ABSTRACT

KALENDRA, ERIC JAMES. Space-time Modeling of Health Effects while Controlling for Spatially Varying Exposure Surfaces. (Under the direction of Dr. Montserrat Fuentes.)

In fields such as environmental science and epidemiology, data are collected over space and time. In many cases, the spatial information in the data is aggregated. The aggregation reduces the information available for the analysis. However, the goal of the study remains the same, spatial risk assessment. In this thesis, we work with a dataset where the risk is varying relative to a constant exposure, race, and a dataset where we consider the risk to be constant relative to a varying exposure.

Racial and ethnic characteristics usually do not change over time, but environmental conditions may vary dramatically across space. In epidemiological studies, estimating the association between individual level covariates and adverse outcomes, the association is believed to be dependent on the environmental conditions. A common method to incorporate spatial information in the analysis is to use neighborhood level variables. They are usually considered sufficient to explain any spatial variation in risk estimates of an adverse birth outcome. We demonstrate that neighborhood level variables may not contain all relevant information to estimate how risk changes across space. By using a spatially-varying coefficients model for race, we show there is a significant amount of information available which may not be captured in models using only individual level covariates and neighborhood covariates.

In contrast, air pollution exposure changes in space and time. The difficulty in estimating the risk associated with air pollution exposure is that the health data are typically aggregated to some geographic unit. It is well known that this aggregation introduces ecological bias in the estimates. We demonstrate how the bias can affect risk estimates

at different levels of aggregation using a dataset in North Carolina for the years 2001 and 2002. Also, there has been extensive research to reduce the bias by relating the aggregated outcomes to a continuously estimated surface for the exposure. However, due to the additional uncertainty introduced, the bias correction framework usually significantly increases the standard error of the risk estimates. We take a different approach where, instead of using a continuously approximated surface, we estimate the pollution at geographic units which are smaller than the level at which the data are aggregated. The new geographic unit is chosen where we can estimate the exposure within the region without drastically increasing the uncertainty. We demonstrate, using the same North Carolina data, that our method reduces the bias induced through aggregation as if the data were actually aggregated to the smaller geographic units.

Space-time Modeling of Health Effects while Controlling for Spatially Varying Exposure Surfaces

by
Eric James Kalendra

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2010

APPROVED BY:

Dr. Montserrat Fuentes Chair of Advisory Committee	Brian J. Reich Co-chair of Advisory Committee
Dr. Huixia Wang	Dr. Jerry M. Davis

DEDICATION

To all of my family and friends.

BIOGRAPHY

Eric Kalendra was born in Florida in 1986. He entered the Statistics department at North Carolina State University in 2004 and earned his B.S. in 2006 in Statistics and Applied Mathematics. He received his M.S. in 2008 in Statistics. He received his PhD in 2010 in Statistics.

ACKNOWLEDGEMENTS

I would like to thank my advisor for her help and endless patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction and Overview

In fields such as environmental science and epidemiology, data are collected over space and time. For example, datasets collected daily include pollution and weather data across the United States, mortality data in North Carolina, and birth data in New York City. Typically, when data are collected on individuals, in addition to collecting the outcome of interest more information is collected at the same time. These additional pieces of information are referred to as individual specific variables. Information can be collected separately and linked to the data, such as data collected in a census. The preferred analysis is to use the individual specific information. While data are usually collected at the individual level, sometimes the data are aggregated to protect individual privacy at the cost of losing some individual covariates that are not taken into account when the aggregation is performed. Aggregating the data loses information as we lose the ability to identify specific individuals, and this loss of information is expected to affect regression estimates. In this thesis, we explore using multiple spatially varying individual specific variables instead of census based variables to analyze birth outcomes. We also investigate how aggregation effects bias in risk estimates of pollution on mortality.

In order to model spatial-temporal data, we first discuss the different types of data. In general, there are three basic types: point-referenced data, areal data, and point pattern data (Banerjee et al., 2004). *Point-referenced* data are data whose location $s \in \Re^d$ can vary continuously over a region, $D \subset \Re^d$ but is fixed. These data are often referred to as *geostatistical* data. For example, daily $PM_{2.5}$, particulate matter less than 2.5 microns in diameter, concentrations are measured at generally fixed monitoring locations across the United States. *Areal* data are data whose measurement referrers to a region D instead of a point. The regions can be regular or irregular in shape. Shapes are typically defined by political boundaries such as zip codes, counties, census tracts, etc. The data are also referred to as *lattice* data, and statistical models for these data usually incorporate neighborhood, adjacency, information of the areal units. For example, some numerical models provide output which represents the average of grid boxes. *Point pattern* data are data where the locations, s, are random. The events of interest, such as a preterm birth or an individual dying, occurs at random locations. In many instances, other information is also collected when the event occurs. For example, if we record a preterm birth we would also record information about the mother, such as her age and race. With such data, we are usually interested in studying the pattern that is exhibited. The usual patterns that are tested for include complete spatial randomness, clustering, and regularity. For more information, see the books by Gelfand et al. (2010), Banerjee et al. (2004), and Schabenberger and Gotway (2004) for many examples and additional discussion.

Time series data are classified into one of two types: continuous or discrete. When observations are made at every instant in time, we consider it to be *continuous* in time. Polygraphs and electrocardiograms are examples of continuous time series data. *Discrete* time series data are data where observations are made at distinct time intervals. Daily pollution measurements and ten year censuses are examples of discrete time series. Many times in practice, a continuous time series is approximated by a discrete time series.

In this thesis, we consider point-referenced data or areal data for space and discrete time series data. In this chapter, we provide some general concepts in statistical methodology for spatially varying regression in Section 1.1, and statistical models for the health effects of air pollution in Section 1.2.

1.1 Spatially varying regression

In many epidemiological models that study adverse health outcomes, it is common to allow the risk to vary across space. There are multiple different methods to allow for a spatially varying risk in an analysis (Fuentes et al., 2006). The most common methods are to include neighborhood level covariates and to use an autoregressive component.

Neighborhood level covariates are characteristics of the subject's location that describe the geographic location. The typical variables used to describe a location are taken from the United States ten year census. The census contains information about individuals living in a geographic unit and provides information about racial composition, income, family structure, ownership of property, and many others. It has been shown that the use of census neighborhood level variables can dramatically improve prediction accuracy as compared to studies that use only individual covariates (English et al., 2003; O'Campo et al., 1997). The assumption is that exposure to location is associated with the measured outcome. Including neighborhood level covariates in epidemiological models provides a basic spatially varying regression as the covariates themselves describe a spatial pattern. However, it is difficult to include multiple census based covariates because many of the variables are highly correlated.

One effective method for using neighborhood information is to use an index instead of the raw data. In epidemiological studies, an index is generally composed of proxy or surrogate measures rather than direct measurements. In general, the components of the index must be additive and the components should be weighted so that the component weights are proportional to the probability each group suffers from the index condition. A historical method for weighting index components is to calculate the weights for a specific index using logistic regression (Gordon, 1995). More recently, a neighborhood deprivation index is calculated using principle component analysis, PCA (Messer et al., 2006). See Jolliffe (2002) for more information on PCA. Both of these methods for creating indexes are data reduction techniques. Data reduction consolidates any number of variables into a smaller set of variables which describes the variation of the larger set. In the case of PCA, the components are orthogonal. Compared to the large number of variables that are used to construct the index, the index itself is much easier to interpret.

Using an index provides a structured spatial framework. A spatial surface constructed from an index must be a multiple of the index used. Using an index does not provide a flexible spatial model. Two more flexible models for areal data, such as census geographic units, are the simultaneous autoregressive model, SAR, and the conditional autoregressive model, CAR (Besag, 1974; Banerjee et al., 2004; Gelfand et al., 2010). Besag (1974) was the first to introduce the CAR model. The CAR model has become a convenient and flexible model for areal data because of the recent work in the context of Gibbs sampling and more general Markov Chain Monte Carlo (MCMC) methods. The difference between the CAR and the SAR is that for the CAR we let the data induce the error distribution, whereas for the SAR we let the errors induce a distribution for the data. The CAR model provides structure through the relationship of one geographic unit and its neighbors. The intrinsic CAR is a CAR model where the conditional distribution of a geographic unit is normal with mean equal to the average of its neighbors and variance equal to prior variance over the number of neighbors. For a more complete summary of CAR and SAR models, see Banerjee et al. (2004).

In Chapter 2, we explore the use of spatial regression to help explain the spatial pattern of adverse pregnancy outcomes in New York City. We compare using a neighborhood deprivation index and the more flexible CAR model to explain spatial variation. We show that the CAR model explains significantly more variability compared to the neighborhood deprivation index.

1.2 Statistical models for the health effects of air pollution

Particulate matter (PM) is a mixture of small particles found in ambient air. PM is classified by its size. The typical size categories currently used are PM_{10} and $PM_{2.5}$. PM_{10} includes all particles less or equal to 10μm in diameter, and $PM_{2.5}$ includes only particles less than 2.5μm in diameter. $PM_{2.5}$ is also known as fine PM, and the difference PM_{10} - $PM_{2.5}$ is known as coarse PM. It is believed that $PM_{2.5}$ is capable of causing a variety of serious health effects because of the ability of small particles to penetrate the respiratory tract. This belief led to the U.S. Environmental Protection Agency (EPA) setting a standard for $PM_{2.5}$ levels.

The health impact of air pollution has received much attention over the past few years. Numerous epidemiological studies investigating the association between exposure to PM and adverse human health outcomes have been conducted in a number of U.S. communities. Several studies have shown that PM exposure may result in tens of thousands of excess deaths per year and many more cases of illness in the U.S. population (Bates et al., 1990; Ostro et al., 1991; Dockery et al., 1992; Schwartz, 1994; Pope et al., 1995a; American Thoracic Society and Bascom, 1996a,b). Two types of studies have been used in the analysis of air pollution: time series studies and cohort studies. Time series studies focus on the association between short-term exposure to PM and adverse health outcomes, and the association between PM and acute health effects (Dockery et al., 1992; Schwartz, 1994; Dominici et al., 2000). Cohort studies are used to estimate the long-term association between exposure to PM and adverse health outcomes (Dockery et al., 1993; Pop et al., 1995b). Cohort studies suffer from lack of data because of the inherent difficulties in conducting these studies. In this thesis, we focus solely on time series studies.

Over the past 20 years, there have been many developments in the analyses for pollution data. Originally, time series studies were conducted at a single site (Bates et al., 1990; Schwartz, 1995). These single site studies demonstrated the association of exposure to air pollution with health effects. Conducting studies at a single location, prevents a standard to the analyses making it difficult to generalize the results to the whole population. To overcome this deficiency in single site studies, researchers move to using multi-site studies. Using much of the available data from the single site analyses, a common framework was developed and the analyses were carried out using a common structure (Burnett and Krewski, 1994; Katsouyanni et al., 1997; Dominici et al., 2000). Dominici et al. (2000) introduced the use of a two-stage linear regression model to estimate the association between exposure to PM_{10} and daily mortality for the 20 largest cities in the United States. Daniels et al. (2000) extended the previous study to estimate the PM_{10} mortality dose-response curve. Using the largest 90 U.S. cities, the National Morbidity, Mortality and Air Pollution Study (NMMAPS) (Samet et al., 2000a,b) is the largest multi-site study. The NMMAPS study was the first to estimate city-specific, regional, and national effects of PM_{10} on mortality. The NMMAPS database has since become popular for other health effect analyses. Dominici et al. (2002) develops a three-stage Bayesian hierarchical model and uses it to analyze the NMMAPS database. They use 88 cities in the database from 1987 to 1994 to obtain results showing a positive association between previous day, lag 1, PM_{10} concentrations and morality in most of the cities.

Studies using PM before 2000 concentrated on PM_{10} as $PM_{2.5}$ were only available at a very limited number of monitoring stations until 2001. Fuentes et al. (2006) showed spatial association between $PM_{2.5}$ and mortality in the entire United States. However, the data were aggregated in time to the month level. While covering a vast spatial region, the temporal aggregation is seen as a draw back in this case. More recently, Choi et al. (2009) introduces a spatial temporal model to estimate the risk between $PM_{2.5}$ and mortality. Their results continue to show an association between PM exposure and mortality.

An important issue when studying the association between ambient PM concentrations and daily mortality counts is spatial aggregation. Aggregating health data, utilizing data at the group level, in a ecological study leads to the introduction of *ecological bias*. For an in depth discussion of some of the potential biases, see Greeland (1992). There is a vast literature discussing ecological bias and ecological fallacy. See Greenland and Robins (1994) for a discussion or Wakefield (2003) for a more recent review of the topic. As the methodologies for analyzing the effect of PM on mortality have become more sophisticated and more data have become available, there has been a renewed interest in correcting for the ecological bias that arises when aggregating a nonlinear individual level model over the within-area distribution of covariates. Wakefield and Shaddick (2006) propose a model to correct for ecological bias. The model depends on a large number of exposure data within each aggregated region. Salway and Wakefield (2008) then extend the model to be used with a smaller number of exposure observations within each aggregated region. Lee and Shaddick (2010) further extend the bias correction framework to use a spatio-temporal model which takes into account both spatial and temporal correlations. All of these methods and extensions are common in that they try to correct for bias by relating back to pollution surface estimated at the individual level.

With interest at modeling the exposures at the individual level, epidemiological studies are integrating the use of exposure simulators. In the absence of actual personal exposures, it may be possible to use simulated exposures. There several exposure simulators available, some of which are Stochastic Human Exposure and Dose Simulation Models (SHEDS)(Burke et al., 2001), Air Pollutants Exposure Model (APEX) (Efroymson and Murphy, 2001), Probabilistic National Ambient Air Quality Standards (NAAQS) Exposure Model (pNEM) (Johnson et al., 1992), and an extension to pNEM (pCNEM) (Zidek et al., 2005). There are a number of significant differences among the simulators, but all the simulators attempt to estimate the cumulative exposure experienced by individuals. The cumulative exposure depends on what environments the individual passes through, such as a car, house, or street. Originally, the exposure simulators were developed for policy development. For example, APEX was used in the development of the ozone criterion (Ozone 2006). Shaddick et al. (2007) uses pCNEM to estimate the distribution for personal exposures for PM_{10}, and Reich et al. (2008) uses SHEDS to estimate the risk of exposure to ultra fine particles, diameter less than 0.1μm. The exposure distributions used in the analysis are highly variable and lead to risk estimates with dramatically larger standard errors. Sources of uncertainty in exposure simulators include penetration factors, air exchange rates, deposition rates, diaries used to predict human activity, and the actual ambient air pollution level. A primary concern in the use of the exposure simulators is that with the uncertainty inherent in the simulator, the exposure simulators may not accurately estimate the variability of individual exposures.

In Chapter 3, we demonstrate that aggregation can lead to significant ecological bias. Then, in Chapter 4, we extend previous bias adjustments to correct for ecological bias relative to a less aggregated level rather than by using a exposure surface.

Chapter 2

Spatially-varying Association between Race, Social Factors, and Birth Outcomes, New York, 1995-2003

2.1 Introduction

Preterm birth (PTB, gestational age less than 37 completed weeks) and low weight birth (LBW, less than 2,500 g) have been linked with infant mortality and morbidity in numerous studies (Callaghan et al., 2006; Kramer et al., 2000). This link has spurred a great amount of research into the risk factors of adverse pregnancy outcomes in order to understand and reduce infant morbidity and mortality.

Multiple individual-level risk factors have been linked with preterm and low weight birth, including stress (Hobel and Culhane, 2003), race/ethnic group (Mathews et al., 2004; Ma, 2008), and whether the mother is foreign-born. Whether foreign-born mothers have the same risk as native-born Americans has been studied extensively in recent years for first-generation Mexican-American women with local-level analyses (Collins and David, 2004; Leslie et al., 2005) and aggregated national studies including multiple other racial groups (Singh and Yu, 2005). Both of the locally-based studies have relatively small sample sizes, and the results appear to have high uncertainty. The analysis presented by Singh and Yu (2005) uses a large sample size but is pooled across the entire United States. When estimating the risk for foreign-born mothers, the risk is usually given relative to native-born mothers of the same race/ethnic group. This means the effects of race are taken into account when estimating the difference in risk for whether a mother was born in the United States.

Individual-level maternal risk factors are only able to explain a small proportion of the variation in LBW and PTB. By including neighborhood-level variables in the model, predictive accuracy can be greatly increased (English et al., 2003; O'Campo et al., 1997). The variables can be transformed into indexes such as education, employment, housing, occupation, poverty and residential stability (Messer et al, 2008), or they can be transformed to create an overall index such as a neighborhood deprivation index (NDI) (Messer et al, 2006). A NDI may be constructed using principal component analysis, which attempts to explain a region's relative social deprivation with a reduced number of variables, often one. While neighborhood-level variables are spatial variables and attempt to explain spatial patterns, these neighborhood-level variables do not capture all the spatial variation in the risk.

Flexible spatial models are introduced in order to capture effects of unmeasured confounders. The most common model for areal data is the conditional autoregressive (CAR) model (Besag, 1974; Banerjee et al., 2004). The CAR model has been used in a number of different epidemiological studies, cancer in particular, to explain the variation across space. The CAR model is typically used to adjust spatially the intercept, with predictor effects treated as non-spatial additive effects.

This study uses census tract geocoded birth data from New York City, focusing on the outcomes PTB and LBW. We propose a mixed effects probit regression model with spatially-varying race effects. Our model allows the effect for race to vary from census tract to census tract in order to account for unmeasured confounders that interact with both the birth outcome and racial distribution within the census tract. We use a CAR prior to smooth estimates across nearby census tracts. In essence, this prior compels neighboring tracts to borrow strength from each other to improve the estimates of the race effects in each census tract. Using a different CAR for each race allows us to evaluate how well a neighborhood-level variable such as NDI is able to characterize a race's spatial variation. Most importantly, the analysis demonstrates the need and importance of using spatially-varying race effects to control for unknown variables that have not been included in the model.

The concept of using multiple independent CAR priors in a model for LBW has been presented in Norton and Niu (2009), where independent intrinsic CAR priors are used to allow for spatiotemporal changes. They show that multiple independent CAR priors are identifiable and do not create any estimation difficulties. An intrinsic CAR implies the conditional distribution of the value in a location is centered around the average of the location's neighbors. The model has a single spatial effect for different time periods forcing every group to vary in the same way across space. Instead of using the CAR priors across time, we use the CAR priors across different ethnic groups, which Norton and Niu (2009) show the framework is estimable. The analysis is also performed with aggregated data, so only a limited number of individual covariates are used in the model.

We demonstrate the novel idea to allow a separate spatially-varying effect for each race while still using the individual level covariates such as the mothers age, education, and others.

The plan of the paper is as follows. In Section 2.2, we introduce the data used in the analysis. Section 2.3 describes the model. Section 2.4 provides the results. We conclude with a summary and a discussion of possible extensions in Section 2.5. Section 2.6 includes tables of the results and graphical representations of the spatial effects.

2.2 Data

We use New York City singleton live birth record data with linked Statewide Planning and Research Cooperative System, SPARCS, hospital discharge data for the years 1995 to 2003. Birth certificate data are collected, geocoded, and cleaned by the New York City Vital Statistics Department. The birth certificates are geocoded to the census tract where the mother lives. After collection, the New York State Department of Health links the birth certificate and the SPARCS hospital discharge data.

There are a total of 1,133,020 singleton live births in vital records, of which, 1,067,356 (94.2%) were successfully linked to the mother/infant SPARCS record. Matched and unmatched birth records were compared on several maternal and infant characteristics. Matched records were more likely than unmatched records to indicate white maternal race (58.1% vs. 52.9%) but were similar regarding infant sex and maternal education. Infants with matched birth records had higher mean birth weight than those with unmatched birth records (3289g vs. 3196g) as well as longer mean duration of gestation (38.9 weeks vs. 38.5 weeks).

After exclusions, there are 897,241 records remaining. Demographic statistics are presented by racial group in Table 2.6. Roughly 55.8% of births were from white mothers, 32.1% from black mothers, and 12% from mothers of other races. Because of the large Hispanic population in New York City, we consider Hispanic to be an exclusive category. After separating the Hispanic mothers, about 37% of the mothers were Hispanic, 25% were non-Hispanic white mothers, 26% were non-Hispanic black mothers, 11% were mothers of other races. Throughout the rest of the paper, the ethnic categories white and black will refer to non-Hispanic white and non-Hispanic black mothers respectively. About 54.8% of the mothers were foreign-born, 42.3% of the mothers were on a federal food supplement program (WIC), 3.4% of the mothers reported smoking during the pregnancy, and 51.3% of the children born were male. Education is reported as the number of years of schooling. We use education as a categorical variable by converting years into the usual levels of elementary school (including middle school), high school, college, and graduate school. About 7.9% of the mothers have elementary school education, 50.7% have high school education, 27.8% have college education, and 9.5% have graduate school education. The mother's age was treated as a categorical variable in order to allow a non-linear effect of age. About 2.6% of the mothers were less than 18 years old, 32.7% were between 18 and 25 years old, 26.7% were between 26 and 30 years old, 23.4% were between 31 and 35 years old, 11.9% were between 36 and 40 years old, and 2.7% were older than 41 years old. The outcomes of interest are PTB and LBW. LBW controls for gestational age, in weeks, through the use of a linear and a quadratic term.

The mothers were geocoded, by the New York Vital Statistics Department, to the census tract level. Census tract locations are determined by the 2009 TIGER/Line Shapefiles. These shape files are retrieved from the United States Census Bureau. These locations, plotted in Figure 2.1, include 2,217 tracts from the 5 NYC counties: Bronx, Kings, New York, Queens, and Richmond. Due to a recording error, some of the mothers were assigned non-unique census tract locations and were therefore excluded from the study, n=1,980. A small number, n=68, of the tracts have zero population in the 2000 census. The variable neighborhood deprivation index, NDI, is indicated in Figure 2.1. NDI was constructed from a number of census tract level variables using principal component analysis. Larger values of NDI indicate a census tract which is considered more deprived, meaning the residents of the census tract are expected to be living in worse conditions than someone living in a census tract with a smaller value of NDI. Since some tracts have no population, the components of NDI cannot be measured in these regions, and are therefore treated as missing.

2.3 Statistical Methods

Here we describe our approach for studying the spatial variation in racial disparity for birth outcomes. In Section 2.3.1, we introduce the model, and in Section 2.3.2, we describe the framework for fitting and inference.

2.3.1 Model Details

In this section, we provide the details of the proposed mixed effects probit regression model with spatially-varying race effects. The model is implemented using an independent CAR prior for each of the racial effects to allow each race effect to vary independently across space, while controlling for individual level covariates. Our model allows the effect for race to vary from census tract to census tract to account for unmeasured confounders. In essence, the CAR prior compels neighboring tracts to borrow strength from each other to improve the estimates of the race effects in each census tract.

Let $y_i \in \{0, 1\}$ indicate that the i^{th} baby had the outcome of interest, either preterm birth or low weight birth, and let s_i be the mother's census tract. The model is:

$$P\left(y_i = 1|\alpha, \beta\right) = \Phi\left[Z_i'\alpha + \beta_0(s_i) + \sum_{j=1}^{k=3} R_{ij}\beta_j(s_i)\right], \tag{2.1}$$

where Z_i is the vector of confounders described above (e.g., age, NDI, education, etc.), α are the corresponding fixed effects, the race covariate R_{ij} indicates whether the i^{th} mother is race j (white being the reference group with $j = 0$), and $\beta_j(s)$ the effect of race j in census tract s. k is an integer between 0 and the number of racial groups being compared to white mothers, which is 3 in this paper (black, other, and Hispanic). All covariates are the same between the two analyses (PTB and LBW) with the exception of gestational age (weeks at delivery) is given a linear and a quadratic term in the LBW model.

Assumptions that are placed on the $\beta_j(s)$'s determine the flexibility of the model. We can assume that the racial effects are constant across space, $\beta_j(s) = \beta_j$, which leads to a standard fixed effects model. Because we want to model the racial effects varying across space, the fixed effects model is too restrictive.

By relaxing the assumptions and only assuming that racial effects are independent, sum to zero, and are normally distributed with a common variance, we have a basic random effects model with a separate random effect for each race. The random effects model ignores the spatial location of the census tracts. Even if two mothers are in neighboring tracts, the simple random effects model would still treat their effects are independent of one another, which does not follow how we believe the effect of race should vary across space. We generally assume that two census tracts close together have more similar characteristics than two census tracts far apart, and we would like to use this knowledge in estimation.

To accomplish this goal, we introduce spatial random effects, modeled using conditionally autoregressive (CAR) priors (Banerjee et al., 2004). The CAR model can be defined through its conditional distribution for $\beta_j(s)$ given β_j at all census tracts other than tract s, denoted $\beta_j(-s)$. The conditional distribution is:

$$\beta_j(s)|\beta_j(-s) \sim N(\phi_j \beta'_j(s), \sigma_j^2/m_s) \tag{2.2}$$

where $0 < \phi < 1$ controls the strength of spatial association for race $j = 0, ..., k$, $\beta'_j(s) = \Sigma_{t \neq s} w_{st} \beta_j(t)$ is a weighted average of the race effects for all tracts other than tract s, $m_s = \Sigma_{t \neq s} w_{st}$, and σ^2 controls the variance. We assume the weights w_{st} are equal to 1 if two tracts are neighbors and 0 if the two tracts are not neighbors. In this case, $\beta'_j(s)$ is simply the average of $\beta_j(s)$'s neighbors. First-order neighbors are defined in the usual sense, where neighbors of s are all regions which share either a vertex or an edge with s. The shape of the census tract comes from the 2009 TIGER/Line Shapefiles provided by the United States Census Bureau. We compare the first order model with the second order model that also includes second order neighbors, defined to be all of the first order neighbors and all of the first order neighbors of the first order neighbors excluding s. The CAR model allows the coefficient to vary spatially while using spatial structure to increase the available information at a given location.

2.3.2 Fitting the model

To complete the specification of the Bayesian model, priors are needed for α, ϕ_j, and σ_j^2. The value of ϕ_j is set equal to 1, making the CAR model the intrinsic CAR (Banerjee et al., 2004). This implies the conditional distribution the coefficient corresponding to a location is centered around the average coefficients of the location's neighbors. The other priors are given by

$$\alpha \sim N_p(0, \frac{1}{\sigma_\alpha^2} I_p), \tag{2.3}$$

$$\sigma_j^2 \sim \text{Inverse Gamma}(a, b), \tag{2.4}$$

with $\sigma_\alpha^2 = 100$ to give relatively uninformative priors for the fixed effects, $a = 0.5$, and $b = 0.005$ following Kelsall and Wakefield (1999). This design has a prior mean variance for $\sigma_j^2 = 100$, which provides very little information to the posterior distribution.

The model comparison between the spatial random effects and the fixed effects model is carried out using the deviance information criterion, DIC, as given in Spiegelhalter et al. (2002). DIC is a penalized deviance criteria, where the penalty is the effective number of parameters, p_D, relative to the average deviance value $\bar{D}$, where deviance is -2 times the log likelihood. DIC is calculated as DIC $= p_D + \bar{D}$. The model with the smallest DIC value is considered to be the model with the best fit. In the spatial setting, DIC is preferable to AIC or BIC because DIC penalizes based on the effective number of parameters not the total number of estimated parameters as in AIC and BIC. In a spatial model which uses a CAR prior, the effective number of parameters is almost always smaller than the total number of estimated parameters because of the correlation structure defined by the CAR prior.

In order to maintain the comparable likelihood values for both the fixed effects model and the spatially-varying model, the likelihood is calculated based on (2.1), the highest level of the hierarchical model. The likelihood used in the calculations for the DIC is

$$\prod_{i=1}^{n} [P(y_i = 1)]^{y_i} [P(y_i = 0)]^{1-y_i} \tag{2.5}$$

where $P(y_i = 1)$ is in (2.1).

2.4 Results

The probit scale is not as easy to interpret as probabilities, so the regression coefficients are presented in the probability scale. For example, taking as the referent covariate pattern a native-born white mother, age 26-30, with high school education and who is not a WIC recipient, with an address in tract s, the probability of having a PTB is

$$P(Y_s = 1) = \Phi\left[\alpha_0 + \alpha_{NDI(s)} NDI(s) + \beta_0(s)\right]. \tag{2.6}$$

This probability is dependent on the tract, and we present a one number summary which averages across space, represented as

$$P(Y = 1) = \frac{1}{|S|} \sum_{s \in S} \Phi\left[\alpha_0 + \alpha_{NDI(s)} NDI(s) + \beta_0(s)\right], \tag{2.7}$$

where S is the collection of all tracts, and $|S|$ is the number of tracts in the collection. Because the baseline probability of an outcome is different for each race, the effect of non-racial covariates such as education and age will not be the same for every race. However, all the outcomes are relatively rare, and the differences in effects will be small. For example, the education effect will be approximately equal for white and black mothers because the baseline probabilities are both small.

$$P(Y = 1|W, College) - P(Y = 1|W, HS) \approx P(Y = 1|B, College) - P(Y = 1|B, HS) \tag{2.8}$$

Because the results are nearly identical for all groups, effect estimates are given assuming the baseline, native-born white mothers, and the effects for the other racial groups would be slightly smaller, since every other category has a greater baseline rate for both PTB and LBW.

The results in the tables and plots are given in percentages. The results are in Table 2.2 and Table 2.3 for PTB and LBW respectively. For PTB, an example interpretation for the covariate smoking in Model 1 (the spatial model) would be that the estimated risk increases by 3.43% with 95% credible interval (3.20%, 3.67%) for mothers who smoke. This implies that a white mother, otherwise in the referent categories, who smokes during pregnancy has an estimated 9.80% chance of having a PTB (6.37% + 3.43%).

The results of the spatial and non-spatial analyses are very similar for the estimates of the fixed effects for both PTB and LBW. However, there is a change in the estimated effect and in the interpretation of the race effects. For the non-spatial analyses, the racial effects are the average effect averaged over all mothers. In the spatial analysis, the racial effects are the average effect averaged across space. While very similar, the spatial analysis gives equal weighting to space rather than mothers. The estimates would only be expected to be equal in the case in which the mothers are spread evenly across space.

To show that the estimates for the fixed effects (age, education, smoking status, and if the child is male) are similar between the spatial and non-spatial models we compare some of the 95% credible intervals. For PTB, we see that the estimated change for mothers who are less than 18 years old when compared to the referent category of mothers who are 26 to 30 years old is 1.49% (1.24%, 1.77%) in the spatial analysis and 1.48% (1.25%, 1.73%) in the non-spatial analysis. We also see overlapping intervals for the education variables such as College which is (-0.73%, -0.58%) for the spatial analysis versus (-0.58%, -0.44%) for the non-spatial analysis (model 2). The same holds true for any of the non-spatial models 2-4 and for LBW when comparing the mothers age, education, smoking status, and if the child is male. For these fixed effect variables we will only look into the estimated effects based on the spatial model, as the estimated effects will be very similar to the non-spatial models. Also, all of the non-spatial models have generally overlapping or close to overlapping credible intervals when comparing effects between them, so in this paper when referring to credible intervals for a non-spatial model, we will always be referring to model 2.

The results for PTB and LBW are very similar. In both analyses, age, education, and smoking have the same general estimated impact. There are minor differences between PTB and LBW as to which group has the smallest risk, but in general we see the same overall pattern.

Very young and very old mothers are estimated to have the greatest chances for PTB and LBW. The referent group for the age is 26 to 30 years old. The mothers with the smallest estimated risk are in the age range 18 to 25 years old for PTB and in the age range 26 to 35 for LBW. In both PTB and LBW, we see the largest increase in estimated risk in both the younger and older mothers. For PTB, mothers largest increase, relative to mothers who are 26 to 30 years old, are mothers who are over 40, and they are estimated to have about a 3.76% (3.45%, 4.09%) increase in risk. Whereas, for LBW we see comparable increases in risk for both the younger and older mothers. The mothers less than 18 years old are estimated to have a 1.01% (0.84%, 1.20%) increase in risk and mothers over 40 are estimated to have a 0.77% (0.59%, 0.96%) increase in risk.

Mothers who have a high school education level have the worst outcomes of any education level. This is seen with all other levels of education being protective. PTB shows strong ordering with non-overlapping credible intervals: Graduate -1.19% (-1.33%, -1.05%), College -0.66% (-0.73%, -0.58%), Elementary -0.34% (-0.47%, -0.19%), and High school (control group). This is in contrast to ordering of moderate strength for LWB which has overlapping credible intervals in the general order: College -0.36% (-0.41%, -0.30%), Graduate -0.30% (-0.39%, -0.21%), Elementary -0.17% (-0.26%, -0.07%), and

High school (control group).

The last of the fixed effect variables follow the expected trend. Comparing back to all referent groups, smoking increases the estimated risk of both PTB by 3.43% (3.20%, 3.67%) and LWB by 2.17% (1.97%, 2.35%). If the child is male, there is an increase in the estimated risk for PTB of 0.39% (0.32%, 0.47%) and a decrease in the estimated risk for LWB of -0.85% (-0.88%, -0.81%).

Because of the difference in the estimation of the racial effects between the spatial and non-spatial models, the estimates of the variables that control for race also have slightly different interpretations. The difference is typically not very large, and the direction of the effect is always the same.

Foreign-born mothers, generally having lower estimated risk. We present the estimates and credible intervals for the spatial model, and estimates for the non-spatial models are very similar. White foreign born mothers which are estimated lower risk in the spatial model than the non-spatial model. In the spatial model, we see an reduction of risk estimated to be -0.92% (-1.14%, -0.67%) for PTB and -0.33% (-0.47, -0.17%) for LBW. In contrast, foreign-born mothers in other racial groups are estimated to have more similar risks in the spatial model than in the non-spatial model, but it is generally estimated to be smaller once the spatial effects are added to the model. This shows a condensing of the estimated effects after spatial effects are introduced in the model. In comparison to white foreign-born mothers, we see the estimated effects decrease for black foreign-born mothers of -0.99% (-1.34%, -0.71%) for PTB and -0.44% (-0.63%, -0.25%) for LBW. The estimated reduction of risk for other mothers is estimated to be -0.89% (-1.77%, -0.25%) for PTB and -0.36% (-0.83%, 0.67%) for LBW, and Hispanic mothers have estimated risks of -0.65% (-1.00%, -0.36%) for PTB and -0.05% (-0.28%, 0.13%) for LBW.

In contrast to the pattern observed with the effect for foreign-born mothers, we see a decrease in the significance in the estimated effect for mothers who receive WIC support. We see the effect of WIC decrease in significance for the control group after the spatial effects are added in the model [PTB 0.13% (-0.17%, 0.45%); 0.27 % LBW (0.07, 0.51%)] when compared to the non-spatial models. With a large shift for the control group, we see shifts in the estimates for the other groups towards zero, suggesting all ethnic groups are becoming more similar in their estimated effects. The estimated effects for PTB in the spatial analysis are: black mothers -1.25% (-1.64%, -0.90%), other mothers -0.25% (-0.76%, 0.22%), and Hispanic mothers -1.19% (-1.55%, -0.85%). The effects for LBW generally move towards zero in the spatial setting, more specifically the estimated intervals are: black mothers -0.26% (-0.52%, 0.00%), other mothers 0.12% (-0.21%, 0.44%), and Hispanic mothers -0.39% (-0.67%, -0.15%). A general decrease in the magnitude of the estimates is seen in spatial models when compared to the non-spatial models.

Lastly, NDI is a function of location. After adding spatial effects, we have two sets of variables that change across space. NDI which changes across space by construction and the effect can be scaled relative to the constructed format, and the flexible racial effects which vary smoothly across space through the CAR model. The credible intervals are presented to show the difference in the estimates between the spatial and non-spatial analyses. Adding the spatial effects significantly changes the effects for NDI. We see the effect of NDI increase in significance for the control group after the spatial effects are added in the model [PTB 1.09% (0.87%, 1.31%); LBW 0.39% (0.20, 0.53%)] when compared to the non-spatial models [PTB 0.16% (0.08%, 0.25%); LBW 0.19% (0.12%, 0.27%)]. With a large shift for the control group, we see shifts in the estimates for the other groups. The estimated effects for PTB are: black mothers [Spatial -0.31% (-0.63%, 0.06%); non-spatial 0.34% (0.19, 0.49%)], other mothers [Spatial -0.58% (-0.97%,

-0.05%); non-spatial -0.08% (-0.30, 0.14%)], and Hispanic mothers [Spatial -0.65% (-0.55%, -0.02%); non-spatial 0.40% (0.26, 0.53%)]. The credible intervals for black and Hispanic mothers do not overlap between the spatial and non-spatial analyses. The effects for LBW generally contain zero in the spatial setting, more specifically the estimated intervals are: black mothers [Spatial -0.08% (-0.28%, 0.11%); non-spatial 0.07% (-0.04, 0.18%)], other mothers [Spatial -0.20%(-0.41%, 0.02%); non-spatial -0.31% (-0.44, -0.18%)], and Hispanic mothers [Spatial -0.15% (-0.27%, -0.01%); non-spatial -0.01% (-0.10, 0.08%)]. The addition of the spatial effects has a dramatic impact on the estimated risks related to NDI.

As mentioned previously, the interpretation of the effects for race is different for the non-spatial and spatial analyses. As the estimates presented in the tables are averaged across space, we now look at the spatial patterns present in the results. The spatial plots show the probabilities at the spatial level for each racial group with NDI included. Because NDI is spatially dependent in the way it is constructed, it is included in the spatial plots in order to provide the overall spatial effect for each racial group. Figure 2.1 presents the spatial pattern for NDI. Results are presented for PTB for white mothers, Figures 2.2 and 2.3, black mothers, Figures 2.4 and 2.5, and the estimated difference between the effect for black mothers and white mothers, Figures 2.6 and 2.7. In both Figures 2.2 and 2.4 show the percent chance a mother in a given tract will have a PTB, and Figures 2.3 and 2.5 show the standard error for the estimates. We see that black mothers have an estimated increase risk across space, and they also have larger standard errors than the white mothers. The difference in Figure 2.6 shows the difference between black mothers and white mothers in percent. Figure 2.7 shows the estimated difference divided by the estimated standard error (z-score). Spatial variation in the values of the z-scores is considerable, with the largest values (> 3) generally being observed in central

Kings, New York, and Richmond counties. There is also a section where the z-scores are in the range of 0 to 1 in central Bronx county and 1 to 2 in eastern Queens county. The difference presented is different than the pattern presented in Figure 2.1 for NDI. While the plots are not displayed here, we see a similar pattern for Hispanic mothers which also have large z-values in central Kings, New York, and Richmond counties, but the value of the z-values are slightly smaller in the 2 to 4 range. For other mothers, we see the same general pattern again, but the z-values are smaller in comparison. The largest z-values are in Richmond county with values in the 2 to 4 range, and the z values in central Kings and New York counties are in the 1 to 2.5 range. There are also some z-values which range into the negative values in the Bronx and eastern Kings counties.

The pattern for LBW is different from the pattern seen for PTB. In general, the z-values are very close to zero with very few values greater than 2. There is clustering of values near zero. For black mothers, these pockets are in the range of 0 to 1 and are in western Bronx, northern New York, and Kings counties. The same pattern is true for Hispanic mothers except the z-values are in the range of -0.5 to 1. The pattern also applies to other mothers with the pockets having z-values -1 to 0.5. We do not see any regions where the z-values are large in the LBW analysis.

To evaluate model preference, DIC is used. The values of DIC are presented in Tables 2.4 and 2.5 for PTB and LBW outcomes respectively. The difference in DIC is large between the spatial and non-spatial models. This is after penalizing for the large number of spatial variables that have been added to the model. Based on DIC, the spatial model is the best fitting model.

An extra model is included for LBW which only includes full term births. These results are seen in Table 2.3. Model 0 is the same as Model 1 except that PTB were excluded from the analysis. The estimated risks are very similar between the two analyses, suggesting that the results are not strongly impacted by PTB in the analysis. Model 0 has not been included in the DIC results because of the change in the number of observations makes comparison using likelihood unequal between the two models. Because the results are very similar, we elected to present the results for the analysis which used all available data.

The results were tested against different priors to evaluate sensitivity. In the spatial model the values of a and b from (2.4) where varied, and no significant difference in the results was observed. A second order neighborhood structure was also used in the spatial analysis without changing the results substantially.

2.5 Discussion

In this paper, we present a model which allows for spatially-varying race effects to account for unmeasured confounders that interact with the birth outcomes, preterm birth and low weight birth. Based on DIC, the spatial model captures more variation than any of the non-spatial models. This suggests that while NDI is capable of capturing some spatial variation, the structure imposed by NDI is not flexible enough to characterize the change in risk across space.

When switching from the non-spatial to spatial model, some estimates were altered. The variables which are invariant to the model used are age, education, smoking, and if the child is male. The other covariates which seem to have some spatial dependence include the race effects, foreign nativity, NDI, and WIC. We see a shrinkage in the effect for WIC, and the effect of foreign born mothers and NDI increased in the spatial model. While the interpretation of the racial effects in the spatial model is different, we can see the effect across space, which is of more interest than just knowing the population average effect.

The spatial race effects provide more insight into the spatial variation than the NDI variable alone. Just using NDI, we would believe that there would be pockets of individuals at an increased risk in only two locations, Bronx and Kings counties. With spatial effects, we see that the areas of increased risk are not localized to just Bronx and Kings counties. It is especially evident for the black PTB effect where there is an increase in risk for New York county. This shows that more variation is characterized through the use of the flexible spatial effects. Because there is a separate spatial structure for each of the race effects, we can see how the spatial structure changes between groups. This is evident comparing the black and Hispanic effects. The estimated black effect does not have regions of small z-values in Queens county, whereas the estimated Hispanic and other effects show regions in Queens county that are clusters of z-values that are close to zero. This shows it is much more informative to use a flexible spatial structure. We also see for LBW that rather than clustering of regions with larger significance, the estimated spatial pattern is clusters of regions where the z-values are close to zero, suggesting regions where there are not any significant racial differences.

While outside the scope of this paper, the model can be further expanded. So far, we have been using probit regression with indicator variables for preterm birth and low weight birth with independent analyses. Birth weight and gestational age at delivery are more naturally viewed as a continuous bivariate response, which can be flexibly characterized using a mixture distribution. Using a bivariate response model would also take into account the interaction between birth weight and gestational age by automatically allowing the distribution of birth weight to shift with increasing gestational age. Also, by avoiding dichotomization, one can assess covariate effects on risk of premature delivery and early premature delivery within one coherent model.

2.6 Tables and Figures

Table 2.1: Summary of the data used in the analysis. The counts are of individuals used in the analysis. Percentages given are based on columns.

	White	Black	Other	Hispanic	Total
PTB	11,649	25,487	5,727	27,075	69,938
	(5.2%)	(10.7%)	(5.9%)	(8.1%)	(7.8%)
Regular	214,121	212,131	92,117	308,934	827,303
Gestational Age	(94.8%)	(89.3%)	(94.1%)	(91.9%)	(92.2%)
LBW	9,461	23,580	5,472	22,285	60,798
	(4.2%)	(9.9%)	(5.6%)	(6.6%)	(6.8%)
Regular	216,309	214,038	92,372	313,724	836,443
Weight	(95.8%)	(90.1%)	(94.4%)	(93.4%)	(93.2%)
WIC	33,203	120,307	27,834	198,284	379,628
	(14.7%)	(50.6%)	(28.4%)	(59%)	(42.3%)
Not on WIC	192,567	117,311	70,010	137,725	517,613
	(85.3%)	(49.4%)	(71.6%)	(41%)	(57.7%)
< 18 Years old	1,191	8,716	265	13,030	23,202
	(0.5%)	(3.7%)	(0.3%)	(3.9%)	(2.6%)
18-25	48,988	86,004	20,103	138,615	293,710
	(21.7%)	(36.2%)	(20.5%)	(41.3%)	(32.7%)
26-30	59,886	58,928	33,084	87,461	239,359
	(26.5%)	(24.8%)	(33.8%)	(26%)	(26.7%)
31-35	67,581	50,485	29,297	62,567	209,930
	(29.9%)	(21.2%)	(29.9%)	(18.6%)	(23.4%)
36-40	38,030	27,465	12,646	28,570	106,711
	(16.8%)	(11.6%)	(12.9%)	(8.5%)	(11.9%)
>40	10,094	6,020	2,449	5,766	24,329
	(4.5%)	(2.5%)	(2.5%)	(1.7%)	(2.7%)
Elementary	4,094	8,199	9,246	49,161	70,700
	(1.8%)	(3.5%)	(9.4%)	(14.6%)	(7.9%)
High School	89,880	139,177	46,434	198,774	474,265
	(39.8%)	(58.6%)	(47.5%)	(59.2%)	(52.9%)
College	83,903	77,834	29,627	74,706	266,070
	(37.2%)	(32.8%)	(30.3%)	(22.2%)	(29.7%)
Grad School	47,893	12,408	12,537	13,368	86,206
	(21.2%)	(5.2%)	(12.8%)	(4%)	(9.6%)
Smoking	7,312	13,059	404	9,340	30,115
	(3.2%)	(5.5%)	(0.4%)	(2.8%)	(3.4%)
Not Smoking	218,458	224,559	97,440	326,669	867,126
	(96.8%)	(94.5%)	(99.6%)	(97.2%)	(96.6%)
Male Child	116,529	121,076	51,006	171,507	460,118
	(51.6%)	(51%)	(52.1%)	(51%)	(51.3%)
Female Child	109,241	116,542	46,838	164,502	437,123
	(48.4%)	(49%)	(47.9%)	(49%)	(48.7%)
Native	156,233	137,644	4,904	106,435	405,216
	(69.2%)	(57.9%)	(5%)	(31.7%)	(45.2%)
Foreign-Born	69,537	99,974	92,940	229,574	492,025
	(30.8%)	(42.1%)	(95%)	(68.3%)	(54.8%)
	225,770	237,618	97,844	336,009	897,241

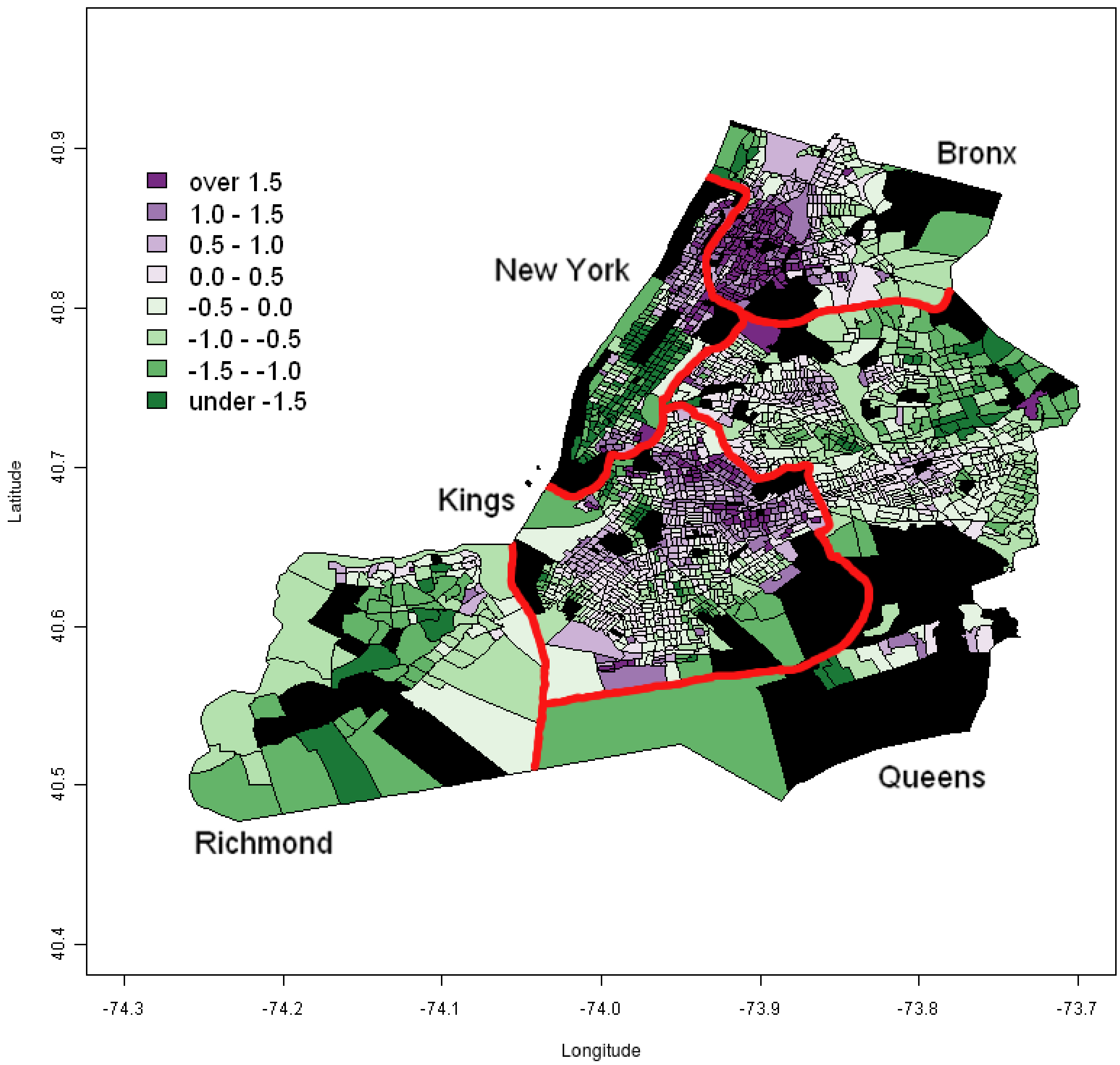

Figure 2.1: Neighborhood deprivation index, NDI, across space at the census tract level, which follows an approximate N(0,1) distribution. The black regions have missing values for NDI.

Table 2.2: Table for PTB effect on the probability scale as given in (2.7) and (2.8) times 100 to be in the unit percent (0-100). The age, education, smoking, and male child categories are increases in probabilities when compared to native-born white mothers not on WIC with a NDI value of 0. The probabilities for other races will be slightly different due to the use of the probit model.

	Model 1		Model 2		Model 3		Model 4	
Spatial	Yes		No, $\beta(s) = \beta$		No		No	
NDI	Yes		Yes		No		Yes	
WIC	Yes		Yes		Yes		No	
	Est	SE	Est	SE	Est	SE	Est	SE
Intercept	6.37	0.21	4.76	0.08	4.72	0.08	4.83	0.08
$<$ 18 years old	1.49	0.16	1.48	0.15	1.51	0.14	1.45	0.15
18 - 25	-0.32	0.06	-0.26	0.05	-0.24	0.05	-0.30	0.05
26 - 30								
31 - 35	0.88	0.07	0.84	0.06	0.81	0.06	0.87	0.06
36 - 40	2.19	0.09	2.04	0.08	2.00	0.08	2.08	0.09
$>$ 40	3.76	0.20	3.43	0.16	3.37	0.16	3.52	0.17
Elementary	-0.34	0.08	-0.35	0.07	-0.35	0.07	-0.40	0.08
High School								
College	-0.66	0.05	-0.51	0.05	-0.57	0.04	-0.45	0.05
Grad School	-1.19	0.08	-0.99	0.08	-1.08	0.07	-0.95	0.07
Smoking	3.43	0.15	3.37	0.14	3.44	0.14	3.35	0.13
Male Child	0.39	0.04	0.36	0.04	0.37	0.04	0.37	0.04
White								
Black	3.80	0.27	6.09	0.14	6.50	0.13	5.53	0.12
Other	1.58	0.56	2.81	0.36	2.88	0.37	2.71	0.39
Hispanic	2.62	0.25	4.15	0.13	4.53	0.11	3.69	0.12
Mforeign	-0.92	0.16	-0.41	0.10	-0.37	0.10	-0.37	0.09
Mforeign * Black	-0.99	0.19	-1.69	0.16	-1.91	0.16	-1.73	0.16
Mforeign * Other	-0.89	0.50	-1.93	0.38	-1.95	0.39	-1.91	0.41
Mforeign * Hispanic	-0.65	0.20	-1.40	0.13	-1.53	0.14	-1.50	0.15
NDI	1.09	0.29	0.16	0.11			0.25	0.11
NDI * Black	-0.31	0.21	0.34	0.09			0.12	0.09
NDI * Other	-0.58	0.27	-0.08	0.13			-0.12	0.13
NDI * Hispanic	-0.28	0.16	0.40	0.08			*0.13	0.07
WIC	0.13	0.19	0.46	0.14	0.55	0.14		
WIC * Black	-1.25	0.22	-1.59	0.19	-1.60	0.18		
WIC * Other	-0.25	0.30	-0.23	0.26	-0.31	0.26		
WIC * Hispanic	-1.19	0.21	-1.49	0.17	-1.38	0.16		

Table 2.3: Table for LBW effect on the probability scale as given in (2.7) and (2.8) times 100 to be in the unit percent (0-100). The age, education, smoking, and male child categories are increases in probabilities when compared to native-born white mothers not on WIC with a NDI value of 0. The probabilities for other races will be slightly different due to the use of the probit model. Model 0 is fit using only full term births.

	Model 0		Model 1		Model 2		Model 3		Model 4	
Preterm Births	No		Yes		Yes		Yes		Yes	
Spatial	Yes		Yes		No, $\beta(s) = \beta$		No		No	
NDI	Yes		Yes		Yes		No		Yes	
WIC	Yes		Yes		Yes		Yes		No	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Intercept	2.92	0.21	2.88	0.16	2.40	0.06	2.32	0.06	2.50	0.06
< 18 years old	1.13	0.14	1.01	0.11	1.14	0.11	1.11	0.11	1.16	0.12
18 - 25	0.27	0.06	0.33	0.05	0.40	0.04	0.41	0.04	0.42	0.04
26 - 30										
31 - 35	0.13	0.06	-0.01	0.05	0.02	0.04	0.02	0.04	0.02	0.04
36 - 40	0.44	0.07	0.18	0.06	0.24	0.05	0.22	0.05	0.24	0.06
> 40	1.13	0.14	0.77	0.11	0.88	0.12	0.84	0.11	0.90	0.13
Elementary	-0.07	0.06	-0.17	0.05	-0.24	0.06	-0.24	0.05	-0.26	0.06
High School										
College	-0.41	0.04	-0.36	0.03	-0.34	0.04	-0.36	0.04	-0.36	0.04
Grad School	-.044	0.07	-0.30	0.05	-0.25	0.05	-0.31	0.06	-0.29	0.06
Smoking	2.58	0.14	2.17	0.12	2.21	0.12	2.20	0.11	2.30	0.12
Male Child	-0.79	0.02	-0.85	0.02	-0.85	0.03	-0.83	0.03	-0.88	0.03
White										
Black	1.39	0.26	1.22	0.17	1.79	0.09	2.07	0.09	1.70	0.08
Other	0.24	0.19	0.50	0.17	0.48	0.24	0.67	0.22	0.54	0.26
Hispanic	0.52	0.17	0.25	0.19	0.74	0.08	0.92	0.07	0.60	0.07
Mforeign	-0.41	0.13	-0.33	0.11	-0.09	0.07	-0.08	0.08	-0.06	0.07
Mforeign * Black	-0.54	0.17	-0.44	0.11	-0.77	0.11	-0.89	0.11	-0.81	0.11
Mforeign * Other	0.83	0.19	-0.36	0.20	0.27	0.25	0.18	0.24	0.29	0.26
Mforeign * Hispanic	-0.06	0.14	-0.05	0.13	-0.37	0.09	-0.41	0.09	-0.40	0.09
NDI	0.51	0.31	0.39	0.22	0.19	0.09			0.25	0.09
NDI * Black	-0.06	0.12	-0.08	0.12	0.07	0.07			0.01	0.07
NDI * Other	-0.35	0.17	-0.20	0.13	-0.31	0.08			-0.30	0.08
NDI * Hispanic	-0.21	0.13	-0.15	0.08	-0.01	0.06			-0.10	0.06
WIC	0.18	0.17	0.27	0.15	0.47	0.11	0.61	0.10		
WIC * Black	-0.28	0.20	-0.26	0.16	-0.48	0.13	-0.56	0.13		
WIC * Other	0.14	0.21	0.12	0.20	0.20	0.19	0.01	0.17		
WIC * Hispanic	-0.48	0.15	-0.39	0.15	-0.57	0.13	-0.64	0.12		

Table 2.4: Table of DIC for PTB analyses.

Model	Spatial	NDI	WIC	M. For. Born	DIC	p_D	$\bar{D}$
Model 1	Yes	Yes	Yes	Yes	481393.6	1649.0	479744.7
Model 1a	Yes	No	Yes	Yes	481419.7	1680.4	479739.3
Model 1b	Intercept	No	Yes	Yes	481498.2	1205.9	480292.3
Model 2	No	Yes	Yes	Yes	481745.3	25.7	481719.5
Model 3	No	No	Yes	Yes	481860.4	21.6	481838.8
Model 4	No	Yes	No	Yes	481920.6	21.8	481898.8
Model 5	No	No	No	Yes	482001.1	17.7	481983.4
Model 6	No	No	No	No	482720.8	13.7	482707.1

Table 2.5: Table of DIC for LBW analyses.

Model	Spatial	NDI	WIC	M. For. Born	DIC	p_D	$\bar{D}$
Model 1	Yes	Yes	Yes	Yes	265330.7	1376.4	263954.3
Model 2	No	Yes	Yes	Yes	267637.9	28.0	267609.8
Model 3	No	No	Yes	Yes	267706.0	23.6	267682.5
Model 4	No	Yes	No	Yes	267680.8	24.1	267656.7
Model 5	No	No	No	Yes	267757.9	19.0	267738.9
Model 6	No	No	No	No	267944.8	15.7	267929.1

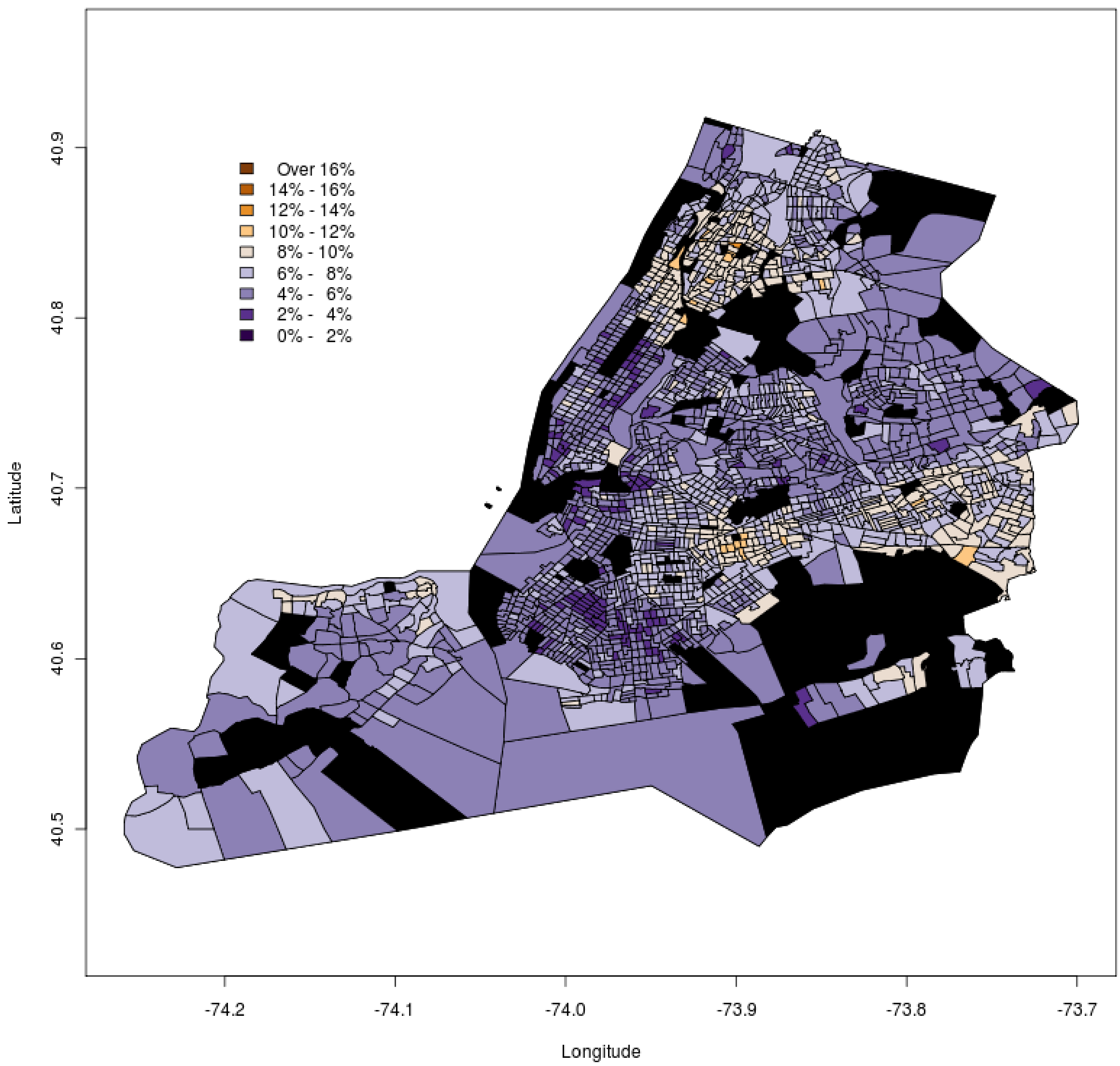

Figure 2.2: The percent chance a white mother has a PTB across space.

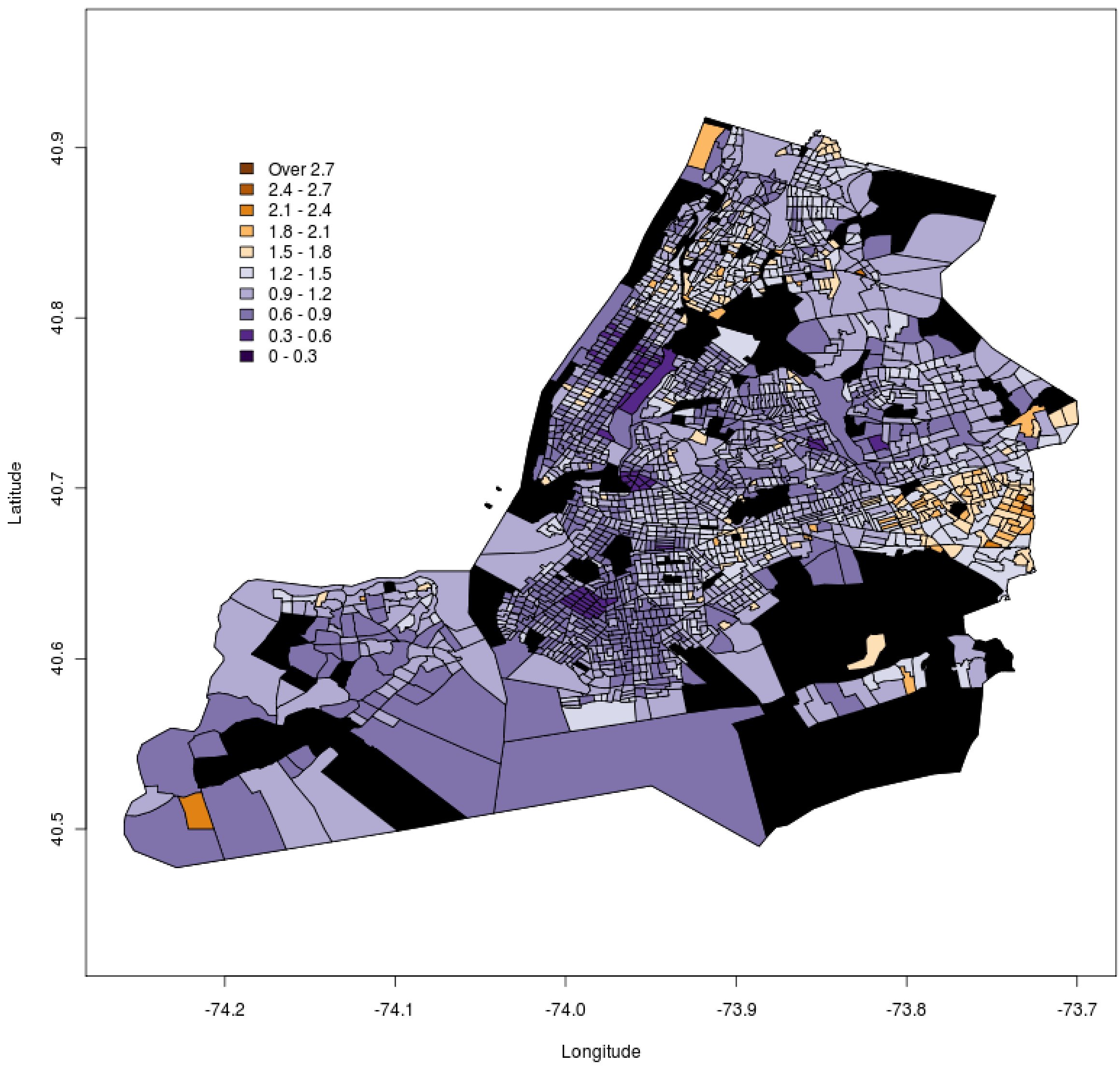

Figure 2.3: The standard error for the percent chance a white mother has a PTB across space.

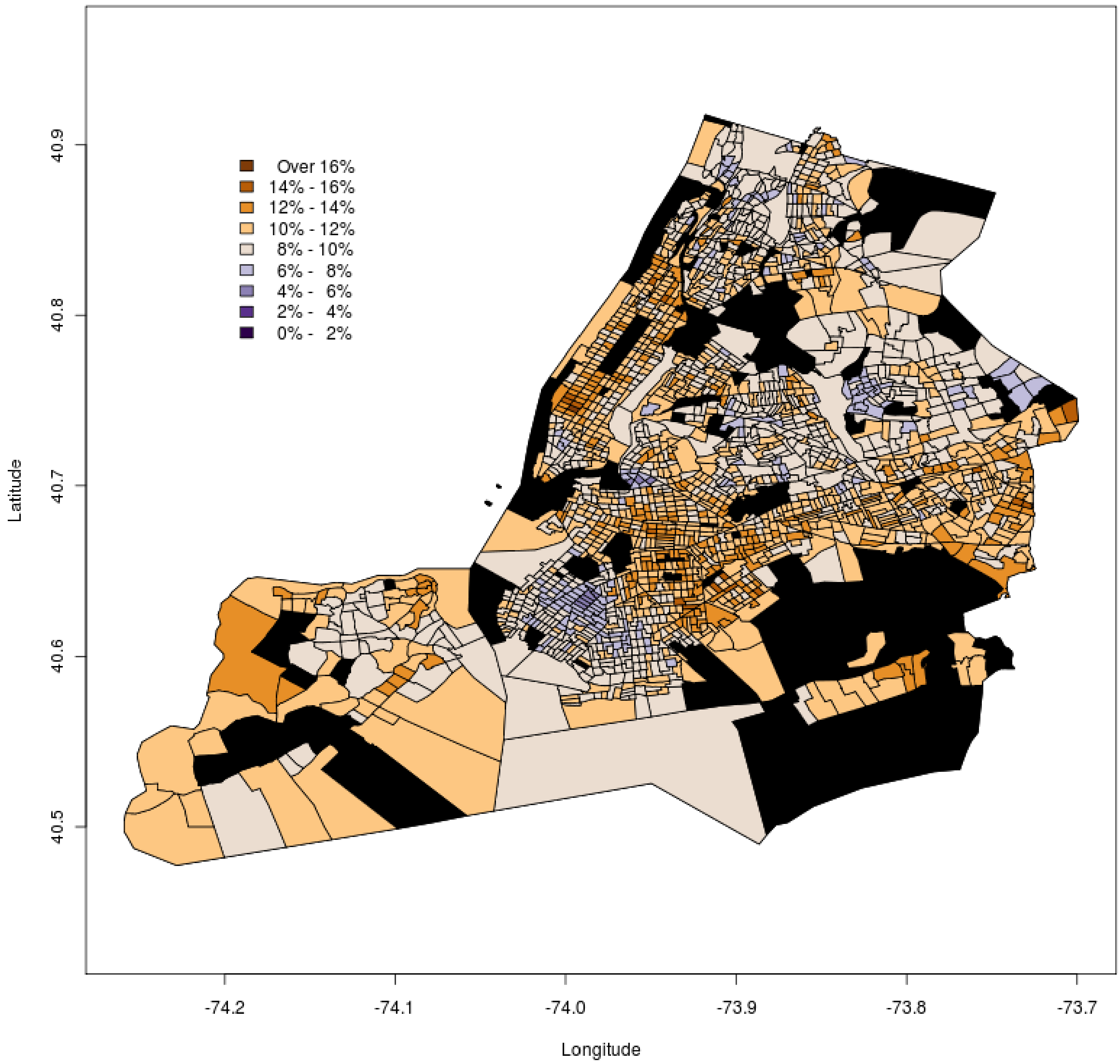

Figure 2.4: The percent chance a black mother has a PTB across space.

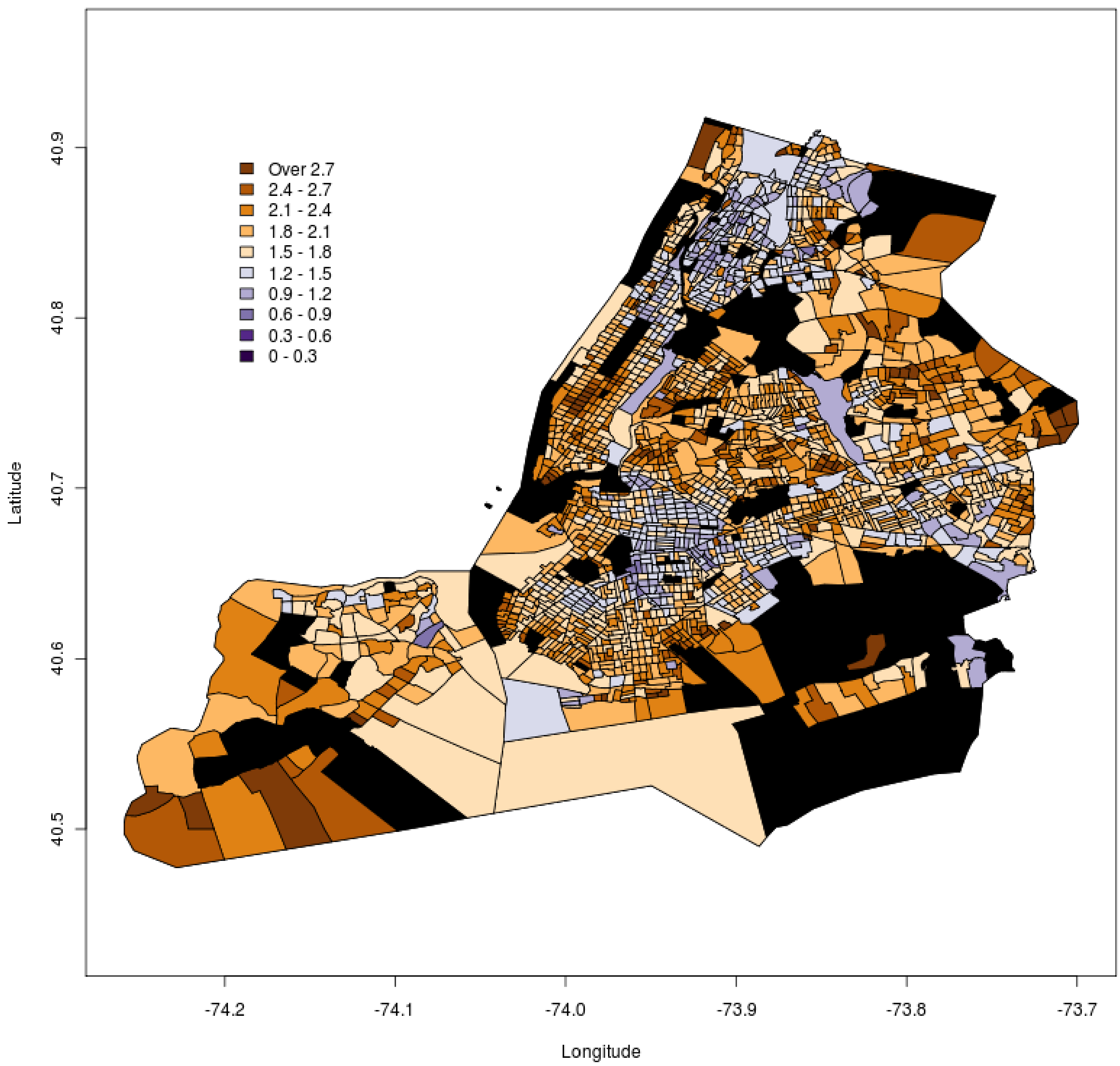

Figure 2.5: The standard error for the percent chance a black mother has a PTB across space.

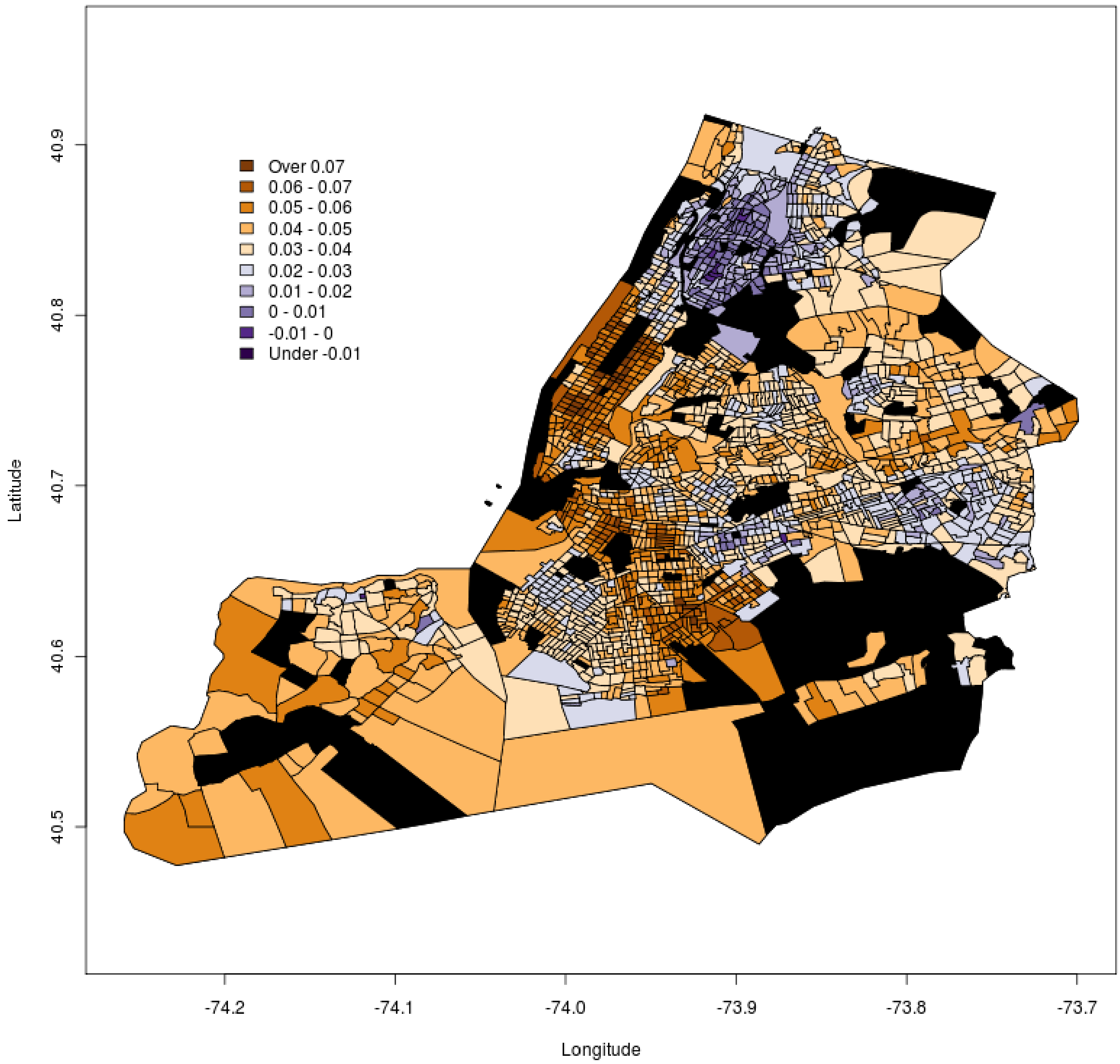

Figure 2.6: The difference between percent change of PTB for white and black mothers.

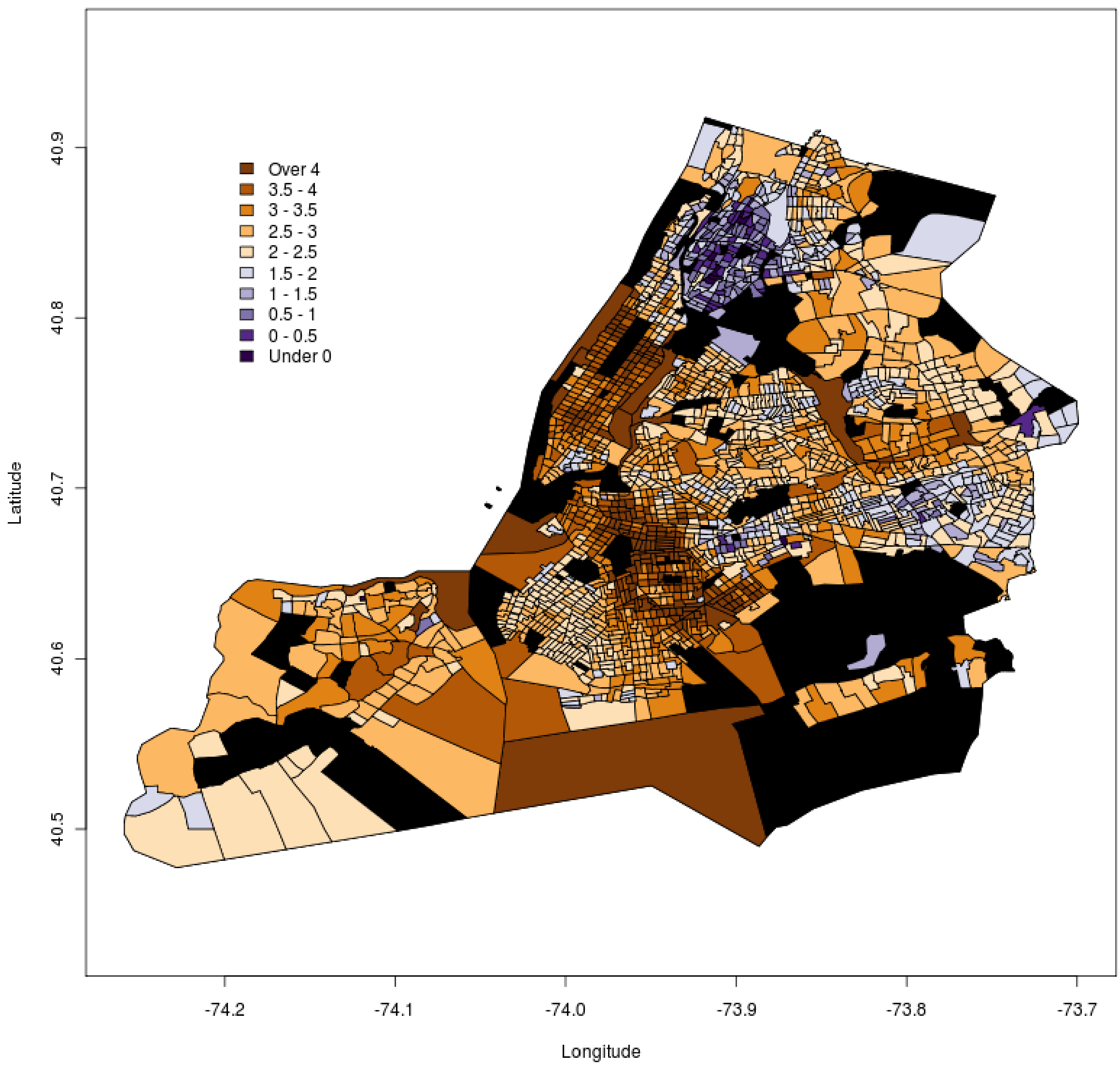

Figure 2.7: The z-values (difference / standard error of the difference) for the difference between percent change of PTB for white and black mothers.

Chapter 3

Studying the Effects of Aggregation and General Interaction of Multiple Pollutants with North Carolina Mortality for 2001-2002

3.1 Introduction

Multiple studies have linked ambient air pollution, particularly ozone and particulate matter (PM) to adverse health outcomes like asthma, birth defects, and mortality (Teague and Bayer, 2001; Schwartz et al., 1993). Some recent epidemiologic studies suggest that exposures to PM may result in tens of thousands of excess deaths per year, and many more cases of illness among the U.S. population (e.g., Bates, Baker-Anderson, and Sizto 1990; Dockery, Schwartz, and Spengler, 1992, Ostro et al., 1991; Schwartz 1994; Pope, Dockery, and Schwartz, 1995a, American Thoracic Society and Bascom 1996a, 1996b).

Analyses of the health effects of ambient air pollution often rely on data collected from a network of air quality monitors maintained by the U.S. Environmental Protection Agency (U.S. EPA 1995a,1995b).

Most community-based studies relate outdoor pollution levels measured at a small number of monitoring stations with aggregated counts of health outcomes (Schwartz 1994, Pope et al. 1995a, Bell et al. 2005). However, monitoring data are usually limited as monitors are sparse and $PM_{2.5}$ (particulate matter less than $2.5\mu m$ in diameter) especially suffers, as most monitors typically measured once every 3 to 5 days. Also, monitoring of $PM_{2.5}$ was initiated recently in 2000, and a significant portion of the monitoring network started in 2001. In contrast, PM_{10} (particulate matter less than $10\mu m$ in diameter) has been monitored much longer. Using the largest 90 U.S. cities, the National Morbidity, Mortality and Air Pollution Study (NMMAPS) (Samet et al., 2000a,b) is the largest multi-site study. The NMMAPS study was the first to estimate city-specific, regional, and national effects of PM_{10} on mortality. When using the monitoring data, typically a nearest monitor or basic averaging approach is taken to represent the pollution exposure for every person inside of a greater areal unit, such as county or community. Although aggregating data clearly introduces ecological bias (Wakefield and Shaddick, 2006), the bias is typically not corrected. There is limited information available as to how much the bias introduced through aggregation affects the estimation. One of the primary interests of this paper is to study the effects of aggregation on estimated risk of $PM_{2.5}$ on mortality.

We use a never before analyzed mortality dataset for North Carolina for 2001 and 2002, where individuals are street geocoded. Street geocoding offers the unique opportunity to choose the level of aggregation, and in the analysis of the data, we compare the aggregation levels of census tract, county, region (coastal, piedmont, and mountain), and state wide. The census tract level is spatially much smaller than the resolution provided by the monitoring network. In order to compensate for the lack of information in the monitoring data, we use other metrics for $PM_{2.5}$.

We use the Community Multiscale Air Quality (CMAQ) model and a recently released EPA product, Statistically Fused Air & Deposition Surfaces, to characterize $PM_{2.5}$ exposures. We will refer to the Statistically Fused Air & Deposition surfaces as the EPA fused data. CMAQ uses meteorological data, emissions data, and boundary values of air pollution as inputs. The CMAQ approach for pollution mapping may be unsatisfactory for two reasons. First, the underlying emissions data are often not of high quality (Dolwick et al., 2001). Second, the underlying models involve a variety of sources of uncertainty, such as unknown initial and boundary conditions; "model parameterizations" (the treatment of relevant chemistry phenomena varying at scales smaller than the grid size of the model); and numerical stability issues. It seems clear that combining the two main approaches and sources of information for $PM_{2.5}$ would lead to better estimates of $PM_{2.5}$ fields across the U.S. and hence to a better assessment of the association between mortality and PM exposure. The EPA fused data are the fusion of monitoring data with CMAQ output. In general, we expect the amount of information contained in the different metrics to be, in increasing order, monitoring data, CMAQ, and EPA fused data.

Distance to roadways, assessed using GIS, can be measured to any precision required. Aggregation of distance to roadways should induce a large loss of information as it is highly variable at even the census tract level. We expect the distance to roadways to characterize small scale pollution, such as pollution from mobile sources, which is not characterized well by the traditional $PM_{2.5}$ metrics. As traffic pollution has been linked to heart variability, it is important to be included in the analysis (Schwartz et al., 2005).

In our analysis, we use multiple metrics for $PM_{2.5}$ and distance to roadways to understand the effects of aggregation. With new data which are point geocoded, we are not only able to ascertain when bias is introduced, but we can determine when aggregation will introduce larger bias. The relationship between $PM_{2.5}$ and the distance to roadways is unclear, so we introduce a general interaction model that allows for understanding the relationship between $PM_{2.5}$ and the distance to roadways. The general interaction model is also applied to $PM_{2.5}$ and ozone, O_3.

The structure of the paper is as follows: In Section 3.2, we describe the data. In Section 3.3, we develop the model used in the analysis. In Section 3.4, we discuss fitting the model. In Section 3.5, we discuss the results, and Section 3.6 provides a concluding discussion.

3.2 Data

3.2.1 Mortality Data

Through a negotiated data sharing agreement with the NC State Center for Health Statistics, we have access to identified mortality data. The mortality dataset contains demographic, address, and cause of death data for every decedent in the State of North Carolina for 2001 and 2002. The address information is used to street geocode the data. Street geocoding assigns a point in space on the centerline of the street segment corresponding to the residential address of the decedent. Statewide, the match rate for street geocoding was 81.75%, and all geocoding was performed using ESRI ArcGIS 9.3. Six counties with a geocoding rate less than 50% were excluded from the study.

The sample size is further reduced by restricting to individuals who are at least 65 years old at the time of their death and who died of natural causes. Cause of death is recorded using the International Statistical Classification of Diseases and Related Health

Problems 10th Revision, ICD10. In most cases, multiple ICD10 codes are listed. We define an individual to have an unnatural death if any of the listed codes are related to an unnatural death. In total, there are 78,961 of deaths after exclusions. See Figure 3.4 for a map of North Carolina which includes the population density for people 65 years old and older at the tract level.

3.2.2 Pollution Data

Pollution data are provided through different metrics. The monitoring data for $PM_{2.5}$ and O_3, obtained from Air Quality System (AQS) using the Interagency Monitoring of Protected Visual Environments (IMPROVE), Speciation Trends Network (STN), and FRM networks. Monitoring data are observed at point locations, whereas numerical models are calculated on a grid. In order to avoid edge effects, we include monitors from the states neighboring NC. These include South Carolina, Georgia, Tennessee, and Virgina. In general, there are 105 to 135 monitors reporting every three or five days and approximately 20 stations reporting every day. The numerical model we use is the Community Multiscale Air Quality (CMAQ) models at a 36km resolution (Binkowski and Roselle 2003; Byun and Schere 2006). CMAQ uses as inputs meteorological data, emissions data, and initial and boundary values of air pollution (Binkowski and Roselle 2003; Byun and Schere 2006). The CMAQ model is a deterministic numerical model used to simulate chemical and physical interactions in the atmosphere based on various inputs such as weather, emissions, and initial conditions. CMAQ exposure estimates do not incorporate actual observed levels of pollutions. The EPA fusion surface combines the sparse monitoring data with the CMAQ surface. A newly released product provided by the EPA is fusion data. The EPA fusion data are a combination of CMAQ and monitoring data and is reported on a 12km grid. The fused data are constructed using a hierarchical Bayesian model that takes into account the change in support to fused data (McMillan et al., 2009). See the spatial interpolation method proposed by Fuentes and Raftery (Biometrics 2005) for more information.

Linking the pollution data to the spatial regions is performed through the census tract units. For the monitoring based data, each census tract for each day is linked to the closest monitor that is reporting. For the CMAQ and fused data which are on a grid, each census tract is linked to the centroid of the closest grid cell. For higher levels of aggregation, area weighted averages of the census tracts are used. The results are similar for the county level, but for the region level and state level, the average of the census tracts provides a more realistic average.

We use the distance to nearest roadway as a surrogate for exposure to $PM_{2.5}$ and other pollutants. We calculate the distance to nearest roadway using a statewide point grid created using Hawths Analysis Tools in GIS. The grid was normal, with 1000 feet between points in each direction. Each region in the study has at least two measurements. A mean for each region is calculated by averaging all points contained in the region.

3.2.3 Weather Data

Weather data for average daily temperature, pressure, and dewpoint are obtained from NNDC climate online website, which is available at:

`(http://www7.ncdc.noaa.gov/CDO/cdo).`

In total, 135 weather stations within North Carolina provide data during 2001 and 2002. There are some instances of missing values. The missing values are assumed to be random.

3.3 Model

3.3.1 Individual-level

The geocoding of mortality data allows us to investigate the impact of aggregation, going from models at the individual level to modeling counts of mortality at tract, county, and regional levels. Here, we directly model Y_{ikt}, the indicator of whether an individual i in region k died on day t, as a function of x_{ikt}, exposure data at location s_i (residence of individual i), and z_{ikt}, other covariates included in the model (e.g. weather). The individual-level model is given by

$$Y_{ikt}|x_{ikt}, z_{ikt} \sim \text{Bernoulli}(p(x_{ikt}, z_{ikt}, \beta)) \tag{3.1}$$

where the probability of death for individual i at time t in region k is

$$p(x_{ikt}, z_{ikt}, \beta) = \exp(f(x_{ikt}, z_{ikt}, \beta)). \tag{3.2}$$

As we are modeling deaths, we know that p is very small and $exp(f) < 1$.

3.3.2 Count Data

The expected number of people who die is the product of the number of people who live in a given region and the probability that each person will die on a given day. This can be represented as

$$E\left[Y_{kt}|x_{kt}, z_{kt}, \beta\right] = N_{kt}q_{kt}, \tag{3.3}$$

where

$$q_{kt} = \frac{1}{N_{kt}} \sum_{i=1}^{N_{kt}} \exp\left(f(x_{ikt}, z_{ikt}, \beta)\right), \tag{3.4}$$

and where Y_{kt} is the mortality counts in region k and N_{kt} is population data. While we have geocoded death recorded data, it by definition does not contain data on those who are still living. To address this issue, we use the number of people who live within a given region (e.g. census tract) based on the 2000 U.S. Census.

Because N_{kt} is typically large and $p(x_{ikt}, z_{ikt}\beta)$ is small, and we assume that the probability of death is the same for everyone in region k at time t, the binomial can be approximated with a Poisson distribution.

$$Y_{kt} \sim \text{Poisson}\left\{N_{kt}\exp(f(x_{kt}, z_{kt}), \beta)\right\} \tag{3.5}$$

where $f(x_{kt}, z_{kt})$ defines the mean function for the Poisson model.

3.3.3 General Interaction Model

Typically, $f(\cdot)$ is structured to have no interaction terms. In order to properly characterize the relationship between $PM_{2.5}$ and other variables, we break $f(\cdot)$ into two parts, the interaction portion and the non-interaction portion:

$$f(x_{kt}, z_{kt}, \beta) = g_k(x_{kt}, \beta_1) + f^*(z_{kt}, \beta_2) \tag{3.6}$$

where $f^*(z_{kt}, \beta_2)$ is the non-interaction portion of the model, $g_k(\cdot)$ is the interaction portion of the model which is a nonlinear function that depends on location. If we let $x_{kt} = (x_{kt}^{(1)}, x_{kt}^{(2)})$, then $x_{kt}^{(1)}$ and $x_{kt}^{(2)}$ are the variables we would like to characterize with a non linear interaction. We consider $x_{kt}^{(1)}$ to be $PM_{2.5}$ and $x_{kt}^{(2)}$ to be either O_3, distance to primary roadways, or distance to secondary roadways.

We consider $g(\cdot)$ a member of the flexible class of functions which can be expressed through a finite basis representation given by

$$g_k(x_{kt}^{(1)}, x_{kt}^{(2)}, \beta_1) = \sum_{m=1}^{M} w_m(k) b_m(x_{kt}^{(1)}, x_{kt}^{(2)}), \tag{3.7}$$

where $\beta_1 = \{b_m\}_{m=1}^{M}$ are two-dimensional basis functions (e.g., thin plate splines or polynomials) and $w_m(k)$ are spatially-varying coefficients. In this application, $w_m(k)$, and thus g_k, are constant across space ($w_m(k) = w_m$), but for a larger spatial domain, $w_m(k)$ and g_k could be spatially varying. One possible method is to use a multivariate CAR (MCAR) prior on $w_m(k)$ (Jin et al., 2005). The MCAR prior would allow the coefficients to vary spatially while allowing for structure among the coefficients.

Obtaining an estimate which is similar to the main effect reported when using a non-interaction model is not possible when using the basis function representation. To summarize the joint effect of two pollutants, we present the marginal (or instantaneous rate of) change for a given coordinate pair $(x^{(1)}, x^{(2)})$. If we let

$$b_m^{(x^{(1)})} = \frac{\partial b_m}{\partial x^{(1)}}$$

and

$$b_m^{(x^{(2)})} = \frac{\partial b_m}{\partial x^{(2)}},$$

then the gradient vector on the surface g_k is

$$\nabla g_k(x^{(1)}, x^{(2)}) = \left(\sum_{m=1}^{M} w_m(k) b_m^{(x^{(1)})}(x^{(1)}, x^{(2)}), \sum_{m=1}^{M} w_m(k) b_m^{(x^{(2)})}(x^{(1)}, x^{(2)}) \right). \tag{3.8}$$

$\nabla g_k(x^{(1)}, x^{(2)})$ gives both the direction and magnitude of the greatest increase in risk from location $(x^{(1)}, x^{(2)})$. As the gradient vector, $\nabla g_k(a, b)$, is additive, it is also easy to obtain the standard error. However, it is not meaningful to define a significance value to $\nabla g_k(x^{(1)}, x^{(2)})$, because even if all gradients are not statistically significant from point $(x_1^{(1)}, x_1^{(2)})$ to point $(x_2^{(1)}, x_2^{(2)})$, the difference in the function values, $g_k(x_2^{(1)}, x_2^{(2)}) - g_k(x_1^{(1)}, x_1^{(2)})$, may still be significant. This suggests that it may be informative to look at the gradient vectors to obtain a visual sense of which component is contributing more to the risk, actual significance should be checked for the difference between function values.

For this paper, we implement polynomial basis functions. Table 3.1 gives the 14 basis functions, excluding the intercept, that are used for a 4th-order general polynomial, along with the associated partial derivatives.

3.4 Model Fitting

3.4.1 One-stage vs Two-stage

The Poisson model can either be fit using all the data simultaneously or in two stages where the data are broken into separate analyses. Using all data in a single stage will be referred to as the one-stage model. The two-stage approach has been used extensively in the National Morbidity, Mortality, and Air Pollution Study, NMMAPS (Samet et al., 2000a,b). The two-stage approach fits the model in two stages. The data are only used in the first stage, where the analysis is broken into pieces based on the geographic domain. In the NMMAPS study, a separate analysis was carried out for each of the communities in the study and the regression coefficients are collected for each local level analysis. These local analyses are pooled in the second stage to estimate the overall effect of a covariate (PM or ozone). If the parameter we are interested in estimating is β, the two level normal model which is typically used is

$$\beta_k|\theta_k \text{ Normal}(\theta_k, V_k) \tag{3.9}$$

$$\theta_k \sim \text{Normal}(\gamma, A) \tag{3.10}$$

where β_k is the local estimate, of pollution from the first stage with its associated standard error V_k, and γ is considered to be the pooled effect. In the two-stage model, we denote the local effect for an analysis performed at a spatial subregion to be β_k versus β the estimated effect from the model using all the data simultaneously.

Fitting the model in two stages works if there are large enough groups for the first stage. The NMMAPS study used communities, usually multiple counties. For example, we cannot run a two stage analysis where the first stage is census tracts, because there are not enough mortalities observed in a census on a day to day basis to provide enough information for the Poisson model to provide reasonable estimates for the effect of pollution. However, the two-stage approach can be advantageous when the dataset is large in size where working with all the data at one time becomes difficult.

3.4.2 Analysis Framework

In analyzing the data, we use a general mean function $f(x_{kt}, z_{kt})$ throughout all the analyses.

$$\begin{aligned} f(x_{kt}, z_{kt}) &= \beta_0 + t\beta_t + \text{h}(t, 7) + \text{ns}(\text{TAVG}, df = 6) + \text{ns}(\text{PRES}, 6) + \text{ns}(\text{DPTP}, 6) \\ &+ g_k(\text{PM}_{2.5}, \text{O}_3, \text{Dist. Pri}, \text{Dist Sec}, \beta) \end{aligned} \tag{3.11}$$

where $h(t,7)$ is a temporal trend defined using a Fourier expansion with 7 sines and 7 cosines, $ns(\cdot, df = 6)$ are natural splines with 6 degrees of freedom, TAVG is the average daily temperature, PRES is the location based pressure, DPTP is the dew point, and β_1 is the effect of $PM_{2.5}$, the quantity of interest. Dist. is the distance to the nearest roadway, with Pri. meaning primary and Sec. meaning secondary. A 1-day lag is used for the pollution measurements.

Initially, g_k will be an additive function, with no interaction, (e.g. $g_k(PM_{2.5}, O_3, \beta) = \beta_1^{(1)} PM_{2.5} + \beta_1^{(2)} O_3$). We will use the additive form to evaluate and compare the different pollution metrics, aggregation levels, and including the distance to nearest roadways. We will also use a polynomial basis for g_k as outlined in the previous section to compare the interaction between $PM_{2.5}$, distance to roadways and O_3.

When using the additive form for g_k, we will conduct the analysis using the one-stage and two-stage models. In the single stage analysis, the different levels of aggregation represent the area unit by which the data have been aggregated. For example, the tract level includes tracts as the spatially aggregated unit. This is different than the two stage analysis where the first stage is always run using a county grouping. The aggregation can then be performed at either the county level or the tract level within this grouping. The county groupings are used in order to maintain enough data in each first stage analysis to have the model converge. Using only a single tract over a period of two years does not provide enough deaths to allow the Poisson model to converge. This means that for each of the two-stage analyses, the model is run using the 94 counties that are included in the model.

3.5 Results

This section analyzes the effect of $PM_{2.5}$ under various circumstances. First in Section 3.5.1, we investigate the relationship between multiple metrics for $PM_{2.5}$, as described in Section 3.2.2, at different aggregation levels. In section 3.5.2, we additively include distance to roadways in the analysis with $PM_{2.5}$. And, in section 3.5.3 we extend the analysis to the general interaction model of Section 3.3.3.

3.5.1 Metric and Aggregation Analysis

In this section, we examine the estimated risk $PM_{2.5}$ on mortality when using different metrics at different levels of aggregation. For all the analyses presented in this section, we use the framework described in Section 3.4.2 where $g_k(\mathrm{PM}_{2.5}, \beta^{(1)}) = \beta^{(1)}\mathrm{PM}_{2.5}$. In this model, we have a single covariate which is capturing the effect of pollution on health.

One of the most important parts of a health effects study is the metric chosen to characterize the exposure variable. We use monitoring stations, CMAQ, and fusion data to investigate the effect different metrics have on the estimated risk. For each of the different $PM_{2.5}$ metrics, we carry out the analysis at the state level, region level, county level, and tract level. Table 3.2 shows the results of using the different metrics at different levels of aggregation. The estimated effect varies dramatically depending on which $PM_{2.5}$ metric is used. In general, we see the estimated risk, at any given aggregation level, is consistently in the same order based on the metric. The order is monitoring data, CMAQ, then fusion. The significance of the estimated risks follows the same ordering. We expect this ordering because of the amount of information each metric contains. Due to sparsity, monitoring data contains the least amount of information of the three metrics we investigate, and it is also consistently estimated to be the least significant of the metrics. The fusion data, which contain the information from both the monitoring data and CAMQ, contain the most information and always have the largest estimated risk at any of the aggregation levels.

For comparison, the results of the two-stage modeling approach are presented in Table 3.3. The two-stage model is presented based on the level of the first stage. When the first stage is county, the risk of $PM_{2.5}$ is estimated at each county. Then the risk estimates are combined to estimate a single risk estimate for North Carolina. For the tract level analysis, rather than the first level being individual tracts, the first stage is the set of tracts for a given county with the data aggregated at the tract level. This results in the same number of first stage estimates as county version. The grouping is necessary because multiple individual tracts do not contain a sufficient number of deaths to provide stable estimates. The first stage estimates are combine in the second stage. The second stage is a normal model as described in Section 3.4.1. This is the same two-stage model typically fit in the NMMAPS papers.

The two-stage results are slightly different from using all the data in a single analysis. The standard error is increased slightly. This increase is expected, as the two level model does include a second layer of uncertainty. For the analysis performed with monitoring data, we notice that the estimated effect is increased proportionally to increase of the standard error. This results in estimates with the same level of statistical significance. However, the same is not true for the fused data. The estimated effects have decreased slightly while the standard errors have increased to the same level as the analysis using monitoring data. These changes result in greatly reduced statistical significance of the fused data.

We see that, when $PM_{2.5}$ is the only exposure covariate in the model, the estimated effect for the risk of $PM_{2.5}$ on mortality is influenced by the metric and level of aggregation used. However, it is rare that only the effect of $PM_{2.5}$ is estimated.

3.5.2 Additive Analysis of $PM_{2.5}$ and Distance to Roadways

In this section, we investigate the relationship of mortality and $PM_{2.5}$ while additively including distance to roadways. It is believed that monitoring data and other pollution surfaces are unable to characterize the exposure near roadways because monitors sparse relative to the size of the state. See Section 3.4.2 for more details about the analysis framework. Including the distance to roadways can be interpreted as allowing the intercept to vary, with structure, at the tract level.

To show the effects of aggregation, we first investigate the relationship of $PM_{2.5}$ and secondary distance to roadways at the tract and county levels. The results for these analyses are presented in Table 3.6. Due to the large bias introduced, we continue working with only the tract level.

Focusing on the tract level, we see that by adding distance to roadway to the model, the estimates for $PM_{2.5}$ are essentially the same as the model without distance to roadway. This suggests that the distance to roadway is providing additional explanatory power beyond that accounted for by the AQS, CMAQ, and fused data variables. Even the fusion data, which is at a much higher resolution than the monitoring network, is unable to explain the information provided by the distance to roadways. The analyses suggest that living further away from roadways is associated with decreased risk. This supports prior belief that roadways are a source of pollution that will locally increase risk for those who live close to the roadways.

Results from the two-stage frame work are presented in Table 3.5. We find that the standard errors are dramatically larger for the two-stage analyses compared to using all the data in a single stage. This shows that when using the two-stage model it is possible to dramatically attenuate the signal. For the distance to secondary roadways, we see z-scores of -1.9 for the two-stage model whereas we find z-scores of -9.2 for the single stage analysis. Statistically, the results show that distance to roadways are highly significant if we use the single stage analysis but they are less significant than $PM_{2.5}$ if we use two-stage analysis.

So far, we have only demonstrated the outcome if only distance to secondary roadways is included in the model. When both distance to secondary and distance to primary roadways are included in the model, the results are in Table 3.7. We see that the effect and significance have been reduced by including distance to primary roadways in the model. However, the direction of the effect is the same. In this case, we see negative effects for both distance from primary and distance from secondary roadways. This is consistent with the analysis with only secondary roadways. The reduced effect size and significance suggest that both of the distance to roadways are explaining the same information. Without allowing for interaction terms in the model, it is impossible to tell if the relationship is more complex.

3.5.3 Non-Additive Analysis

In order to understand the relationship between $PM_{2.5}$ and other variables, we move to non-additive functions that include $PM_{2.5}$ with Distance to roadways and O_3. See Section 3.4.2 for more details about the analysis framework. In this section, g_k in (3.11) is a non-additive function using polynomial basis functions, as described in Section 3.3.3. It is important to note that multiple methods for displaying the results are required because of the difference between interpreting a surface in a non-additive model and interpreting the effect of β in an additive model. The added complexity by including the interaction terms allows for added flexibility in the modeling that will help to explain the relationship between $PM_{2.5}$ and the other covariates.

Three different combinations of covariates are fit using the non-additive model. All of these analyses are performed at the tract level to maximize available power and minimize the introduction of aggregation bias. The combinations are $PM_{2.5}$ and O_3, $PM_{2.5}$ and distance to primary roads, and $PM_{2.5}$ and distance to secondary roads. Each combination has a figure and corresponding table. Akaike information criterion, AIC, values for the various models are given in Table 3.8. When using both secondary roads and primary roads, we see that the AIC decreases the most when 3^{rd} and 4^{th} order polynomials are used. This suggests that $PM_{2.5}$ and distance to roadways have a complex nonlinear relationship. We decide to use 4^{th} order polynomials for all model fits included here for consistency. The figure has contours of the surface which can be interpreted as 1% increase in the risk of death per unit of the corresponding variables (km for distance to roadways and 10 ppb for O_3). The plots also overlay gradient vectors to visualize the direction of greatest increase of risk at any given point. The arrows are not scaled, but rather simply indicate the direction of increasing risk. Also, the corresponding tables have selected changes in location on the surface to indicate how the surface is changing in effect size and statistical significance.

The combination $PM_{2.5}$ and O_3 corresponds with Figure 3.1 and Table 3.9. The figure shows that the surface is relatively linear in nature. This suggests that O_3 does not have a strong effect when compared with $PM_{2.5}$. This is reinforced by Table 3.9. We see that moving from left to right on the thin blue line is statistically significant (z-score = 3.95). However, as we move to thicker blue lines, the change on the surface becomes less significant. Also, the vertical lines presented suggest increasing mortality with increasing O_3, but none are statistically significant. This suggests that the significance of the change varies with location on the surface, and the significance depends on the direction of change, not the location itself.

The combination $PM_{2.5}$ and distance to primary roadways has corresponding Figure 3.2 and Table 3.10. The resulting surface is nonlinear in nature. We see that at approximately 1km in distance there is a strong increase in risk for increasing $PM_{2.5}$. Away from 1km in distance, we see that the distance to roadway dominates the change in risk. Through further exploratory analysis, we are unable to find any variables that present as confounding with the distance to primary roadways. Restricting the analysis to a smaller subset leads to results that are expected based on the presented results. The $z - scores$ for the distance to primary roadways are marginally significant compared to $PM_{2.5}$ in some locations on the surface.

The combination $PM_{2.5}$ and distance to secondary roadways has corresponding Figure 3.3 and Table 3.11. While the gradients appear to be relatively constant in direction, we can see in the table that as we move further away from secondary roads, the significance of the effect of $PM_{2.5}$ decreases. Also, the effect of distance to secondary roadways is highly significant across all the values of $PM_{2.5}$, but the significance is decreasing with increasing $PM_{2.5}$. In general, we see that the effect of $PM_{2.5}$ more significant more for closer distances to secondary roads. This suggests that while distance to secondary roads appears to dominate the risk of mortality, pollution is still statistically significant in some regions of the surface.

Overall, the non-linear surfaces provide greater insight into the interdependence between $PM_{2.5}$ and other covariates in the model as compared to the simple additive models.

3.5.4 Sensitivity Analysis

The analyses were evaluated for sensitivity to lag and the degrees of freedom in the smooth functions of the time and weather covariates. A sensitivity table for the lag is presented in Table 3.12. We see the largest effect using lag one suggesting the greatest correlation of any of the lags. Also, previous work supports this selection (Dominici et al., 2003). Table 3.13 and 3.14 presents the sensitivity analysis for the degrees of freedom for weather and time function, respectively. Changing the degrees of freedom does not impact the analyses significantly.

At the county level, we explore using county specific fixed effects. The results of the exploratory analyses are presented in Table 3.4. After comparing these results to the analyses with a fixed intercept across space (Table 3.2), we conclude the additional model complexity is not necessary to effectively estimate the effect of $PM_{2.5}$.

3.6 Discussion

This paper presents multiple analyses showing the relationship between aggregation and bias. We demonstrate that bias is dependent on both the pollution metric as well as the amount of spatial aggregation in the health data. Analyses where risk of pollution and other covariates are estimated using non-linear interaction models. The interaction models indicate that there is strong interaction between the covariates.

In all the comparisons presented, we compare back to the tract level, because tracts are small enough units that decreasing the size further does not change the pollution estimates assigned to individual locations. Tracts are a higher resolution than the resolution of the pollution covariates. This means that even if tracts were broken down further the sub regions would not have different estimates of pollution assigned to them making the two analyses approximately equivalent for the analysis of pollution. Also, moving the analysis to a individual based model would still require the use of areal units. Because we only observe the geocoded location of the people who died, we would have to estimate the locations of the people who are alive. This means an estimate of the population is required, which is only available at the areal level. In order to provide comparable analyses, we decided to use the tract level as the best possible analysis.

Aggregation can result in significant changes in the estimated effects. We know that by aggregating health outcomes, ecological bias is introduced (Wakefield and Shaddick 2006). However, the difference depends on the covariate used in the analysis. The analyses using monitoring data were affected much less than the analyses using CMAQ or fused data. This implies that as the amount of information in the covariate increases, the bias is more evident. The resolution and information contained in the pollution covariate is important in estimating the effect of pollution.

Distance to roadways contains more information than any of the pollution surfaces because we directly measure the roads to any precision we choose. The magnitude of the effect of the distance to secondary roadway is drastically reduced in the analysis at the county level when compared with the analysis at the tract level. This shows that variables that explain variation at one level of aggregation can appear drastically different after changing the aggregation level. This also suggests that the true effect of the distance to roadways may in fact be much greater than estimated at the tract level, because the distance to a roadway is fixed for an individual house, which is more of an individual covariate.

There is also significant difference when using the two-stage model versus using all the data in a single analysis. The significance of the estimates when using the fused data were drastically reduced. But, the size of the effect was increased for the analyses using the monitoring data. Using the two stage model is only possible for the additive effects model, as there is no clear way to use the two stage model on coefficients of a function represented with a basis.

The general interaction results clearly demonstrates nonlinear behavior. While the surfaces show nonlinear behavior, there are no model comparison statistics to report at this time to suggest a preferred model. The surface provides insight into how the $PM_{2.5}$ interacts with other variables. In the case of $PM_{2.5}$ and the distance to primary roadways, the surface raises the question why a signal emerges at 1km. Further investigation is required to evaluate possible explanations for this unexplained result.

The model presented in this paper does not take advantage of the spatial structure inherent in census tracts. A more complete spatial temporal model, such as what Choi et al. (2009) uses would increase the available information. This would allow for spatial variation within the model. However, it would dramatically increase the computational complexity of the analysis.

3.7 Tables and Figures

Table 3.1: Basis functions for a 4th-order polynomial, and associated derivatives. The partial derviatives of the basis functions are also given, where $b_m^{(a)} = \frac{\partial b_m}{\partial a}$ and $b_m^{(b)} = \frac{\partial b_m}{\partial b}$.

m	b_m	$b_m^{(x^{(1)})}$	$b_m^{(x^{(2)})}$
1	$(x^{(1)})^4$	$4(x^{(1)})^3$	0
2	$(x^{(1)})^3$	$3(x^{(1)})^2$	0
3	$(x^{(1)})^3(x^{(2)})$	$3(x^{(1)})^2(x^{(2)})$	$(x^{(1)})^3$
4	$(x^{(1)})^2$	$2(x^{(1)})$	0
5	$(x^{(1)})^2(x^{(2)})$	$2(x^{(1)})(x^{(2)})$	$(x^{(1)})^2$
6	$(x^{(1)})^2(x^{(2)})^2$	$2(x^{(1)})(x^{(2)})^2$	$2(x^{(1)})^2(x^{(2)})$
7	$(x^{(1)})$	1	0
8	$(x^{(1)})(x^{(2)})$	$(x^{(2)})$	$(x^{(1)})$
9	$(x^{(1)})(x^{(2)})^2$	$(x^{(2)})^2$	$2(x^{(1)})(x^{(2)})$
10	$(x^{(1)})(x^{(2)})^3$	$(x^{(2)})^3$	$3(x^{(1)})(x^{(2)})^2$
11	$(x^{(2)})$	0	1
12	$(x^{(2)})^2$	0	$2(x^{(2)})$
13	$(x^{(2)})^3$	0	$3(x^{(2)})^2$
14	$(x^{(2)})^4$	0	$4(x^{(2)})^3$

Table 3.2: Estimated effect of $PM_{2.5}$ for different levels of aggregation and different metrics. Each effect is estimated individually. Notation is Estimate$_{(Standard Error)}$[Z-score]. Restrictions for distance to the closest monitor are presented. Restricting to a distance closer than 75km results in significant data being excluded from the analysis.

	NC	Region	County	Tract
$PM_{2.5}$ Monitor	$0.022_{(.008)}$[2.9]	$0.018_{(.007)}$[2.7]	$0.0146_{(.0060)}$[2.5]	$0.0155_{(.0058)}$[2.7]
$PM_{2.5}$ Monitor <100km			$0.014_{(.006)}$[2.2]	$0.015_{(.006)}$[2.3]
$PM_{2.5}$ Monitor <75km			$0.014_{(.007)}$[2.0]	$0.015_{(.007)}$[2.2]
$PM_{2.5}$ Monitor <36km			$0.006_{(.009)}$[0.7]	$0.012_{(.008)}$[1.4]
$PM_{2.5}$ Monitor <12km			$0.002_{(.011)}$[0.3]	$0.008_{(.013)}$[0.6]
$PM_{2.5}$ CMAQ	$0.022_{(.008)}$[2.7]	$0.024_{(.007)}$[3.5]	$0.0190_{(.0060)}$[3.1]	$0.0216_{(.0059)}$[3.7]
$PM_{2.5}$ Fusion	$0.019_{(.007)}$[2.8]	$0.021_{(.006)}$[3.5]	$0.0273_{(.0055)}$[5.0]	$0.0322_{(.0053)}$[6.0]

Table 3.3: Two-stage results at the county and tract level using $PM_{2.5}$ monitoring data and fused data. Notation is $\text{Estimate}_{(StandardError)}$[Z-score].

	County	Tract
Monitor	$0.020_{(.007)}$[2.7]	$0.019_{(.007)}$[2.6]
Fused	$0.025_{(.007)}$[3.7]	$0.030_{(.007)}$[4.3]

Table 3.4: Using a spatially varying intercept which is independent and constant at the county level. Notation is $\text{Estimate}_{(StandardError)}$[Z-score].

	County
Monitor	$0.0146_{(.0060)}$[2.5]
CMAQ	$0.0193_{(.0061)}$[3.2]
Fused	$0.0203_{(.0055)}$[3.7]

Table 3.5: Two-stage results for the effect of $PM_{2.5}$ and distance to secondary roadways at the tract level using monitoring data and fused data. Tract level groups the tracts into counties in order to estimate the effect of distance to secondary roadways. Notation is Estimate $_{(StandardError)}$[Z-score].

	$PM_{2.5}$	Distance to Secondary
Tract Monitor	$0.019_{(.007)}$[2.6]	$-.448_{(.231)}$[-1.9]
Tract Fused	$0.027_{(.007)}$[4.1]	$-.448_{(.230)}$[-1.9]

Table 3.6: Estimated effects of $PM_{2.5}$ and distance to secondary roads at tract and county level of aggregation. Notation is $\text{Estimate}_{(StandardError)}$[Z-score].

	$PM_{2.5}$	Distance to Secondary
Tract Monitor	$0.016_{(.006)}$[2.7]	$-.097_{(.010)}$[-9.2]
County Monitor	$0.014_{(.006)}$[2.4]	$-.024_{(.009)}$[-2.8]
Tract Fused	$0.029_{(.005)}$[5.5]	$-.094_{(.011)}$[-8.9]
County Fused	$0.026_{(.005)}$[4.7]	$-.019_{(.009)}$[-2.2]

Table 3.7: Estimated effects of $PM_{2.5}$ and distance to roadway when in the model in a additive non-interactive design using monitoring data and fused data. Notation is Estimate$_{(StandardError)}$[Z-score].

	Monitor	Fused
PM2.5	$0.016_{(.006)}$[2.7]	$0.029_{(.005)}$[5.4]
Dist Sec	$-.065_{(.011)}$[-5.7]	$-.062_{(.011)}$[-5.5]
Dist Pri	$-.044_{(.006)}$[-7.3]	$-.044_{(.006)}$[-7.2]

Table 3.8: AIC values for various interaction models.

	Additive	Linear Interaction	2^{nd} Order Polynomial	3^{rd} Order Polynomial	4^{th} Order Polynomial
$PM_{2.5}$ and O_3	572139.3	572139.6			572142.1
$PM_{2.5}$ and Primary	572028.8	572023.1	572021.3	572001.3	571957.7
$PM_{2.5}$ and Secondary	572047.9	572046.6	572049.7	571893.8	571895.2

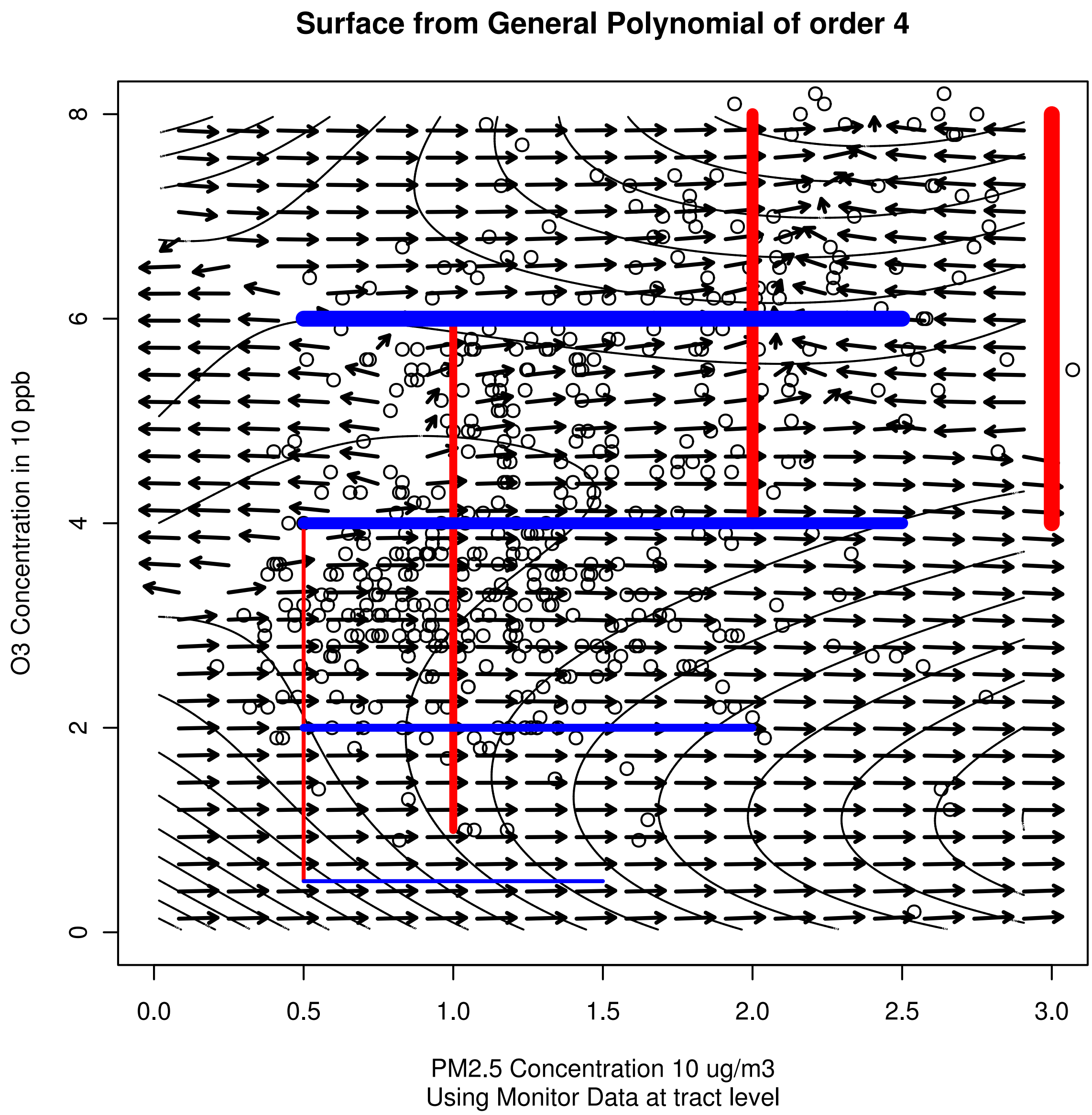

Figure 3.1: Resulting surface from non-linear function g_k using a polynomial basis function of order 4 for the joint effect of $PM_{2.5}$ and O_3. The lines represent points changes which are given in Table 3.9. The circles are a sub sample of the corresponding $PM_{2.5}$ and O_3 measurements in the analysis.

Table 3.9: The estimates and standard errors for the movement from the bottom to the top or left to right on the lines in Figure 3.1

Line	$\Delta(g)$	Δ_{SE}	$\frac{\Delta(g)}{SE}$
Thin Red	0.03	0.02	1.57
(vertical)	0.01	0.04	0.27
	0.06	0.04	1.46
Thick Red	0.04	0.04	0.99
Thin Blue	0.05	0.01	3.95
(horizontal)	0.04	0.02	1.84
	0.01	0.04	0.37
Thick Blue	0.01	0.05	0.12

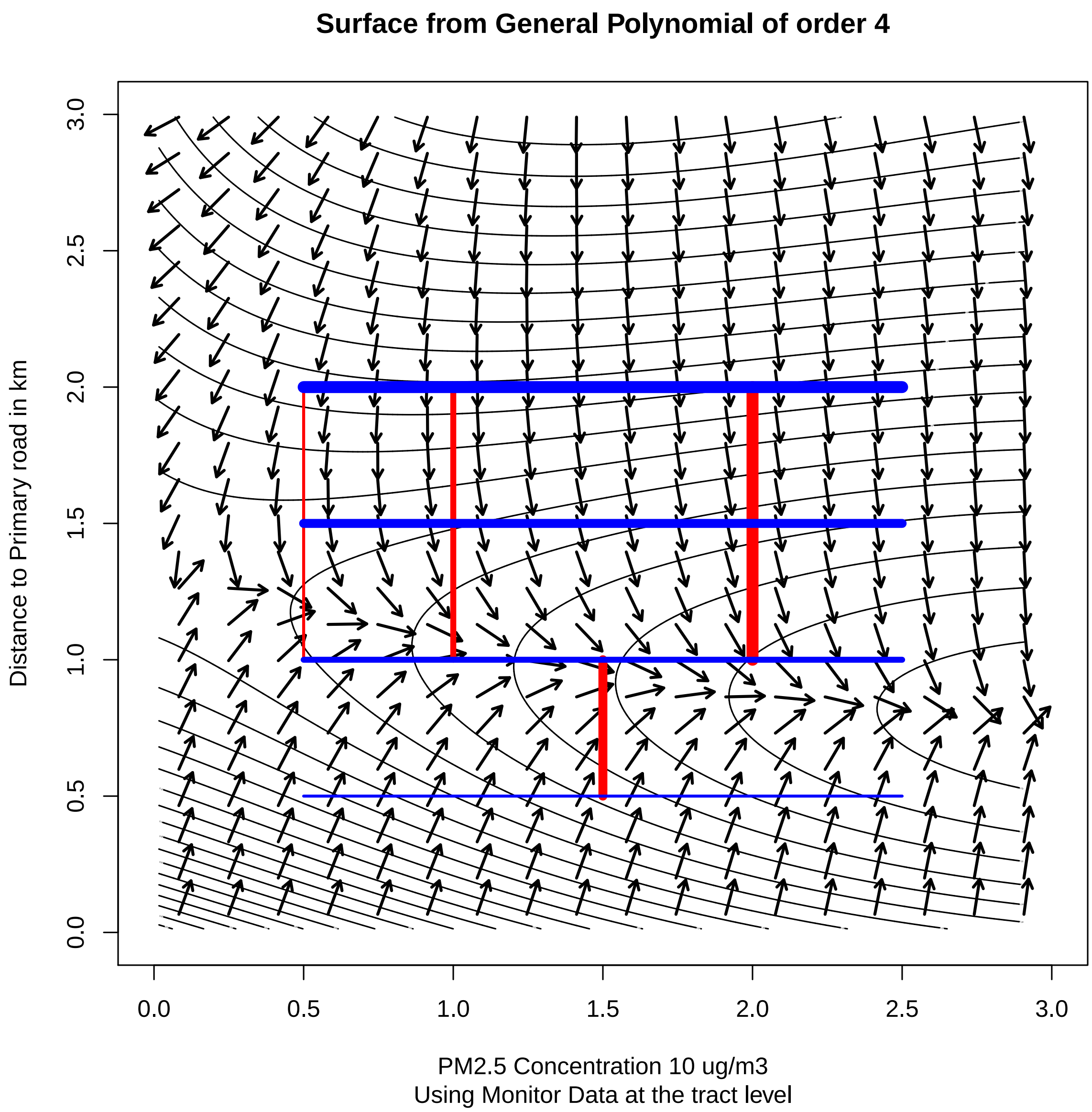

Figure 3.2: Resulting surface from non-linear function g_k using a polynomial basis function of order 4 for the joint effect of $PM_{2.5}$ and distance to primary roadways. The lines represent points changes which are given in Table 3.9.

Table 3.10: The estimates and standard errors for the movement from the bottom to the top or left to right on the lines in Figure 3.2

Line	$\Delta(g)$	Δ_{SE}	$\frac{\Delta(g)}{SE}$
Thin Red	-0.03	0.03	-1.10
(vertical)	-0.05	0.03	-1.61
	0.02	0.01	1.13
Thick Red	-0.07	0.04	-1.95
Thin Blue	0.08	0.02	3.89
(horizontal)	0.05	0.02	2.06
	0.03	0.03	0.96
Thick Blue	0.01	0.03	0.32

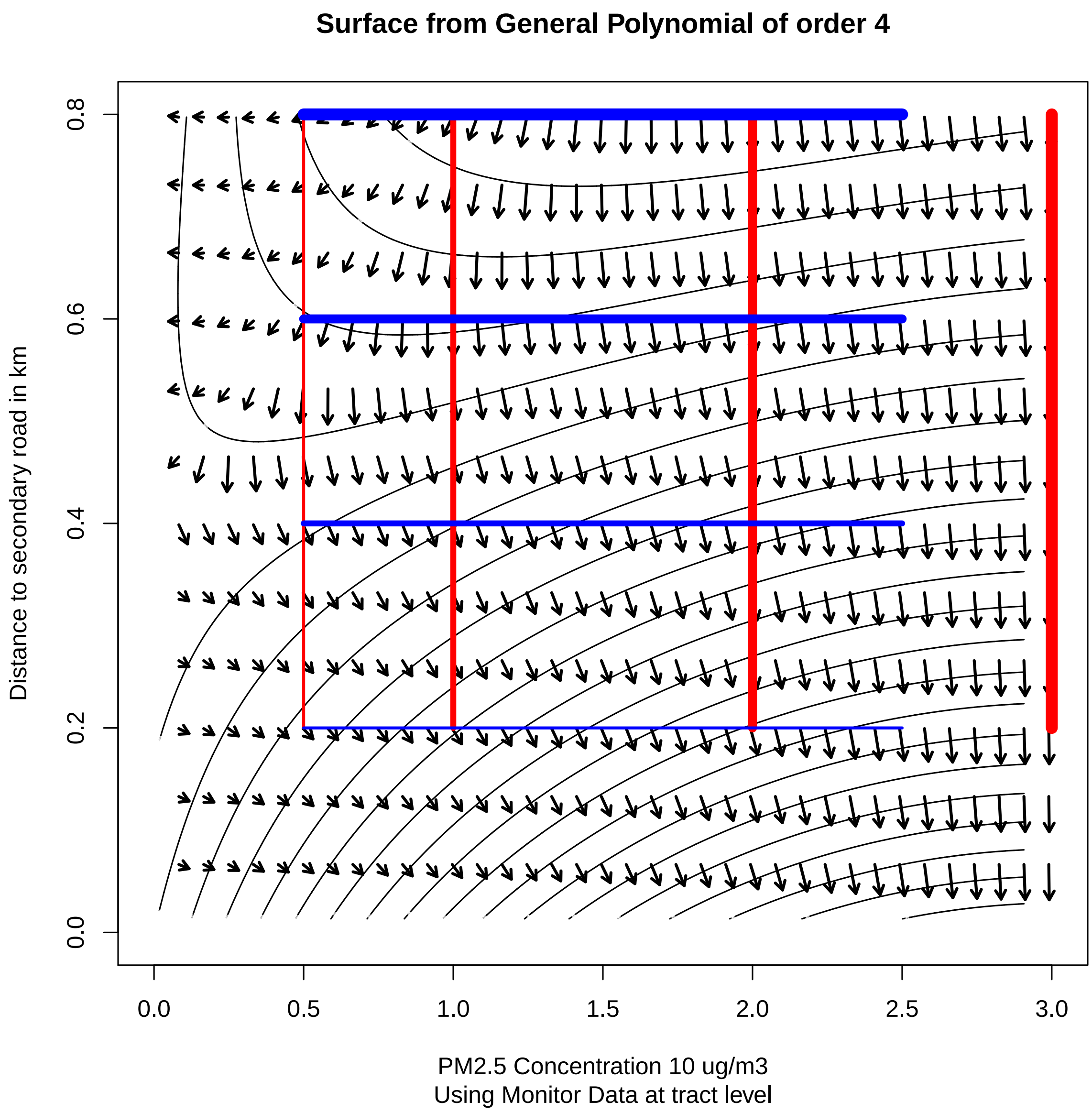

Figure 3.3: Resulting surface from non-linear function g_k using a polynomial basis function of order 4 for the joint effect of $PM_{2.5}$ and distance to secondary roadways. The lines represent points changes which are given in Table 3.9.

Table 3.11: The estimates and standard errors for the movement from the bottom to the top or left to right on the lines in Figure 3.3

Line	$\Delta(g)$	Δ_{SE}	$\frac{\Delta(g)}{SE}$
Thin Red	-0.05	0.00	-12.89
(vertical)	-0.09	0.01	-10.76
	-0.14	0.02	-6.76
Thick Red	-0.15	0.03	-4.36
Thin Blue	0.08	0.02	4.75
(horizontal)	0.04	0.02	1.84
	0.01	0.03	0.40
Thick Blue	-0.02	0.04	-0.38

Table 3.12: Estimated effects of of $PM_{2.5}$ and O_3 for different lags at the county level.

	PM2.5	O_3
Lag 0	$0.012_{(.006)}[2.1]$	$0.005_{(.003)}[1.5]$
Lag 1	$0.015_{(.006)}[2.5]$	$0.008_{(.003)}[2.4]$
Lag 2	$0.006_{(.006)}[1.0]$	$0.006_{(.003)}[1.9]$
Lag 3	$-.003_{(.006)}[-0.4]$	$0.002_{(.003)}[0.8]$
Lag 4	$-.007_{(.006)}[-1.2]$	$-.001_{(.003)}[-.2]$

Table 3.13: Estimated effects of of $PM_{2.5}$ and O_3 for different degrees of freedom for the natural splines at the county level.

	PM2.5	O_3
DF = 3	$0.015_{(.006)}[2.5]$	$0.008_{(.003)}[2.5]$
DF = 5	$0.015_{(.006)}[2.5]$	$0.008_{(.003)}[2.5]$
DF = 6*	$0.015_{(.006)}[2.5]$	$0.008_{(.003)}[2.4]$
DF = 7	$0.014_{(.006)}[2.4]$	$0.007_{(.003)}[2.4]$
DF = 10	$0.014_{(.006)}[2.4]$	$0.007_{(.003)}[2.4]$
DF = 15	$0.015_{(.006)}[2.5]$	$0.007_{(.003)}[2.4]$

Table 3.14: Estimated effects of of $PM_{2.5}$ and O_3 for different number of periods for the temporal effect at the county level.

	PM2.5	O_3
Periods = 3	$0.013_{(.006)}[2.2]$	$0.008_{(.003)}[2.5]$
Periods = 5	$0.014_{(.006)}[2.4]$	$0.008_{(.003)}[2.6]$
Periods = 7*	$0.015_{(.006)}[2.5]$	$0.008_{(.003)}[2.4]$
Periods = 9	$0.014_{(.006)}[2.4]$	$0.007_{(.003)}[2.4]$
Periods = 11	$0.014_{(.006)}[2.4]$	$0.009_{(.003)}[2.7]$
Periods = 13	$0.013_{(.006)}[2.2]$	$0.008_{(.003)}[2.6]$

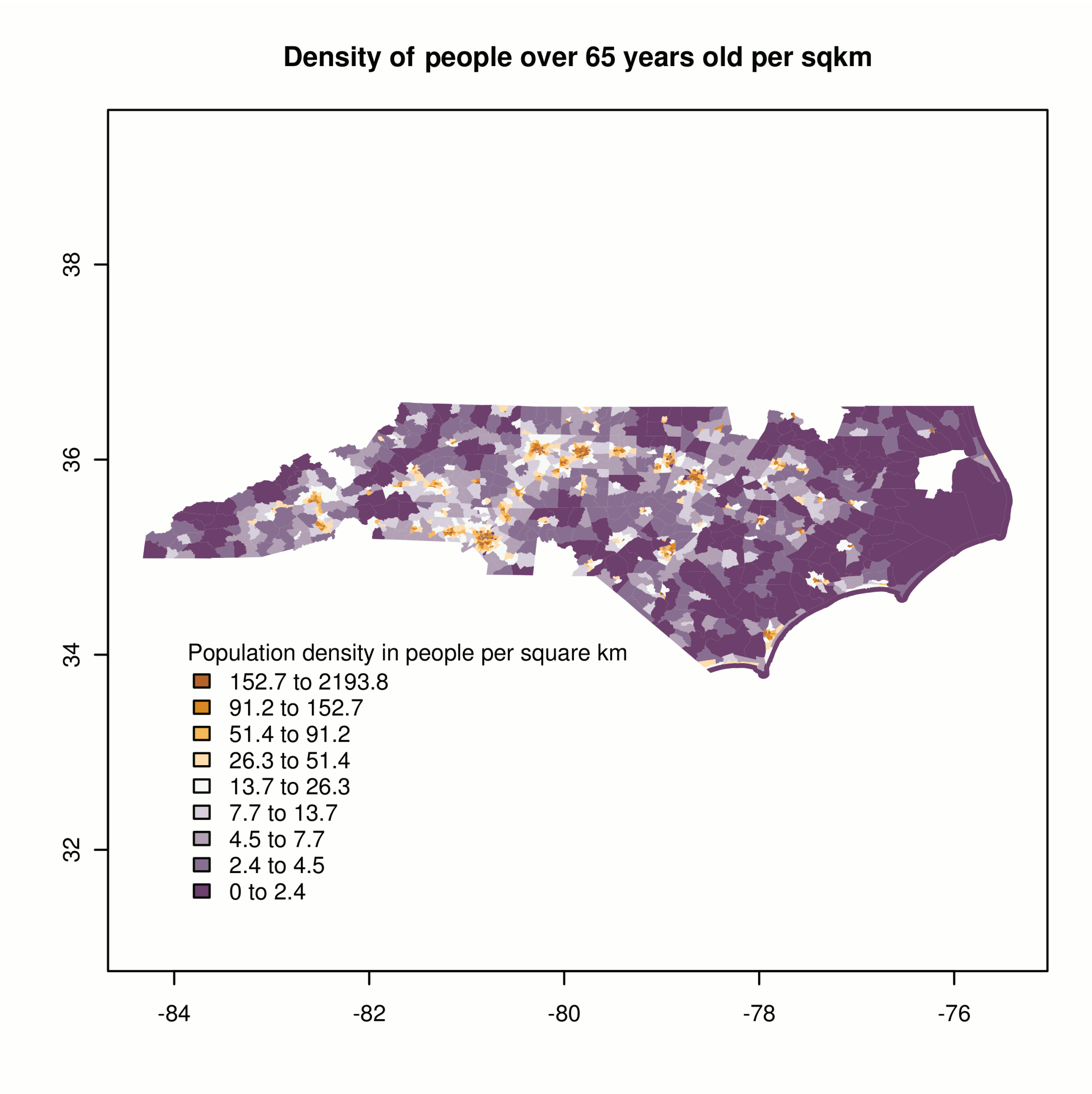

Figure 3.4: Population density for individuals over 65 at the tract level in North Carolina. The four blank counties are the counties which are excluded from the analyses.

Chapter 4

Correcting for Ecological Bias in a Mortality Study, North Carolina, 2001-2002

4.1 Introduction

Many epidemiological studies use health data that have been aggregated to some geographic unit. Using aggregated data leads to the introduction of ecological bias. This is especially true with studies involving the estimation of risk associated with exposure to pollution. It is well accepted that individual exposures can be highly variable within an aggregated geographic unit. The individual exposures have been estimated by multiple exposure simulators, some of which are Stochastic Human Exposure and Dose Simulation Models (SHEDS)(Burke et al., 2001), Air Pollutants Exposure Model (APEX) (Efroymson and Murphy, 2001), Probabilistic National Ambient Air Quality Standards (NAAQS) Exposure Model (pNEM) (Johnson et al., 1992), and an extension to pNEM (pCNEM)

(Zidek et al., 2005). However, instead of measuring individual exposures, typically pollution data are available from monitors that record ambient concentrations at a small number of sparse locations. Due to the variability across space, an individual's exposure will almost certainly be different from the measured observations. As Wakefield (2003) and Wakefield and Shaddick (2006) show, this underlying variability in the pollution surface leads to the introduction of ecological bias.

There are some situations in which there are several measurements taken within an aggregated geographical unit. In these situations, the usual approach is to perform some method of averaging on the measurements to provide a representative pollution level for the region of interest. These methods are prone to introducing ecological bias into the results. If the exposures within the region of aggregation are not constant, then averaging measurements ignores the fact that different individuals are exposed to different levels of pollution. Salway and Wakefield (2008) and Lee and Shaddick (2010) have developed methods to decrease the ecological bias by taking into account the individual level variability. Instead, the proposed methods use information from the multiple monitoring locations within the aggregated area to estimate the variability within a region through the use of a Bayesian spatiotemporal model.

One common feature of the bias correction methods is that they assume that a sufficiently large number of monitors exist within the study region to appropriately characterize the spatial variability within the aggregated region. In the case of Lee and Shaddick (2010), the level of aggregation is large relative to the number of monitors. They require greater than five monitors within the aggregated region otherwise the estimates will be prone to a significantly larger amount of bias. A current trend in health analyses is to aggregate health data at the county level. In rural and non-major cities, limited monitoring coverage exists. Most counties which are located away from major cities, typically have one or zero monitors. Since the number of monitors at the county level is often limited, less than five, the data aggregated at the county level will not effectively function in the framework developed by Lee and Shaddick (2010). A different approach to bias correction is to use exposure simulators to characterize the individual level variability of the exposure (Shaddick et al., 2007; Reich et al., 2008). In the framework introduced by Shaddick et al. (2007), bias correction is included when using the exposure distributions given by the exposure simulator. However, exposure simulators are based on accurately representing the population of interest on a number of levels, such as activity patterns and time spent in different environments. Also, it is currently unclear if the variability in exposures represented by the exposure simulator accurately correspond to the variability of individuals in the real world.

We extend previous bias adjustments to correct for ecological bias relative to a less aggregated geographic units rather than by using a exposure surface or exposure simulator. For example, we relate data aggregated at the county level to pollution estimated at census tracts instead of at the individual level. To this end, we evaluate how well the bias correction framework performs when we are unable to accurately estimate the spatial variability. By choosing the level of aggregation, we use for the exposure distribution we avoid introducing additional uncertainty. The monitoring data alone may be unable to characterize the variability in the pollution surface, so to improve the available information, we use a pollution surface which integrates numerical model output with the monitoring data. We use CMAQ for the numerical model, and the pollution estimates we use come from the Fused Air Surface product from the Environmental Protection Agency (EPA) which uses monitoring data to correct for bias in the CMAQ model. The CMAQ model alone is believed to be prone to bias as only the initial conditions are based on physical measurements. The fused data surface uses the monitoring station data to correct for any bias in CMAQ surface. The resulting surface contains the additional information provided by the numerical model with the centering of physical measurements.

The structure of the paper is as follows: In Section 4.2, we describe the data. In Section 4.3, we discuss the ecological models and ecological bias. In Section 4.4, we demonstrate the models through a simulation. In Section 4.5, we present results to the motivating data, and in Section 4.6, we provide a discussion to the overall framework.

4.2 Description of the Data

Through a negotiated data sharing agreement with the NC State Center for Health Statistics, we have access to identified mortality data. The mortality dataset contains demographic, address, and cause of death data for every decedent in the State of North Carolina for 2001 and 2002. The address information is used to street geocode the data. Street geocoding assigns a point in space on the centerline of the street segment corresponding to the residential address of the decedent. Statewide, the match rate for street geocoding was 81.75%, and all geocoding was performed using ESRI ArcGIS 9.3. Six counties with a geocoding rate less than 50% were excluded from the study. We restrict the population included in the study to only individuals over the age of 65 and who die of natural causes. Cause of death is determined through codes listed on the death certificate in international Statistical Classification of Diseases and Related Health Problems 10th Revision, ICD10, format. In cases where more than one cause of death is listed, we exclude the decedent if any of the codes listed indicate that cause of death is not natural. In total, there are 78,961 of deaths after exclusions. See Figure 3.4 for a map of North Carolina which includes the population density for people 65 years old and older at the tract level.

The monitoring data for $PM_{2.5}$ are obtained from Air Quality System (AQS) using the Interagency Monitoring of Protected Visual Environments (IMPROVE), Speciation Trends Network (STN), and FRM networks. In order to avoid edge effects, we include monitors from the states neighboring NC. These include South Carolina, Georgia, Tennessee, and Virgina. In general, there are 105 to 135 monitors reporting every three or five days and approximately 20 stations reporting every day. A newly released product provided by the EPA is the Statistically Fused Air & Deposition surfaces. We will refer to the Statistically Fused Air & Deposition surfaces as the EPA fused data. The EPA fused data are a combination of CMAQ and monitoring data and are reported on a 12km grid. The fused data are constructed using a hierarchical Bayesian model that takes into account the change in support to fused data (McMillan et al., 2009). CMAQ, Community Multiscale Air Quality models, is a numerical model providing output on a grid (Binkowski and Roselle 2003; Byun and Schere2006). CMAQ uses as inputs meteorological data, emissions data, and initial and boundary values of air pollution (Binkowski and Roselle 2003; Byun and Schere 2006). The CMAQ model is a deterministic numerical model used to simulate chemical and physical interactions in the atmosphere based on various inputs such as weather, emissions, and initial conditions. CMAQ exposure estimates do not incorporate actual observed levels of pollutions. The EPA fusion surface combines the sparse monitoring data with the CMAQ surface. See the spatial interpolation method proposed by Fuentes and Raftery (Biometrics 2005) for more information. The fused data provide estimates of the mean in the grid cell as well as the standard deviation.

Weather data are included in the analysis to control for environmental conditions. The weather data are from NNDC climate online website, which is available at:

`(http://www7.ncdc.noaa.gov/CDO/cdo).`

4.3 Ecological Models and Ecological Bias

4.3.1 Individual Level Model

The most basic model is an individual level model. Here we directly model Y_{ikt}, the indicator of whether an individual i in region k died on day t, as a function of x_{ikt}, exposure data at location s_i (residence of individual i), and z_{ikt}, other covariates included in the model (e.g. weather). The individual level model is given by

$$Y_{ikt}|x_{ikt}, z_{ikt} \sim \text{Bernoulli}(p(x_{ikt}, z_{ikt}, \beta)) \tag{4.1}$$

where the probability of death for individual i at time t in region k is

$$p(x_{ikt}, z_{ikt}, \beta) = \exp(f(x_{ikt}, zikt, \beta)). \tag{4.2}$$

As we are modeling deaths, it is safe to assume that p is a statistically rare event and $\exp(f) < 1$.

The expected number of people who die is the product of the number of people who live in a given region, N_{kt} and the probability that each person will die on a given day. This can be represented as

$$E\left[Y_{kt}|x_{kt}, z_{kt}, \beta\right] = N_{kt}q_{kt}, \tag{4.3}$$

where

$$q_{kt} = \frac{1}{N_{kt}} \sum_{i=1}^{N_{kt}} \exp\left(f(x_{ikt}, z_{ikt}, \beta)\right), \tag{4.4}$$

and where $Y_{kt} = \sum_{i=1}^{N_{kt}} Y_{ikt}$, the number of people who died in region k on day t.

Because N_{kt} is typically large and $p(x_{ikt}, z_{ikt}, \beta)$ is small, the binomial can be approximated with a Poisson distribution.

$$Y_{kt} \sim \text{Poisson}\left\{N_{kt}\exp(f(x_{kt}, z_{kt}, \beta))\right\} \tag{4.5}$$

where $f(x_{kt}, z_{kt}, \beta)$ defines the mean function for the Poisson model.

4.3.2 Approximation

In this section, we propose an adjustment that approximates the convolution model proposed by Wakefield and Shaddick (2006) without requiring the use of Bayesian modeling or MCMC. We propose this approximation because working with the convolution model for large populations may not be computationally possible in a reasonable amount of time. The approximation can be thought of as using information from a higher resolution, census tract, in an analysis at a lower resolution, county. For example, if the heath data are aggregated at the county level, and other covariates are available at, or can be estimated at, the tract level, then we can use the following approximation to adjust the county level analysis to regain some of the information that is present at the tract level which is lost in aggregating to the county level. This methodology does not assume a structure for pollution surface, but instead we are finding the best deterministic approximation.

The convolution model, for a single covariate, proposed by Wakefield and Shaddick (2006) estimates $N_{kt}q_{kt}$ with the equation

$$\sum_{i=1}^{N_{kt}} \exp(f(x_{ikt}, z_{ikt}, \beta)) = e^{\beta_0} \sum_{i=1}^{N_{kt}} \exp(\beta x_{ikt}). \tag{4.6}$$

The convolution model requires information about the exposures for all the individuals in region k without knowing which individuals the exposures correspond to. It should be noted that even this assumption is unrealistic in a mortality study, because we only observe those individuals who die, not all individuals in the study location. We expand the equation using a Taylor series expansion, centered around $\beta = 0$,

$$\sum_{i=1}^{N_{kt}} \exp(\beta x_{ikt}) = (N_{kt}) \left(1 + \beta_1 \frac{\sum_{i=1}^{N_{kt}} x_{ikt}}{N_{kt}} + \frac{1}{2}\beta_1^2 \frac{\sum_{i=1}^{N_{kt}} x_{ikt}^2}{N_{kt}} + \sum_{a=3}^{\infty} \frac{\beta^a \sum_{i=1}^{N_{kt}} x_{ikt}^a}{a! N_{kt}} \right). \quad (4.7)$$

From the Taylor series expansion we make approximations, without need for any distributional assumptions. Based on exploratory analysis and prior knowledge that the effect of pollution is small, we see no need to extend the approximation past the second order term. The first order approximation

$$\sum_{i=1}^{N_{kt}} \exp(\beta x_{ikt}) \approx (N_{kt}) \left(1 + \beta_1 \frac{\sum_{i=1}^{N_{kt}} x_{ikt}}{N_{kt}} \right), \quad (4.8)$$

and second order approximation

$$\sum_{i=1}^{N_{kt}} \exp(\beta x_{ikt}) \approx (N_{kt}) \left(1 + \beta_1 \frac{\sum_{i=1}^{N_{kt}} x_{ikt}}{N_{kt}} + \frac{1}{2}\beta_1^2 \frac{\sum_{i=1}^{N_{kt}} x_{ikt}^2}{N_{kt}} \right). \quad (4.9)$$

This leads to the new means for the Poisson model:

$$\lambda_{kt}^{(1)} = a_{kt} \exp(\beta_0 + \tilde{x}_{kt}\beta_1) \quad (4.10)$$

and

$$\lambda_{kt}^{(2)} = a_{kt} \exp(\beta_0 + \tilde{x}_{kt}\beta_1 + \frac{1}{2}\beta_1^2 \tilde{S}_{x,kt}^2). \quad (4.11)$$

where we break the region k into J_k subregions and $a_{kt} = \sum_{j=1}^{J_k} N_{jkt}$,

$$\tilde{x}_{kt} = \frac{\sum_{j=1}^{J_k} x_{jkt} N_{jkt}}{a_{kt}}, \tag{4.12}$$

$$\tilde{x}^2_{kt} = \frac{\sum_{j=1}^{J_k} x^2_{jkt} N_{jkt}}{a_{kt}}, \tag{4.13}$$

and

$$\tilde{S}^2_{x,kt} = \tilde{x}^2_{kt} - (\tilde{x}_{kt})^2. \tag{4.14}$$

Instead of computing the computationally expensive sum in the convolution model, we only need to store $\tilde{x}_{kt}$ and $\tilde{S}^2_{x,kt}$. $\tilde{x}_{kt}$ is the population weighted average of the mean pollution of the subregions, and $\tilde{S}^2_{x,kt}$ is the population weighted variance of mean pollution level of the subregions. In the case where the sub regions are exactly equal to the original regions, $\tilde{S}^2_{x,kt} = 0$. Neither $\tilde{x}_{kt}$ or $\tilde{S}^2_{x,kt}$ requires information of the pollution surface at the individual level. Instead of estimating the individual level distribution, we are in effect estimating the distribution for a different level of geographic aggregation. In the limiting case, $J_k = N_k$ every individual has an estimated level of pollution. Estimating the pollution for every individual may not be practical or possible. Therefore, the population weighted mean and variance are calculated based on assuming that the subregions are constant. This is equivalent to attempting to correct for the bias that is introduced by aggregating from the subregion level to the region level.

Only the first component is needed for the first order approximation and both components are required for the second order approximation. The first order approximation can be fit with regular Poisson regression, and the second order approximation requires using a restricted Poisson regression model.

Wakefield and Shaddick (2006), Shaddick and et al. (2007), and the other literature arrive at very similar expressions. The motivation in the rest of the literature is to correct for the bias by relating the aggregated health data back to distributions for individuals. Where as we take the approach to only compare back to a set of sub regions within the aggregated geographic units. Using the approximations to find appropriate weights for the sub regions is different from the other proposed methods that require estimating continuous surfaces or individual exposure distributions.

In order to understand the gain from using the second order approximation over the first order approximation, we first calculate the information matrices. The information matrix for the first order approximation is

$$I^{(1)}(\beta) = \sum_{t=1}^{T} \sum_{k=1}^{K} \begin{bmatrix} \lambda_{kt}^{(1)} & \lambda_{kt}^{(1)} \tilde{x}_{kt} \\ \lambda_{kt}^{(1)} \tilde{x}_{kt} & \lambda_{kt}^{(1)} (\tilde{x}_{kt})^2 \end{bmatrix}, \tag{4.15}$$

and the information matrix for the second order approximation is

$$I^{(2)}(\beta) = \sum_{t=1}^{T} \sum_{k=1}^{K} \begin{bmatrix} \lambda_{kt}^{(2)} & \lambda_{kt}^{(2)} \left[\tilde{x}_{kt} + \beta \tilde{S}_{x,kt}^2\right] \\ \lambda_{kt}^{(2)} \left[\tilde{x}_{kt} + \beta \tilde{S}_{x,kt}^2\right] & \lambda_{kt}^{(2)} \left[\tilde{x}_{kt} + \beta \tilde{S}_{x,kt}^2\right]^2 \end{bmatrix}. \tag{4.16}$$

To compare the difference in information, we develop a new metric to compare the two approximations. We look at the difference between the $(2, 2)$ element of the information matrices which describes the information of β. The difference in information is

$$\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_{kt}^{(1)} (\tilde{x}_{kt})^2 \left(\left[\exp(\frac{1}{2}\beta^2 \tilde{S}_{x,kt}^2)\right] \left[1 + 2\frac{\tilde{S}_{x,kt}^2}{\tilde{x}_{kt}}\beta + \left(\frac{\tilde{S}_{x,kt}^2}{\tilde{x}_{kt}}\right)^2 \beta^2\right] - 1 \right). \tag{4.17}$$

This shows that as the variation in of the regions increases relative to the mean, the difference in the information increases. To obtain a lower bound estimate for the percent increase in information, we assume $\beta^2 = 0$ and the ratio of the variance over the mean is constant, $\frac{\tilde{S}^2_{x,kt}}{\tilde{x}_{kt}} = \frac{\tilde{S}^2_x}{\tilde{x}}$. This implies the percent increase in information is greater than

$$\left[2\frac{\tilde{S}^2_x}{\tilde{x}}\beta\right][100\%]\,. \tag{4.18}$$

Typically, the largest beta is estimated to be is approximately 0.03, and if we assume that within region variance is equal to the mean, we would obtain an increase in information of about 6%. Also, the nature of the approximation, an upward adjustment of the sum, implies that overestimating S^2_x would result in an attenuated estimate of β. This actually presents a problem in estimation. As we are assuming that we have a limited number of sparse monitors, if the ratio of variation to the mean increases we loose the ability to estimate the mean pollution level effectively as well as the small scale variability. This implies that when the approximations are the most beneficial is also when the limited monitoring network provides the least amount of information.

4.4 Simulation

4.4.1 Simulation Design

In this section we design a simulation to compare the effectiveness of the first and second order approximations relative to using the usual analysis. The four models used in the simulation study are:

1. The generated mean of the region (the usual area based measure),

2. The first order approximation using generated subregion means,

3. The second order approximation using generated subregion means, and

4. Using the generated subregion values, no aggregation after the data were generated.

With the 4 given models we use data that are either based on being collected from a monitor, equivalent to an individual trajectory, or true exposure means of the subregions, the best case.

The data are generated using the typical Poisson model:

$$y_{itk} \sim Poisson\left(N_k \exp(\beta_0 + \beta_1 x^*_{itk})\right), \tag{4.19}$$

where x^*_{itk} is the generated pollution data at the individual level.

In the pollution design, we sample (n= 20) tracts from North Carolina which are considered our sub regions. These sub regions provide both the population size (over 65 years of age) (N_k) and tract areas. The pollution data are a single year of the values for a single cell from the EPA fused data. Assuming the variance of the pollution is a linear function of the mean of the pollution in the region, we generate exposures for each individual from a gamma distribution. ($x^*_{itk} \sim Gamma(mean = x^*_{tk}, var = s_f(x^*_{tk}))$), where $s_f(x) = \sigma_{sub} x$. Based on exploratory analysis, the Gamma distribution provided a reasonable fit to the fused data, and for large S^2_x a normal distribution is not very practical. Individual deaths are aggregated to the sub region level. Measurements for the pollution can be taken to be either the mean for each region or an individuals exposure path (which represents a monitoring station). In order to simulate the actual level of between subregion variation, the fused data are restricted to North Carolina, and to simulate the greatest amount of variability between subregions possible, the fused data are not restricted to North Carolina.

In the simulation the parameters that are changing are β_1 and σ_{sub}. These parameters take the values given in the following sets:

- $\beta_1 \in \{-.5, -.25, -.1, 0, .1, .25, .5\}$
- $\sigma_{sub} \in \{0.001, .1, .25, .5, .75\}$

While $\beta_0 = -9$.

The data are aggregated across the census tracts to generate one observation per a day y_t. The exposures are aggregated using the different methods described earlier, (area based, population based, and population based with inner region variability).

4.4.2 Simulation Results

Figures 4.1 – 4.4 and Figures 4.5 – 4.8 provide results for the bias when the pollution surface is smooth when we have monitor type profiles and subregion means respectively. Figures 4.17 – 4.20 and Figures 4.21 – 4.24 provide results for the bias when the pollution surface is variable between subregions when we have monitor type profiles and subregion means respectively.

There is a dramatic difference between the bias profiles when we have monitor type data versus true subregion means. The monitor type data results in a negatively correlated pattern and the subregion mean data results in U-shaped profiles. The largest different is for the typical values of β, the small positive values.

When the pollution surface is relative smooth, Figures 4.1 – 4.4 and Figures 4.5 – 4.8, there is relatively little bias for any when β is between -0.1 to 0.1 and we know the true mean within the sub regions. When we are using monitoring type data, we see that the increases slightly with the first order approximation performing the best. The analysis at the sub region level is prone to greater amounts of bias if there is within sub region variability without between sub region variability of the mean. This suggests that bias can be reduced by performing an analysis at an aggregated level if the mean trend is relatively smooth but small scale variability is large relative to the mean when monitoring type data are available.

When profiles were generated using the complete fused surface, representing significant between region variability, the profiles for the bias are dramatically different. The bias is again smaller when we know the true mean trend, Figures 4.17 – 4.20, versus if we have monitoring type data, Figures 4.21 – 4.24. The smallest bias is seen when analysis is performed at the sub region level using the mean trend data. With monitoring type data, the first order approximation performs the best when there is within region variability.

Like many other cases in statistics, we see a trade off between bias and standard error. We see in the mean $z-$scores, Figures 4.9 – 4.12 and Figures 4.13 – 4.16 for monitor profiles and true means respectively when using the subset data. Also, for the full fused data, we have Figures 4.25 – 4.28 and Figures 4.29 – 4.32 showing the mean z scores for monitor profiles and true means respectively when using the subset data. All the curves are essentially identical, within the suggesting that as bias is corrected the standard error increases proportionally.

4.5 Results

In this section, we demonstrate the effectiveness of using the approximations introduced in Section 4.3.2 using the aggregated mortality data from North Carolina. The general mean equation used for the analysis is

$$\begin{aligned} f(x_{kt}, z_{kt}) &= \beta_0 + t\beta_t + \mathrm{h}(t, 7) + \mathrm{ns}(\mathrm{TAVG}, df = 6) + \mathrm{ns}(\mathrm{PRES}, 6) + \mathrm{ns}(\mathrm{DPTP}, 6) \\ &+ \mathrm{PM}^*_{2.5}\beta_1, \end{aligned} \tag{4.20}$$

where $\mathrm{h}(t, 7)$ is a temporal trend defined using a Fourier expansion with 7 sines and 7 cosines, $\mathrm{ns}(\cdot\ , df = 6)$ are natural splines with 6 degrees of freedom, TAVG is the average daily temperature, PRES is the location based pressure, DPTP is the dew point, and β_1 is the effect of $PM_{2.5}$, the quantity of interest. In the above equation, $PM^*_{2.5}$ is the estimate of $PM_{2.5}$ used for the analysis, and this is the only component changing across the different analyses.

In this paper, we have introduced an approximation to correct the bias introduced by aggregating from the tract level to the county level. In relation to the simulation presented in Section 4.4, the analysis using the monitoring data is equivalent to the monitoring data used in the simulation, and the fused data are equivalent to the mean data used in the simulation. The fused data are estimates of the average pollution within grid cells, so there is a level of uncertainty which is not present in the simulation.

The results of the analysis are provided in Table 4.1. We compare the results at the county level to the results at the tract level, the baseline. Both the monitoring data and the fused data produce more significant results at the tract level compared to the county level. This demonstrates that bias is being introduced by aggregating the data.

All of the analyses have similar estimated standard errors. This is important as we would expect the analyses at the tract level to have smaller standard errors than the analyses at the county level. The standard error for the analyses using the fused data are smaller than the analyses using the monitoring data (0.005 versus 0.006). This suggests that relative to the pollution data available for $PM_{2.5}$, aggregating the data from the tract level to the county level does not result in a significant loss of information. It is important to note that the standard errors are slightly different in the fourth decimal place. However, the spread of the standard errors for both the monitoring data and the fused data are less than 0.0007.

Next, we investigate the size of the estimated risk of $PM_{2.5}$ on mortality. The tract analysis for both the fused data and the monitoring data have an estimated risk larger than any of the estimated risks at the county level. The same is also true for the significance, z-score, of the analyses. The difference in the analyses can be attributed to the aggregation of the health data. We obtain different results depending on if we use monitoring data or fused data. For the monitoring data, the size of the effect and significance are both smaller than the county area based estimator. For the fused data, the size of the effect and significance are both larger than the county area based estimator.

The difference between the performance of the approximations is likely a result of how much bias is introduced with the different pollution data. As there is almost no bias introduced when monitoring data are used, there is not a significant amount of information lost in the aggregation. Without a loss of information, the approximations cannot improve the estimates. The loss for the fused data are much larger; the risk decreases 0.005 when using the fused data compared with 0.002 when using the monitoring data. The larger loss of information allows the approximations to recover some of the loss and the estimates are corrected by 0.001 towards the tract level analysis. If the same ratio is true for the monitoring data, we would expect a change less than 0.0005. Therefore, the slight increase in standard error is more than enough to offset any improvement the approximations could have with the monitoring data.

In summary, the effectiveness of the bias correction approximations are dependent on the scale of β and the within region variability of the pollution. When β is close to zero or when the within region variability is small, the amount of bias introduced to the estimates is negligible. Understanding how these factors contribute to the potential ecological bias provides insight into which situations would benefit the most from using bias correction methods. The methods we propose provide an easy to implement framework, both computationally and statistically, that allows for bias correction, even in the particular case when there is no bias. In situations where there is no bias, the methods we propose do not increase the standard errors of the estimates. This means that the methodology can be used to evaluate if there is in fact significant bias in the risk estimates from spatial aggregation without the risk of significantly changing the risk estimates.

4.6 Discussion

This paper presents a new variation on bias correction for health effects analysis for exposure data. The current trend in the literature has been to correct for ecological bias by relating the health data to continuously estimated exposure surfaces or exposure distributions from exposure simulators. We have taken a different approach by linking the health data to exposure distributions estimated at smaller geographical units which are not at the individual level.

Rather than developing the model relative to the individual specific model, we develop the approximations based on a smaller level of aggregation, for instance, at the census tract level instead of at the county level. As the simulation results show, when the data available are representative of monitoring data and the mean trend is very smooth across space but highly variable on the small scale, aggregating the data will reduce the bias on average. However, if there is little to no small scale variability then the analysis with the least amount of aggregation always has the smallest amount of bias on average.

As we see with our data analysis, North Carlina has a relatively smooth pollution surface, and as a result find little difference between the different analyses. Because of the limited number of monitoring stations, it is unclear what the within region variability is for $PM_{2.5}$. The analysis suggests that there is very limited small scale variability in $PM_{2.5}$. By using the fused data, we see aggregating from the tract level to the county level does introduce bias into the risk estimates. However, even with the same fused data which was used at the tract level, we were unable to significantly reduce the bias.

The lack of bias reduction can be attributed to a few contributing factors, most importantly the magnitude of β. In the analyses, the largest β we estimate is approximately 0.03. This implies we can obtain an increase in information of about

$$6\frac{S_x^2}{\tilde{x}}\%. \tag{4.21}$$

Based on data exploration, we estimate $\frac{S_x^2}{\tilde{x}}$ to be much smaller than one. Even if we assume the ratio is equal to 1, this implies that the decrease in the standard error would be at most 0.001 and not practically affect the results.

Given a study where the effect size is larger, the approximations we propose provide simple tools that can be used to quickly correct for some bias introduced by aggregation without requiring complicated exposure surface modeling.

4.7 Tables and Figures

Table 4.1: Results from using the different models. The tract level analysis is using the data at the tract level aggregation. All other models are with county level aggregation with different models. County area has no correction for ecological bias, and the approximations attempt to correct for bias introduced by aggregating the tract level data to the county level.

	$PM_{2.5}$ Monitor	$PM_{2.5}$ Fused data
Tract level analysis	$0.016_{(.006)}[2.7]$	$0.032_{(.005)}[6.0]$
County Area	$0.014_{(.006)}[2.4]$	$0.027_{(.005)}[5.0]$
County First order approximation	$0.013_{(.006)}[2.2]$	$0.028_{(.005)}[5.1]$
County Second order approximation	$0.014_{(.006)}[2.3]$	$0.028_{(.005)}[5.1]$

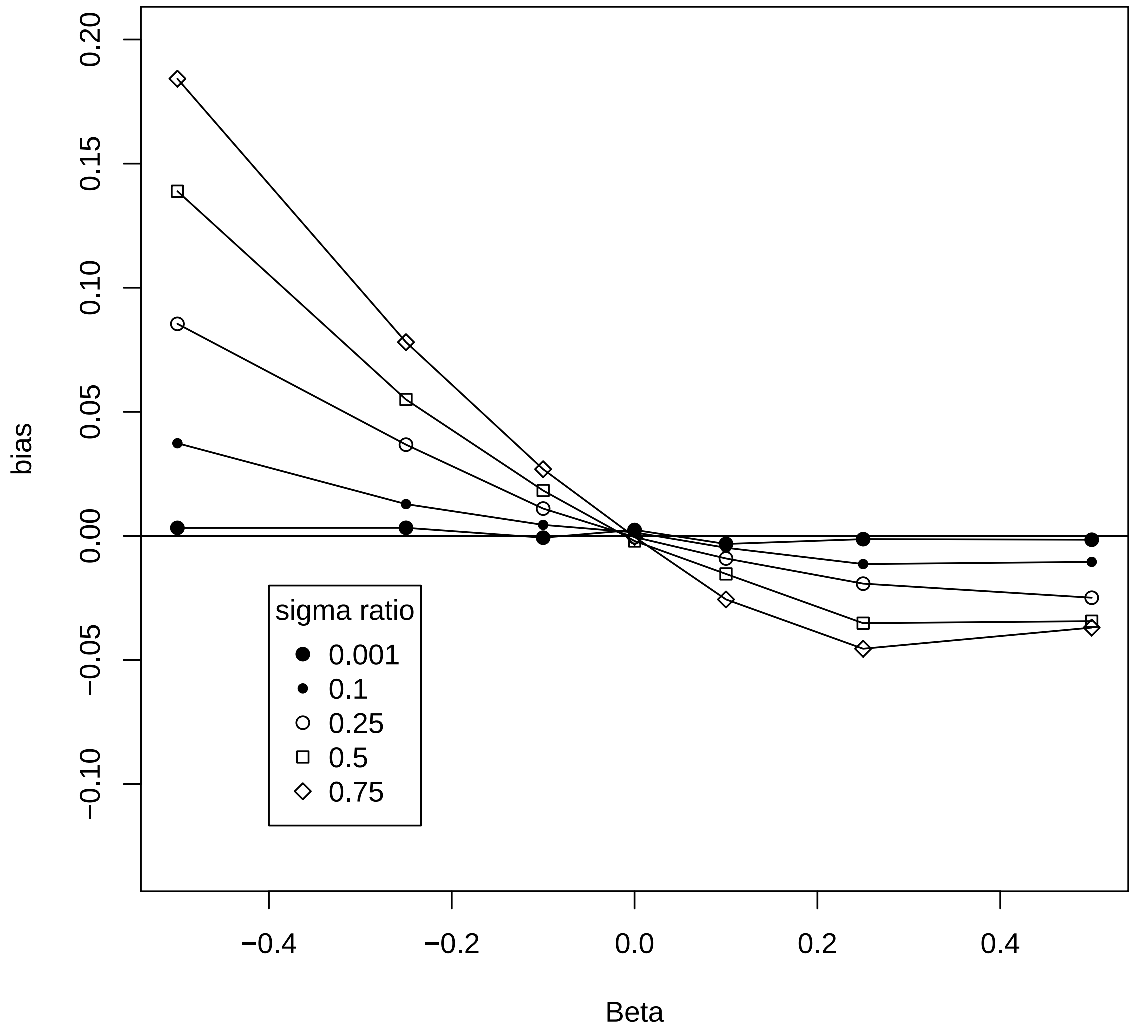

Figure 4.1: Simulation bias results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station using typical area aggregation to estimate the exposures.

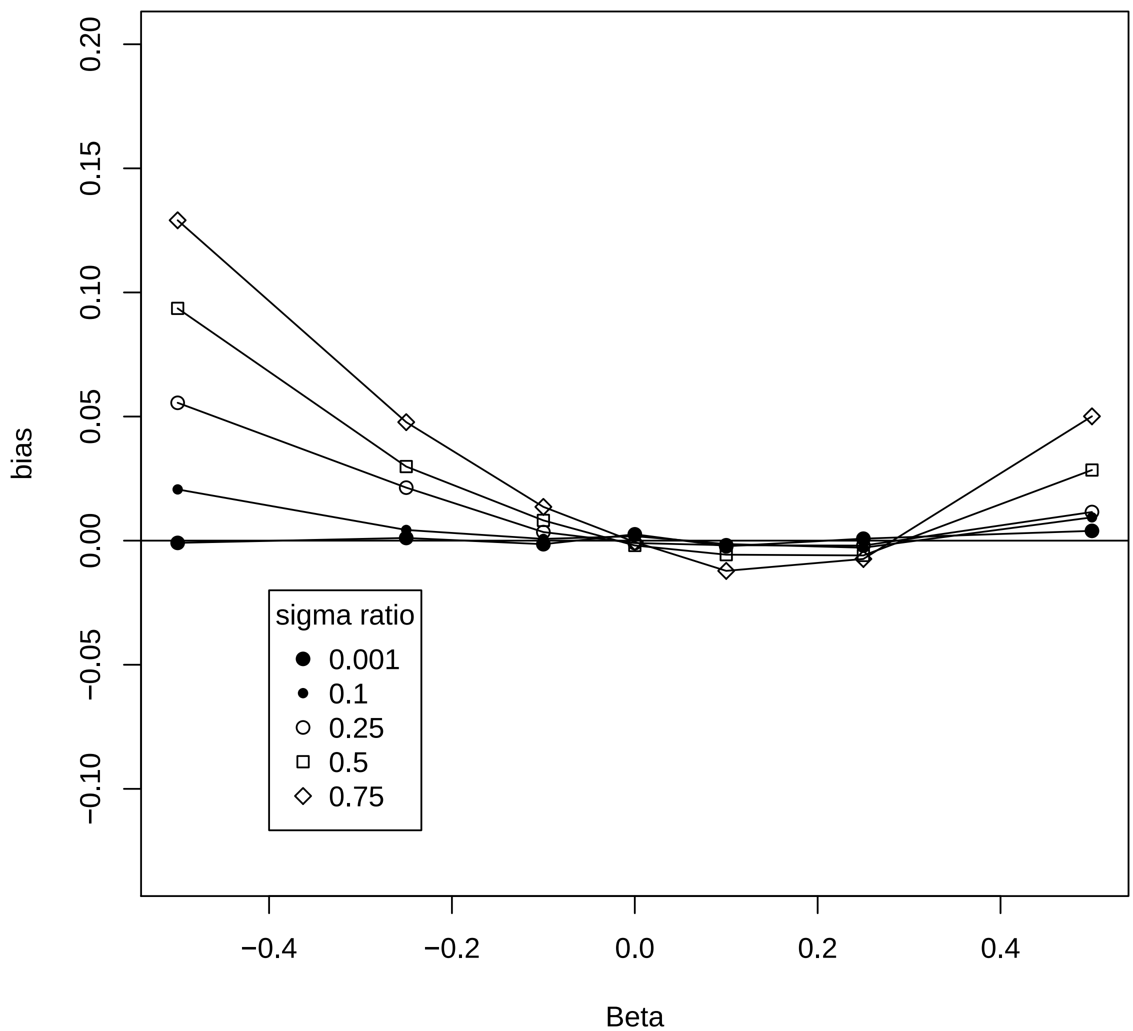

Figure 4.2: Simulation bias results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station using the first order approximation to estimate the exposures.

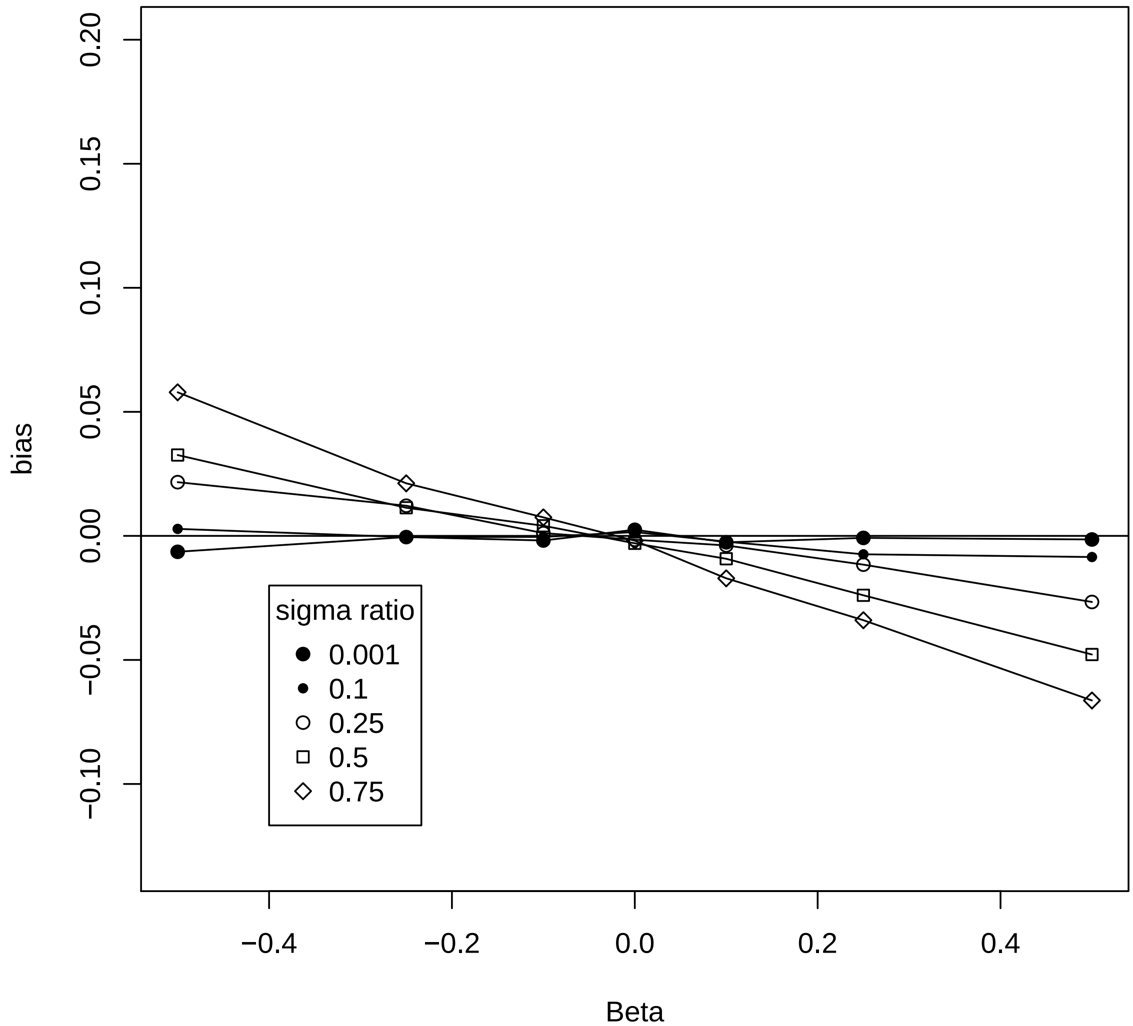

Figure 4.3: Simulation bias results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station using the second order approximation to estimate the exposures.

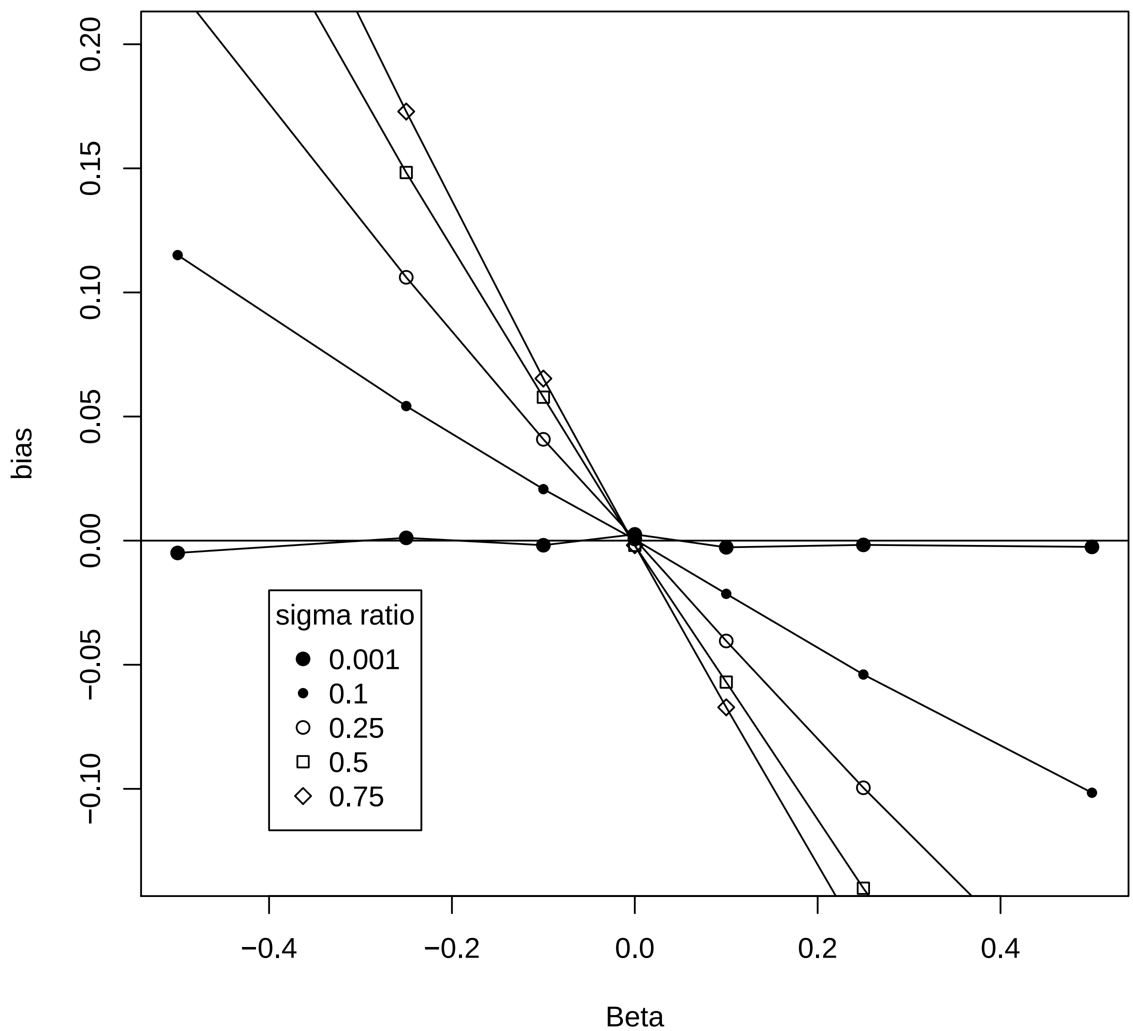

Figure 4.4: Simulation bias results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station using the non aggregated data at the sub region level.

Table 4.2: Simulation results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station. M-xβ refers to mean beta estimate for method x, and M-x-Z refers to the mean z-score for method x.

β	Sigma ratio	M-1β	M-1-Z	M-2β	M-2-Z	M-3β	M-3-Z	M-4β	M-4-Z
-0.50	0.001	-0.27	-1.05	-0.39	-1.33	-0.46	-1.43	-0.50	-4.30
-0.50	0.100	-0.21	-0.95	-0.33	-1.21	-0.40	-1.37	-0.38	-3.83
-0.50	0.250	-0.17	-0.80	-0.28	-1.10	-0.33	-1.21	-0.29	-3.30
-0.50	0.500	-0.11	-0.66	-0.20	-0.91	-0.26	-1.10	-0.19	-2.68
-0.50	0.750	-0.09	-0.59	-0.17	-0.85	-0.21	-1.03	-0.15	-2.42
-0.25	0.001	-0.15	-0.62	-0.21	-0.77	-0.27	-0.83	-0.25	-2.59
-0.25	0.100	-0.12	-0.55	-0.18	-0.72	-0.24	-0.76	-0.20	-2.30
-0.25	0.250	-0.10	-0.51	-0.16	-0.68	-0.21	-0.74	-0.15	-2.00
-0.25	0.500	-0.07	-0.42	-0.11	-0.52	-0.15	-0.57	-0.11	-1.64
-0.25	0.750	-0.07	-0.45	-0.11	-0.57	-0.15	-0.69	-0.09	-1.49
-0.10	0.001	-0.06	-0.25	-0.09	-0.32	-0.13	-0.32	-0.10	-1.15
-0.10	0.100	-0.05	-0.25	-0.07	-0.29	-0.11	-0.30	-0.08	-1.02
-0.10	0.250	-0.04	-0.19	-0.06	-0.27	-0.09	-0.26	-0.06	-0.91
-0.10	0.500	-0.02	-0.14	-0.05	-0.26	-0.09	-0.29	-0.05	-0.83
-0.10	0.750	-0.02	-0.13	-0.04	-0.22	-0.08	-0.27	-0.04	-0.71
0.00	0.001	0.02	0.10	0.03	0.11	-0.01	0.14	0.00	0.05
0.00	0.100	0.01	0.06	0.01	0.04	-0.03	0.06	0.00	0.06
0.00	0.250	0.01	0.06	0.01	0.05	-0.03	0.09	-0.00	0.00
0.00	0.500	-0.00	0.00	0.00	0.02	-0.04	0.03	-0.01	-0.05
0.00	0.750	0.00	0.03	0.01	0.05	-0.03	0.01	-0.00	0.02
0.10	0.001	0.07	0.36	0.11	0.50	0.08	0.57	0.10	1.44
0.10	0.100	0.06	0.35	0.10	0.49	0.07	0.56	0.08	1.32
0.10	0.250	0.04	0.28	0.09	0.44	0.05	0.50	0.06	1.14
0.10	0.500	0.04	0.32	0.08	0.44	0.04	0.47	0.05	1.03
0.10	0.750	0.02	0.22	0.05	0.34	0.02	0.37	0.04	0.93
0.25	0.001	0.20	1.08	0.31	1.42	0.25	1.70	0.25	4.20
0.25	0.100	0.17	1.05	0.27	1.36	0.22	1.64	0.21	3.82
0.25	0.250	0.13	0.94	0.23	1.26	0.18	1.51	0.17	3.52
0.25	0.500	0.10	0.84	0.19	1.14	0.15	1.44	0.13	3.24
0.25	0.750	0.08	0.76	0.16	1.10	0.12	1.30	0.11	2.91
0.50	0.001	0.48	3.13	0.69	4.06	0.53	5.87	0.49	11.84
0.50	0.100	0.38	2.68	0.62	3.59	0.48	5.03	0.43	10.80
0.50	0.250	0.31	2.49	0.55	3.60	0.42	5.25	0.36	10.46
0.50	0.500	0.25	2.61	0.48	3.78	0.34	5.71	0.29	10.66
0.50	0.750	0.23	2.55	0.45	3.68	0.31	4.98	0.26	10.36

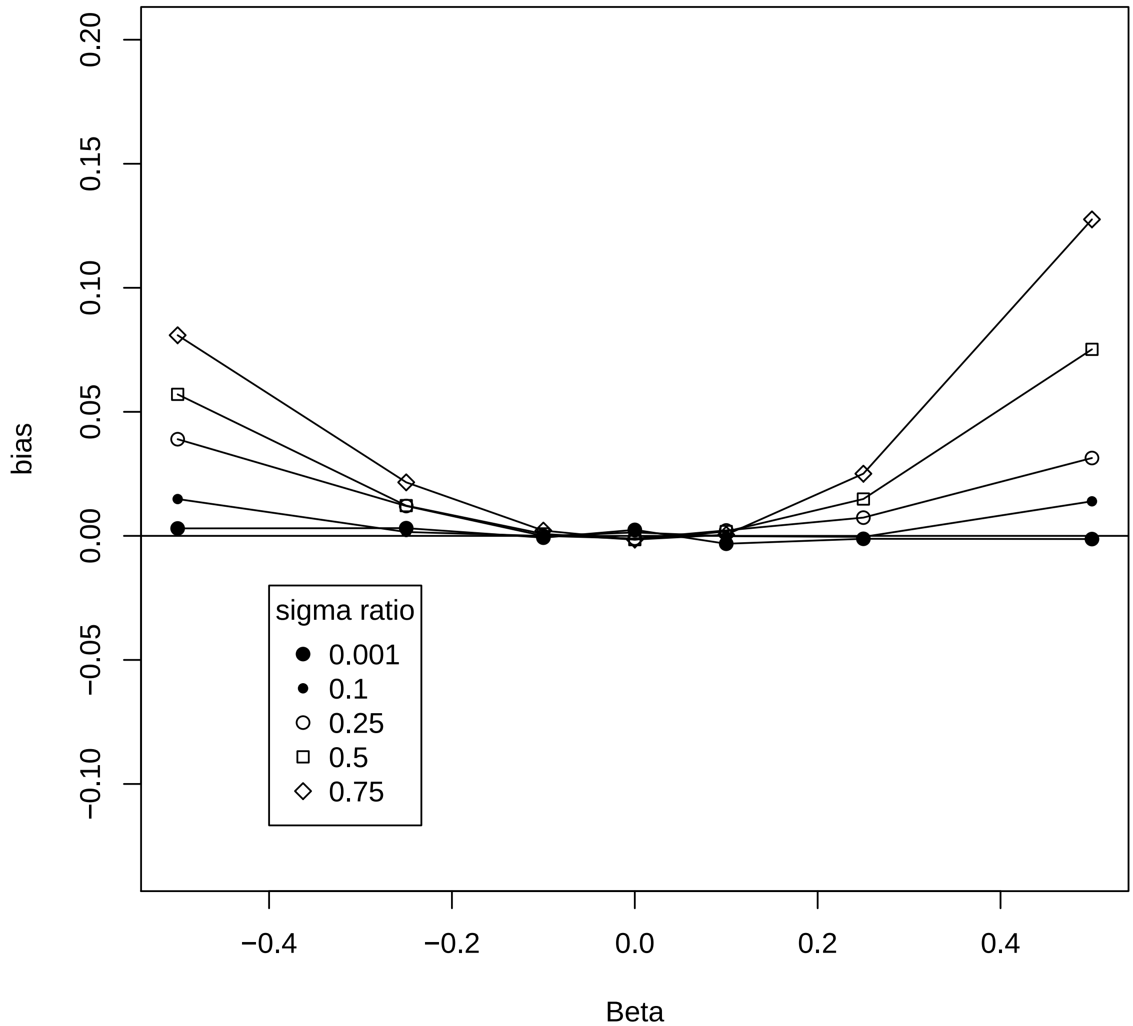

Figure 4.5: Simulation bias results using the fused data subset of North Carolina under the condition that the data available are the true means of the subregions using typical area aggregation to estimate the exposures.

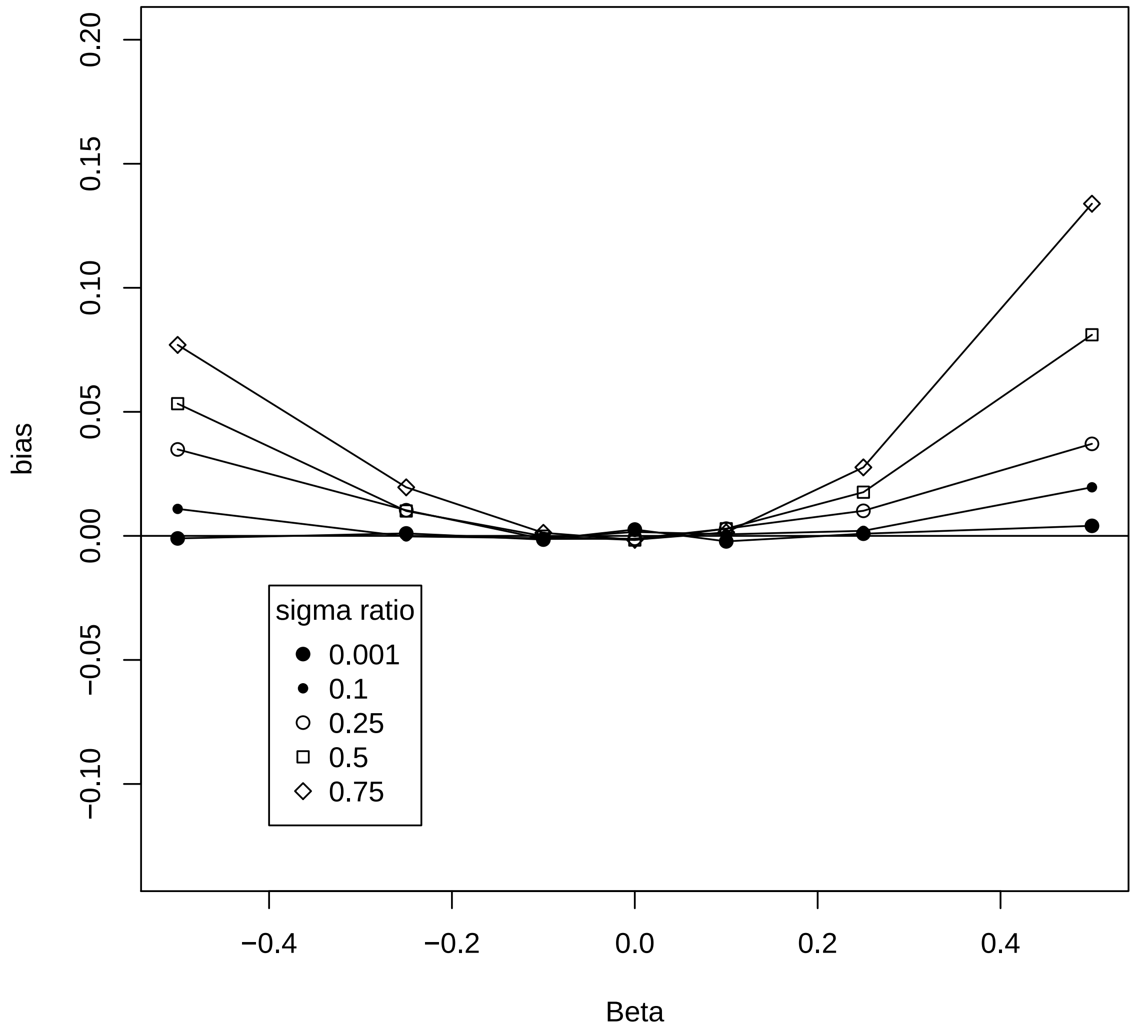

Figure 4.6: Simulation bias results using the fused data subset of North Carolina under the condition that the data available are the true means of the subregions using the first order approximation to estimate the exposures.

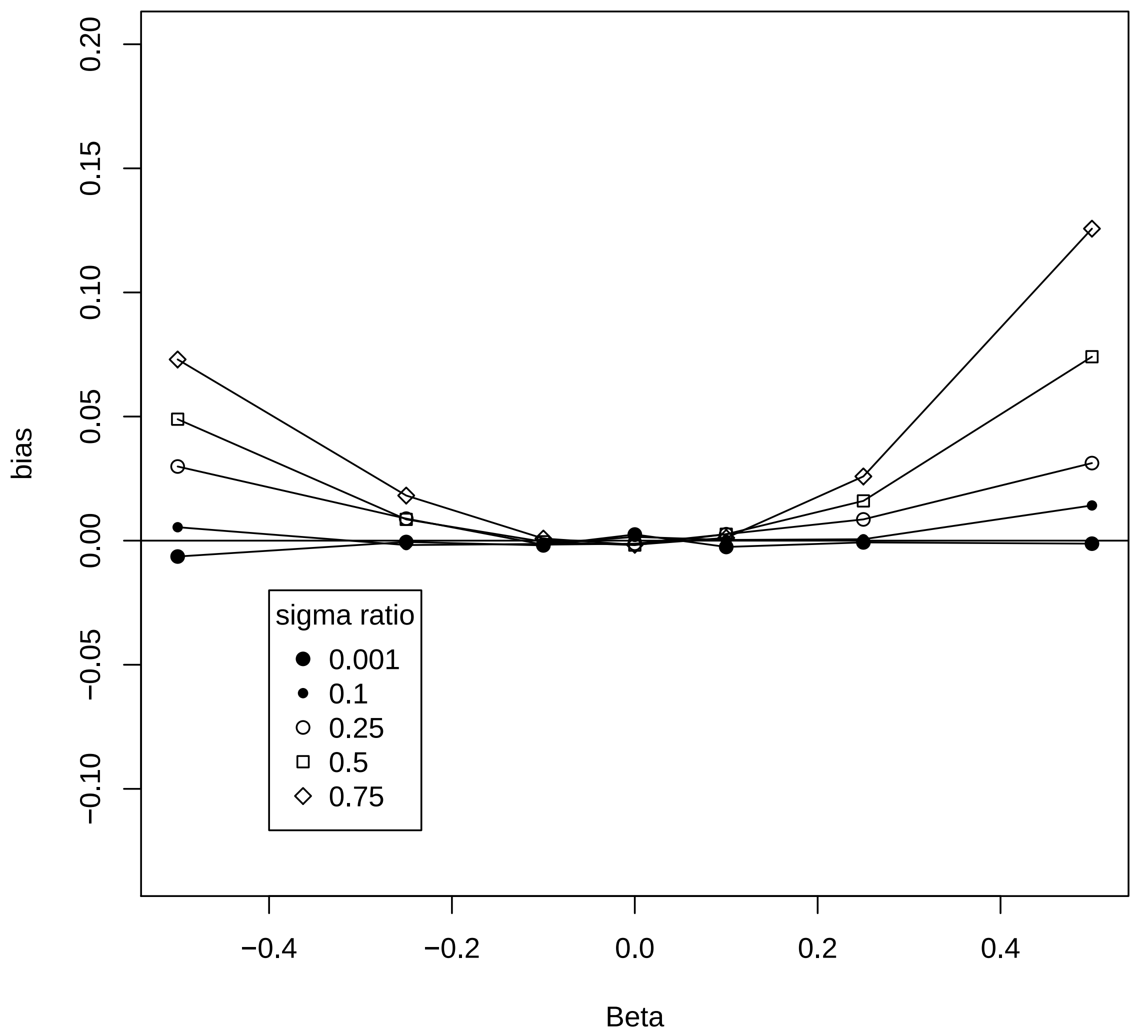

Figure 4.7: Simulation bias results using the fused data subset of North Carolina under the condition that the data available are the true means of the subregions using the second order approximation to estimate the exposures.

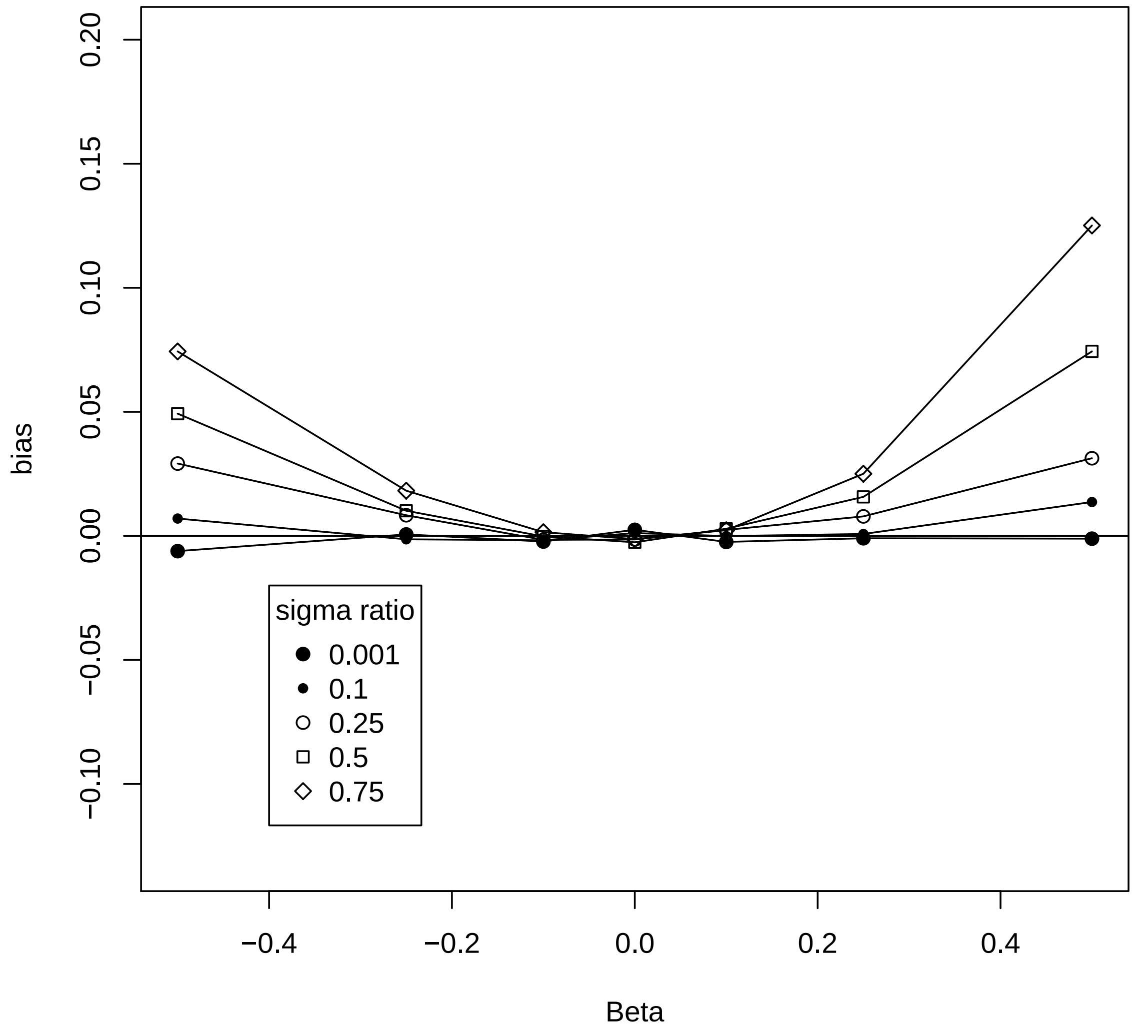

Figure 4.8: Simulation bias results using the fused data subset of North Carolina under the condition that the data available are the true means of the subregions using the non aggregated data at the sub region level.

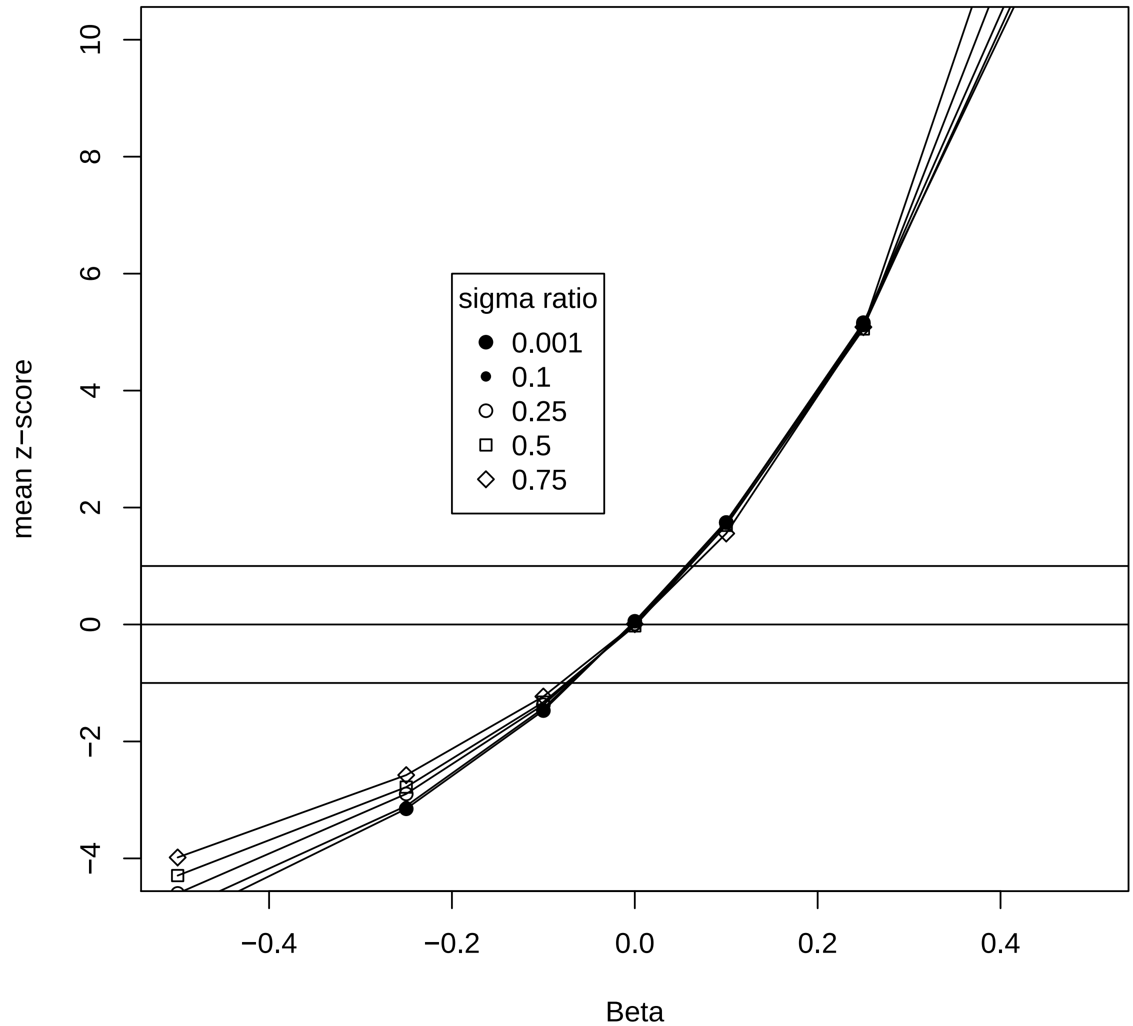

Figure 4.9: Simulation z-score results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station using typical area aggregation to estimate the exposures.

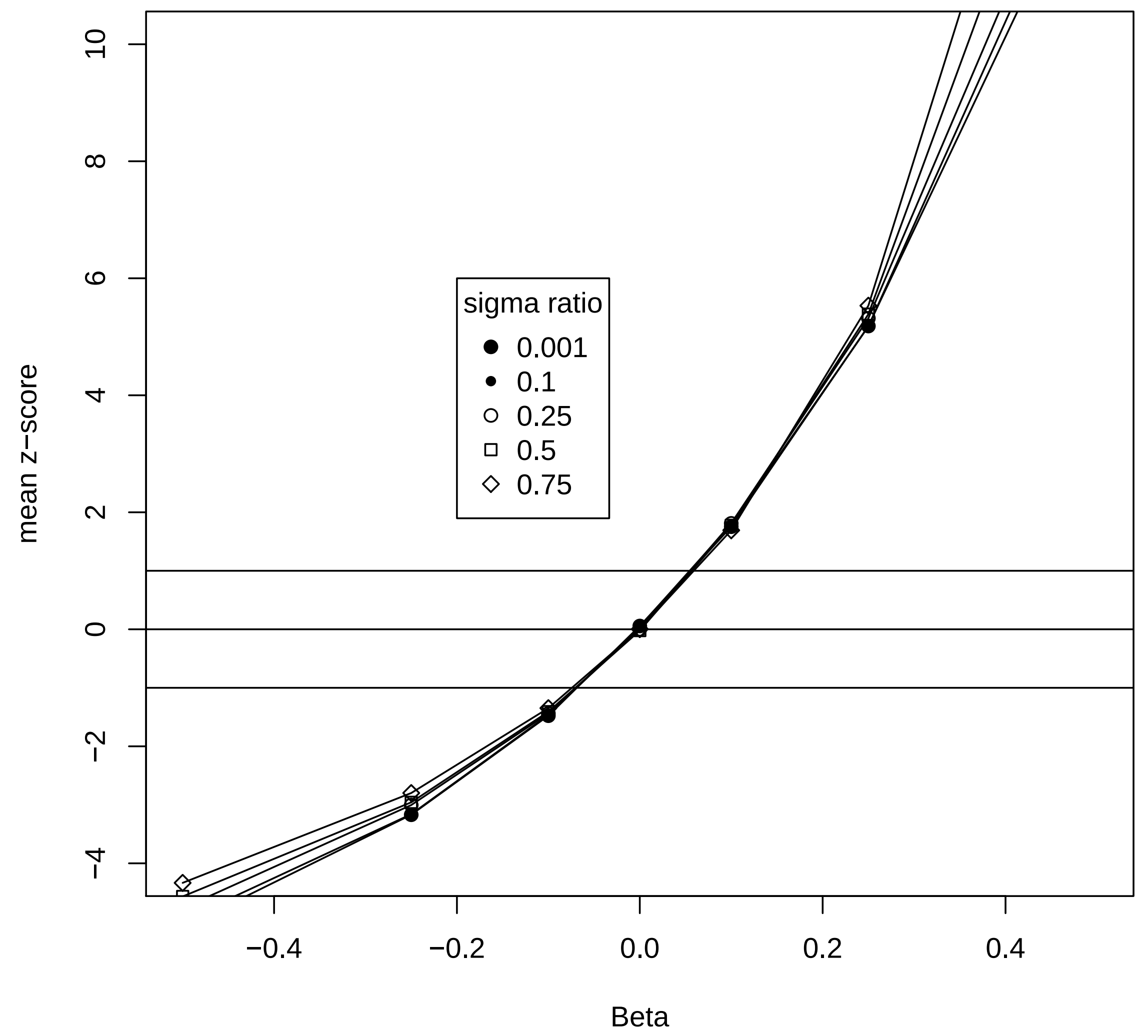

Figure 4.10: Simulation z-score results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station using the first order approximation to estimate the exposures.

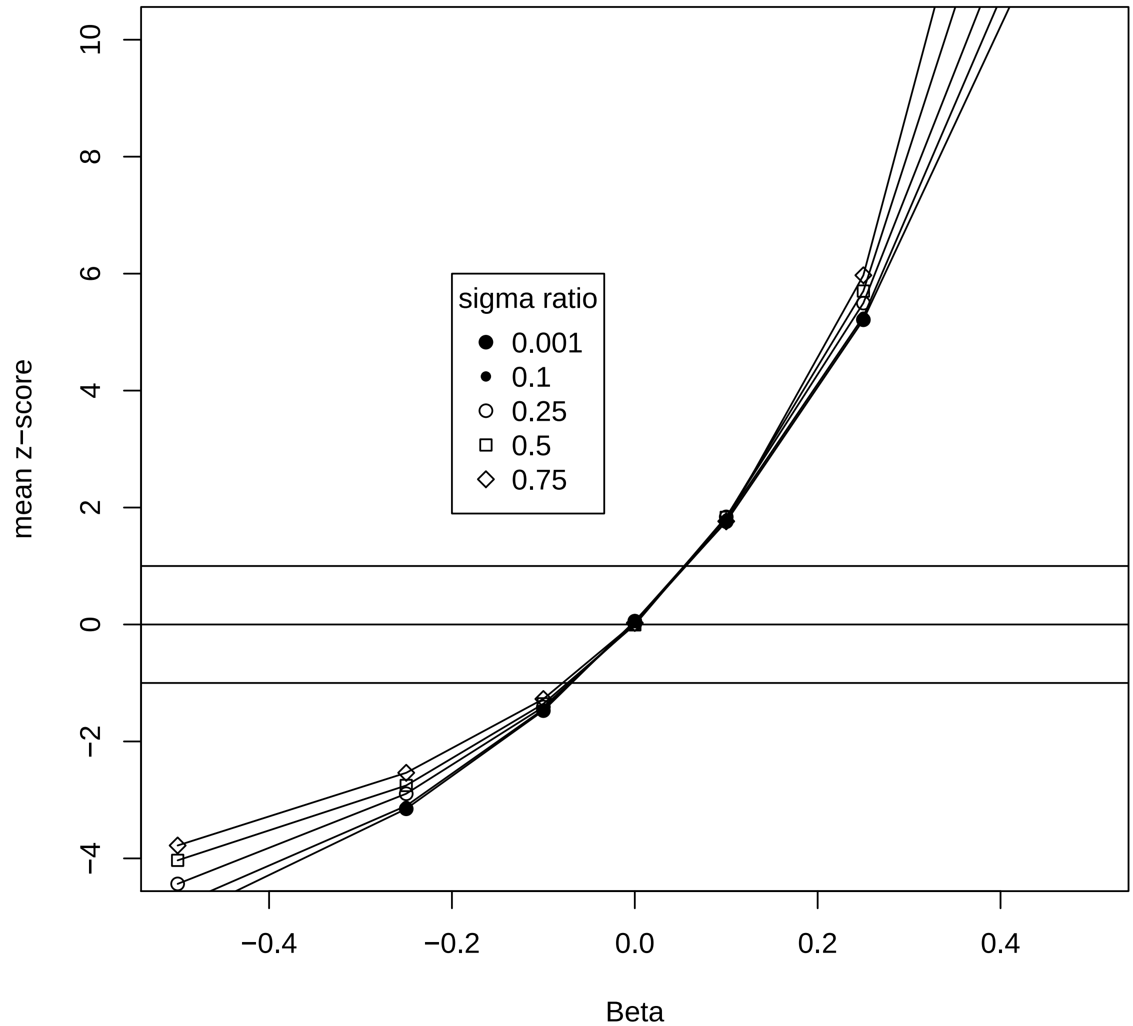

Figure 4.11: Simulation z-score results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station using the second order approximation to estimate the exposures.

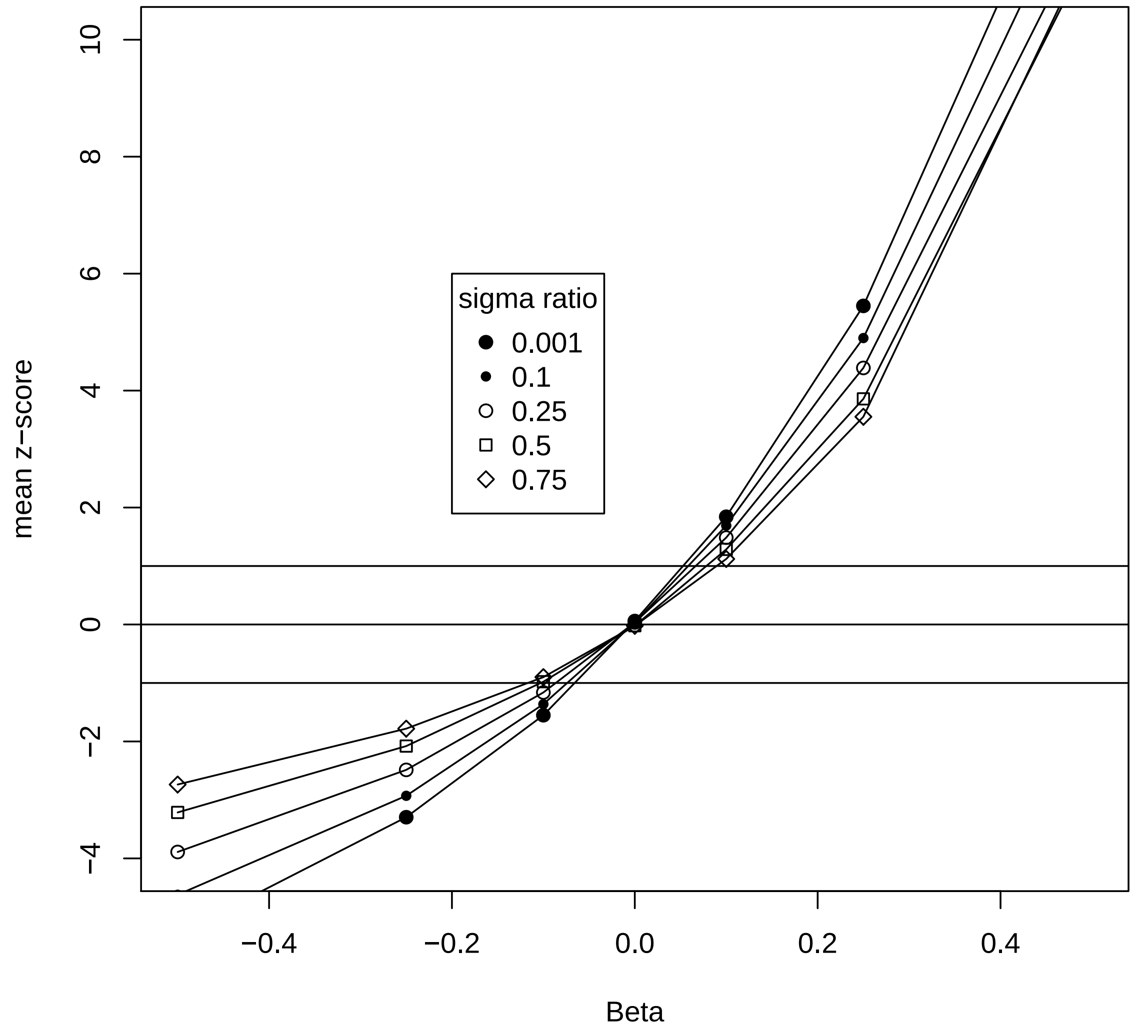

Figure 4.12: Simulation z-score results using the fused data subset of North Carolina under the condition that the data available are equivalent to a monitoring station using the non aggregated data at the sub region level.

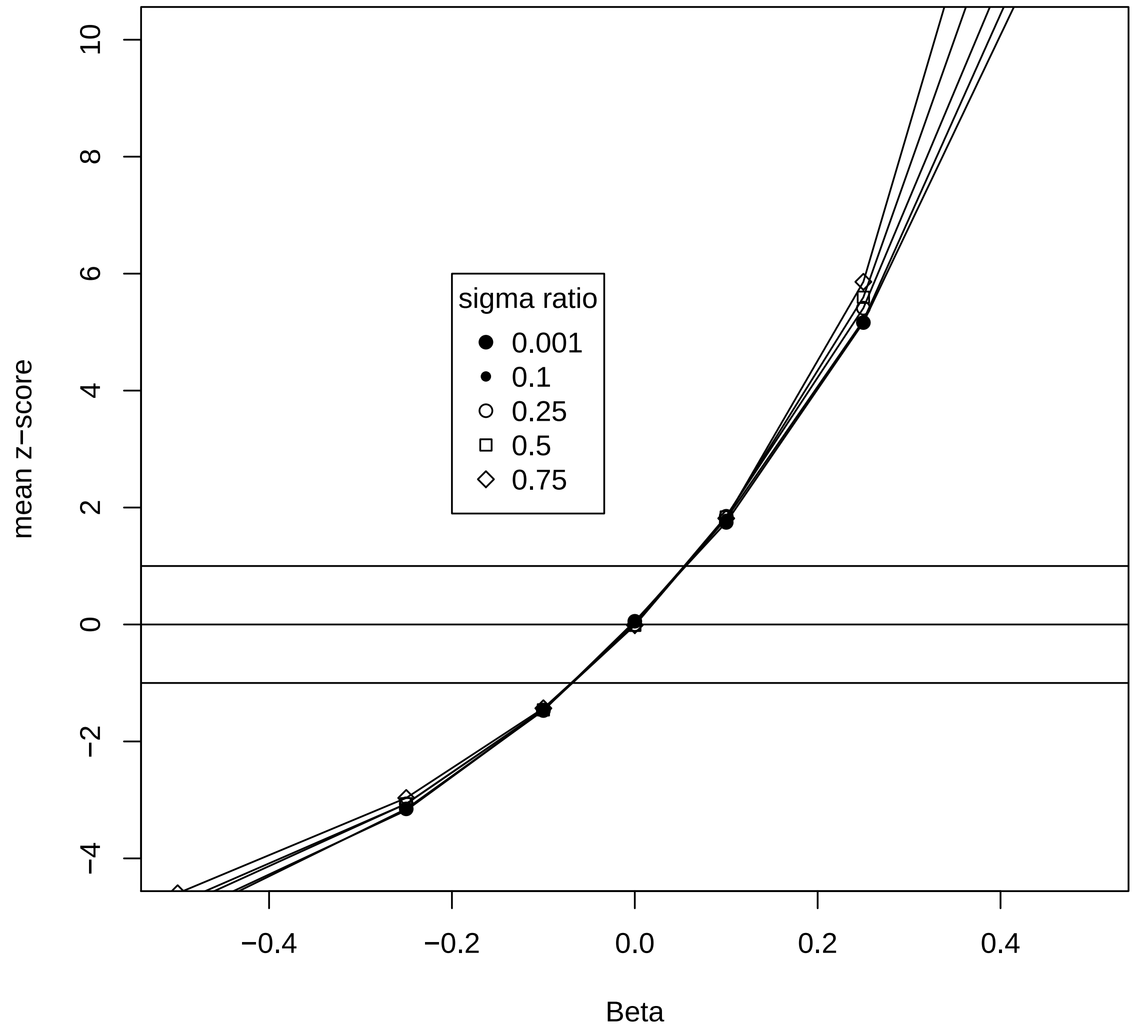

Figure 4.13: Simulation z-score results using the fused data subset of North Carolina under the condition that the data available are the true means of the subregions using typical area aggregation to estimate the exposures.

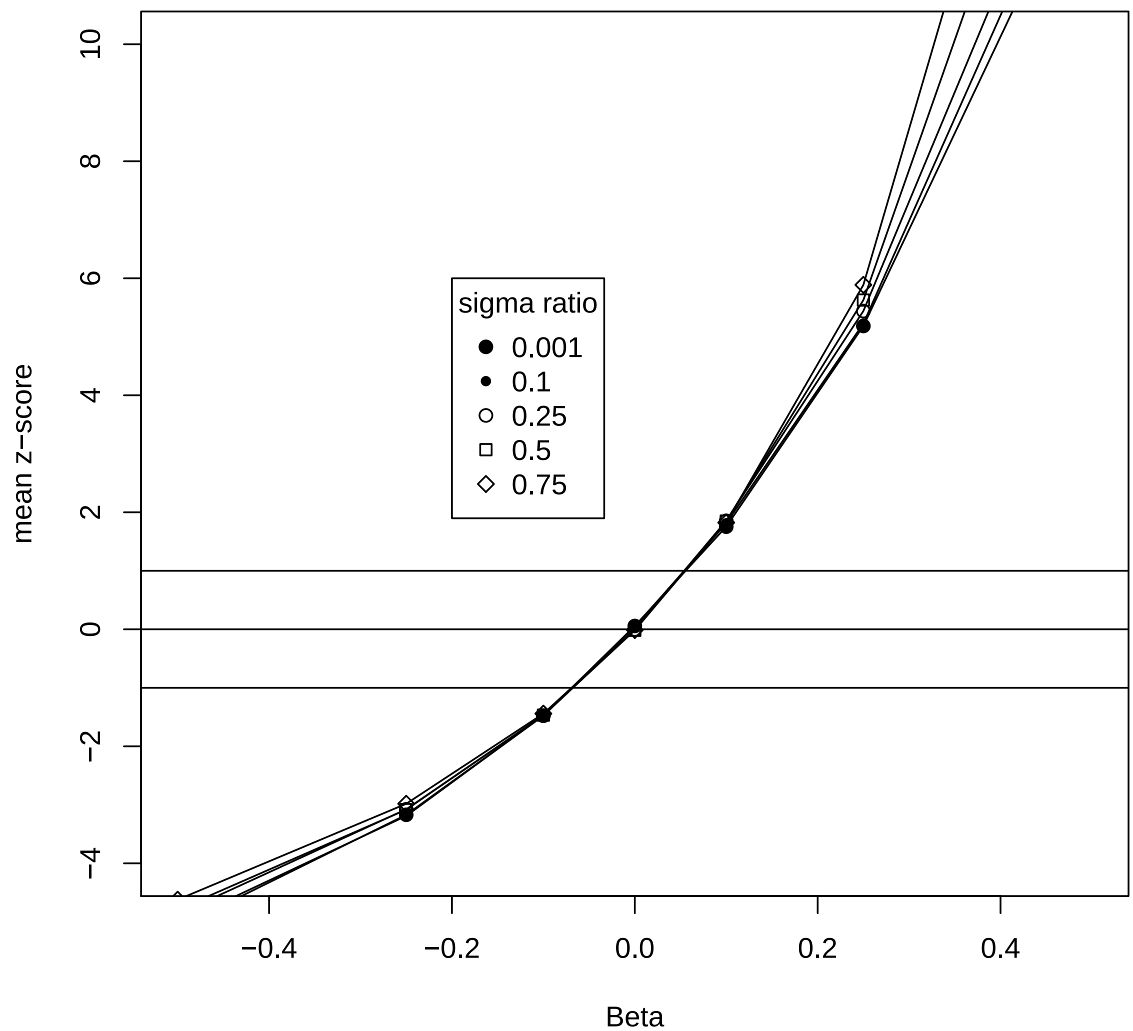

Figure 4.14: Simulation z-score results using the fused data subset of North Carolina under the condition that the data available are the true means of the subregions using the first order approximation to estimate the exposures.

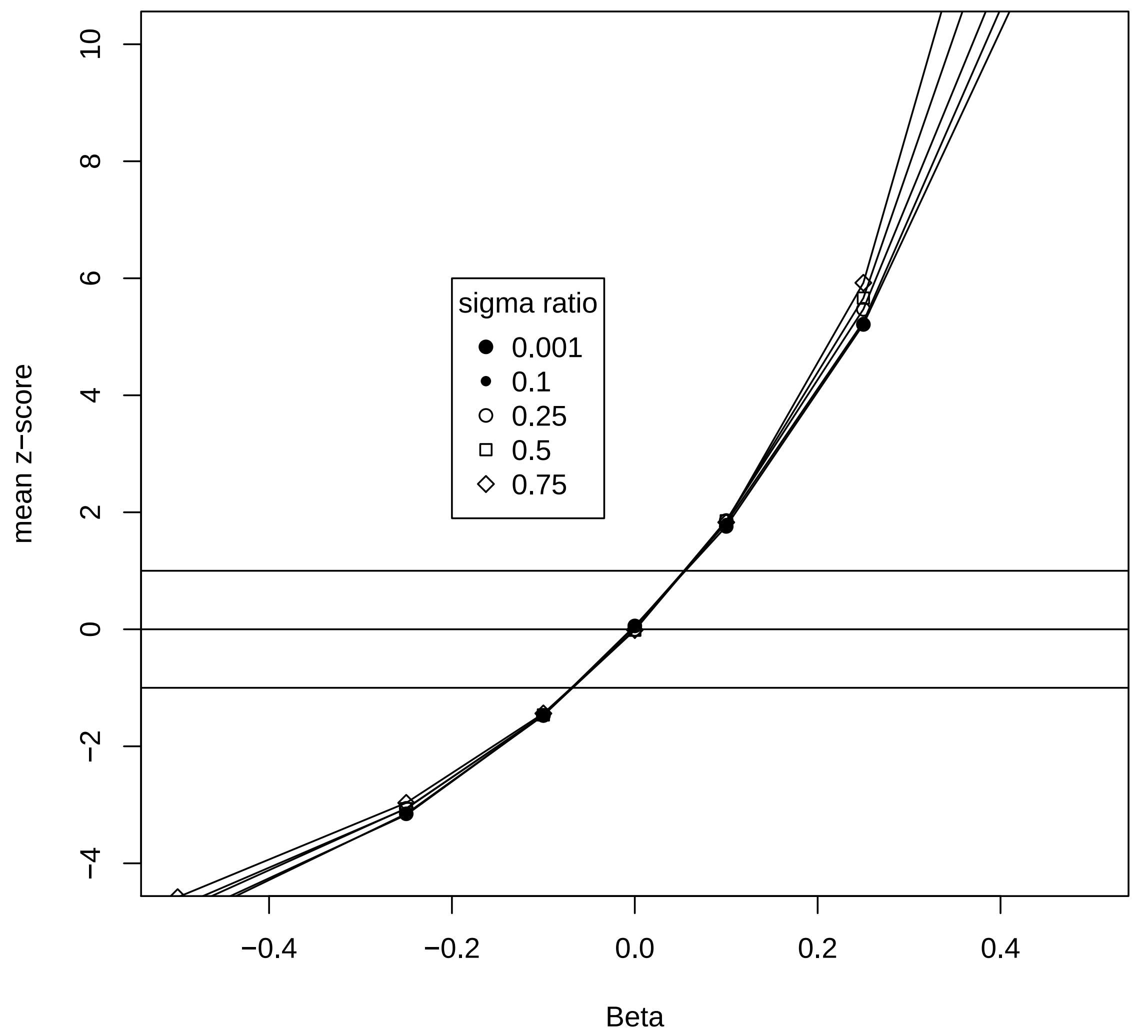

Figure 4.15: Simulation z-score results using the fused data subset of North Carolina under the condition that the data available are the true means of the subregions using the second order approximation to estimate the exposures.

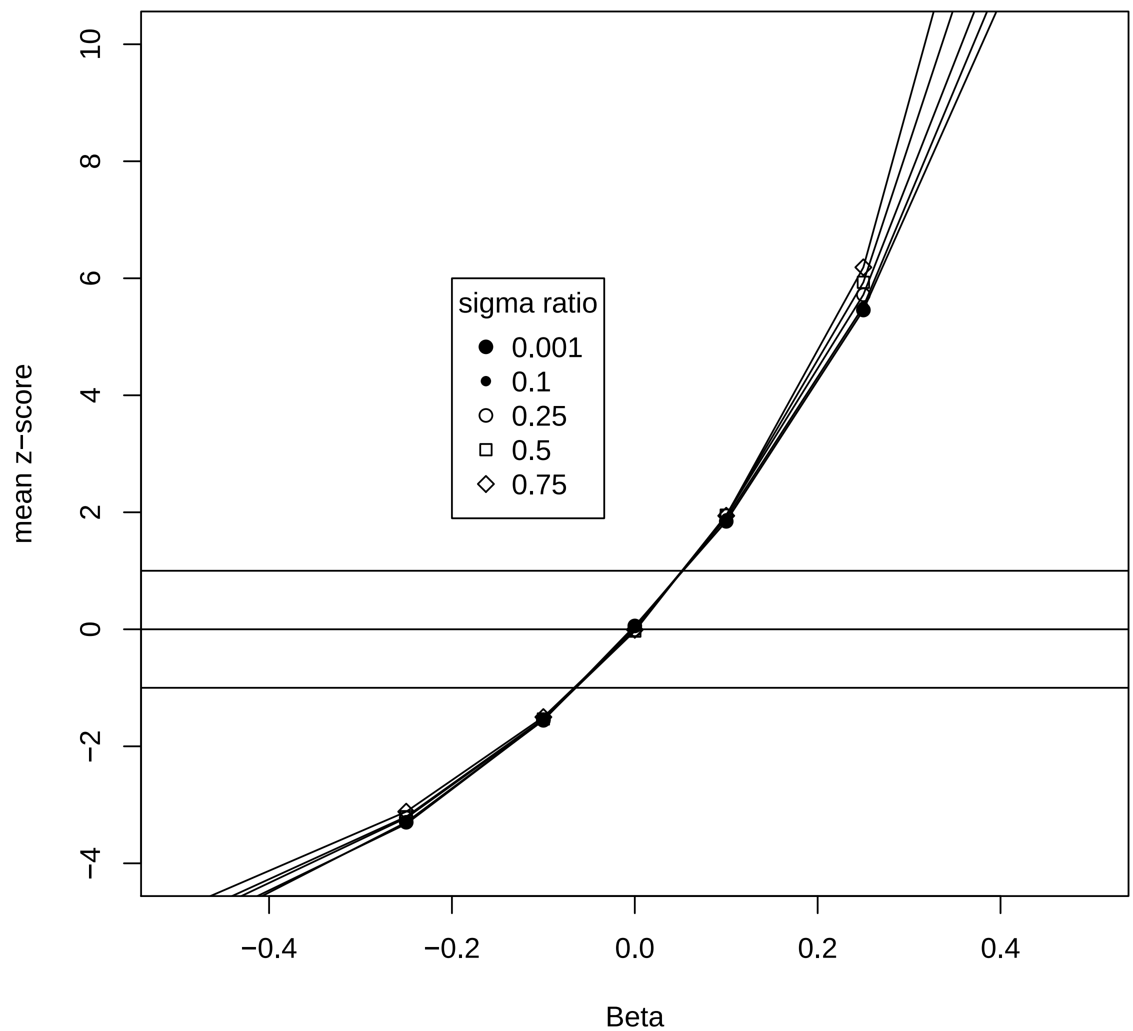

Figure 4.16: Simulation z-score results using the fused data subset of North Carolina under the condition that the data available are the true means of the subregions using the non aggregated data at the sub region level.

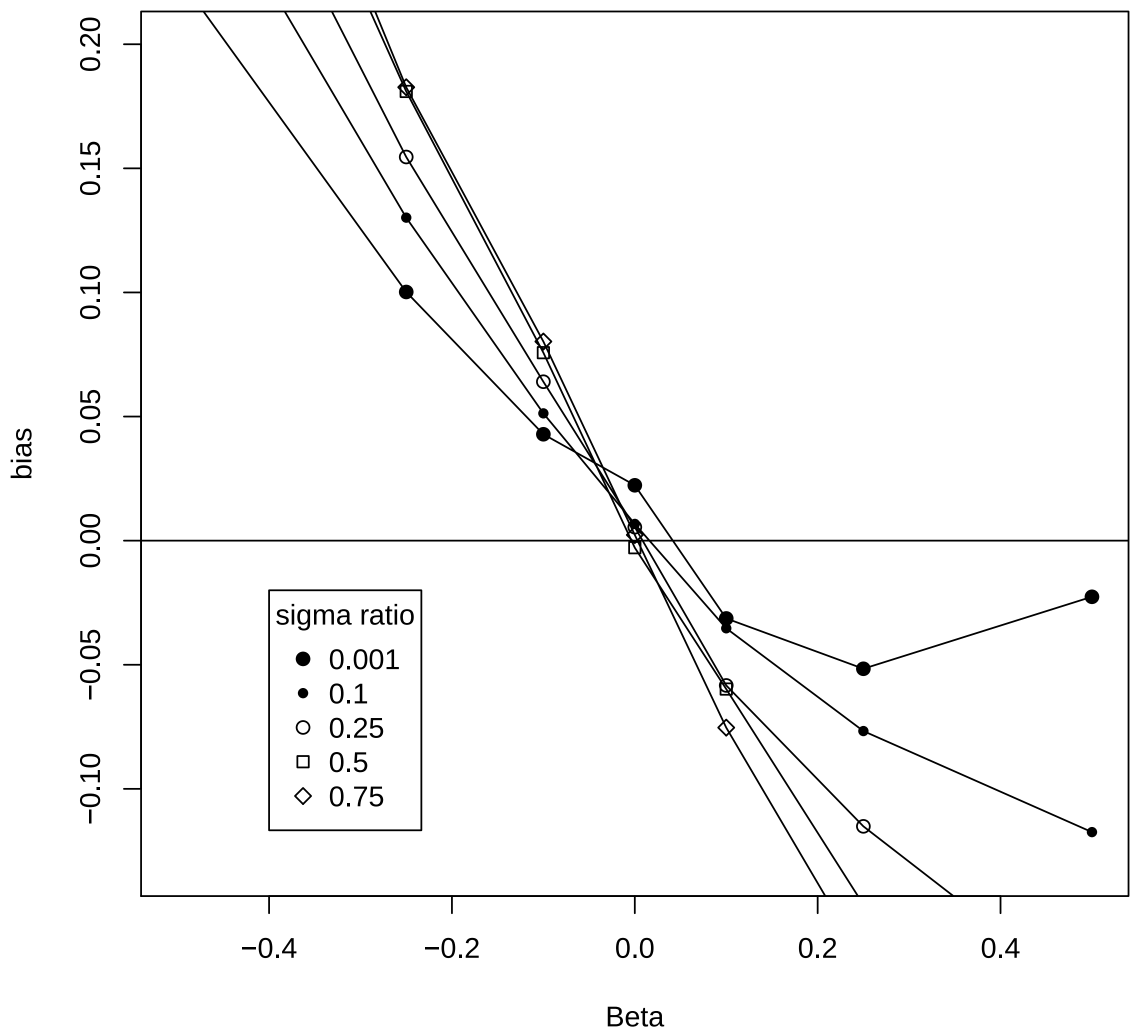

Figure 4.17: Simulation bias results using the fused across the United States under the condition that the data available are equivalent to a monitoring station using typical area aggregation to estimate the exposures.

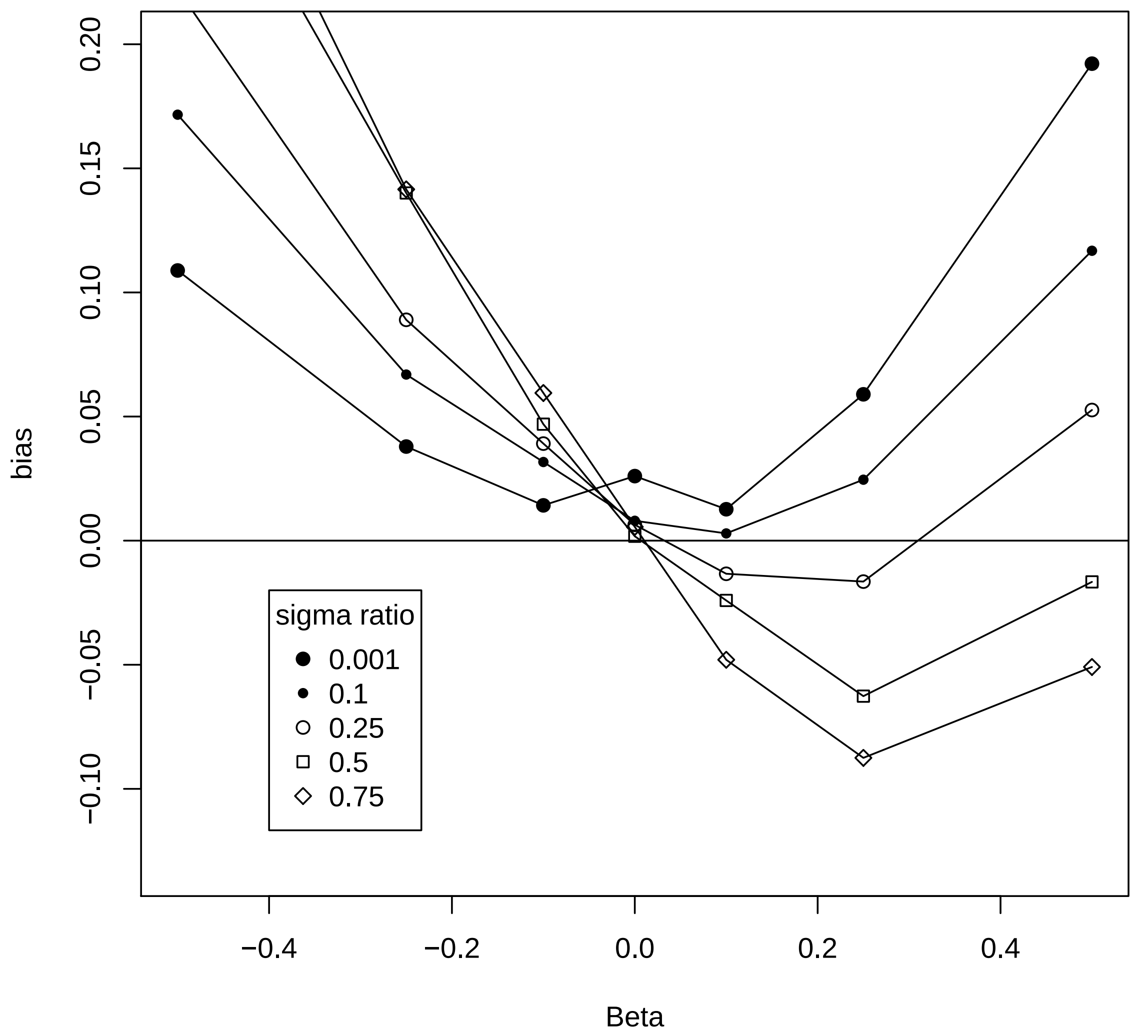

Figure 4.18: Simulation bias results using the fused across the United States under the condition that the data available are equivalent to a monitoring station using the first order approximation to estimate the exposures.

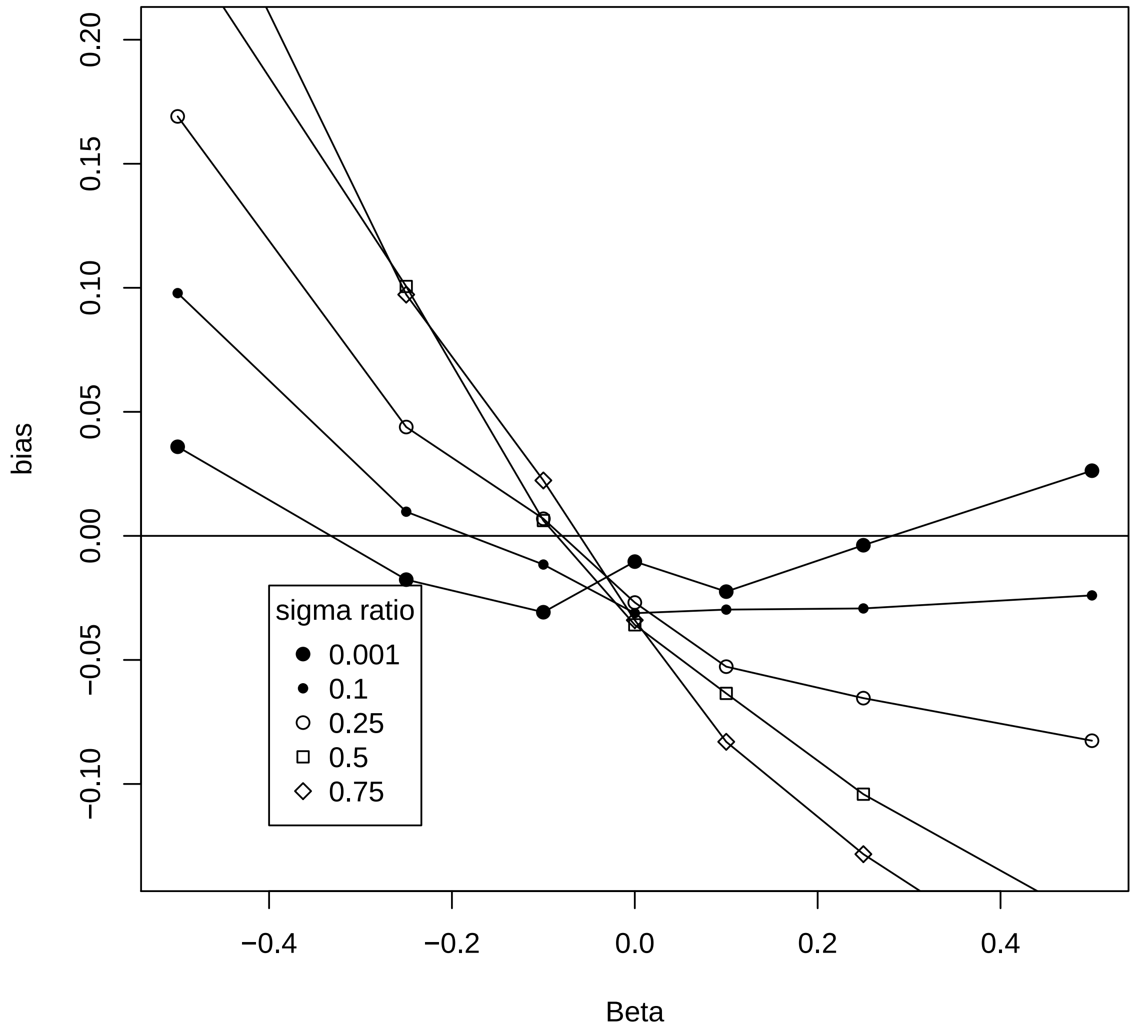

Figure 4.19: Simulation bias results using the fused across the United States under the condition that the data available are equivalent to a monitoring station using the second order approximation to estimate the exposures.

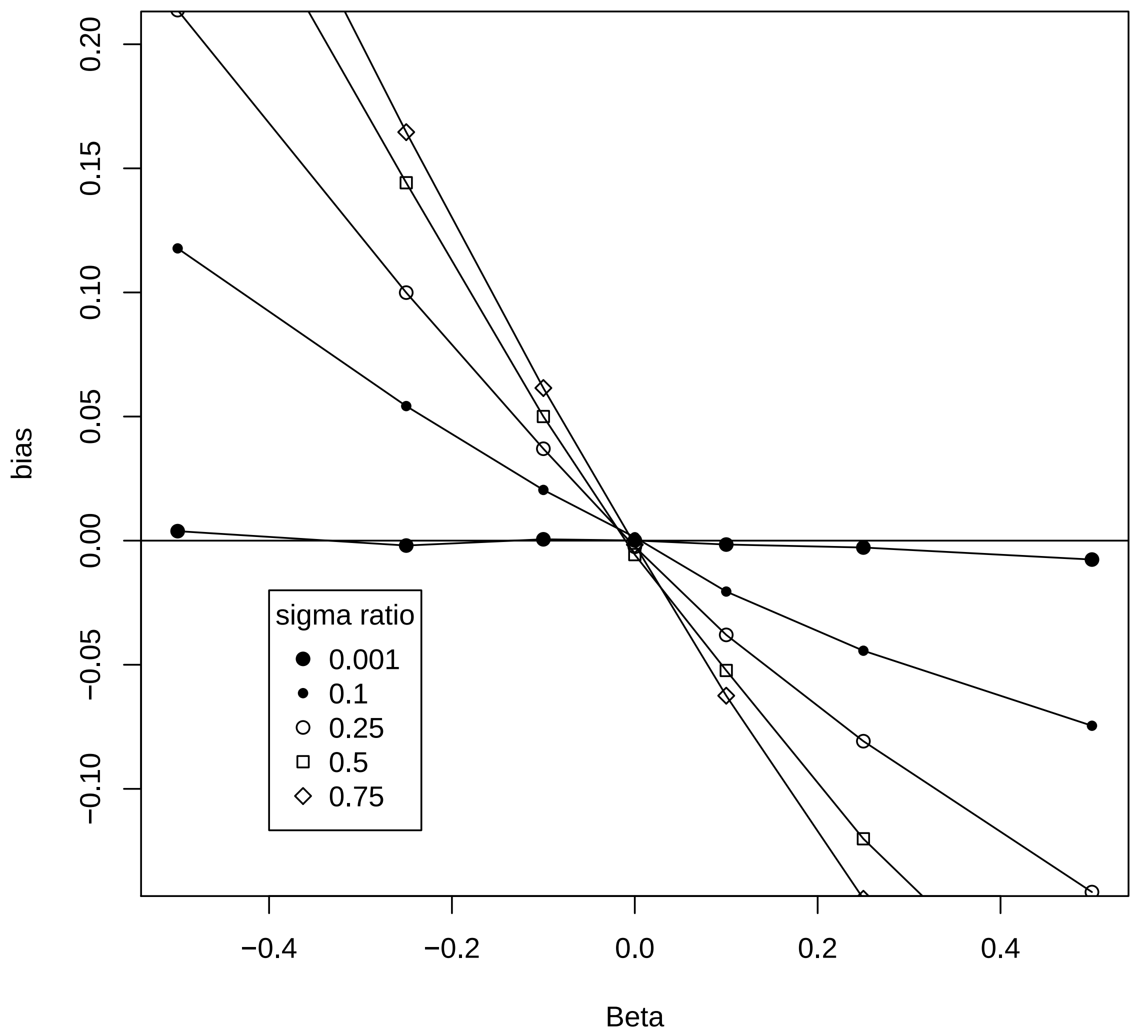

Figure 4.20: Simulation bias results using the fused across the United States under the condition that the data available are equivalent to a monitoring station using the non aggregated data at the sub region level.

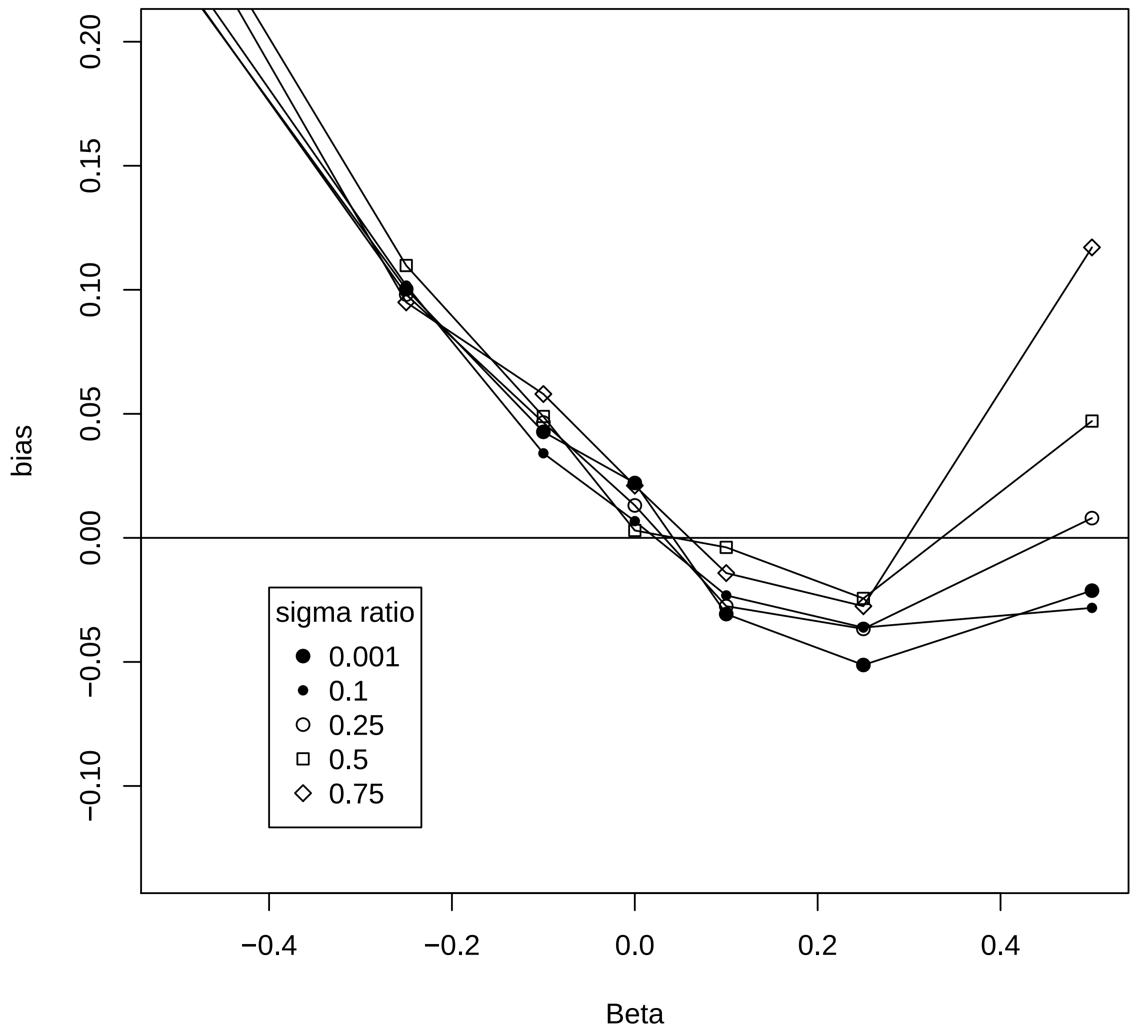

Figure 4.21: Simulation bias results using the fused across the United States under the condition that the data available are the true means of the subregions using typical area aggregation to estimate the exposures.

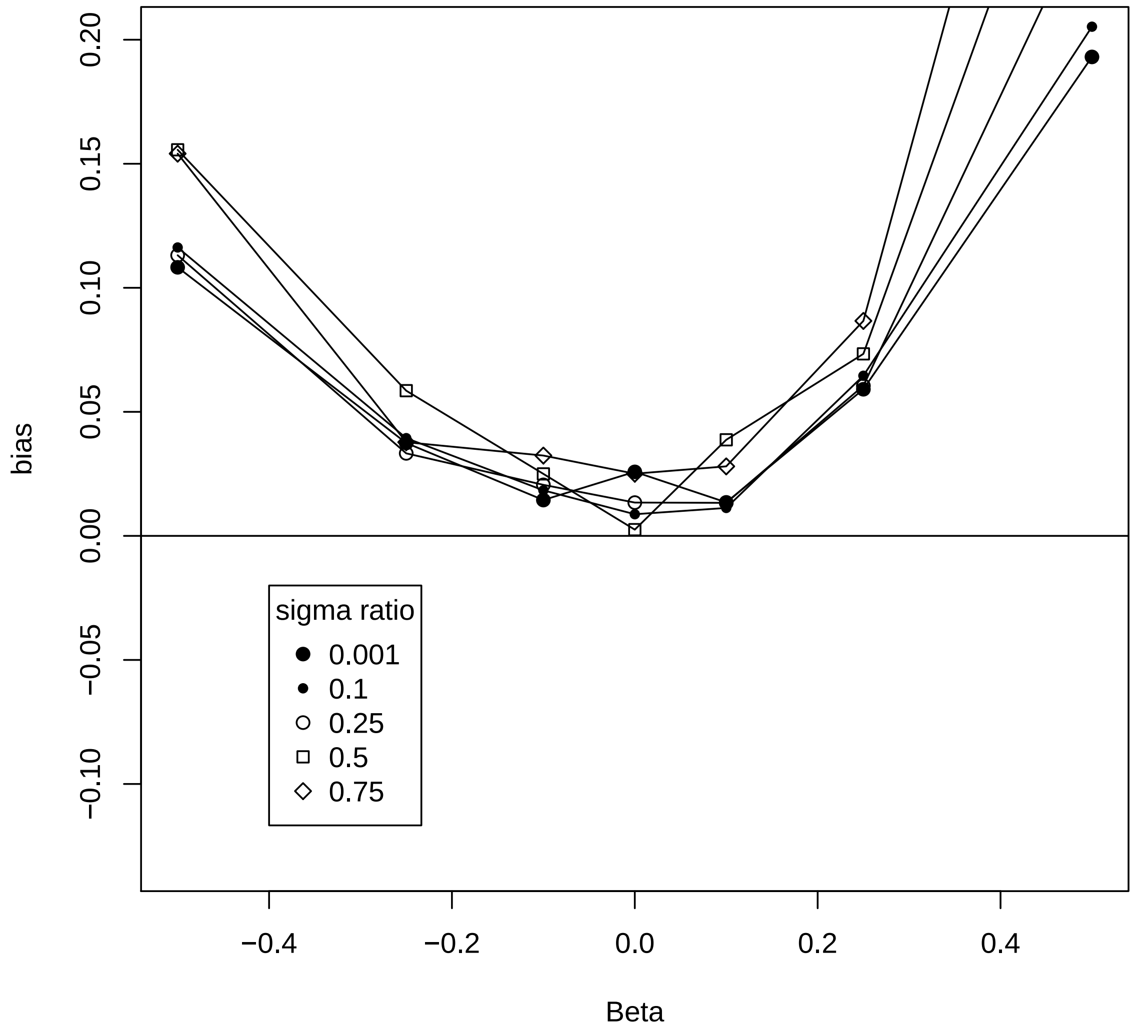

Figure 4.22: Simulation bias results using the fused across the United States under the condition that the data available are the true means of the subregions using the first order approximation to estimate the exposures.

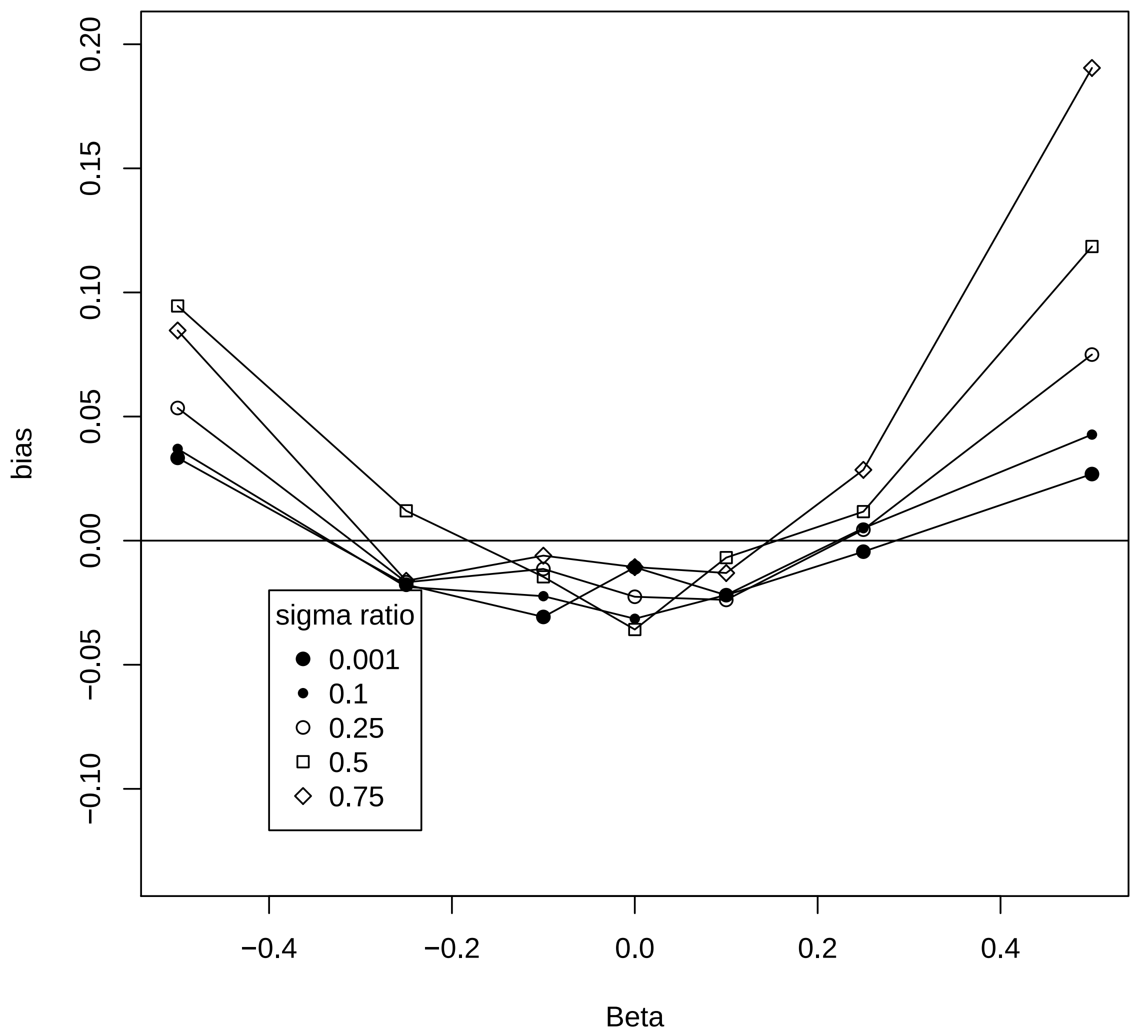

Figure 4.23: Simulation bias results using the fused across the United States under the condition that the data available are the true means of the subregions using the second order approximation to estimate the exposures.

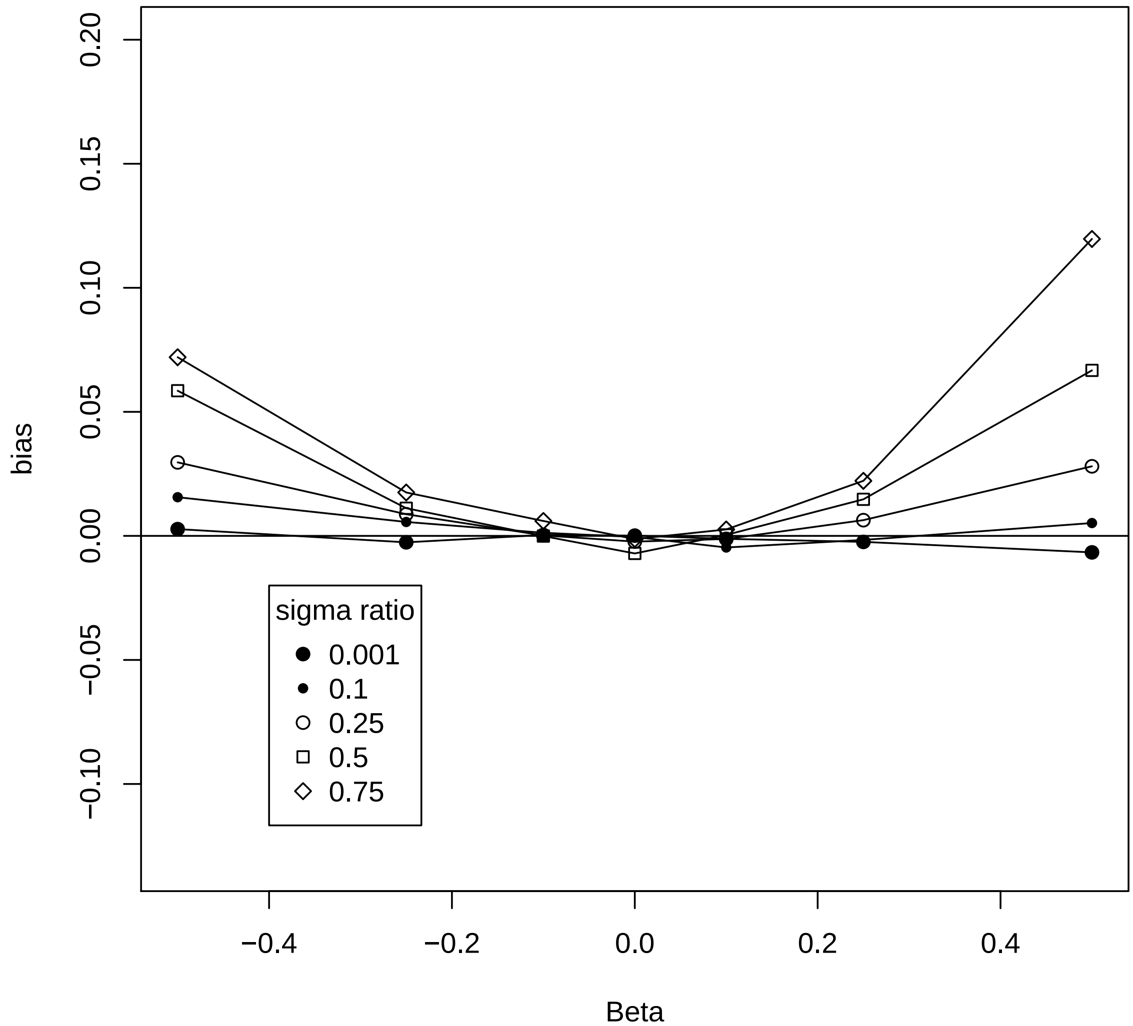

Figure 4.24: Simulation bias results using the fused across the United States under the condition that the data available are the true means of the subregions using the non aggregated data at the sub region level.

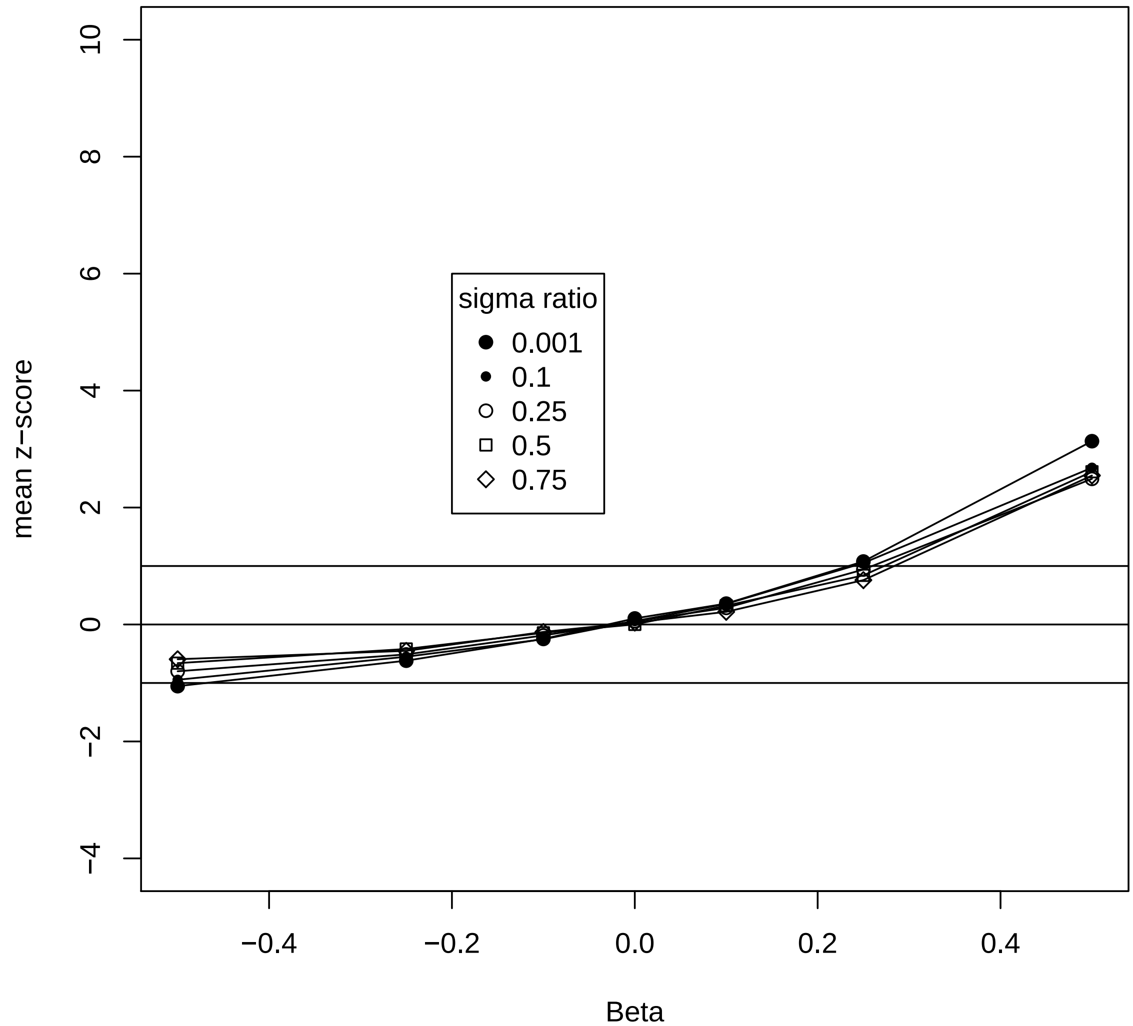

Figure 4.25: Simulation z-score results using the fused across the United States under the condition that the data available are equivalent to a monitoring station using typical area aggregation to estimate the exposures.

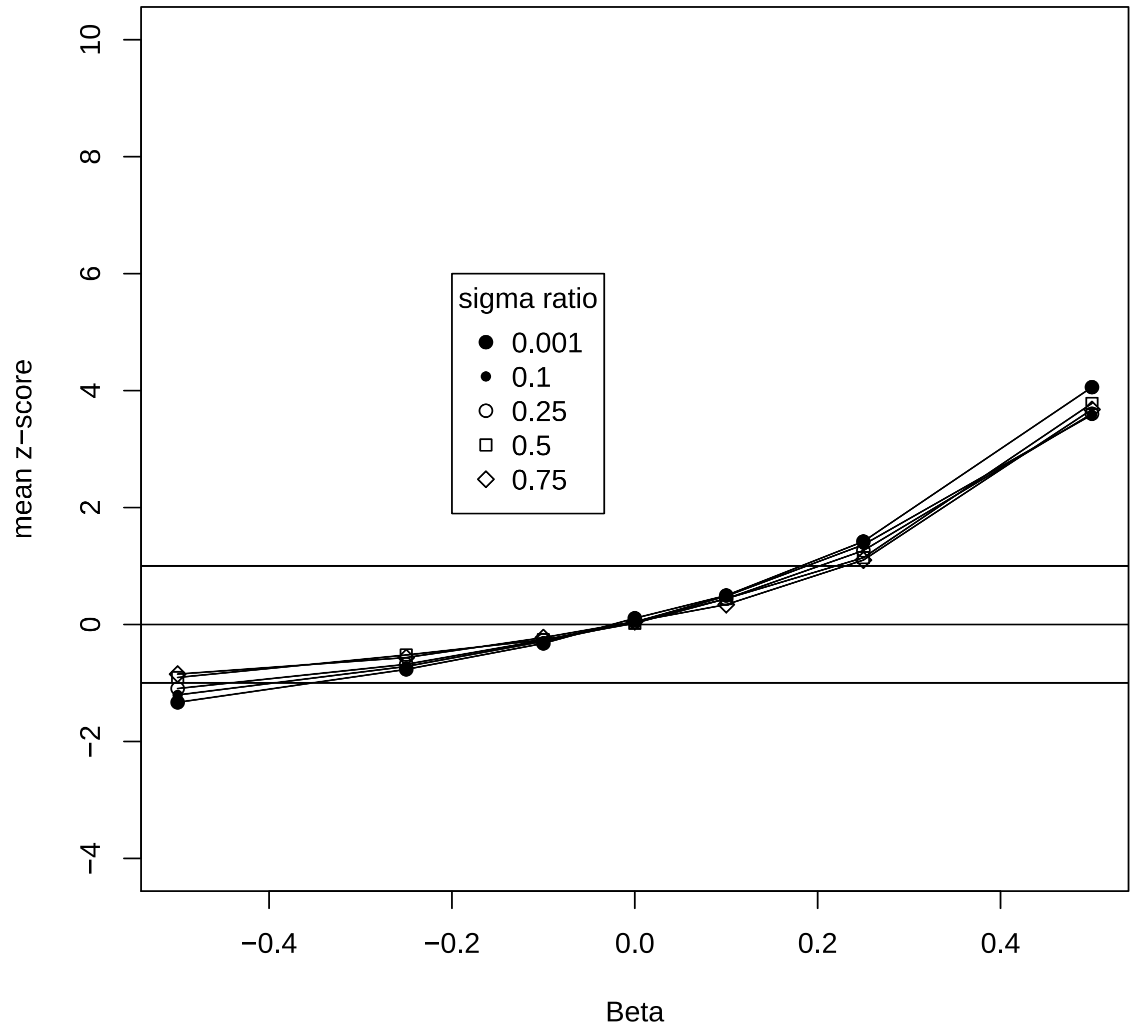

Figure 4.26: Simulation z-score results using the fused across the United States under the condition that the data available are equivalent to a monitoring station using the first order approximation to estimate the exposures.

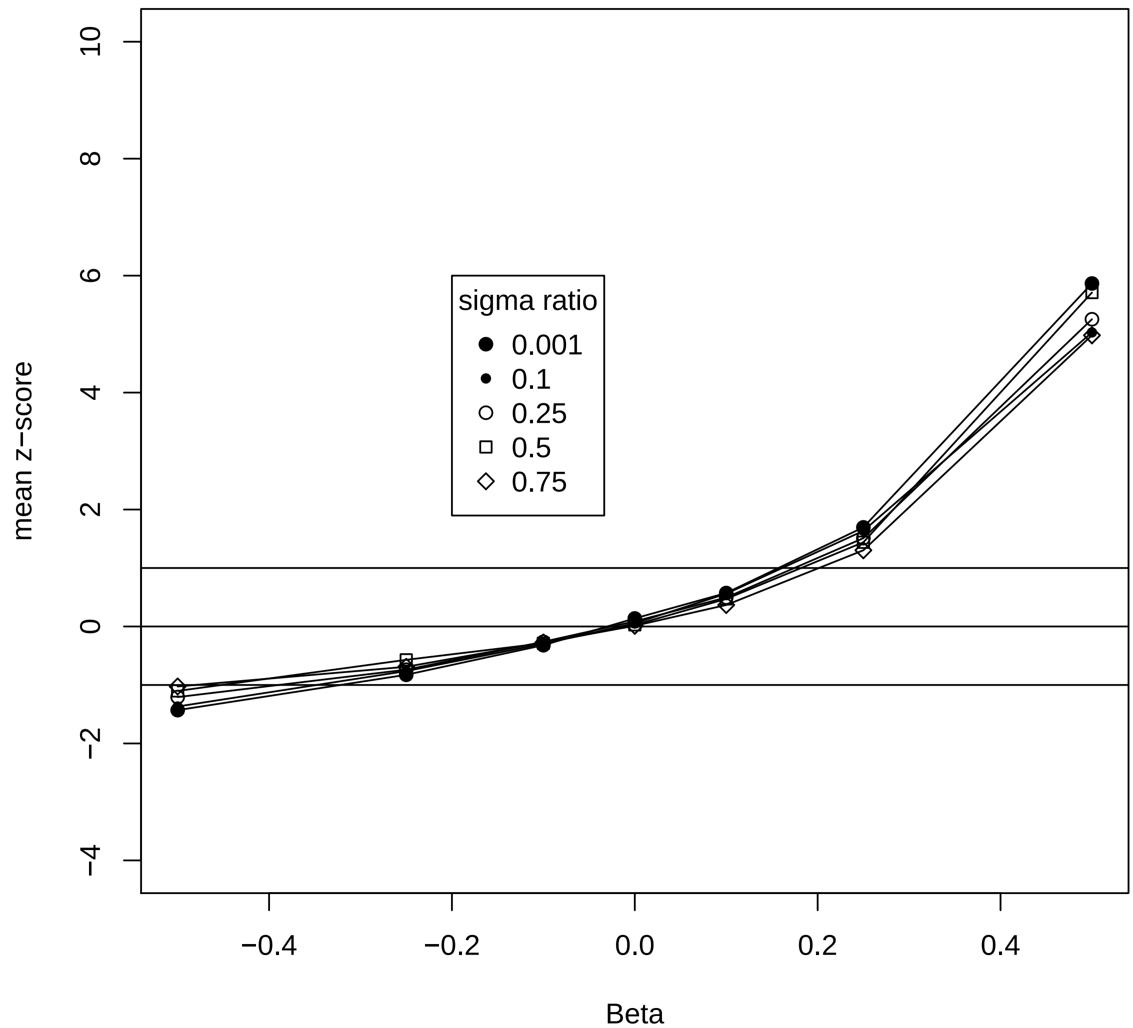

Figure 4.27: Simulation z-score results using the fused across the United States under the condition that the data available are equivalent to a monitoring station using the second order approximation to estimate the exposures.

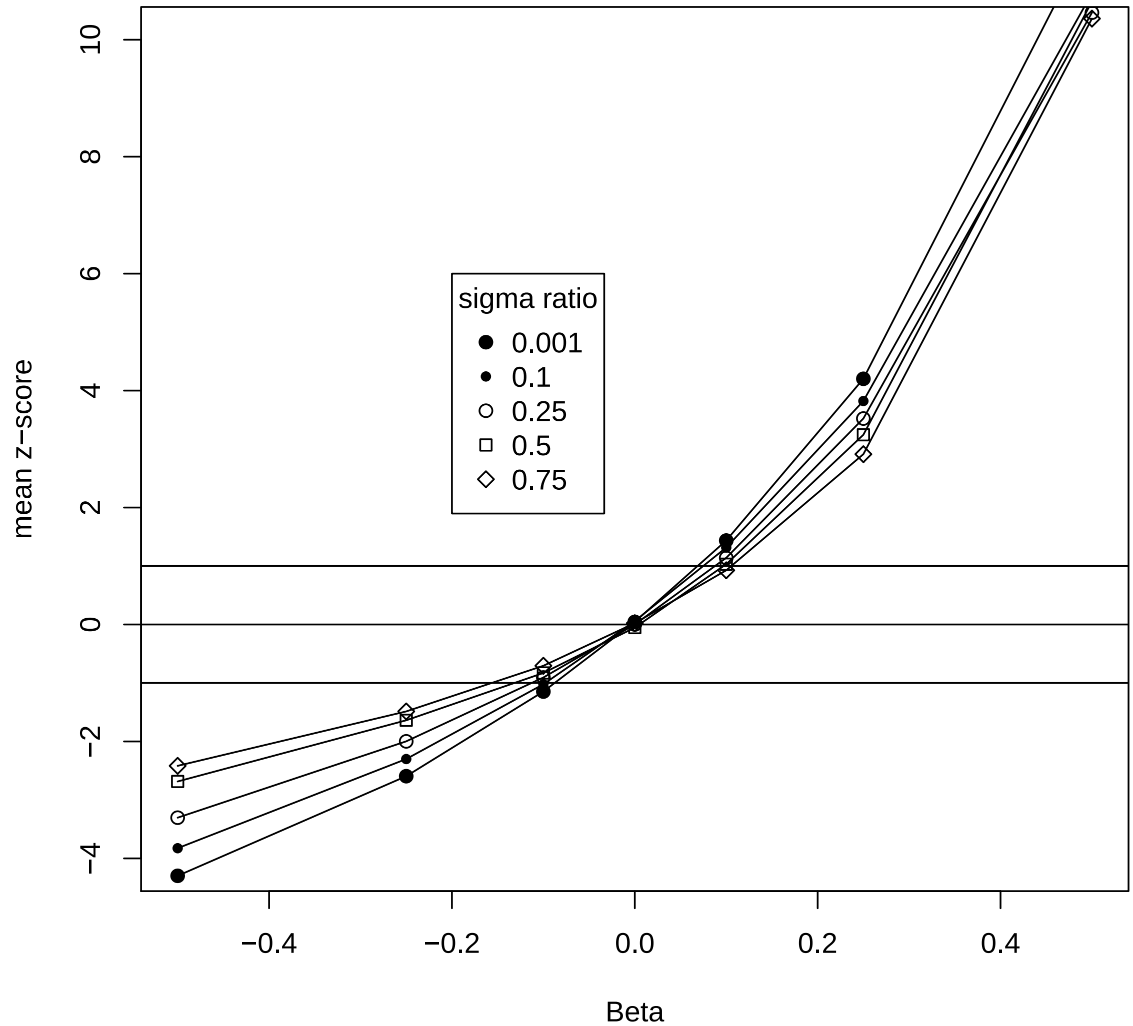

Figure 4.28: Simulation z-score results using the fused across the United States under the condition that the data available are equivalent to a monitoring station using the non aggregated data at the sub region level.

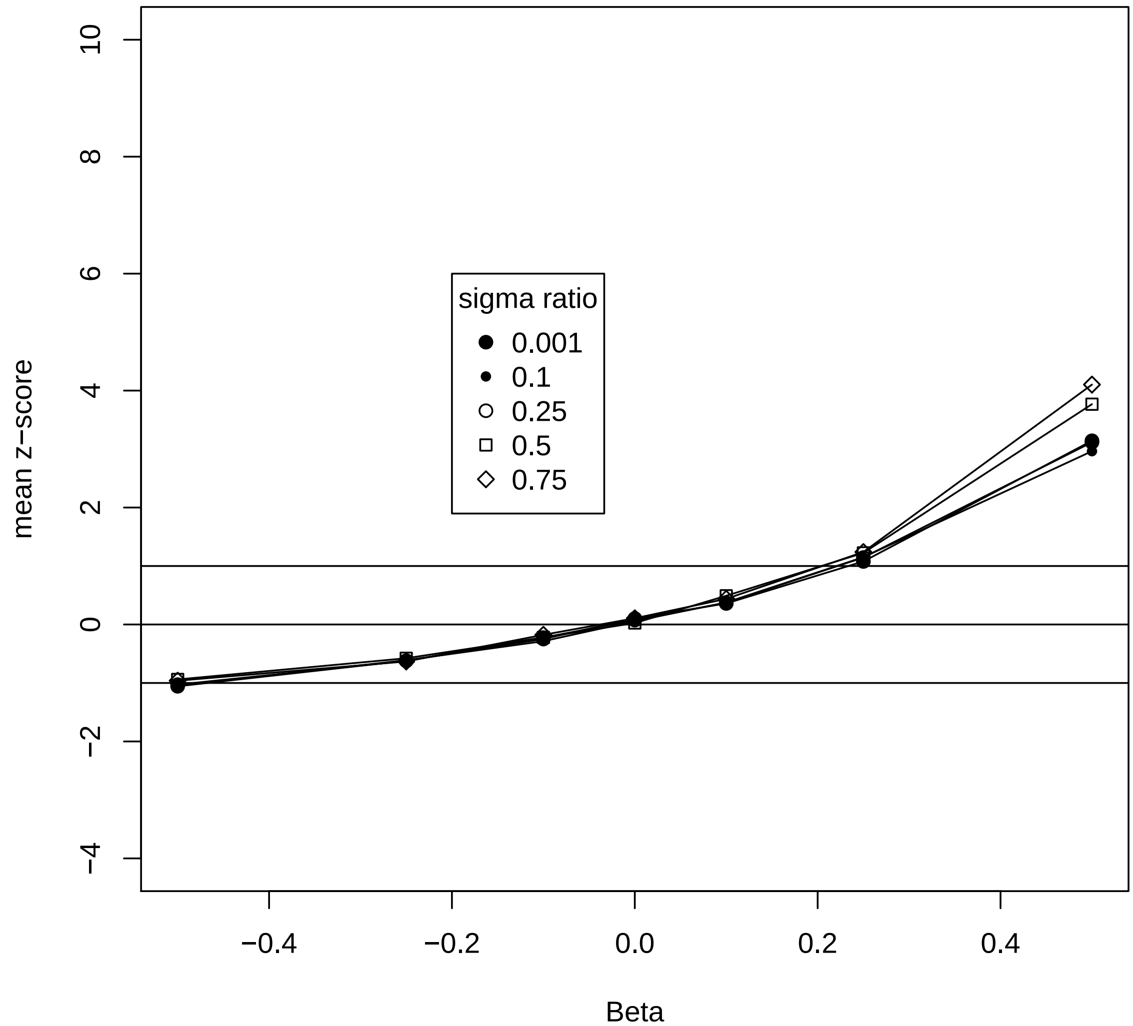

Figure 4.29: Simulation z-score results using the fused across the United States under the condition that the data available are the true means of the subregions using typical area aggregation to estimate the exposures.

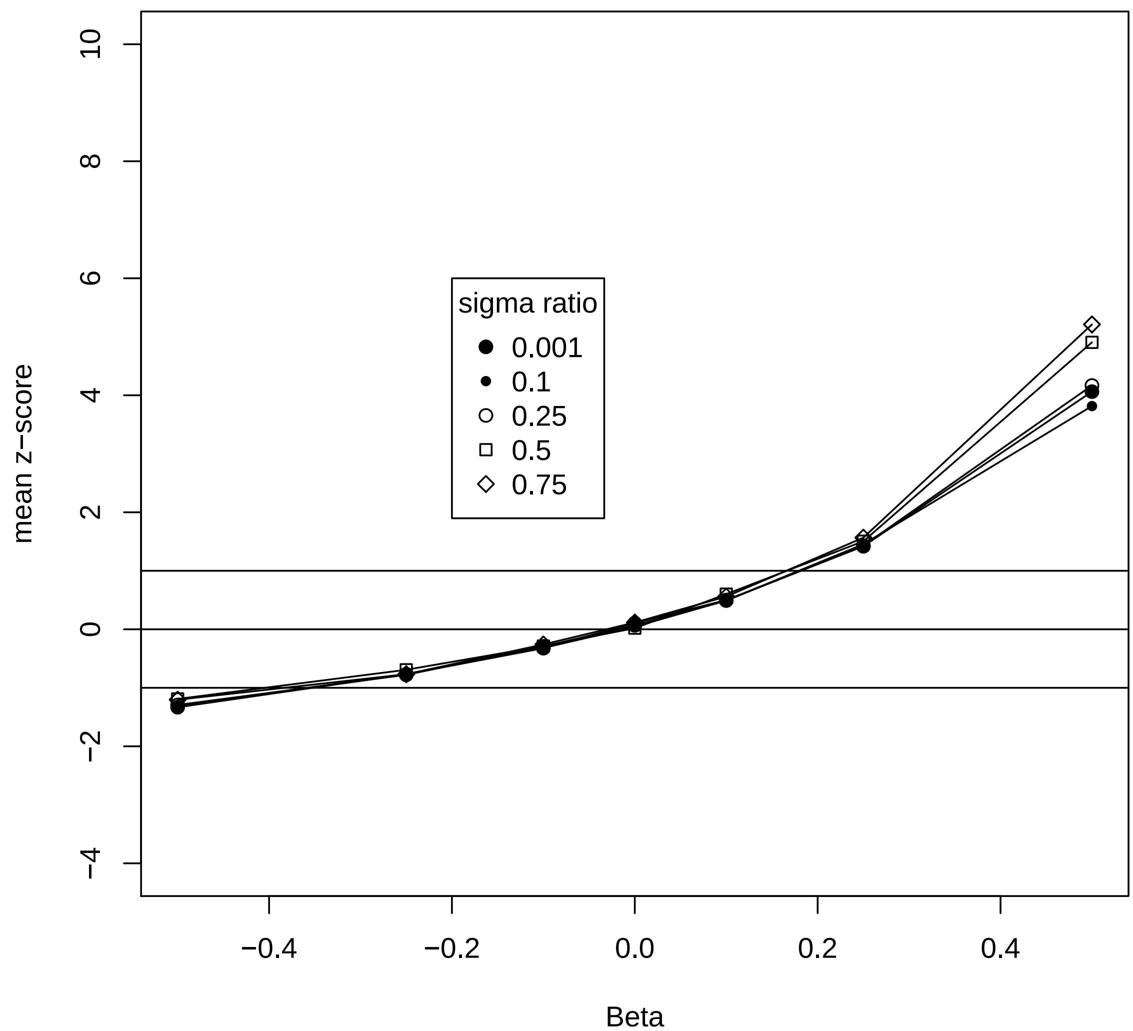

Figure 4.30: Simulation z-score results using the fused across the United States under the condition that the data available are the true means of the subregions using the first order approximation to estimate the exposures.

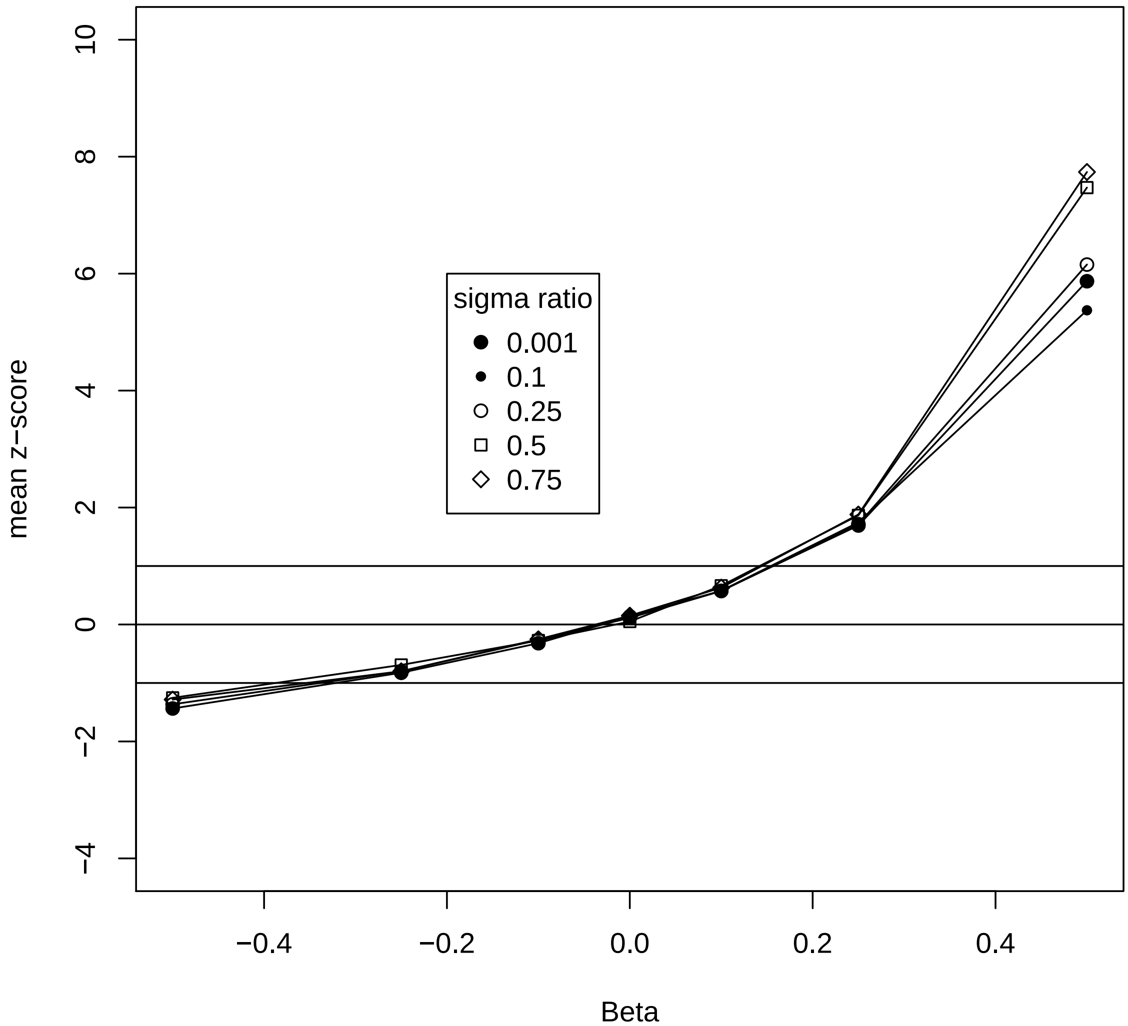

Figure 4.31: Simulation z-score results using the fused across the United States under the condition that the data available are the true means of the subregions using the second order approximation to estimate the exposures.

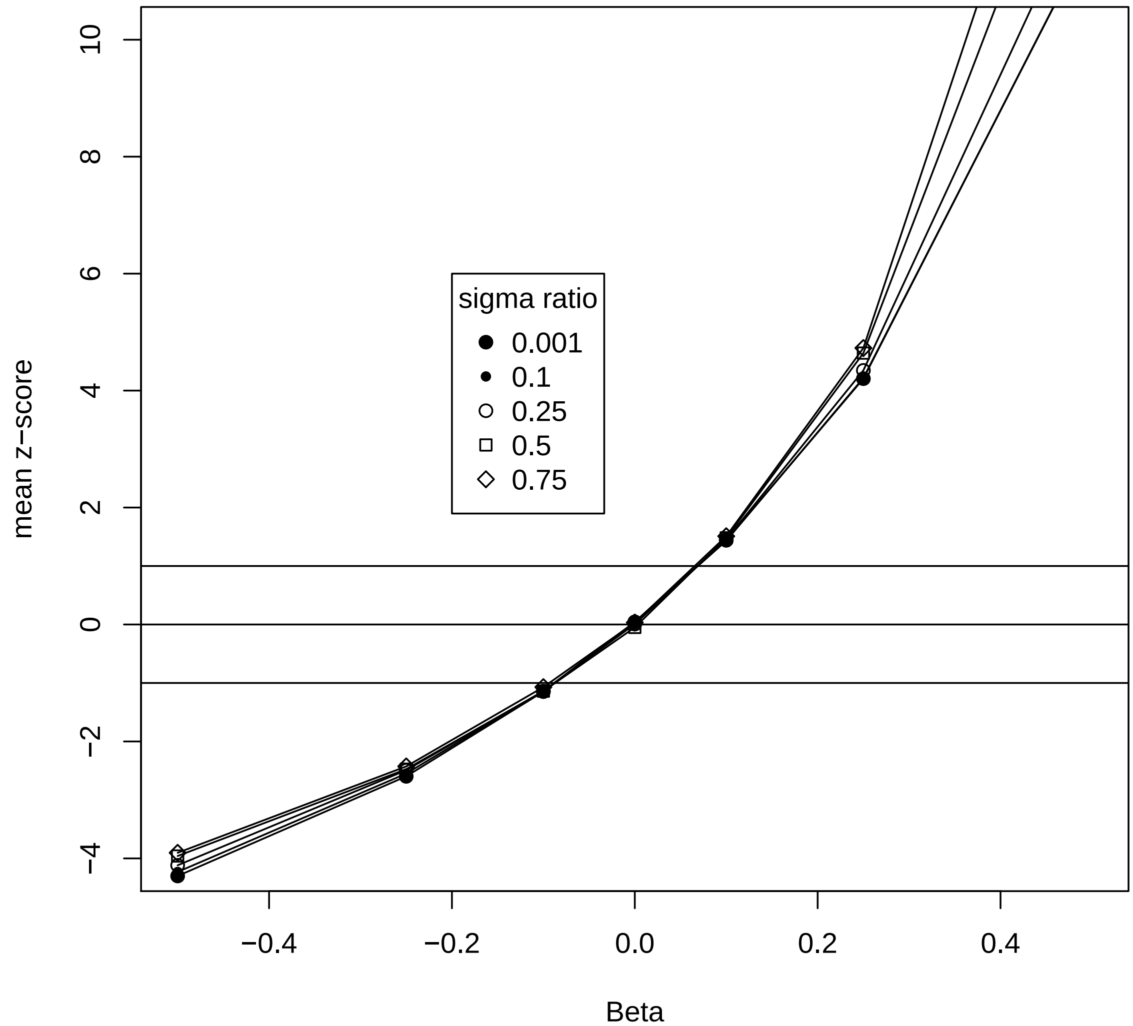

Figure 4.32: Simulation z-score results using the fused across the United States under the condition that the data available are the true means of the subregions using the non aggregated data at the sub region level.

REFERENCES

American Thoracic Society, and Bascom, R. (1996a). Health effects of outdoor air pollution. Part 1. *American Journal of Respiratory and Critical Care Medicine* **153**, 3-50.

American Thoracic Society, and Bascom, R. (1996b). Health effects of outdoor air pollution. Part 2. *American Journal of Respiratory and Critical Care Medicine* **153**, 477-498.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical modeling and analysis for spatial data.* Chapman and Hall, New York.

Bates, D.V., Baker-Anderson, M., and Sizto, R. (1990). Asthma attack periodicity: A study of hospital emergency visits in Vancouver. *Environmental Research* **51**, 51-70.

Bell, M.L., McDermott, A., Zeger, S.L., Samet, J.M., and Dominici, F. (2005). Ozone and Short-Term Mortality in 95 US Urban Communities, 1987-2000. *Journal of American Medical Association* **292**, 2371-2378.

Besag J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological* **36**, 192-236.

Binkowski, F.S. and Roselle, S.J. (2003). Models-3 community multiscale air quality (CMAQ) model aerosol component, 1. Model description. J Geophys Res 108:4183. doi:10.1029/2001JD001409

Burke, J., Zufall, M., and Ozkaynak, H. (2001). A population exposure model for particulate matter: case study results for $PM_{2.5}$ in Philadelphia, PA. *Journal of Exposure*

Analysis and Environmental Epidemiology **11**, 470-489.

Burnett, R., and Krewski, D. (1994). Air pollution effects of hospital admission rates: A random effects modelling approach. *The Canadian Journal of Statistics* **22**, 441-458.

Byun, D.W. and Schere, K.L. (2006). Review of the governing equations, computational algorithms and other components of models-3 community multiscale air quality (CMAQ) modeling system. Appl Mech Rev 59:51-77. ENVIRON, 2006, CAMx User's Guide, ENVIRON International Corporation, Novato, CA.

Callaghan, W., MacDorman, M., Rasmussen, S., Qin, C., and Lackritz, E. (2006). The contribution of preterm birth to infant mortality rates in the United States. *Pediatrics* **118**, 1566-1573.

Choi, J., Fuentes, M., and Reich, B. (2009). Spatial-temporal association between fine particulate matter and daily mortality. *Journal of Computational Statistics and Data Analysis* **53**, 2989-3000.

Collins, J.W., and David, R.J. (2004). Pregnancy outcome of mexican-american women: The effect of generational residence in the United States. *Ethnicity & Disease* **14**, 317-321.

Daniels, M., Dominici, F., Samet, J.M., and Zeger, S.L. (2000). Estimating PM_{10}-mortality dose-response curves and threshold levels: An analysis of daily time series for the 20 largest U.S. cities. *American Journal of Epidemiology* **152**, 397-412.

Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., and Speizer, F.E. (1993). An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* **329**, 1753-1759.

Dockery, D.W., Schwartz, J., and Spengler, J.D. (1992). Air pollution and daily mortality: Associations with particulates and acid aerosols. *Environmental Research* **59**, 362-373.

Dolwick, P. D., Jang, C., Possiel, N., Timin, B. Gipson, G., and Godowitch, J. (2001). Summary of results from a series of Models-3/CMAQ simulations of ozone in the Western United States. 94th Annual A&WMA Conference and Exhibition, Orlando, FL.

Dominici, F., Daniels, M., Zeger, S., and Samet, J. (2002). Air pollution and mortality: Estimating regional and national doseresponse relationships. *Journal of the American Statistical Association* **97**, 100111.

Dominici, F., McDermott, A., Zeger, S.L., and Samet, J.M. (2003). National maps of the effects of particulate matter on mortality: Exploring geographical variation. *Environmental Health Perspectives* **111**, 39-43.

Dominici, F., Samet, J.M., and Zeger, S.L. (2000). Combining evidence on air pollution and daily mortality from the twenty largest U.S. cities: A hierarchical modeling strategy, (with discussion). *Journal of the Royal Statistical Society, Series A* **163**, 263-302.

Efroymson, R. E., and Murphy, D. L. (2001). Ecological Risk Assessment of Multimedia Hazardous Air Pollutants: Estimating Exposure and Effects. *Sci Total Environ* **274**, 219-230.

English, P.B., Kharrazia, M., Daviesb, S., Scalfb, R., Wallerc, L., and Neutraa, R. (2003). Changes in the spatial pattern of low birth weight in a souther California county: the role of individual and neighborhood level factors. *Social Science & Medicine*, **56**, 2073-2088.

Fuentes, M. and Raftery, A. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with output from numerical models. *Biometrics* **61**, 36-45.

Fuentes, M., Song, H., Ghosh, S.K., Holland, D.M., and Davis, J.M. (2006). Spatial association between speciated fine particles and mortality. *Biometrics* **62**, 855-863.

Gelfand A.E., Diggle P.J., Fuentes M., and Guttorp P. (2010). *Handbook of Spatial Statistics*. Chapman and Hall, Florida.

Gordon, D. (1995). Census based deprivation indices: their weighting and validation. *Journal of Epidemiology and Community Health* **49**, S39-S44.

Gouveia, N. and Fletcher, T. (2000). Respiratory diseases in children and outdoor air pollution in Sao Paulo, Brasil: A time series analysis. *Occupational and Environmental Medicine* **57**, 477483.

Greeland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine* **11**, 1209-1223.

Greenland, S. and Robinds, J. (1994). Ecological studies-biases, misconceptions and counter examples. *American Journal Epidemiology* **139**, 747-760.

Hobel, C., and Culhane, J. (2003). Role of psychosocial and nutritional stress on poor pregnancy outcome. *Journal of Nutrition* **133**, 1709S - 1717S.

Jin, X., Carlin, B.P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics* **61**, 950-961.

Johnson, T.R., Capel, J., Paul, R., Wijnberg, L. (1992). Estimation of carbon monoxide exposure and associated carboxyhemoglobin level in Denver residents using a probabilistic version of NEM. Prepared by IT/Air Quality Services for US EPA, RTP, NC.

Jolliffe, I.T. (2002). *Principal component analysis*. Springer.

Katsouyanni, K., Touloumi, G., Spix, C., Schwartz, J., Balducci, F., Medina, S., Rossi, G., Wojtniak, B., Sunyer, J., Bacharova, L., Schouten, J., Ponka, A., and Anderson, H. (1997). Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities. *British Medical Journal* **314**, 1658-1663.

Kelsall, J.E., and Wakefield, J.C. (1999). *Discussion of Bayesian models for spatially correlated disease and exposure data*. Oxford: Oxford University Press.

Kramer, M., Demissie, K., and Yang, H. (2000). The contribution of mild and moderate preterm birth to infant mortality. *Journal of the American Medical Association* **284**, 843-849.

Lee D. and Shaddick G. (2010). Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics*. In press.

Leslie, J.C., Diehl, S.J., and Galvin, S.L. (2005). A comparison of birth outcomes among US-born and non US-born Hispanic women in North Carolina. *Maternal and Child Health Journal* **10**.

Ma, S. (2008). Paternal race/ethnicity and birth outcomes. *American Journal of Public Health* **98**, 2285-2292.

Mathews, T.J., Menacker, F., and MacDorman, M.F. (2004). Infant mortality statistics from the 2002 period linked birth/infant death data set. *National Vital Statistics Reports* **53**.

McMillan, N.J., Holland, D.M., Morara, M., and Feng, J. (2009). Combining numerical model output and particulate data using Bayesian space-time modeling, *Environ-*

metrics, 48-65

Messer, L.C., Laraia, B.A., Kaufman, J.S., Eyster, J., Holzman, C., Culhane, J., Elo, I., Burke, J.G., and O'Campo, P. (2006). The development of a standardized neighbourhood deprivation index. *Journal of Urban Health* **83**, 1041-1062.

Messer, L.C., Vinikoor, L.C., Laraia, B.A., Kaufman, J.S., Eyster, J., Holzman, C., Culhane, J., Elo, I., Burke, J.G., and O'Campo, P. (2008). Socioeconomic domains and associations with preterm birth. *Social Science & Medicine* **67**, 1247-1257.

Norton, J.D. and Niu, X. (2009). Intrinsically autoregressive spatiotemporal models with application to aggregated birth outcomes. *Journal of the American Statistical Association* **104**, 638-649.

O'Campo, P., Xue, X., Wang, M.C., and Caughy, M. (1997). Neighborhood risk factors for low birth weight in Baltimore: a multilevel analysis. *American Journal of Public Health* **87**, 1113-1118.

Ostro, B.D., Lipsett, M.J., Wiener, M.B., and Selner, J.C. (1991). Asthmatic responses to airborne acid aerosols. *American Journal of Public Health* **81**, 694-702.

Ozone (2006). Air quality criteria for ozone and related photochemical oxidants (final). *U.S. Environmental Protection Agency, Washington, DC*, EPA/600/R-05/004aF-cF.

Peng, R.D., Welty L.J., and McDermott A. (2004). The National Morbidity, Mortality, and Air Pollution Study Database in R. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 44.

Pope, C.A., Dockery, D., and Schwartz, J. (1995a). Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation Toxicology* **47**, 1-18.

Pope, C.A., Thun, M.J., Namboodiri, M.M., Dockery, D., Evans, J., Speizer, F., and

Health, C.M. (1995b). Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal Respiratory and Critical Care Medicine* **151**, 669-674.

Reich, B., Fuentes, M., and Burke, J. (2008). Analysis of the effects of ultrafine particulate matter while accounting for human exposure. *Environmetrics* **20**, 131-146.

Salway, R. and Wakefield, J. (2008). A hybrid model for reducing ecological bias. *Biostatistics* **9**, 1-17.

Samet, J., Zeger, S., Dominici, F., Curriero, F., Coursac, I., Dockery, D., Schwartz, J., and Zanobetti, A. (2000a). *The National Morbidity, Mortality, and Air Pollution Study Part II: Morbidity and Mortality from Air Pollution in the United States.* Health Effects Institute, Cambridge, MA.

Samet, J., Zeger, S., Dominici, F., Dockery, D., and Schwartz, J. (2000b). *The National Morbidity, Mortality, and Air Pollution Study Part I: Methods and Methodological Issues.* Health Effects Institute, Cambridge, MA.

Schabenberger, O., and Gotway, C.A. (2004). *Statistical Methods for Spatial Data Analysis.* Chapman and Hall, Florida.

Schwartz, J., Slater, D., Larson, T.V., Pierson, W.E., and Koenig, J.Q. (1993). Particulate air pollution and hospital emergency room visits for asthma in Seattle. *The American review of respiratory disease* **147**, 826-831.

Schwartz, J. (1994). Air pollution and daily mortality: A review and meta analysis. *Environmental research* **64**, 36-52.

Schwartz, J. (1995). Air pollution and daily mortality in Birmingham, Alabama. *American Journal of Epidemiology* **137**, 1136-1147.

Schwartz, J., Litonjua, A., Suh, H., Verrier, M., Zanobetti, A., Syring, M., Nearing, B., Verrier, R., Stone, P., MacCallum, G., Speizer, F.E., and Gold, D.R. (2005). Traffic related pollution and heart rate variability in a panel of elderly subjects. *Thorax* **60**, 455-461.

Schwartz, J. and Neas, L. (2000). Fine Particles Are More Strongly Associated Than Coarse Particles with Acute Respiratory Health Effects in Schoolchildren. *Epidemiology* **11**, 6-10.

Shaddick, G., Lee, D., Zidek, J. V., and Salway, R. (2007). Estimating exposure response functions using ambient pollution concentrations. arXiv:0710.5805v1 [stat.AP].

Singh, G.K. and Yu, S.M. (2005). Adverse pregnancy outcomes: differences between US- and foreign-born women in major US racial and ethnic groups. *American Journal of Public Health*, **86**, 837-843.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society* **64**, 583-639.

Teague, W.G. and Bayer, C.W. (2001). Outdoor air pollution. Asthma and other concerns. *Pediatric clinics of North America* **48**, 1167-1183.

U.S. EPA (1995a). Air Quality Criteria for Particulate Matter. Vol 1. EPA/600/p-95/oo1aF. Research Triangle Park, NC: U.S. Environmental Protection Agency, Environmental Criteria and Assessment Office.

U.S. EPA (1995b). Office of Air Quality Planning and Standards: Aerometric Information Retrieval System (AIRS). Vol 2. Research Triangle Park, NC: U.S. Environmental Protection Agency.

Wakefield, J. (2003). Sensitivity Analyses for Ecological Regression. *Biometrics* **59**, 9-17.

Wakefield, J. and Shaddick, G. (2006). Health-exposure modeling and the ecological fallacy. *Biostatistics* **7**, 438-455.

Zidek, J., Shaddick G., White R., Meloche, J., and Chatfield, C. (2005). Using a probabilistic model (pCNEM) to estimate personal exposure to air pollution. *Environmetrics* **16**, 481-493.