line in Figure 4.1, corresponds to a unique value of σ_U^2, the measurement error variance. The x-axis in the figure varies according to the value of the range parameter τ_x, which reflects the strength of the spatial correlation in the true covariate $\mathbf{X}$. All calculations in the figure assume that there is zero spatial correlation in the measurement error term, $\mathbf{U}$. Note that, as the range parameter goes to zero ($\tau_x \to 0$), the attenuation factor becomes identical to that which would be obtained if OLS were used instead of GLS. Of course, these results could change in the presence of other covariates in the model (see Zeger et al. 2000).

From equation (4.6), it is clear that the two attenuation factors, ρ^{gls} and ρ^{ols}, will equal the familiar attenuation factor under OLS modeling for independent data, namely $\rho = \sigma_X^2/(\sigma_X^2 + \sigma_U^2)$, under a variety of circumstances, including:

1. $\boldsymbol{\Sigma}_X = \sigma_X^2 \boldsymbol{I}$ and $\boldsymbol{\Sigma}_U = \sigma_U^2 \boldsymbol{I}$. That is, there is no spatial correlation in X or U and both random variables have homogeneous variance.

2. When the degree of spatial correlation is the same for $\mathbf{X}$ and the measurement error, $\mathbf{U}$. That is, $\boldsymbol{\Sigma}_X = \sigma_X^2 \boldsymbol{R}$ and $\boldsymbol{\Sigma}_U = \sigma_U^2 \boldsymbol{R}$, where $\boldsymbol{R}$ is a spatial correlation matrix.

Note that these results hold regardless of the value of $\boldsymbol{\Sigma}_a$, the assumed correlation for the residuals in regression models. In practice, ρ^{gls} and ρ^{ols} may differ depending on how well the assumed spatial correlation structure resembles the true process of the underlying covariance structure.

In the next section, we propose several approaches to adjust for measurement error in spatial regression settings.

4.4 Bias Correction

In the previous section, we have shown that presence of measurement error in covariates attenuates estimated regression coefficient to the null. A consistent estimate of the true regression coefficient can be obtained if we can estimate the various parameters that govern the measurement error process. This is possible if we have access to a validation data set without measurement error (Carroll et al., 2006). In the context of spatial epidemiology, however, validation data are rarely available. Therefore, we need additional assumptions to estimate the components of the attenuation factor. Without such assumptions or validation data, the measurement error and the true residual error variance are not identifiable in both case. In this paper we considered two different sets of assumptions that lead to the model identifiability. The first approach assumes that the true covariate $\mathbf{X}$ is smooth and that any observed nugget effect must be measurement error. The second assumes that measurement error variance is fixed and

known over a feasible range. A sensitivity analysis is then carried out over the feasible range of known measurement error variance. Similar to Li et al. (2009), we assume that the underlying covariate process $\{X_i\}$ defined in section 2 contains all the spatial correlation and that the measurement error is pure noise i.e., $\mathbf{\Sigma}_U = \sigma_U^2 \boldsymbol{I}$. Under this assumption, the attenuation factor, from equation (4.6) becomes

$$\rho^{gls} = \left[1 + \sigma_U^2 trace\left(\mathbf{\Sigma}_a^{-1}\right) / trace\left(\mathbf{\Sigma}_a^{-1}\mathbf{\Sigma}_X\right)\right]^{-1} \quad . \tag{4.7}$$

The OLS version can be obtained from the special case where $\mathbf{\Sigma}_a = \sigma_\epsilon^2 \boldsymbol{I}$.

We examine two different bias correction strategies to obtain a consistent estimate of the regression coefficient. The first approach deals with estimation of each of the components of ρ^{gls} and the second uses a linear transformation of the error prone covariate, $\mathbf{W}$.

Both methods require knowledge of $\mathbf{\Sigma}_X$ and σ_U^2 or their estimated values. We estimated $\mathbf{\Sigma}_X$ and σ_U^2 by fitting the error prone covariate ($\mathbf{W}$) in an intercept only model with an assumed spatial correlation structure (i.e., spatial geostatistical model). Under the assumption that measurement error is pure noise and $\mathbf{\Sigma}_X$, is a smooth spatial covariate with no nugget, the above model gives us a maximum likelihood estimate of the nugget effect in $\mathbf{W}$, which corresponds to σ_U^2. And Σ_X was estimated from the estimated covariance matrix by subtracting the measurement error variance. Similarly, fitting $\boldsymbol{Y}$ on $\mathbf{W}$ with spatial correlation structure give us a maximum likelihood estimate of the underlying residual covariance structure, $\mathbf{\Sigma}_\epsilon$. The first method additionally requires an estimate of $\mathbf{\Sigma}_a$.

4.4.1 Method I: Method of Moments

This method involves post analysis adjustment of the estimated regression coefficient using an estimate of the attenuation factor. Ignoring the measurement error and performing a likelihood analysis under the assumed covariance structure of $\boldsymbol{Y}|\mathbf{X}$ using $\mathbf{W}$ instead of $\mathbf{X}$ results in estimates denoted by $\hat{\boldsymbol{\beta}}^{ols}$ or $\hat{\boldsymbol{\beta}}^{gls}$ depending on whether ordinary or generalized least squares has been used. Let $\hat{\beta}_1$ be the estimate of the corresponding slope from the above regression, where for the ease of exposition we leave

off the superscript 'ols' or 'gls'. We have shown that its limiting value is $\rho\beta_1$. Denote its variance by σ_*^2. We then define an adjusted estimate, $\hat{\beta}_1^{adj} = \hat{\beta}_1/\hat{\rho}$ where $\hat{\rho}$ is an estimate of the attenuation factor defined at equation (4.7) and where the estimated variance $\hat{Var}(\hat{\beta}_1^{adj}) = \hat{\rho}^{-2}\sigma_*^2$. An estimate of ρ is obtained by substituting $\hat{\sigma}_U^2$, $\hat{\mathbf{\Sigma}}_X$ and $\hat{\mathbf{\Sigma}}_a$ in equation (4.7).

4.4.2 Method II: Transformation Method

Recall from equation (4.2) that $E(\boldsymbol{Y}|\mathbf{W}) = \beta_0\mathbf{1} + \beta_1(\boldsymbol{I} - \mathbf{\Lambda})\boldsymbol{\mu}_X + \beta_1\mathbf{\Lambda}\mathbf{W}$, where $\mathbf{\Lambda} = \mathbf{\Sigma}_X(\mathbf{\Sigma}_X + \mathbf{\Sigma}_U)^{-1}$. This suggests the use of a linear transformation of $\mathbf{W}$ to achieve an appropriate linear regression model that can be fitted to yield a consistent estimate of $\boldsymbol{\beta}$. Specifically, letting $\boldsymbol{T} = \boldsymbol{\mu}_X + \mathbf{\Lambda}(\mathbf{W} - \boldsymbol{\mu}_X)$, it follows that $\boldsymbol{T} \sim (\boldsymbol{\mu}_X, \mathbf{\Lambda}\Sigma_X)$ and $Cov(\boldsymbol{Y}, \boldsymbol{T}) = \beta_1\mathbf{\Lambda}\Sigma_X$. Hence using the joint normality of $\mathbf{W}$ and $\boldsymbol{Y}$, we have $E(\boldsymbol{Y}|\boldsymbol{T}) = \beta_0 + \beta_1\boldsymbol{T}$, and $Var(\boldsymbol{Y}|\boldsymbol{T}) = Var(\boldsymbol{Y}|\mathbf{W}) = \Sigma_\epsilon + \beta_1^2(\boldsymbol{I} - \mathbf{\Lambda})\Sigma_X$.

Define $\tilde{\boldsymbol{T}}$ as the estimator of $\boldsymbol{T}$, obtained by substituting in consistent estimates of $\boldsymbol{\mu}_X$ and $\mathbf{\Lambda}$. The outcome $\boldsymbol{Y}$ can then be regressed on $\tilde{\boldsymbol{T}}$, with an assumed spatial correlation structure, via a linear mixed model to obtain a consistent estimate of β_1 and corresponding standard error.

4.5 Simulation Study

We conducted a simulation study to evaluate the finite sample properties of two methods proposed in the previous section to adjust for measurement error. We simulated 100 sample locations randomly within a $d \times d$ rectangular grid, where d is taken to have a value of either 40 or 80. Specifically, the i^{th} random sample location s_i was generated by simulating two coordinates (e.g., latitude and longitude) from a Uniform$[0, d]$ distribution. Given the set of s_i's, the unobserved true covariate $\mathbf{X}$ was generated with mean 0 and covariance matrix $\mathbf{\Sigma}_X$, where $\mathbf{\Sigma}_X$ was assumed to have an exponential correlation structure with unit variance. This implies that the correlation between two observations distance h units apart is $(1 - \eta_x) * \exp(-h/\tau_x)$, where τ_x is the range parameter and η_x characterizes the so called nugget effect. We considered three different

range parameters ($\tau_x = 1, 5, 10$) resulting in minimal, moderate and high correlation among the values of X's with a nugget effect of $\eta_x = 0.1$.

The observed error-prone versions, $\mathbf{W}$, of the true covariate were generated by adding Gaussian noise with variance σ_U^2 to $\mathbf{X}$. Outcome data, $\boldsymbol{Y}$, were then generated according to equation (4.1), the slope and intercept parameter are taken as $(\beta_0, \beta_1)^T = (1, 2)^T$ and the error variances were generated using a similar exponential correlation structure as $\boldsymbol{\Sigma}_X$, but with different range parameters. We also add a random Gaussian noise to the residual error variance (nugget effect).The variance parameter and the nugget for the residual error was taken as 0.5 and 0.1, respectively.

To generate simulated data with exponential spatial correlation and also in model fitting, we used the *nlme* package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2013). To extract the covariate matrices from the object of lme fit we used the *mgcv* package (S. Wood, 2006) in *R* (R Core Team, 2013).

To study the performance of our proposed methods under various degree of correlations within the rectangular grid and for various values of the measurement error variance, we simulated data based on various combinations of measurement error variances (σ_U^2=0.0, 0.2 and 0.5). To simplify our presentation, only the results with measurement error variance $\sigma_U^2 = 0.2$ in an 80×80 grid scales are illustrated in Table 4.1. In general, the results obtained by varying the measurement error and/or size of the grid are similar.

Table 4.1 shows the average of the estimated regression coefficients, empirical standard errors and average of the estimated standard errors under 9 different combinations of spatial correlation in the covariate $\mathbf{X}$ and in the error for the model $\boldsymbol{Y}$ given $\mathbf{X}$, based on 1000 simulations. The first column of Table 4.1 specifies the combination of range parameters $(\tau_X, \tau_{\boldsymbol{\epsilon}})$ used in that particular simulation. The 2nd and 3rd columns shows the estimated regression parameters under ordinary least squares based on using the true covariate $\mathbf{X}$ and the error prone covariates $\mathbf{W}$, respectively (see equation 4.4). The 4th column shows the results from fitting a linear mixed model (using the 'lme' function in R) with assumed exponential correlation structure, but without adjusting for measurement error. The 5th and 6th columns present the bias corrected estimates of the regression parameter $\boldsymbol{\beta}_1$ using Method I and Method II, respectively. The next two

Table 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.

Range*	OLS		lme	Bias corrected lme using				
				estimated σ_u^2		true σ_u^2		true $\mathbf{\Lambda}$
$(\tau_X, \tau_{\boldsymbol{\epsilon}})$	using X	using W	using W	Method I	Method II	Method I	Method II	Method II
Estimated coefficient								
(1,1)	1.999	1.689	1.683	1.867	1.838	2.039	2.000	1.995
(1,5)	2.000	1.692	1.682	1.886	1.844	2.048	2.001	1.997
(1,10)	2.001	1.691	1.681	1.889	1.849	2.038	2.001	1.995
(5,1)	1.999	1.682	1.665	2.075	1.987	2.039	1.990	1.995
(5,5)	2.002	1.687	1.641	2.106	1.998	2.051	1.988	1.998
(5,10)	2.004	1.687	1.630	2.106	2.000	2.039	1.986	1.997
(10,1)	2.000	1.666	1.638	2.113	2.013	2.040	1.990	1.996
(10,5)	2.002	1.668	1.584	2.151	2.028	2.050	1.984	1.996
(10,10)	2.005	1.670	1.562	2.173	2.048	2.037	1.982	1.997
Empirical Standard error								
(1,1)	0.075	0.097	0.099	0.473	0.321	0.138	0.126	0.115
(1,5)	0.077	0.099	0.099	0.614	0.331	0.147	0.127	0.115
(1,10)	0.077	0.099	0.095	0.676	0.349	0.142	0.123	0.112
(5,1)	0.079	0.104	0.107	0.583	0.409	0.145	0.131	0.120
(5,5)	0.091	0.110	0.113	0.753	0.508	0.178	0.139	0.123
(5,10)	0.098	0.116	0.115	0.768	0.512	0.172	0.143	0.127
(10,1)	0.083	0.114	0.121	0.494	0.334	0.176	0.139	0.125
(10,5)	0.102	0.126	0.142	0.616	0.418	0.247	0.159	0.137
(10,10)	0.117	0.134	0.145	0.692	0.469	0.210	0.165	0.142
Average of estimated standard errors								
(1,1)	0.075	0.100	0.099	0.110	0.109	0.120	0.119	0.118
(1,5)	0.074	0.100	0.096	0.108	0.107	0.118	0.115	0.114
(1,10)	0.073	0.099	0.093	0.105	0.104	0.113	0.112	0.110
(5,1)	0.076	0.101	0.101	0.127	0.124	0.125	0.120	0.120
(5,5)	0.075	0.100	0.101	0.130	0.126	0.126	0.121	0.121
(5,10)	0.073	0.099	0.098	0.127	0.124	0.123	0.119	0.119
(10,1)	0.078	0.103	0.104	0.135	0.129	0.131	0.124	0.123
(10,5)	0.077	0.103	0.106	0.145	0.137	0.138	0.129	0.129
(10,10)	0.075	0.102	0.104	0.146	0.139	0.137	0.128	0.128

Range*- $(\tau_X, \tau_{\boldsymbol{\epsilon}})$values of the range parameter following exponential correlation in $\mathbf{X}$ and the error term in the model on $\boldsymbol{Y}$ respectively.

columns of the table represent the results from Method I and Method II when true measurement error variances were used instead of estimated values. That is, results in column 5 use the estimated measurement error and column 7 uses the true value of the measurement error variance under Method I. Similarly, columns 6 and 8 represent the results obtained using Method II based on the *estimated* and *true* measurement error variances, respectively. The last column of the table shows results from Method II when all the components of $\mathbf{\Lambda}$ were calculated using true values (i.e., values used in data generation).

The simulation results confirmed that the degree of bias for linear mixed model with

error prone covariate varies with the strength of the spatial correlation structure of covariate as well as residuals. However, our proposed bias correction methods perform well in terms of providing consistent estimates of the regression coefficient. Both methods under-estimate the true regression coefficient when measurement error is estimated and there is very low correlation in $\mathbf{X}$. This makes sense because the nugget effect in $\mathbf{X}$ is non-identifiable in that setting. This is because the assumption that the true covariate $\mathbf{X}$ is smooth is no longer valid, hence estimates are not reliable in such situations.

To assess the sensitivity of the true spatial correlation structure on parameter estimation, we run a simulation with misspecified spatial correlation structure. In this simulation we generated data using an exponential covariance structure, but fitted under the assumption of a Gaussian covariance structure. Figure 4.2 shows the distribution of estimated coefficients when estimated and true values of σ_u^2 are used with Method I (a-b) and Method II (c-d), under different range parameters combined with true and misspecified covariate structure. For each combination of range parameters, the first boxplot (from left) represents results from the misspecified covariance structure. The results obtained for the other combination of range parameters (not shown in this figure) are similar.

Our simulation results illustrate that proposed methods are quite robust in case of misspecification of underlying covariance structure. However the accuracy of the methods depends largely on the value of σ_u^2. Therefore, a close estimate of σ_u^2 to the true value is more important than having a good estimates of underlying covariance structure. Hence, we recommend a sensitivity analysis be used in practice.

To evaluate the performance of the proposed method under small samples, we also conducted a simulation with a sample size of 50. In this case, the estimates obtained from Method I are slightly upwardly biased with higher standard errors. However Method II adjusts for bias quite well and provides a reliable estimates of standard errors. In general, considering all the simulation scenarios, the transformation method (Method II) outperforms the method of moments (Method I) in terms of standard errors.

We also run another set of simulations to ascertain whether the spatial configuration of the point locations is an important feature in determining the effectiveness of bias

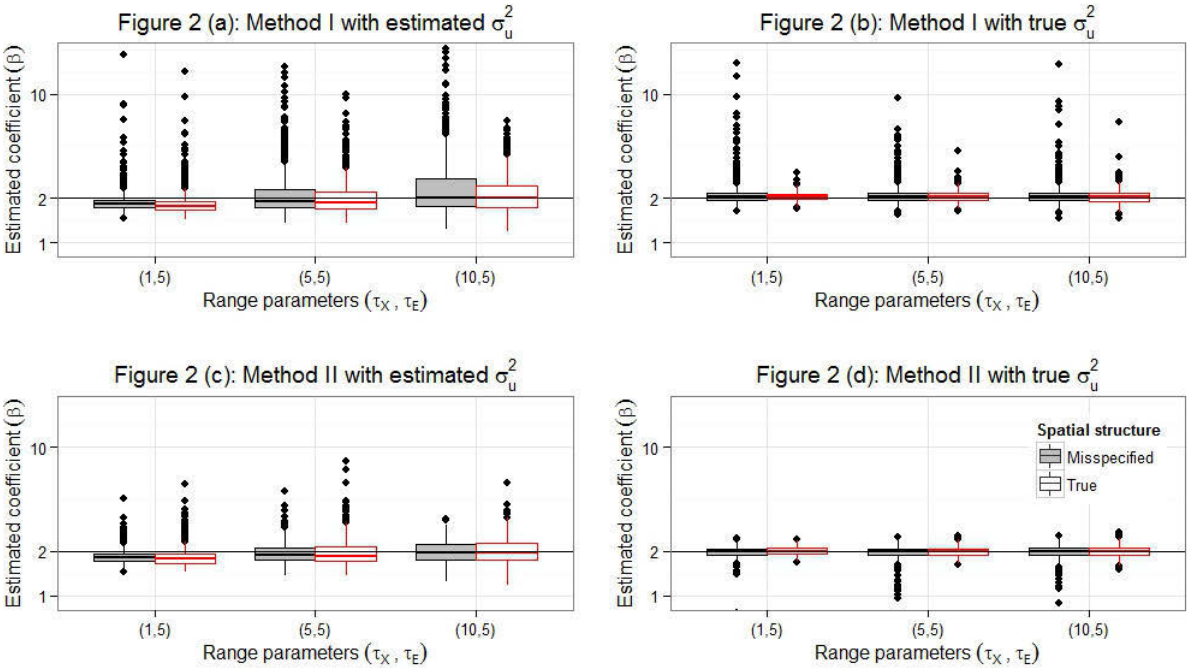


Figure 4.2: Distribution of estimated coefficient when estimated and true value of σ_u^2 used with Method I (a-b) and Method II (c-d) under different range parameters combinations with true and misspecified covariate structure.

correction. We generated data based on a cubic function with Gaussian noise in a 80×80 grid rather than uniformly distributed within the grid. Our simulation results (not shown) show that the spatial configuration does affect the estimates of measurement error variance, but not the estimated spatial structure. Thus the bias correction methods remains consistent when true measurement error is known. In practice, if measurement error is unknown, the best way is to run a sensitivity analysis within the reasonable range of measurement error variance.

4.6 Analysis of Ischemic Heart Disease Data

Data on Ischemic Heart Disease (IHD) were collected from all hospitals in New South Wales (NSW), Australia between July 1, 1994 and June 30, 2002. A detailed description of the data has been given elsewhere (Burden et al., 2005). Briefly, patients who were admitted to the hospitals via the emergency room and discharged with a diagnosis of IHD were considered as acute IHD cases. Data also includes patient age, gender and geographic location reported via postcode of residence. Data from 579 postcodes were

included in the analysis. IHD event data were linked with the Census data which contains age and gender-specific population counts. SEIFA (Socio-Economic Indexes For Areas) scores and centroid co-ordinates (latitude and longitude) for each postcode were obtained from Australian Bureau of Statistics (ABS). Since temporal patterns were not our main concern in this study, we averaged the 8 year SEIFA scores and aggregated values of the population size and number of IHD admitted cases for each postcode. We then calculated age-sex adjusted standardized incidence ratios (SIR) by dividing the observed number of IHD cases by the age-sex adjusted expected IHD cases (Breslow & Day, 1987).

Li et al. (2009) analyzed square root transformed standard mortality ratios (SMR) to make them more normal distributed. However, we found that untransformed SIR values more closely approximated the normally distribution and hence we did not transform. We fit model (4.1) assuming an exponential correlation structure for data observed for each postcode,with distance based on latitude and longitude of each postcode centroid. As Burden et al. (2005) noted, the principal component analysis that was used to derive the SEIFA score only accounts for about 30 percent of the total variation of the component used. Therefore, it is likely that the SEIFA score is subject to substantial measurement error. We standardized the SEIFA scores to have a mean of zero.

Table 4.2: Analysis of Ischemic Heart Disease Data in NSW, Australia under different specification of measurement error

Methods	Estimates for SEIFA	
Ignoring measurement error	$\hat{\boldsymbol{\beta}}$	se($\hat{\boldsymbol{\beta}}$)
Ordinary Least Squares	-0.062	0.014
LME with spatial correlation	-0.141	0.015
Accounting for measurement error bias		
Method I	-0.377	0.041
Method II	-0.278	0.015

The results of our analysis are given in Table 4.2. The naive analysis ignoring spatial correlation, suggests a significant protective effect associated with higher SEIFA values ($\hat{\boldsymbol{\beta}}_{SEIFA}$=-0.062, SE=0.014). Analysis via a linear mixed model accounting for spatial correlation also suggests that the effect is very strong ($\hat{\boldsymbol{\beta}}_{SEIFA}$=-0.141 with SE=0.015). However, the magnitude of the effect is much larger.

We applied our bias correction methods on the result obtained from the linear mixed

model. The linear mixed model of SEIFA based on a intercept only model with assumed exponential spatial correlation suggests the estimate of measurement error variance as 0.28. Both methods suggest a strongly significant effect of SEIFA ($\hat{\boldsymbol{\beta}}^{adj}_{SEIFA}$=-0.377 and -0.278 with SE=0.041 and 0.015 respectively). A large difference in the estimated standard error for Method I and Method II is observed. The Method II account for less uncertainty than Method I.

In practice, a bootstrap procedure can be used to calculate the standard error for Method I. We implement a block bootstrap procedure by leaving one block at each iteration. Blocks are automatically selected using the cluster separation method *"clara"* (Kaufman & Rousseeuw, 2005) in *R* (R Core Team, 2013). Specifically, this method selects k representative objects in the data set, where k is the number of clusters. The remaining objects are then assigned to the nearest representative object to form a cluster. The representative objects are selected in such a way that the average distance of the representative objects to all other objects in the same cluster is minimized. Our results show that the difference in estimated standard error reduces with large number of blocks while estimated standard error for Method II remain unchanged (result not shown).

Since we do not have a validation dataset and thus cannot test the assumption underlying the bias correction methods, we conduct a sensitivity analysis to help in the interpretation of our results. We conducted sensitivity analysis varying measurement error variance, σ^2_U from 0.0 (naive) to 0.40. The result of the sensitivity analysis is presented in Figure 4.3.

As measurement error variance, σ^2_U increases, the estimates obtained by method of moments also decreases. The estimates obtained using transformation methods also decreases until when the assumed measurement error variance is less than the estimated measurement error variance and then increases. We note that the transformation method appears to give stable results over the range of σ^2_U.

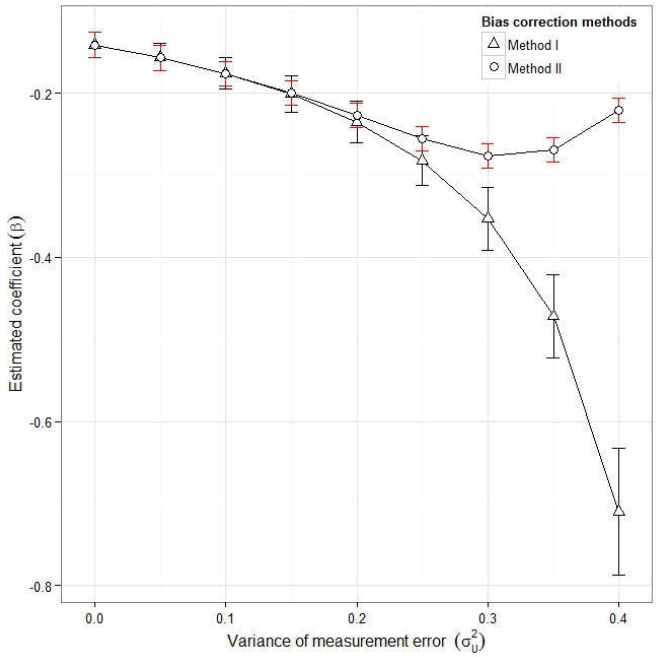


Figure 4.3: Sensitivity analysis for IHD data. The assumed measurement error variance varied between 0 (naive) and 0.40

4.7 Discussion

In this paper, we have developed a framework to quantify the bias induced in estimated regression coefficients when covariates are measured with error in spatial regression settings. Both analytic and simulations results suggest that naive analysis that ignore measurement error will attenuate estimated regression coefficients towards the null hypothesis of no effect. Our results extend classical measurement error theory in that the amount of attenuation depends on the degree of spatial correlation in both the covariate of interest as well as the assumed random error from the regression model. These results explain why the results from spatial regression modeling are often so sensitive to the assumed model error structure. We proposed two different strategies to obtain consistent estimates of the regression coefficients of interest in the presence of covariate measurement error. These strategies include 1) post hoc adjustment of estimated regression coefficient via an estimate of the attenuation factor and 2) a linear transformation of the error prone covariates that can then be analyzed to yield consistent results. We found that both methods perform well, though the second method tends to be less variable and hence preferable in practice. We illustrated the proposed approaches using the analysis of Ischemic Heart Disease data. There are a number of

areas where future study would be useful.

Our analytic results are similar to those of Li et al. (2009) who also consider asymptotic bias associated with spatial regression analysis involving covariate measurement error. They also propose an adjusted analysis based on an EM algorithm. However, their approach is difficult to apply, especially for large data sets. In contrast, our proposed approaches can be easily implemented using readily available packages such as lme in R (R Core Team, 2013). While Li et al. (2009) demonstrate via simulation that their method works well, they use the true values of the measurement error variances and did not consider the setting where measurement error variances are estimated. While our simulations confirm the reliability of our proposed methods in settings where the measurement error variance can be assumed known, we also suggest an approach to estimating the measurement error variance. As in the classical measurement error context, estimation of measurement error variance requires either additional assumptions or additional information such as validation or replicate data. We showed how the measurement error variance could be estimated under the assumption that the true covariate of interest is smooth and hence that any estimated nugget effect can be interpreted as measurement error. As expected, our simulations suggest that the performance of our proposed bias correction methods decline when the measurement error variance is estimated instead. Method I performs more poorly than Method II since that latter requires the estimation of fewer model parameters. Our results suggest that having some knowledge of the magnitude of measurement error is important and in practice we suggest the use of a sensitivity analysis that varies the assumed values of the measurement error variance over a feasible range.

One observation from our simulation is that the use of an estimated measurement error variance estimates from the data leads to under estimation of the regression coefficients when there is minimal spatial correlation in the covariate. This makes sense since most of the covariate variability will be absorbed in to the estimated nugget effect. As expected, we found that the situation improved when we used a smaller grid size (see also, Bell and Grunwald 2004). Use of an estimated measurement error variance also led to much greater discrepancies between the average of our estimated standard errors and the empirical standard errors derived from the simulation. In contrast, these were much closer when the true values of the measurement error variances were used. These

observations suggest that having knowledge of the true measurement error variance is crucial not only in obtaining consistent estimates of the regression coefficients, but also in terms of estimating standard errors and conducting appropriate inference. Again, we recommend use of a sensitivity analysis in practice.

The estimated standard errors of regression coefficients underestimate the empirical standard errors. This might be because our formulae for the standard errors of the adjusted estimated regression coefficients do not fully account for the uncertainty associated with the estimation of the variance component. It would be of future research interest to conduct an asymptotic bias analysis on estimation of variance components in spatial models when measurement error is ignored and the error-prone covariate follows a spatial model. In practice, a block bootstrap procedure can be used to obtain appropriate standard errors.

Our heart disease example demonstrated a substantial increase in the rates of Ischemic Heart Disease as the level of SEIFA (Socioeconomic Indicators for Areas) measured at the postcode level decreased. The magnitude of the effect increased after adjusting for measurement error. Our results are consistent with social epidemiology literature (see systematic review by Pickett and Pearl 2001) that suggests that low socio-economic status leads to increased rates of a wide variety of health outcomes. While it is tempting to interpret these results at an individual level, it is important to remember that doing so may result in ecological bias (Sheppard, 2003). Prentice and Sheppard (1995) showed that using group level covariates in the analysis reduces the effects of error in the measurement of covariates at the individual level. However, Greenland (2001) and Jackson et al. (2006) noted that ecological covariates are subject to non-random survey errors and may not be addressed by aggregation of group level analysis of covariates. Moreover, in many research areas, group level data are the only available source for analysis. Air pollution epidemiology provides a classic example, since individual measurements of air pollution studies rarely collected and instead are estimated based on neighborhood monitoring and other sources (Sheppard et al., 2012). Consequently, air pollution exposures are typically measured with error and it would be useful to consider the impact of this error on subsequent effect size estimates.

In our simulation we have considered only a single covariate measured with error in a

spatial linear mixed model with Gaussian error. It would be of interest to explore the effect of covariate measurement error in the presence of multiple covariates and also omitted covariates. Future work can also be done on extending our formulation to the spatial generalized linear mixed model with non-Gaussian outcomes. However, such explorations are beyond the scope of this present paper. In this paper we consider a full parametric approach to adjust for covariate measurement error. Indeed, Ruppert, Wand, and Carroll (2009) argued that penalized splines are the most effective method for correcting the covariate measurement error in case of independent data. So it is of natural interest to extent the spatial regression model with measurement error to a semi-parametric framework. In the next chapter we have considered such as approach.

In light of the increasing popularity of multi-level models that include both individual and area-specific covariates, it is important that practitioners be aware of the importance, not only of careful modeling of the mean function, but also of accounting for measurement error and appropriate spatial structure of their data.

Appendix 4A

The ordinary least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{ols} = (\mathbf{W}_*^T\mathbf{W}_*)^{-1}\mathbf{W}_*^T\boldsymbol{Y}. \tag{4.8}$$

with $\mathbf{W}_*$ defined in the text. Under the true model, $\boldsymbol{Y} = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{ols} &= (\mathbf{W}_*^T\mathbf{W}_*)^{-1}\mathbf{W}_*^T\boldsymbol{Y} \\
&= (\mathbf{W}_*^T\mathbf{W}_*)^{-1}\mathbf{W}_*^T(\mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \mathbf{W}_*^T\mathbf{W}_*)^{-1}\mathbf{W}_*^T\mathbf{X}_*\boldsymbol{\beta} + (\mathbf{W}_*^T\mathbf{W}_*)^{-1}\mathbf{W}_*^T\boldsymbol{\epsilon} \\
&= \left(\frac{\mathbf{W}_*^T\mathbf{W}_*}{n}\right)^{-1}\left(\frac{\mathbf{W}_*^T\mathbf{X}}{n}\right)\boldsymbol{\beta} + \left(\frac{\mathbf{W}_*^T\mathbf{W}_*}{n}\right)^{-1}\left(\frac{\mathbf{W}_*^T\boldsymbol{\epsilon}}{n}\right)
\end{aligned}$$

Now under certain regularity conditions (Zheng & Zhu, 2012) and by the weak law of large numbers, $\left(\frac{\mathbf{W}_*^T\mathbf{W}_*}{n}\right) \xrightarrow{p} cov(\mathbf{W}_*)$, $\left(\frac{\mathbf{W}_*^T\mathbf{X}_*}{n}\right) \xrightarrow{p} cov(\mathbf{X}_*)$ and $\left(\frac{\mathbf{W}_*\boldsymbol{\epsilon}}{n}\right) = \left(\frac{\mathbf{X}_*\boldsymbol{\epsilon}}{n} + \frac{\mathbf{U}_*\boldsymbol{\epsilon}}{n}\right) \xrightarrow{p} 0$. It follows that, $\tilde{\boldsymbol{\beta}}^{ols} \xrightarrow{p} [cov(\mathbf{W}_*)]^{-1}\, cov(\mathbf{X}_*)\boldsymbol{\beta}$. Since $\mathbf{W}_*$ and $\mathbf{X}_*$ have first column equals to $\mathbf{1}$ corresponding to intercept of the model and assuming $\mu_X = 0$, it follows,

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}^{ols} &\xrightarrow{p} \begin{pmatrix} 1 & 0 \\ 0 & trace\,(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U) \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & trace\,(\boldsymbol{\Sigma}_X) \end{pmatrix}\boldsymbol{\beta} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & \rho^{ols} \end{pmatrix}\boldsymbol{\beta},
\end{aligned}$$

where $\rho^{ols} = trace\,(\boldsymbol{\Sigma}_X)\,/trace\,(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U)$.

Chapter 5

Spatial Regression with Covariate Measurement Error: A Semiparametric Approach

Summary

Spatial data have become increasingly common in epidemiology and public health research thanks to advances in GIS (Geographic Information Systems) technology. In health research, for example, it is common for epidemiologists to incorporate geographically indexed data into their studies. In practice, however, the spatially-defined covariates are often measured with error. Naive estimators of regression coefficients are attenuated if measurement error is ignored. Moreover, the classical measurement error theory is inapplicable in the context of spatial modeling because of the presence of spatial correlation among the observations. We propose a semi-parametric regression approach to obtain bias corrected estimates of regression parameters and derive their large sample properties. We evaluate the performance of the proposed method through simulation studies and illustrate using data on Ischemic Heart Disease (IHD). Both simulation and

The content of this chapter is published as: Huque, MH; Bondell, HD; Carroll, RJ and Ryan. LM. (2016). Spatial regression with covariate measurement error: A semiparametric approach. *Biometrics*, DOI: 10.1111/biom.12474. This research was also presented at the Joint Statistical Meetings, Boston, 2014.

practical application demonstrate that the proposed method can be effective in practice.

5.1 Introduction

With the rapid growth of Geographic Information Systems (GIS), it is now common for epidemiologists to incorporate spatially indexed data into their studies (Elliott & Wartenberg, 2004). Analysis of such data, however, is complicated by correlations among neighboring observations. Although there are well known statistical methods to adjust for spatial correlation, relatively little has been done in the context of spatial modeling when the covariate of interest is measured with error. In the case study that motivates this study, Australian researchers explored the relationship between the SEIFA index (an area-based measure of socio-economic status produced by the Australian Bureau of Statistics) and acute hospitalization for Ischemic Heart Disease (IHD) in New South Wales, Australia (Burden et al., 2005). Multivariate regression models suggest a significantly negative association between SEIFA and IHD, implying that heart disease rates increase with social disadvantages. However, the strength of association might be attenuated due to the fact that the SEIFA index is constructed using principal component analysis, therefore, is highly likely to be measured with error (Huque, Bondell, & Ryan, 2014).

Many articles have appeared in the literature over the years on covariate measurement error in the context of independent data (Carroll et al., 2006; Fuller, 1987). However, relatively few have addressed the specific context of spatial modeling. Bernadinelli et al. (1997) and Xia and Carlin (1998) presented a spatio-temporal analysis of spatially correlated data with errors in covariates, in the context of disease mapping. They empirically studied several alternative measurement error models using a Gibbs algorithm. Li et al. (2009) derived asymptotic bias expressions for estimated regression coefficients in the context of a spatial linear mixed model. They showed that the regression estimates obtained from naive use of an error prone covariate are attenuated, while variance component estimates are inflated.

Recently, Huque et al. (2014) confirmed the findings of Li et al. (2009) and derived expressions for the bias when measurement error is ignored. They proposed two different

strategies for obtaining consistent estimates: (i) correcting the estimate using an estimated attenuation factor; and (ii) using an appropriate transformation of the error prone covariate. They showed that both bias correction methods work reasonably well, however, the standard error is underestimated in the case when measurement error variances are estimated from the data. Moreover, their approach is fully parametric. Indeed, Ruppert et al. (2009) argued that penalized splines are the most effective method for correcting the covariate measurement error in case of independent data. So it is of natural interest to extent the spatial regression model with measurement error to a semi-parametric framework.

In this paper we propose a joint modeling approach to assess the relationship between a covariate with measurement error and a spatially correlated outcome in a semi-parametric regression modeling context. Our approach contrasts with what is commonly assumed in the measurement error context, namely that some form of validation data are available. Underlying our approach is the critical assumption that the true, but unobserved covariate is smooth and that any random fluctuations from this smooth surface represent measurement error. This assumption makes our model identifiable by representing the unknown true covariate with a linear combination of spline basis functions (Xun, Cao, Mallick, Maity, & Carroll, 2013; Yu & Ruppert, 2002). We use penalized least squares which makes the estimation of parameters and inference straightforward. We develop asymptotic theory for the estimated parameters and provide both model based and simulation based standard error estimates. Our simulation results reveal that the proposed method works well in obtaining consistent estimates of the true regression coefficient in the presence of measurement error. Our approach is computationally efficient and stable and can be implemented using standard nonlinear least squares software.

The structure of the paper is as follows: Section 5.2 describes our model formulation, estimation and inference procedures. Section 5.3 presents the data generation process and results from the simulation study. In section 5.4 we present an application of the proposed method to data on Ischemic Heart Disease (IHD). We conclude with general discussion in section 5.5. The Appendix 5A gives detailed proofs, as needed.

5.2 Model

Suppose that X_i represents the true covariate of interest measured at geographical location, $S_i \in \mathbb{R}^2$, $i = 1, ..., n$ and suppose that X_i is related to an outcome Y_i, according to a spatial linear model:

$$Y_i = \beta_0 + \beta_1 X_i + G_1(S_i) + \epsilon_i, \tag{5.1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, ...\epsilon_n)^T \sim N(0, \sigma_\epsilon^2)$ and $\{G_1(S_i) : S_i \in \mathbb{R}^2\}$ is an unknown function that captures the spatial correlation, for now kept arbitrary. Further assume that ϵ_i and $G_1(S_i)$ are independent of each other and of the true covariate X_i (Cressie, 1993). In practice, the outcome might also be related to other covariates and it is straight forward to extent model (5.1) to include these. However, for simplicity, we only consider a single covariate in model (5.1).

In the presence of measurement error, measurements on the true covariate X are not observed directly, instead an error contaminated version is available. Let W_i be the observed covariate for location $S_i \in \mathbb{R}^2$, $i = 1, ..., n$, related to the true covariate X_i according to a classical measurement error model:

$$W_i = X_i + U_i, \tag{5.2}$$

where $U_i \sim N\left(0, \sigma_u^2\right)$. Note that in the case of independent data, a consistent estimate of the true regression coefficient β_1 can be obtained if either the measurement error variance is known or can be estimated using a validation data set on the true covariate (X) without measurement error (Carroll et al., 2006). However, in the spatial epidemiology setting such validation data are relatively rare. We develop an alternative approach assuming that the true covariate X is smooth and can be modeled by a second smooth function, $G_2(S_i)$.

Many different choices of smoothers have been discussed in the literature, including locally-weighted running line smoothers (loess), Kernel smoothers or splines (Hastie & Tibshirani, 1990). In general, techniques based on regression splines are robust in approximating the true underling smooth functions and are relatively straight forward

from a computational perspective, but have rigorous mathematical properties (Ruppert et al., 2003; S. Wood, 2006). In this paper we also adopt such a technique using cubic thin plate splines (S. Wood, 2006).

Within this framework, the unknown smooth functions, $G_j(S_i)$, for $j = 1, 2$ are represented by linear combination of thin plate spline basis functions i.e., $G_j(S_i) = B_j^{\mathrm{T}}(S_i)\theta_j$. Here $B_1(S_i)$ and $B_2(S_i)$ are two sets of thin plate splines basis functions with dimensions $(q_1 + 3) \times 1$ and $(q_2 + 3) \times 1$, respectively, where q_1 and q_2 are the corresponding number of knots and θ_1 and θ_2 are vectors of corresponding basis coefficients.

Under the above specifications model (5.1) and (5.2) can be rewritten as

$$Y_i = B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 + B_1^{\mathrm{T}}(S_i)\theta_1 + \epsilon_i; \tag{5.3}$$

$$W_i = B_2^{\mathrm{T}}(S_i)\theta_2 + U_i. \tag{5.4}$$

Since these equations are linear with respect to a set of unknown parameters, we use penalized least squares techniques for estimation (Xun et al., 2013; Yu & Ruppert, 2002). In this method, the data, (Y, W), are fitted to two different sets of spline basis functions $B_1(S_i)$ and $B_2(S_i)$ by least squares where parameters are estimated by minimizing the usual sum of squares plus roughness penalties. That is, we minimize

$$\boldsymbol{J}(\beta, \theta_1) = n^{-1}\textstyle\sum_{i=1}^{n}\{Y_i - B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 - B_1^{\mathrm{T}}(S_i)\theta_1\}^2 + \delta_1\theta_1^{\mathrm{T}}D_1\theta_1; \tag{5.5}$$

$$\boldsymbol{J}(\theta_2) = n^{-1}\textstyle\sum_{i=1}^{n}\{W_i - B_2^{\mathrm{T}}(S_i)\theta_2\}^2 + \delta_2\theta_2^{\mathrm{T}}D_2\theta_2, \tag{5.6}$$

where the terms $\delta_1\theta_1^{\mathrm{T}}D_1\theta_1$ and $\delta_2\theta_2^{\mathrm{T}}D_2\theta_2$ are roughness penalties associated with models (5.3) and (5.4). These involve unknown regression coefficients θ_j, j=1,2, penalty parameters δ_j and penalty matrices D_j of dimension $(q_j + 3) \times (q_j + 3)$. The penalty matrices map the spline basis functions to the data whereas the penalty parameters control the amount of smoothing (Ruppert et al., 2003; S. Wood, 2006). Given knot locations $\{x^*_{j(i)} : 1, 2, ..., q_j\}$, penalty matrices have zeroes everywhere except in its lower right $q_j \times q_j$ block with $D_{j(ik)} = \left\|x^*_{j(i)} - x^*_{j(k)}\right\|^2 \log\left\|x^*_{j(i)} - x^*_{j(k)}\right\|$, for $i, k \leq q_j$.

Note that the intercept term β_0 in the model (5.1) is set to 0 in (5.3), because it is not identifiable in the presence of a non-parametric function $G_1(\cdot)$. Even so, the parameters

of these models are not completely identifiable without some additional assumptions outlined in the next section.

5.2.1 Identifiability

From the above models (5.3) and (5.4), it is evident that if $B_1(\cdot) \equiv B_2(\cdot)$, then these models are not identifiable because in this case (5.3) becomes

$$Y_i = B_2^{\mathrm{T}}(S_i)(\theta_2\beta_1 + \theta_1) + \epsilon_i.$$

Thus, we can identify only θ_2 and $\theta_2\beta_1 + \theta_1$, and cannot separate out β_1 and θ_1. To make these models identifiable, we assume that the asymptotic variability, Λ_1 and Λ_2 of two sets of basis functions $B_1(.)$ and $B_2(.)$, respectively, are different. The asymptotic variability Λ_j for j=1, 2, are the limiting values of Λ_{nj}, where

$$\Lambda_{nj} = \{n^{-1}\textstyle\sum_{i=1}^{n} B_j(S_i)B_j^{\mathrm{T}}(S_i) - \delta_j D_j\}^{-1}. \tag{5.7}$$

In practice, this requirement can be easily achieved by ensuring that the numbers of knots q_1 and q_2 are unequal.

5.2.2 Parameter Estimation

In addition to the assumption that $\Lambda_1 \neq \Lambda_2$, we also assume that the penalty parameters are small relative to the sample size as $n \to \infty$, i.e., $n^{1/2}\delta_j \to 0$ for $j = 1, 2$. This means that with large sample sizes, the estimated regression coefficients obtained using penalized least squares will be close to the OLS estimates. Thus minimizing the penalized sum of squares (5.6) and solving for θ_2, we have

$$\widehat{\theta}_2 = \Lambda_{n2} n^{-1} \textstyle\sum_{i=1}^{n} B_2(S_i)W_i, \tag{5.8}$$

where Λ_{n2} is defined in equation (5.7). A detailed derivation of $\widehat{\theta}_2$ along with it's asymptotic distribution is given in the Appendix 5A. Similarly, we can estimate θ_1 and β_1 by minimizing the corresponding penalized sum of squares (5.5). This yields (see the

Appendix 5A)

$$\widehat{\theta_1} = V_n - R_n\widehat{\theta}_2\widehat{\beta}_1 \tag{5.9}$$

$$\widehat{\beta_1} = \frac{n^{-1}\sum_{i=1}^{n} Y_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2}{\widehat{\theta}_2^{\mathrm{T}}(\mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n)\widehat{\theta}_2}, \tag{5.10}$$

where

$$\begin{aligned} V_n &= \Lambda_{n1}n^{-1}\textstyle\sum_{i=1}^{n} B_1(S_i)Y_i; \\ R_n &= \Lambda_{n1}n^{-1}\textstyle\sum_{i=1}^{n} B_1(S_i)B_2^{\mathrm{T}}(S_i); \\ \mathcal{T}_n &= n^{-1}\textstyle\sum_{i=1}^{n} B_2(S_i)B_2^{\mathrm{T}}(S_i). \end{aligned}$$

Although the above estimator of β_1 was estimated using pseudo-likelihood, it is consistent for β_1. In the next section we will establish the asymptotic properties of the estimator.

5.2.3 Asymptotic Theory

Asymptotic theory for the estimators $\widehat{\beta}_1$ is based on treating the spatial locations $S_i \in \mathbb{R}^2$ as fixed constants. Following Yu and Ruppert (2002), if $\delta_j \to 0$ as $n \to \infty$, then the bias also tends to 0 and consistency can be established. Asymptotic normality is established by the following theorem, whose proof appears in Appendix 5A.

Theorem 1 *Assume that the smoothing parameters are small relative to the sample size, i.e., $n^{1/2}\delta_j \to 0$, and the spatial correlation $G_1(.)$ and unknown covariate X are correctly represented by a finite number of splines basis functions. Then the estimate of β_1 is consistent and asymptotically normally distributed with*

$$n^{1/2}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{d} N\left(0, \sigma^2\right), \tag{5.11}$$

where

$$
\begin{aligned}
\sigma^2 &= \lim_{n\to\infty} n^{-1}\textstyle\sum_{i=1}^{n}(\sigma_\epsilon^2\mathcal{G}_{ni}^2 + \sigma_u^2\mathcal{H}_{ni}^2);\\
\mathcal{G}_{ni} &= \mathcal{D}_{ni}(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-1};\\
\mathcal{H}_{ni} &= \mathcal{A}_n\Lambda_{n2}B_2(S_i)(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-1} - \mathcal{A}_n\theta_2\mathcal{F}_{ni}(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-2};\\
\mathcal{A}_n &= n^{-1}\textstyle\sum_{i=1}^{n}\{G_2(S_i)\beta_1 + G_1(S_i)\}\{B_2(S_i) - R_n^{\mathrm{T}}B_1(S_i)\}^{\mathrm{T}};\\
\mathcal{C}_n &= \mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n;\\
\mathcal{D}_{ni} &= \{B_2(S_i) - R_n^{\mathrm{T}}B_1(S_i)\}^{\mathrm{T}}\theta_2;\\
\mathcal{F}_{ni} &= \theta_2^{\mathrm{T}}\mathcal{C}_2\Lambda_{n2}B_2(S_i) + B_2^{\mathrm{T}}(S_i)\Lambda_{n2}\mathcal{C}_n\theta_2.\\
R_n &= \Lambda_{n1}n^{-1}\textstyle\sum_{i=1}^{n}B_1(S_i)B_2^{\mathrm{T}}(S_i);\\
\mathcal{T}_n &= n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)B_2^{\mathrm{T}}(S_i).
\end{aligned}
\tag{5.12}
$$

Using this asymptotic expression we can also estimate the standard error of the estimated regression coefficient $\widehat{\beta}_1$. The next section will discuss two such options.

5.2.4 Estimating the Standard Error of $\widehat{\boldsymbol{\beta}}_1$

We first consider a model based estimate of the standard error of $\widehat{\beta}_1$ using the asymptotic theorem discussed in the previous section and then suggest a more robust estimate of standard error using simulation.

Model Based Standard Error

The model based standard errors of the estimated $\widehat{\beta}_1$ can be estimated by substituting corresponding consistent estimates of σ_ϵ^2 and σ_u^2 (defined below) into expression (5.12). Specifically,

$$
\begin{aligned}
\widehat{\sigma}_\epsilon^2 &= \frac{\sum_{i=1}^{n}\{Y_i - B_2(S_i)\widehat{\theta}_2\widehat{\beta}_1 - B_1(S_i)\widehat{\theta}_1\}^2}{n - 2\mathrm{trace}\{L_1(\delta_1,\delta_2)\} + \mathrm{trace}\{L_1(\delta_1,\delta_2)L_1^{\mathrm{T}}(\delta_1,\delta_2)\}}\\
\widehat{\sigma}_u^2 &= \frac{\sum_{i=1}^{n}\{W_i - B_2(S_i)\widehat{\theta}_2\widehat{\beta}_1\}^2}{n - 2\mathrm{trace}\{L_2(\delta_2)\} + \mathrm{trace}\{L_2(\delta_2)L_1^{\mathrm{T}}(\delta_2)\}},
\end{aligned}
$$

where the denominators are the residual degrees of freedom associated with model (5.3) and model (5.4) with smoother matrices $L_1(\delta_1,\delta_2)$ and $L_2(\delta_2)$, respectively (Ruppert et al., 2003). Define $\mathbf{B}_j = \{B_j(S_1),...,B_j(S_n)\}^{\mathrm{T}}$ for j=1,2 and $\boldsymbol{D}_n = \{D_{n1},...,D_{nn})\}^{\mathrm{T}}$.

Then the smoother matrices have the following expressions (see the Appendix 5A)

$$L_1(\delta_1, \delta_2) = n^{-1}\left\{\boldsymbol{D}_n \boldsymbol{D}_n^{\mathrm{T}}(\widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2)^{-1} + \mathbf{B}_1\Lambda_{n1}\mathbf{B}_1^{\mathrm{T}}\right\} \tag{5.13}$$

$$L_2(\delta_2) = n^{-1}\mathbf{B}_2\Lambda_{n2}\mathbf{B}_2^{\mathrm{T}}. \tag{5.14}$$

Simulated Standard Error

From (5.10), the expression for $\widehat{\beta}_1$ can be written as (see the Appendix 5A)

$$\widehat{\beta}_1 = \frac{\mathcal{A}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\}}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\mathcal{F}_{ni}U_i} + o_p(n^{-1/2}),$$

where ϵ_i and U_i are the random errors defined in models (5.1) and (5.2). Since these quantities are not directly observed, we can estimate the variance of $\widehat{\beta}_1$ by a residual bootstrap (Carroll et al., 2006).

Let M be a fairly large number, say 100, and for $b = 1, ..., M$, generate independent random samples $\epsilon_{bi} \sim \text{Normal}(0, \widehat{\sigma}_\epsilon^2)$ and $U_{bi} \sim \text{Normal}(0, \widehat{\sigma}_u^2)$ for i = 1, 2,...n. Define the b'th bootstrap estimates of β_1 as

$$\widehat{\beta}_1^b = \frac{\widehat{\mathcal{A}}_n\widehat{\theta}_2 + n^{-1}\sum_{i=1}^{n}\{\widehat{\mathcal{A}}_n\Lambda_{n2}B_2(S_i)U_{bi} + \widehat{\mathcal{D}}_{ni}\epsilon_{bi}\}}{\widehat{\theta}_2^{\mathrm{T}}\widehat{\mathcal{C}}_n\widehat{\theta}_2 + n^{-1}\sum_{i=1}^{n}\widehat{\mathcal{F}}_{ni}U_{bi}},$$

where $\widehat{\mathcal{A}}_n, \widehat{\mathcal{D}}_n, \widehat{\mathcal{C}}_n$ and $\widehat{\mathcal{F}}_{ni}$ can be estimated by substituting the appropriate quantities into expression (5.12). These estimated quantities preserve the underlying spatial structure. Therefore, the sample variance of $\widehat{\beta}_1^1, ..., \widehat{\beta}_1^M$ is a consistent estimate of the variance of $\widehat{\beta}_1$ (Efron & Tibshirani, 1993).

5.2.5 Smoothing Parameter Selection

Our main objective is to obtain a consistent estimate of the regression parameter β_1 such that it accounts for the measurement error in the covariate. However, selecting a suitable combination of the smoothing parameters (δ_1, δ_2) is a prerequisite to a good model fit. All discussion so far has assumed that these parameters are fixed and known.

To choose smoothing parameters that attempt to minimize the mean square error (prediction error), three common approaches have been discussed in the literature (Ruppert et al., 2003) (a) Generalized Cross Validation (GCV); (b) Mallow's C_p; and (c) Akaike Information Criterion (AIC). Among these methods, minimization of GCV scores is more attractive because of being rotational invariant and being computationally efficient as it avoids having to actually refit the model multiple times (S. Wood, 2006). We use the GCV criterion to estimate the smoothing parameters (δ_1, δ_2) in a two-step procedure (S. Wood, 2006). We first obtain an optimum value of δ_2 by minimizing the GCV score based on model (5.2) and then substitute this estimated value of δ_2 into (5.8) to obtain an estimate of θ_2. We then use these estimates of $\widehat{\delta}_2$ and $\widehat{\theta}_2$ in (5.13) to obtain an expression for the smoothing matrix, $L_1(\delta_1, \widehat{\delta}_2)$. Finally, we minimize the following GCV score associated with the outcome model to get an optimum value of δ_1:

$$GCV(\delta_1) = \frac{n^{-1}\sum_{i=1}^{n}\{Y_i - \widehat{Y}_i\}^2}{\{1 - n^{-1}\text{trace}\{L_1(\delta_1, \widehat{\delta}_2)\}\}^2},$$

where L_1 is defined in section 5.2.4.

5.3 Simulation Study

In this section we discuss a simulation study designed to evaluate the finite sample properties of our proposed method in the presence of covariate measurement error in spatial linear regression.

5.3.1 Data Generation

We simulate n sample locations randomly within a square, where n is the sample size. Specifically, the i^{th} random sample location S_i is generated by simulating two coordinates (e.g., latitude and longitude) from a Uniform[0,1] distribution. Given a set of simulated S_i's, the unobserved true covariate X is generated using a bivariate bump function. Specifically, the bivariate bump function is generated using the product of two univariate bump functions generated separately for each co-ordinate. That is, for each coordinate, k, we generate $X_{ik} = \frac{1}{1+a_{ik}} + 3e^{-50(a_{ik}-0.3)^2} + 2e^{-25(a_{ik}-0.7)^2}, k = 1, 2$, where

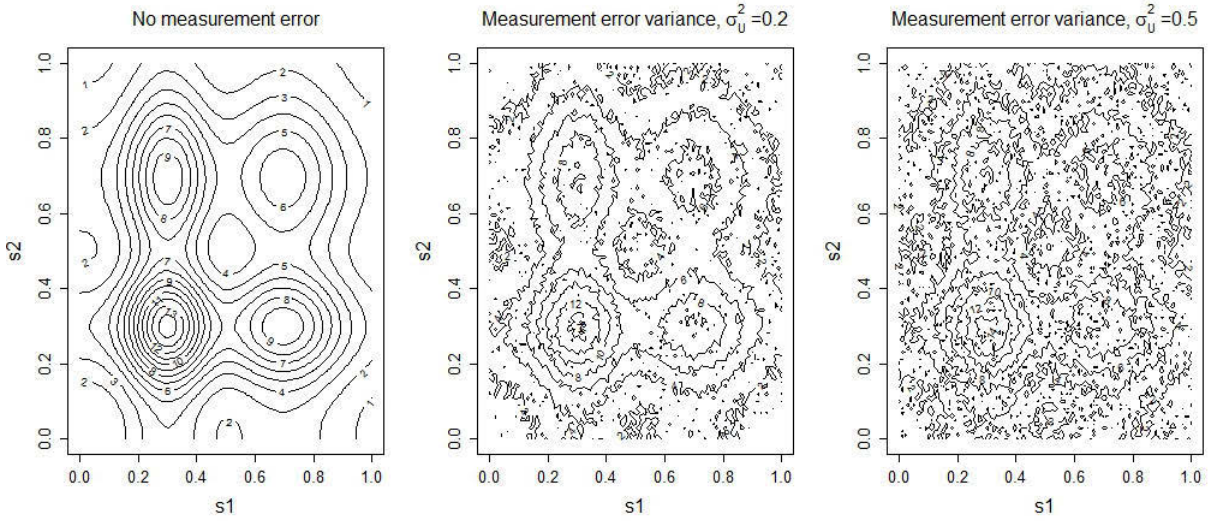


Figure 5.1: Contour plots of covariates (**X** and **W**) with different specification of measurement error variance

a_{i1} and a_{i2} are the first and second coordinates of simulated i^{th} sample location, respectively. The observed error contaminated versions, W, of the true covariate is generated by adding independent Gaussian noise with varying the measurement error variance σ_U^2 as 0, 0.25 and 0.50 to X. The contour plot associated with the true and error prone covariate is given in Figure 5.1.

As shown in the Figure 5.1, presence of measurement error adds noises to the true distribution of the smooth covariate. As a result the underlying true covariate distribution becomes obscured for higher degrees of measurement error.

The smooth spatial surface, $G_1(S_i)$, is generated to have a normal distribution with mean 0 and variance-covariance matrix $\sigma_{G_1}^2 \boldsymbol{R}$, where $\sigma_{G_1}^2 = 0.2$ and $\boldsymbol{R}$ has an exponential correlation structure with range parameter τ_{G_1} (Pinheiro & Bates, 2000). This implies that the correlation between two observations with distance h units apart is $exp(-h/\tau_{G_1})$. We considered three different range parameters (τ_{G_1}= 0.1, 0.3 and 0.5) resulting in minimal, moderate and high correlation among the values of G_1's.

Outcome data, Y, were then generated according to equation (5.1), with intercept and slope parameters are $(\beta_0, \beta_1)^T = (1, 2)^T$ and the variance parameter for the independent residual error assumed to be 0.5. We used the *nlme* package (Pinheiro et al., 2013) in ***R*** to generate exponential spatial correlation for our simulated data and in model fitting.

5.3.2 Generating Bi-variate Splines Basis Functions

We now describe the steps used to fit our proposed semi-parametric model. We generated two sets of basis functions $B_1(\cdot)$ and $B_2(\cdot)$ using bivariate thin plate spline regression basis with 125 and 150 knots for response and covariate model, respectively. We choose thin plate splines because they are not sensitive to knot locations, perform reasonably well for a basis of any given lower rank, are computationally efficient and more importantly rotational invariant (Ruppert et al., 2003; S. Wood, 2006). Unequal number of knots were chosen for $B_1(.)$ and $B_2(.)$ to make the model identifiable, (see Section 5.2.1). The number of knots for the response model (5.1) were analogous to the default number of knots [max{20,min(n/4,150)}] suggested by Ruppert et al. (2003). For the covariate model (5.2) we increased the default number of knots by 20%. Knot positions were automatically selected using the cluster separation method *"clara"* (Kaufman & Rousseeuw, 2005) in *R* (R Core Team, 2013).

Of course one could select the number of knots by another algorithm such as space filling algorithm (Nychka & Saltzman, 1998). However, implementation of this algorithm is computationally intensive. Nychka and Saltzman (1998, page-169) argued that the number of knots is flexible in the context of geo-spatial model and one needs to select large enough knots to accurately represent the underlying function while keeping the computational burden as low as possible. Furthermore, Ruppert (2002) suggest that given the GCV criteria, the number of knots is not crucial for penalized regression splines once it reaches a certain minimum value.

5.3.3 Simulation Results

The average of estimated regression coefficients along with their estimated standard errors based on 1000 simulation runs are presented in Table 5.1, assuming a sample size of 500 and varying the measurement error variance σ_U^2 between 0, 0.25 and 0.50. We estimated three different standard errors of the estimated regression coefficients, including, (i) empirical standard errors obtained by taking the standard deviation of the 1000 simulated regression coefficient estimates, (ii) the average of model based standard errors and (iii) the average of simulated standard errors defined in section 5.2.4. We

considered three different range parameters (τ_{G_1}=0.1, 0.3 and 0.5) to represent minimal, moderate and high level of spatial correlation in $G_1(S_i)$. The first column of Table 5.1 specifies the range parameter used in that particular simulation. The next four columns list the estimated regression coefficient using ordinary least squares (OLS), linear mixed models with spatial correlation structure (LME), generalized additive models (GAM) and our proposed method when the true covariate is measured without error. The second and thirds sets of four columns also list estimates obtained using the above four methods (OLS, LME, GAM and proposed method) with measurement error variances 0.25 and 0.50, respectively. Except for our proposed method, all of these methods produce naive estimates of regression coefficient.

In the absence of measurement error, OLS, LME, GAM and our method all give similar results. As the degree of measurement error increases, OLS, LME and GAM all exhibit bias, though the degree of bias varies. All naive standard error estimates ignoring covariate measurement error severely underestimate the empirical standard errors. In contrast, our proposed bias correction method performs well even if the degree of bias for generalized additive model with error prone covariate varies (range: 0.99-1.32) with the strength of the spatial correlation structure. Both model based and simulation based estimates of the standard error appear to be working well. In all cases, the average of the estimated measurement error variances are very similar to the true values (not shown in the table).

To evaluate the performance of the proposed method under small sample settings, we also conducted simulations with sample sizes of 250 and 100 assuming a measurement error variance σ_U^2 of 0.5. The results are given in Table 5.2.

With the size of 250 and 100 samples our proposed method still provides slightly biased (around 1.95) estimates of the true regression coefficient of 2. However, with moderate sample sizes (say, n=100) the variance of estimated regression coefficients tends to be slightly inflated. To explore the impact of number of knots on our proposed method we conducted additional simulation study by varying the number of knots for covariate model as 130, 140 and 170 with measurement error 0.025, sample size of 500 and varying range parameters, where the number of knots for the residual error model was fixed as 125. The results are presented in the Appendix Table 5.4. These results indicate that the

Table 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.

Range* (τ_{G_1})	No measurement error				Measurement error variance, $\sigma_u^2 = 0.25$				Measurement error variance, $\sigma_u^2 = 0.5$			
	OLS	LME	GAM	Proposed	OLS	LME	GAM	Proposed	OLS	LME	GAM	Proposed
	Estimated coefficient											
0.1	2.001	2.001	2.002	1.991	1.928	1.439	1.332	2.066	1.858	1.274	0.986	2.034
0.3	1.999	1.999	1.999	1.988	1.927	1.574	1.327	2.096	1.858	1.343	0.987	2.036
0.5	2.001	2.001	2.001	1.991	1.926	1.599	1.312	2.064	1.857	1.389	0.999	2.035
	Empirical standard error											
0.1	0.029	0.028	0.040	0.029	0.032	0.609	0.216	0.056	0.035	0.751	0.216	0.069
0.3	0.035	0.030	0.030	0.032	0.036	0.554	0.211	0.045	0.038	0.730	0.219	0.058
0.5	0.031	0.027	0.026	0.029	0.035	0.543	0.223	0.051	0.039	0.712	0.210	0.052
	Average of estimated standard errors											
0.1	0.014	0.021	0.030	0.015	0.022	0.040	0.058	0.051	0.027	0.037	0.057	0.053
0.3	0.014	0.020	0.026	0.014	0.022	0.035	0.057	0.041	0.027	0.035	0.056	0.052
0.5	0.014	0.018	0.023	0.014	0.022	0.034	0.057	0.049	0.026	0.034	0.056	0.051
	Average of simulated standard errors											
0.1	—	—	—	0.015	—	—	—	0.050	—	—	—	0.068
0.3	—	—	—	0.014	—	—	—	0.041	—	—	—	0.052
0.5	—	—	—	0.014	—	—	—	0.049	—	—	—	0.051

τ_{G_1}: values of the range parameter following exponential correlation in $G_1(s_i)$.

Table 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.

Range*	Sample size 250				Sample Size 100			
(τ_{G_1})	OLS	LME	GAM	Proposed	OLS	LME	GAM	Proposed
Estimated coefficient								
0.1	1.860	1.511	0.976	1.952	1.859	1.831	1.037	1.947
0.3	1.861	1.495	0.975	1.951	1.859	1.824	1.045	1.948
0.5	1.860	1.522	0.980	1.950	1.860	1.831	1.036	1.949
Empirical standard error								
0.1	0.045	0.536	0.217	0.046	0.066	0.088	0.344	0.069
0.3	0.047	0.541	0.207	0.048	0.067	0.099	0.349	0.072
0.5	0.046	0.530	0.209	0.046	0.066	0.095	0.342	0.068
Average of estimated standard errors								
0.1	0.038	0.051	0.083	0.046	0.061	0.064	0.132	0.099
0.3	0.038	0.051	0.081	0.045	0.060	0.064	0.130	0.099
0.5	0.037	0.050	0.081	0.045	0.060	0.063	0.130	0.098
Average of simulated standard errors								
0.1	—	—	—	0.046	—	—	—	0.101
0.3	—	—	—	0.046	—	—	—	0.101
0.5	—	—	—	0.045	—	—	—	0.099
τ_{G_1}: values of the range parameter following exponential correlation in $G_1(s_i)$.								

proposed methods is robust for the selection of number of knots for covariates models.

5.4 Analysis of Ischemic Heart Disease Data

We applied our proposed methodology to re-analyse data on Ischemic Heart Disease (IHD). One of the key objectives of the analysis is to assess the relationship between IHD rates and area level measures of socio-economic status. These data were collected from all hospitals in New South Wales, Australia between July 1, 1994 to June 30, 2002. A detailed description of the data has been given elsewhere (Burden et al., 2005). Briefly, patients who were admitted to the hospitals via the emergency room and discharged with IHD were defined as acute IHD cases. Data also includes patient age, gender and geographic location reported via postcode of residence. Data from 579 postcodes were included in the analysis. IHD event data were linked with the Census data which contains age and gender-specific population counts. SEIFA (Socio-Economic Indexes For Areas) scores and centroid co-ordinates (latitude and longitude) for each postcode were obtained from Australian Bureau of Statistics. We calculated age-sex adjusted standardized incidence ratios (SIR) by dividing the observed number of IHD

Table 5.3: Analysis of Ischemic Heart Disease Data in NSW, Australia under different specification of measurement error.

Methods	Estimates for SEIFA		
		model based	simulated
	$\hat{\boldsymbol{\beta}}$	se($\hat{\boldsymbol{\beta}}$)	se($\hat{\boldsymbol{\beta}}$)
Ordinary Least Squares	-0.062	0.014	—
Generalized additive model	-0.145	0.014	—
Proposed semiparametric approach	-0.273	0.045	0.045
Huque et al. (2014) approach			
Method I: Method of Moments	-0.377	0.041	—
Method II: Transformation of covariate	-0.278	0.015	—

cases by the age-sex adjusted expected number of IHD cases (Breslow & Day, 1987).

The results of our analysis are given in Table 5.3. The naive analysis ignoring spatial correlation, suggests a significant protective effect associated with higher SEIFA values ($\hat{\boldsymbol{\beta}}_{SEIFA}$=-0.062, SE=0.014). Our proposed semi-parametric approach that account for measurement error in the covariates result in an estimated slope parameter β_1 of -0.273 with measurement error variance estimated as 0.52. We choose 145 knots to represent the spatial correlation in the outcome model and 180 knots to represent the covariate model. The model and simulation based standard errors were estimated as 0.045 and 0.045, respectively. Thus, accounting for the measurement error in the covariate reflects a high magnitude of protective effect of higher SEIFA scores on IHD rates, compared with naive analysis.

5.5 Discussion

In this chapter, we develop a semi-parametric framework to obtain a consistent estimate of the true regression coefficients when covariates are measured with error in spatial regression modeling settings. Asymptotic theory establishes that our approach provides consistent, asymptotically normal estimates for the regression coefficient. The theory yields both model based and simulation based standard error estimates. Our empirical simulation results confirm that ignoring measurement error and conducting naive analysis using both generalized additive model and linear mixed model attenuates the estimated regression coefficient towards the null hypothesis of no effect. Our results also

confirm the results of Huque et al. (2014) that the degree of measurement error bias depends on the assumed correlation structure. It is interesting that the bias appears to be least with OLS. This is likely because the covariate spatial structure and residual spatial structure compete to explain the variability in the response (Waller & Gotway, 2004). Our proposed semiparameteric bias correction method performs very well and provides comparable estimates of the regression parameters to the parametric methods described by Huque et al. (2014) when applied to Ischemic Heart Disease (IHD) data. Our approach is computationally efficient and stable because it involves direct estimation using least squares and can be implemented using standard nonlinear least squares software.

Although Huque et al. (2014) and Li et al. (2009) reported similar results for the bias associated with covariate measurement error in spatial regression settings, their approaches requires correct specification of the true covariate measurement error variance. In addition, Huque et al. (2014) reported under estimation of standard error when measurement error variances are estimated from the data. In contrast, our approach is robust because it neither assumes that the covariate measurement error is known nor depends on any particular kind of spatial correlation structure. Our method is analogous to the popular regression calibration method where we estimate the true underlying covariate following smoothing assumption and replace the error prone covariate with this estimate in the outcome model.

Measurement error theory makes it very clear that without some kind of information regarding the magnitude of measurement error, models will not be identifiable. Broadly speaking there are two possibilities: (i) measurement error variance is known or can be estimated using some form of validation data (ii) assumptions are made regarding the nature of the measurement error process. By assuming that the true unobserved covariate is smooth, our paper is using the second approach. Because our approach is assumption based and not an empirical measurement error adjustment, our solution will not be robust to this particular assumption. Nevertheless, because we use a semi-parametric approach to quantifying the spatial correlation in our regression model, our approach should be more robust than parametric alternatives, such as those proposed by Huque et al. (2014). In practice, there will often be situations where it makes sense that spatially-defined covariates are smooth. Air pollution epidemiology

might be a good example. In general, however, we recommend that our proposed method be used in the spirit of sensitivity analysis to assess the impact of measurement error.

One of the additional assumptions required by our approach is that the basis functions for the covariate and the spatial residual term are unequal. In practice, this can be achieved through ensuring more knots for the basis function representing covariate than the spatial residuals. This ensures estimation of variability in covariate in a smaller scale than the residual error. In many spatial epidemiology contexts, measurement error becomes an increasing concern at small scales because of limitations in measurement resources. As a result, the covariate measurement bias reduction relies in estimating variability in covariate at scale smaller than the residual error (Paciorek, 2010).

In our simulation, we have considered only a single covariate measured with error in a spatial linear mixed model with Gaussian error. It would be of interest to explore the effect of covariate measurement error in the presence of multiple covariates and also omitted covariates. Future work should also consider extensions of our formulation to the setting of spatial generalized linear mixed model with non-Gaussian outcomes. One can also consider a Bayesian approach that may relax the smooth covariate assumption. However such an exploration is beyond the scope of the present thesis. We further note that although our approach improved the estimation of standard error compare to Huque et al. (2014) approach further improvement may be obtained using block-bootstrap type estimation techniques or higher order asymptotic. Future study need to be done in this regard.

Our heart disease example demonstrated a substantial increase in the rates of IHD as the level of SEIFA measured at the postcode level decreased, with the magnitude of the effect increasing after adjustment for measurement error. Our results are consistent with broader literature suggesting a relationship between low socio-economic status and adverse health outcomes (see systematic review by Pickett and Pearl 2001).

Because the SEIFA Index is measured at a group level, it is tempting to think that Berkson measurement error theory should be in operation. However, this argument doesn't apply since we are considering measurement error in a group level covariate applied at a group level analysis. It is also important to note that our results can only be

interpreted at a group level. Interpretation at the individual level may result in ecological bias (Sheppard, 2003). While it might be ideal to use individual level data, in many research areas, group-level data are the only available source for analysis. Air pollution epidemiology provides a classic example, because individual measurements of air pollution studies are rarely collected, instead, they are estimated based on neighborhood monitoring and other sources (Sheppard et al., 2012). Consequently, air pollution exposures are typically measured with error.

In spatial data settings, for example, in environmental epidemiology, with the increasing popularity of the semi parametric/multilevel models to account for the observed data correlations, it is important that practitioners be aware of the consequences of measurement error. Furthermore, it is useful to quantify its potential effect on the estimating exposure-outcome relationship. The approach presented in this paper provides one way of achieving this.

Appendix 5A

Estimation of $\widehat{\theta}_2$ and its asymptotic distribution

We assume that the smoothing parameters are small relative to the sample size, i.e., $n^{1/2}\delta_j \to 0$ for $j = 1, 2$. From (5.4) we estimate θ_2 by minimizing the penalized sum of squares $\boldsymbol{J}(\theta_2) = n^{-1}\sum_{i=1}^{n}\{W_i - B_2^{\mathrm{T}}(S_i)\theta_2\}^2 + \delta_2\theta_2^{\mathrm{T}}D_2\theta_2$. We have

$$\begin{aligned}
&\frac{\partial \boldsymbol{J}(\theta_2)}{\partial \theta_2} = 0 \\
&\Rightarrow 2n^{-1}\textstyle\sum_{i=1}^{n}\{W_i - B_2^{\mathrm{T}}(S_i)\theta_2\}B_2(S_i) + 2\delta_2 D_2\theta_2 = 0 \\
&\Rightarrow n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)W_i - \{n^{-1}\sum_{i=1}^{n}B_2(S_i)B_2^{\mathrm{T}}(S_i) - \delta_2 D_2\}\theta_2 = 0 \\
&\Rightarrow n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)W_i - \Lambda_{n2}^{-1}\theta_2 = 0 \\
&\Rightarrow \widehat{\theta}_2 = \Lambda_{n2}n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)W_i.
\end{aligned}$$

Now, the asymptotics of $\widehat{\theta}_2$ can be given by

$$\begin{aligned}
n^{1/2}(\widehat{\theta}_2 - \theta_2) &= n^{1/2}(\Lambda_{n2}n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)W_i - \theta_2) \\
&= n^{1/2}(\Lambda_{n2}n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)[B_2^{\mathrm{T}}(S_i)\theta_2 + U_i] - \theta_2) \\
&= n^{1/2}\Lambda_{n2}(n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)B_2^{\mathrm{T}}(S_i) - \Lambda_{n2}^{-1})\theta_2 \\
&\qquad + \Lambda_{n2}n^{-1/2}\textstyle\sum_{i=1}^{n}B_2(S_i)U_i \\
&= \Lambda_{n2}n^{-1/2}\textstyle\sum_{i=1}^{n}B_2(S_i)U_i + o_p(1).
\end{aligned}$$

Thus,

$$(\widehat{\theta}_2 - \theta_2) = n^{-1}\Lambda_{n2}\textstyle\sum_{i=1}^{n}B_2(S_i)U_i + o_p(n^{-1/2}). \tag{5.15}$$

This is because as, $n^{1/2}\delta_j \to 0$, then

$$n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)B_2^{\mathrm{T}}(S_i) - \Lambda_{n2}^{-1} = \delta_2 D_2 = o_p(n^{-1/2}). \tag{5.16}$$

Estimation of $\widehat{\theta}_1$

From (5.3), we estimate β_1 and θ_1 by minimizing the penalized sum of squares $\boldsymbol{J}(\beta,\theta_1) = n^{-1}\sum_{i=1}^n\{Y_i - B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 - B_1^{\mathrm{T}}(S_i)\theta_1\}^2 + \delta_1\theta_1^{\mathrm{T}}D_1\theta_1$. That is,

$$\begin{aligned}
&\frac{\partial \boldsymbol{J}(\beta,\theta_1)}{\partial\beta_1} = 0\\
&\Rightarrow 2n^{-1}\textstyle\sum_{i=1}^n\{Y_i - B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\beta_1 - B_1^{\mathrm{T}}(S_i)\theta_1\}B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 = 0\\
&\Rightarrow n^{-1}\textstyle\sum_{i=1}^n\{B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2Y_i - \widehat{\theta}_2B_2^{\mathrm{T}}(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\beta_1 - \widehat{\theta}_2B_2^{\mathrm{T}}(S_i)B_1^{\mathrm{T}}(S_i)\theta_1\} = 0\\
&\Rightarrow n^{-1}\textstyle\sum_{i=1}^n\{B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2Y_i - \widehat{\theta}_2B_2^{\mathrm{T}}(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\beta_1 - \widehat{\theta}_2B_2^{\mathrm{T}}(S_i)B_1^{\mathrm{T}}(S_i)\theta_1\} = 0,
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial \boldsymbol{J}(\beta,\theta_1)}{\partial\theta_1} = 0\\
&\Rightarrow 2n^{-1}\textstyle\sum_{i=1}^n\{Y_i - B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\beta_1 - B_1^{\mathrm{T}}(S_i)\theta_1\}B_1(S_i) + 2\delta_1D_1\theta_1 = 0\\
&\Rightarrow n^{-1}\textstyle\sum_{i=1}^n\{B_1(S_i)Y_i - B_1(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\beta_1 - B_1(S_i)B_1^{\mathrm{T}}(S_i)\theta_1\} + \delta_1D_1\theta_1 = 0\\
&\Rightarrow n^{-1}\textstyle\sum_{i=1}^n B_1(S_i)Y_i - n^{-1}\sum_{i=1}^n B_1(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\beta_1 -\\
&\qquad\qquad (n^{-1}\textstyle\sum_{i=1}^n B_1(S_i)B_1^{\mathrm{T}}(S_i) - \delta_1D_1)\theta_1 = 0\\
&\Rightarrow n^{-1}\textstyle\sum_{i=1}^n B_1(S_i)Y_i - n^{-1}\sum_{i=1}^n B_1(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\beta_1 - \Lambda_{n1}^{-1}\theta_1 = 0\\
&\Rightarrow n^{-1}\{\textstyle\sum_{i=1}^n B_1(S_i)Y_i - n^{-1}\sum_{i=1}^n B_1(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\beta_1 - \Lambda_{n1}^{-1}\theta_1\} = 0.
\end{aligned}$$

This leads to the estimating equations

$$\begin{aligned}
0 \;=\; & n^{-1}\textstyle\sum_{i=1}^n \begin{Bmatrix} B_1(S_1)Y_i \\ B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2Y_i \end{Bmatrix}\\
& -n^{-1}\textstyle\sum_{i=1}^n \begin{Bmatrix} \Lambda_{n1}^{-1} & B_1(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 \\ B_1^{\mathrm{T}}(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 & \widehat{\theta}_2^{\mathrm{T}}B_2^{\mathrm{T}}(S_i)B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 \end{Bmatrix} \begin{pmatrix} \theta_1 \\ \beta_1 \end{pmatrix}.
\end{aligned}$$

Writing $\widehat{G}_2(S_i) = B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2$, we see that the estimating equations are

$$\begin{aligned}
0 \;&=\; \textstyle\sum_{i=1}^n\{B_1(S_i)Y_i - \Lambda_{n1}^{-1}\widehat{\theta}_1 - B_1(S_i)\widehat{G}_2(S_i)\widehat{\beta}_1\}; &(5.17)\\
0 \;&=\; \textstyle\sum_{i=1}^n\widehat{G}_2(S_i)\{Y_i - B_1^{\mathrm{T}}(S_i)\widehat{\theta}_1 - \widehat{G}_2(S_i)\widehat{\beta}_1\}. &(5.18)
\end{aligned}$$

Now from (5.17) we see that

$$\widehat{\theta}_1 = V_n - R_n \widehat{\theta}_2 \widehat{\beta}_1,$$

where,

$$\begin{aligned} V_n &= \Lambda_{n1} n^{-1} \textstyle\sum_{i=1}^n B_1(S_i) Y_i, \\ R_n &= \Lambda_{n1} n^{-1} \textstyle\sum_{i=1}^n B_1(S_i) B_2^{\mathrm{T}}(S_i). \end{aligned}$$

Estimation of $\widehat{\beta}_1$

Substituting the value of $\widehat{\theta}_1$ in (5.18) we have

$$0 = \textstyle\sum_{i=1}^n \widehat{G}_2(S_i)\{Y_i - B_1^{\mathrm{T}}(S_i)V_n + B_1^{\mathrm{T}}(S_i)R_n\widehat{\theta}_2\widehat{\beta}_1 - \widehat{G}_2(S_i)\widehat{\beta}_1\}.$$

From this,

$$\widehat{\beta}_1 = \frac{n^{-1}\sum_{i=1}^n \widehat{G}_2(S_i)\{Y_i - B_1^{\mathrm{T}}(S_i)V_n\}}{n^{-1}\sum_{i=1}^n \widehat{G}_2(S_i)\{\widehat{G}_2(S_i) - B_1^{\mathrm{T}}(S_i)R_n\widehat{\theta}_2\}}. \tag{5.19}$$

The numerator of (5.19) is

$$\begin{aligned} & n^{-1}\textstyle\sum_{i=1}^n \widehat{G}_2(S_i)\{Y_i - B_1^{\mathrm{T}}(S_i)V_n\} \\ &= n^{-1}\textstyle\sum_{i=1}^n Y_i B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 - n^{-1}\textstyle\sum_{i=1}^n \widehat{\theta}_2^{\mathrm{T}} B_2(S_i) B_1^{\mathrm{T}}(S_i) V_n \\ &= n^{-1}\textstyle\sum_{i=1}^n Y_i B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 - n^{-1}\textstyle\sum_{i=1}^n \widehat{\theta}_2^{\mathrm{T}} \{B_1(S_i) B_2^{\mathrm{T}}(S_i)\}^{\mathrm{T}} V_n \\ &= n^{-1}\textstyle\sum_{i=1}^n Y_i B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 - \widehat{\theta}_2^{\mathrm{T}} R_n^{\mathrm{T}} \Lambda_{n1}^{-1} V_n \\ &= n^{-1}\textstyle\sum_{i=1}^n Y_i B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 - V_n^{\mathrm{T}} \Lambda_{n1}^{-1} R_n \widehat{\theta}_2 \\ &= n^{-1}\textstyle\sum_{i=1}^n \left\{Y_i B_2^{\mathrm{T}}(S_i) - V_n^{\mathrm{T}} \Lambda_{n1}^{-1} R_n\right\} \widehat{\theta}_2 \\ &= n^{-1}\textstyle\sum_{i=1}^n Y_i \{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i) R_n\} \widehat{\theta}_2. \end{aligned}$$

The denominator of (5.19) is

$$\begin{aligned}
&n^{-1}\textstyle\sum_{i=1}^{n}\widehat{G}_2(S_i)\{\widehat{G}_2(S_i) - B_1^{\mathrm{T}}(S_i)R_n\widehat{\theta}_2 \\
&= \widehat{\theta}_2^{\mathrm{T}} n^{-1}\textstyle\sum_{i=1}^{n} B_2(S_i)\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 \\
&= \widehat{\theta}_2^{\mathrm{T}}\{n^{-1}\textstyle\sum_{i=1}^{n} B_2(S_i)B_2^{\mathrm{T}}(S_i) - n^{-1}\textstyle\sum_{i=1}^{n} B_2(S_i)B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 \\
&= \widehat{\theta}_2^{\mathrm{T}}(\mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n)\widehat{\theta}_2.
\end{aligned}$$

where, $\mathcal{T}_n = n^{-1}\sum_{i=1}^{n} B_2(S_i)B_2^{\mathrm{T}}(S_i)$ (say).

Hence,

$$\widehat{\beta}_1 \;=\; \frac{n^{-1}\sum_{i=1}^{n} Y_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2}{\widehat{\theta}_2^{\mathrm{T}}(\mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n)\widehat{\theta}_2}.$$

Proof of Theorem 1

To ease exposition, let us define the following expressions:

$$\begin{aligned}
\mathcal{A}_n &= n^{-1}\textstyle\sum_{i=1}^{n}\{G_2(S_i)\beta_1 + G_1(S_i)\}\{B_2(S_i) - R_n^{\mathrm{T}}B_1(S_i)\}^{\mathrm{T}}; \\
\mathcal{C}_n &= \mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n; \\
\mathcal{D}_{ni} &= \{B_2(S_i) - R_n^{\mathrm{T}}B_1(S_i)\}^{\mathrm{T}}\theta_2; \\
\mathcal{F}_{ni} &= \theta_2^{\mathrm{T}}\mathcal{C}_n\Lambda_{n2}B_2(S_i) + B_2^{\mathrm{T}}(S_i)\Lambda_{n2}\mathcal{C}_n\theta_2; \\
\mathcal{G}_{ni} &= \mathcal{D}_{ni}/\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2; \\
\mathcal{H}_{ni} &= \mathcal{A}_n\Lambda_{n2}B_2(S_i)/\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 - \mathcal{A}_n\theta_2\mathcal{F}_{ni}/(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^2.
\end{aligned} \tag{5.20}$$

Consider that the S_i's are fixed and known and recall that, $Y_i = G_2(S_i)\beta_1 + G_1(S_i) + \epsilon_i$. Substituting the expression for Y_i into the numerator of (5.19) and simplifying using the

expression from (5.20), we have

$$
\begin{aligned}
& n^{-1}\textstyle\sum_{i=1}^{n} Y_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 \\
&= n^{-1}\textstyle\sum_{i=1}^{n}(B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 + B_1^{\mathrm{T}}(S_i)\theta_1 + \epsilon_i)\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 \\
&= n^{-1}\textstyle\sum_{i=1}^{n}(B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 + B_1^{\mathrm{T}}(S_i)\theta_1\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 + \\
& n^{-1}\textstyle\sum_{i=1}^{n}\epsilon_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 \\
&= \mathcal{A}_n\widehat{\theta}_2 + n^{-1}\textstyle\sum_{i=1}^{n}\mathcal{D}_{ni}\epsilon_i \\
&= \mathcal{A}_n\theta_2 + \mathcal{A}_n(\widehat{\theta}_2 - \theta_2) + n^{-1}\textstyle\sum_{i=1}^{n}\mathcal{D}_{ni}\epsilon_i.
\end{aligned}
$$

Applying (5.15) to the above equation, we have

$$
\mathcal{A}_n\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\} + o_p(n^{-1/2}).
$$

Again, the denominator of (5.19) is

$$
\widehat{\theta}_2^{\mathrm{T}}(\mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n)\widehat{\theta}_2 = \widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2.
$$

Now applying (5.15), the denominator becomes,

$$
\begin{aligned}
& (\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\Lambda_{n2}B_2(S_i)U_i) + o_p(n^{-1/2})^{\mathrm{T}}\mathcal{C}_n(\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\Lambda_{n2}B_2(S_i)U_i) + o_p(n^{-1/2}) \\
&= \theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\theta_2^{\mathrm{T}}\mathcal{C}_n\Lambda_{n2}B_2(S_i)U_i + n^{-1}\textstyle\sum_{i=1}^{n}U_i^{\mathrm{T}}B_2(S_i)^{\mathrm{T}}\Lambda_{n2}\mathcal{C}_n\theta_2 + \\
& (n^{-1}\textstyle\sum_{i=1}^{n}\Lambda_{n2}B_2(S_i)U_i)^{\mathrm{T}}(n^{-1}\textstyle\sum_{i=1}^{n}\Lambda_{n2}B_2(S_i)U_i) + o_p(n^{-1/2}) \\
&= \theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\mathcal{F}_{ni}U_i + o_p(n^{-1/2}).
\end{aligned}
$$

Then, by a Taylor series expansion,

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\mathcal{A}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\}}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\mathcal{F}_{ni}U_i} + o_p(n^{-1/2}) \\
&= \left\{\frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} + \frac{n^{-1}\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\}}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2}\right\}\left\{1 + \frac{\sum_{i=1}^{n}\mathcal{F}_{ni}U_i}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2}\right\}^{-1} + o_p(n^{-1/2}) \\
&= \frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} + \frac{n^{-1}\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\}}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} - \frac{\mathcal{A}_n\theta_2}{(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^2}n^{-1}\sum_{i=1}^{n}\mathcal{F}_{ni}U_i + o_p(n^{-1/2}).
\end{aligned}
$$

Thus

$$
\widehat{\beta}_1 - \frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} = n^{-1}\textstyle\sum_{i=1}^{n}(\mathcal{G}_{ni}\epsilon_i + \mathcal{H}_{ni}U_i) + o_p(n^{-1/2}).
$$

Now considering the fact from equation (5.16) that
$n^{-1}\sum_{i=1}^{n} B_2(S_i)B_2^{\mathrm{T}}(S_i) = \Lambda_{n2}^{-1} + o(n^{-1/2})$ and using this in the expression for $\mathcal{A}_n\theta_2$, we have

$$\begin{aligned}
\mathcal{A}_n\theta_2 &= n^{-1}\textstyle\sum_{i=1}^{n}\beta_1\theta_2^{\mathrm{T}}B_2(S_i)\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\theta_2 \\
&\qquad +\theta_1^{\mathrm{T}}n^{-1}\textstyle\sum_{i=1}^{n}B_1(S_i)\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\theta_2 \\
&= \beta_1\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + \theta_1^{\mathrm{T}}\{\Lambda_{n1}^{-1} - n^{-1}\textstyle\sum_{i=1}^{n}B_1(S_i)B_1^{\mathrm{T}}(S_i)\}R_n\theta_2 \\
&= \beta_1\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + o_p(n^{-1/2}).
\end{aligned}$$

Therefore,

$$\frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} = \beta_1 + o_p(n^{-1/2}).$$

Hence,

$$n^{1/2}(\widehat{\beta}_1 - \beta_1) \sim \mathrm{Normal}(0, \sigma^2),$$

where

$$\sigma^2 = \lim_{n\to\infty} n^{-1}\textstyle\sum_{i=1}^{n}(\sigma_\epsilon^2\mathcal{G}_{ni}^2 + \sigma_u^2\mathcal{H}_{ni}^2). \tag{5.21}$$

Thus, $\hat{\beta}_1$ is a consistent estimate for β_1.

Derivation of smoothing matrix

From (5.3), we have

$$\begin{aligned}
\widehat{Y}_i &= B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\widehat{\beta}_1 + B_1^{\mathrm{T}}(S_i)\widehat{\theta}_1 \\
&= B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\widehat{\beta}_1 + B_1^{\mathrm{T}}(S_i)[V_n - R_n\widehat{\theta}_2\widehat{\beta}_1] \\
&= [B_2(S_i) - B_1^{\mathrm{T}}(S_i)R_n]^{\mathrm{T}}\widehat{\theta}_2\widehat{\beta}_2 + B_1^{\mathrm{T}}(S_i)V_n \\
&= [B_2(S_i) - B_1^{\mathrm{T}}(S_i)R_n]^{\mathrm{T}}\widehat{\theta}_2 \left(\frac{n^{-1}\sum_{i=1}^n Y_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2}{\widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2} \right) \\
&\qquad + B_1^{\mathrm{T}}(S_i)\Lambda_{n1}n^{-1}\textstyle\sum_{i=1}^n B_1(S_i)Y_i \\
&= \frac{D_{ni}n^{-1}\sum_{i=1}^n D_{ni}Y_i}{\widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2} + B_1^{\mathrm{T}}(S_i)\Lambda_{n1}n^{-1}\textstyle\sum_{i=1}^n B_1(S_i)Y_i.
\end{aligned}$$

Similarly from (5.4), we have

$$\begin{aligned}
\widehat{W_i} &= B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 \\
&= B_2^{\mathrm{T}}(S_i)\Lambda_{n2}n^{-1}\textstyle\sum_{i=1}^n B_2(S_i)W_i.
\end{aligned}$$

Appendix Table 1

Table 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.

Range*	Number of knots for covariate model		
(τ_{G_1})	130	140	170
Estimated coefficient			
0.1	2.016	2.013	2.007
0.3	2.019	2.016	2.010
0.5	2.025	2.023	2.018
Empirical standard error			
0.1	0.060	0.061	0.068
0.3	0.058	0.059	0.066
0.5	0.052	0.053	0.060
	Average of estimated standard errors		
0.1	0.048	0.048	0.049
0.3	0.047	0.047	0.047
0.5	0.045	0.045	0.046
	Average of simulated standard errors		
0.1	0.048	0.048	0.049
0.3	0.047	0.047	0.047
0.5	0.045	0.045	0.046

τ_{G_1}: values of the range parameter following exponential correlation in $G_1(s_i)$.

R Code

Function for data generation

```
dataSimulation<-function(nsamp=500,b0=1,b1=2, sigma.Delta=0.25,
sigma.G=0.2,sigma.Y=0.5,range.par.Y=0.1)
{
# nsamp= Sample size for data generation.
# range.par.Y= Value of the range parameter for outcome variable.
# sigma.Delta= Measurement error variance.
# sigma.G=Variance for smooth spatial surface.
# sigma.Y=Residual error variance.
# b0=Intercept of the regression model.
# b1= Slope parameter for regression model.
# Load necessary R library
require(nlme)
# generate grid of nsamp points uniformly on square
s1 <- runif(nsamp)
s2 <- runif(nsamp)
spatDat <- data.frame(cbind(s1,s2))
#Covariance structure for Y following exponential correlation
cs1Exp.Y <- corExp(range.par.Y, form = ~ s1 + s2)
cs1Exp.Y <- Initialize(cs1Exp.Y, spatDat)
R.Y.matrix <- corMatrix(cs1Exp.Y)
# Get true inverse covariance matrices, square roots, eigenvalues,
# used to generate data and also to compute bias correction factors
eigen.decomp.corr.Y <- eigen(R.Y.matrix)
gamma.Y.mat <- cbind(eigen.decomp.corr.Y$vectors)
lambda.Y.vec <- eigen.decomp.corr.Y$val
decomp.Y <- gamma.Y.mat%*%diag(sqrt(lambda.Y.vec))%*%
t(gamma.Y.mat)
# generate true covariate X is is generated using a bivariate bump function
X1 <- 1/(1+s1) + 3*exp(-50*(s1-.3)^2) + 2*exp(-25*(s1-.7)^2)
X2 <- 1/(1+s2) + 3*exp(-50*(s2-.3)^2) + 2*exp(-25*(s2-.7)^2)
X <-X1*X2
# generate observed covariate measured with error
# Z ~ N (X, variance = sigma.Delta)
Z <- X + sqrt(sigma.Delta)*rnorm(nsamp)
# Generate smooth spatial surface
G1<-sqrt(sigma.G)*decomp.Y%*%rnorm(nsamp)
# generate outcome Y | X ~ N (b0 + b1 X+G1(si)+variance)
Y <- b0 + b1*X+G1+sqrt(sigma.Y)*rnorm(nsamp)
data<-as.data.frame(cbind(s1=s1,s2=s2,X=X,Z=Z,Y=as.vector(Y)))
data
}
```

Necessary functions for implementing the proposed method

```
### Function for selecting knots in two dimensional space
defaultKnots2D <- function(x1,x2,num.knots)
{
require("cluster")
# Set default value for num.knots
if (missing(num.knots))
num.knots <- max(10,min(50,round(length(x1)/4)))
# Delete repeated values from x
X <- cbind(x1,x2)
dup.inds <- (1:nrow(X))[dup.matrix(X)==T]
if (length(dup.inds) > 0)
X <- X[-dup.inds,]
# Obtain and output knots chosen using
# coverage design principles
knots <- clara(X,num.knots)$medoids
return(knots)
}
### Function for penalty matrices using thin plate basis
# Set up thin plate spline generalised covariance function:
tps.cov <- function(r,m=2,d=1)
{
r <- as.matrix(r)
num.row <- nrow(r)
num.col <- ncol(r)
r <- as.vector(r)
nzi <- (1:length(r))[r!=0]
ans <- rep(0,length(r))
if ((d+1)%%2!=0)
ans[nzi] <- (abs(r[nzi]))^(2*m-d)*log(abs(r[nzi])) # d is even
else
ans[nzi] <- (abs(r[nzi]))^(2*m-d)
if (num.col>1) ans <- matrix(ans,num.row,num.col)     # d is odd
return(ans)
}
# Set up function for matrix square-roots:
matrix.sqrt <- function(A)
{
sva <- svd(A)
if (min(sva$d)>=0)
Asqrt <- t(sva$v %*% (t(sva$u) * sqrt(sva$d)))
else
stop("Matrix square root is not defined")
return(Asqrt)
}
Ztps <- function(x,knots)
{
# Obtain  matrix of inter-knot distances:
numKnots <- nrow(knots)
dist.mat <- matrix(0,numKnots,numKnots)
dist.mat[lower.tri(dist.mat)] <- dist(as.matrix(knots))
dist.mat <- dist.mat + t(dist.mat)
Omega <- tps.cov(dist.mat,d=2)
# Obtain preliminary Z matrix of knot to data covariances:
x.knot.diffs.1 <- outer(x[,1],knots[,1],"-")
x.knot.diffs.2 <- outer(x[,2],knots[,2],"-")
x.knot.dists <- sqrt(x.knot.diffs.1^2+x.knot.diffs.2^2)
prelim.Z <- tps.cov(x.knot.dists,m=2,d=2)
# Transform to canonical form:
sqrt.Omega <- matrix.sqrt(Omega)
Z <- t(solve(sqrt.Omega,t(prelim.Z)))
output<-list(basis=prelim.Z, penalty=Omega ,Z=Z)
return(output)
}
#Function for generalized cross validation for delta2
gcv.delta2<-function(nsamp,Z,B2,BTB.2, D2, delta.2)
{
Lambda.2.inv<-1/nsamp*BTB.2+delta.2*D2
eigen.decomp.Lambda.2 <- eigen(Lambda.2.inv)
```

```
eigen.Lamdba.2.mat <- cbind(eigen.decomp.Lambda.2$vectors)
eigen.Lambda.2.vec <- eigen.decomp.Lambda.2$val
Lambda.2<-eigen.Lamdba.2.mat%*%diag(1/eigen.Lambda.2.vec)%*%
t(eigen.Lamdba.2.mat)
theta.2.Hat<-1/nsamp*Lambda.2%*%(t(B2)%*%Z)
RSS.delta2<-t(Z-B2%*%theta.2.Hat)%*%(Z-B2%*%theta.2.Hat)
smooth.delta2<-1/nsamp*B2%*%Lambda.2%*%t(B2)
gcv.delta2<-1/nsamp*
as.vector(RSS.delta2)/(1-1/nsamp*sum(diag(smooth.delta2)))^2
gcv.delta2
}
#Function for generalized cross validation for delta1
gcv.delta1<-function(nsamp,Y,Z,B1,B2,BTB.1,BTB.2,D1, D2, delta.2,delta.1)
{
Lambda.2.inv<-1/nsamp*BTB.2+delta.2*D2
eigen.decomp.Lambda.2 <- eigen(Lambda.2.inv)
eigen.Lamdba.2.mat <- cbind(eigen.decomp.Lambda.2$vectors)
eigen.Lambda.2.vec <- eigen.decomp.Lambda.2$val
Lambda.2<-eigen.Lamdba.2.mat%*%diag(1/eigen.Lambda.2.vec)%*%
t(eigen.Lamdba.2.mat)
theta.2.Hat<-1/nsamp*Lambda.2%*%(t(B2)%*%Z)
RSS.delta2<-t(Z-B2%*%theta.2.Hat)%*%(Z-B2%*%theta.2.Hat)
smooth.delta2<-1/nsamp*B2%*%Lambda.2%*%t(B2)
gcv.delta2<-1/nsamp*
as.vector(RSS.delta2)/(1-1/nsamp*sum(diag(smooth.delta2)))^2
Lambda.1.inv<-1/nsamp*BTB.1+delta.1*D1
eigen.decomp.Lambda.1 <- eigen(Lambda.1.inv)
eigen.Lamdba.1.mat <- cbind(eigen.decomp.Lambda.1$vectors)
eigen.Lambda.1.vec <- eigen.decomp.Lambda.1$val
Lambda.1<-eigen.Lamdba.1.mat%*%diag(1/eigen.Lambda.1.vec)%*%
t(eigen.Lamdba.1.mat)
### Generating the required components for equation (13)
Vn<-1/nsamp*Lambda.1%*%(t(B1)%*%Y)
Rn<-1/nsamp*Lambda.1%*%crossprod(B1,B2)
Tn<-1/nsamp*BTB.2
Cn<-(Tn-t(Rn)%*%Lambda.1.inv%*%Rn)
Dn<-(B2-B1%*%Rn)%*%theta.2.Hat
numerator<-1/nsamp*t(Y)%*%Dn
denominator<-t(theta.2.Hat)%*%Cn%*%theta.2.Hat
# Estimates of the regression parameter
b1.hat.basis<-numerator/denominator
theta.1.Hat<-Vn-Rn%*%theta.2.Hat*as.vector(b1.hat.basis)
RSS.Y<-
t(Y-B2%*%theta.2.Hat*as.vector(b1.hat.basis)-B1%*%theta.1.Hat)%*%
(Y-B2%*%theta.2.Hat*as.vector(b1.hat.basis)-B1%*%theta.1.Hat)
smooth.mat<-1/(nsamp*as.vector(denominator))*Dn%*%t(Dn)+
1/nsamp*(B1%*%Lambda.1%*%t(B1))
gcv.delta1<-as.vector(RSS.Y)/(1-1/nsamp*sum(diag(smooth.mat)))^2
gcv<-gcv.delta1
return(gcv)
}
```

Codes for Data analysis

```
# Set the number of knots
q1=125
q2=150
# Get the data
set.seed(123456)
data<-dataSimulation()
# Extract the coordinates (s1,s2), covariate Z and outcome Y
s1<-data$s1
s2<-data$s2
Z<-data$Z
Y<-data$Y
nsamp<-nrow(data)
```

```
# Load relevant R library
require(cluster)
require(SemiPar)
# Section of knot locations
Knots1<-defaultKnots2D(s1,s2,q1)
Knots2<-defaultKnots2D(s1,s2,q2)
# Generating thin plate spline basis B1(.) & B2(.) with q1 and q2 knots
# respectively
B1<-cbind(rep(1,nsamp),s1,s2,Ztps(cbind(s1,s2),Knots1)$basis)
B2<-cbind(rep(1,nsamp),s1,s2,Ztps(cbind(s1,s2),Knots2)$basis)
### Generating penulty matrices
k1<-nrow(Knots1)+3
k2<-nrow(Knots2)+3
D1<-matrix(0,k1,k1)
D2<-matrix(0,k2,k2)
D1[4:k1,4:k1]<-Ztps(cbind(s1,s2),Knots1)$penalty
D2[4:k2,4:k2]<-Ztps(cbind(s1,s2),Knots2)$penalty
### Generating Lambda1 and Lamda2
BTB.1 <- crossprod(B1,B1)
BTB.2 <- crossprod(B2,B2)
# Estimating delta.2
delta.range<-seq(-11, 11, length.out =60)
V<-rep(0,60)
for (i in 1:60){
V[i]<-gcv.delta2(nsamp,Z,B2,BTB.2, D2,exp(delta.range[i]))}
index.delta2<-(1:60)[V==min(V)]
delta.2<-10^delta.range[index.delta2]
# Estimating delta.1
U<-rep(0,60)
for (i in 1:60){
U[i]<-gcv.delta1(nsamp,Y,Z,B1,B2,BTB.1,BTB.2,D1,D2,delta.2,
exp(delta.range[i]))}
index.delta1<-(1:60)[U==min(U)]
delta.1<-10^delta.range[index.delta1]
#Get lambda.1 using eigen value decomposition
Lambda.1.inv<-1/nsamp*BTB.1+delta.1*D1
eigen.decomp.Lambda.1 <- eigen(Lambda.1.inv)
eigen.Lamdba.1.mat <- cbind(eigen.decomp.Lambda.1$vectors)
eigen.Lambda.1.vec <- eigen.decomp.Lambda.1$val
Lambda.1<-eigen.Lamdba.1.mat%*%diag(1/eigen.Lambda.1.vec)%*%
t(eigen.Lamdba.1.mat)
#Get lambda.2 using eigen value decomposition
Lambda.2.inv<-1/nsamp*BTB.2+delta.2*D2
eigen.decomp.Lambda.2 <- eigen(Lambda.2.inv)
eigen.Lamdba.2.mat <- cbind(eigen.decomp.Lambda.2$vectors)
eigen.Lambda.2.vec <- eigen.decomp.Lambda.2$val
Lambda.2<-eigen.Lamdba.2.mat%*%diag(1/eigen.Lambda.2.vec)%*%
t(eigen.Lamdba.2.mat)
### Generating the required components for equation (13)
theta.2.Hat<-1/nsamp*Lambda.2%*%(t(B2)%*%Z)
Vn<-1/nsamp*Lambda.1%*%(t(B1)%*%Y)
Rn<-1/nsamp*Lambda.1%*%crossprod(B1,B2)
Tn<-1/nsamp*BTB.2
Cn<-(Tn-t(Rn)%*%Lambda.1.inv%*%Rn)
Dn<-(B2-B1%*%Rn)%*%theta.2.Hat
numerator<-1/nsamp*t(Y)%*%Dn
denominator<-t(theta.2.Hat)%*%Cn%*%theta.2.Hat
# Estimates of the regression parameter
b1.hat.basis<-numerator/denominator
# Estimates of the empirical variance for the beta estimates
theta.1.Hat<-Vn-Rn%*%theta.2.Hat*as.vector(b1.hat.basis)
RSS.Y<-
t(Y-B2%*%theta.2.Hat*as.vector(b1.hat.basis)-B1%*%theta.1.Hat)%*%
(Y-B2%*%theta.2.Hat*as.vector(b1.hat.basis)-B1%*%theta.1.Hat)
smooth.mat<-1/nsamp*(1/as.vector(denominator)*Dn%*%t(Dn)+
B1%*%Lambda.1%*%(t(B1)))
sigma.error.est<-as.vector(RSS.Y)/(nsamp-2*sum(diag(smooth.mat))+
sum(diag(crossprod(smooth.mat,smooth.mat))))
RSS.Z<-t(Z-B2%*%theta.2.Hat)%*%(Z-B2%*%theta.2.Hat)
```

```
smooth.delta2<-1/nsamp*B2%*%Lambda.2%*%t(B2)
sigma.u.est<-as.vector(RSS.Z)/(nsamp-2*sum(diag(smooth.delta2))+
sum(diag(smooth.delta2%*%t(smooth.delta2))))

Gn<-Dn/(as.vector(denominator))
Fn<-2*B2%*%Lambda.2%*%Cn%*%theta.2.Hat
An<-(1/nsamp)*t(B2%*%theta.2.Hat*as.vector(b1.hat.basis)+
B1%*%theta.1.Hat)%*%(B2-B1%*%Rn)
Hn<-t(An%*%Lambda.2%*%t(B2)/as.vector(denominator))-
as.vector(An%*%theta.2.Hat)*Fn/(as.vector(denominator))^2
sigma.beta.est<-1/(nsamp^2)*(sigma.error.est*t(Gn)%*%Gn+
sigma.u.est*t(Hn)%*%Hn)

#Codes for the simulated standard error
nboot<-100
sim.b1.hat<-NULL
for (i in 1:nboot)
{
sim.sigma.Y<-rnorm(nsamp,mean=0,sd=sqrt(sigma.error.est))
sim.sigma.Delta<-rnorm(nsamp,mean=0,sd=sqrt(sigma.u.est))
sim.numerator<-An%*%theta.2.Hat+
1/nsamp*(An%*%Lambda.2%*%t(B2)%*%sim.sigma.Delta+
t(Dn)%*%sim.sigma.Y)
sim.denominator<-t(theta.2.Hat)%*%Cn%*%theta.2.Hat+
1/nsamp*(t(Fn)%*%sim.sigma.Delta)
sim.new.b1.hat<-sim.numerator/sim.denominator
sim.b1.hat<-rbind(sim.b1.hat, sim.new.b1.hat)
}
#Estimated regression coefficient
b1.hat<-b1.hat.basis
b1.hat
#Estimation of standard error
se.b1.hat<-sqrt(as.vector(sigma.beta.est))
se.b1.hat
#Estimation of simulated standard error
sim.se.b1.hat<-sqrt(apply(sim.b1.hat,2,var))
sim.se.b1.hat
```

Chapter 6

Exposure enriched case-control (EECC) design for the assessment of gene-environment interaction

Summary

Genetic susceptibility and environmental exposure both play an important role in the aetiology of many diseases. Case-control studies are often the first choice to explore the joint influence of genetic and environmental factors on the risk of developing a rare disease. In practice, however, such studies may have limited power, especially when susceptibility genes are rare and exposure distributions are highly skewed. We propose a variant of the classical case-control study, the exposure enriched case-control (EECC) design, where not only cases, but also high (or low) exposed individuals are oversampled, depending on the skewness of the exposure distribution. Of course, a traditional logistic regression model is no longer valid and results in biased parameter estimation. We show that addition of a simple covariate to the regression model removes this bias and yields reliable estimates of main and interaction effects of interest. We also discuss optimal

The content of this chapter is published as: Huque, MH; Carroll, RJ; Diao N; Christiani DC; and Ryan LM. (2016). Exposure Enriched Case-Control (EECC) Design for the Assessment of GeneEnvironment Interaction. *Genetic Epidemiology*, doi: 10.1002/gepi.21986. This research was also presented at the ENAR 2016 Spring Meeting, Austin, Texas, USA.

design, showing that judicious oversampling of high/low exposed individuals can boost study power considerably. We illustrate our results using data from a study involving arsenic exposure and detoxification genes in Bangladesh.

6.1 Introduction

Many common diseases are now believed to be the result of interdependence between genetic and environmental factors (Chatterjee & Carroll, 2005; Liu, Maity, Lin, Wright, & Christiani, 2012; Mukherjee, Ahn, Gruber, Ghosh, & Chatterjee, 2010). Gene-environment interaction refers to the setting where the effects of an environmental exposure are enhanced in a particular genetic subgroup. Consequently, identification of gene-environment (GE) interactions plays an important role in understanding the aetiology of underlying diseases and hence, developing disease prevention and intervention strategies. However, the classic case-control design can have limited power for studying gene-environment interaction, especially in the case of rare genetic variants and also when exposure distributions are skewed (Foppa & Spiegelman, 1997; García-Closas & Lubin, 1999; Luan et al., 2001). To address this, various complex sampling strategies have been proposed (see Thomas (2010) for a recent review). In one of the first such approaches, White (1982) proposed a two stage design where exposure (or an appropriate surrogate) is first measured in a large number of case and control subjects (Stage I). At Stage II, detailed covariate information is obtained for a subset from each strata defined by case/control and exposure status.

Breslow and Cain (1988) formalized and generalized White's approach to a general two-stage design with analysis proceeding via logistic regression applied to stage II data, but including an offset terms that reflects the stage I sampling probabilities. Weinberg and Wacholder (1990) suggest a slightly simpler approach to the analysis of two stage designs based on a so-called pseudo-likelihood approach that condition on being sampled in the second stage. Their method also requires inclusion of an offset reflecting sampling probabilities into the logistic regression. While these approaches all provide consistent estimate of main and interaction effects, they require knowledge of the screening variable specific disease rates. In this paper, we propose an alternative approach that does not require knowledge of these probabilities.

Our work is motived by a study designed to explore the relationship between drinking water arsenic levels, genetic polymorphisms and skin lesions in Pabna, Bangladesh (Breton et al., 2007). Because the distribution of arsenic exposure generally high and right skewed in Bangladesh (Ravenscroft, Burgess, Ahmed, Burren, & Perrin, 2005), study investigators had over-sampled low exposed individuals (< 50 micro-grams per liter) among controls. Consequently, traditional logistic analysis was no longer valid, since the sampling mechanism has then violated the key assumption for a case-control study that sampling should be independent of exposure status.

Our approach is designed for settings where interest lies in characterizing a dose response relationship and associated interactions based on a continuous exposure. Our exposure enriched case control (EECC) design over samples subjects based on case/control status, as well as a categorical assessment of exposure (e.g. high versus low). We show that as expected, selection of individuals based on high (or low) exposure results in biased estimation of the regression coefficients when standard logistic regression is used. However, we further show that valid statistical inference can be achieved simply by the addition of a single covariate that reflects this exposure-related category. We illustrate via computer simulations that judicious oversampling of individuals based on exposure can significantly boost study power. We also investigate the relative importance of each of the parameters that determine power for detecting interaction effects.

6.2 Methods

Suppose the probability of disease occurrence in the general population satisfies a logistic regression model

$$logit[Pr(D = 1|E, G)] = \beta_0 + \beta_E E + \beta_G G + \beta_{GE} EG, \tag{6.1}$$

where $D = 1$ denotes the diseased and $D = 0$ the non-diseased state, E denotes the level of a continuous environmental exposure, G is a binary indicator of genetic-susceptibility, EG is the gene-environment interaction and where β_0, β_E, β_G, β_{GE} are the associated regression coefficients. Genetic susceptibility is defined as the presence of one or more gene mutations thought to be associated with the disease of interest. In practice, the

susceptible group will generally correspond to those with the less common variant of the allele of interest.

It is well known that while ordinary logistic regression analysis of case control study data results in incorrect estimation of the intercept (β_0), all other regression coefficients are estimated correctly. This is due to the fact that instead of selecting a random sample from the source population, a biased sample based on case control status was recruited. There are many approaches to understanding why ordinary logistic regression works, despite the fact that sampling is biased in the case control setting (Prentice & Pyke, 1979; Weinberg & Wacholder, 1990). We find it particularly useful to consider a derivation based on Bayes rule (Hosmer, David, & Lemeshow, 2004). We use the same principle to show that it is possible to boost study power to detect an interaction effect by over-sampling not only cases, but also high (or low) exposed individuals. Similar probabilistic logic was also used by Weinberg and Wacholder (1990).

Define Δ as the sampling indicator with $\Delta = 1$ if the individual is selected into the study sample and 0 otherwise. Also denote the probability of selecting an individual into the sample who has disease status D, exposure level E and genetic characteristic G by $\rho(D, E, G)$ that is, $\rho(D, E, G) = Pr(\Delta = 1|D, E, G)$. Then some simple algebra and an application of Bayes rule leads to the probability of being diseased conditional on exposure, gene and being included in the study sample as (see technical Appendix 6A for details)

$$logit[Pr(D = 1|E, G, \Delta = 1)] = \ln\left\{\frac{\rho(D = 1, E, G)}{\rho(D = 0, E, G)}\right\} + \beta_0 + \beta_E E + \beta_G G + \beta_{GE} GE. \quad (6.2)$$

Note that the additional term in model (6.2), compared with model (6.1), is the log odds of selection for cases versus controls, conditional on environmental exposure status and genetic susceptibility. It is associated with the mechanism of recruitment of the study sample, is within the control of investigators and is to be fixed as part of the design. Depending on the recruitment of individuals in the sample, equation (6.2) leads to a variety of familiar designs, including:

1. If the sampling probability $\rho(D, E, G)$ is constant, i.e., a simple random sample of subjects is chosen from the population then model (6.2) will estimate the true population intercept, β_0. This design is popularly known as the prospective cohort

study (Prentice & Pyke, 1979).

2. If the sampling probability $\rho(D, E, G)$ depends on the disease status, D but is independent of genetic status G or environmental exposure E, i.e., $\rho(D, E, G) = \rho(D)$, then the above model (6.2) represents ordinary case-control design with the intercept of the model corresponding to $\beta_0^* = \ln\{\rho(D = 1)/\rho(D = 1)\} + \beta_0$. That is, the estimated intercept of the model (6.2) will be incorrect without the knowledge of the disease prevalence. This result explains the well-known fact that standard logistic regression applied to case-control data yields valid estimates of all regression coefficients except the intercept.

3. If the sampling probability $\rho(D, E, G)$ depends on disease and level of exposure, then its effect on equation (6.2) depends on the nature of the relationship. It is Case (3) that we examine in more detail in this chapter.

Consider a situation where the selection of individuals depends on a certain cut-off value, k, of the observed exposure, E, that characterizes the high or low exposure. Let p_{11} and p_{10} denote the probability of selecting a case in the sample with high (that is, a subject with $D = 1$ and $E > k$) and low (that is, a subject with $D = 1$ and $E < k$) exposure, respectively. Similarly, let p_{01} and p_{00} denote the probability selecting a control subject in the sample with high (that is, $D = 0$ and $E > k$) and low (that is, $D = 0$ and $E < k$) exposure, respectively. If we only select low exposed individuals (or high exposed individuals) in the sample. Then, since equation (6.1) holds in the population, we have a logistic regression with different intercept for high and low exposed group, respectively. Consequently, equation (6.2) can be re-expressed as

$$logit[Pr(D = 1|E, G, \Delta = 1)] = \begin{cases} \ln(\frac{p_{11}}{p_{01}}) + \beta_0 + \beta_E E + \beta_G G + \beta_{GE} GE & , E \geq k \\ \ln(\frac{p_{10}}{p_{00}}) + \beta_0 + \beta_E E + \beta_G G + \beta_{GE} GE & , E < k \end{cases}$$

This means that the intercept in the model varies according to whether or not $E \geq k$. Thus we can succinctly write:

$$logit[Pr(D = 1|E, G, \Delta = 1)] = \beta_0^{**} + \lambda I[E \geq k] + \beta_E E + \beta_G G + \beta_{GE} GE, \qquad (6.3)$$

where $I[E \geq k]$ is an indicator representing whether the environmental exposure E is

above the exposure level k. The parameter β_0^{**} represents the intercept in the low exposed group and can be expressed as a function of true intercept β_0 and the log odds of selection for cases versus controls in the low exposed group, i.e., $\beta_0^{**} = \ln(p_{11}/p_{01}) + \beta_0$. Similarly, λ represents difference between the log odds of the selection probabilities of cases versus controls in the high versus the low exposed group, i.e., $\lambda = \ln(p_{11}/p_{01}) - \ln(p_{10}/p_{00})$. As the cutoff is related to the exposure status, ignoring the indicator variable is like having an omitted variable in the model. Hence, both main effect and interaction will be biased.

Equation (6.3) suggests the simple approach of adding an indicator variable for high versus low exposure into the logistic regression on case/control status. Hence, prospective analysis (Prentice & Pyke, 1979) via logistic regression with the addition of this covariate will yield consistent maximum likelihood estimation and inference of regression coefficients associated with exposure, gene and interaction. For simplicity, we have ignored confounder variables in model (6.2). Additional covariates can also be included if desired. Throughout this paper, we assume that the sampling depends only on the case-control status and environmental exposure, but is independent of genetic susceptibility.

6.3 Simulation Study

In this section we discuss a simulation study designed to evaluate the finite sample properties of the estimated parameters under the EECC design. We also explore power properties of the proposed methods under various alternatives.

6.3.1 Data Generation

We generated a hypothetical population of one million subjects. A genetic susceptibility covariate, G, was generated as a Bernoulli random variable with prevalence 0.2. The environmental exposure, E, was generated as exponentially distributed random variable with rate 1. The outcome data were generated using model (6.1) where parameters β_0, β_E, β_G and β_{GE} were set to -4.60, 1.15, 0.8 and 0.406, respectively. The intercept

parameter ensures the rare disease assumption with 1% disease prevalence in the control population with non-susceptible gene. We defined high (or low) exposed individuals if its exposure level was greater (or smaller) than an exposure level of 2 corresponding to 13.5% of the exposure data in the upper tail area. We then selected a total sample of 1100 observations from the above population equally stratified by exposure level and case status. An equal number of high and low exposed subjects in the sample thus result in oversampling from high exposed group.

6.3.2 Parameter Estimation

We estimated relevant parameters of our proposed method applying logistic regression with an indicator variable (6.3) to the pooled data. We also compare these estimates with naive analysis that ignores indicator variable. The sampling and estimation procedure is repeated 1000 times.

6.3.3 Power Calculation

We use a simulation based technique to calculate power for testing $H_0 : \beta_{GE} = 0$ vs $H_A : \beta_{GE} \neq 0$ via a standard Wald test (Cox & Hinkley, 1974). To estimate the power of the test, we simulated data under the alternative hypothesis, H_A and EECC design, fitted model (6.3). Again, there were 1,000 simulated data sets. The calculated power is the proportion of the 1000 replicates whose test statistics exceeds the relevant critical value of 1.96 (at 5% level of significance). Though we use 5% level of significance, a smaller level of significance can also be incorporated in testing the hypothesis. All calculations were performed using R (R Core Team, 2013).

6.4 Simulation Results

We first show that ignoring the sampling scheme and performing standard logistic regression results in biased estimation. We then show that the bias can be removed through addition of an indicator variable, indicating high exposure, in the logistic

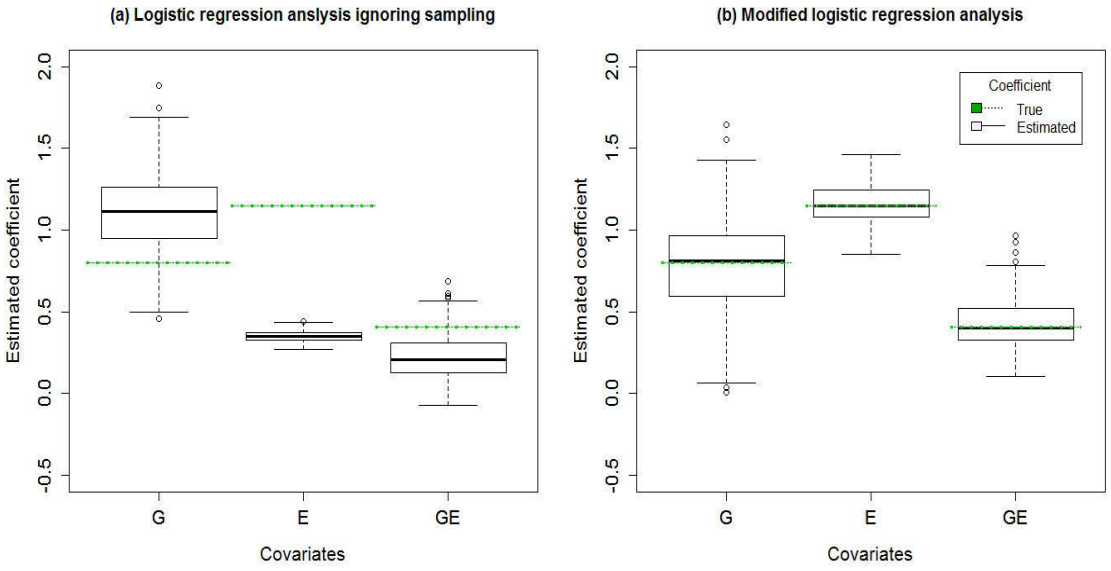


Figure 6.1: Comparison of estimated coefficients obtained using usual logistic regression ignoring sampling and EECC.

regression model (6.3). We later calculate power of our proposed method by varying different parameters that govern the power to detect gene-environment interaction.

6.4.1 Estimation of Parameters

Figure 6.1 compares the distribution of the estimated regression coefficients using logistic regression ignoring the over sampling of high exposed subjects (Figure 6.1a) and accounting for high exposed subject in the sample (Figure 6.1b). In this case, data were generated according to the EECC design and using the parameters values describe in data generation section. The dotted lines indicate the true value of the corresponding parameters. As the boxplots indicate, performing a standard logistic regression results in incorrect estimates of all the parameters of the logistic regression model (Figure 6.1a). However, adding an extra covariate indicating high exposure yields reliable estimates of the true parameters (Figure 6.1b).

We also conducted additional simulations to exposure the nature of bias in estimated regression coefficients when there is no interaction ($\gamma = 0$) or a negative interaction ($\gamma = -0.406$) in the true model. The results are given in the Figure 6.2. The results are quite similar to the results with positive interaction parameters, i. e., traditional analysis

ignoring sampling provides biased estimates, however, analysis by adding an indicator variable in the model provides reliable estimates of the true parameters.

We further conduct a simulation study similar to the data collection in our motivating example where cases were randomly selected without stratification and low exposed controls are oversampled. Specifically, we set a lower cut off value ($k = 0.4$) for exponential exposure distribution, corresponding to 33% exposure information in the low exposed group. We then sampled equal number of controls from high/low exposed group based on above cut-off. All other parameters involved in this simulation were similar to values presented in Figure 6.1. The results are given in Figure 6.3. The EECC design in this case also provides reliable estimates of the true regression coefficients.

6.4.2 Estimation of Power

In this section we will compare power to detect gene environment interaction effect employing EECC design and traditional case control design (sampling from case and control population independent of exposure status). Given a particular value of the gene environment interaction parameters under the alternative hypothesis, the power is a function of the following parameters: the magnitude of the type I error (α), the sample size (n), the gene frequency (P_G), the exposure distribution (E), the control to case ratio(r_c), the ratio of high and low exposed sample(r_H), and the cut-off, k above (or below) which the exposure is considered to be high (or low). Therefore, we estimate power by varying one of the aforementioned parameters, the remaining parameters were held fixed. For all the power comparisons, unless stated otherwise the exposure distribution is considered as exponential (rate=1) and distribution of gene prevalence is considered as Binomial ($P_G = 0.2$).

Relation between Power and Gene Frequency

Figure 6.4 illustrates the power comparison to detect the interaction parameter β_{GE} using traditional case control design and EECC design for different values of the prevalence of genetic susceptibility, P_G and sample size of 2000. As expected, power to

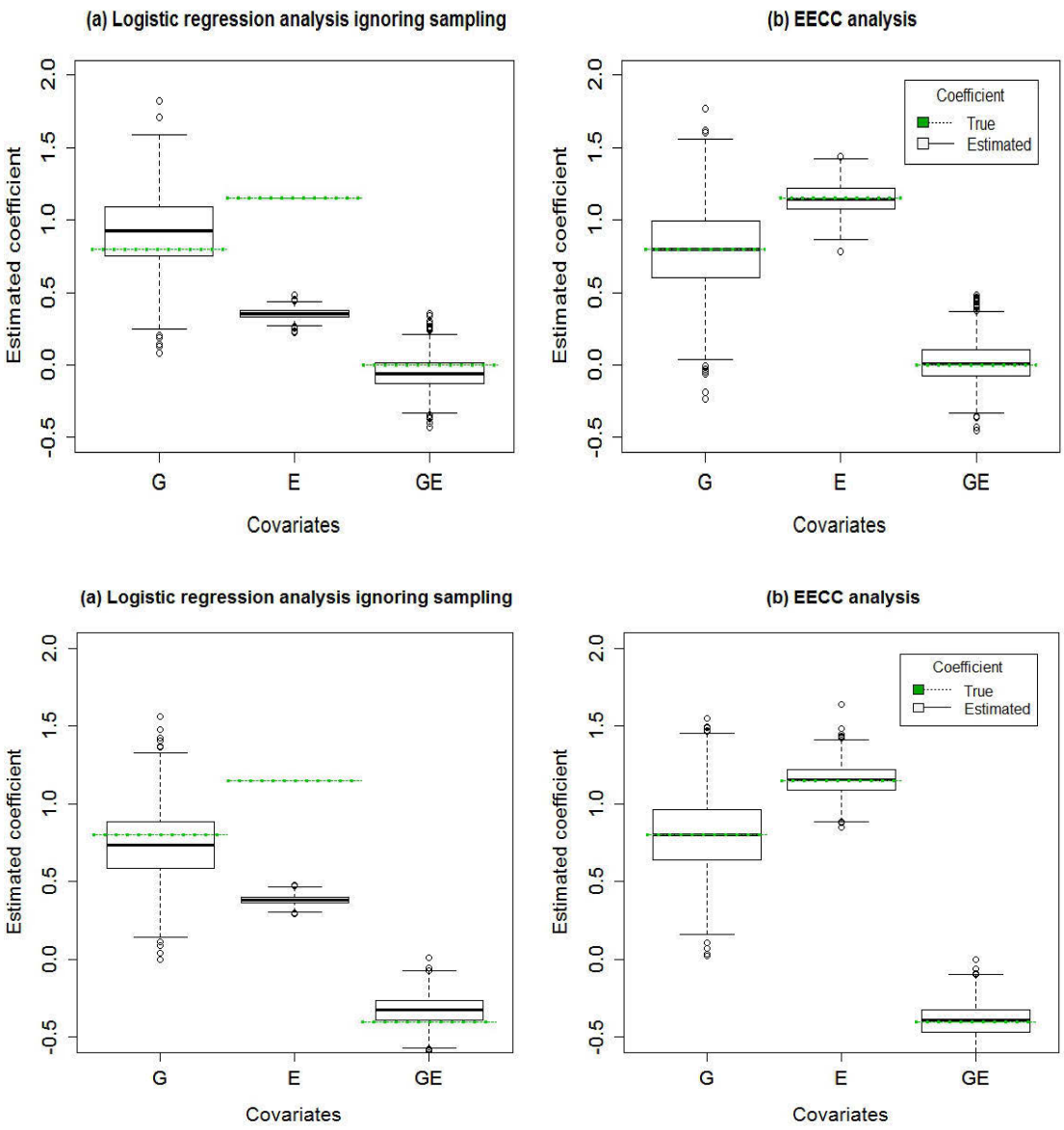


Figure 6.2: Comparison of the estimated parameters for (a) Logistic regression analysis ignoring sampling and (b) EECC design when true interaction parameters are 0 (upper panel) and -0.406 (lower panel), respectively.

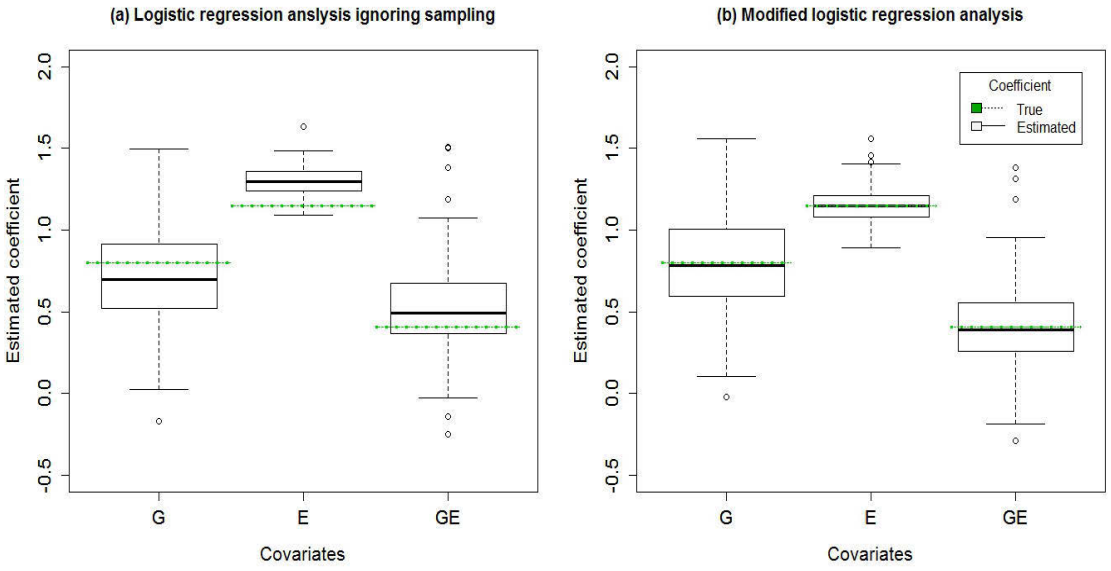


Figure 6.3: Comparison of Estimated parameters for (a) Logistic regression analysis ignoring sampling and (b) EECC design when cases are randomly selected and controls are selected by oversampling of low exposed subjects.

detect interaction parameter decreases with low gene prevalence. However, the EECC design yields better power compare to traditional design for all cases with various probability of gene susceptibilities (result not shown).

Relation between Power and Case-Control Ratio

In Figure 6.5, the power is shown as a function of control-case ratio (r_C) for a given number of cases using EECC design, see Figure 6.5(a), and using traditional case control design; see Figure 6.5(b). Power increases as the number of controls increases for a fixed number of cases in both the design. Similar to the classical case control studies (Taylor, 1986), most of the gain is evident if the control to case ratio is at most 4 in EECC design. However, the EECC design outperforms the traditional case control design in obtaining power to detect gene-environment interaction.

Relation between power and ratio of high and low exposed sample

In Figure 6.6, the power is shown as a function of high to low exposure ratio (r_H) in the sample with equal number of case and control. Sampling of more high exposed subjects

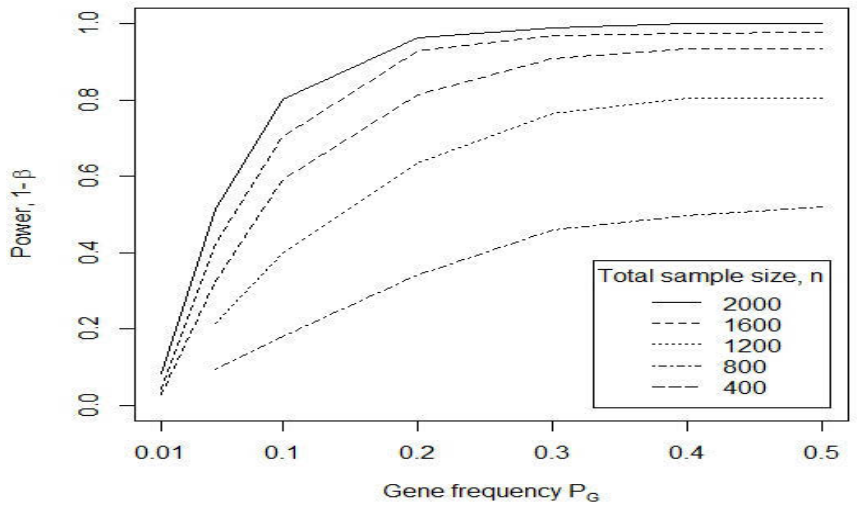


Figure 6.4: Comparison of estimated power to detect gene-interaction effect obtained via a traditional case-control design and EECC design for a sample size of 2000.

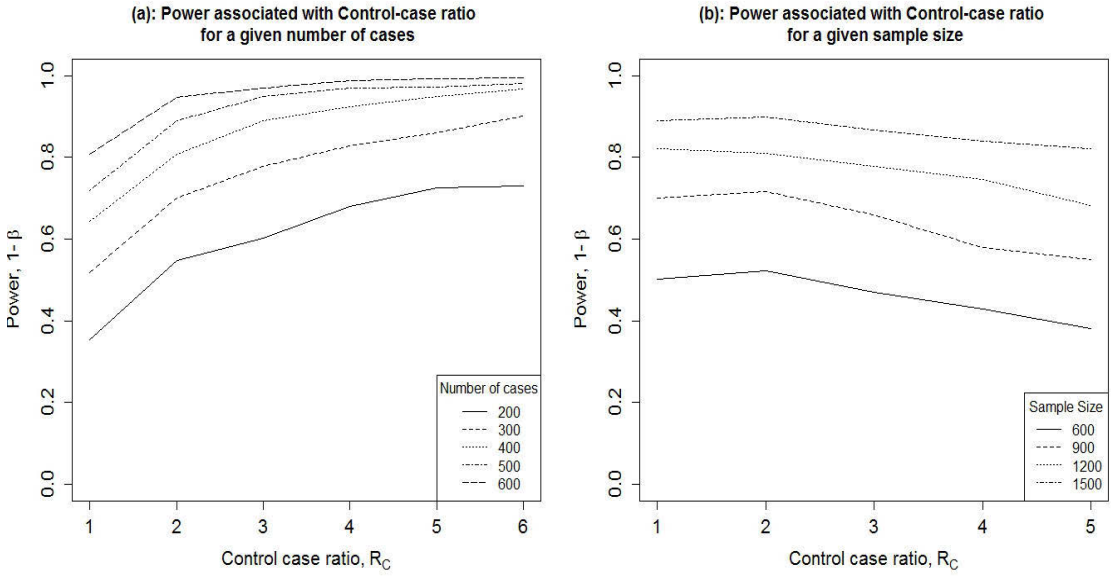


Figure 6.5: Power as a function of control-case ratio and sample sizes: (a) EECC design (b) Traditional case control design.

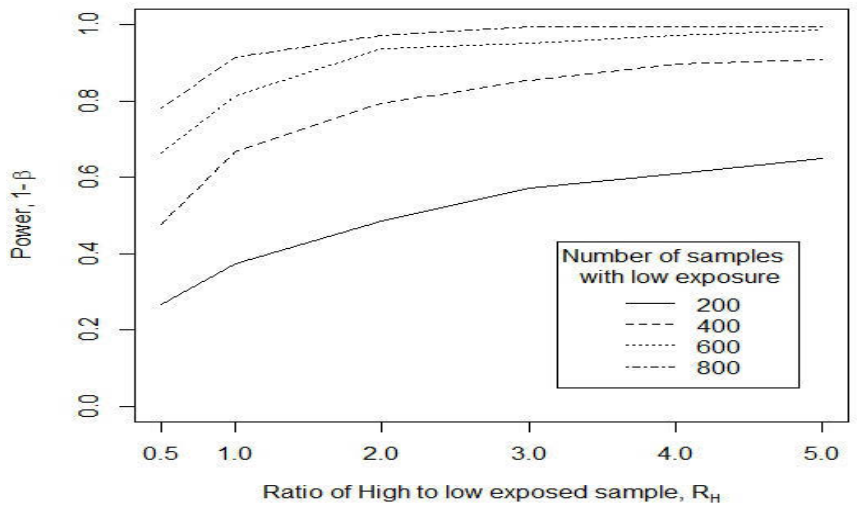


Figure 6.6: Power as a function of ratio of high and low exposed sample. Different values of low exposed samples result in a different trajectory for power.

compared to low exposed subjects resulted in increased power for exponential (rate=1) distribution. Most of the gain in terms of power is achieved if the ratio of high to low exposed subjects is between 1 and 3.

Relation between Power and Asymmetry of the Exposure Distribution

In Table 6.1, we evaluate the estimated regression coefficients, their standard errors and power for detecting the gene-environment interaction effect as a function of the asymmetry of the exposure distribution and level of exposure cut off, k. Various values of cut off were examined to ensure varying proportions (5% to 95%) of exposure information lies above (or below) the cut off for EECC design, whereas, the samples for the traditional case control design were selected independent of exposure status. Specifically, we simulated exposure distribution following $Beta(6,2)$, $Beta(6,6)$ and $Beta(2,6)$, where $Beta(\alpha,\beta)$ represents Beta distribution with shape parameters α and β. Under the above specification the exposure distribution is either negatively skewed, symmetric or positively skewed. The true parameter values used in this particular simulation is given in the first row of the table. As expected, the power to detect a significant interaction effect increases considerably with oversampling from the skewed tail area of the exposure distribution. In the case of a symmetric exposure distribution,

oversampling from lower tail areas boosts in power. The estimated regression coefficients remain consistent to the true values. Furthermore, if the cut off for oversampling is selected appropriately, there is a gain in efficiency for estimating gene and gene-environment interaction effect. However, addition of an indicator variable related to exposure status decrease efficiency for the continuous exposure effect.

Relation between Power and Exposure Distributions

To examine the performance of our proposed method with other exposure distribution, we compared the performance of EECC design with traditional case control design for various right tailed exposure distributions e.g., exponential (rate=1), Weibull (shape=2.5,scale=1) and gamma (shape=3,rate=2). We select suitable cut off values so that a 10% of the total exposure information lies above these cut off. The results are given in the Table 6.2. For all the sample sizes and exposure distribution compared, the EECC design resulted in higher power for detecting the interaction effect than the traditional case-control study design.

Relation between Power/Probability of Type I Error with the Signs of Interaction Parameter

We also estimate the power and probability of type I error in simulation studies corresponding to the true interaction parameters -0.406 and 0, respectively with exponential exposure distribution. The power to detect negative interaction parameter remains similar to that of positive interaction parameters with the same cut off for oversampling. This indicates that power doesnt depend on the sign of the interaction parameters rather the skewness of the true exposure distribution. Moreover, type I error probability corresponding to testing interaction parameter 0, for the EECC design remained closed to 0.05 (results not shown in table).

Table 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.

Exposure Distribution	% of exposure in the left tail	Cut off	Traditional case-control design $\beta_E(se)$	$\beta_G(se)$	$\beta_{GE}(se)$	Power	EECC $\beta_E(se)$	$\beta_G(se)$	$\beta_{GE}(se)$	Power
True regression coefficient			1.15	0.8	1.5		1.15	0.8	1.5	
Beta (6,2)	5%	0.48	1.167 (0.533)	0.830 (0.711)	1.480 (0.920)	0.372	1.132 (0.658)	0.793 (0.376)	1.153 (0.603)	0.729
(Left skewed)	10%	0.55					1.181 (0.687)	0.823 (0.465)	1.487 (0.708)	0.592
	20%	0.63					1.175 (0.693)	0.803 (0.506)	1.511 (0.736)	0.503
	30%	0.69					1.127 (0.783)	0.792 (0.593)	1.518 (0.829)	0.439
	40%	0.73					1.164 (0.753)	0.825 (0.660)	1.476 (0.885)	0.406
	60%	0.81					1.161 (0.793)	0.799 (0.763)	1.519 (0.974)	0.393
	80%	0.88					1.133 (0.738)	0.784 (0.751)	1.527 (0.900)	0.380
	90%	0.92					1.145 (0.719)	0.801 (0.730)	1.509 (0.857)	0.381
	95%	0.95					1.128 (0.693)	0.774 (0.785)	1.543 (0.904)	0.410
Beta (6,6)	5%	0.27	1.178 (0.502)	0.823 (0.469)	1.474 (0.886)	0.363	1.182 (0.650)	0.799 (0.279)	1.516 (0.678)	0.595
(Symmetric)	10%	0.32					1.170 (0.679)	0.792 (0.337)	1.540 (0.762)	0.534
	20%	0.38					1.167 (0.722)	0.796 (0.377)	1.527 (0.835)	0.450
	30%	0.42					1.134 (0.705)	0.767 (0.405)	1.562 (0.845)	0.442
	40%	0.46					1.183 (0.732)	0.814 (0.450)	1.488 (0.904)	0.375
	60%	0.54					1.131 (0.739)	0.793 (0.470)	1.531 (0.868)	0.393
	80%	0.62					1.160 (0.726)	0.801 (0.500)	1.510 (0.851)	0.430
	90%	0.68					1.143 (0.716)	0.806 (0.495)	1.502 (0.792)	0.460
	95%	0.73					1.162 (0.652)	0.833 (0.491)	1.459 (0.751)	0.499
Beta (2,6)	5%		1.156 (0.448)	0.802 (0.255)	1.531 (0.851)	0.433	1.162 (0.605)	0.800 (0.173)	1.532 (0.780)	0.498
	10%	0.08					1.139 (0.640)	0.786 (0.193)	1.601 (0.838)	0.471
Right skewed	20%	0.12					1.149 (0.660)	0.800 (0.191)	1.512 (0.832)	0.428
	30%	0.16					1.162 (0.685)	0.797 (0.235)	1.525 (0.908)	0.446
	40%	0.19					1.181(0.670)	0.799 (0.228)	1.555 (0.836)	0.420
	60%	0.27					1.152 (0.699)	0.786 (0.265)	1.562 (0.825)	0.467
	80%	0.37					1.130 (0.692)	0.806 (0.284)	1.508 (0.744)	0.518
	90%	0.45					1.147 (0.648)	0.800 (0.301)	1.519 (0.715)	0.619
	95%	0.52					1.146 (0.593)	0.805 (0.290)	1.512 (0.626)	0.701

Table 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.

			Traditional case-control design				EECC			
Sample Size, n	Exposure Distribution	Cut-off	β_E se(β_E)	β_G se(β_G)	β_{GE} se(β_{GE})	Power	$\beta_E(se)$ se(β_E)	$\beta_G(se)$ se(β_G)	$\beta_{GE}(se)$ se(β_{GE})	Power
True regression coefficient			1.15	0.8	0.406		1.15	0.8	0.406	
1600	Exp (1)	2.3	1.149	0.780	0.418	0.654	1.153	0.797	0.408	0.938
			(0.072)	(0.253)	(0.181)		(0.090)	(0.241)	(0.119)	
	Weibull	1.74	1.152	0.786	0.430	0.499	0.149	0.791	0.425	0.761
	(1.5, 1)		(0.099)	(0.258)	(0.220)		(0.130)	(0.279)	(0.170)	
	Gamma	2.66	1.155	0.784	0.424	0.574	1.156	0.810	0.408	0.838
	(3, 2)		(0.079)	(0.335)	(0.198)		(0.100)	(0.332)	(0.143)	
1200	Exp (1)	2.3	1.156	0.781	0.427	0.530	1.158	0.794	0.412	0.861
			(0.082)	(0.291)	(0.215)		(0.097)	(0.289)	(0.141)	
	Weibull	1.74	1.155	0.769	0.437	0.405	1.156	0.792	0.415	0.620
	(1.5,1)		(0.120)	(0.300)	(0.254)		(0.148)	(0.311)	(0.188)	
	Gamma	2.66	1.156	0.789	0.424	0.459	1.161	0.811	0.409	0.725
	(3, 2)		(0.097)	(0.397)	(0.233)		(0.121)	(0.371)	(0.163)	

6.5 Application to the Arsenic Data from Bangladesh

We re-analyze data from a case-control study designed to evaluate the joint effect of genetic polymorphisms and drinking water arsenic exposure on skin lesions (Breton et al., 2007). These data were collected from 23 villages of the Pabna district in Bangladesh, where a range of high and low well water arsenic levels were suspected due to their proximity to the Ganges river. Therefore, to ensure a sufficient range of drinking water arsenic exposure and to prevent over-matching on exposure, the study investigators made sure that 80% of the controls were selected from communities having suspected low exposed arsenic contamination ($< 50\mu_g/l$). More detailed descriptions of the data collection have been given elsewhere (Breton et al., 2007; McCarty et al., 2006).

Previously, analyzing these data, Breton et al. (2007), reported that the X-ray repair cross complementing group 1 (XRCC1 Arg194Trp) polymorphism has a significant interaction with toenail arsenic concentrations. We evaluate this relationship employing our proposed method. We defined an indicator variable that indicates whether or not the sample is obtained from high exposed communities ($\geq 50\mu_g/l$). We assessed the gene-environment interaction in a crude and adjusted logistic regression model, with the latter accounting for the potential confounders such: age, sex, village, body mass index (BMI), education, ever smoked status and ever chewed betel nuts status. These models were then compared with crude and adjusted modified logistic regression model (6.3). Of the 1800 participating cases and controls, 1756 (98%) were genotyped successfully for XRCC1 Arg194Trp genotypes. Complete information on well water arsenic, toenail arsenic, BMI, education, smoking and betel nut chewing was available for 1676 of these participants. Crude and adjusted odds ratios along with 95% CI and p-values for both the traditional case-control analysis as well as our EECC analysis are displayed in Table 6.3. The results are qualitatively similar in that we see a significant association between arsenic exposure and skin lesions, as well as a significant gene-environment interaction. Specifically, participants with the Trp/Arg genotype have a significantly stronger dose response associated with arsenic exposure. However, the magnitudes of estimated effects are different between the traditional and EECC analyses. Whereas the traditional analysis suggested an adjusted odds ratio of 4.12 ($95\%CI : 3.17 - 5.36$), our EECC analysis reduced the estimated dose response coefficient to 3.15 ($95\%CI : 2.37 - 4.20$). The EECC analysis has much greater efficiency (smaller CI lengths) compare to

Table 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.

	Traditional case-control design						EECC					
	Crude analysis			Adjusted analysis			Crude analysis			Adjusted analysis		
Exposure	OR	95% CI	P-value	aOR	95% CI	P-value	OR	95% CI	P-value	aOR	95% CI	P-value
log(Toenail)	4.01	(3.11-5.16)	<0.001	4.12	(3.17-5.36)	<0.001	3.10	(2.35-4.09)	<0.001	3.15	(2.37-4.20)	<0.001
XRCC1Arg194Trp												
Arg/Arg	Ref			Ref			Ref			Ref		
Trp/Arg	1.10	(0.80-1.51)	0.56	1.10	(0.80-1.51)	0.56	1.08	(0.79-1.48)	0.63	1.08	(0.79-1.47)	0.65
Trp/Trp	0.97	(0.33-2.86)	0.96	1.01	(0.34-2.99)	0.99	1.22	(0.44-3.35)	0.70	1.29	(0.46-3.61)	0.63
log(Toenail) * XRCC1Arg194Trp												
lnTA*Trp/Arg	0.53	(0.31-0.91)	0.02	0.52	(0.31-0.90)	0.02	0.56	(0.33-0.94)	0.03	0.55	(0.33-0.94)	0.03
lnTA*Trp/Trp	0.37	(0.06-2.21)	0.28	0.35	(0.06-2.07)	0.25	0.28	(0.05-1.58)	0.15	0.26	(0.05-1.45)	0.12

traditional design. The estimates for genetic effect and gene-environment interaction effect are similar for the traditional and EECC design.

6.6 Discussion

In this paper, we have introduced the Exposure Enriched Case Control (EECC) study which over-sampled subjects according to case/control status, as well as a categorization of exposure (e.g. high versus low) into the study. We have shown via simulations that the EECC can significantly boost the power to detect gene-environment interaction, especially in the case of rare genetic variants and skewed exposure distributions. Stenzel et al. (2015) also use the term Exposure Enriched and argue that oversampling high exposure can boost the power to detect gene-environment interactions. However, they analyzed the resulting data via ordinary logistic regression method and did not suggest any analysis strategy to remove the biased induced by oversampling highly exposed individuals. Our EECC method removes the bias induced by oversampling high exposed individuals through the addition of a simple indicator covariate that reflects high versus low exposure. Our approach assumes that case/control status will be modeled as a function of a continuous exposure variable, genetic susceptibility and their interactions. Our approach differs from other analysis methods for data collected via biased sampling in that it does not require knowledge of the explicit sampling probabilities (Breslow & Cain, 1988; Breslow & Chatterjee, 1999; Weinberg & Wacholder, 1990; White, 1982). Our proposed EECC method has the advantage of simplicity since no specialized software is required.

Although existing two stage case control designs (Breslow & Cain, 1988; Breslow & Chatterjee, 1999) and their matched variant, counter matching (Andrieu et al., 2001), are known to have higher power than traditional case control design, they can only be used if surrogate information on gene, exposure or both is available. The efficiency obtained from these two designs though similar, counter matching designs are complex and require two specific and sensitive surrogates for the risk factor of interest. Our EECC design is simpler and use similar underlying probability principle as pseudo-likelihood analysis based on a two stage design, hence will results in similar efficiency for an appropriate oversampling of high exposed individuals. Breslow and

Chatterjee (1999) compared efficiency of two stage case control design with traditional case control design using weighted likelihood, pseudo-likelihood and non-parametric maximum likelihood approaches. They noted that pseudo-likelihood analysis provides better results in a balanced two stage design (similar numbers of exposed and unexposed individuals within case and control samples) but are generally worse than non-parametric maximum likelihood approach. The estimation of the EECC design using non-parametric maximum likelihood is beyond the scope of the present thesis. Future research studies can be carried out in this regards. Other designs such as family based designs (see Thomas 2010 for a recent review) are appealing in gene-environment interaction studies. However, they generally have less power to test main effects, relative to case-control studies using unrelated controls (Thomas, 2010) Moreover, they are very sensitive to the independence assumptions of gene and environment effects (Albert et al., 2001). The empirical comparison of the above designs with our proposed EECC design is beyond the scope of the current paper. However, interested reader might consider recent review (Thomas, 2010) for a detailed comparison among some of these methods.

While our paper has focussed primarily on introducing the EECC method, we have also presented a re-analysis of data from a case-control study from Bangladesh, where low exposed control subjects had been over-sampled (McCarty et al., 2006). They used traditional logistic regression with a sensitivity analysis to explain the effect of this biased sampling. However, the authors reported that they were not able to make a succinct conclusion about the observed exposure-response relationship between arsenic levels in tube-well drinking water and skin lesions, due to oversampling of controls from the low exposed area. Our EECC approach rectifies the analysis with the addition of an extra covariate indicating the oversampling rules in the model.

While our proposed EECC methodology has a number of appealing features, there are some limitations that could be addressed in future studies. Our proposed EECC design is currently allows only a binary cut off variable to represent oversampling from tail area. However, in application more than one exposure levels in the tail area might be of interest. Future work need to accommodate such extension. In environmental epidemiology, exposure is often susceptible to measurement error (Huque et al., 2014). In the case of exposure misclassification, it is well known that the estimates of the regression coefficients will be attenuated (Stefanski & Carroll, 1985) and may distorts

the power gain of exposure enriched design. Although various methods have been proposed in the literature to correct the effect exposure measurement error in gene-environment interaction studies (Lobach, Fan, & Carroll, 2010; Lobach, Mallick, & Carroll, 2011; Spiegelman, Rosner, & Logan, 2000; Zhang, Mukherjee, Ghosh, Gruber, & Moreno, 2008), however, further research is needed to evaluate and incorporate such extension into our proposed EECC methodology.

Despite these potential limitations, our EECC design can be regarded as a simple alternative to traditional two-stage designs. Furthermore the EECC methodology enhances power to detect the joint influence of genetic and environment exposure for a given sample size compare to traditional case-control studies. Therefore, it has a very strong potential to be used in practice. This design also has potential to be used in context of risk analysis where interest lies in quantifying dose response relationships (Piegorsch, 2010).

Appendix 6A

Let Δ is the sampling indicator with $\Delta = 1$ if the individual is selected in the sample and 0 if not selected and the probability of selecting an individual in the sample is denoted by $\rho(D, E, G)$ that is, $\rho(D, E, G) = Pr(\Delta = 1|D, E, G)$. Then following Bayes rule (Hosmer et al., 2004) we have $Pr(D = 1|E, G, \Delta = 1) = \frac{Pr(D=1,\Delta=1|E,G)}{Pr(\Delta=1|E,G)}$ and $Pr(D = 0|E, G, \Delta = 1) = \frac{Pr(D=0,\Delta=1|E,G)}{Pr(\Delta=1|E,G)}$.

Therefore,

$$
\begin{aligned}
logit[Pr(D = 1|E, G, \Delta = 1)] &= \ln\left\{\frac{Pr(D = 1|E, G, \Delta = 1)}{Pr(D = 0|E, G, \Delta = 1)}\right\} \\
&= \ln\left\{\frac{Pr(D = 1, \Delta = 1|E, G)}{Pr(D = 0, \Delta = 1|E, G)}\right\} \\
&= \ln\left\{\frac{Pr(\Delta = 1|D = 1, E, G) \times Pr(D = 1|E, G)}{Pr(\Delta = 1|D = 0, E, G) \times Pr(D = 0|E, G)}\right\} \\
&= \ln\left\{\frac{Pr(\Delta = 1|D = 1, E, G)}{Pr(\Delta = 1|D = 0, E, G)}\right\} + \ln\left\{\frac{Pr(\mathcal{D} = 1|E, G)}{Pr(D = 0|E, G)}\right\} \\
&= \ln\left\{\frac{Pr(\Delta = 1|D = 1, E, G)}{Pr(\Delta = 1|D = 0, E, G)}\right\} + \beta_0 + \beta_E E + \beta_G G + \beta_{GE} EG \\
&= \ln\left\{\frac{\rho(D = 1, E, G)}{\rho(D = 0, E, G)}\right\} + \beta_0 + \beta_E E + \beta_G G + \beta_{GE} EG
\end{aligned}
$$

Chapter 7

Conclusion

This thesis has made several original contributions. The first contribution is the extension of the Conditional Auto-Regressive (CAR) model to incorporate individual level covariate data in spatial modeling. We have proposed both a linear and a semiparametric adjustment of a continuous individual level covariate effect on outcome of interest in Chapter 2 and 3, respectively. In both cases, we have shown that parameter estimation can be carried out in a distributed computing framework, thus achieving a helpful reduction in computational cost and memory requirements. These contributions also provide a convenient way to extend recent developments in Big Data for independent responses to spatially correlated response. However, the estimated parameters of the CAR based ecological regression method strongly depends on the assumed spatial correlation structure. We have shown that such sensitivity is especially likely when the important ecological covariate is measured with error.

In Chapter 4, we have extended classical measurement error theory to show that the amount of attenuation depends on the degree of spatial correlation in both the covariate of interest as well as the assumed random error from the regression model. These results explain why the results from spatial regression modeling are often so sensitive to the assumed model error structure. Based on these results, we have proposed two different strategies to obtain consistent estimates of the regression coefficients of interest in the presence of covariate measurement error. These strategies include 1) post hoc adjustment of the estimated regression coefficient via an estimate of the attenuation

factor and 2) a linear transformation of the error prone covariates that can then be analyzed to yield consistent results. We have found that both methods perform well, though the second method tends to be less variable and hence preferable in practice. We have presented formulas for the standard errors of the adjusted estimated regression coefficients, though these do not fully account for the uncertainty associated with the estimation of unknown parameters. In practice, a bootstrap procedure can be used to obtain appropriate standard errors. However, this approach is fully parametric and the standard error is underestimated in the case when measurement error variances are estimated from the data.

In Chapter 5, we have developed a semi-parametric framework to obtain a consistent estimate of the true regression coefficients when covariates are measured with error in spatial regression modeling settings. Asymptotic theory establishes that our approach provides consistent, asymptotically normal estimates for the regression coefficient. The theory yields both model based and simulation based standard error estimates. Our empirical simulation results confirm that ignoring measurement error and conducting naive analysis using both generalized additive model and linear mixed model attenuates the estimated regression coefficient towards the null hypothesis of no effect. These results also confirm the result presented in the previous chapter that the degree of measurement error bias depends on the assumed correlation structure. Our work on measurement error in spatial linear regression model is an important addition in the literature as methods to treat error in correlated covariates are lacking.

In the ecological studies group level exposure may not reflect each individual's exposure experience in the group. Case-control studies with detailed individual level exposure information are often desired to support and test hypotheses generated from spatial correlation studies. Recent exploration of the human genome presents new opportunities to understand how genetic and environmental factors interplay to cause disease. The statistical power to detect an interaction effect from a case control study, however, is limited when exposure distribution is highly skewed and disease is rare.

In Chapter 6, we have proposed a variant of the classical case-control study, the exposure enriched case-control (EECC) design, where not only cases, but also individuals with high (or low) exposure are over-sampled, depending on the skewness of the exposure

distribution. We show that judicious over-sampling of high/low exposed individuals can boost study power considerably. Of course, a traditional logistic regression model is no longer valid and results in biased estimates. We have shown that addition of a simple covariate to the regression model removes this bias and yields reliable estimates of main and interaction effects of interest.

We have studied a broad range of methodological issues in this thesis and provide some solutions. However, there are a number of areas where future study would be useful. In the Chapter 2 and 3, we assume individual outcome follows a Poisson distribution with rare disease assumption. In many application, the disease may be binary, hence a Bernoulli assumption for the individual level outcome might be appropriate, but the group level outcome can be well approximated by the Poisson model (Guthrie et al., 2002).

In our analysis of geographical variation of neutropenia admission rates, we modeled the regional counts of neutropenia admission rates adjusting for various types of cancer in multivariable models. However, it might be interesting to investigate neutropenia admission rates in various different types of cancer on the same regional grid. Hence, joint modeling (Held, Natário, Fenton, Rue, & Becker, 2005) or generalized multivariate CAR (Jin, Banerjee, & Carlin, 2007) might be of interest. Future work can explore the possibility of incorporating individual level covariates in these context.

In the Chapter 4 and 5, we have studied the consequences of measurement error in the ecological covariates in a spatial linear regression framework. Future work is needed in order to extend and generalize these approaches in the context of a generalized spatial linear regression model. Moreover, we only studied measurement error biases in a single ecological covariate. Future studies need to accommodate the consequence of measurement error in spatial linear regression when there are are other covariates/confounders in the model, that also may be measured with error. It might also be interesting to study the consequence of measurement error in the individual level covariate adjusted conditional auto-regressive model, where either or both individual and group level covariates are susceptible to measurement error biases.

In Chapter 6, we have shown that judicious oversampling can boost study power and can

provide unbiased estimates of regression parameters of the model. However, we didn't present a sample size calculation. Furthermore, EECC design currently allows only a binary cut off variable to represent oversampling from the tail area. In application, more than one exposure level in the tail area might be of interest. Further work need to accommodate such an extension. In environmental epidemiology, exposures are often susceptible to measurement error (Huque et al., 2014). In the case of exposure misclassification, it is well known that the estimates of the regression coefficients will be attenuated (Stefanski & Carroll, 1985) and may distort the power gain of exposure enriched design. Although various methods have been proposed in the literature to correct the effect of exposure measurement error in gene-environment interaction studies (Lobach et al., 2010, 2011; Spiegelman et al., 2000; Zhang et al., 2008), further research is needed to evaluate and incorporate such extension into our proposed EECC methodology.

Our contributions in the methodological development of environmental epidemiology will be a significant addition to the existing methodology. We believe our work in developing effective algorithms for spatio-temporal modeling and research designs, and applying these to real world cancer registry data has the potential for major real-world impact.

References

Aapro, M., Bohlius, J., Cameron, D., Dal Lago, L., Donnelly, J. P., Kearney, N., ... others (2011). 2010 update of EORTC guidelines for the use of granulocyte-colony stimulating factor to reduce the incidence of chemotherapy-induced febrile neutropenia in adult patients with lymphoproliferative disorders and solid tumours. *European Journal of Cancer*, *47*(1), 8–32.

Ainsworth, L., & Dean, C. (2006). Approximate inference for disease mapping. *Computational Statistics & Data Analysis*, *50*(10), 2552–2570.

Albert, P. S., Ratnasinghe, D., Tangrea, J., & Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, *154*(8), 687–693.

Andrieu, N., & Goldstein, A. (1996). Use of relatives of cases as controls to identify risk factors when an interaction between environmental and genetic factors exists. *International Journal of Epidemiology*, *25*(3), 649–657.

Andrieu, N., Goldstein, A., Thomas, D., & Langholz, B. (2001). Counter-matching in studies of gene-environment interaction: efficiency and feasibility. *American Journal of Epidemiology*, *153*(3), 265–274.

Anselin, L. (2002). Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, *27*(3), 247–267.

Armstrong, B., & Doll, R. (1975). Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International Journal of Cancer*, *15*(4), 617–631.

Ash, A., Fienberg, S. E., Louis, T. A., Normand, S., Stukel, T. A., & Utts, J. (2012). Statistical issues in assessing hospital performance. *COPPS-CMS White Paper*.

Australian Bureau of Statistics. (2015, March). *Regional Population Growth, Australia, 2013-14 (cat. no. 3218.0)*.

Australian Institute of Health and Welfare (AIHW). (2014). *Cancer in Australia: an*

overview 2014. Cancer series No 90. Cat. no. CAN 88. Canberra: AIHW.

Australian Institute of Health and Welfare (AIHW) & Australasian Association of Cancer Registries (AACR). (2012). *Cancer in Australia: an overview 2012.* Cancer series no. 74. Cat. no. CAN 70. Canberra: AIHW.

Baden, L., Bensinger, W., Angarone, M., Casper, C., Dubberke, E., Freifeld, A., ... Shead, D. (2012). Prevention and treatment of cancer-related infections. *JNCCN Journal of the National Comprehensive Cancer Network*, *10*(11), 1412–1445.

Baker, D., Kjellstrom, T., Calderon, R., & Pastides, H. (1999). *Environmental epidemiology : a textbook on study methods and public health applications.* World Health Organization. Retrieved from `http://apps.who.int/iris/handle/10665/66026#sthash.HWLLY3gY.dpuf`

Beale, L., Abellan, J. J., Hodgson, S., & Jarup, L. (2008). Methodologic issues and approaches to spatial epidemiology. *Environmental Health Perspectives*, *116*(8), 1105.

Bell, M. L., & Grunwald, G. K. (2004). Mixed models for the analysis of replicated spatial point patterns. *Biostatistics*, *5*(4), 633–648.

Bernadinelli, L., Pascutto, C., Best, N., & Gilks, W. (1997). Disease mapping with errors in covariates. *Statistics in Medicine*, *16*(7), 741–752.

Bernardinelli, L., & Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, *11*(8), 983–1007.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20.

Best, N. G., Ickstadt, K., & Wolpert, R. L. (2000). Spatial poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association*, *95*(452), 1076–1088.

Bodey, G. P. (2009). Fever and neutropenia: the early years. *Journal of Antimicrobial Chemotherapy*, *63*(suppl 1), i3–i13.

Breslow, N., & Cain, K. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*(1), 11–20.

Breslow, N., & Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(4), 457–468.

Breslow, N., & Clayton, D. (1993). Approximate inference in generalized linear mixed

models. *Journal of the American Statistical Association*, *88*(421), 9-25.

Breslow, N., & Day, N. (1987). *Statistical Methods in Cancer Research. Volume II–The Design and Analysis of Cohort Studies*. New York, U.S.A.: International Agency for Research on Cancer, Oxford University Press.

Breton, C. V., Zhou, W., Kile, M. L., Houseman, E. A., Quamruzzaman, Q., Rahman, M., ... Christiani, D. C. (2007). Susceptibility to arsenic-induced skin lesions from polymorphisms in base excision repair genes. *Carcinogenesis*, *28*(7), 1520–1525.

Burden, S., Guha, S., Morgan, G., Ryan, L., Sparks, R., & Young, L. (2005). Spatio-temporal analysis of acute admissions for ischemic heart disease in NSW, Australia. *Environmental and Ecological Statistics*, *12*(4), 427-448.

Cameron, D. (2009). Management of chemotherapy-associated febrile neutropenia. *British Journal of Cancer*, *101*(Suppl 1), S18–S22.

Cancer Institute NSW. (2011). *Cancer Institute NSW Annual Report.* Sydney.

Carroll, R. J., Chen, R., George, E., Li, T., Newton, H., Schmiediche, H., & Wang, N. (1997). Ozone exposure and population density in Harris County, Texas. *Journal of the American Statistical Association*, *92*(438), 392–404.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Florida, U.S.A.: Chapman and Hall/CRC.

Chatterjee, N., & Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, *92*(2), 399–418.

Chatterjee, N., Kalaylioglu, Z., & Carroll, R. J. (2005). Exploiting gene-environment independence in family-based case-control studies: Increased power for detecting associations, interactions and joint effects. *Genetic Epidemiology*, *28*(2), 138–156.

Chen, J., Kang, G., VanderWeele, T., Zhang, C., & Mukherjee, B. (2012). Efficient designs of gene–environment interaction studies: implications of Hardy–Weinberg equilibrium and gene–environment independence. *Statistics in Medicine*, *31*(22), 2516–2530.

Clayton, D., & Bernardinelli, L. (1992). *Bayesian Methods for Mapping Disease Risk in Geographical and Environmental Epidemiology Methods for Small-Area Studies* (P. Elliott, J. Cuzick, D. English, & R. Stern, Eds.). World Health Organization.

Clayton, D., Bernardinelli, L., & Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, *22*(6), 1193–1202.

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

Clayton, D., & McKeigue, P. M. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *The Lancet*, *358*(9290), 1356–1360.

Cook, D. G., & Pocock, S. J. (1983). Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics*, *39*(2), 361–371.

Cox, D. R. (1990). Role of models in statistical analysis. *Statistical Science*, *5*(2), 169–174.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London, UK: Chapman and Hall.

Crawford, J., Dale, D. C., & Lyman, G. H. (2004). Chemotherapy-induced neutropenia. *Cancer*, *100*(2), 228–237.

Cressie, N. (1993). *Statistics for spatial data* (G. Watson, Ed.). New York, U.S.A.: John Wiley & Sons.

Cressie, N., & Chan, N. H. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association*, *84*(406), 393–401.

David, J. (1995). *Cancer care: prevention, treatment and palliation*. Nelson Thornes.

Devine, O. J., Louis, T. A., & Halloran, M. E. (1994). Empirical Bayes estimators for spatially correlated incidence rates. *Environmetrics*, *5*(4), 381–398.

Diggle, P. J., Tawn, J., & Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(3), 299–350.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., & Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International Journal of Health Geographics*, *6*(1), 54-65.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Florida, U.S.A.: Chapman & Hall.

Elliot, P., Wakefield, J., Best, N., & Briggs, D. (2000). *Spatial epidemiology: methods and applications*. New York, U.S.A: Oxford University Press.

Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives*, *112*(9), 998-1006.

Enea, M. (2012). speedglm: Fitting linear and generalized linear models to large data sets. *R package version 0.1*.

Farrow, D. C., Hunt, W. C., & Samet, J. M. (1992). Geographic variation in the treatment of localized breast cancer. *New England Journal of Medicine*, *326*(17), 1097–1101.

Foppa, I., & Spiegelman, D. (1997). Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. *American Journal of Epidemiology*, *146*(7), 596–604.

Fox, P., & Boyce, A. (2014). Cancer health inequality persists in regional and remote Australia. *Medical Journal Australia*, *201*(8), 445–446.

Fuller, W. (1987). *Measurement error models*. New York, U.S.A: John Wiley & Sons.

García-Closas, M., & Lubin, J. H. (1999). Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *American Journal of Epidemiology*, *149*(8), 689–692.

Gardner, M. (1973). Using the environment to explain and predict mortality. *Journal of the Royal Statistical Society. Series A (General)*, *136*(3), 421–440.

Gelman, A. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, *55*(3), 245–259.

Greenland, S. (2001). Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, *30*(6), 1343–1350.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J., & Coull, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, *10*(2), 258–274.

Guha, S., Ryan, L., & Morara, M. (2009). Gauss–Seidel Estimation of Generalized Linear Mixed Models With Application to Poisson Modeling of Spatially Varying Disease Rates. *Journal of Computational and Graphical Statistics*, *18*(4), 818–837.

Guthrie, K. A., Sheppard, L., & Wakefield, J. (2002). A hierarchical aggregate data model with spatially correlated disease rates. *Biometrics*, *58*(4), 898–905.

Haneuse, S., & Bartell, S. (2011). Designs for the combination of group-and individual-level data. *Epidemiology (Cambridge, Mass.)*, *22*(3), 382-389.

Haneuse, S., & Wakefield, J. (2008). The combination of ecological and case–control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 73–93.

Haneuse, S., & Wakefield, J. C. (2007). Hierarchical models for combining ecological and case–control data. *Biometrics*, *63*(1), 128–136.

Hannan, T. J. (1999). Variation in health care: the roles of the electronic medical record. *International Journal of Medical Informatics*, *54*(2), 127–136.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320–338.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Florida, U.S.A: Chapman and Hall/CRC.

Held, L., Natário, I., Fenton, S. E., Rue, H., & Becker, N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research*, *14*(1), 61–82.

Henderson, H. V., & Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *Siam Review*, *23*(1), 53–60.

Hosmer, J., David, W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

Howe, G. M. (1964). National atlas of disease mortality in the United Kingdom. *The Geographical Journal*, *130*(1), 15-22.

Huque, M. H., Bondell, H. D., & Ryan, L. M. (2014). On the impact of covariate measurement error on spatial regression modelling. *Environmetrics*, *25*(8), 560-570. doi: 10.1002/env.2305

Jackson, C., Best, N., & Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in Medicine*, *25*(12), 2136–2159.

Jemal, A., Center, M. M., DeSantis, C., & Ward, E. M. (2010). Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiology Biomarkers & Prevention*, *19*(8), 1893–1907.

Jin, X., Banerjee, S., & Carlin, B. P. (2007). Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(5), 817–838.

Jolley, D., Jarman, B., & Elliott, P. (1992). *Socio-economic confounding in Geographical and Environmental Epidemiology Methods for Small-Area Studies* (P. Elliott, J. Cuzick, D. English, & R. Stern, Eds.). World Health Organization.

Jong, K. E., Smith, D. P., Xue, Q. Y., O'Connell, D. L., Goldstein, D., & Armstrong, B. K. (2004). Remoteness of residence and survival from cancer in New South Wales. *Medical Journal of Australia*, *180*(12), 618-622.

Kammann, E., & Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*(1), 1–18.

Kaufman, L., & Rousseeuw, P. (2005). *Finding groups in data: An introduction to cluster analysis*. New Jersey, U.S.A.: John Wiley & Sons.

Khoury, M. J. (1994). Case-parental control method in the search for disease-susceptibility genes. *American Journal of Human Genetics*, *55*(2), 414.

Khoury, M. J., Adams Jr, M., & Flanders, W. D. (1988). An epidemiologic approach to ecogenetics. *American Journal of Human Genetics*, *42*(1), 89.

Khoury, M. J., & Flanders, W. D. (1996). Nontraditional epidemiologic approaches in the analysis of gene environment interaction: Case-control studies with no controls! *American Journal of Epidemiology*, *144*(3), 207–213.

Klastersky, J., Paesmans, M., Rubenstein, E. B., Boyer, M., Elting, L., Feld, R., ... others (2000). The multinational association for supportive care in cancer risk index: a multinational scoring system for identifying low-risk febrile neutropenic cancer patients. *Journal of Clinical Oncology*, *18*(16), 3038–3051.

Knorr-Held, L., & Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *164*(1), 73–85.

Kraft, P., Yen, Y.-C., Stram, D. O., Morrison, J., & Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity*, *63*(2), 111–119.

Kuderer, N. M., Dale, D. C., Crawford, J., Cosler, L. E., & Lyman, G. H. (2006). Mortality, morbidity, and cost associated with febrile neutropenia in adult cancer patients. *Cancer*, *106*(10), 2258–2266.

Langford, I. H., Leyland, A. H., Rasbash, J., & Goldstein, H. (1999). Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(2), 253–268.

Langholz, B., & Borgan, Ø. R. (1995). Counter-matching: a stratified nested case-control sampling method. *Biometrika*, *82*(1), 69–79.

Lawson, A. (2013). *Statistical methods in spatial epidemiology* (2nd ed.). West Sussex, England: John Wiley & Sons.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, *2*(2), 79–89.

Leroux, B. G., Lei, X., & Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (p. 179-191).

New York: Springer.

Leyland, A. H., Langford, I. H., Rasbash, J., & Goldstein, H. (2000). Multivariate spatial models for event data. *Statistics in Medicine*, *19*(17-18), 2469–2478.

Li, Y., Tang, H., & Lin, X. (2009). Spatial linear mixed models with covariate measurement errors. *Statistica Sinica*, *19*(3), 1077-1093.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 381–400.

Lingaratnam, S., Slavin, M., Mileshkin, L., Solomon, B., Burbury, K., Seymour, J., ... others (2011). An Australian survey of clinical practices in management of neutropenic fever in adult cancer patients 2009. *Internal Medicine Journal*, *41*(1b), 110–120.

Liu, C.-y., Maity, A., Lin, X., Wright, R. O., & Christiani, D. C. (2012). Design and analysis issues in gene and environment studies. *Environmental Health*, *11*(1), 1-15.

Lobach, I., Fan, R., & Carroll, R. J. (2010). Genotype-based association mapping of complex diseases: gene-environment interactions with multiple genetic markers and measurement error in environmental exposures. *Genetic Epidemiology*, *34*(8), 792–802.

Lobach, I., Mallick, B., & Carroll, R. J. (2011). Semiparametric Bayesian analysis of gene-environment interactions with error in measurement of environmental covariates and missing genetic data. *Statistics and Its Interface*, *4*(3), 305-316.

Luan, J., Wong, M., Day, N., & Wareham, N. (2001). Sample size determination for studies of gene-environment interaction. *International Journal of Epidemiology*, *30*(5), 1035–1040.

Lumley, T. (2011). biglm: Bounded memory linear and generalized linear models. *R package version 0.8,*. Retrieved from `URLhttp://cran.r-project.org/web/packages/biglm`

Lyman, G. H., Michels, S. L., Reynolds, M. W., Barron, R., Tomic, K. S., & Yu, J. (2010). Risk of mortality in patients with cancer who experience febrile neutropenia. *Cancer*, *116*(23), 5555–5563.

MacNab, Y. C. (2003). Hierarchical bayesian modeling of spatially correlated health service outcome and utilization rates. *Biometrics*, *59*(2), 305–315.

MacNab, Y. C. (2004). Bayesian spatial and ecological models for small-area accident and

injury analysis. *Accident Analysis & Prevention*, *36*(6), 1019–1028.

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, *39*(1), 53–74.

Marshall, R. J. (1991). Mapping disease and mortality rates using empirical bayes estimators. *Applied Statistics*, 283–294.

Martínez, J. M., Benach, J., Benavides, F. G., Muntaner, C., Clèries, R., Zurriaga, O., ... Yasui, Y. (2009). Improving multilevel analyses: the integrated epidemiologic design. *Epidemiology*, *20*(4), 525–532.

Martinez, J. M., Benach, J., Ginebra, J., G Benavides, F., & Yasui, Y. (2007). An integrated analysis of individual and aggregated health data using estimating equations. *The International Journal of Biostatistics*, *3*(1).

McCarty, K. M., Houseman, E. A., Quamruzzaman, Q., Rahman, M., Mahiuddin, G., Smith, T., ... Christiani, D. C. (2006). The impact of diet and betel nut use on skin lesions associated with drinking-water arsenic in Pabna, Bangladesh. *Environmental Health Perspectives*, 334-340.

McKenzie, D. (2003). Measure inequality with asset indicators, in bread working paper no. 042.

Miettinen, O. S. (1985). The case-control study: valid selection of subjects. *Journal of Chronic Diseases*, *38*(7), 543–548.

Miller, R. A. (1994). Medical diagnostic decision support systems: Past, present, and future a threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association*, *1*(1), 8–27.

Mitchell, K., Fritschi, L., Reid, A., McEvoy, S., Ingram, D., Jamrozik, K., ... Byrne, M. (2006). Rural–urban differences in the presentation, management and survival of breast cancer in Western Australia. *The Breast*, *15*(6), 769–776.

Molitor, J., Jerrett, M., Chang, C.-C., Molitor, N.-T., Gauderman, J., Berhane, K., ... others (2007). Assessing uncertainty in spatial exposure models for air pollution health effects assessment. *Environmental Health Perspectives*, *115*(8), 1147-1153.

Mukherjee, B., Ahn, J., Gruber, S. B., Ghosh, M., & Chatterjee, N. (2010). Case–control studies of gene–environment interaction: Bayesian design and analysis. *Biometrics*, *66*(3), 934–948.

Mukherjee, B., & Chatterjee, N. (2008). Exploiting Gene-Environment Independence for Analysis of Case–Control Studies: An Empirical Bayes-Type Shrinkage Estimator to Trade-Off between Bias and Efficiency. *Biometrics*, *64*(3), 685–694.

Nattinger, A. B., Gottlieb, M. S., Veum, J., Yahnke, D., & Goodwin, J. S. (1992). Geographic variation in the use of breast-conserving treatment for breast cancer. *New England Journal of Medicine*, *326*(17), 1102–1107.

Nattinger, A. B., Kneusel, R. T., Hoffmann, R. G., & Gilligan, M. A. (2001). Relationship of distance from a radiotherapy facility and initial breast cancer treatment. *Journal of the National Cancer Institute*, *93*(17), 1344–1346.

Ngo, L., & Wand, M. P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, *9*(1), 1–54.

Nychka, D., & Saltzman, N. (1998). Design of air-quality monitoring networks. In D. Nychka, W. Piegorsch, & L. Cox (Eds.), *Case studies in environmental statistics* (Vol. 132, p. 51-76). Springer, U.S.A. Retrieved from `http://dx.doi.org/10.1007/978-1-4612-2226-2_4` doi: 10.1007/978-1-4612-2226-2_4

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, *70*(349), 120–126.

Pacione, M. (2013). *Medical geography (Routledge Revivals): Progress and prospect*. New York, USA: Routledge.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, *25*(1), 107-125.

Parkin, D. M., Bray, F., Ferlay, J., & Pisani, P. (2005). Global cancer statistics, 2002. *CA: a cancer journal for clinicians*, *55*(2), 74–108.

Pickett, K. E., & Pearl, M. (2001). Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology & Community Health*, *55*(2), 111–122.

Pickle, L., Mungiole, M., Jones, G., & White, A. (1996). *Atlas of United States mortality*. Hyattsville, Maryland: National Center for Health Statistics.

Piegorsch, W. W. (2010). Translational benchmark risk analysis. *Journal of Risk Research*, *13*(5), 653–667.

Piegorsch, W. W., Weinberg, C. R., & Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, *13*(2), 153–162.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2013). nlme: Linear and nonlinear mixed effects models [Computer software manual]. (R package version 3.1-109)

Pinheiro, J., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, U.S.A: Springer Science & Business Media.

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *66*(3), 403–411.

Prentice, R. L., & Sheppard, L. (1990). Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. *Cancer Causes & Control*, *1*(1), 81–97.

Prentice, R. L., & Sheppard, L. (1995). Aggregate data studies of disease risk factors. *Biometrika*, *82*(1), 113-125.

Prüss-Üstün, A., Wolf, J., Corvalán, C., Bos, R., & Neira, M. (2016). *Preventing disease through healthy environments.* Retrieved from `http://apps.who.int/iris/bitstream/10665/204585/1/9789241565196_eng.pdf?ua=1`

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ravenscroft, P., Burgess, W. G., Ahmed, K. M., Burren, M., & Perrin, J. (2005). Arsenic in groundwater of the Bengal Basin, Bangladesh: Distribution, field relations, and hydrogeological setting. *Hydrogeology Journal*, *13*(5-6), 727–751.

Rich, E., & Romero, M. (2005). Exposure to chronic stress downregulates corticosterone responses to acute stressors. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *288*(6), R1628–R1636.

Richardson, S., & Gilks, W. R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, *12*(18), 1703–1722.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*(6), 15–32.

Rosenbaum, P. R., & Rubin, D. B. (1984). Difficulties with regression analyses of age-adjusted rates. *Biometrics*, *40*(2), 437–443.

Rothman, K., Greenland, S., & Lash, T. (2008). *Modern epidemiology, 3rd edition.* U.S.A: Philadelphia, PA: Lippincott, Williams & Wilkins.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, *11*(4), 735-757.

Ruppert, D., Wand, M., & Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, *3*, 1193-1256. doi: http://doi.org/10.1214/09-EJS525

Ruppert, D., Wand, P., & Carroll, R. (2003). *Semiparametric regression.* New York, U.S.A.: Cambridge University Press.

Schlattmann, P., & Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine*, *12*(19-20), 1943–1950.

Sheppard, L. (2003). Insights on bias and information in group-level studies. *Biostatistics*, *4*(2), 265-278.

Sheppard, L., Burnett, R. T., Szpiro, A. A., Kim, S.-Y., Jerrett, M., Pope III, C. A., & Brunekreef, B. (2012). Confounding and exposure measurement error in air pollution epidemiology. *Air Quality, Atmosphere & Health*, *5*(2), 203-216.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., ... others (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*(2), 203–209.

Smith, T. J., Khatcheressian, J., Lyman, G. H., Ozer, H., Armitage, J. O., Balducci, L., ... others (2006). 2006 update of recommendations for the use of white blood cell growth factors: an evidence-based clinical practice guideline. *Journal of Clinical Oncology*, *24*(19), 3187–3205.

Snow, J. (1855). *On the mode of communication of cholera.* London, England.: John Churchill.

Spiegelman, D., Rosner, B., & Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, *95*(449), 51–61.

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, *13*(4), 1335–1351.

Stenzel, S. L., Ahn, J., Boonstra, P. S., Gruber, S. B., & Mukherjee, B. (2015). The impact of exposure-biased sampling designs on detection of gene–environment interactions in case–control studies with potential exposure misclassification. *European Journal of Epidemiology*, *30*(5), 413–423.

Stern, H., & Cressie, N. (1999). *Disease mapping and risk assessment for public health.* (A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, & R. Bertollini, Eds.). John Wiley & Sons.

Szpiro, A. A., Sheppard, L., & Lumley, T. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics*, *12*(4), 610–623.

Taylor, J. M. (1986). Choosing the number of controls in a matched case-control study,

some sample size, power and efficiency considerations. *Statistics in Medicine*, *5*(1), 29–36.

Thomas, D. (2010). Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, *11*(4), 259–272.

Thompson, D., & Easton, D. (2001). Variation in cancer risks, by mutation position, in brca2 mutation carriers. *The American Journal of Human Genetics*, *68*(2), 410–419.

Tierney, W. M., & McDonald, C. (1996). Testing informatics innovations: the value of negative trials. *Journal of the American Medical Informatics Association*, *3*(5), 358-359.

Tierney, W. M., Overhage, J. M., & McDonald, C. J. (1997). Demonstrating the effects of an IAIMS on health care quality and cost. *Journal of the American Medical Informatics Association*, *4*(2), s41-s46.

U. S. Department of Health and Human Services. National Institutes of Health. National Cancer Institute. (2015, February). *What is cancer?* Retrieved from `http://www.cancer.gov/about-cancer/what-is-cancer`

Umbach, D. M., & Weinberg, C. R. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, *16*(15), 1731–1743.

Wakefield, J. (2004a). A critique of statistical aspects of ecological studies in spatial epidemiology. *Environmental and Ecological Statistics*, *11*(1), 31–54.

Wakefield, J. (2004b). Ecological inference for 2× 2 tables (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *167*(3), 385–445.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, *8*(2), 158–183.

Wakefield, J., & Sebastien, H. (2008). Overcoming ecologic bias using the two-phase study design. *American Journal of Epidemiology*, *167*(8), 908–916.

Waller, L. A., Carlin, B. P., Xia, H., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, *92*(438), 607–617.

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data* (Vol. 368). John Wiley & Sons, New Jersey, U.S.A.

Walter, S. (2000). *Disease mapping: a historical perspective, in Spatial Epidemiology Methods and Application* (E. Paul, J. Wakefield, N. Best, & D. Briggs, Eds.). Oxford

University Press.

Wand, M. (2002). Vector differential calculus in statistics. *The American Statistician*, *56*(1), 55–62.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, *18*(2), 223–249.

Wansbeek, T. J., & Meijer, E. (2000). *Measurement error and latent variables in econometrics*. Elsevier, North-Holland, Amsterdam.

Weinberg, C. R., & Umbach, D. M. (2000). Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *American Journal of Epidemiology*, *152*(3), 197–203.

Weinberg, C. R., & Wacholder, S. (1990). The design and analysis of case-control studies with biased sampling. *Biometrics*, 963–975.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1–25.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, *115*(1), 119–128.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114.

Wood, S. (2006). *Generalized additive models: an introduction with R*. Florida, U.S.A.: Chapman and Hall/CRC.

Wood, S., Goude, Y., & Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *64*(1), 139–155.

Wood, S. N. (2012). On p-values for smooth components of an extended generalized additive model. *Biometrika*, ass048.

Xia, H., & Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: mapping ohio lung cancer mortality. *Statistics in Medicine*, *17*(18), 2025–2043.

Xun, X., Cao, J., Mallick, B. K., Maity, A., & Carroll, R. J. (2013). Parameter estimation of partial differential equation models. *Journal of the American Statistical Association*, *108*, 1009-1020.

Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, *97*(460), 1042-1054.

Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., & Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution:

concepts and consequences. *Environmental Health Perspectives*, *108*(5), 419-426.

Zhang, D., Lin, X., Raz, J., & Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, *93*(442), 710–719.

Zhang, L., Mukherjee, B., Ghosh, M., Gruber, S., & Moreno, V. (2008). Accounting for error due to misclassification of exposures in case–control studies of gene–environment interaction. *Statistics in Medicine*, *27*(15), 2756–2783.

Zheng, Y., & Zhu, J. (2012). On the asymptotics of maximum likelihood estimation for spatial linear models on a lattice. *Sankhya A*, *74*(1), 29–56.