Some Advanced Semiparametric Single-Index Modeling for Spatially-Temporally Correlated Data

Hamdy Fayez Farahat Mahmoud

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Inyoung Kim, Chair
Eric P. Smith
George R. Terrell
Pang Du

September 26, 2014
Blacksburg, Virginia

KEYWORDS: Change Point; Generalized Linear Model; Generalized Additive Model; Markov Chain Expectation Maximization; Mixed model, Permutation Test; Semiparametric regression; Single Index model; Spatially correlated data; Spatio-temporal data.

Some Advanced Semiparametric Single-Index Modeling for Spatially-Temporally Correlated Data

Hamdy Fayez Farahat Mahmoud

(ABSTRACT)

Semiparametric modeling is a hybrid of the parametric and nonparametric modelings where some function forms are known and others are unknown. In this dissertation, we have made several contributions to semiparametric modeling based on the single index model related to the following three topics: the first is to propose a model for detecting change points simultaneously with estimating the unknown function; the second is to develop two models for spatially correlated data; and the third is to further develop two models for spatially-temporally correlated data.

To address the first topic, we propose a unified approach in its ability to simultaneously estimate the nonlinear relationship and change points. We propose a single index change point model as our unified approach by adjusting for several other covariates. We nonparametrically estimate the unknown function using kernel smoothing and also provide a permutation based testing procedure to detect multiple change points. We show the asymptotic properties of the permutation testing based procedure. The advantage of our approach is demonstrated using the mortality data of Seoul, Korea from January, 2000 to December, 2007.

On the second topic, we propose two semiparametric single index models for spatially correlated data. One additively separates the nonparametric function and spatially correlated random effects, while the other does not separate the nonparametric function and spatially correlated random effects. We estimate these two models using two algorithms based on Markov Chain Expectation Maximization algorithm. Our approaches are compared using simulations, suggesting that the semiparametric single index nonadditive model provides more accurate estimates of spatial correlation. The advantage of our approach is demonstrated using the mortality data of six cities, Korea from January, 2000 to December, 2007.

The third topic involves proposing two semiparametric single index models for spatially and temporally correlated data. Our first model has the nonparametric function which can separate from spatially and temporally correlated random effects. We refer it to "semiparametric spatio-temporal separable single index model (SSTS-SIM)", while the second model does not separate the nonparametric function from spatially correlated random effects but separates the time random effects. We refer our second model to "semiparametric nonseparable single index model (SSTN-SIM)". Two algorithms based on Markov Chain Expectation Maximization algorithm are introduced to simultaneously estimate parameters, spatial effects, and times effects. The proposed models are then applied to the mortality data of six major cities in Korea. Our results suggest that SSTN-SIM is more flexible than SSTS-SIM because it can estimate various nonparametric functions while SSTS-SIM enforces the similar nonparametric curves. SSTN-SIM also provides better estimation and prediction.

To my Mother and my Father's soul

Acknowledgments

I would like to express the deepest appreciation to my supervisor, Dr. Inyoung Kim, who has provided guidance, support, and encouragement throughout the time of my dissertation research. She has always taken great care of her students and helped them to the best of her abilities and beyond. I would like also to thank my committee members, Dr. Eric Smith, Dr. Pang Du, and Dr. George Terrell for their help, valuable comments, and positive input. Without their ongoing support, I could not have finished this work.

I would also like to extend gratitude to all the professors in the Department of Statistics for their inspiring courses and guidance. Special thanks goes to Dr Birch for his continued moral support, advice, and assistance. He was always there in times of need.

Lastly, I would like to thank my beloved parents, for their great and never-ending support accompanying me all the way. A special thank you to my brothers Mohamed, Shaban, Kamal, Abdallah, Mohsen, my sister Shaimaa, my wife Esraa, and my little beloved sons Muhammad and Ibrahim. None of this would have been possible without the love and support of you all.

Always yours,

Hamdy Mahmoud

(Fall 2014)

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Background

1.1.1 The Temperature and Mortality Relationship

It has been showed that there is a nonlinear relationship between mortality and temperature. The human body can adapt to exposure to extreme temperatures across the thermoregulatory functions. High temperatures lead to an increase in the heart rate because of increase the flow of blood from the body to the skin, which can lead to sweating in high temperatures or shaking in the cold temperatures. Within certain temperature ranges, human body responses allow individuals to follow some physical and mental activities but body exposing to temperatures outside these ranges or exposure to temperature extremes for a long periods of time makes human health in danger and can result in mortality. Basu and Samet (2002) showed that hot temperatures are associated with excess mortality due to some kind of diseases, such as cardiovascular, respiratory, and cerebrovascular. That is because hot temperatures are associated with increases in blood viscosity and blood cholesterol levels.

On the other side, Deschenes and Moretti (2009) showed exposure to cold days has also been found to be a significant factor for mortality.

1.1.2 Semi/nonparametric Regression Models

Nonparametric regression is a form of regression analysis in which the conditional mean function does not take a predetermined form but it is constructed according to information derived from the data. Nonparametric regression requires larger sample sizes than regression based on parametric models because the data must supply the model structure as well as the model estimates. Semiparametric models are a mix between parametric and non parametric models. Semiparametric model is more flexible comparing to the parametric model and more appropriate to the real situation where the functional form is possibly neither linear nor nonlinear, see Ruppert et al. (2003). Any applications area uses regression modeling analysis can benefit from semiparametric regression modeling. Semiparametric models has been used in may areas but it is still quit limited in spatially and spatially-temporally correlated data. As a result, the main goal of this dissertation is to address many problems semiparametrically by proposing different semiparametric models for different problems. All the proposed models are based on semiparametric single index model.

1.1.3 Semiparametric Single Index Model

Semiparametric single index model (SIM) assumes the conditional mean function is unknown. It takes the following form:

$$E(Y|X = x) = g(x\boldsymbol{\alpha}),$$

where Y is a scalar-dependent variable, X is a vector of explanatory variables, $\boldsymbol{\alpha}$ is a vector of parameters whose values are unknown, and g is an unknown function.

SIM is popular in many scientific fields such as biostatistics, medicine, economics and financial econometrics and has been extensively studied in the statistical literature; see Li (1991); Naik and Tsai (2000); Stute and Zhu (2005); Xia and Li (1999); Xia et al. (2002); Zhu and Ng (1995); Zhu and Xue (2006); Lin and Kulasekera (2007); Xia (2006); Zhu and Zhu (2009a,b); Wang et al. (2010); Hridtache et al. (2001); Chang et al. (2010).

SIM outperforms the parametric models, such as linear models and generalized linear model in terms of flexibility because it assumes the conditional mean function is unknown which makes this model more flexible for the conditional mean function. Also SIM does not assume a specific distribution for error which avoids the misleading results of using incorrect distribution for errors (Horowitz and Hardle, 1996).

It has some advantages over the nonparametric models, such as the precision of nonparametric estimation decreases when X dimension increases, curse of dimensionality, and to overcome this problem a large sample is needed. In SIM we have only one dimension which is the index $X\boldsymbol{\alpha}$ that enables SIM to avoid the curse of dimensionality and $\boldsymbol{\alpha}$ can be estimated with rate of convergence $n^{-1/2}$ (Li et al., 2007). There are several methods for estimating the single coefficients vector, $\boldsymbol{\alpha}$, such as average derivative estimator (Stoker, 1986; Powell et al., 1989) and semiparametric M-estimation estimator (Ichimura, 1993; Klein and Spady, 1993).

1.1.4 Change Point Detection

In statistical analysis, change detection or change point detection tries to identify when the probability distribution of a stochastic process or time series changes. In general the problem

concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes. It is very common in time series analysis but can be used in many applications. In mortality data, we need to detect at which temperature degree the slope of mortality changes.

Many articles have studied the relationship between temperature and mortality (El-Zein et al., 2004; Hashizume et al., 2009; Chung et al., 2009; Armstrong, 2006; Son et al., 2011; Kan et al., 2007). These methods consist of two steps: they first estimate the models using either generalized linear model or generalized additive model and then detect change points. However, methods for simultaneously identifying the nonlinear relationship and detecting the number of change points are quite limited. In addition, there are no studies in the mortality analysis tried to detect more than one change points or used a testing procedure to test either the change point is significant or not.

1.1.5 Spatial Data

The first law of geography according to Tobler (1970) is "Everything is related to everything else, but near things are more related than distant things". In general terms, spatial analysis can be considered to be the formal quantitative study of phenomena that manifest themselves in space. This includes a focus on location, area, distance and interaction. Spatial data are data that have a spatial component, it means that data are connected to a place on the Earth. Spatial dependency leads to the spatial autocorrelation problem in statistics since this violates standard statistical techniques that assume independence among observations. Li et al. (2007) showed that when statistical modeling procedure ignores spatial correlation, the standard errors of the parameter estimates is deflated and therefore the statistical significance gets inflated.

The parametric approach in spatial data analysis remains the most common one. However, the traditional assumptions of normality and of perfect knowledge of the model specification are often rather crude abstractions of reality. As a result, the relevance of a strict parametric approach has been questioned, and a nonparametric approach is needed. Modeling these kind of data semiparametrically is one of this dissertation objectives.

1.1.6 Spatial Covariance Functions

In probability theory and statistics, covariance is a measure of how much two variables change together and the covariance function describes the spatial covariance of a random variable process or field. For a random field or stochastic process, say u(s), on a domain S, a covariance function $C(s_i, s_j)$ gives the covariance of the values of the random field at any two locations, say s_i and s_j:

$$C(s_i, s_j) = cov[u(s_i), u(s_j)].$$

In spatial analysis, the covariance function needs to have a parametric form to guarantee positive definite property. The common covariance functions are in Table 1.1

Table 1.1: Covariance functions; d = the distance between any two locations, s_i and s_j, σ^2=scale parameter giving the overall variability of the process, ρ= dependence range, ν= rate of decay, m and θ= maximum range of dependence, and K_η= modified Bessel function of the second kind, of order η.

	Covariance function	Parameters
Exponential	$\sigma^2 e^{\frac{-\lvert d\rvert}{\rho}}$	$\sigma^2 > 0,\ \rho > 0$
Gaussian	$\sigma^2 e^{\frac{-\lvert\lvert d\rvert\rvert^2}{\rho}}$	$\sigma^2 > 0,\ \rho > 0$
Spherical	$\sigma^2(1 - \frac{3\lvert\lvert d\rvert\rvert}{2m} + \frac{\lvert\lvert d\rvert\rvert^3}{2m^3})$	$\sigma^2 > 0,\ \lvert\lvert d\rvert\rvert \leq m$
Tend	$\sigma^2(1 - \frac{\lvert\lvert d\rvert\rvert}{\theta})$	$0 \leq \lvert\lvert d\rvert\rvert \leq \theta$
Matern	$\frac{\sigma^2}{\Gamma(\eta+1/2)}(\frac{\lvert\lvert d\rvert\rvert}{2\nu})^\eta K_\eta(\nu\lvert\lvert d\rvert\rvert)$	$\eta > 0,\ \nu > 0$

1.1.7 Spatio-Temporal Data

Spatial-temporal data arise when data are collected across time as well as space. The earliest papers on spatial-temporal data analysis were a series of papers by Bilonick and Nichols (1983) and Bilonick (1983, 1985, 1988). An example of spatio-temporal data would be that of Bilonick and Nichols (1983) who considered the analysis of rainfall data from 22 stations in or near New York State, collected from 1965 to 1979, and pooled into monthly values at each station. Thus the data analysis has to take account of spatial dependence among the stations, but also that the observations at each station typically are not independent but form a time series. Therefore one must take account of temporal correlations as well as spatial correlations.

The common approach in analyzing spatio-temporal data is the parametric approach. However, the traditional assumptions of normality and of perfect knowledge of the model specification are often rather crude abstractions of reality. As a result, the relevance of a strict parametric approach has been questioned, and a nonparametric approach is needed for spatially-temporally correlated data. Introducing semiarametric models for these kind of data is one of our goals in this dissertation.

1.1.8 Spatio-Temporal Covariance Functions

The product of valid covariance functions is a valid covariance function. So, a nonnegative definite spatio-temporal covariance function can be written as a product of a spatial covariance function and a temporal covariance function as follows:

$$Cov[Z(s,t), Z(s+d, t+h)] = Cov[Z(s), Z(s+d)] \times Cov[Z(t), Z(t+h)],$$

where d is the distance between two locations and h is the difference between two time points. This gives a separable spatio-temporal covariance function.

The well-known covariance functions for spatial effects are presented in Table 1.1. For the temporal random effects, $\nu(t)$, it is commonly used the random walk (RW), autoregressive (AR), moving average (MA) or autoregressive integrated moving average (ARIMA). Clayton (1996) applied a first order random walk, denoted by RW(1), to model temporal trend. RW(1) means the difference between any two consecutive time points follows a normal distribution and the relationship between any two consecutive time points, say $\nu(t)$, and $\nu(t-1)$, has the form:

$$\nu(t) = \nu(t-1) + \epsilon(t),$$

where $\epsilon(t)$ is the random noise term that accounting for the difference from one observation to next observation within each location and follows a normal distribution.

Several nonseparable spatio-temporal covariance functions have been introduced (Jones and Zhang, 1997; Cressie and Huang, 1999; Gneiting, 2002; Stein, 2005). For example, Gneiting's nonseparable covariance function takes the form:

$$Cov(d, h, \theta) = \frac{\sigma^2}{(|d|^{2\gamma}+1)^{\tau}} \exp[\frac{-\rho||h||^{2\gamma}}{(|d|^{2\gamma+1})^{\beta\gamma}}],$$

where ρ is the spatial dependence range, τ controls the smoothness of temporal correlation, $\gamma \in (0, 1]$ controls the smoothness of the spatial correlation, and $\beta \in [0, 1)$ controls the strength of the interaction between space and time.

1.2 Motivation

The main goals of this dissertation are to address three semiparametric problems. The first problem is introducing a flexible model for modeling the nonlinear relationship between mortality and temperature and simultaneously incorporate the change point. In addition a testing procedure for the change points is needed. To address this problem, single index model has been developed by including the change point, this model called single index change point model (SICM). The second problem is proposing a semiparametric model for spatially correlated data. We developed the single index model to incorporate the spatial correlated data. Including spatial correlated data is in two different formats: one is additively separable and the other is nonseparable from the unknown function. The last problem is introducing a semiparametric model to spatially-temporally correlated data. We introduced that model by elaborating the single index model to include spatio-temporal dependence. The research of developing these models is motivated by a real data set. This data has daily non accident mortality and weather variables; mean temperature, humidity, mean pressure, day, and day of week were recorded for six cities in Korea from January, 2000 to December, 2007. The sample size for each city is 2922 observations. Those cities are the major cities in South Korea: Seoul, Busan, Daegu, Incheon, Gwangju, and Daejeon.

1.2.1 Change Points Detection in Single Index Model

Generalized linear models with log link (GLM) or generalized additive models (GAM) have been used to describe the nonlinear relationship between mortality and temperature. The current available methods to detect change point consist of two steps: they first estimate the models and then detect change points. However, the methods for simultaneously identifying the nonlinear relationship and detecting the number of change points are quite limited.

Therefore in this dissertation, we propose a unified approach in its ability to simultaneously estimate the nonlinear relationships and detect the change points. We propose a single index change point model (SICM) as our unified approach by adjusting for several other covariates. Currently there are no testing procedures for testing whether the change points are significant or not. We provide a permutation based testing procedure to detect multiple change points and test whether they are significant or not.

1.2.2 Semiparametric Spatial Modeling

Parametric models for spatially correlated data (Anselin and Florax, 1995; Cressie, 1991; Guyon, 1995; Possolo, 1991; Ripley, 1981; Zhang, 2002) were developed to take account spatial correlation into statistical analysis. However, there are quite limited approaches (Gu and Ma, 2005; Pang and Xue, 2012) on nonparametic model although nonparametric regression has become a standard statistical method when the functional form is possibly neither linear nor nonlinear of a specific type. We model the relationship between the response variable Y and some covariates, $X's$, by incorporating correlated spatial random effects in single index nonparametric model, that is, we want to estimate $E[Y|X]$, nonparametrically for spatially correlated data.

Pang and Xue (2012) introduced the single index model with random effects to incorporate the random effects into the model and to solve the curse of dimensionality problem. This model can be written

$$y_{ij} = m(X_{ij}^T \boldsymbol{\beta}) + Z_{ij}^T b_i + \epsilon, \ i = 1, \ldots, I, \ j = 1, , \ldots, J,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ index coefficients vector associated with the covariates X_{ij} and $\|\beta\| = 1$ is used as single index restriction for identifiability, the $b_i's$ are independent random effects

with mean $\mathbf{0}$ and covariance matrix $\sigma_b^2 I$, $m(\cdot)$ is an unknown function, $\epsilon_i' s$ are independent mean $\mathbf{0}$ and variance $\sigma_\epsilon^2 I$. Pang and Xue (2012) used the Generalized Estimating Equation (GEE) to estimate the single-index model with the random effect. However there are several limitations of this model: (1) it assumes that random effects are independent of each other; (2) $\mathbf{Y}$ follows normal distribution; and (3) it assumes that nonparametric $m(\cdot)$ and random effects can be additively separated. Because of those limitations, we are not able to use this model to our motivated data.

1.2.3 Semiparametric Spatio-Temporal Modeling

Some literatures (Cressie and Hawkins, 1980; Cressie and Huang, 1999; Cressie, 1993; Chiles and Delfiner, 1999; Isaaks and Srivastava, 1999; Stein, 1999) on spatio-temporal analyses have extensively studied covariance functions, while others (Lekdee and Ingsrisawang, 2013; Arcuti et al., 2013; Hayn et al., 2009; Landagan and Barrios, 2007) have focused on estimating the mean function. However, the most approaches are developed using parametric model with strong model assumption which can not be applicable to real application. Perhaps, semiparametric model is more flexible and appropriate to the real situation where the functional form is possibly neither linear nor nonlinear. Since semiparametric modeling for for spatially and temporally correlate data is quit limited, in this dissertation we propose a semiparametric model based on single index model for allow several covariates X. A semiparametric model is developed by incorporating spatial effects and time effects into single index model. To the best of our knowledge, there is no semiparametric spatio-temporal model proposed before.

1.3 Overview

The rest of this dissertation is organized as follows. In Chapter 2, we introduce a flexible semiparametric model to simultaneously estimate the unknown relationship between daily/weekly mortality and temperature and detect the change points by proposing a single index change points model. also we propose a permutation based test to test the significance of the detected change points. In Chapter 3, we propose two models to incorporate the correlated spatial random effects with the nonparametric function in two different formats; additively separable and nonseparable. Tow algorithms based on EM algorithm to estimate the models parameters has been introduced. In addition we apply the proposed models to South Korea data. In Chapter 4, two proposed models have been introduced to incorporate spatio-temporal dependence with the nonparametric function. Two algorithms based on EM algorithm have been proposed to estimate the models parameters and spatial and time effects. In addition the two proposed models have been applied to South Korea data. In chapter 5, we give a general review on the contributions of this dissertation, as well as discuss directions for future research.

Chapter 2

Single Index Change Point Model with an Application of Environmental Health Study on Mortality and Temperature

2.1 Background

The Generalized Linear Model (GLM) generalizes linear regression by allowing the linear model to be related to the response variable ($\mathbf{Y}$) and model matrix (X) via a link function, $E(\mathbf{Y}) = g^{-1}(X\boldsymbol{\beta})$, where g function is known and $\boldsymbol{\beta}$ is an unknown vector of parameters. The generalized linear additive model takes the form $E(\mathbf{Y}) = g^{-1}\{\beta_0 + f_1(x_1) + f_1(x_2) + ... + f_p(x_p)\}$, where x_j is the jth explanatory variable, $j = 1, \ldots, p$, and p is the total number of explanatory variables in the model. The function $f_j(x_j)$ may be estimated using parametric

or nonparametric means, thus providing the potential for better fits to data than GLM. To make this analysis more flexible, a single index model (SIM) can be used. It assumes the link between the mean response and the linear model is unknown and estimates it nonparametrically. Two main advantages of SIM over GLM and GAM are as follows; (1) it avoids misspecifying the link function which could cause misleading results (Horowitz and Hardle, 1996) and (2) it reduces the dimension of the problem by assuming the link function to be a univariate function applied to the projection of the explanatory covariate vector onto some direction. SIM is popular in many scientific fields such as biostatistics, medicine, economics and financial econometrics and has been extensively studied in the statistical literature; see Zhu and Xue (2006); Lin and Kulasekera (2007); Xia (2006); Zhu and Zhu (2009a,b); Wang et al. (2010); Hridtache et al. (2001); Chang et al. (2010). However, there are few applications of SIM in environmental epidemiology. Environmental health studies are of great interest in human research, especially in evaluating the relationship between daily/weekly mortality and temperature. Many articles have studied the relationship between temperature and mortality. Methods that have been used include generalized linear Poisson regression model (El-Zein et al., 2004; Hashizume et al., 2009), and generalized additive models (Chung et al., 2009; Armstrong, 2006; Son et al., 2011; Kan et al., 2007). These methods consist of two steps: they first estimate the models using either GLM or GAM and then detect change points. However, methods for simultaneously identifying the nonlinear relationship and detecting the number of change points are quite limited. We propose the single index change point model, $E(\mathbf{Y}) = g^{-1}([X - \boldsymbol{\theta}]_+\boldsymbol{\alpha})$, with a link function $g(\cdot)$, single index vector of coefficient $\boldsymbol{\alpha}$, and vector of change points $\boldsymbol{\theta}$ which are unknown and need to be estimated. This single index change point model can simultaneously identify the nonlinear relationship and detect the number of change points.

The testing procedure for detecting change points is also limited for GLM and GAM. Kim et al. (2000) provided a permutation based procedure for testing change points in the segmented linear regression model and Kim et al. (2009) studied the asymptotic properties of the number of change points selected by this permutation test. They showed that, under some regularity conditions, the number of change points selected by the permutation test is consistent in the segmented line regression model. Therefore, the goals are to contribute to the single index model in its ability to detect the change points and also to provide a permutation based testing procedure for detecting change points in SICM and compare its performance with GLM and GAM.

The motivation of our study is based on mortality data in Korea. This data has daily non accident mortality and weather variables; mean temperature, humidity, mean pressure, day, and day of week were recorded for seven cities in Korea from January, 2000 to December, 2007. The sample size for each city is 2922 observations. One of these cities is Seoul, the capital and largest metropolis of South Korea. Home to over 10 million citizens, Seoul is one of the largest cities in the world. It is classified as having a temperate climate with four distinct seasons, but temperature differences between the hottest part of summer and the depths of winter are extreme. Two primary questions are whether there exists a nonlinear relationship between weekly mortality and temperature after adjusting for other covariates and how many change points of temperature exist. We would like to answer these questions using our single index change point model (SICM) and a permutation testing procedure on SICM.

The reminder of this chapter is organized as follows. In Section 2.2, we introduce the single index change point model. In Section 2.3, we describe a permutation based testing procedure; the permutation based testing procedures for detecting one and multi change points and its asymptotic properties. In Section 2.4, we perform a simulation study to compare our

approach with the generalized linear model and generalized additive model. In Section 2.5, we apply our approach to real data. Concluding remarks are provided in Section 2.6.

2.2 Single Index Change Point Model (SICM)

Suppose we have n observations and p covariates. Without loss of generality, suppose we are interested in change points in variable x_1. The SICM with k change points can be written as

$$\begin{aligned} \mathbf{Y} &= g(\alpha_{01}x_1 + \alpha_{11}[x_1 - \theta_1]_+ + ... + \alpha_{1k}[x_1 - \theta_k]_+ + \alpha_2 x_2 + \alpha_3 x_3 + \ldots + \alpha_p x_p) + \boldsymbol{\epsilon} \\ &= g(X\boldsymbol{\alpha}) + \boldsymbol{\epsilon} \end{aligned}$$

where $g(\cdot)$ is an unknown function, $X = \{x_1, [x_1 - \theta_1]_+, [x_1 - \theta_2]_+, \ldots, [x_1 - \theta_k]_+, x_2, x_3, \ldots, x_p\}$ is a $n \times (p + k)$ matrix of regressors values, $\boldsymbol{\alpha}$ is a $(p + k) \times 1$ vector of parameters, and $E(\boldsymbol{\epsilon}|X) = \mathbf{0}$. We define $[x_1 - \theta_l]_+ = \max[0, x_1 - \theta_l]$, $l = 1, ..., k$, where θ_l is the unknown change point.

In SIM, some restrictions on $\boldsymbol{\alpha}$ are needed in order for it to be identifiable. One approach is to set $\|\boldsymbol{\alpha}\|=1$ (Xia et al., 2004; Hardle et al., 1993; Lin and Kulasekera, 2007) while another is to set one component of $\boldsymbol{\alpha}$ to be equal to one (Ichimura, 1993; Sherman, 1994). In this study, we use the second approach for identifiability in SICM. The variable whose coefficient is set to one is required to have a non-zero coefficient. In addition, it is required that $p \geq 2$. If $p = 1$, then $\boldsymbol{\alpha}$ is simply normalized to one. Identification of $\boldsymbol{\alpha}$ and g also requires having at least one continuous variable, also having a non-zero coefficient.

Note that when $g(\cdot)$ is known, the model reduces to a class of GLM. If it is the identity function, it reduces to the linear model. In SIM, we are interested in modeling the relation

between $E(\mathbf{Y}|X)$ and a single linear combination $(X\boldsymbol{\alpha})$. In a semiparametric single index model, the object of interest depends on X through the function $g(X\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} \in R^{p+k}$ and $g : R \longmapsto R$ are unknown.

We use Ichimura (1993)'s approach to simultaneously estimating $\boldsymbol{\alpha}$ and $g(\cdot)$. We also use Lerman (1980)'s grid search method to estimate the change points. Lerman proposed a method to estimate the change points in linear and non-linear regression models. The estimated change point is the one that minimize the residual sum of squares function. Ichimura (1993) suggested a two-step estimator. First g is estimated using the leave-one-out Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964):

$$\hat{g}_i(x_i^T\boldsymbol{\alpha}) = \left[\frac{\sum_{j\neq i} f(\frac{(x_j - x_i)^T\boldsymbol{\alpha}}{h})y_j}{\sum_{j\neq i} f(\frac{(x_j - x_i)^T\boldsymbol{\alpha}}{h})}\right]$$

where $f(.)$ is the kernel function and h is the bandwidth. Then, $\boldsymbol{\alpha}$ is estimated using nonlinear least-squares by minimizing $S_n(\boldsymbol{\alpha}, g)=\sum_{i=1}^n (y_i - g(x_i^T\boldsymbol{\alpha}))^2$ for given $g = \hat{g}$. Note that $g(x_i^T\boldsymbol{\alpha})$ is the conditional mean of y_i given $x_i^T\boldsymbol{\alpha}$ and $g(\cdot)$ depends on $\boldsymbol{\alpha}$. We use multiple grid search to estimate $\boldsymbol{\theta}$, the vector of change points.

2.3 Permutation Test and Its Asymptotic Properties

In Section 2.3.1, we provide a permutation based testing approach for one and multiple change points in the SICM. We also describe its asymptotic properties in Section 2.3.2.

2.3.1 Permutation Test

First, we consider to test one change point. The hypothesis can be written as follows;

$$H_0 : \text{there is no change point} \quad \text{versus} \quad H_1 : \text{there is one change point.}$$

The SICM under H_0 becomes a simple SIM

$$y_i = g(\alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + ... + \alpha_p x_{pi}) + \epsilon_i^{(0)} = \mu_i^{(0)} + \epsilon_i^{(0)}$$

and the SICM under H_1 can be written as

$$y_i = g(\alpha_1 x_{1i} + \alpha_2 [x_{1i} - \theta]_+ + \alpha_3 x_{2i} + ... + \alpha_{p+1} x_{pi}) + \epsilon_i^{(1)} = \mu_i^{(1)} + \epsilon_i^{(1)}$$

where $\epsilon_i^{(k)}$ is the error and $\mu_i^{(k)}$ is the mean value, and $k = 0, 1$, where $k = 0$ represents the null hypothesis and $k = 1$ represents the alternative hypothesis.

Under the null hypothesis, the assumptions for the permutation test are $E(\epsilon_i^{(0)}) = 0$, $\text{var}(\epsilon_i^{(0)}) = \sigma^2$ for all i, and $\text{cov}(\epsilon_i^{(0)}, \epsilon_j^{(0)}) = 0$ for all $i \neq j$, where $\epsilon_i^{(0)}$ is the residual under the null hypothesis. These assumptions are the same as those given in Kim et al. (2009).

The permutation based testing procedures are following steps for one change point:

Step 1 Fit the model under the null hypothesis and obtain the parameter estimates, $(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p)$, predicted mean response, $\hat{\mu}_i^{(0)}$, and residuals, $\hat{\epsilon}_i^{(0)} = y_i - \hat{\mu}_i^{(0)}$. Then, at each step of the grid search, fit the model under the alternative hypothesis and find the parameter estimates. That is, we find the values of

$$\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \ldots, \hat{\alpha}_{p+1}, \hat{\theta}$$

that minimize

$$\sum_{i=1}^{n} (y_i - \mu_i^{(1)})^2$$

and obtain the residuals, $\hat{\epsilon}_i^{(1)}$;

Step 2 Compute the test statistic

$$T(\mathbf{y}_{(0)}) = \frac{[\hat{\boldsymbol{\epsilon}}^{(0)}(\mathbf{y}_{(0)})]^T[\hat{\boldsymbol{\epsilon}}^{(0)}(\mathbf{y}_{(0)})]}{[\hat{\boldsymbol{\epsilon}}^{(1)}(\mathbf{y}_{(0)})]^T[\hat{\boldsymbol{\epsilon}}^{(1)}(\mathbf{y}_{(0)})]},$$

where $\mathbf{y}_{(0)}$ denotes to the original data set and $\hat{\boldsymbol{\epsilon}}^{(0)}(\mathbf{y}_{(0)})$ and $\hat{\boldsymbol{\epsilon}}^{(1)}(\mathbf{y}_{(0)})$ are the residuals under the null and alternative hypotheses of the original data set respectively;

Step 3 Permute the residuals, $\hat{\boldsymbol{\epsilon}}^{(0)}$, and add them back to the null model means, $\mathbf{y}_{(m)} = \hat{\boldsymbol{\mu}}^{(0)} + \hat{\boldsymbol{\epsilon}}_m^{(0)}$, where m represents the m^{th} permutation, $\hat{\boldsymbol{\epsilon}}_m^{(0)}$ is the permuted $n \times 1$ vector of residuals and $\mathbf{y}_{(m)}$ is $n \times 1$ vector of permuted responses;

Step 4 For the permuted data set, $\mathbf{y}_{(m)}$, fit the null and alternative hypothesis the same way as in step 1 and compute the test statistic

$$T(\mathbf{y}_{(m)}) = \frac{[\hat{\boldsymbol{\epsilon}}^{(0)}(\mathbf{y}_{(m)})]^T[\hat{\boldsymbol{\epsilon}}^{(0)}(\mathbf{y}_{(m)})]}{[\hat{\boldsymbol{\epsilon}}^{(1)}(\mathbf{y}_{(m)})]^T[\hat{\boldsymbol{\epsilon}}^{(1)}(\mathbf{y}_{(m)})]}$$

where, $\hat{\boldsymbol{\epsilon}}^{(0)}(\mathbf{y}_{(m)})$ and $\hat{\boldsymbol{\epsilon}}^{(1)}(\mathbf{y}_{(m)})$ are the residuals from fitting the m^{th} permuted data set under the null and alternative hypothesis respectively;

Step 5 For a large number, N_p of values of $T(\mathbf{y}_{(m)}), m = 1, 2, ..., N_p$, and the value of the original data set, $T(\mathbf{y}) = T(\mathbf{y}_{(0)})$, compute the empirical p-value as

$$\frac{\text{number of times that}[T(\mathbf{y}_{(m)}) \geq T(\mathbf{y})]}{N_p + 1},\ m = 0, 1, 2, \ldots, N_p,$$

which measures how extreme the $T(\mathbf{y})$ value is.

For the multiple change points case, our permutation based testing procedure can be extended as follows. In this case, the hypothesis is written as

$$H_0 : \text{there are } k_0 \text{ change points} \quad \text{versus} \quad H_1 : \text{there are } k_1 \text{ change points,}$$

where $k_0 < k_1$ and the SICM can be written as

$$\begin{aligned} y_i &= g(\alpha_{01}x_{1i} + \alpha_{11}[x_{1i} - \theta_1]_+ + ... + \alpha_{1k}[x_{1i} - \theta_k]_+ + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \ldots + \alpha_p x_{pi}) + \epsilon_i^{(k)} \\ &= \mu_i^{(k)} + \epsilon_i^{(k)} \end{aligned}$$

where $\epsilon_i^{(k)}$ is the error and $\mu_i^{(k)}$ is the mean value. Under the null hypothesis $k = k_0$ and under the alternative hypothesis $k = k_1$.

The permutation based testing procedure for multi change points is as follows:

Step 1 Test H_0: $k_0 = 0$ versus H_1: $k_1 = K$, where K is the maximum number of change points presumed;

Step 2 If H_0 is rejected, we then test H_0: $k_0 = 1$ versus H_1: $k_1 = K$, otherwise we test H_0: $k_0 = 0$ versus H_1: $k_1 = K - 1$;

Step 3 We proceed in a similar manner, where we increase the number of change points under the null model by one if the null hypothesis is rejected and decrease the number of change points under the alternative model by one if we failed to reject until we reach testing H_0: $k_0 = k$ versus H_1: $k_1 = k + 1$;

Step 4 If we rejected H_0, the estimated number of change points is $k + 1$ otherwise it is k. Bonferroni correction for the significance level is used for each test to be α/K.

2.3.2 Asymptotic Properties of Permutation Test

Let H be a $n \times (p + K)$ matrix as described in Appendix A. We assume two Assumptions 1 and 2.

Assumption 1

A.1: $\{H_t\}$ is a strictly stationary, ergodic process with positive definite matrices $\mathrm{E}\{H_1\acute{H}_1 1_{h_{1d}\in(\theta_i^0-\delta,\theta_i^0)}\}$ and $\mathrm{E}\{H_1\acute{H}_1 1_{h_{1d}\in(\theta_i^0,\theta_i^0+\delta)}\}$ in a small δ-neighborhood of each of the true change points $\theta_1^0, ..., \theta_{l_0}^0$.

A.2: The ϵ_i are independent and identically distributed with mean zero and variance σ_0^2, and for some constants B_0 and T_0 in $(0, \infty)$, $E(e^{t\epsilon_i}) \leq e^{B_0 t^2}$ for all $\|t\| \leq T_0$.

The conditions in Assumption 1 are the same ones in Lui et al. (1997) and Kim et al. (2009). We can show that our model can be a local linear approximation model. The details are given in Appendix A.

Under Assumption 1, we have the following Theorem:

Theorem 1: *Suppose that Assumption 1 is satisfied and the maximum number of change points, M, is fixed. Then the estimated number of change points, $\hat{k}$, converges to the true number of change points, k^*, in probability as $n \longrightarrow \infty$.*

The proof is similar to that given for Theorem 1 in Kim et al. (2009) since our model can be approximately expressed as a local linear model after replacing the matrix X in Kim et al. (2009) by H.

Assumption 2

B.1: There are at least $n/\ln(n)$ observations in each segment of $[\hat{\theta}_j, \hat{\theta}_{j+1})$ and so of $[\theta_j, \theta_{j+1})$ for $j = 0, ..., k^*$, $k^* < M$ for a positive fixed constant.

B.2: The ϵ_i are independently and identically distributed with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_0^2$, and $E|\epsilon_i|^{2(1+\delta)} < \infty$ for some $\delta > 0$.

B.3: lim $sup_n K_n/n < 1$, and K_n is increasing slowly such that $lim_{n\longrightarrow\infty} K_n/\sqrt{\eta^*} = 0$, where $\eta^* = \mu^{*T}(1 - G_{k^*-1}(\boldsymbol{\theta}_{k^*}))\mu^*$.

The conditions in Assumption 2 are also the same ones in Lui et al. (1997) and Kim et al. (2009).

Under the Assumption 2, we have the following Theorem:

Theorem 2: *If the maximum number of the change points, K, depends on the number of observations, n, and Assumption 2 is satisfied, then* $\hat{k}$ *convergences to* k^* *in probability as* $n \longrightarrow \infty$.

The proof is similar to that given for Theorem 2 in Kim et al. (2009) since, once again, our model can be approximately expressed as a local linear model (1), in Appendix A, after replacing the symbol G by H

Theorem 1 and 2 mean that the estimated number of change points converges to the true number of change points. Theorem 1 considers a fixed maximum number of change points, while Theorem 2 considers a random maximum number of change points depending on n.

2.4 Simulations

We conduct a simulation to assess the performance of the SICM in detecting the change point(s) and compare its performance with the two alternative models, GAM and GLM.

We simulated 100 data sets for each of the three models under H_1 and H_0, respectively. We then computed power and type I error using permutation based testing procedure. We estimated Power and Type I error which are the relative frequency of p-values. For H_1, we consider two cases: (1) there is one change point and (2) there are two change points.

2.4.1 Case 1: One Change Point

We consider two variables, x_1, and x_2. Both are simulated from uniform $[\pi, 2\pi]$ with sample size $n = 500$.

The following three models were considered:

- Single index change point model

$$y_i = \sin(\alpha_1 x_{1i} + \alpha_2 [x_{1i} - \theta]_+ + \alpha_3 x_{2i}) + \epsilon_i$$

  where ϵ_i is generated from Normal distribution with 0 mean and 0.05 standard deviation and θ is the change point. The first component of the index coefficients, $\boldsymbol{\alpha}^T = (\alpha_1, \alpha_2, \alpha_3)$, is fixed at 1 for identifiability reasons;

- Generalized additive model

$$\log(\mu_i) = \alpha_1 x_{1i} + \alpha_2 [x_{1i} - \theta]_+ + ns(x_{2i})$$

  where ns stands for natural splines smoothing function. The response variable y_i is generated from a Poisson distribution with μ_i;

- Generalized linear model

$$\log(\mu_i) = \alpha_1 x_{1i} + \alpha_2 [x_{1i} - \theta]_+ + \alpha_3 x_{2i}$$

  where the response variable y_i is generated from a Poisson distribution with μ_i.

For power, we set the change point θ equal to 4.7 and the single index coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ equal to $(1, -0.5, 1)^T$. For each model, we simulated 100 data sets and fit the

three different models which are SICM, GLM and GAM. When we generate data sets from SICM, $\mathbf{Y}$ follows a normal distribution in this case, so we fit GLM and GAM as a multiple linear regression. We then estimated the power using a permutation based testing procedure for each model. The number of permutations was 1000 and was on the scaled residuals. The ith scaled residual is

$$\tilde{\epsilon}_i^{(0)} = \hat{\epsilon}_i^{(0)} / \sqrt{\hat{\mu}_i^{(0)}}.$$

For Type I error, we used the same $\boldsymbol{\alpha}$ except that $\alpha_2 = 0$. We then simulated 100 data sets for each model with $\alpha_2 = 0$ and fit three different models which are SICM, GLM and GAM. Type I error was obtained using a permutation based testing procedure.

The results for power and Type I error are shown in Table 2.1 and Table 2.2, respectively. Table 2.1 shows the power of the test and the mean estimate of θ along with its standard errors. The mean estimate of θ is the average of $\hat{\theta}$ the estimates of θ from all the simulation runs and the standard error is the standard deviation of these estimates.

From this Table 2.1, we have the following results. First, when the true model is SICM which means the data simulated from the SICM under H_1, the power of SICM to detect the change point is 1 which is larger than those of GAM (0.81) and GLM (0.78), respectively. In addition, the mean estimate of θ based on the SICM is much closer to the true value than the other two models. The standard error of θ based on the SICM is also smaller. Second, for the 100 data sets simulated from GAM, SICM performs well with power 1 which are comparable to the other two models. In terms of the mean estimate of θ, SICM is again much closer to true value than the other two models. The standard error of SICM is comparable to those of other two models. Finally, when the true model is GLM, the power of SICM is 0.86 which is smaller than the other two models. The mean estimate of θ is comparable to the other two

models, while the standard error is larger than the other two models. Consequently, if the model is correctly specified, our model is the best in detecting the change point. However, if the model is misspecified or there is uncertainty on the correct model, choosing SICM is still a good choice in terms of power.

Table 2.1: Power of the three models based on 100 data sets simulated from each model and the estimates of mean and standard error (SE) of θ in case H_0 : there is no change point $k_0 = 0$, H_1: there are two change points, $k_1 = 2$; SICM= single index change point model; GAM= generalized additive model; GLM= generalized linear model. When true model is SICM with normal error, we fit the GAM and GLM using identity link function.

Data simulated from	Estimated model	Power	$\theta \pm$ SE
SICM	SICM	1	4.73±0.08
	GAM	0.79	4.82±0.62
	GLM	0.78	4.81±0.62
GAM	SICM	1	4.70±0.19
	GAM	1	4.72±0.12
	GLM	1	4.72±0.12
GLM	SICM	0.86	4.80±0.44
	GAM	1	4.76±0.06
	GLM	1	4.76±0.06

Table 2.2 summarizes the results for Type I error. First, for the 100 data sets simulated from the SICM, Type I error for SICM is 0.06, while type I errors of other two models are 0.78 and 0.76 which are much larger. In this case, SICM performs much better than GAM and GLM as its Type I error is close to the nominal value of 0.05. Second, for GAM data case, SICM is still better, with a Type I error of 0.08, while the Type I errors of the other two models are too small, that is 0. Finally, when the data follows GLM, the three models are comparable. Their Type I errors are 0.06, 0.05, and 0.05.

From Table 2.1 and Table 2.2, when the data are simulated from SICM, one can see that the power and Type I error for GLM and GAM are almost the same. That is, the probabilities

Table 2.2: Type I error rate of the three models based on 100 data sets simulated from each model in case H_0 : there is no change point, $k_0 = 0$, H_1 : there is one change point, $k_1 = 1$; SICM= single index change point model; GAM= generalized additive model; GLM= generalized linear model. When true model is SICM with normal error, we fit the GAM and GLM using identity link function.

Data simulated from	Estimated model	Type I error rate
SICM	SICM	0.06
	GAM	0.78
	GLM	0.67
GAM	SICM	0.08
	GAM	0
	GLM	0
GLM	SICM	0.06
	GAM	0.05
	GLM	0.05

of rejecting H_0 are similar whether H_0 is true or not. This means if we misspecified the model and used GAM or GLM as the model for our data set, our decision regarding H_0 is incorrect. In addition, it can also be seen that the performance of GAM and GLM are the same in terms of power and Type I error.

To see how much the SICM fits a data well, we also conducted a small simulation. 50 data sets were simulated from the single index model with one change point. We used a grid search 0.01 for estimating θ. Table 2.3 shows the mean estimates, the standard errors, and the mean square error for $\boldsymbol{\alpha}$ and θ. The mean estimates are very close to the true ones, the standard errors and MSE are small. Figure 2.1 and Figure 2.2 shows the true and estimated single index function as a function of x_1 and x_2. It is clear that they are identical. In addition, Figure 2.3 reveals that the SICM detects the change points very well. The scatter plot of index versus actual and fitted $\mathbf{y}$ is shown in Figure 2.4

Table 2.3: Mean, standard error (SE), and mean square error (MSE) of the parameter estimates in a single index change point model

Parameter	True	Mean	SE	MSE
θ	4.7	4.7060	0.0527	0.0019
α_1	1	1	0	0
α_2	1	1.0007	0.0142	0
α_3	-0.5	-0.5020	0.0134	0

2.4.2 Case 2: Two Change Points

In this case, x_1, and x_2 were simulated from uniform $[\pi, 3\pi]$ distribution with sample size $n = 500$. The following three models are considered:

- SICM

$$y_i = \sin(\alpha_1 x_{1i} + \alpha_2[x_{1i} - \theta_1]_+ + \alpha_3[x_{1i} - \theta_2]_+ + \alpha_4 x_{2i}) + \epsilon_i$$

  where ϵ_i is generated from Normal distribution with 0 mean and 0.05 standard deviation and θ_1 and θ_2 are two change points. The first component of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$, α_1, is fixed at 1 for identifiability reasons;

- GLM

$$\log(\mu_i) = \alpha_1 x_{1i} + \alpha_{2i}[x_{1i} - \theta_1]_+ + \alpha_3[x_{1i} - \theta_2]_+ + ns(x_{2i})$$

  where ns stands for natural splines smoothing function. The response variable y_i is generated from a Poisson distribution with μ_i;

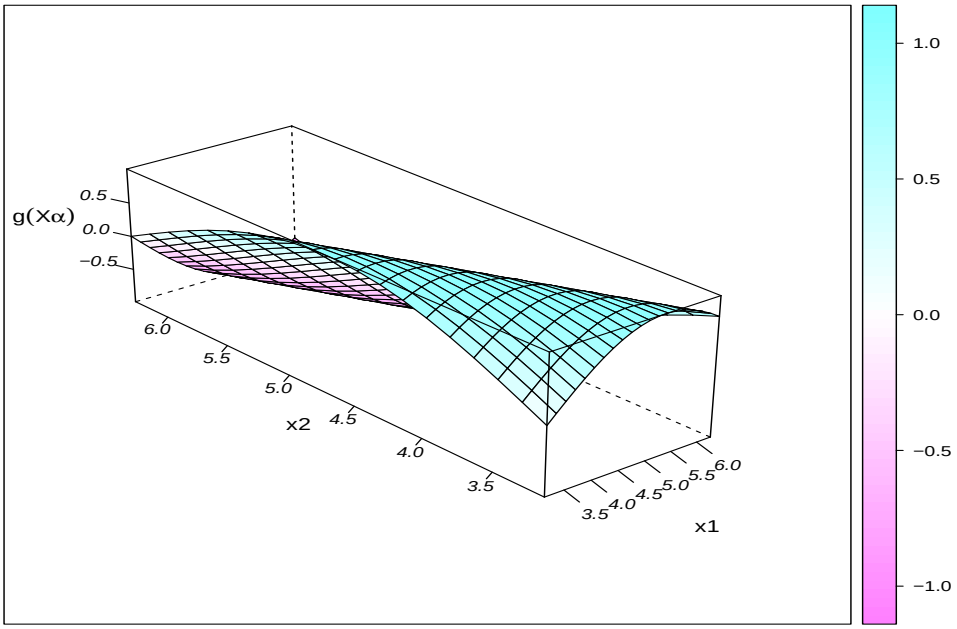


Figure 2.1: True single index function as a function of x_1 and x_2

- GAM

$$\log(\mu_i) = \alpha_1 x_{1i} + \alpha_2[x_{1i} - \theta_1]_+ + \alpha_3[x_{1i} - \theta_2]_+ + \alpha_4 x_{2i}$$

where $\alpha_1, \alpha_2, \alpha_3$, and α_4 are the coefficients and θ_1 and θ_2 are two change points. The response variable y_i is generated from a Poisson distribution with μ_i.

For power, $\theta_1 = 4$, $\theta_2 = 6$, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T = (1, -2, 1.5, 1)^T$. For Type I error, we used the same $\boldsymbol{\alpha}$ except that $\alpha_2 = \alpha_3 = 0$. We estimated power and Type I error using permutation based testing procedures. These values are summarized in Table 2.4 and Table 2.5.

Table 2.4 shows the power of the test and the mean estimates of θ_1 and θ_2 along with their standard errors. When the data were simulated from SICM, GAM and GLM have a very

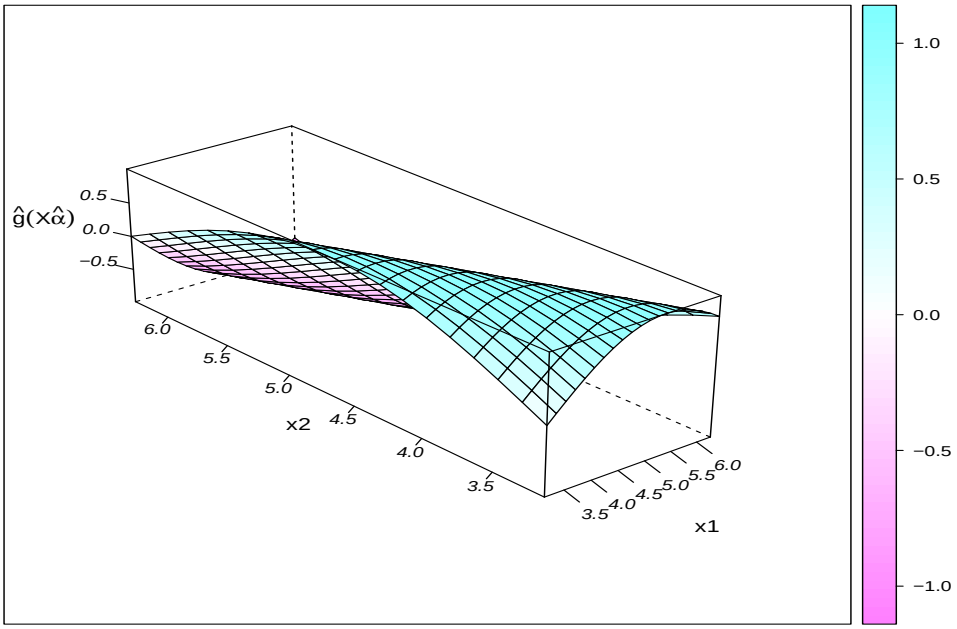


Figure 2.2: Estimated single index function as a function of x_1 and x_2

small power, 0.02 and 0.01 respectively, but SICM has power of one. Compared to one change point, one can conclude that the more change points there are under a misspecified model, the worse power there is for GAM and GLM. The standard errors for the θ_1 and θ_2 are small for SICM compared to the other two models. For data generated from GAM and GLM, the three models performs well in terms of power. In terms of standard error, SICM performs better for θ_2 than the other two models, which perform better standard error for θ_1.

Table 2.5 reveals that SICM performs well in terms of Type I error. Again as in Case 1, when the data is generated from SICM, the power and Type I error for GAM and GLM in Table 2.4 and Table 2.5 are almost the same. This means the probabilities of rejecting H_0 are similar whether H_0 is true or not. This means if we misspecified and used GAM or GLM as the model for our data set, our decision regarding H_0 is not correct.

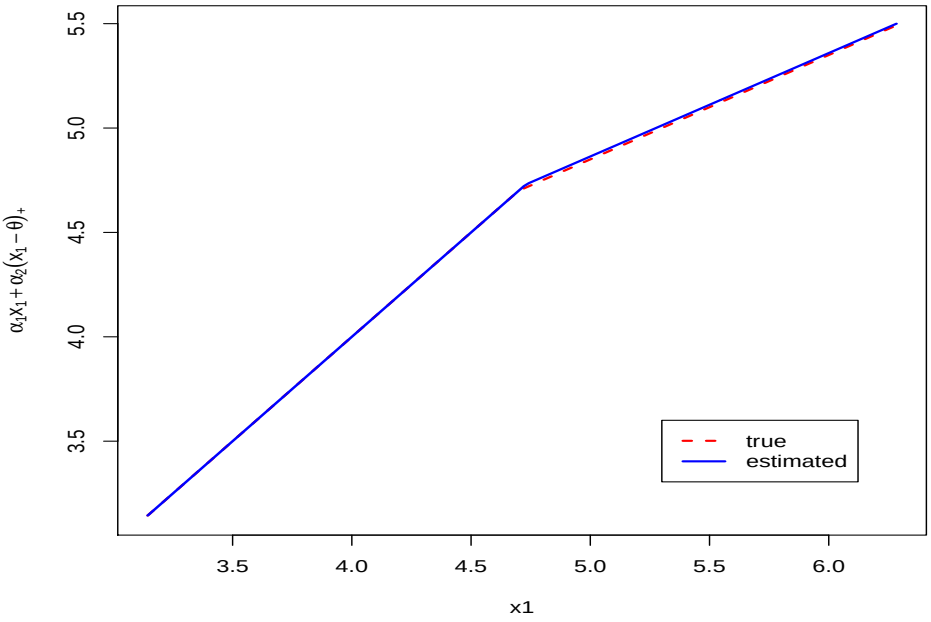


Figure 2.3: True and estimated change point

2.5 Real Data Application

We apply our approach to the mortality data obtained from Seoul, Korea from January, 2000 to December, 2007. Our goals are (1) to model the relationship between the weekly non accident mortality $\mathbf{Y}$ and mean temperature ($meantemp$) adjusting for other covariates such as mean humidity ($meanhumi$), mean pressure ($meanpress$), and month as a factor (m) and (2) to test whether there are change points on temperature in the nonlinear relationship between temperature and mortality. In this weekly data, we have four explanatory variables, $p = 4$, and the sample size is 417 observations. The model without change points has the form

$$\mathbf{Y} = g[\alpha_1 x_1 + \alpha_2 lag(x_2, 1) + \alpha_3 x_3 + \alpha_4 x_4] + \boldsymbol{\epsilon},$$

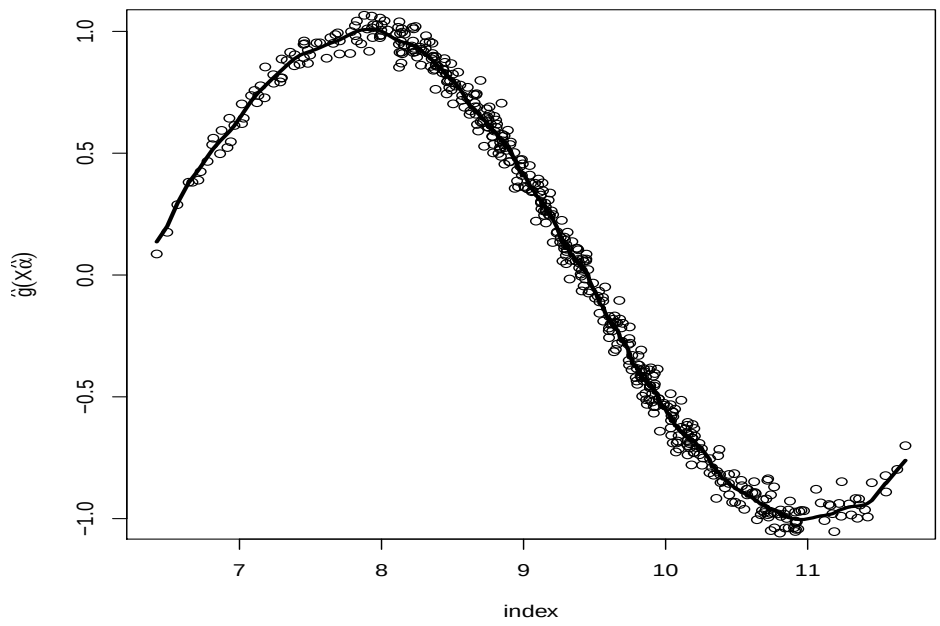


Figure 2.4: Scatterplot of index versus actual and fitted y

where x_1: mean pressure, x_2: mean temperature, x_3: month, and x_4: mean humidity. $\boldsymbol{\alpha}^T = (1, \alpha_2, \alpha_3, \alpha_4)$ is the single index vector of coefficients. α_1 set equal 1 for identifiability. ϵ_i is assumed to have mean 0 and variance σ^2 for all i. The model with two change points has the form

$$\mathbf{Y} = g[\alpha_1 x_1 + \alpha_2 lag(x_2, 1) + \alpha_3 (x_2 - \theta_1)_+ + \alpha_4 (x_2 - \theta_2)_+ + \alpha_5 x_3 + \alpha_6 x_4] + \boldsymbol{\epsilon}$$

where θ_1 and θ_2 are two change points.

Figure 2.5, the scatter plot of mean temperature and non accident mortality for the weekly data, shows a clear nonlinear pattern for the relationship with some outliers. After removing these potential outliers, we have Figure 2.6 which gives a clear nonlinear pattern for the

Table 2.4: Power of the three models based on 100 data sets simulated from each model and the estimates of mean and standard error (SE) of two change points, θ_1 and θ_2 in case H_0 : there is no change point $k_0 = 0$, H_1: there are two change points, $k_1 = 2$; SICM= single index change point model; GAM= generalized additive model; GLM= generalized linear model. When true model is SICM with normal error, we fit the GAM and GLM using identity link function.

Data simulated from	Estimated model	Power	$\theta_1 \pm SE$	$\theta_2 \pm SE$
SICM	SICM	1	4.09±0.03	6.004±0.02
	GAM	0.02	4.13±0.67	5.79±0.66
	GLM	0.01	4.13±0.67	5.79±0.66
GAM	SICM	1	4.07±0.15	6.03±0.02
	GAM	1	4.001±0.03	6.001±0.03
	GLM	1	4.002±0.06	6.001±0.03
GLM	SICM	1	4.07±0.27	6.01±0.03
	GAM	1	4.003±0.05	6.006±0.07
	GLM	1	4.003±0.05	6.006±0.07

relationship. One can see, in figure 2.6, that there is a decreasing pattern between non accident mortality and the mean temperature less than about 15 °C, a sharp decreasing in the interval between 15 °C and 23 °C, and then rapidly the non accident mortality increases after about 23 °C.

The permutation test is used to test whether there are change points for the weekly data with outliers and the weekly data without outliers. The permutation test results are shown in Table 2.6 which summarizes the change points detected in the alternative hypothesis, the p-value of the test, and the R-square of the model associated with the change points in the alternative model.

For the weekly data with outliers (WDWO), we failed to reject $H_0 : K_0 = 0$ vs. $H_1 : k_1 = 2$ with p-value=0.092, so we tested $H_0 : K_0 = 0$ vs. $H_1 : k_1 = 1$. H_0 is rejected with p-value=0.006. So we conclude that there is one change point which is located at -2.4°C. But one can see, from the index coefficient estimates in Table 2.7, the non accident mortality

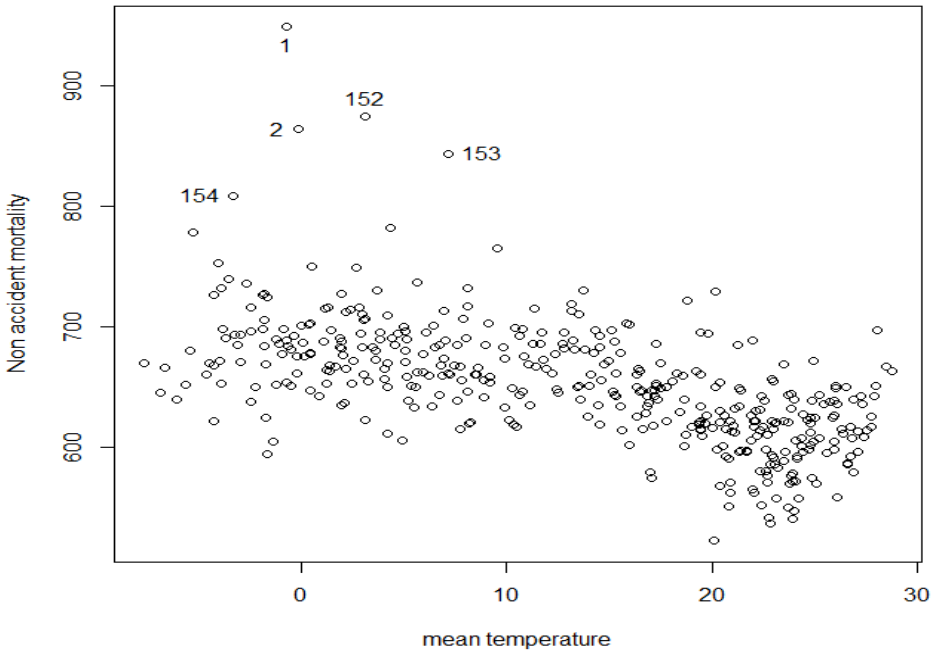


Figure 2.5: Scatter plot of mean temperature and non accident mortality of weekly data with the outliers points

Table 2.5: Type I error rate of the three models based on 100 data sets simulated from each model in case H_0 : there is no change point $k_0 = 0$, H_1: there are two change points, $k_1 = 2$; SICM= single index change point model; GAM= generalized additive model; GLM= generalized linear model. When true model is SICM with normal error, we fit the GAM and GLM using identity link function.

Data simulated from	Estimated model	Type I error rate
SICM	SICM	0.06
	GAM	0.03
	GLM	0.02
GAM	SICM	0.06
	GAM	0.01
	GLM	0.01
GLM	SICM	0.07
	GAM	0.04
	GLM	0.04

is increasing before -2.4 °C with slope 6.2149 and decreasing after that point which is not supported by the scatter plot, in Figure 2.5. It seems that the outliers are highly influential points. As a result, we removed those points and re-run the permutation test again for the weekly data without outliers (WDWOO). For WDWOO, we rejected $H_0 : K_0 = 0$ versus $H_1 : k_1 = 2$ with p-value=0.001. Then, we tested $H_0 : K_0 = 1$ versus $H_1 : k_1 = 2$. Again, H_0 is rejected with p-value=0.021. So, we have two change points in the WDWOO with R^2=0.61. These two change points are given in Table 2.6 as 15.8°C and 23.1°C.

Table 2.6: P-value, change points detected under H_1, and R^2 for weekly data with outliers (WDWO) and without outliers (WDWOO)

	WDWO		WDWOO	
No. of change points in H_0 and H_1	0 vs 2	0 vs 1	0 vs 2	1 vs 2
p-value	0.092	0.006	0.001	0.021
Change points detected in H_1	3.8, 18.6	-2.4	15.8,23.1	15.8,23.1
R^2	0.778	0.715	0.729	0.610

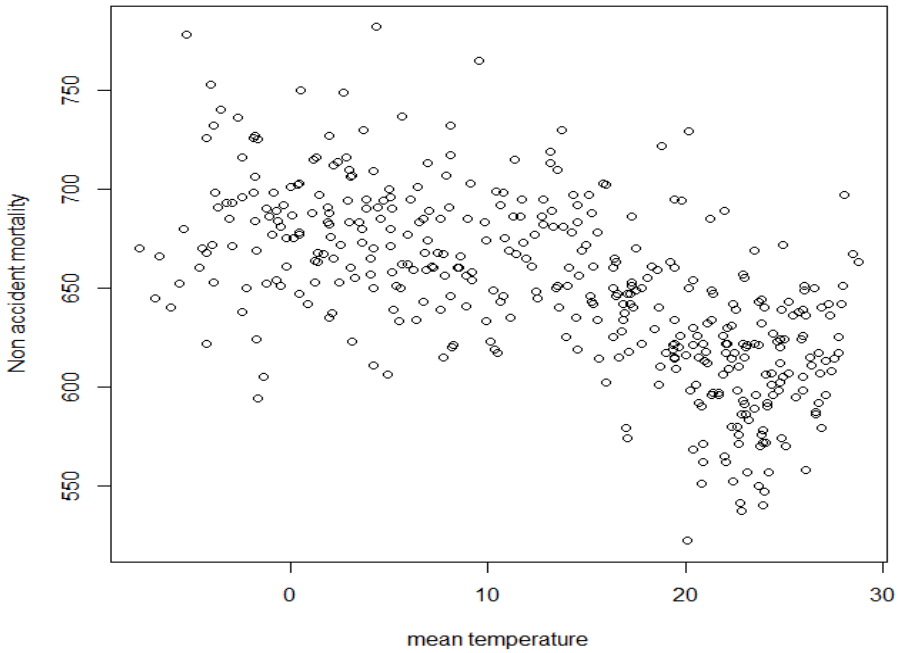


Figure 2.6: Scatter plot of mean temperature and non accident mortality of weekly data without the outliers points

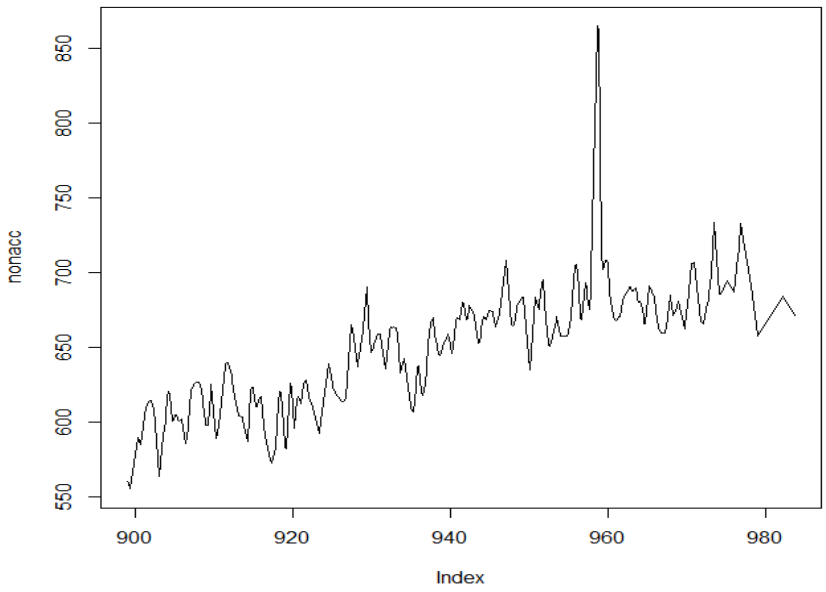


Figure 2.7: Estimated non accident mortality (nonacc) function for weekly data with the outliers points

Table 2.7 shows that using WDWOO, the non accident mortality decreases in the first interval before 15.8 °C, rapidly decreases in the second interval (15.8 °C, 23.1 °C), and increases after 23.1 °C. Figure 2.7 and 2.8 are for the single index function for both the weekly data with and without outliers points respectively. They are different which also supports that the four points removed from the data are high influential points and these points masked the pattern. In this study, we found two change points using SICM, while the previous studies for different countries data sets found only one change point using different models from ours, for example, Armstrong (2006), Chang et al. (2010), El-Zein et al. (2004), Hashizume et al. (2009), and Kan et al. (2007).

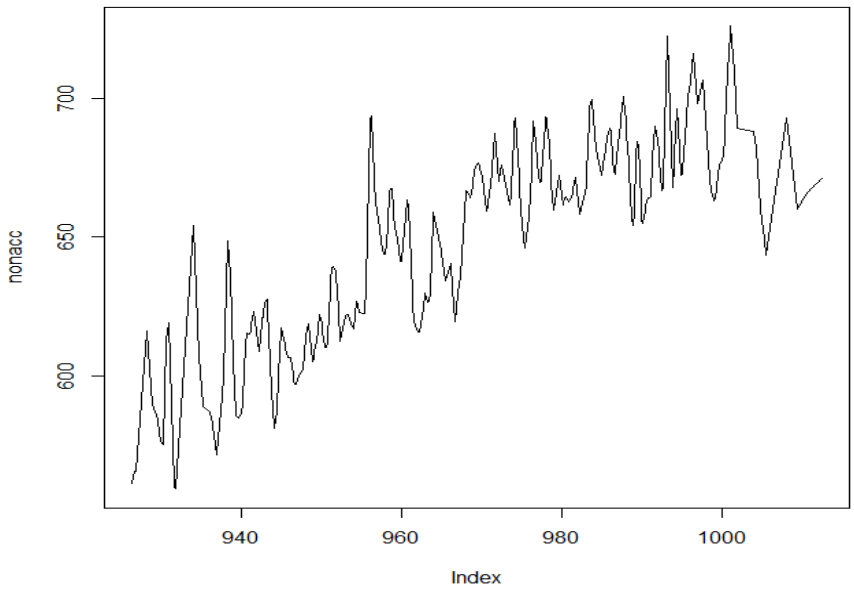


Figure 2.8: Estimated non accident mortality (nonacc) function for weekly data without the outliers points

2.6 Summary

We have proposed the single index change point model to simultaneously estimate the non-linear relationship and detect change points. The permutation based testing procedure is applied to identify the number of change points. We have investigated the performance of our model in detecting the change point(s) and compared it to the performance of the generalized linear model and additive generalized linear model. Simulation result suggests that the SICM is better than the other two models in terms of Type I and power. We examined the asymptotic efficiency of the permutation procedure in selecting the number of change points in the single index model. Under regularity conditions, we proved that the test results would converge in probability to the true number of change points.

Our procedure is illustrated with the real data example of Seoul, Korea collected from January, 2000 to December, 2007. Using the permutation test procedure, two change points

Table 2.7: Index coefficients estimates for weekly data with outliers (WDWO) and weekly data without outliers (WDWOO); Using WDWO, one change point at -2.4°C is detected. Using WDWOO, two change points at 15.8°C and 23.1°C are detected

	WDWO	WDWOO
x_1	1	1
$lag(x_2, 1)$	6.2150	-1.0351
$(x_2 - \theta_1)_+$	-7.4918	-0.8974
$(x_2 - \theta_2)_+$	–	5.0487
x_3	0.1409	0.4527
x_4	-0.6860	-0.5962

were detected at 15.8 °C and 23.1 °C, while the previous results found only one change point. In this chapter we worked with data from only one city. In the next chapter, Chapter 3, we develop single index model to incorporate spatial dependency from several other cities. And in Chapter 4 we will include the time effects into the model in addition the spatial correlated effects.

Chapter 3

Semiparametric Spatial Single Index Models

3.1 Background

Spatially correlated data can arise in many fields, such as econometrics, epidemiology, environmental science, image analysis, oceanography and many others. A main question of interest was how to incorporate this spatial correlation into statistical analysis. Cressie (1993) and Sherman (2011) showed that when statistical modeling procedure ignores spatial correlation, the standard error of the parameter estimates is deflated, and therefore the statistical significance gets inflated. Parametric models for spatially correlated data (Anselin and Florax, 1995; Cressie, 1991; Guyon, 1995; Possolo, 1991; Ripley, 1981; Zhang, 2002) were developed to take account spatial correlation into statistical analysis. However, there are quite limited approaches (Gu and Ma, 2005; Pang and Xue, 2012) on nonparametic models although nonparametric regression has become a standard statistical method when the functional form is possibly neither linear nor nonlinear of a specific type. In this chapter, we

want to model the relationship between the response variable Y and some covariates, $X's$, by incorporating correlated spatial random effects in single index nonparametric model, that is, we want to estimate $E[Y|X]$, nonparametrically for spatially correlated data.

In the nonparametric regression without considering spatial correlation, $Y = m(x)+\epsilon$, where $m(x)$ is unknown function of univariate variable x. One of the most common approach to estimate $m(x)$ is the local constant estimator, known as Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964),

$$\hat{m}(x) = \frac{\Sigma_{j=1}^{n} f(\frac{x-X_j}{h})}{\Sigma_{j=1}^{n} f(\frac{x-X_j}{h})},$$

where $\hat{m}(x)$ is the estimation of the unknown function at value x, $f(\cdot)$ is the kernel function and h is the bandwidth which controls the degree of smoothness. Under the assumption that error is independently and identically distributed, this estimator is consistent and is following asymptotically normal distribution under regularity conditions. In addition, consistency and asymptotic theory for this estimator has been driven by Robinson (2009).

For correlated data such as longitudinal data and repeated measures, mixed effects models are extensively studied. The linear and nonlinear mixed effects models were explained in detail (Harville, 1977; Lindstrom and Bates, 1990; Ke and Wng, 2001; McCulloch, 2003). For cross-sectional data, Gu and Ma (2005) proposed the nonparametric mixed effects model

$$\mathbf{Y} = m(X) + Z^T\mathbf{u} + \epsilon,$$

where the regression function $m(\cdot)$ is assumed to be a smooth function on a domain X. This unknown function $m(X)$ is estimated for univariate variable. Hence, when the dimension of X is high, the "curse of dimensionality" will occur. Pang and Xue (2012) introduced the single index model with random effects to incorporate the random effects into the model and

to solve the curse of dimensionality problem. This model can be written as

$$y_{ij} = m(X_{ij}^T \boldsymbol{\beta}) + Z_{ij}^T b_i + \epsilon,\ i = 1, \ldots, I,\ j = 1, , \ldots, J,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ index coefficients vector associated with the covariates X_{ij} and $\|\beta\| = 1$ is used as single index restriction for identifiability, the $b_i's$ are independent random effects with mean $\mathbf{0}$ and covariance matrix $\sigma_b^2 I$, $m(\cdot)$ is an unknown function, and $\epsilon_i's$ are independent mean $\mathbf{0}$ and variance $\sigma_\epsilon^2 I$. Pang and Xue (2012) used the Generalized Estimating Equation (GEE) method to estimate the single-index model with the random effect. However there are several limitations of this model: (1) it assumes that random effects are independent of each other; (2) $\mathbf{Y}$ follows normal distribution; and (3) it assumes that nonparametric $m(\cdot)$ and random effects can be additively separated. Because of those limitations, we are not able to use this model for our motivated example.

In our motivated example, we would like to model the relation between the non-accident mortality and the weather variables: mean temperature, mean humidity, mean pressure, and the month as a factor from six cities. The data cover the time from January, 2000 to December, 2007. One of questions of interests is to find the relationship between the non-accident mortality and the weather variables by incorporating spatial correlation. Hence we consider generalized single index model with the random effects. However our spatial random effects are not independent. They are spatially dependent having a parametric form. Since our Y is non accident mortality, we assume that it follows Poisson distribution with mean $\mu_{ij} = m(X_{ij}^T \beta) + Z_{ij}^T b_i$. To incorporate spatial dependency into statistical analysis, we propose two models. One is to extend the generalized single index additive model which can be separated with spatially correlated random effects with $m(\cdot)$. However the other model we propose in this chapter is the nonparametric single index model which can not be separated with spatially correlated random effect with $m(\cdot)$. We propose this model

because nonparametric function $m(\cdot)$ is different from that of each spatial location. Hence the nonparametric function $m(\cdot)$ can not be separated from the spatially correlated random effects, that is, the mean function of this proposed model is $\mu_{ij} = m(X_{ij}^T\beta + Z_{ij}^T b_i)$. To the best of our knowledge, there is no such nonparametric single index model which can not be separated with spatial random effects. In later of this chapter, we show this nonseparable model provides not only accurate parameter estimation but also better prediction accuracy. Our two models, separable and nonseparable models, were estimated via the Markov Chain Expectation Maximization (MCEM) algorithm. After estimating the nonparametric function, we also provide the prediction accuracy to predict "unobserved" mortality at location s.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce the proposed two models. In Section 3.3, we describe how to estimate these two models using the MCEM algorithm. We explain the proposed two algorithms for estimating the two models. Section 3.4 describes bandwidth selection. Simulation studies are conducted in Section 3.5. Real data application is in Section 3.6. Concluding remarks and discussion are provided in Section 3.7.

3.2 Semiparametric Spatial Single Index Random Effects Models

In this section, two models are proposed to find the relationship between the response variable and the covariates for spatially correlated data. One is to separate unknown function $m(\cdot)$ and random effect, while the other can not be. We refer the first model to "Semiparametric spatial-separable single index model" and the second model to "Semiparametric spatial-

nonseparable single index model".

3.2.1 Semiparametric Spatial-Separable Single Index Model (SSS-SIM)

For a spatial location s, let $Y_i(s)$ denote the response variable and $x_{1i}(s), x_{2i}(s), ..., x_{pi}(s)$ be the p observable explanatory variables at location s, $(i = 1, \ldots, r)$, where r is the number of observation at location s. For simplicity, we will use $Y(s)$ and $X(s)$ and let $\{u(s), s \in R^2\}$ be an unobservable spatial random process such that $u(s)$ represents the random effect at site s of unknown or unobservable causes unaccounted for by the explanatory variables. Semiparametric spatial-separable single index random effects model is defined as follows:

$$\begin{aligned} Y(s)|\mu(s) &\sim \text{Pois}[\mu(\text{s})|\text{u(s)}] \\ \mu(s)|u(s) &= m[X(s)\boldsymbol{\alpha}] + Zu(s), \end{aligned}$$

where

1. X(s) is the explanatory variable matrix at location s, $m(\cdot)$ is an unknown function, and $\boldsymbol{\alpha}$ is a vector of single index coefficients parameters. We use a restriction on $\boldsymbol{\alpha}$ to solve the identifiability problem which is $\alpha_1 = 1$;

2. $\{u(s), s \in R^2\}$ is a Gaussian stationary process with $E[u(s)] = 0$ for all s and cov$[u(s+d), u(s)] = C(d)$ for all $s, d \in R^2$, where $C(\cdot)$ is some parametric covariance function and d is the distance between two locations;

3. Conditionally on $\{u(s), s \in R^2\ \}$, $\{Y(s), s \in R^2\ \}$ is independent Poisson process and the distribution of $Y(s)$ is specified by the conditional mean $E\{Y(s)|u(s)\}$.

In this model, the relationship between $E[Y(s)|u(s)]$ and $u(s)$ is linear. To estimate this model, we need to estimate $\boldsymbol{\alpha}$, $\mathbf{u}$, and also estimate the unknown function $m(\cdot)$. We estimate them simultaneously. In the single index model, some restrictions on $\boldsymbol{\alpha}$ are needed in order for it to be identifiable. One approach is to set one component of $\boldsymbol{\alpha}$ to be equal to one ((Ichimura, 1993; Sherman, 1994)). In estimating SSS-SIM, we also use $\alpha_1 = 1$.

The covariance function describes the spatial association between the random effects at any two locations in space, say $u(s)$ and $u(s')$:

$$\begin{aligned} \text{cov}(u(s), u(s')) &= \sigma_u^2 \Sigma(\rho_u) \\ \mathbf{u} &\sim \text{MN}[0, \sigma_{\text{u}}^2 \Sigma(\rho_{\text{u}})], \end{aligned}$$

where σ_u^2 is the variance of the random effects $\mathbf{u}$, ρ_u is the dependence range, and Σ is a parametric covariance function that depends only on the distance between any two locations s and s'.

3.2.2 Semiparametric Spatial-Nonseparable Single Index Model (SSN-SIM)

Unlike the Seimparametric spatial-separable single index model which has the linear relationship between $E[Y(s)|u(s)]$ and $u(s)$, Semiparametric spatial-nonseparable single index random effects model does not have the linear relationship. One of the advantages of this model is that it does not need restriction on the single index coefficients parameters for the identifiability problem, such as $\alpha_1 = 1$ or $\|\boldsymbol{\alpha}\| = 1$. We use $\mathbf{u}$ as the variable which has its coefficient is equal to be one. This enables us to estimate all the single index coefficients without having identifiability problem and estimate the unknown function without such assumptions.

With the same definition of $Y(s)$, $x(s)$'s, and $u(s)$, the model can be written as

$$\begin{aligned} \mathbf{Y}(s)|\boldsymbol{\mu}(s) &\sim \text{Pois}[\boldsymbol{\mu}(\text{s})|\text{u(s)}] \\ \boldsymbol{\mu}(s)|u(s) &= m[X(s)\boldsymbol{\alpha} + Zu(s)], \end{aligned}$$

where

1. X(s) is the explanatory variable matrix at location s, $m(\cdot)$ is unknown function, and $\boldsymbol{\alpha}$ is the single index coefficients parameters;

2. $\{u(s), s \in R^2\}$ is a Gaussian stationary process with $E[u(s)] = 0$ for all s and cov(u(s+d), u(s) = C(d) for all $s, d \in R^2$, where the covariance function is $C(\cdot)$ and d is the distance between the two models;

3. Conditionally on $\{u(s), s \in R^2\}$, $\{Y(s), s \in R^2\}$ is independent process and the distribution of $Y(s)$ is specified by the conditional mean $E[Y(s)|u(s)]$.

In this model, the random effects of the location s is included in the unknown function $m(\cdot)$ which needs to be estimated.

We also use the same covariance function described in Section 3.2.1 for the spatial association between the random effects at any two locations in space, say $u(s)$ and $u(s')$ as follows:

$$\begin{aligned} \text{cov}(u(s), u(s')) &= \sigma_u^2 \Sigma(\rho_u) \\ \mathbf{u} &\sim \text{MN}[0, \sigma_\text{u}^2 \Sigma(\rho_\text{u})], \end{aligned}$$

where σ_u^2 is the variance of the random effects $\mathbf{u}$, ρ_u is the dependence range, and Σ is a parametric covariance function that depends only on the distance between any two locations s and s'.

To guarantee that the covariance matrix is positive definite, the spatial covariance matrix $\Sigma(\rho_u)$ is assumed to be a known parametric form; see the common ones in Table 3.1. It often assumes that this process is stationarity and isotropy. This means that the covariance between any two points in this process depends on only the distance between two points.

Table 3.1: Correlation functions; $d =$ the distance between any two locations, s and s'.

	Correlation function
Exponential	$e^{-\|d\|}$
Gaussian	e^{-d^2}
Triangular	$(1 - \|d\|)_+$
Spherical	$(1 - \frac{3}{2}\|r\| + \frac{1}{2}\|r\|^3)I_{[-1,1]}(d)$

3.3 SSS-SIM and SSN-SIM Estimation

In this section, we first briefly explain MCEM algorithm. We provide how to choose our candidate distributions for the Metropolish-Hastings (M-H) step and then propose two MCEM algorithms; one is for SSS-SIM, and the other is for SSN-SIM.

3.3.1 MCEM Algorithm

This algorithm is commonly used in the GLMM estimation (McCulloch, 1994, 1997; Booth and Hobert, 1999; Caffo et al., 2005; Tan et al., 2007; An and Bentler, 2012). The EM algorithm consists of two steps; expectation (E-step) and maximization (M-step). Iterating between the two steps until convergence satisfies. In many cases, the E-step involves analytically intractable integrals, one approach is to approximate E-step using some Monte Carlo method. Incorporating the Monte Carlo step into EM algorithm gives MCEM algorithm. In our proposed models, the spatial random effects are not independent. They have mean

$\mathbf{0}$ and variance-covariance matrix $\sigma_u^2\Sigma(\rho_u)$ so that no closed form is available for the expectation. Hence we need to incorporate Bayesian MCMC to generate a random sample from the full conditional distribution of $\mathbf{u}$. We use the Metropolis-Hastings algorithm. Choosing the candidate or proposal function is very important in M-H algorithm. The complete-data log-likelihood for our both models, in general, is given by

$$\log f[\mathbf{Y}, \mathbf{u}|\boldsymbol{\mu}, \sigma_u^2\Sigma(\rho_u)] = \log f_{Y|u}[\mathbf{Y}|\mathbf{u}, \boldsymbol{\mu}] + \log f_u[\mathbf{u}|\sigma_u^2\Sigma(\rho_u)]$$

If we used a candidate as multivariate normal distribution $\mathrm{MN}(\mathbf{0}, \sigma_0^2\Sigma)$=$f(.|\sigma_0^2\Sigma)$, the probability of accepting a new value $\mathbf{u}^*$ with the current value being $\mathbf{u}$ is

$$\min\left\{\frac{f[\mathbf{Y}|\mathbf{u}^*, \boldsymbol{\mu}]f_u[\mathbf{u}^*|\sigma_u^2\Sigma(\rho_u)]}{f[\mathbf{Y}|\mathbf{u}, \boldsymbol{\mu}]f_u[\mathbf{u}|\sigma_u^2\Sigma(\rho_u)]}, 1\right\}.$$

If we use the single-component Metropolis-Hastings algorithm, i.e., at each iteration, we only update a single component, say the sth component $u(s)$, we will generate the candidate values from the conditional normal distribution of $N(0, \sigma_0^2\bar{\Sigma})$, where σ_0^2 is a proposal variance of the spatial random effects. This conditional normal distribution can be derived as follows:

Let $\mathbf{v}$=$(v_1, v_2, \ldots, v_n) = \left[\, v_1 \ \mathbf{v_2} \,\right]^T$ has a multivariate normal distribution with mean $\boldsymbol{\theta} = \left[\, \theta_1 \ \boldsymbol{\theta}_2 \,\right]^T$ and variance-covariance matrix $\sigma_0^2\Sigma$, where

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then the distribution of v_1 conditioned on $\mathbf{v_2} = \mathbf{a}$ is multivariate normal $(v_1|\mathbf{v_2} = \mathbf{a}) \sim \mathbf{N}(\bar{\theta}, \bar{\mathbf{\Sigma}})$, where $\bar{\theta} = \theta_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{a} - \boldsymbol{\theta_2})$ and covariance matrix $\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

In our models, spatial random effects have multivariate normal with mean $\mathbf{0}$. Hence the

conditional normal distribution of $u(s)$ given the other random effects has $N(\bar{\theta}, \sigma_u^2\bar{\Sigma})$, where $\bar{\theta} = \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{a})$ and $\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. The proposal distribution is conditional normal distribution $N(0, \sigma_0^2\bar{\Sigma})$, where σ_0^2 is the proposed variance of the random effects.

Because of the conditional independence, the acceptance probability can be simplified to

$$\min\left\{\frac{f[\mathbf{Y}(s)|u^*(s), \boldsymbol{\mu}(s)]f_u[u^*(s)|\bar{\theta}, \sigma_u^2\bar{\Sigma}(\rho_u)]}{f[\mathbf{Y}(s)|u(s), \boldsymbol{\mu}(s)]f_u[u(s)|\bar{\theta}, \sigma_u^2\bar{\Sigma}(\rho_u)]}, 1\right\},$$

where $f_u[u(s)|\bar{\theta}, \sigma_u^2\bar{\Sigma}(\rho_u)]$ is the conditional distribution of $u(s)$ given all the other spatial random effects.

For S spatial locations, assume $\boldsymbol{\mu}(s), (s = 1, \ldots, S)$, ρ_u, σ_u^2 $\Sigma(\rho_u)$, and the estimated unknown function, $\hat{m}(\cdot)$ are known, we perform a subroutine for the M-H algorithm to generate N samples for each spatial random effect from $f[\mathbf{Y}, \mathbf{u}|\boldsymbol{\mu}, \sigma_u^2\Sigma(\rho_u)]$. This subroutine procedure are summarized as follows:

Subroutine for M-H algorithm:

Step 0 Generate initial value for $\mathbf{u}$ and set $s = 1$ and $t = 1$;

Step 1 Generate a candidate value for spatial random effect at location, s, from $N(\bar{\theta}, \sigma_0^2\bar{\Sigma})$, $u^*(s)$ and generate a uniform(0,1) random value U,

$$\text{If U} < \min\left\{\frac{f[\mathbf{Y}(s)|u^*(s), \boldsymbol{\mu}(s)]f_u[u^*(s)|\bar{\theta}, \sigma_u^2\bar{\Sigma}(\rho_u)]}{f[\mathbf{Y}(s)|u(s), \boldsymbol{\mu}(s)]f_u[u(s)|\bar{\theta}, \sigma_u^2\bar{\Sigma}(\rho_u)]}, 1\right\},$$

then set $\mathbf{u}^{(t)} = [u^*(s), u(2), \ldots, u(S)]$. Otherwise, $\mathbf{u}^{(t)} = \mathbf{u}$ stays unchanged.

Step 2 Set s=s+1 and repeat Step 1 until all locations are visited;

Step 3 Set the current value of $\mathbf{u} = \mathbf{u}^{(t)}$ and set $s = 1$ and $t = t + 1$. Repeat Step 1-2;

Step 4 Repeat Step 1-3, N times.

Note that here we take a sample only after each coordinate has been visited and the first N_0 burn-in samples should be discarded. Geyer (1992) suggested using an N_0 that is between 1% and 2% of the run length N.

In our simulations in Section 3.5, we generate initial values for spatial effects from $N(0, 0.1)$. We choose initial values from normal distribution for $\mathbf{u}$ because if we started the algorithm using the same values for all $u_i's$, such as $\mathbf{u}^T = (u_1, \ldots, u_n) = (0, \ldots, 0)$, we will have an identifiability problem as explained in Section 3.3.3.

3.3.2 Estimation for Spatial-Separable Single Index Model

The complete-data log-likelihood for the first proposed model takes the form
$\log f\{\mathbf{Y}(s), u(s)|\boldsymbol{\mu}(s) = m[X(s)\boldsymbol{\alpha}]+Zu(s), \sigma_u^2, \Sigma(\rho_u)\} = \log f[\mathbf{Y}(s)|\boldsymbol{\mu}(s), u(s)]+\log f_u[u(s)|\sigma_u^2\Sigma(\rho_u)]$,
where $\mathbf{Y}(s) \sim \text{Pois}[\boldsymbol{\mu}(s)|\text{u(s)}]$, $u(s) \sim GP[0, \sigma_u^2\Sigma(\rho_u)]$, $\boldsymbol{\mu}(s)|u(s) = m[X(s)\boldsymbol{\alpha}] + Zu(s)$, and $\sigma_u^2\Sigma(\rho_u) = Cov[u(s+d), u(s)] = \sigma_u^2 exp(||d||^2/\rho_u)$ for all $s, d \in R^2$.

Proposed algorithm I

To run MCEM algorithm, we need to initialize $\boldsymbol{\alpha}$, $u(s)$, $(s = 1, \ldots, S)$, σ_u^2, $m(\cdot)$, and estimate ρ_u, $\hat{\rho}_u$. The proposed algorithm to estimate the first model is as follows:

Step 0 Initialize parameters:

1. Initialize $\sigma_u^{2(0)}$, $u(s)^{(0)}$;
2. $\mathbf{Y}(s)^* = \mathbf{Y}(s) - u(s)^{(0)}$;
3. By using Ichimura method, estimate $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(0)}$, and smooth $m(\cdot)$ using some bandwidth to get $\hat{m}(\cdot)^{(0)}$.

E-step Given all initials, generate N values for each location s, $u(s)_1, u(s)_2, ...,u(s)_N$ from $\log f\{\mathbf{Y}(s), u(s)|\boldsymbol{\mu}(s) = \hat{m}(s)^{(0)} + Zu(s), \sigma_u^{2(0)}, \Sigma(\hat{\rho}_u)\}$ via Subroutine for M-H algorithm which is described in 3.3.1

M-step Maximize $\frac{1}{N}\sum_{k=1}^{N} \log f\{\mathbf{u}_k|\sigma_u^{2(0)}\Sigma(\hat{\rho}_u)\}$:

1. Get $\sigma_u^{2(1)}$;
2. Calculate $u(s)^{(1)} = \frac{1}{N}\sum u(s)_k$ and $\mathbf{Y}(s)^* = \mathbf{Y}(s) - u(s)^{(1)}$;
3. Estimate $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(1)}$, and smooth the function $m(\cdot)$ to get $\hat{m}(\cdot)^{(1)}$.

Iterate E-step and M-step until convergence.

3.3.3 Estimation for Spatial-Nonseparable Single Index Model

The complete-data log-likelihood for the second proposed model takes the form $\log f\{\mathbf{Y}(s), u(s)|\boldsymbol{\mu} = m[X(s)\boldsymbol{\alpha}+Zu(s)], \sigma_u^2, \Sigma(\rho_u)\} = \log f[\mathbf{Y}(s)|\boldsymbol{\mu}(s), u(s)]+\log f_u[u(s)|\sigma_u^2\Sigma(\rho_u)]$, where $\mathbf{Y}(s) \sim \text{Pois}[\boldsymbol{\mu}(s)|\text{u(s)}]$, $u(s) \sim \text{GP}[0, \sigma_\text{u}^2\Sigma(\rho_\text{u})]$, $\boldsymbol{\mu}(s)|u(s) = m[X(s)\boldsymbol{\alpha} + Zu(s)]$, and $\sigma_u^2\Sigma(\rho_u) = Cov[u(s+d), u(s)] = \sigma_u^2\text{exp}(||\text{d}||^2/\rho_\text{u})$ for all $s, d \in R^2$.

Single index nonseparable model enables us to estimates all the parameters in the model without having restrictions on the parameters because the random effects already fix the problem of the identifiability where we consider the random effect is the variable which its coefficient is equal to one.

The previous proposed algorithm I does not work for this model because two issues arise: (1) intensive calculations: if we run the M-H algorithm 1000 times using a single component in case of having seven components, we need to fit the single index model 7000 times to run the MCEM algorithm only one time; (2) we can not separate the random effects from the single index coefficient parameters estimates in the M-H step, while we are comparing the

current and the last single component. This point can be explained as follows:
The acceptance ratio

$$\min\left\{\frac{f[\mathbf{Y}|\mathbf{u}^*,\boldsymbol{\mu}]f_u[\mathbf{u}^*|\sigma_u^2\Sigma(\rho_u)]}{f[\mathbf{Y}|\mathbf{u},\boldsymbol{\mu}]f_u[\mathbf{u}|\sigma_u^2\Sigma(\rho_u)]},1\right\}.$$

can be re-written for the SSN-SIM model as

$$\min\left\{\frac{f[\mathbf{Y}(\mathbf{s})|\mathbf{u}^*(\mathbf{s}),\hat{m}^*[X(s)\hat{\boldsymbol{\alpha}}^*+Z\mathbf{u}^*(s)]f_u[\mathbf{u}^*|\sigma_u^2\Sigma(\rho_u)]}{f[\mathbf{Y}(s)|u(s),\hat{m}[X(s)\hat{\boldsymbol{\alpha}}+Z\mathbf{u}(s)]]f_u[\mathbf{u}|\sigma_u^2\Sigma(\rho_u)]},1\right\}.$$

In this case, if the ratio is greater than one, it is not known whether $u(s)^*$ is better than $u(s)$ or it is due to the difference between $\hat{m}^*$ and $\hat{m}^*$ or due to the difference between $\hat{\boldsymbol{\alpha}}^*$ and $\hat{\boldsymbol{\alpha}}$. Hence we can not compare the two spatial random effects at the same values of the other model parameters.

To solve this problem, we use a linear approximation for the unknown function at value $[X(s)\alpha^{(0)}+Z(s)u^{(0)}]$ to separate the spatial random effect and the other model parameters which will be

$$\boldsymbol{\mu}(s) \;=\; m[X(s)\boldsymbol{\alpha}+Zu(s)]=\hat{m}(\cdot)+\hat{m'}(\cdot)[X(s)\boldsymbol{\alpha}+Zu(s)-X(s)\boldsymbol{\alpha}^{(0)}-Zu(s)^{(0)}],$$

where $\hat{m}(\cdot)$ is the estimate of the unknown function using a smoothing method such as p-spline, kernel smoothing or any other basis function and $\hat{m'}(\cdot)$ is the estimate of the first derivative of the unknown function. We use local linear kernel regression to estimate the function and its first derivative. In general, the estimator of the j^{th} derivative $m^{(j)}(x)$ at a point x is given by $\hat{m}^{(j)}(x)=j!\hat{\beta}_j(x)$ for the local polynomials of degree d of the form

$$m(x_i)=\beta_0+\beta_1(x_i-x)+\ldots+\beta_1(x_i-x)^d.$$

Finally we propose algorithm II for estimating SSN-SIM which is summarized as follows:

Proposed algorithm II

Step 0 Initialize parameters:

1. Initialize $\sigma_u^{2(0)}$, $u(s)^{(0)}$;
2. By using Ichimura method, estimate $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(0)}$, and then smooth $m(\cdot)$ using some bandwidth to get $\hat{m}(\cdot)^{(0)}$ and $\hat{m'}(\cdot)^{(0)}$, where $\hat{m'}(\cdot)^{(0)}$ is the estimate of the first derivative of $m(\cdot)$.

E-step Given $\hat{m}(\cdot)$ and $\hat{m'}(\cdot)$, and using the Taylor approximation of $m(\cdot)$, generate N values for each location s, $u(s)_1, u(s)_2, ...,u(s)_N$ from $\log f\{\mathbf{Y}(s), u(s)|\boldsymbol{\mu}(s) = \hat{m}(\cdot) + \hat{m'}(\cdot)[X(s)\boldsymbol{\alpha} + Zu(s) - X(s)\boldsymbol{\alpha}^{(0)} - Zu(s)^{(0)}], \sigma_u^{2(0)}, \hat{\rho}_u\}$ via subroutine of MH which is described in 3.3.1.

M-step Maximize $\frac{1}{N}\sum_{k=1}^{N} \log f\{\mathbf{u}_k|\sigma_u^{2(0)}\Sigma(\hat{\rho}_u)\}$:

1. Obtain $\sigma_u^{2(1)}$;
2. Calculate $u(s)^{(1)} = \frac{1}{N}\sum u(s)_k$;
3. Estimate $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(1)}$, and also get $\hat{m}(\cdot)^{(1)}$ and $\hat{m'}(\cdot)^{(1)}$.

Iterate E-step and M-step until convergence.

3.4 Bandwidth Selection

The smoothing parameter, h, plays an important role in estimating the unknown function. In this section we focus on bandwidth selection. For the local polynomial smoother, there are two parameters affect the smoothness: (1) d, the order of the local polynomial which

is related to the number of derivatives assumed of the regression function; and (2) h, the bandwidth which controls the amount of local averaging. When we increase the order of the local polynomial and decrease the bandwidth that reduces the bias of the estimate at the cost of increasing its variance. In this study we consider local linear regression, i.e. we set the order $d = 1$, and concentrate on selecting the bandwidth h. In SSS-SIM, we need to estimate the unknown function, $m(X\boldsymbol{\alpha})$. The smoothing parameter in this model is estimated simultaneously with the unknown function estimation by minimizing the $\sum_{i=1}^{n}[y_i - \hat{m}_h(X\boldsymbol{\alpha})]^2$ over bandwidths h in a given finite grid search. On the other hand, to estimate SSN-SIM we do not only need estimating the unknown function but also the first derivative estimation, $\hat{m}(\cdot)$ and $\hat{m}'(\cdot)$, because we use local linear approximation for the unknown function. Hence it is important to select a good bandwidth which is suitable for estimating the function and its derivative simultaneously. We use local linear Gaussian kernel regression to estimate the unknown function and its first derivative.

Let S_h to be $n \times n$ smoother matrix associated with $\hat{m}(\cdot)$, where the vector of fitted values is $\hat{m}(X\boldsymbol{\alpha} + Zu) = S_h\mathbf{Y}$. The bandwidth can be selected with several criteria, such as Cross-validation (CV), Generalized Cross-validation (GCV), mellow C_p, Akaike Information Criterion (AIC), and Corrected AIC. The forms of the bandwidth selectors are as follows:

$$\begin{aligned}
CV(h) &= \sum_{i=1}^{n}\left[\frac{y_i - \hat{m}_{h,i}(X\hat{\alpha})}{1 - S_{h,ii}}\right]^2, \\
GCV(h) &= \frac{RSS(h)}{[1 - n^{-1}df_{fit}(h)]^2}, \\
C_p(h) &= RSS(h) + 2\hat{\sigma}^2_{error}df_{fit}(h), \\
AIC(h) &= \log[RSS(h)] + \frac{2df_{fit}(h)}{n}, \\
AIC_c(h) &= \log[RSS(h)] + \frac{2[df_{fit}(h) + 1]}{n - df_{fit}(h) - 2},
\end{aligned}$$

where $RSS(h) = \sum_{i=1}^{n}[y_i - \hat{m}_h(X\hat{\alpha})]^2$, $\hat{\sigma}^2_{error} = RSS(h)/(df_{error}(h))$, $df_{fit} = tr(S_h)$, and

$df_{error} = n - tr(S_h)$.

We conduct a simulation study to find which criterion provides the best performance to choose bandwidth for smoothing the unknown function and its derivative simultaneously in SSN-SIM.

3.5 Simulations

In this section, four simulation studies were performed; the first is to evaluate which known bandwidth selector criterion selects the most appropriate bandwidth for SSN-SIM in subsection 3.5.1; the second is how well our proposed algorithm I estimates parameters of SSS-SIM which are described in subsection 3.5.2, while the third is how well our proposed algorithm II estimates parameters of SSN-SIM in subsection 3.5.3. The last is to compare SSS-SIM and SSN-SIM in case the unknown function is identity function, in subsection 3.5.4.

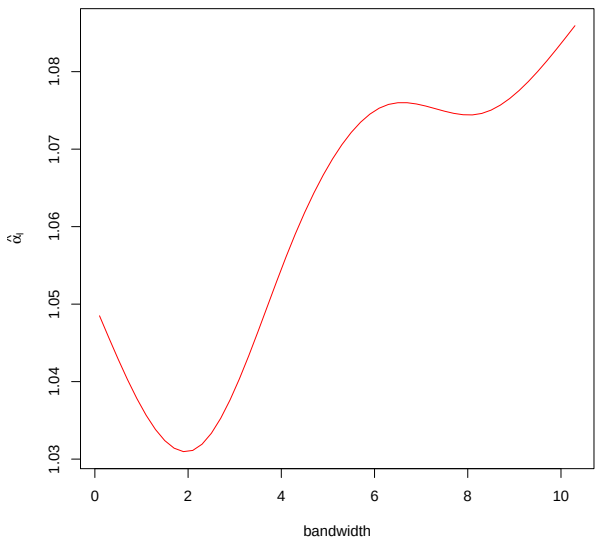
3.5.1 Comparison of Bandwidth Selection Criteria

We assume seven different locations which are the same number of locations in our motivated example. We then generate 60 observations for each location. We consider two variables x_1 and x_2, where x_1 is simulated from $Uniform(5, 20)$ and x_2 is simulated from $N(0, 1)$. We let $\boldsymbol{\alpha} = (1, 2)^T$, $\rho_u = 3$, and the variance of the random effects $\sigma_u^2 = 0.5$. The mean function $\mu(s)|u(s)$ is then $[x_1+2x_2+Zu(s)]^2$ and $Y(s)|u(s)$ generates from a Poisson distribution with mean $\mu(s)|u(s)$. We use a grid search for h in range (0,10.5) with increment 0.2. At each value of the grid, we replicated the simulation setting 100 times and estimated the model parameters for each simulated data. For the 100 estimates of each parameter, the mean, bias, variance, and mean square error were calculated to find the best bandwidth according

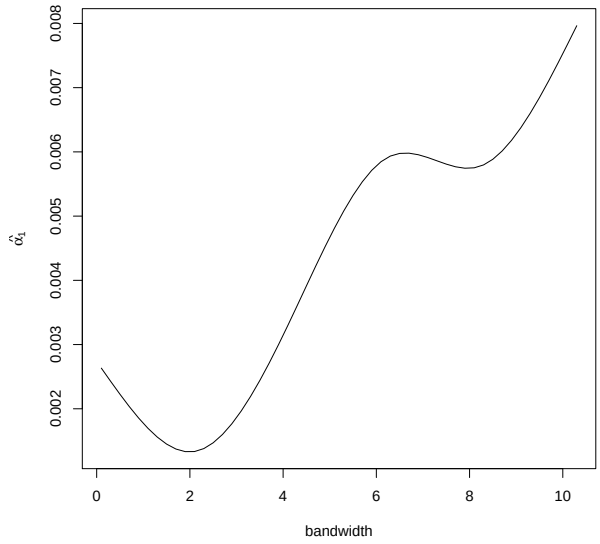
to these criteria. We then compare our selected bandwidth with the bandwidth selected from several known criteria to discover which known bandwidth selector criterion is more appropriate for our SSN-SIM.

After estimating the model 100 times, we removed the outliers from estimates of parameters, α_1, α_2, and σ_u^2, using IQR method. According to the mean criterion, plot (a) in Figures 3.1-3.3 show that a bandwidth 2 provides the most appropriate estimates for α_1, α_2, and σ_u^2, where the mean estimates are very close to the true values at this bandwidth. It is also can be seen from plot (b), in Figures 3.1-3.3, that the best bandwidth for α_1, α_2, and σ_u^2 is also bandwidth equal to 2 in terms of the bias criterion. At that bandwidth, the bias is the smalled for all the parameters. On the other hand, plots (c) and (d), in figures 3.1-3.3, reveal that variance and mean squares error (MSE) are minimum at bandwidth equal to 4.6 because bias is relatively much smaller than variance so that variance is dominated in MSE plots, plot (d) in figures 3.1-3.3. Hence we think that bandwidth equal to 2 is appropriate in the proposed algorithms, in terms of the mean estimate, to estimate the unknown function and its derivative.

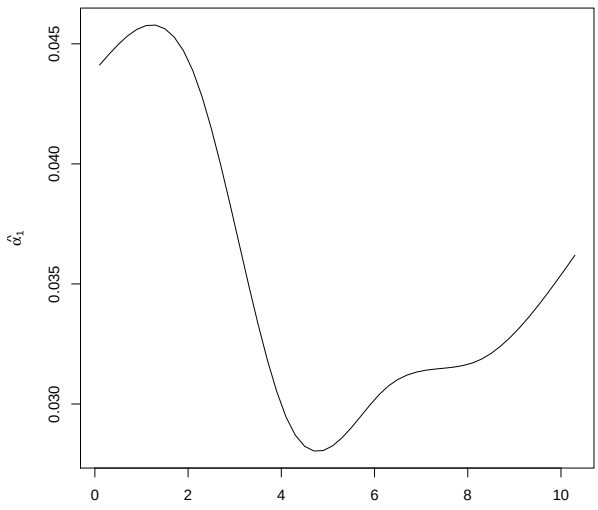
Next we need to know what is the best known bandwidth selector among cross-validation (CV), Akaike's information criterion (AIC), corrected Akaike's information criterion (AICc), and Mallows' Cp. This means that which one gives a value close to 2. Figure 3.4(a)-(e) are scatter plots of bandwidth versus the selectors values. We can see that they have the same performance. The best bandwidth for these methods is equal to 2. The best method which gives a minimum value at 2 is CV method.



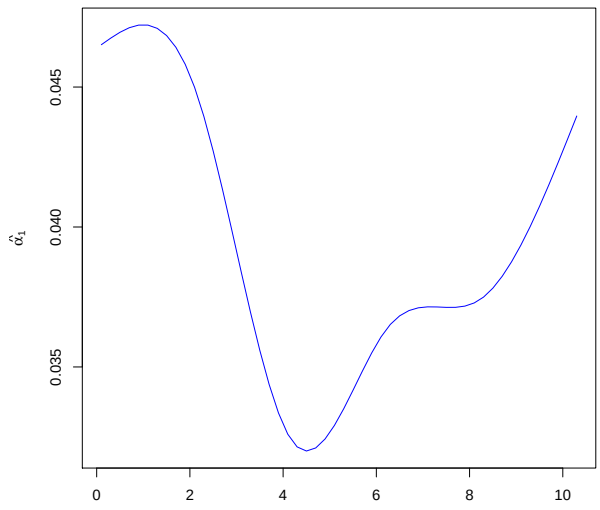
(a) Mean



(b) Bias

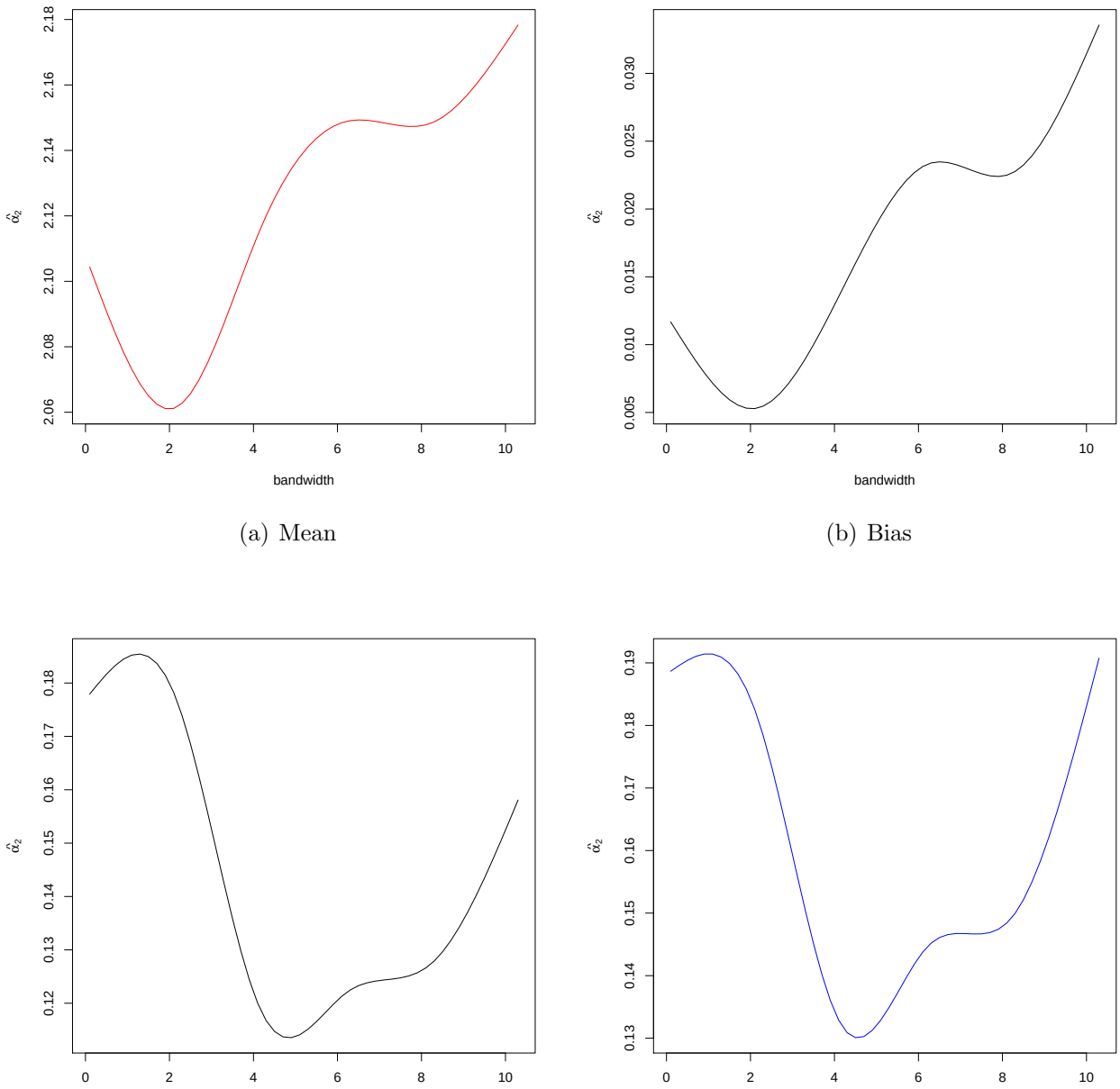


(c) Variance



(d) MSE

Figure 3.1: Plots of Mean (a), Bias (b), Variance (c), and MSE (d) for $\hat{\alpha}_1$ estimates as a function of bandwidth. 100 data sets were simulated at each value in grid range (0, 10.5) with increment equal to 0.2.



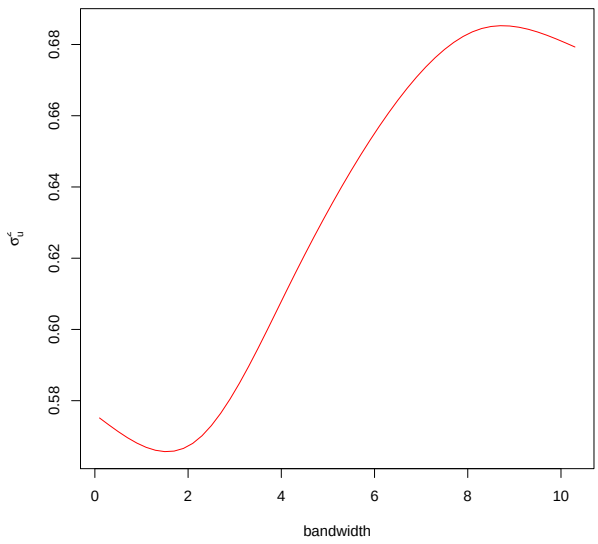
(a) Mean

(b) Bias

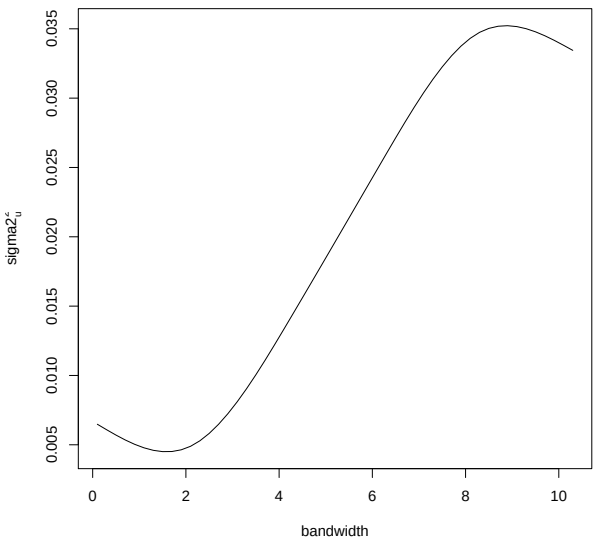
(c) Variance

(d) MSE

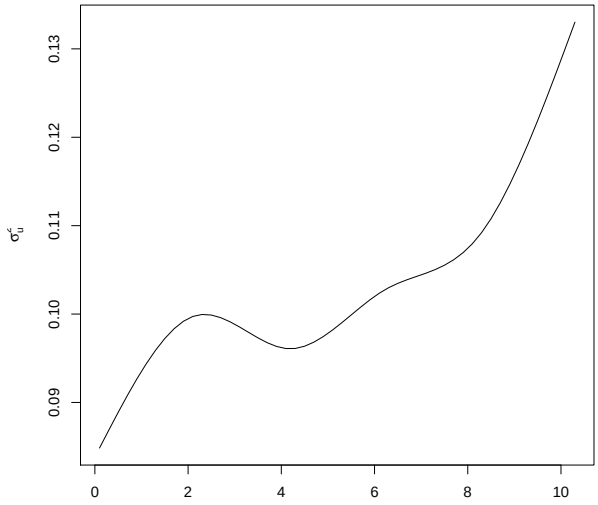
Figure 3.2: Plots of Mean (a), Bias (b), Variance (c), and MSE (d) for $\hat{\alpha}_2$ estimates as a function of bandwidth. 100 data sets were simulated at each value in grid range (0, 10.5) with increment equal to 0.2.



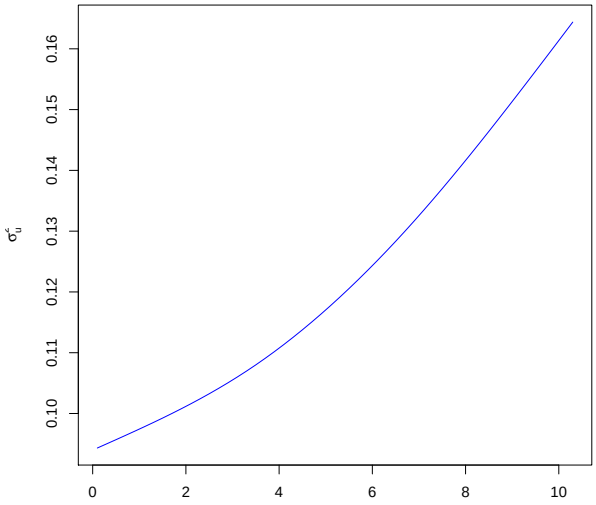
(a) Mean



(b) Bias



(c) Variance



(d) MSE

Figure 3.3: Plots of Mean (a), Bias (b), Variance (c), and MSE (d) for $\hat{\sigma}_u^2$ estimates as a function of bandwidth. 100 data sets were simulated at each value in grid range (0, 10.5) with increment equal to 0.2.

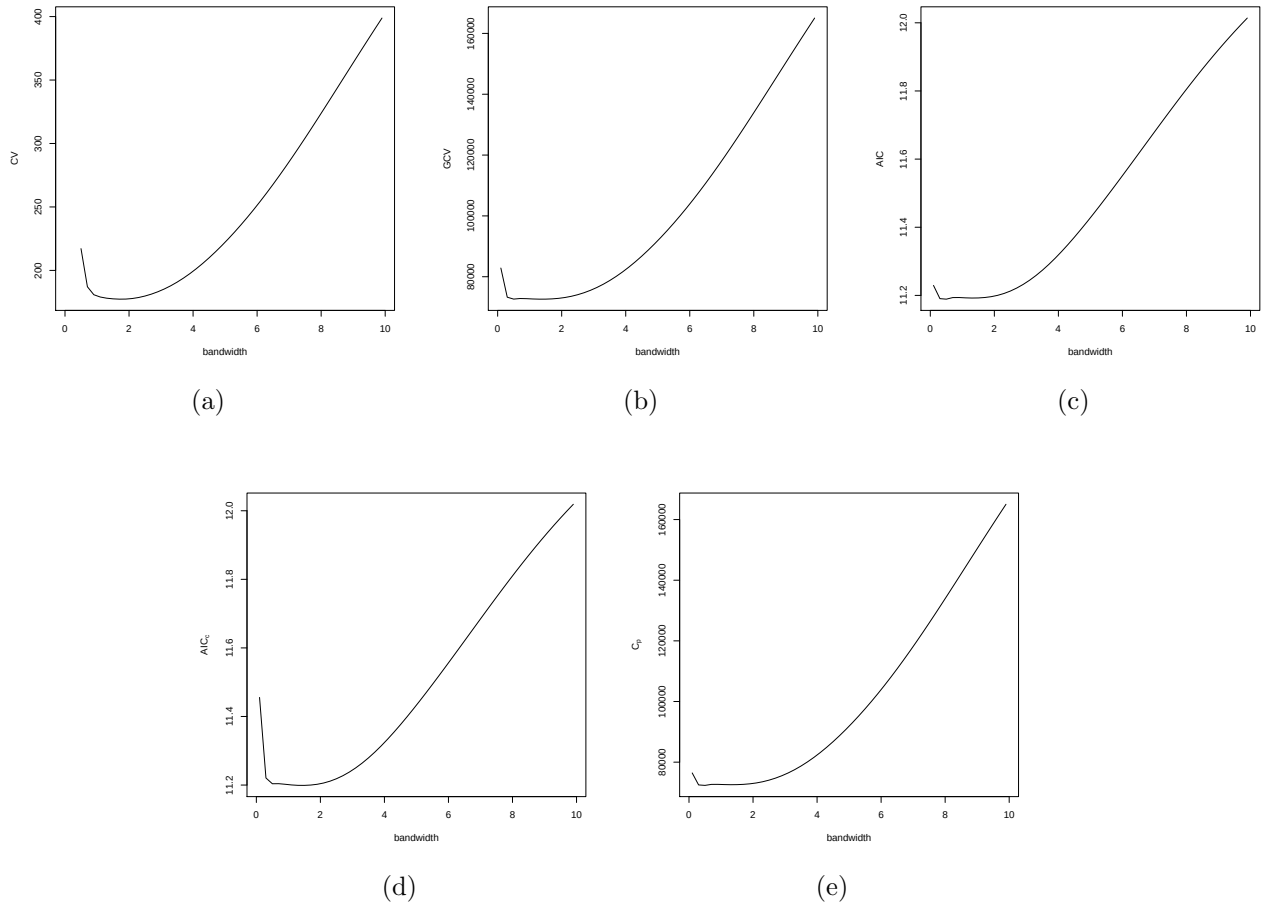


Figure 3.4: $CV(a), GCV(b), AIC(c), AIC_c(d)$ and $C_p(e)$ for bandwidth selection criterions

3.5.2 Parameters Estimation of SSS-SIM Using Proposed Algorithm I

We also conduct a simulation study to asses how well our proposed algorithm I estimates the parameters in SSS-SIM. We consider the same setting as before except for the mean function $\boldsymbol{\mu}(s)|u(s) = [x_1 + 2x_2]^2 + Zu(s)$.

In the proposed algorithm I for SSS-SIM, N is number of samples after 2% of the MCMC run length discarded. We run the M-H algorithm for 5000 iterations and discard 2% of

the run length. We noticed the convergence of the random effects happens quickly, so 2% is enough. The initial values were chosen to be from normal distribution, $N(0, 0.1)$. The complete MCEM algorithm was run for 15 times and the estimates have been taken that corresponding to the maximum likelihood function value.

Table 3.2 shows summary results of the 100 simulated data sets. One can see that estimated value of α_1 is equal to one because we set it to be one for the identifiability. For α_2, its mean is close to the true one and the standard error is small. Moreover, the mean and the median are equal which means the distribution of α_1 and α_2 estimates are approximately normal. For the variance of spatial random effects, the mean and the median are less than the true value. The standard error of the variance of spatial random effects estimates is larger than the standard error of the estimate of the single index coefficients parameters. It is also can be seen that the algorithm does not give many outlying values. These all results are summarized in Table 3.2. Figure 3.5 shows the boxplots of estimates of the parameters which supports the findings from Table 3.2.

Table 3.2: Summary results for parameters estimates of 100 data sets simulated from SSS-SIM; Mean, Standard Error (SE) Mean, Minimum, Median and Maximum

Parameter	True value	Mean	SE Mean	Minimum	Median	Maximum
α_1	1	1	0	1	1	1
α_2	2	2.026	0.010	1.681	2.022	2.302
σ_u^2	0.5	0.374	0.028	0.878	0.270	1.659

3.5.3 Parameter Estimation of SSN-SIM Using Proposed Algorithm II

We also conduct a simulation study to asses how well our proposed algorithm II estimates the parameters in SSN-SIM. We consider the same setting as Section 5.1 except for the mean

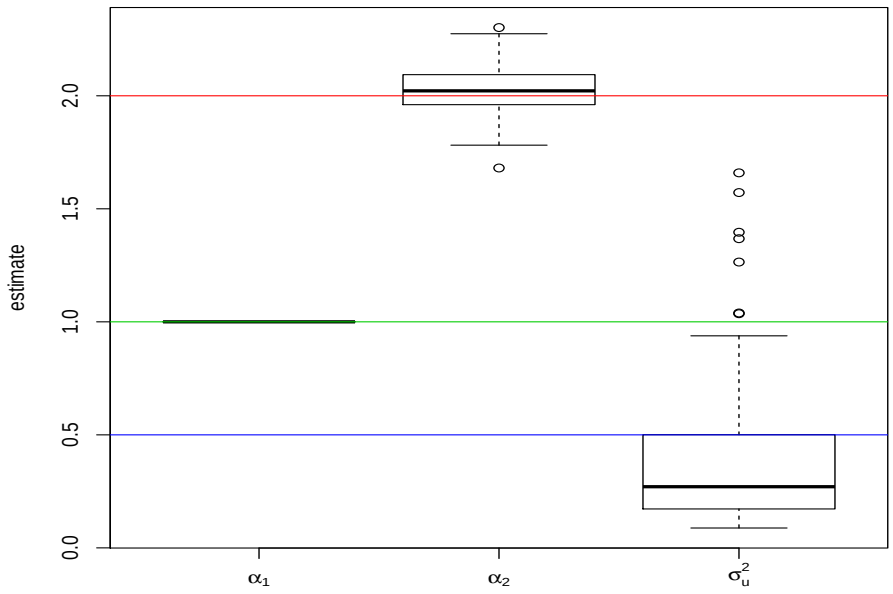


Figure 3.5: Boxplots of estimates of parameters; $\alpha_1, \alpha_2,$ and σ_u^2 in SSS-SIM using 100 simulated data sets. Red line represents the true value for α_2, green for true value of α_1, and blue for true value of σ_u^2

function $\boldsymbol{\mu}(s)|u(s) = [x_1 + 2x_2 + Zu(s)]^2$. We used a bandwidth equal to 2, $h = 2$, which is selected as the best according to the several criteria and the mean criteria as seen in Section 3.5.1.

The initial values were chosen from the normal distribution, $N(0, 0.1)$, for $u(s)$, because if we started the algorithm using the same values for all $u(s)'s$, such as $\mathbf{u}^T = (0, \ldots, 0)$, we will have an identifiability problem. We use $\mathbf{u}$ as the variable which has its coefficient is equal to one which can be used to be a vector of constants. This enables us to estimate all the single index coefficients without having identifiability problem and estimate the unknown function without more assumptions.

Table 3.3 has the results of the simulation study. This Table has estimates for all the parameters because we have no restrictions on the parameters. One can see that the mean estimates of all the parameters are greater than the true values. That is because this algorithm gives

some outliers which can be seen from the maximum values in Table 3.3 and Figure 3.6. If we removed those outliers using the Boxplot method, we will get a better estimates. By looking at the medians, it is clear that the estimates are very close to the true ones. From Table 3.2 and Table 3.3, the standard error of estimates for SSS-SIM are less than the standard errors for the estimates of SSN-SIM. On the other hand, it can be seen that the median (0.5564) of the variance for random effects for SSN-SIM is better than the median (0.27055) of SSS-SIM where it is close to the true one (0.5). This means SSN-SIM is better than SSS-SIM in terms of estimating the spatial random effects.

Table 3.3: Summary results for parameters estimates of 100 data sets simulated from SSN-SIM; Mean, Standard Error(SE) Mean, Minimum, Median and Maximum

Parameter	True value	Mean	SE Mean	Minimum	Median	Maximum
α_1	1	1.2920	0.0951	0.5036	1.0467	8.6407
α_2	2	2.5965	0.1921	0.9953	2.0937	17.3663
σ_u^2	0.5	0.8413	0.1100	0.0709	0.5564	9.9992

3.5.4 Parameters Estimation of SSS-SIM and SSN-SIM When the Unknown Function is the Identity Function

The main difference between the two models is in the mean function. Hence, we first investigate whether the two proposed algorithms give the same estimates under the identity function. We fit the two models in case of that the unknown function is the identity function. We generated 200 data sets from the same simulation settings as before except for the mean function is $\boldsymbol{\mu}|u(s) = X(s)\boldsymbol{\alpha} + Zu(s)$. In this case the two models should give the same results.

Table 3.4 shows that the mean estimates of α_1, α_2 and the variance of the spatial random effects σ_u^2 are close to the true values for SSS-SIM and SSN-SIM except for α_2 of SSN-SIM

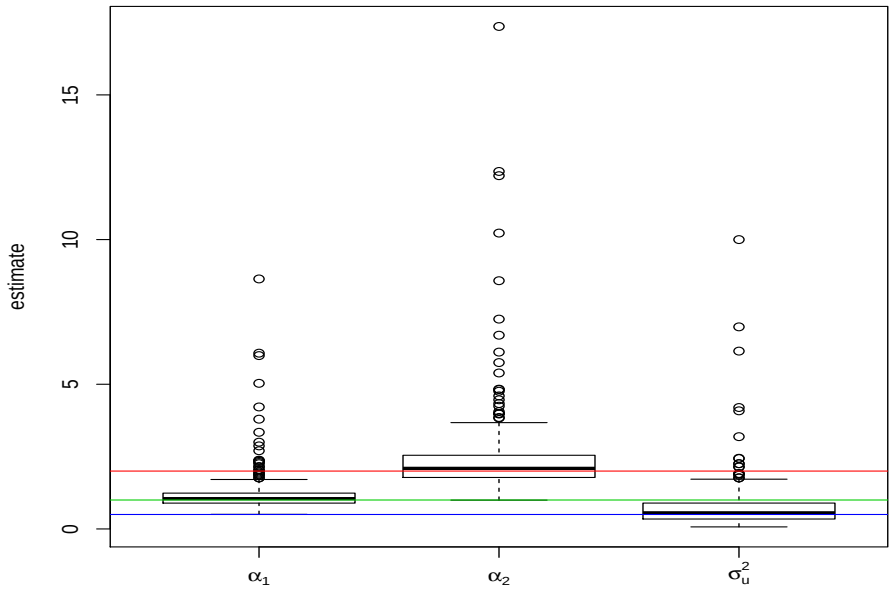


Figure 3.6: Boxplots of estimates of parameters; $\alpha_1, \alpha_2,$ and σ_u^2 in SSN-SIM using 100 simulated data sets. Red line represents the true value for α_2, green for true value of α_1, and blue for true value of σ_u^2

which is somewhat higher than the true value because of some outliers. The standard errors for both are small but the standard error of the mean for SSN-SIM is smaller than that of SSS-SIM. The median values are also close to the true values for both models. Figure 3.7 reveals that the distribution of estimates of the parameters is almost symmetric after discarding the outliers for both models. Figure 3.7 also shows that there are some outliers for both model parameters estimates. Figure 3.8 and Figure 3.9 are the scatterplots of the true spatial random effects and the estimated spatial random effects for the 200 simulated data sets from each model. One can see that the proposed algorithm II for SSN-SIM can estimate the spatial random effects better than the proposed algorithm I for SSS-SIM. The coefficients of determination between the true spatial random effects and the estimated are 0.97 and 0.99 for SSS-SIM and SSN-SIM, respectively.

Table 3.4: Summary results for parameters estimates of 200 data sets simulated from SSS-SIM and 200 data sets simulated from SSN-SIM; Mean, Standard Error (SE) Mean, Minimum, Median and Maximum when the unknown function is the identity function

Mean function	Model	Parameter	True value	Mean	SE Mean	Minimum	Median	Maximum
$X(s)\boldsymbol{\alpha} + Zu(s)$	SSS-SIM	α_1	1	1	0	1	1	1
		α_2	2	2.0004	0.0002	1.9910	2.0001	2.0001
		σ_u^2	0.5	0.5092	0.0185	0.0601	0.5255	0.4487
$X(s)\boldsymbol{\alpha} + Zu(s)$	SSN-SIM	α_1	1	1.0918	0.0134	0.5210	1.0667	1.8998
		α_2	2	2.1841	0.0269	1.0440	2.1320	3.7982
		σ_u^2	0.5	0.5836	0.0227	0.0367	0.5255	2.1845

3.6 Real Data Application

In this section, the two proposed models, SSS-SIM and SSN-SIM, will be applied to the South Korea data set; the dependence range, the models parameters and their confidence intervals, and the unknown semiparametric functions for each city will be estimated. In

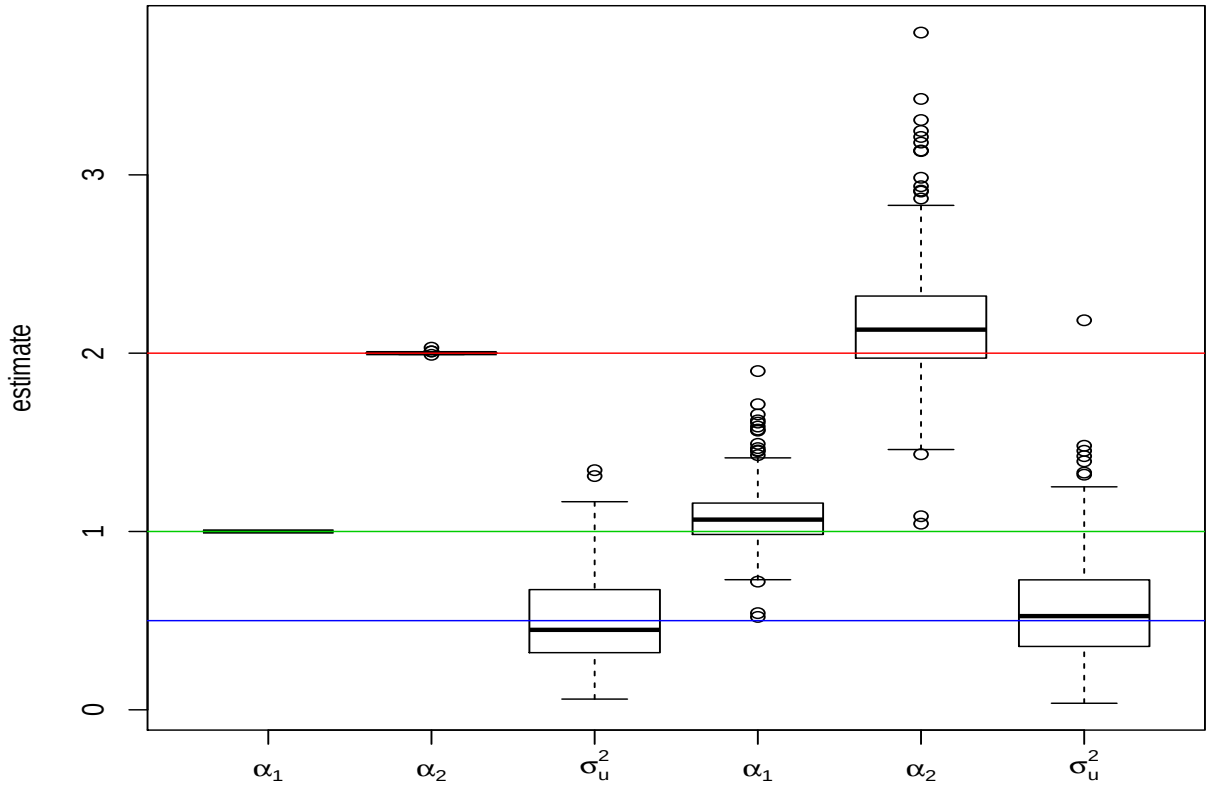


Figure 3.7: Boxplots of estimates of parameters; α_1, α_2, and σ_u^2 in SSN-SIM and SSS-SIM using 200 simulated data sets from each model. We draw six plots: the first three are for SSS-SIM and the next three are for SSN-SIM. Red line represents the true value for α_2, green for true value of α_1, and blue for true value of σ_u^2. Boxplot of α_1 for SSS-SIM appears as a horizontal line because it is set to be one for the identifiability problem in SSS-SIM

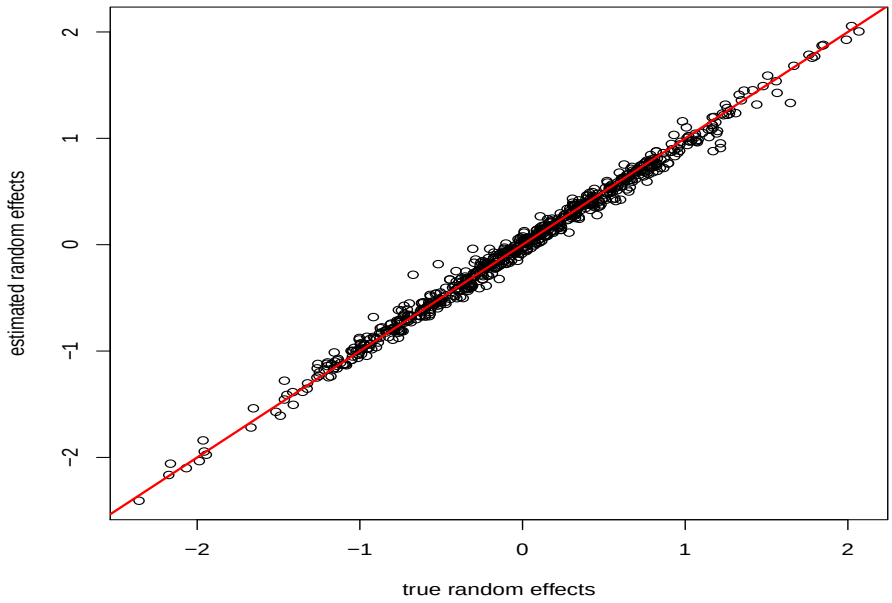


Figure 3.8: Scatterplot of the true random effects and the estimated random effects from SSS-SIM of 200 simulated data sets when the unknown function is identity function with 45^o line

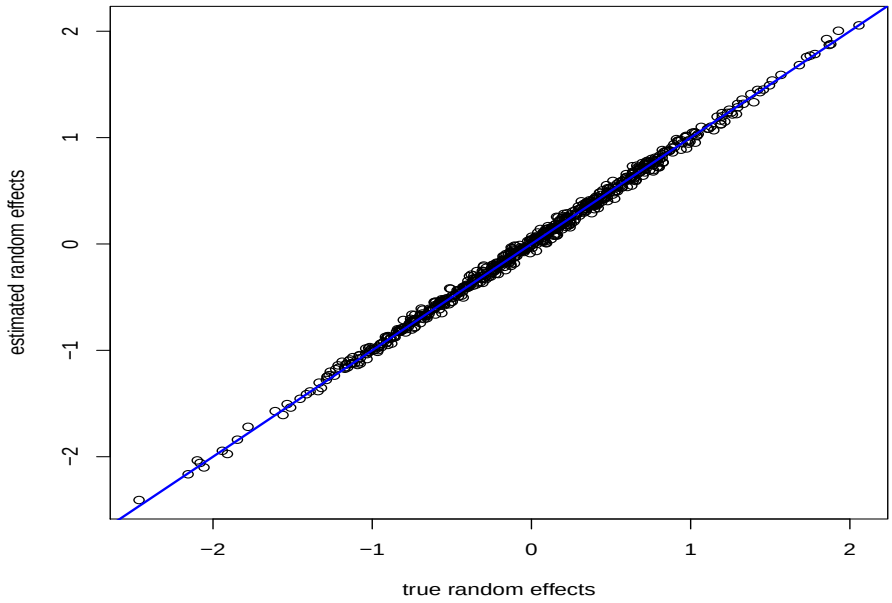


Figure 3.9: Scatterplot of the true random effects and the estimated random effects from SSN-SIM of 200 simulated data sets when the unknown function is identity function with 45^o line

addition, model selection will be done to select the suitable model for South Korea mortality data.

3.6.1 Data and Model

In South Korea non accident mortality and other covariates, mean temperature, mean humidity, mean pressure, time, has been recorded daily during the period from January, 2000 to December, 2007 for six major cities: Seoul, Incheon, Daejeon, Daegu, Gwangju, and Busan. So, we have in total 2922 observations for each city. The weekly data are created by calculating the mean of the weather variables per week. In this case we have 417 observation for each city. Those cities are different in population size. So, to get ride of the population size effect on non accident mortality, we divided weekly non accident mortality by population size of each city and multiplied by 1 million, so we get weekly non accident mortality per 1 million persons in each city. Figure 3.10 shows the characteristics of the 6 metropolitan areas in South Korea. Our goals are (1) including the correlated spatial random effects into single index model (2) estimating the two proposed models, SSS-SIM and SSN-SIM, (3) studying how spatial random effects affects to the estimation (4) addressing the prediction performance of the proposed models, and (5) selecting the best model for our data set.

The SIM and the proposed models are written as the following:

- SIM

$$\begin{aligned} Y(s)|\mu(s) &\sim \text{Pois}[\mu(\text{s})|\text{u(s)}], \\ \mu(s)|u(s) &= m[x_1(s)\alpha_1 + x_2(s)\alpha_2 + x_3(s)\alpha_3 + x_4(s)\alpha_4], \end{aligned}$$

- SSS-SIM

$$\begin{aligned} Y(s)|\mu(s) &\sim \text{Pois}[\mu(\text{s})|\text{u(s)}], \\ \mu(s)|u(s) &= m[x_1(s)\alpha_1 + x_2(s)\alpha_2 + x_3(s)\alpha_3 + x_4(s)\alpha_4] + u(s), \end{aligned}$$

- SSN-SIM

$$\begin{aligned} Y(s)|\mu(s) &\sim \text{Pois}[\mu(\text{s})|\text{u(s)}], \\ \mu(s)|u(s) &= m[x_1(s)\alpha_1 + x_2(s)\alpha_2 + x_3(s)\alpha_3 + x_4(s)\alpha_4 + u(s)], \end{aligned}$$

where

1. $x_i(s)$ is the explanatory variable x_i vector at location s, $m(\cdot)$ is unknown function, and $\alpha_1, \alpha_2, \alpha_3,$ and α_4 are the single index coefficients parameters.

2. $\{u(s), s \in R^2\}$ is a Gaussian stationary process with $E[u(s)] = 0$ for all s and cov(u(s+ d), u(s) = C(d) for all $s, d \in R^2$, where the covariance function is $C(\cdot)$ and d is the distance between the two locations.

3. Conditionally on $\{u(s), s \in R^2\}$, $\{Y(s), s \in R^2\}$ is independent process and the distribution of $Y(s)$ is specified by the conditional mean $E[Y(s)|u(s)]$.

4. x_1 : weekly mean temperature, x_2 :, weekly mean humidity, x_3 : weekly mean pressure, and x_4 : month.

3.6.2 Dependence Range, ρ_u

The semivariogram will be used to estimate the dependence range. A semivariogram is usually characterized by three measures: (1) nugget, (2) sill, and (3) range where the semivariogram plot stops increasing. First the nugget which refers to the variability in the field data cannot be explained by distance between the observations. There are many factors influence the magnitude of the nugget including imprecision in sampling techniques and underlying variability of the attribute that is being measured. In addition, the minimum spacing between observations can influence the nugget because if there are no observations located close to each other, it is impossible to estimate close-range spatial dependence. Second, the sill refers to the maximum observed variability in the data. In theory, the sill corresponds to the variance of the data as normally estimated in statistics. The difference between the sill and the nugget represents the amount of observed variation that can be explained by distance between observations. An ideal situation would consist of a small nugget and a large sill (i.e., there is much spatial dependence and a lot could be inferred about an unobserved location based on its distance from an observed site). Third, the range is the point at which the semivariogram plot stops increasing. The range represents the distance at which two observations are unrelated (i.e., independent). A model is often fit to the empirical variogram to aid in interpretation and in order to make use of the spatial dependence in other statistical techniques.

The dependence range, ρ_u, is the range which represents that if two observations or locations are far apart from this distance range, the correlation will be negligible. The moment estimator or sample semivariogram is given by:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \{[Y(s_i) - Y(s_j)]^2\},$$

where $N(\mathbf{h}) = \{(s_i, s_j) : s_i - s_j = \mathbf{h}\}$, the set of all pairs of locations separated by vector $\mathbf{h}$, of interest, sufficient pairs of points separated by vector $\mathbf{h}$ should be existed, and $Y(s_i)$ and $Y(s_j)$ are the responses at locations s_i and s_j, respectively Sherman (2011).

Before calculating the dependence range, it would be better to see the relationship between latitude and the longitude and the non accident mortality. From Figure 3.11, one can see that the weekly mean of non accident mortality for each city and its longitude and latitude. It shows that the highest mean mortality is of Busan and then Daegu, while the smallest mortality is for Seoul, the capital of South Korea. Figure 3.12 reveals that there is a decreasing pattern of mortality with longitude and increasing pattern with latitude, which means there is a location effect on non accident mortality. The range of dependence can be measured by drawing the semivariogram.

The semivariogram in Figure 3.13 shows that the dependence range is about distance 2. It is conventionally taken to be the distance when the semivariance first reaches 95% of the sill. Hence, in estimation of SSS-SIM and SSN-SIM, we use 2 as a dependence range. We can also see that there is a nugget effect value about 52.

3.6.3 SSS-SIM Estimation

Using the proposed algorithm I in Subsection 3.3.2, SSS-SIM parameters are estimated. Estimation and standard errors of the model parameters are shown in Table 3.5. Mean pressure coefficient parameter has been chosen to be equal 1 to fix the identifiability problem. A dependence range of 2 is used to calculate the covariance function. Table 3.5 shows that all the parameters are significance where zero value is not included in any 95% confidence intervals. Table 3.6 shows the estimates of spatial effects ($\hat{u}$), spatial variance ($\hat{\sigma}_u^2$), the bootstrapped-based standard errors (SE) of the estimates, and 95% confidence intervals.

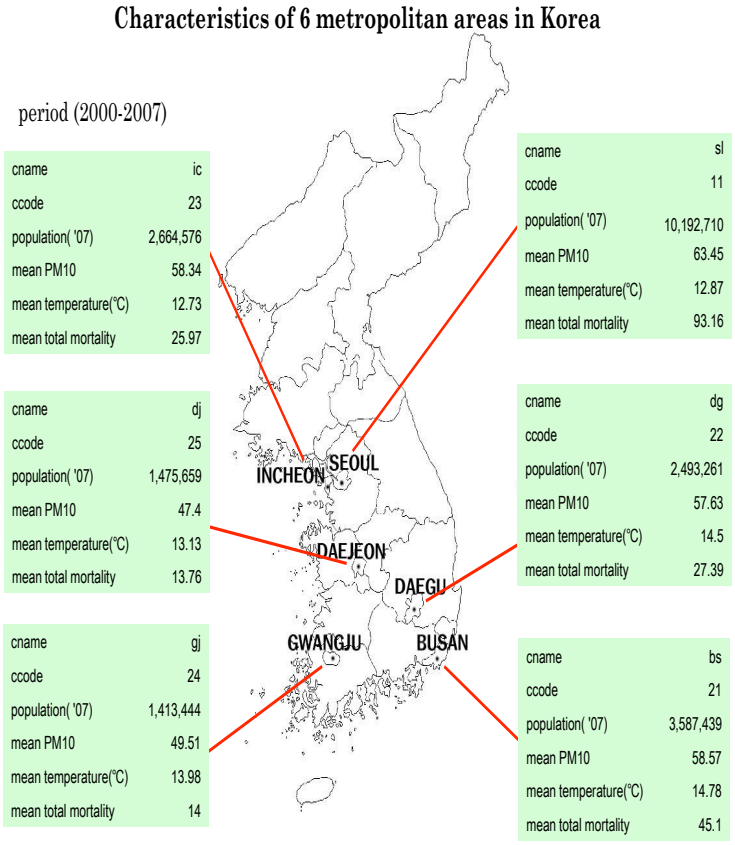


Figure 3.10: South Korea map shows the major 6 cities and their characteristics

The correlated spatial random effects estimates ($\hat{u}$) can be considered as the adjustments of the common non accident mortality function, $m(X\boldsymbol{\alpha})$. Table 3.6 reveals that the highest spatial random effect is for Busan (15.428) which located in the south east of South Korea as it can be seen in Figure 3.10. This means Busan has the highest mortality among the 6 major cities in South Korea. Daejeon has the lowest non accident mortality spatial effect (-5.357). The spatial effect estimates for the other cities in order are for Gwangju, Seoul, Incheon, and Daegu, respectively.

The variation among the spatial correlated random effects value is large ($\hat{\sigma}_u^2$=103.81). The

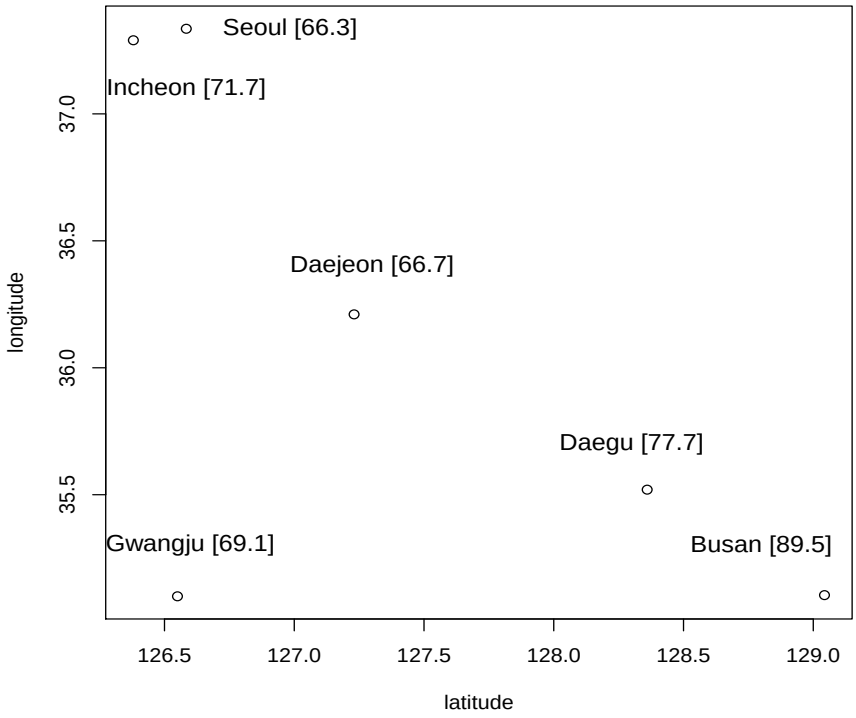


Figure 3.11: Cities locations (longitude and latitude) and their weekly mean of non accident mortality

95% confidence interval of σ_u^2 has no zero value, (102.195, 105.425), which means there is a location effect on non accident mortality. All the 95% confidence intervals of the spatial effects do not contain zeros which means the location effects are significant.

To calculate the standard error of the parameters estimates of SSS-SIM and the spatial effects, a bootstrapped method is used. The steps of the bootstrap procedure are described as follows:

1. For each location, 417 observations have been randomly selected with replacement.

2. Fit SSS-SIM using the bootstrapped data and estimate parameters ($\hat{\boldsymbol{\alpha}}$), spatial random effects estimates ($\hat{\mathbf{u}}$), and spatial variance ($\hat{\sigma}_u^2$).

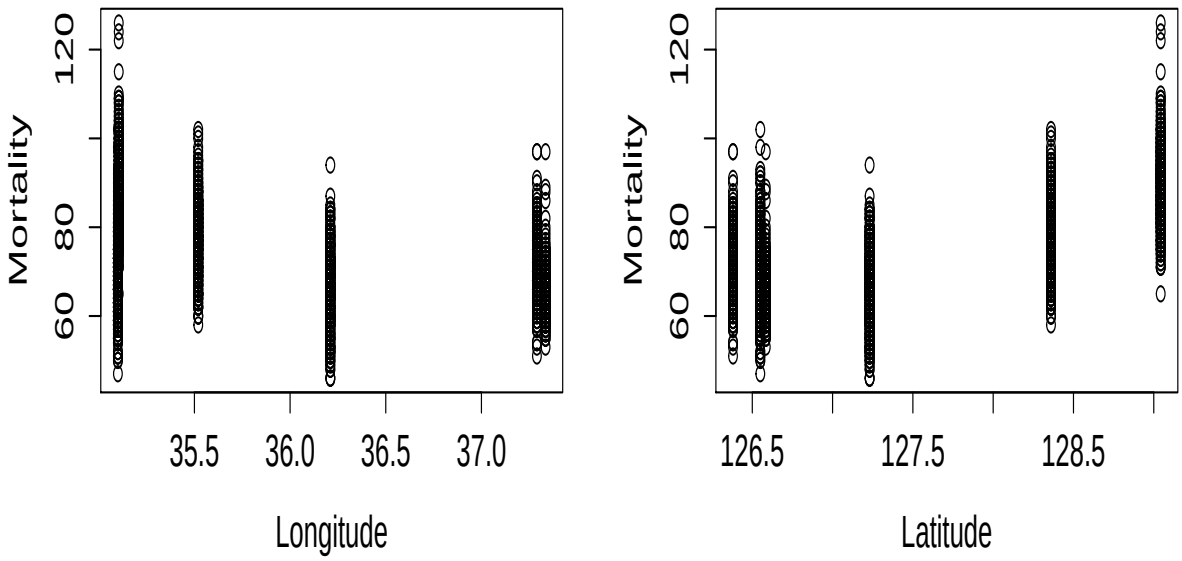


Figure 3.12: Scatterplot of longitude versus mortality (left) and latitude versus mortality (right)

3. Repeat step 1 and 2 for 250 times

4. Calculate the standard error of the 250 estimates of each parameter.

The boxplots of the 250 parameters estimates and the spatial random effects are displayed in Figure 3.14 and Figure 3.15. It seems that the distributions of parameter estimates ($\hat{\boldsymbol{\alpha}}$) are skewed and the distribution of estimate of the correlated spatial random effects ($\hat{\mathbf{u}}$) are close to symmetric.

The estimated common non-accident mortality function for all 6 South Korea cities, $\hat{m}(X\boldsymbol{\alpha})$, and its 95% pointwise confidence interval is displayed on Figure 3.16. One can see that it increases as the single index value increases until it reaches some point and then it decreases. Figure 3.17 shows the non accident mortality functions which are obtained by adding the spatial effects to the common single index function. It can be seen the highest non accident

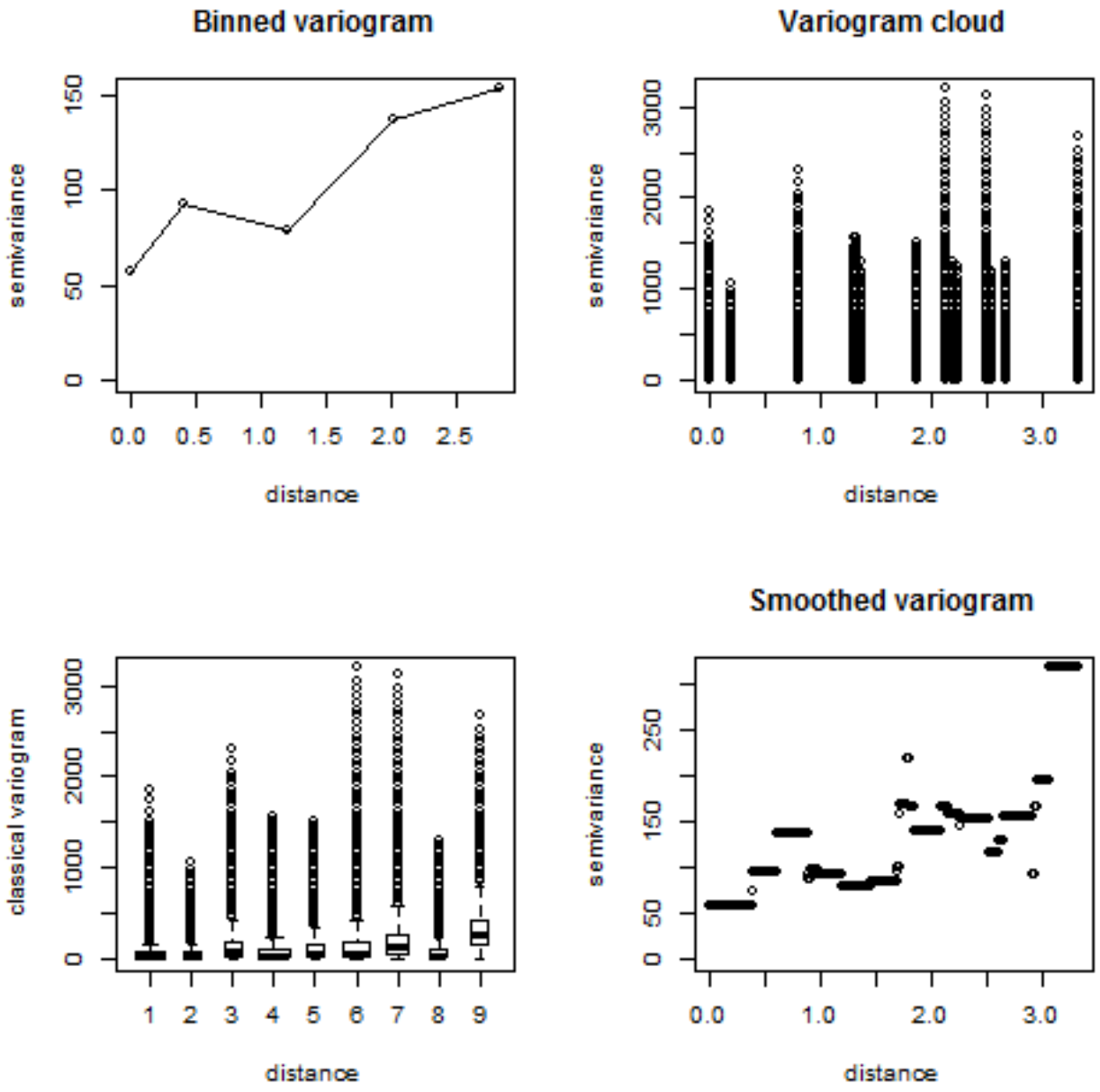


Figure 3.13: Different types of semivariogram: binned (top left), cloud (top right), cloud for binned (bottom left), and smoothed (bottom right) semivariogram

mortality function is for Busan and the lowest one is for Daejeon.

Table 3.5: Parameters estimation and their standard error (SE) of SSS-SIM=$m(X\boldsymbol{\alpha}) + Zu$ based on 250 bootstrapped data sets

	Parameter estimate ($\hat{\boldsymbol{\alpha}}$)	Bootstrapped-based SE	95% CI
Mean temperature	-1.1330	0.0332	(-1.1980, -1.0679)
Mean humidity	-0.1160	0.0068	(-0.1293, -0.1027)
Mean pressure	1	-	-
Month	-0.1010	0.0169	(-0.1341, -0.0678)

Table 3.6: Correlated spatial effects estimation, their standard error (SE), and 95% confidence interval of SSS-SIM=$m(X\boldsymbol{\alpha}) + Zu$ based on 250 bootstrapped data sets

	Spatial estimate ($\hat{u}$)	Bootstrap-based SE	95% CI
Seoul	-3.2261	0.0511	(-3.3261, -3.1258)
Busan	15.4280	0.0311	(14.3670, 14.4889)
Daegu	4.3510	0.0280	(4.2961, 4.4058)
Incheon	-3.0221	0.0377	(-3.0958, -2.9481)
Gwangju	-4.6050	0.0426	(-4.6886, -4.5213)
Daejeon	-5.3572	0.0425	(-5.4403, -5.2737)
σ_u^2	103.8100	0.8240	(102.1950, 105.4250)

3.6.4 SSN-SIM Estimation

The algorithm II is used to estimate SSN-SIM parameters and the spatial random effects. The bootstrap procedure described in Section 3.6.3 is also used to estimate the standard error of the estimates. There are two main differences between SSS-SIM and SSN-SIM: (1) SSS-SIM needs a restriction on $\boldsymbol{\alpha}$ but SSN-SIM does not. That is, in SSS-SIM estimation, we set the mean pressure coefficient parameter to be equal to 1 but in SSN-SIM we do not

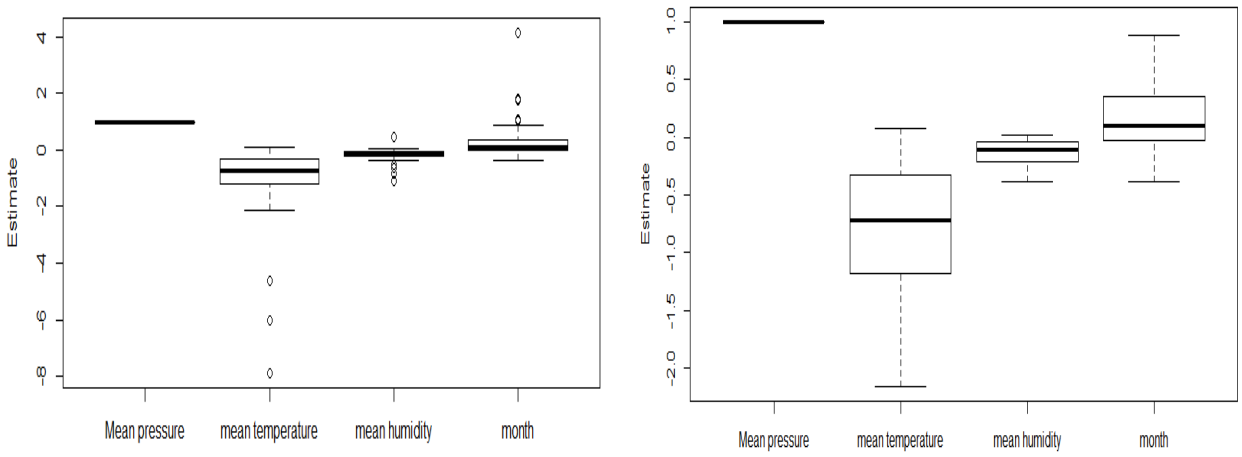


Figure 3.14: Boxplots of SSS-SIM parameters estimates from 250 bootstrap samples with outliers (left) and without outliers (right)

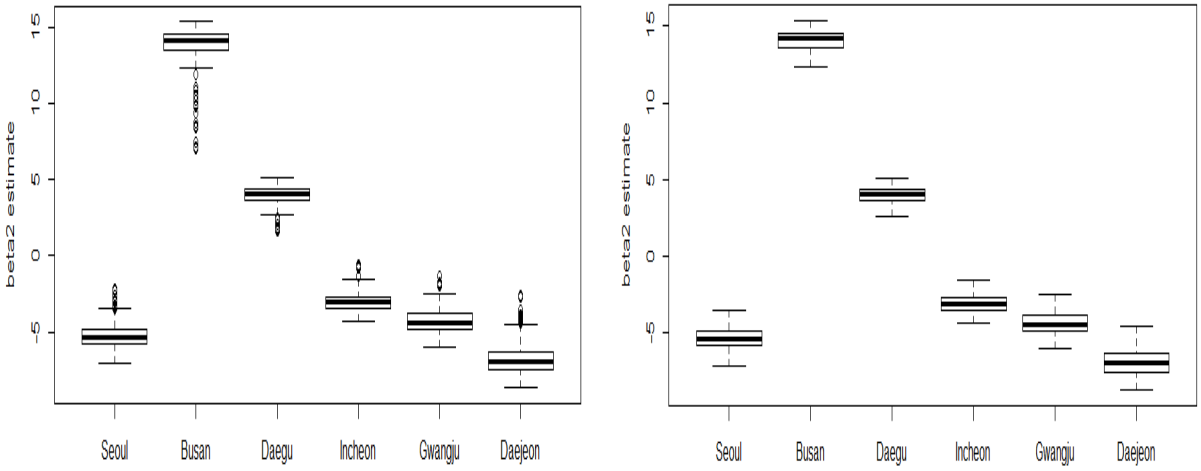


Figure 3.15: Boxplots of spatial random effects estimates from 250 bootstrap samples with outliers (left) and without outliers (right)

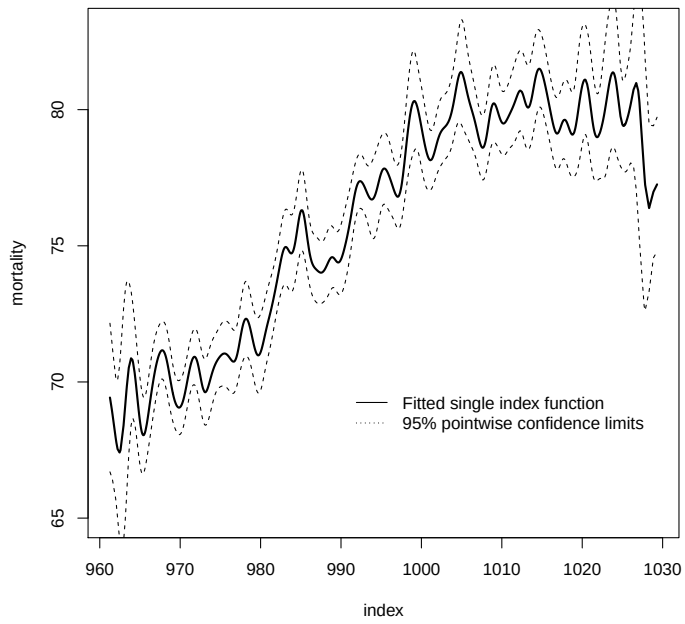


Figure 3.16: The estimated common non accident mortality single index function, $\hat{m}(X\boldsymbol{\alpha})$, and corresponding 95% pointwise CIs

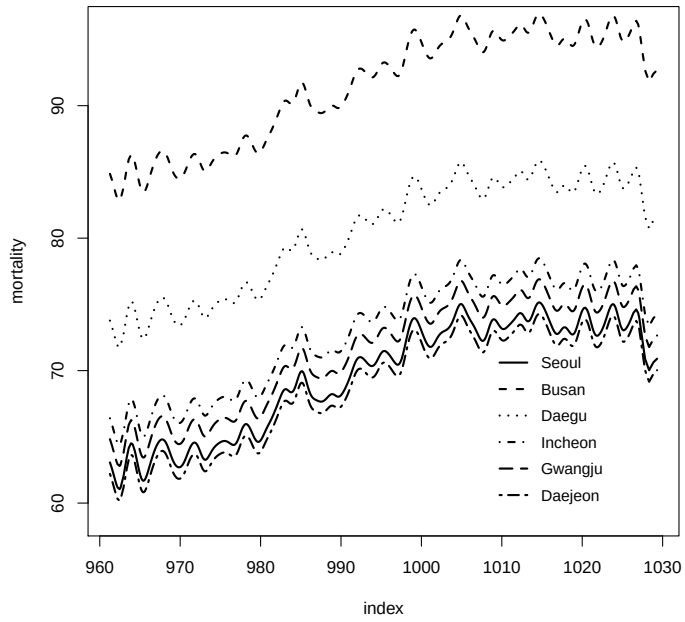


Figure 3.17: The estimated mortality single index function for each city

need this restriction and (2) SSN-SIM's correlated spatial effects are in the unknown single index mortality function but SSS-SIM's correlated spatial random effects are additive to the unknown single index function. As a result, the parameters estimates and their standard errors values are expected to be different. Results are shown in Table 3.7.

Because we do not have an identifiability problem in SSN-SIM, we could estimate all the parameters. One can notice that the parameters have smaller values compared to those of SSS-SIM. Bootstrapped-based standard error of the parameter estimates, $\hat{\boldsymbol{\alpha}}$, are smaller than that of SSS-SIM. It can be also seen that the variance of spatial random effects, $\hat{\sigma}_u^2$, of SSN-SIM is much smaller than that of SSS-SIM. They are 103.82 and 1.42, respectively. From 95% confidence interval in Table 3.7, all the parameters are significance because all the confidence intervals do not include zero. Bootstrap procedure has been used to estimate the standard error of the model parameters and spatial random effects estimates. These results are also summarized in Figure 3.18 which shows the boxplots of the 250 estimates of the model parameters with and without outliers.

The correlated spatial random effects estimates of SSN-SIM, bootstrapped-based standard error and 95% confidence intervals are reported in Table 3.8. The spatial effects estimates of SSN-SIM are smaller than of SSS-SIM but they have the same sign. All of 95% confidence interval do not cover zero values. Figure 3.19 shows the boxplots of the 250 estimates of spatial random effects estimates with and without outliers.

The estimated non accident mortality functions depending on mean temperature, mean pressure, mean humidity and month covariates and spatial random effects, $\hat{m}(X\hat{\boldsymbol{\alpha}} + Z\hat{u})$ with their 95% pointwise confidence intervals, are displayed in Figure 3.20. The non accident mortality function of Busan is the highest. However, the lowest non accident mortality function is in Seoul. Figure 3.21 shows the this findings. However using SSS-SIM, we found that Daejeon is estimated as the lowest mortality which is different from SSN-SIM. Figure

3.21 has been obtained after standardizing the single index for each city. It can be seen that the mortality functions are not identical which means that SSN-SIM is more flexible than SSS-SIM. The estimated mortality functions of SSN-SIM are similar to each other.

Table 3.7: Parameters estimates and their bootstrapped-based standard error (SE) of SSN-SIM=$m(X\boldsymbol{\alpha} + Zu)$

	Parameter estimate ($\hat{\boldsymbol{\alpha}}$)	Bootstrapped-based SE	95% CI
Mean temperature	-0.0244	7.09e-04	(-2.57e-02, -2.30e-02)
Mean humidity	-0.0047	6.49e-04	(-5.97e-03, 3.34e-03)
Mean pressure	-0.0152	2.34e-04	(-1.56e-02, 1.47e-02)
Month	0.0063	7.06e-04	(4.91e-03, 7.68e-03)

Table 3.8: Correlated spatial effects estimation, their standard error (SE), and 95% confidence interval of SSN-SIM=$m(X\boldsymbol{\alpha} + Zu)$ based on 250 bootstrapped data sets

	Spatial estimate ($\hat{u}$)	Bootstrap-based SE	95% CI
Seoul	-0.790	0.017	(-0.823, -0.756)
Busan	1.247	0.024	(1.199, 1.294)
Daegu	0.351	0.006	(0.339, 0.362)
Incheon	-0.275	0.007	(-0.288, -0.261)
Gwangju	-0.525	0.011	(-0.527, -0.522)
Daejeon	-0.775	0.016	(-0.806, -0.522)
σ_u^2	1.420	0.059	(1.3020, 1.5370)

3.6.5 Prediction and Model Selection

Our question of interest in our spatial data analysis is to predict a future event at a specific location. This section describes the performance of both models in terms of prediction

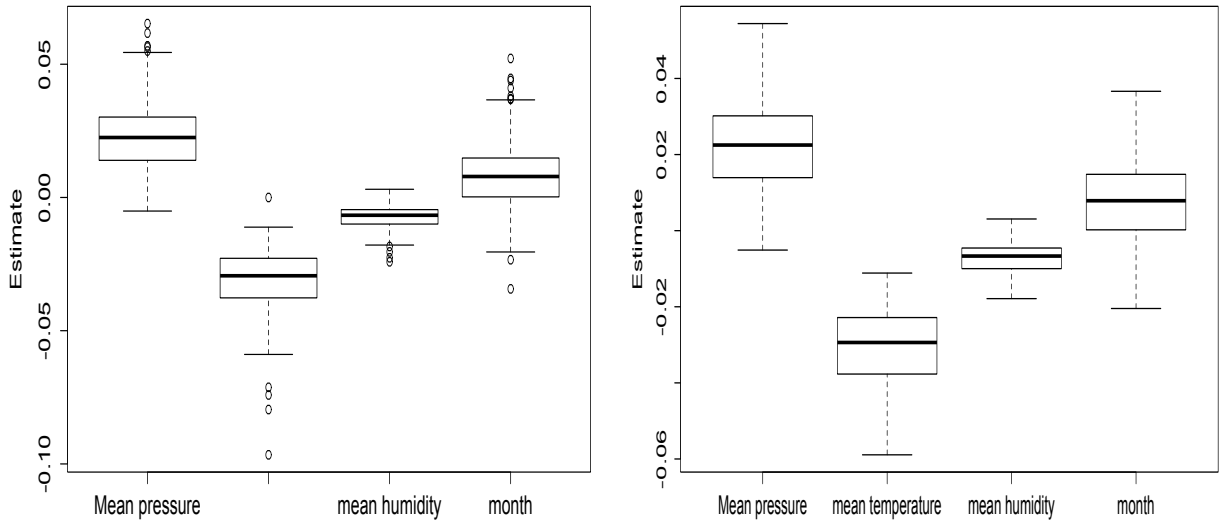


Figure 3.18: Boxplots of SSN-SIM parameters estimates obtained from 250 bootstrap samples with outliers (left) and without outliers (right).

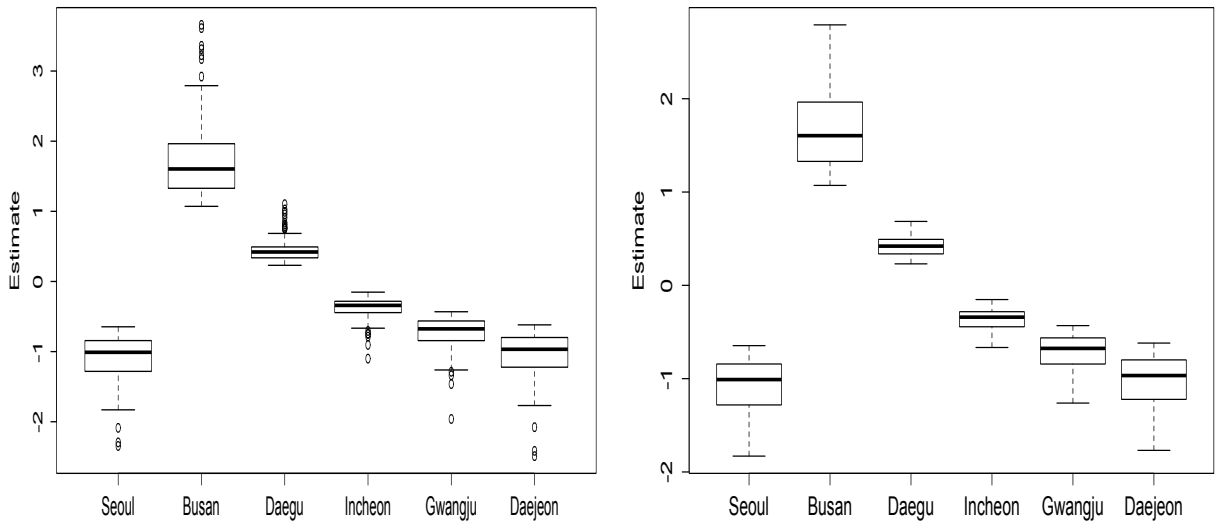


Figure 3.19: Boxplots of spatial random effects estimates of SSN-SIM obtained from 250 bootstrap samples with outliers (left) and without outliers (right)

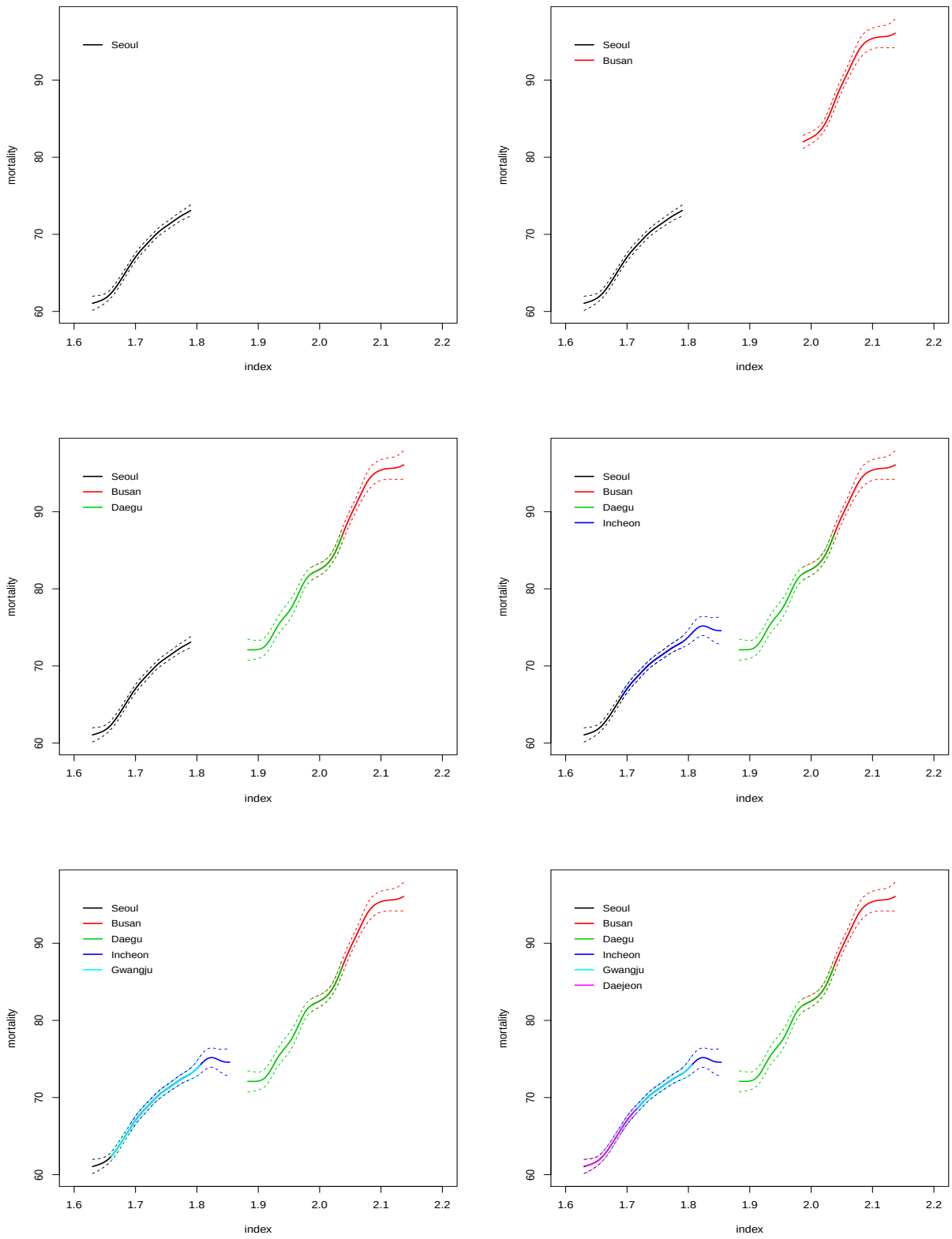


Figure 3.20: Estimated Non accident mortality functions of the six cities in South Korea and their 95% pointwise confidence intervals of SSN-SIM.

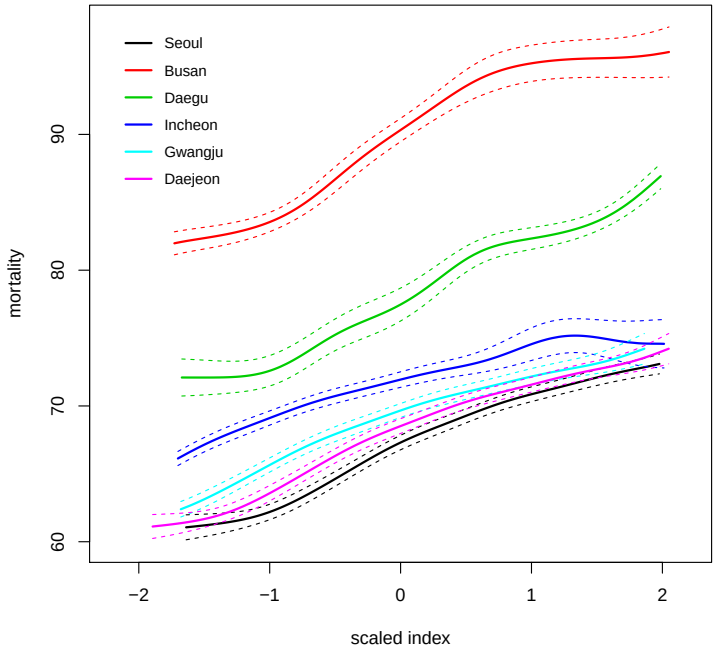


Figure 3.21: Estimated non accident mortality functions of the six cities in South Korea and their 95% pointwise confidence intervals after scaling the linear index variable of SSN-SIM

and estimation. The performance of the two models will also be compared to single index model (SIM) which does not include spatial random effects. Several criteria are used to compare the three model performance. We compare their prediction performances in terms of mean and median of predicted mean square error (PMSE) and predicted log likelihood value of test data given estimated parameters obtained from training data (PLogLE). We also compare their estimation performances in terms of MSE, R^2, and log likelihood value for given estimated parameters obtained from training data (LogLE). The following steps are conducted to calculate these criteria:

1. Set $n = 1$, $j = 1$

2. Randomly select n observations from each location (city) and consider it as a test data set and the rest will be considered as a training data set.

3. Fit the SSS-SIM using the training data and calculate R^2_j, MSE_j, and the log likelihood

value for given estimated parameters, which is defined as

$$\mathrm{LogLE_j} = \log \mathrm{f}[\mathbf{Y}|\hat{\boldsymbol{\mu}} = \hat{\mathrm{m}}(\mathrm{X}\hat{\boldsymbol{\alpha}}) + \mathrm{Z}\hat{\mathbf{u}}, \hat{\mathbf{u}}] + \log \mathrm{f_u}[\hat{\mathbf{u}}|\hat{\sigma}^2_{\mathrm{u}}\Sigma(\hat{\rho}_{\mathrm{u}})],$$

where $\mathbf{Y}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\mathbf{u}}$ are the vectors of the response of the training data set, estimated conditional mean, estimated spatial effects and $\hat{\sigma}^2_u$ and $\hat{\rho_u}$ are the estimated variance of spatial effects and dependence range.

4. Use the test data set to calculate $\mathrm{PMSE_j}$,

$$\mathrm{PMSE}_j = \sum_{s=1}^{6}\sum_{i=1}^{n} \frac{(y_i(s) - \hat{y_i^*}(s))^2}{6n},$$

where $y_i(s)$ and $\hat{y_i^*}(s)$ are the i^{th} actual and predicted response values at location (s), respectively; and also calculate the predicted log likelihood value of test data given estimated parameters, which is defined as

$$\mathrm{PLogLE_j} = \log \mathrm{f}[\mathbf{Y}^*|\hat{\boldsymbol{\mu}} = \hat{\mathrm{m}}(\mathrm{X}\hat{\boldsymbol{\alpha}}) + \mathrm{Z}\hat{\mathbf{u}}, \hat{\mathbf{u}}] + \log \mathrm{f_u}[\hat{\mathbf{u}}|\hat{\sigma}^2_{\mathrm{u}}\Sigma(\hat{\rho}_{\mathrm{u}})],$$

where $\mathbf{Y}^*$ is the vector of response in testing data, $\hat{\boldsymbol{\mu}}$, and $\hat{\mathbf{u}}$ are the estimated vectors of the conditional mean and spatial effects and $\hat{\sigma}^2_u$ and $\hat{\rho_u}$ are the estimated variance of spatial effects and dependence range.

5. Repeat 1-4 for 500 times (j=1,. . . , 500).

6. Calculate the average and median of the 500 estimates of $\mathrm{PMSE_j}$ (APMSE and MPMSE, respectively). Also calculate R^2, MSE, PLogLE, and LogLE, where

$$R^2 \;=\; \sum_{j=1}^{500} \frac{R_j^2}{500}$$

$$
\begin{aligned}
MSE &= \sum_{j=1}^{500} \frac{MSE_j}{500} \\
\text{PLogLE} &= \sum_{j=1}^{500} \frac{PLogLE_j}{500} \\
\text{LogLE} &= \sum_{j=1}^{500} \frac{LogLE_j}{500}.
\end{aligned}
$$

7. Repeat 2-5 for different n (=5, 10, 20, 50, 100) and $j = 1, \ldots, 500$.

8. Repeat 1-7 for SSN-SIM and SIM.

In Table 3.9, the results are shown for SSS-SIM, SSN-SIM, and SIM. Estimating criteria, LogLE, MSE, and R^2 for all data have been also reported. For all data (all the data are training data), it can be seen that SSS-SIM and SSN-SIM are much better than SIM in terms of R^2, LogLE, and MSE. This means including spatial random effects in SSS-SIM and SSN-SIM improved the model fitting of our data. In other words, there is an effect of location on the non accident mortality function of each city. For the two models include spatial effects, SSS-SIM and SSN-SIM, SSN-SIM has higher value for LogLE and R^2 and smaller value for MSE. As a result, SSN-SIM fits the data better than SSS-SIM and SIM.

For prediction, in case $n = 1$ where one observation has been selected randomly from each location, we use total 6 observations from all the locations as a testing data set. We can see that SSS-SIM and SSN-SIM perform better than SIM in terms of all the criteria: estimation and prediction. And again, SSN-SIM works better than SSS-SIM in terms of prediction. The distributions of PMSE_j ($j = 1, \ldots, 500$) of all the three models are skewed to right. It can be seen from the mean and median of PMSE_j of all the 500 estimates, (APMSE > MPMSE).

In case $n = 5$ in which we have 30 observations as a testing data set and 2472 observations as training data. One can see similar results as before; SSN-SIM is better than the other

two models (SSS-SIM and SIM). SIM is the worst model in terms of all the criteria.

The larger number of observations for predicting, the better prediction performance. In case $n = 100$ in which we have 600 observations as evaluation data and 1902 observations as training data, SSN-SIM still the best model fitting the mortality data and SIM is the worst. For the three models, PMSE_j distribution is symmetric where APMSE is equal to MPMSE. In conclusion, although both models SSS-SIM and SSN-SIM are semiparametric models incorporating spatial correlated random effects into, SSN-SIM performs much better than SSS-SIM in terms of prediction and estimation.

Figure 3.22 shows the boxplots of the 500 estimates of PMSE_j for each model at each test data set size ($n = 1, 5, 10, 20, 50, 100$). It shows that the variation of SIM prediction estimates is significantly higher than the other two models, SSS-SIM and SSN-SIM, which have spatial random effects in different forms. The prediction of SSN-SIM and SSS-SIM have almost the same variation, however the median of SSN-SIM is less than SSS-SIM at all the test sample sizes as shown in Table 3.9. Figure 3.22 also reveals that the prediction variation decreases as the sample size of test data set increases.

For the PMSE_j distributions, when n is small, they are fairly skewed to right. However as n becomes larger, the distributions approach to symmetric distributions with a few outliers.

Table 3.9: Several criteria (APMSE, MPMSE, PLogLE, LogLE, MSE, and R^2) to compare between SIM, SSS-SIM, SSN-SIM, SSTS-SIM, and STSN-SIM in estimation and prediction based on 500 testing and training data at different sizes of testing data set ($n = 1, 5, 10, 20, 50, 100$)

	Model	MPMSE	APMSE	PLogLE	PLogLE	MSE	R^2
All Data	SIM	—	—	—	-9371.27	102.81	17.43
	SSS-SIM	—	—	—	-8398.76	41.93	66.84
	SSN-SIM	—	—	—	-8352.20	40.13	67.46
n=1	SIM	95.56	105.40	-22.55	-10005.66	101.83	17.43
	SSS-SIM	39.74	46.17	-44.64	-8373.41	41.53	67.04
	SSN-SIM	33.64	40.18	-22.80	-8332.00	40.09	67.50
n=5	SIM	101.30	104.90	-225.80	-9929.91	101.78	17.47
	SSS-SIM	41.78	44.43	-126.71	-8300.38	41.86	66.41
	SSN-SIM	38.34	40.77	-103.11	-8250.31	40.04	67.52
n=10	SIM	102.90	104.90	-226.73	-9822.00	101.32	17.91
	SSS-SIM	42.62	43.63	-224.82	-8200.48	42.09	66.70
	SSN-SIM	41.12	42.27	-204.90	-8154.72	40.17	67.53
n=20	SIM	104.20	104.90	-451.50	-9577.34	101.67	17.59
	SSS-SIM	42.99	42.99	-426.37	-7992.01	41.47	66.81
	SSN-SIM	40.95	41.46	-404.20	-7949.00	39.97	67.51
n=50	SIM	104.80	105.30	-1129.00	-8891.04	101.53	17.72
	SSS-SIM	42.83	42.96	-1033.70	-7385.90	40.86	66.96
	SSN-SIM	40.92	41.17	-1006.27	-7348.12	39.93	67.62
n=100	SIM	105.60	105.70	-2260.75	-7582.92	101.32	17.92
	SSS-SIM	44.15	44.15	-2039.70	-6390.27	42.02	66.75
	SSN-SIM	41.07	41.42	-2010.00	-6349.00	40.05	67.48

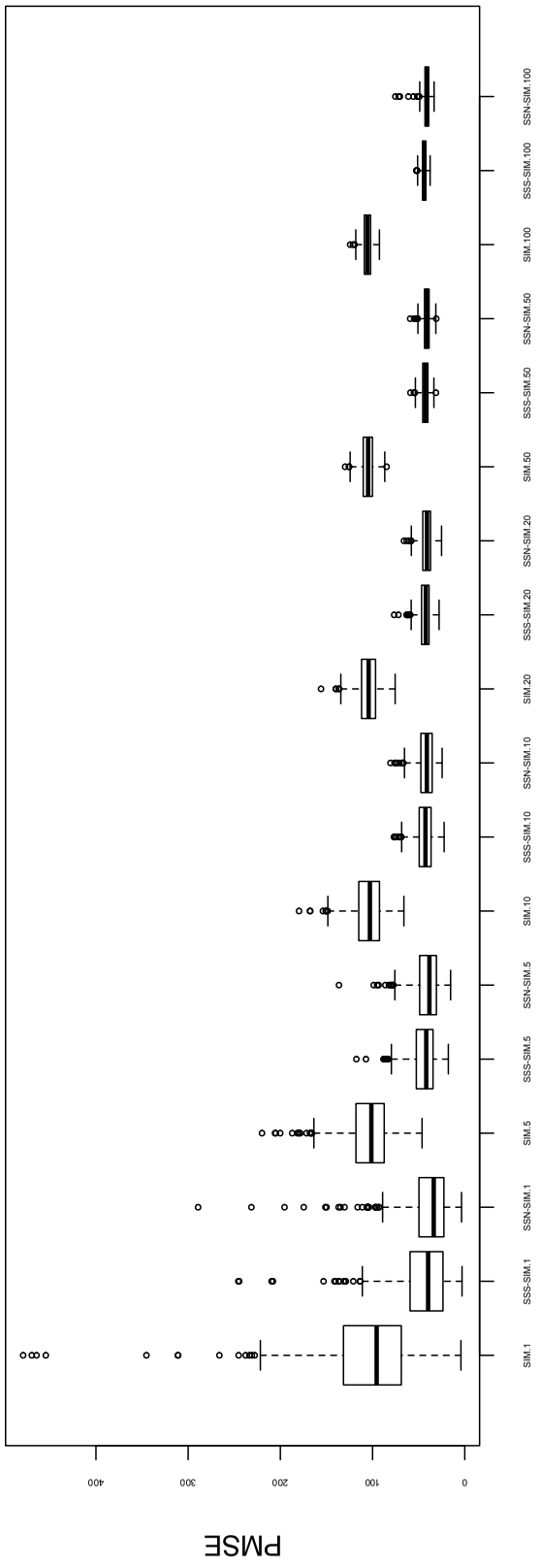


Figure 3.22: Boxplots of Prediction Mean Square Error (PMSE_j) ($j = 1, \ldots, 500$) of the three models; SIM= single index model, SSS-SIM=semiparametric spatial-separable single index model and SSN-SIM= semiparametric spatial-nonseparable single index model at different testing data set sizes ($n = 1, 5, 10, 20, 50, 100$)

3.7 Summary

We have proposed two models to incorporate the spatially correlated random effects into the single index model. One is based on spatial random effects that are separated additively from the unknown function and the relationship between the mean response and the random effects is linear. In the other one, the spatially correlated random effects are included in the unknown function so that the spatial random effects can not be separated from the unknown function. The relationship between the spatial random effects and the mean response is unknown. We have proposed a nonseperable model because the nonparametric function $m(\cdot)$ is different from that of each spatial location. To the best of our knowledge, there is no such nonparametric single index model which can not be separated with spatial random effects. We showed this nonseperable model provides not only accurate parameter estimation but also better prediction accuracy.

For each model we have proposed an algorithm to simultaneously estimate the unknown function, the single index coefficients parameters, the variance of the spatial effects, and the spatial random effects based on MCEM algorithm. Simulation studies were performed to understand the performance of these two models. In terms of estimating the spatially correlated random effects, the SSN-SIM performs better than SSS-SIM. In SSN-SIM we do not need to have a restriction on the single index coefficients parameters which enable us to estimate all the parameters and the unknown function without additional constraints on identifiability. Additional simulation study was done to make sure the two models have the same performance when the unknown function is the identity function. It is found that they give almost the same estimates. The two models have been applied to the South Korea data, SSS-SIM and SSN-SIM. It is found that Busan city has highest non accident mortality when we fit SSS-SIM or SSN-SIM, however the city which has lowest mortality is different in both models: Seoul city has the lowest mortality using SSN-SIM and Daejeon using SSS-SIM.

Also, we found that the shape of mortality functions of the 6 cities are the same in SSS-SIM models, but they are different in SSN-SIM. To see which model fits the data better, SSS-SIM, SSN-SIM, and SIM has been compared. We found that SSN-SIM is the most suitable model for our mortality data. This means including spatial effect in single index model improved the model and including it in nonadditive format is much better for both the prediction and estimation performance.

Chapter 4

Semiparametric Spatio-Temporal Single Index Model

4.1 Background

During over the last decades, the need for analyzing spatio-temporal data has been greatly increased due to the increasing amount of spatio-temporal data in various areas (Kanevski and Maignan, 2004; Genton et al., 2006; Li et al., 2007; Landagan and Barrios, 2007; Nelson et al., 2009; Hayn et al., 2009; Sherman, 2011; Lekdee and Ingsrisawang, 2013; Arcuti et al., 2013) involving data across time and space. The early research works on spatio-temporal analysis can be also found in a series of papers by Bilonick and Nichols (1983) and Bilonick (1983, 1985, 1988). Numerous models have been provided for modeling spatio-temporal data using generalized linear mixed model (GLMM) (Lekdee and Ingsrisawang, 2013), generalized linear additive model (GAM) (Hayn et al., 2009; Arcuti et al., 2013), and spatio-temporal autoregressive model (Landagan and Barrios, 2007). In generalized linear additive model, time and spatial variables can be nonparametrically and additively modeled. Alternatively,

the mixed model frameworks can be used by treating them as correlated random effects. One of the main differences between the spatio-temporal random models and mixed effects models is the construction of covariance matrix of time and spatial effects. Covariance matrix can be constructed using some parametric form which mainly depends on the distance between the observations. Some literatures (Cressie and Hawkins, 1980; Cressie and Huang, 1999; Cressie, 1993; Chiles and Delfiner, 1999; Isaaks and Srivastava, 1999; Stein, 1999) on spatio-temporal analyses have extensively studied covariance functions, while others (Lekdee and Ingsrisawang, 2013; Arcuti et al., 2013; Hayn et al., 2009; Landagan and Barrios, 2007) have focused on estimating the mean function. However, the most approaches are developed using parametric model with strong model assumption which can not be applicable to real application. Perhaps, semiparametric model is more flexible and appropriate to the real situation where the functional form is possibly neither linear nor nonlinear. Since semiparametric modeling for for spatially and temporally correlated data is quit limited, in this chapter we propose a semiparametric model based on single index model to allow several covariates X.

Semiparametric single index model (SIM) is also popular in many scientific fields such as biostatistics, medicine, economics and financial econometrics and has been extensively studied in the statistical literature; see Li (1991); Naik and Tsai (2000); Stute and Zhu (2005); Xia and Li (1999); Xia et al. (2002); Zhu and Ng (1995); Zhu and Xue (2006); Lin and Kulasekera (2007); Xia (2006); Zhu and Zhu (2009a,b); Wang et al. (2010); Hridtache et al. (2001); Chang et al. (2010). SIM outperforms the parametric models, such as linear models and generalized linear model in terms of flexibility because it assumes the conditional mean function is unknown which makes this model more flexible for the conditional mean function. SIM also does not assume a specific distribution for error which avoids the misleading results of using incorrect distribution for errors (Horowitz and Hardle, 1996).

Semiparametric single index model has some advantages over the nonparametric models: (1) the precision of nonparametric estimation decreases when X dimension increases, curse of dimensionality, and to overcome this problem a large sample is needed. In SIM we have only one dimension which is the index $X\boldsymbol{\alpha}$ that enables SIM to avoid the curse of dimensionality and $\boldsymbol{\alpha}$ can be estimated with rate of convergence $n^{-1/2}$ (Li et al., 2007), (2) using nonparametric estimation does not allow us to to make prediction of the conditional mean at some points are not in X range which is important in taking decision and forecasting, on the other hand SIM permits predictions at points not in X range but in range of the index $\boldsymbol{\alpha}$. Pang and Xue (2012) introduced single index model with random effects for longitudinal data but they assumed independent random effects and used generalized estimating equations (GEE) to estimate single index coefficient parameters.

Therefore, in this chapter, we propose two semiparametric single index models simultaneously incorporating spatial effects and time effects. One model is developed by separating both spatial and time effects from the nonparametric function of covariates, we call it "Semiparametric Spatio-Temporal Separable Single Index model (SSTS-SIM)". However the other model we propose is the semiparametric single index model whose spatial effects can not be separated from the nonparametric function of covariates but time effects can be separated from it, we call it "Semiparametric Spatio-Temporal Nonseparable Single Index (SSTN-SIM)". We propose this model because nonparametric function is different from that of each spatial location and time point. To the best of our knowledge, there is no such nonparametric single index model which can not be separated with spatial random effects. In later of this chapter, we show this nonseparable model provides not only accurate parameter estimation but also better prediction accuracy. Our two models, separable and nonseparable models, were estimated via Markov Chain Expectation Maximization (MCEM) algorithm. After estimating the nonparametric function, we also provide the prediction accuracy to predict

"unobserved" mortality at location s.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce the proposed two models. In Section 4.3, we describe how to estimate these two models using the MCEM algorithm. We explain the proposed two algorithms for estimating the two models. We apply our models to real application 4.4. Evaluating the performance of the proposed models in prediction and model selection is presented in Section 4.5. Concluding remarks and discussion are provided in Section 4.6.

4.2 Semiparametric Spatio-Temporal Single Index Random Effects Models

In this section, two models are proposed for spatio-temporal data. One is to additionally separate unknown single index function from spatial and time random effects, while the other can only separate time effect but spatial effect can not be separated from unknown function. We refer the first model to "semiparametric spatio-temporal separable Single Index model" and the second model to "semiparametric spatio-temporal nonseparable Single Index model". The covariance functions of spatial and time effects are introduced.

4.2.1 Semiparametric Spatio-Temporal Separable Single Index Model (SSTS-SIM)

For a spatial location s and time point t, let $Y(s,t)$ denotes the response variable, $x_1(s,t)$, $x_2(s,t)$, ..., $x_p(s,t)$ are the p observable explanatory variables at location s and time t, $\{u(s), s \in R^2\}$ be an unobservable spatial random process such that $u(s)$ represents the

random effect at site s of unknown or unobservable causes unaccounted for by the explanatory variables, and $\nu(t)$ is the time effect at time point t and location s.

Semiparametric Spatio-Temporal Separable Single Index model is defined as follows:

$$\begin{aligned} Y(s,t)|\mu(s,t) &\sim \text{Pois}[\mu(s,t)|u(s),s(t)], \\ \mu(s,t)|u(s),s(t) &= m[X(s,t)\boldsymbol{\alpha}] + u(s) + \nu(t), \end{aligned}$$

where

1. X(s,t): explanatory variable matrix at location s and time point t, m is an unknown function, and $\boldsymbol{\alpha}$ is single index coefficients parameters;

2. $\{u(s), s \in R^2\}$: a Gaussian stationary process with $E[u(s)] = 0$ for all s and cov[u(s + d), u(s)] = C(d) for all $s, d \in R^2$;

3. $\nu(t)$: time effect which is a random process follows some parametric form;

4. Conditionally on $\{u(s), s \in R^2$ and $\nu(t)\}$, $\{Y(s,t), s \in R^2\}$ is independent Poisson process.

In this model, $E[Y(s,t)|u(s),\nu(t)]$, $u(s)$ and $\nu(t)$ are assumed to be additive. We simultaneously estimate $\boldsymbol{\alpha}$, $\mathbf{u}$, $\nu(t)$ and the unknown function $m(\cdot)$. In single index model, some restrictions on $\boldsymbol{\alpha}$ is needed in order for it to be identifiable. Our approach is to set one component of $\boldsymbol{\alpha}$ to be equal to one (Ichimura, 1993; Sherman, 1994).

For modeling spatial variation, we use Gaussian covariance function. The covariance function describes the spatial association between the random effects at any two locations in space,

say $u(s)$ and $u(s')$ takes the form:

$$\begin{aligned} \text{cov}(u(s), u(s')) &= \sigma_u^2 \Sigma_u(\rho_u) \\ \mathbf{u} &\sim \text{MN}[\mathbf{0}, \sigma_{\text{u}}^2 \Sigma_{\text{u}}(\rho_{\text{u}})], \end{aligned}$$

where σ_u^2 is the variance of the random effects $\mathbf{u}$, Σ_u is a parametric covariance function that depends only on the distance between any two locations s and s', and ρ_u is the dependence range which can be estimated by semivariogram. The range represents the distance at which the smeivariogram plot stops increasing and after that range the two points are considered unrelated. The spatial covariance matrix Σ_u is assumed to be in a known parametric form to guarantee the covariance matrix is positive definite, some well-known parametric covariance functions are displayed in Table 1.1.

For modeling time variation, we first consider a random walk model. The random walk (RW), autoregressive (AR), moving average (MA) or autoregressive integrated moving average (ARIMA) are commonly used to model the temporal random effects, $\nu(t)$, in time series. Clayton (1996) applied a first order random walk, denoted by RW(1), to model temporal trend. RW(1) implies that the difference between any two consecutive time points follows a normal distribution and the relationship between any two consecutive time points, say $\nu(t)$, and $\nu(t-1)$, has the form:

$$\nu(t) = \nu(t-1) + \epsilon(t),$$

where $\epsilon(t)$ is the random noise term that accounts for difference from one observation to the next observation within each location and follows a normal distribution. Following Knorr-Held (2000), assuming we have n locations, the vector form of the temporal random effects

is $\boldsymbol{\nu} = (\nu(1), \nu(2), ..., \nu(n))^T$ and the joint distribution of $\boldsymbol{\nu}$ has the following form

$$f(\boldsymbol{\nu}|\sigma_\nu^2) \propto \exp(-\frac{\sigma_\nu^2}{2}\boldsymbol{\nu}^{\mathrm{T}}\Sigma_\nu^{-1}\boldsymbol{\nu}),$$

where Σ_ν^{-1} is the temporal precision matrix. As an example if we have 5 time points, the corresponding Σ_ν^{-1} matrix will be as follows:

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

However, this precision matrix is singular, so we can not get the covariance matrix from it. In Bayesian frame work, it can be used as a prior but in frequentest approach, it can not be used because it is not invertible. Hence, we propose to overcome this problem using a diagonal of 2's which means $\nu(1)$ and $\nu(5)$ are random numbers and keep the other time points follow first order random walk.

Another covariance function for temporal effects in our study is a Gaussian process with $\rho_\nu = 2$ which means that each time effect at time point t, $\nu(t)$, depends only on the previous and next time random effect, $\nu(t-1)$ and $\nu(t+1)$. The association between any two time points at location s, say $\nu(t)$ and $\nu(t')$ takes the form:

$$\begin{aligned} \mathrm{Cov}(\nu(t), \nu(t')) &= \sigma_\nu^2 \Sigma_\nu(\rho_\nu = 2); \\ \boldsymbol{\nu} &\sim \mathrm{MN}[\mathbf{0}, \sigma_\nu^2 \Sigma_\nu(\rho_\nu = 2)]. \end{aligned}$$

4.2.2 Semiparametric Spatio-Temporal Nonseparable Single Index Model (SSTN-SIM)

Unlike SSTS-SIM which has the linear relationship between $E[Y(s,t)|u(s),\nu(t)]$ and $u(s)$ and $\nu(t)$, SSTN-SIM does not have the linear relationship between $E[Y(s,t)|u(s),\nu(t)]$ and $u(s)$ but does have a linear relationship between $E[Y(s,t)|u(s),\nu(t)]$ and $\nu(t)$. One of the advantages of this model is that it does not need restriction on the single index coefficients parameters for the identifiability problem, such as $\alpha_1 = 1$ or $\|\boldsymbol{\alpha}\| = 1$. We use $\mathbf{u}$ as the variable which has its coefficient is equal to one. This enables us to estimate all the single index coefficients without having identifiability problem and estimate the unknown function without more assumptions.

With the same definition of $Y(s,t)$, $X(s,t)$, $\nu(t)$ and $u(s)$, this model can be written as

$$\begin{aligned} Y(s,t)|\mu(s,t) &\sim \text{Pois}[\mu(s,t)|u(s),\nu(t)], \\ \mu(s,t)|u(s),s(t) &= m[X(s,t)\boldsymbol{\alpha} + u(s)] + \nu(t), \end{aligned}$$

where

1. X(s,t): explanatory variable matrix at location s and time point t, m is an unknown function, and $\boldsymbol{\alpha}$ is single index coefficients parameters;

2. $\{u(s), s \in R^2\}$: a Gaussian stationary process with $E[u(s)] = 0$ for all s and cov[u(s + d), u(s)] = C(d) for all $s, d \in R^2$;

3. ν(t): time effect at point t which follows some parametric form;

4. Conditionally on $\{u(s), s \in R^2$ and $\nu(t)\}$, $\{Y(s,t), s \in R^2$ and $t \in R\ \}$ is independent

Poisson process.

We use similar covariance function for spatial variation described in SSTS-SIM. The spatial association between the random effects at any two locations in space, say $u(s)$ and $u(s')$ takes the form:

$$\begin{aligned} \text{cov}(u(s), u(s')) &= \sigma_u^2 \Sigma_u(\rho_u) \\ \mathbf{u} &\sim \text{MN}[\mathbf{0}, \sigma_{\text{u}}^2 \Sigma_{\text{u}}(\rho_{\text{u}})]. \end{aligned}$$

For time variation, we consider both random walk and Gaussian process described in SSTS-SIM.

The vector form of the temporal random effects is $\boldsymbol{\nu} = (\nu(1), \nu(2), ..., \nu(n))^T$ and the joint distribution of $\boldsymbol{\nu}$ has the following form

$$f(\boldsymbol{\nu}|\sigma_\nu^2) \propto \exp(-\frac{\sigma_\nu^2}{2} \boldsymbol{\nu}^{\text{T}} \Sigma_\nu^{-1} \boldsymbol{\nu}),$$

where Σ_ν^{-1} is the temporal precision matrix. As an example if we have 5 time points, the corresponding Σ_ν^{-1} matrix will be as follows:

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

However, this precision matrix is singular, so we can not get the covariance matrix from it.

In Bayesian frame work, it can be used as a prior but in frequentest approach, it can not be used because it is not invertible. Hence, we propose to overcome this problem using a diagonal of 2's which means $\nu(1)$ and $\nu(5)$ are random numbers and keep the other time points follow first order random walk. Another covariance function for temporal effects is Gaussian process (Liu et al., 2008) with $\rho_\nu = 2$ which means that each time effect at time point t, $\nu(t)$, depends only on the previous and next time random effect, $\nu(t-1)$ and $\nu(t+1)$. The association between any two time points at location s, say $\nu(t)$ and $\nu(t')$ takes the form:

$$\begin{aligned} \text{Cov}(\nu(t), \nu(t')) &= \sigma_\nu^2 \Sigma_\nu(\rho_\nu = 2); \\ \boldsymbol{\nu} &\sim \text{MN}[\mathbf{0}, \sigma_\nu^2 \Sigma_\nu(\rho_\nu = 2)]. \end{aligned}$$

In section 4.4, the performance of the two covariance functions of time effects will be shown in terms of the stability of parameters estimation. Other covariance functions for time and spatial effects can be easily extended and used in our proposed models.

4.3 SSTS-SIM and SSTN-SIM Estimation

In this section, we first briefly explain MCEM algorithm. We provide how to choose our candidate distributions for the Metropolish-Hastings (M-H) step and then propose two MCEM algorithms; one is for SSTS-SIM, and the other is for SSIN-SIM.

4.3.1 MCEM Algorithm

This algorithm is commonly used in the GLMM estimation (McCulloch, 1994; Booth and Hobert, 1999; Caffo et al., 2005; Tan et al., 2007; An and Bentler, 2012). EM algorithm

consists of two steps; expectation (E-step) and maximization (M-step). Iterating between the two steps until convergence satisfies. When the E-step involves analytically intractable integrals, one approach is to approximate E-step using some Monte Carlo method. Incorporating the Monte Carlo step into EM algorithm gives MCEM algorithm. In our proposed models, the spatial random effects are not independent and also time effects . Both distributions of spatial and time effects have mean $\mathbf{0}$ and variance-covariance matrices $\sigma_u^2\Sigma_u(\rho_u)$ and $\sigma_\nu^2\Sigma_\nu(\rho_\nu = 2)$, respectively. Because there is no closed form is available for the expectation, we need to incorporate MCMC to generate random samples from full conditional distributions of $\mathbf{u}$ and $\boldsymbol{\nu}$. We use the Metropolis-Hastings algorithm. Choosing the candidate or proposal function is very important in the M-H algorithm. The complete-data log-likelihood for our both models, in general, is given by

$$\begin{aligned}\log f[\mathbf{Y}, \mathbf{u}, \boldsymbol{\nu}|\boldsymbol{\mu}, \sigma_u^2\Sigma_u(\rho_u), \sigma_\nu^2\Sigma_\nu(\rho_\nu = 2)] &= \log f_{Y|u,\nu}[\mathbf{Y}|\boldsymbol{\mu}, \mathbf{u}, \boldsymbol{\nu}] + \log f_u[\mathbf{u}|\sigma_u^2\Sigma(\rho_u)] \\ &+ \log f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu = 2)].\end{aligned}$$

If we used as candidate distributions the multivariate normal distributions $\mathrm{MN}(\mathbf{0}, \sigma_0^2\Sigma_\mathrm{u})$=$f(.|\sigma_0^2\Sigma_u)$ for spatial effects and $\mathrm{MN}(\mathbf{0}, \sigma_0^2\Sigma_\nu)$=$f(.|\sigma_0^2\Sigma_\nu)$ for time effects, the probability of accepting a new value $\mathbf{u}^*$ with the current value being $\mathbf{u}$ given a value of $\boldsymbol{\nu}$ is

$$\min\left\{\frac{f[\mathbf{Y}|\mathbf{u}^*, \boldsymbol{\nu}, \boldsymbol{\mu}]f_u[\mathbf{u}^*|\sigma_u^2\Sigma_u(\rho_u)]f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]}{f[\mathbf{Y}|\mathbf{u}, \boldsymbol{\nu},\boldsymbol{\mu}]f_u[\mathbf{u}|\sigma_u^2\Sigma_u(\rho_u)]f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]}, 1\right\}.$$

If we use the single-component Metropolis-Hastings algorithm, i.e., at each iteration, we only update a single component, say the sth component $u(s)$. We will generate the candidate values from the conditional normal distribution of $N(\bar{\theta}, \sigma_0^2\bar{\Sigma_u})$, where σ_0^2 is a proposal variance of the spatial random effects. This conditional distribution and its parameters can be derived as follows:

Let $\mathbf{v}=(v_1, v_2, \ldots, v_n) = \left[v_1 \ \mathbf{v_2} \right]^T$ has multivariate normal distribution with mean $\boldsymbol{\theta} = \left[\theta_1 \ \boldsymbol{\theta}_2 \right]^T$ and variance-covariance matrix $\sigma_0^2\Sigma$, where

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then the distribution of v_1 conditioned on $\mathbf{v_2} = \mathbf{a}$ is multivariate normal $(v_1|\mathbf{v_2} = \mathbf{a}) \sim \mathbf{N}(\bar{\theta}, \bar{\mathbf{\Sigma}})$, where $\bar{\theta} = \theta_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{a} - \boldsymbol{\theta_2})$ and covariance matrix $\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

In our models, spatial random effects have multivariate normal with mean $\mathbf{0}$ and variance-covariance matrix $\sigma_u^2\Sigma_u(\rho_u)$. Hence the conditional normal distribution of $u(s)$ given the other random effects has $N(\bar{\theta}_u, \sigma_u^2\bar{\Sigma}_u)$, where $\bar{\theta}_u = \Sigma_{u_{12}}\Sigma_{u_{22}}^{-1}(\mathbf{a})$ and $\bar{\Sigma}_u = \Sigma_{u11} - \Sigma_{u_{12}}\Sigma_{u_{22}}^{-1}\Sigma_{u_{21}}$. The proposal distribution is conditional normal distribution $N(\bar{\theta}, \sigma_0^2\bar{\Sigma}_u)$, where σ_0^2 is the proposed variance of the random effects.

Because of the conditional independence, the acceptance probability can be simplified to

$$\min\left\{ \frac{f[\mathbf{Y}(s)|u^*(s), \boldsymbol{\nu}, \boldsymbol{\mu}(s)] f_u[u^*(s)|\bar{\theta}_u, \sigma_u^2\bar{\Sigma}_u(\rho_u)] f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]}{f[\mathbf{Y}(s)|u(s), \boldsymbol{\nu}, \boldsymbol{\mu}(s)] f_u[u(s)|\bar{\theta}_u, \sigma_u^2\bar{\Sigma}_u(\rho_u)] f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]}, 1 \right\},$$

where $f_u[u(s)|\bar{\theta}_u, \sigma_u^2\bar{\Sigma}_u(\rho_u)]$ is the conditional distribution of $u(s)$ given all the other random effects.

Proceeding the same manner one can drive the acceptance ratio of time effects as

$$\text{If U} < \min\left\{ \frac{f[\mathbf{Y}(t)|\nu^*(t), \mathbf{u}, \boldsymbol{\mu}(t)] f_\nu[\nu^*(t)|\bar{\theta}_\nu, \sigma_\nu^2\bar{\Sigma}_\nu(\rho_\nu)] f_u[\mathbf{u}|\sigma_u^2\Sigma_u(\rho_u)]}{f[\mathbf{Y}(t)|\nu^(t), \mathbf{u}, \boldsymbol{\mu}(t)] f_\nu[\nu^(t)|\bar{\theta}_\nu, \sigma_\nu^2\bar{\Sigma}_\nu(\rho_\nu)] f_u[\mathbf{u}|\sigma_u^2\Sigma_u(\rho_u)]}, 1 \right\}.$$

Assuming we have n spatial locations, $\boldsymbol{\mu}(s), (s = 1, \ldots, n)$, σ_0^2 Σ_u, and the estimated unknown function, $\hat{m}(\cdot)$, we perform a subroutine for the M-H algorithm to generate N samples

for each spatial random effect and each time effect from $\log f[\mathbf{Y}, \mathbf{u}, \boldsymbol{\nu}|\boldsymbol{\mu}, \sigma_u^2\Sigma(\rho_u), \sigma_\nu^2\Sigma_\nu(\rho_\nu = 2)]$. This subroutine procedure is summarized as follows:

Subroutine for M-H algorithm:

Step 0 Given σ_u^2, σ_ν^2, $\boldsymbol{\mu}(s,t)$, and initial values for $\mathbf{u^{(0)}}$, and $\boldsymbol{\nu}^{(0)}$, set $s = 1$, $t = 1$ and $i = 1$;

Step 1 Given $\boldsymbol{\nu}$= mean($\boldsymbol{\nu}^{[0:(i-1)]}$), generate a candidate value for spatial random effect at location s from $N(\bar{\theta_u}, \sigma_0^2\bar{\Sigma_u})$, $u^*(s)$ and generate a uniform(0,1) random value U,

$$\text{If U} < \min\left\{\frac{f[\mathbf{Y}(s)|u^*(s), \boldsymbol{\nu}, \boldsymbol{\mu}(s)]f_u[u^*(s)|\bar{\theta}_u, \sigma_u^2\bar{\Sigma}_u(\rho_u)]f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]}{f[\mathbf{Y}(s)|u(s), \boldsymbol{\nu}, \boldsymbol{\mu}(s)]f_u[u(s)|\bar{\theta}_u, \sigma_u^2\bar{\Sigma}_u(\rho_u)]f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]}, 1\right\},$$

then set $\mathbf{u}^{(i)} = [u^*(s), u(2), \ldots, u(S)]$. Otherwise, $\mathbf{u}^{(i)} = \mathbf{u}$ stays unchanged.

Step 2 Set s=s+1 and repeat Step 1 until reach location number S;

Step 3 Given $\mathbf{u}$= mean($\mathbf{u}^{[0:(i-1)]}$), generate a candidate value for time effect at time point t from $N(\bar{\theta_\nu}, \sigma_0^2\bar{\Sigma_\nu})$, $\nu^*(t)$, and generate a uniform(0,1) random value U,

$$\text{If U} < \min\left\{\frac{f[\mathbf{Y}(t)|\nu^*(t), \mathbf{u}, \boldsymbol{\mu}(t)]f_\nu[\nu^*(t)|\bar{\theta}_\nu, \sigma_\nu^2\bar{\Sigma}_\nu(\rho_\nu)]f_u[\mathbf{u}|\sigma_u^2\Sigma_u(\rho_u)]}{f[\mathbf{Y}(t)|\nu(t), \mathbf{u}, \boldsymbol{\mu}(t)]f_\nu[\nu(t)|\bar{\theta}_\nu, \sigma_\nu^2\bar{\Sigma}_\nu(\rho_\nu)]f_u[\mathbf{u}|\sigma_u^2\Sigma_u(\rho_u)]}, 1\right\},$$

then set $\boldsymbol{\nu}^{(i)} = [\nu^*(s), \nu(2), \ldots, \nu(T)]$. Otherwise, $\mathbf{u}^{(i)} = \mathbf{u}$ stays unchanged.

Step 4 Set t=t+1 and repeat Step 3 until reach time point number T;

Step 5 Repeat Step 1-4, N times, $i = 1, 2, \ldots, N$.

Note that here we take a sample only after each coordinate has been visited and the first N_0 burn-in samples should be discarded. Geyer (1992) suggested using an N_0 that is between 1% and 2% of the run length N.

4.3.2 Estimation for Semiparametric Spatio-Temporal Separable Single Index Model

The complete-data log-likelihood for the first proposed model takes the form

$$\begin{aligned}\log f[\mathbf{Y},\mathbf{u},\boldsymbol{\nu}|\boldsymbol{\mu}=m[X(s,t)\boldsymbol{\alpha}]+Z\mathbf{u}+W\boldsymbol{\nu},\sigma_u^2\Sigma(\rho_u),\sigma_\nu^2\Sigma_\nu(\rho_\nu)] &= \log f_{Y|u,\nu}[\mathbf{Y}|\boldsymbol{\mu},\mathbf{u},\boldsymbol{\nu}]\\ &+ \log f_u[\mathbf{u}|\sigma_u^2\Sigma(\rho_u)]\\ &+ \log f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)],\end{aligned}$$

where $\mathbf{Y}\sim \text{Pois}[\boldsymbol{\mu}|\mathbf{u},\boldsymbol{\nu}]$, $[\boldsymbol{\mu}|\mathbf{u},\boldsymbol{\nu}]=m[X(s,t)\boldsymbol{\alpha}]+Z\mathbf{u}+W\boldsymbol{\nu}$, $\mathbf{u}\sim GP[\mathbf{0},\sigma_u^2\Sigma_u(\rho_u)]$, $\boldsymbol{\nu}\sim GP[\mathbf{0},\sigma_\nu^2\Sigma_u(\rho_\nu)]$, $\sigma_u^2\Sigma_u(\rho_u)=Cov[u(s+d),u(s)]=\sigma_u^2 exp(||d||^2/\rho_u)$ for all $s,d\in R^2$, and $\sigma_\nu^2\Sigma_\nu(\rho_u)=Cov[\nu(t+\delta),\nu(t)]=\sigma_\nu^2 exp(||\delta||^2/\rho_\nu)$ for all $t,\delta\in R$. The following is the proposed algorithm of estimating SSTS-SIM parameters and spatial and time effects:

Proposed algorithm III

To run MCEM algorithm, we need to initialize $\boldsymbol{\alpha}$, $m(\cdot)$, $u(s)$, $(s=1,\ldots,S)$, $\nu(t)$, $(t=1,\ldots,T)$, σ_ν^2, σ_u^2, and estimates of ρ_u and ρ_ν ($\hat{\rho}_u$ and $\hat{\rho}_\nu$). The proposed algorithm to estimate SSTS-SIM parameters and spatial and time effects is as follows:

Step 0 Initialize parameters:

1. Initialize $\sigma_u^{2(0)}$, $u(s)^{(0)}$, $\sigma_\nu^{2(0)}$, and $\nu(t)^{(0)}$;
2. $Y(s,t)^*=Y(s,t)-u(s)^{(0)}-\nu(t)^{(0)}$;
3. By using Ichimura method, estimate $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(0)}$, and smooth $m(\cdot)$ using some bandwidth to get $\hat{m}(\cdot)^{(0)}$.

E-step Given all initials, generate N values for each location effect $u(s)$, $u(s)_1,u(s)_2,....,u(s)_N$ and N values for each time effect $\nu(t)$, $\nu(t)_1,\nu(t)_2,....,\nu(t)_N$ from

$\log f[\mathbf{Y}, \mathbf{u}, \boldsymbol{\nu}|\boldsymbol{\mu} = m[X(s,t)\boldsymbol{\alpha}] + Z\mathbf{u} + W\boldsymbol{\nu}, \sigma_u^2\Sigma(\hat{\rho}_u), \sigma_\nu^2\Sigma_\nu(\hat{\rho}_\nu)]$ via Subroutine for M-H algorithm which is described in 4.3.1

M-step Maximize $\frac{1}{N}\sum_{k=N_0+1}^{N} \log f\{\mathbf{u}_k|\sigma_u^2\Sigma_u(\hat{\rho}_u)\}$ and $\frac{1}{N}\sum_{k=N_0+1}^{N} \log f\{\boldsymbol{\nu}_k|\sigma_\nu^2\Sigma_\nu(\hat{\rho}_\nu)\}$:

1. Get $\sigma_u^{2(1)}$ and $\sigma_\nu^{2(1)}$;
2. Calculate $u(s)^{(1)} = \frac{1}{N}\sum_{k=1}^{N} u(s)_k$, $\nu(t)^{(1)} = \frac{1}{N}\sum_{k=1}^{N} \nu(t)_k$ and $Y(s,t)^* = Y(s,t) - u(s)^{(1)} - \nu(t)^{(1)}$;
3. Estimate $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(1)}$, and smooth the function $m(\cdot)$ to get $\hat{m}(\cdot)^{(1)}$.

Iterate E-step and M-step until convergence.

4.3.3 Estimation for Semiparametric Spatio-Temporal Nonseparable Single Index Model

The complete-data log-likelihood for the second proposed model takes the form

$$\begin{aligned}\log f[\mathbf{Y}, \mathbf{u}, \boldsymbol{\nu}|\boldsymbol{\mu} = m[X(s,t)\boldsymbol{\alpha} + Z\mathbf{u}] + W\boldsymbol{\nu}, \sigma_u^2\Sigma(\rho_u), \sigma_\nu^2\Sigma_\nu(\rho_\nu)] &= \log f_{Y|u,\nu}[\mathbf{Y}|\boldsymbol{\mu}, \mathbf{u}, \boldsymbol{\nu}] \\ &+ \log f_u[\mathbf{u}|\sigma_u^2\Sigma(\rho_u)] \\ &+ \log f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)],\end{aligned}$$

where $\mathbf{Y} \sim \text{Pois}[\boldsymbol{\mu}|\mathbf{u}, \boldsymbol{\nu}]$, $[\boldsymbol{\mu}|\mathbf{u}, \boldsymbol{\nu}] = m[X(s,t)\boldsymbol{\alpha} + Z\mathbf{u}] + W\boldsymbol{\nu}$, $\mathbf{u} \sim GP[\mathbf{0}, \sigma_u^2\Sigma_u(\rho_u)]$, $\boldsymbol{\nu} \sim GP[\mathbf{0}, \sigma_\nu^2\Sigma_u(\rho_\nu)]$, $\sigma_u^2\Sigma_u(\rho_u) = Cov[u(s+d), u(s)] = \sigma_u^2 exp(||d||^2/\rho_u)$ for all $s, d \in R^2$, and $\sigma_\nu^2\Sigma_\nu(\rho_u) = Cov[\nu(t+\delta), \nu(t)] = \sigma_\nu^2 exp(||\delta||^2/\rho_\nu)$ for all $t, \delta \in R$.

Semiparametric spatio-temporal nonseparable single index model enables us to estimates all the parameters in the model without putting restrictions on the parameters because the spatial random effects already fix the problem of the identifiability where we consider the

spatial random effect is the variable which its coefficient is equal to one.

The previous proposed algorithm III does not work for this model because of two issues arise: (1) intensive calculations: if we run the M-H algorithm 1000 times using a single component with six spatial random effects and six time points for each location, we need to fit the single index model 36000 times to run the MCEM algorithm only one time; (2) we can not separate the spatial random effects from the single index coefficient parameters estimates in the M-H step, while we are comparing the current and the previous single-component. This point can be explained with details as follows:

The acceptance ratio

$$\min\left\{\frac{f[\mathbf{Y}|\mathbf{u}^*,\boldsymbol{\nu},\boldsymbol{\mu}]f_u[\mathbf{u}^*|\sigma_u^2\Sigma_u(\rho_u)]f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]}{f[\mathbf{Y}|\mathbf{u},\boldsymbol{\nu},\boldsymbol{\mu}]f_u[\mathbf{u}|\sigma_u^2\Sigma_u(\rho_u)]f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]},1\right\}$$

can be re-written for the SSTN-SIM as

$$\min\left\{\frac{f[\mathbf{Y}|\mathbf{u}^*,\boldsymbol{\nu},\hat{m^*}[X\hat{\boldsymbol{\alpha}^*}+Z\mathbf{u}^*]+W\boldsymbol{\nu}]f_u[\mathbf{u}^*|\sigma_u^2\Sigma_u(\rho_u)]f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]}{f[\mathbf{Y}|\mathbf{u}^*,\boldsymbol{\nu},\hat{m^*}[X\hat{\boldsymbol{\alpha}}+Z\mathbf{u}]+W\boldsymbol{\nu}]f_u[\mathbf{u}|\sigma_u^2\Sigma_u(\rho_u)]f_\nu[\boldsymbol{\nu}|\sigma_\nu^2\Sigma_\nu(\rho_\nu)]},1\right\}.$$

In this case, if the ratio is greater than one, it is not known whether $\mathbf{u}^*$ is better than $\mathbf{u}$ or because of the difference between $\hat{m}^*$ and $\hat{m}^*$ or the difference between $\hat{\boldsymbol{\alpha}}^*$ and $\hat{\boldsymbol{\alpha}}$. In this case, we can not compare the two spatial random effects at the same values of the other model parameters. To solve this problem, we use a linear approximation for the unknown function at value $[X\boldsymbol{\alpha}^{(0)}+Z\mathbf{u}^{(0)}]$ to separate the spatial random effect and the other model parameters which will be

$$\boldsymbol{\mu}=m[X\boldsymbol{\alpha}+Z\mathbf{u}]+W\boldsymbol{\nu}=\hat{m}+\hat{m'}\times[X\boldsymbol{\alpha}+Z\mathbf{u}-X\boldsymbol{\alpha}^{(0)}-Z\mathbf{u}^{(0)}]+W\boldsymbol{\nu},$$

where $\hat{m}(\cdot)$ is the estimate of the unknown function using a smoothing method such as p-spline (Ruppert et al., 2003), kernel smoothing (Wand and Jones, 1995) or any other basis function and $\hat{m'}$ is the estimate of the first derivative of the unknown function. We use local linear kernel regression to estimate the function and its first derivative. In general, the estimator of the j^{th} derivative $m^{(j)}(x)$ at a point x is given by $\hat{m}^{(j)}(x) = j!\hat{\beta}_j(x)$ for the local polynomials of degree d of the form

$$m(x_i) = \beta_0 + \beta_1(x_i - x) + \ldots + \beta_1(x_i - x)^d.$$

Finally we propose algorithm VI which is summarized as follows;

Proposed algorithm VI

To run MCEM algorithm, we need to initialize $\boldsymbol{\alpha}$, $m(\cdot)$, $u(s)$, $(s = 1, \ldots, S)$, $\nu(t)$, $(t = 1, \ldots, T)$, σ_ν^2, σ_u^2, and estimates of ρ_u and ρ_ν ($\hat{\rho}_u$ and $\hat{\rho}_\nu$). The proposed algorithm to estimate SSTN-SIM parameters and spatial and time effects is as follows:

Step 0 Initialize parameters:

1. Initialize $\sigma_u^{2(0)}$, $u(s)^{(0)}$, $\sigma_\nu^{2(0)}$, and $\nu(t)^{(0)}$;
2. $Y(s,t)^* = Y(s,t) - \nu(t)^{(0)}$;
3. By using Ichimura method, estimate $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(0)}$, and then smooth $m(\cdot)$ using some bandwidth to get $\hat{m}(\cdot)^{(0)}$ and $\hat{m'}(\cdot)^{(0)}$, where $\hat{m'}(\cdot)^{(0)}$ is the estimate of the first derivative of $m(\cdot)$.

E-step Given $\hat{m}(\cdot)$ and $\hat{m'}(\cdot)$, and using the Taylor approximation of $m(\cdot)$, generate N values for each location effect $u(s)$, $u(s)_1, u(s)_2, ...,u(s)_N$ and N values for each time effect $\nu(t)$, $\nu(t)_1, \nu(t)_2, ...,\nu(t)_N$ from

$\log f[\mathbf{Y}, \mathbf{u}, \boldsymbol{\nu}|\boldsymbol{\mu} = \hat{m} + \hat{m}'[X\boldsymbol{\alpha} + Z\mathbf{u} - X\boldsymbol{\alpha}^{(0)} - Z\mathbf{u}^{(0)}] + W\boldsymbol{\nu}, \sigma_u^2\Sigma(\hat{\rho}_u), \sigma_\nu^2\Sigma_\nu(\hat{\rho}_\nu)]$ via Subroutine for M-H algorithm which is described in 4.3.1

M-step Maximize $\frac{1}{N}\sum_{k=N_0+1}^{N} \log f\{\mathbf{u}_k|\sigma_u^2\Sigma_u(\hat{\rho}_u)\}$ and $\frac{1}{N}\sum_{k=N_0+1}^{N} \log f\{\mathbf{u}_k|\sigma_\nu^2\Sigma_\nu(\hat{\rho}_\nu)\}$:

1. Get $\sigma_u^{2(1)}$ and $\sigma_\nu^{2(1)}$;
2. Calculate $u(s)^{(1)} = \frac{1}{N}\sum_{k=1}^{N} u(s)_k$, $\nu(t)^{(1)} = \frac{1}{N}\sum_{k=1}^{N} \nu(t)_k$ and $Y(s,t)^* = Y(s,t) - \nu(t)^{(1)}$;
3. Estimate $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(1)}$, and smooth the function $m(\cdot)$ to get $\hat{m}(\cdot)^{(1)}$.

Iterate E-step and M-step until convergence.

4.4 Real Data Application

In this section, the two proposed models, SSTS-SIM and SSTN-SIM, are applied to the South Korea data set; the dependence range, the models parameters and their confidence intervals, and the unknown semiparametric functions for each city will be estimated. In addition a model selection study is done to select the suitable model for the South Korea mortality data.

4.4.1 Data and Model

In the South Korea data set, non accident mortality and other covariates, such as mean temperature, mean humidity, mean pressure and time, were recorded daily during the period from January, 2000 to December, 2007 for six major cities: Seoul, Incheon, Daejeon, Daegu, Gwangju, and Busan. In total there are 2922 observations for each city. A monthly data set is created by calculating the mean of the weather variables per month. The monthly data

are used for two reasons: (1) to avoid the big "N" problem (Banerjee et al., 2004) because if we used the daily data we will be having a large covariance matrix and require a substantial amount of computing time, however using monthly data has a covariance matrix of spatial effects of order 6×6 and covariance matrix of time effects of order 12×12 and (2) to study the pattern of the mortality during the year.

In monthly data, we have six cities and 12 time points for each city. In total we have 72 observations. Those cities are different in population size. Hence we first divided monthly non accident mortality by population size of each city and multiplied by 100,000 and then obtain monthly non accident mortality per 100,000 persons in each city. Figure 3.11 shows the characteristics of the 6 metropolitan areas in South Korea.

Figure 4.1(a-c) show that the relationship between non accident mortality and temperature and humidity is negative, while the relationship between mortality and mean pressure is positive. Figure 4.1(d) shows that June, July and August have the lowest mean monthly mortality but November-February have the highest mean monthly mortality. We can also observe the effect of location on mortality from scatterplots in Figure 4.1(a-d) where the observations of each city is distinguished from the others. Busan has the highest mortality, while Seoul and Daejeon have the lowest mortality.

Our goals are: (1) including the correlated spatial random effects and time effects into single index model, (2) proposing two algorithms for fitting the two models, SSTS-SIM and SSTN-SIM , (3) studying the effect of including spatial random effects and time effects on estimation and prediction, (4) addressing the prediction performance of the proposed models, and (5) selecting the best model for our data set.

For the South Korea data, SIM, SSS-SIM, SSSN-SIM, and our new models, SSTS-SIM and SSTN-SIM, have the following forms:

- SIM

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\mu} &\sim \text{Pois}[\boldsymbol{\mu}], \\ \boldsymbol{\mu} &= m[\mathbf{x}_1\alpha_1 + \mathbf{x}_2\alpha_2 + \mathbf{x}_3\alpha_3], \end{aligned}$$

- SSS-SIM

$$\begin{aligned} \mathbf{Y}(s)|\boldsymbol{\mu}(s) &\sim \text{Pois}[\boldsymbol{\mu}(\text{s})|\text{u(s)}], \\ \boldsymbol{\mu}(s)|u(s) &= m[\mathbf{x}_1(s)\alpha_1 + \mathbf{x}_2(s)\alpha_2 + \mathbf{x}_3(s)\alpha_3] + u(s), \end{aligned}$$

- SSN-SIM

$$\begin{aligned} \mathbf{Y}(s)|\boldsymbol{\mu}(s) &\sim \text{Pois}[\boldsymbol{\mu}(\text{s})|\text{u(s)}], \\ \boldsymbol{\mu}(s)|u(s) &= m[\mathbf{x}_1(s)\alpha_1 + \mathbf{x}_2(s)\alpha_2 + \mathbf{x}_3(s)\alpha_3 + u(s)], \end{aligned}$$

- SSTS-SIM

$$\begin{aligned} Y(s,t)|\mu(s,t) &\sim \text{Pois}[\mu(\text{s,t})|\text{u(s)}, \nu(\text{t})], \\ \mu(s,t)|u(s),\nu(t) &= m[x_1(s,t)\alpha_1 + x_2(s,t)\alpha_2 + x_3(s,t)\alpha_3] + u(s) + \nu(t), \end{aligned}$$

- SSTN-SIM

$$\begin{aligned} Y(s,t)|\mu(s,t) &\sim \text{Pois}[\mu(\text{s,t})|\text{u(s)}, \nu(\text{t})], \\ \mu(s,t)|u(s),\nu(t) &= m[x_1(s,t)\alpha_1 + x_2(s,t)\alpha_2 + x_3(s,t)\alpha_3 + +u(s)] + \nu(t), \end{aligned}$$

where

1. $\mathbf{x}_i(s)$, (i=1,2,3), is the explanatory variable $\mathbf{x}_i$ vector at location s, $m(\cdot)$ is unknown function, and $\alpha_1, \alpha_2, \alpha_3,$ and α_4 are the single index coefficients parameters. $x_i(s,t)$ is the value of X_i at location s and time point t.

2. $\{u(s), s \in R^2\}$ is a Gaussian stationary process with $E[u(s)] = 0$ for all s and cov(u(s+ d), u(s) = C(d) for all $s, d \in R^2$, where the covariance function is $C(\cdot)$ and d is the distance between the two locations. $\{\nu(t), t \in R\}$ is time random effect at time point t of location s.

3. Conditionally on $\{u(s), s \in R^2$ and $\nu(t), t \in R\}$, Y(s,t) is independent process and the distribution of $Y(s,t)$ is specified by the conditional mean $E[Y(s,t)|u(s), \nu(t)]$.

4. X_1 : monthly mean temperature, X_2 : monthly mean humidity, X_3 : monthly mean pressure.

4.4.2 SSTS-SIM and SSTN-SIM Estimation

In this section, Algorithm III is used to estimate SSTS-SIM parameters and Algorithm VI is used to estimate SSTN-SIM parameters, respectively. Generalized linear mixed models parameters estimates are used as initial values for the two proposed models parameters ($\alpha's$, σ_u^2, σ_ν^2) and spatial effects and time effects ($\mathbf{u}$ and $\boldsymbol{\nu}$). Spatial and time effects samples are obtained from M-H subroutine after discarding 2% of the MCMC run length. We run the M-H algorithm for 5000 times. The complete MCEM algorithm was run for 30 times and the estimates have been taken that corresponding to the maximum likelihood function value. From Figure 4.2 which shows different types of variograms, the dependence range estimates is about 1.5, $\rho_\nu = 1.5$, where the correlation stops increasing after this distance.

4.4.3 SSTS-SIM Estimation

Estimation and asymptotic standard errors of SSTS-SIM parameters are shown in Table 4.1. Ichimura (1993) proved that the estimator of $\boldsymbol{\alpha}$ convergence with rate $1/\sqrt(n)$ and the asymptotic distribution of $1/\sqrt(n)[\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}]$ is normal with mean $\mathbf{0}$ and a variance-covariance matrix has some specific structure. The standard error shown in Table 4.1 is the standard error obtained from Ichimura's asymptotic covariance matrix. It is used here rather than the bootstrapped-based standard error, such as in Chapter 3 to save time because it is found that the bootstrapped-based standard error is comparable to the asymptotic standard error. Table 4.1 shows R^2 of SSTS-SIM is 0.83 which is acceptable and log likelihood value is -313.83. The mean pressure coefficient is set to be equal 1 to fix the identifiability problem. All the 95% confidence intervals do not include zero value.

Table 4.1: Parameters estimates for SSTS-SIM=$m(X\boldsymbol{\alpha})+Zu+W\boldsymbol{\nu}$ model and their standard error (SE) from the asymptotic covariance matrix.

	Mean temperature	Mean humidity	Mean pressure	log likelihood	R^2
Estimate	-2.1734	3.7409	1	-313.8392	0.8301
Asymptotic SE	1.0520	0.1507	0		
95% CI	(-4.2353, -0.1115)	(3.4455, 4.0363)	-		

To calculate the standard error of the spatial random effects and spatial variance, a bootstrapped method is used, The steps of the bootstrapped procedure are described as follows:

1. From each location, 12 observations have been randomly selected with replacement.
2. Estimate SSTS-SIM using the bootstrapped data and obtain the parameters estimates ($\hat{\boldsymbol{\alpha}}$), spatial random effects estimates ($\hat{\mathbf{u}}$), time effects ($\hat{\nu}$) and spatial variance σ_u^2 and time variance ($\hat{\sigma_\nu^2}$).

3. Repeat step 1 and 2 for 250 times and calculate the standard error of the 250 estimates of each parameter.

Table 4.2 reveals the spatial effects estimation of the six cities. It can be seen that Busan has the highest mortality location effect and Daegu has the second highest. The lowest mortality location effects are in Daejeon and Seoul.

Table 4.2: Correlated spatial random effects estimates, their standard error (SE), and 95% confidence interval for SSTS-SIM=$m(X\boldsymbol{\alpha}) + Zu + W\boldsymbol{\nu}$ model of 250 bootstrapped simulated data sets

	Spatial estimate ($\hat{\mathbf{u}}$)	Bootstrap-based SE	95% CI
Seoul	-17.197	0.120	(-17.432, -16.962)
Busan	26.438	0.091	(26.259, 26.616)
Daegu	10.791	0.140	(10.516, 11.065)
Incheon	-13.191	0.112	(-13.410, -12.971)
Gwangju	-12.462	0.130	(-12.717, -12.207)
Daejeon	-18.925	0.081	(-19.083, -18.766)
σ_u^2	231.467	8.501	(214.807, 248.127)

The actual mortality values versus fitted values is displayed in Figure 4.3. Overall the fitting seems reasonable except for small and large mortality. We can see an overestimation for the small values and an underestimation for the large values.

Figure 4.4 shows the actual monthly non accident mortality and the fitted one for each city. The fitted values and the actual values are not very close to each other. The over and under estimations are showed.

Figure 4.5 (left) shows the unknown function estimate, $\hat{m}(X\boldsymbol{\alpha})$, is increasing as single index value increases. Figure 4.5 (right) shows the six estimated unknown functions which are

obtained by adding the spatial effect to the common unknown mortality function. It indicates that Busan has the highest mortality function and Daejeon has the lowest mortality function.

From Figure 4.6 (left), one can see that there are some patterns in the residual plot, there is a decreasing pattern. Figure 4.6 (right) shows the residuals are clustered.

4.4.4 SSTN-SIM Estimation

Using our SSTN-SIM, we can estimate all the parameters because we do not need to specify restriction on $\boldsymbol{\alpha}$ for identifiability problem. The spatial random effect can be treated as the variable which its coefficient is set to be 1. Table 4.3 shows the model parameter estimates, their standard errors, and 95% confidence intervals. One can see that the parameter estimates of SSTN-SIM are smaller than those of SSTS-SIM but the signs are the same. The asymptotic standard error values of SSTN-SIM parameters are much smaller than those of SSTS-SIM parameters. Spatial random effects of SSTN-SIM have the same signs of spatial effects of SSTS-SIM but smaller values. Again Busan has the highest mortality, while Seoul and Daejeon have the smallest spatial location effect on non accident mortality. Table 4.4 shows the variance of spatial random effects, σ_u^2, is very small comparing to SSTS-SIM spatial variance. They are 0.001538 and 231.46, respectively. SSTN-SIM is better than SSTS-SIM using the monthly data in terms of Log likelihood and R^2. log likelihood of SSTS-SIM and SSTN-SIM are -313.83 and -236.54, respectively. R^2 are 0.83 and 0.94, respectively.

In Figure 4.7, the scatter plot between actual mortality values versus fitted values is displayed. There are strong linear relationship. The points are scattered near 45° line and R^2

Table 4.3: Parameters estimates and their standard errors (SE) which are calculated using the asymptotic covariance matrix in SSTN-SIM=$m(X\boldsymbol{\alpha} + Zu) + W\boldsymbol{\nu}$

	Parameter estimate	Asymptotic SE	95% CI
Mean temperature	-0.005342	1.123e-06	(-0.005344, -0.005340)
Mean humidity	0.002597	2.234e-06	(0.002593, 0.002601)
Mean pressure	0.004642	2.592e-06	(0.004638, 0.004648)
log Likelihood	-236.548601		
R^2	0.940102		

Table 4.4: Correlated spatial effects estimates, their standard error (SE), and 95% confidence interval for SSTN-SIM=$m(X\boldsymbol{\alpha} + Zu) + W\boldsymbol{\nu}$ model based on 250 bootstrapped simulated data sets

	Spatial estimate ($\hat{\mathbf{u}}$)	Bootstrap-based SE	95% CI
Seoul	-0.0711	0.0021	(-0.0750, -0.0672)
Busan	0.1973	0.0031	(0.1915, 0.2032)
Daegu	0.1034	0.0012	(0.1015, 0.1054)
Incheon	-0.0307	0.0012	(-0.0327, -0.0287)
Gwangju	-0.0646	0.0021	(-0.0686, -0.0607)
Daejeon	-0.0890	0.0023	(-0.0929, -0.0851)
σ_u^2	0.0015	0.0001	(0.0013, 0.0017)

is 0.94 which is larger than R^2's value, 0.83, obtained from using SSTS-SIM.

Figure 4.8 shows the actual and fitted monthly non accident mortality for each city and each month. The fitted and the actual values are very close to each other which means that SSTN-SIM fit the data much better than SSTS-SIM.

Figure 4.9 shows the unknown mortality function estimate of each city and its 95% confidence interval. It indicates Busan has the highest mortality function but Daejeon has the lowest mortality function.

There is no pattern for the residuals in Figure 4.10 (left) and the residuals are randomly distributed about zero line in Figure 4.10 (right). We conclude that SSTN-SIM could capture the relationship between non accident mortality and the explanatory variables better than SSTS-SIM.

For modeling time effects, Gaussian process or the modified random walk, RW(1), with the first order, are used for covariance function. To see which covariance function gives us more stable results, we repeated the estimation of the two models 10 times. Standard deviation of estimates are summarized in Table 4.5. It shows that both covariance functions estimates are comparable in terms of standard deviation. Both give stable results.

4.5 Prediction and Model Selection

This section describes the performance of SSTS-SIM and SSTN-SIM in terms of prediction and estimation. The performance of the two proposed models will also be compared to single index model without spatial random effects, semiparametric spatial separable single index, and semiparametric spatial nonseparable single index model. Several criteria are used to

Table 4.5: Standard deviation (SD) of 10 runs of estimating spatial effects and its variance, σ_u^2, of SSTN-SIM=$m(X\boldsymbol{\alpha}) + Zu + W\boldsymbol{\nu}$ and SSTS-SIM=$m(X\boldsymbol{\alpha} + Zu) + W\boldsymbol{\nu}$ using random walk, RW(1), with the first order and Gaussian process with $\rho_\nu = 2$

	SSTS-SIM		SSTN-SIM	
	RW(1)	Gaussian	RW(1)	Gaussian
Seoul	0.93	0.76	0.02	0.01
Busan	1.29	1.12	0.07	0.04
Daegu	0.66	0.69	0.03	0.02
Incheon	0.79	0.96	0.02	0.00
Gwangju	0.67	0.74	0.02	0.01
Daejeon	1.23	0.75	0.03	0.01
σ_u^2	10.53	10.8	0.00	0.00
σ_ν	0.01	0.01	0.00	0.00

to compare their performance in estimation, such as R^2, MASE, and log likelihood value for given estimated parameters obtained from training data (LogLE). We also compare the prediction performance of the two proposed models, SSTS-SIM and SSTN-SIM in terms of average and median of predicted mean square error (APMSE and MPMSE, respectively) and predicted log likelihood value of test data given estimated parameters obtatined from training data (PLogLE). The following steps are conducted to calculate these criteria:

1. Set $n = 1$, $j = 1$

2. Randomly select n observations from each location (city) and consider it as a test data set and the rest will be considered as a training data set.

3. Fit the SSTS-SIM using the training data and calculate R^2, MSE_j, and the log likelihood value for given estimated parameters, which is defined as

$$\begin{aligned} \text{LogLE}_j &= \log f_{Y|u,\nu}[\mathbf{Y}|\hat{\boldsymbol{\mu}} = \hat{m}(X\hat{\boldsymbol{\alpha}}) + Z\hat{\mathbf{u}}, \hat{\mathbf{u}}, \hat{\boldsymbol{\nu}}] + \log f_u[\hat{\mathbf{u}}|\hat{\sigma}_u^2\Sigma(\hat{\rho}_u)] \\ &+ \log f_\nu[\hat{\boldsymbol{\nu}}|\hat{\sigma}_\nu^2\Sigma_\nu(\hat{\rho}_\nu)] \end{aligned}$$

where $\mathbf{Y}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\nu}}$ and $\hat{\mathbf{u}}$ are the vectors of the response of the training data set, estimated conditional mean, time effect and spatial effects and $\hat{\sigma}_u^2$, $\hat{\sigma}_\nu^2$ and $\hat{\rho}$ are the estimated variance of spatial effects, variance of time effects and dependence range.

4. Use the test data set to calculate PMSE_j,

$$\text{PMSE}_j = \sum_{s=1}^{6} \sum_{i=1}^{n} \frac{(y_i(s) - \hat{y_i^*}(s))^2}{6n},$$

where $y_i(s)$ and $\hat{y_i^*}(s)$ are the i^{th} actual and predicted response values at location (s), respectively; and also calculate the predicted log likelihood value of test data given estimated parameters, which is defined as

$$\begin{aligned} \text{PLogLE}_\text{j} &= \log f[\mathbf{Y}^* | \hat{\boldsymbol{\mu}} = m(X\hat{\boldsymbol{\alpha}}) + Z\hat{\mathbf{u}} + W\hat{\boldsymbol{\nu}}, \hat{\mathbf{u}}] + \log f_u[\hat{\mathbf{u}} | \hat{\sigma}_u^2 \Sigma(\hat{\rho}_u)] \\ &+ \log f_\nu[\hat{\boldsymbol{\nu}} | \hat{\sigma}_\nu^2 \Sigma_\nu(\hat{\rho}_\nu)], \end{aligned}$$

where $\mathbf{Y}^*$ is the vector of response values in the testing data, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\nu}}$ and $\hat{\mathbf{u}}$ are the estimated vectors of the conditional mean, time effects and spatial effects and $\hat{\sigma}_u^2$, $\hat{\sigma}_\nu^2$ and $\hat{\rho}$ are the estimated variance of spatial effects, variance of time effects and dependence range.

5. Repeat 1-4 for 500 times (j=1,. . . , 500).

6. Calculate the average and median of the 500 estimates of PMSE_j (APMSE and MPMSE, respectively). Also calculate R^2, MSE, PLogLE, and LogLE, where

$$\begin{aligned} \text{R}^2 &= \sum_{j=1}^{500} \frac{R_j^2}{500} \\ \text{MSE} &= \sum_{j=1}^{500} \frac{MSE_j}{500} \end{aligned}$$

$$\text{PLogLE} = \sum_{j=1}^{500} \frac{PLogLE_j}{500}$$
$$\text{LogLE} = \sum_{j=1}^{500} \frac{LogLE_j}{500}.$$

7. Repeat 2-5 for different n (=2, 4, 6) and $j = 1, \ldots, 500$.

8. Repeat 1-7 for SSTN-SIM.

The results are shown in Table 4.6. Several estimation criteria (LogLE, MSE, and R^2) have been reported for SIM, SSS-SIM, SSN-SIM, SSTS-SIM, and SSTN-SIM. Using all data (all the data as training data), it can be seen that SIM without spatial and time random effects does not fit the data well comparing to other models in terms of LogLE, MSE, and R^2. This means including spatial random effects in SSS-SIM and SSN-SIM can improve the model estimation. For the two models including spatial effects only, SSN-SIM is better than SSS-SIM. For the two models including both spatial and time effects, SSTN-SIM fits the data much better than SSTS-SIM. As a result, SIM is the worst and SSTN-SIM is the best in terms of fitting the data.

For prediction, in case $n = 2$ where two observation has been selected randomly form each location and in total we obtain 12 observations from all the locations as a test data set, one can see that SSTN-SIM performs better than SSTS-SIM in terms of R^2, MSE, PLogLE, LogLE, and MPMSE. APMSE of SSTN-SIM is larger than APMSE of SSTS-SIM because SSTN-SIM gives some outliers in prediction. Similarly, in case $n = 4$, SSTN-SIM works better than SSTS-SIM. In case $n = 6$, we have 36 observations as testing data and 36 observations as training data. In this case SSTN-SIM is better than SSTS-SIM in terms of PLogLE and PLogLE, however, SSTS-SIM is a little better than SSTN-SIM.

Figure 4.11 shows the boxplots of the 500 estimates of PMSE for each model and each sample test data set size ($n = 2, 4, 6$). It shows that the variation of SSTN-SIM prediction estimates is almost the same for both models when $n = 2$. However, when $n = 4$, SSTN-SIM prediction estimates variation become larger than SSTS-SIM estimates. SSTN-SIM median is still smaller than SSTS-SIM. In case $n = 6$, where half of the data is training data and half is testing data, PMSE_j variation of SSTN-SIM is much larger than SSTS-SIM and MPMSE of SSTS-SIM model is smaller than MPMSE of SSTN-SIM.

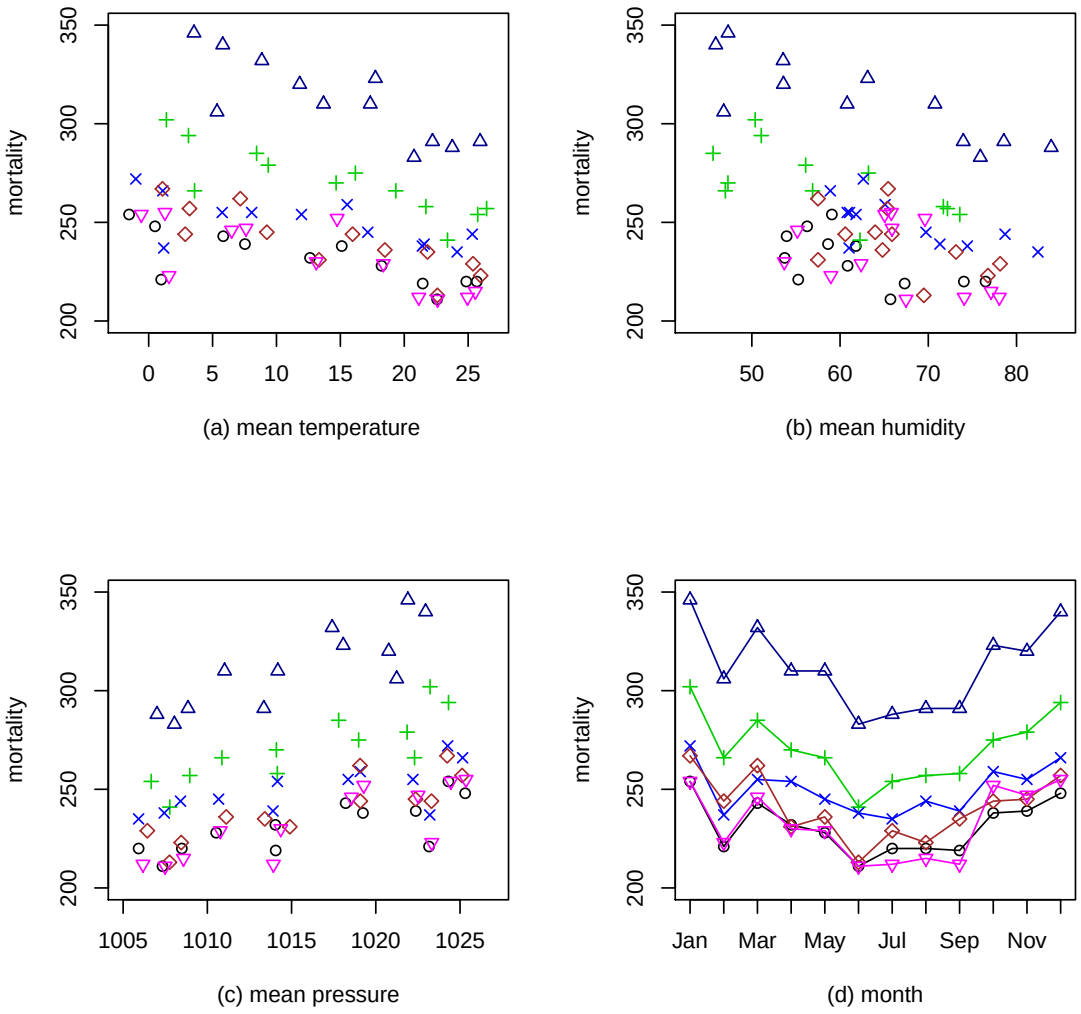


Figure 4.1: Scatterplots of non accident mortality versus the weather explanatory variables and month. △: Busan, ×: Incheon, ▽: Seoul, +: Daegu, ○: Daejeon, and ◇: Gwangju.

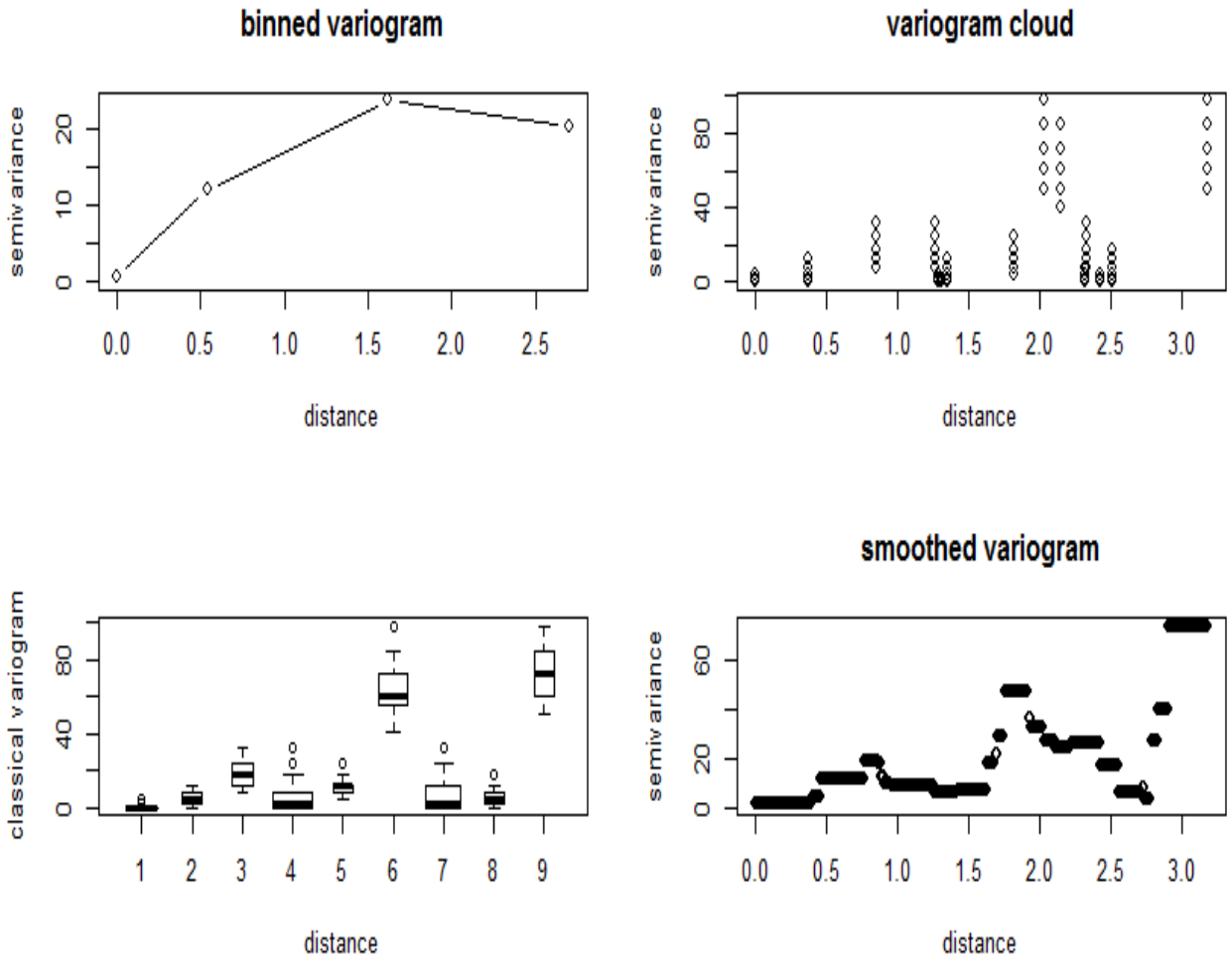


Figure 4.2: Different types of semivariograms: binned (top left), cloud (top right), cloud for binned (bottom left), and smoothed (bottom right) semivariogram

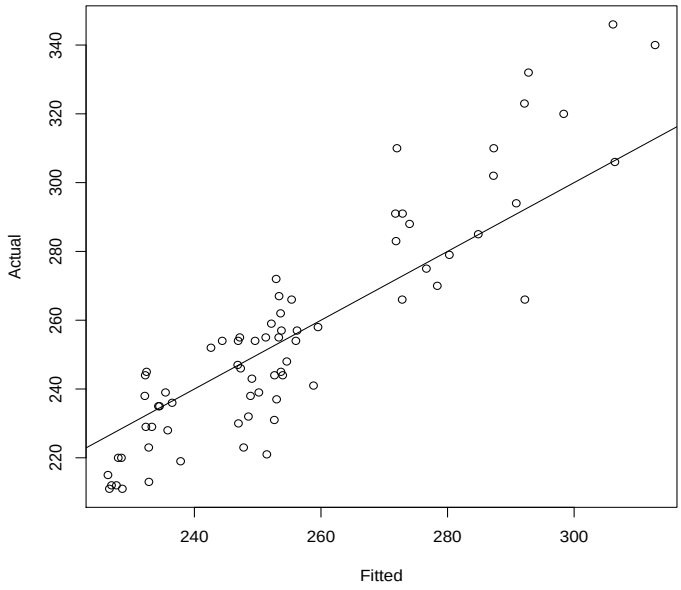


Figure 4.3: Actual versus fitted values of mean non-accident mortality for SSTS-SIM with 45^o line

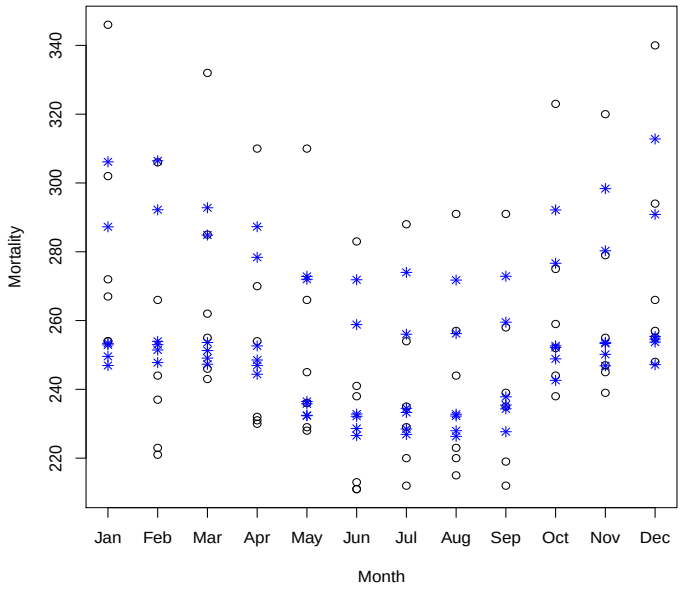


Figure 4.4: Actual values ($*$) and fitted values (o) of mean non accident mortality per month for the six cities for SSTS-SIM

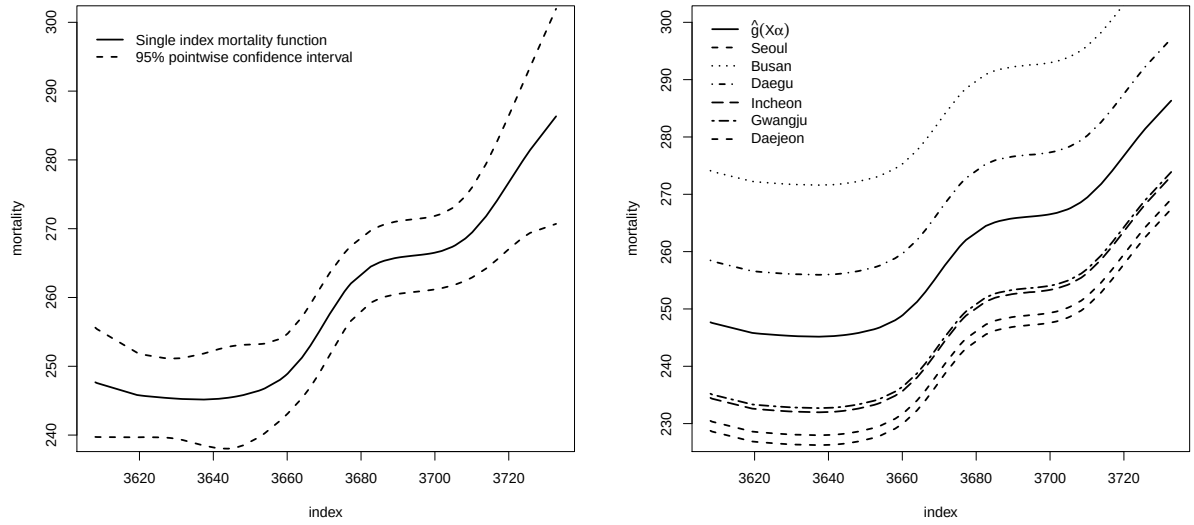


Figure 4.5: Estimated common non accident mortality of six cities of SSTS-SIM model and its 95% pointwise confidence interval (left) and Estimated Mortality functions for the six cities estimated by SSTS-SIM model (right)

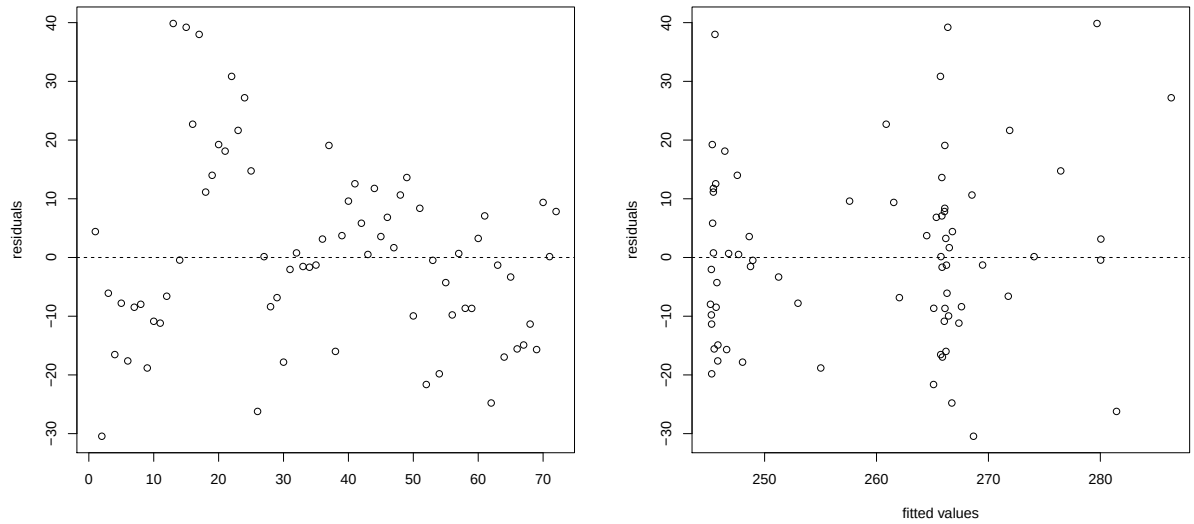


Figure 4.6: Order versus residuals plot (left) and fitted values versus residuals for SSTS-SIM (right)

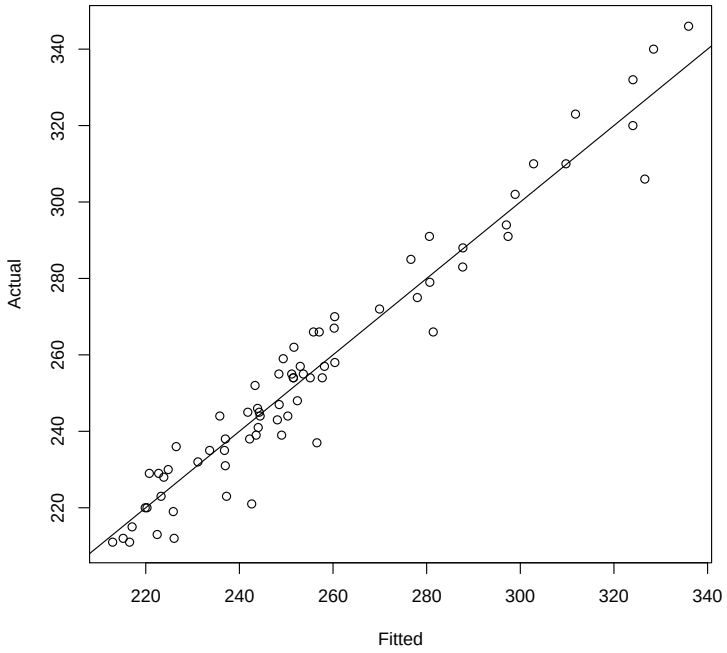


Figure 4.7: Scatter plot between actual versus fitted mean non accident mortality values using SSTN-SIM

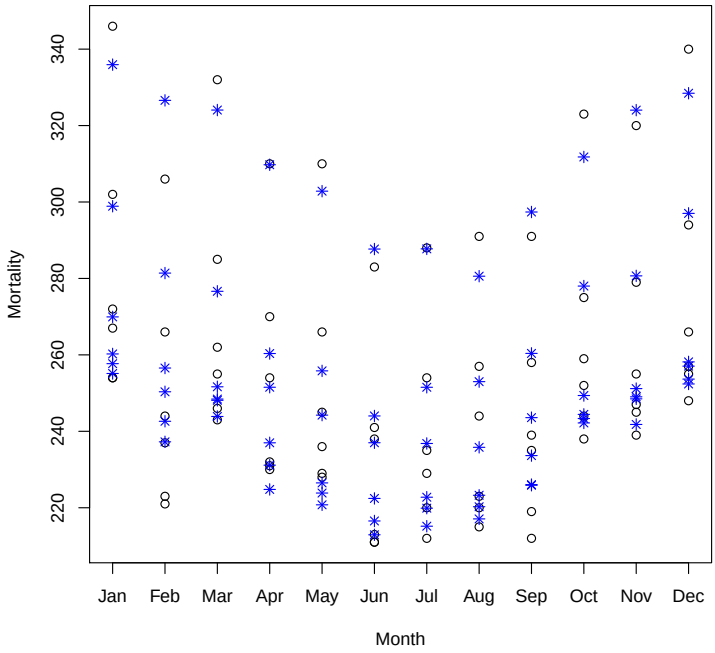


Figure 4.8: Actual ($*$) and fitted (o) mean non accident mortality per month for the six cities using SSTN-SIM

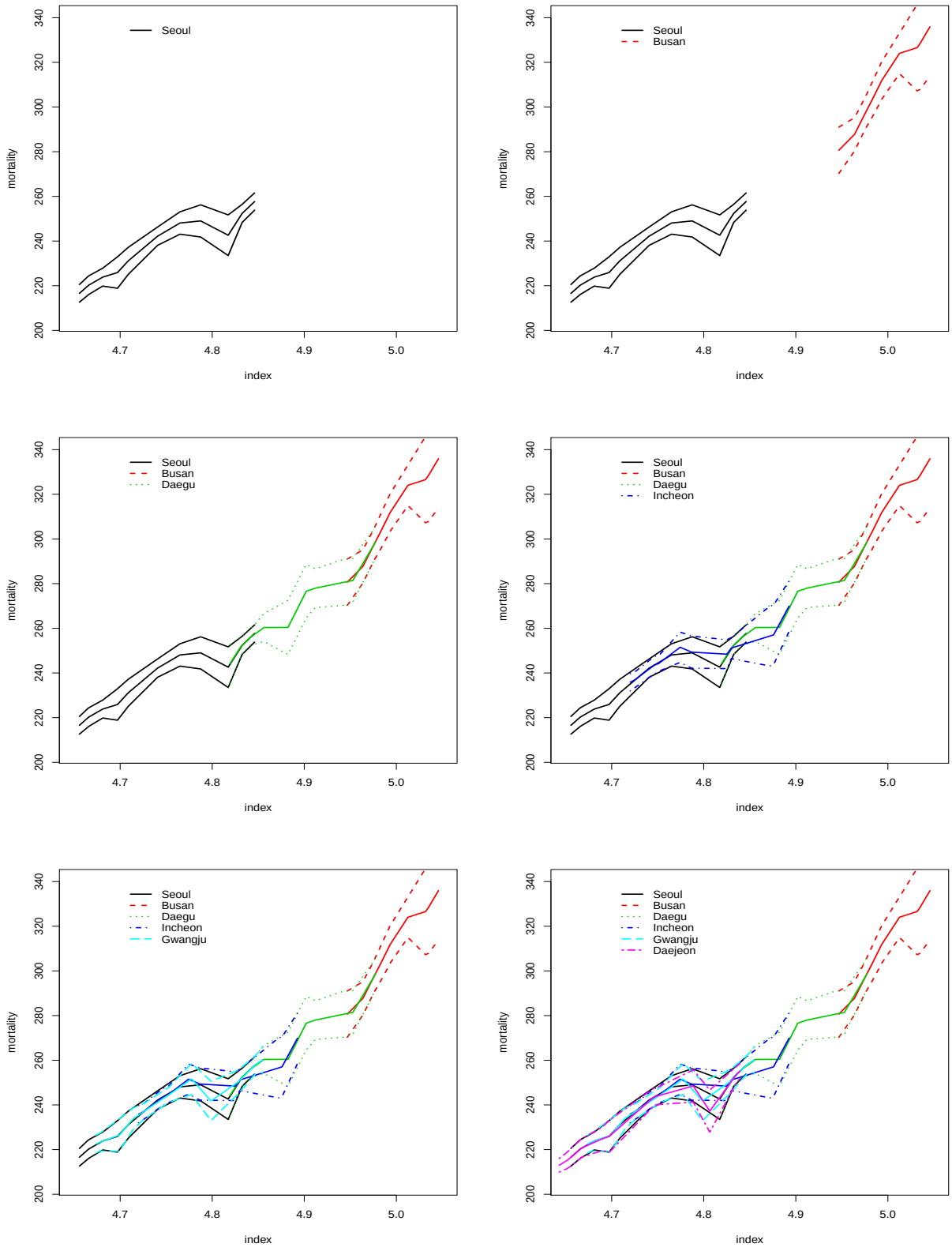


Figure 4.9: Estimated Non accident mortality functions of the six cities in South Korea and their 95% pointwise confidence intervals using SSTN-SIM.

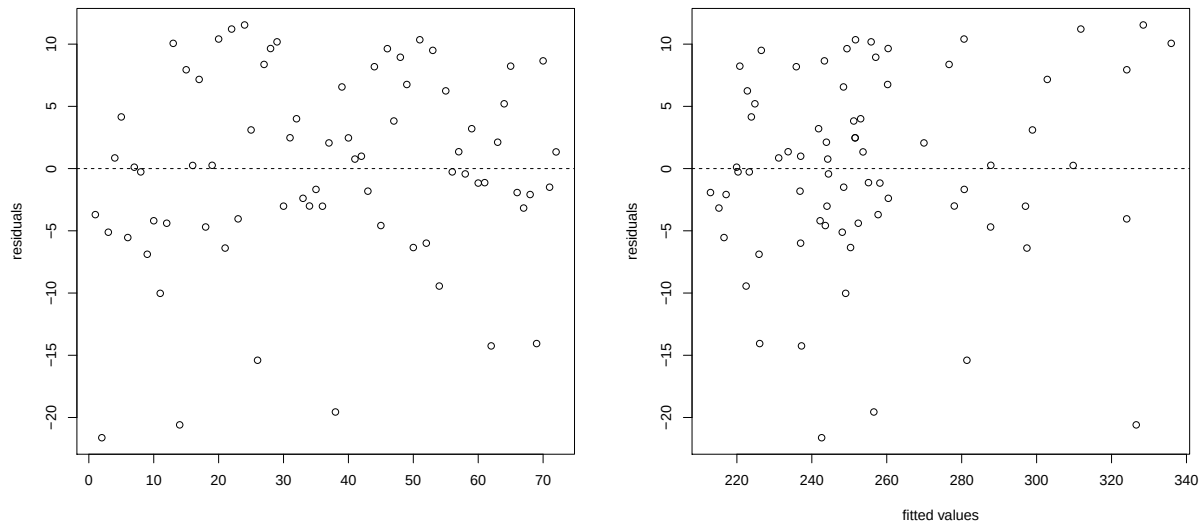


Figure 4.10: Order versus residuals plot (left) and fitted values versus residuals using SSTN-SIM (right)

Table 4.6: Several criteria (APMSE and MPMSE, PLogLE, LogLE, MSE, R^2) to compare SIM, SSS-SIM, SSN-SIM, SSTS-SIM and SSTN-SIM in estimation and prediction calculated from 500 testing and training data at different sizes for testing data (n=2, 4, 6)

	Model	MPMSE	APMSE	PLogLE	LogLE	MSE	R^2
All Data	SIM	—	—	—	-423.21	673.12	0.61
	SSS-SIM	—	—	—	-314.19	186.65	0.87
	SSN-SIM	—	—	—	-272.15	87.27	0.91
	SSTS-SIM	—	—	—	-309.23	180.19	0.87
	SSTN-SIM	—	—	—	-236.54	57.97	0.94
n=2	SSTS-SIM	227.70	253.70	-68.78	-257.20	169.20	0.88
	SSTN-SIM	184.60	385.00	-15.82	-195.70	101.00	0.90
n=4	SSTS-SIM	256.90	296.20	-122.50	-207.10	144.70	0.89
	SSTN-SIM	193.60	543.50	-74.36	-149.80	107.70	0.89
n=6	SSTS-SIM	284.20	331.10	-172.40	-158.70	130.90	0.90
	SSTN-SIM	829.30	893.30	-153.60	-104.70	131.20	0.87

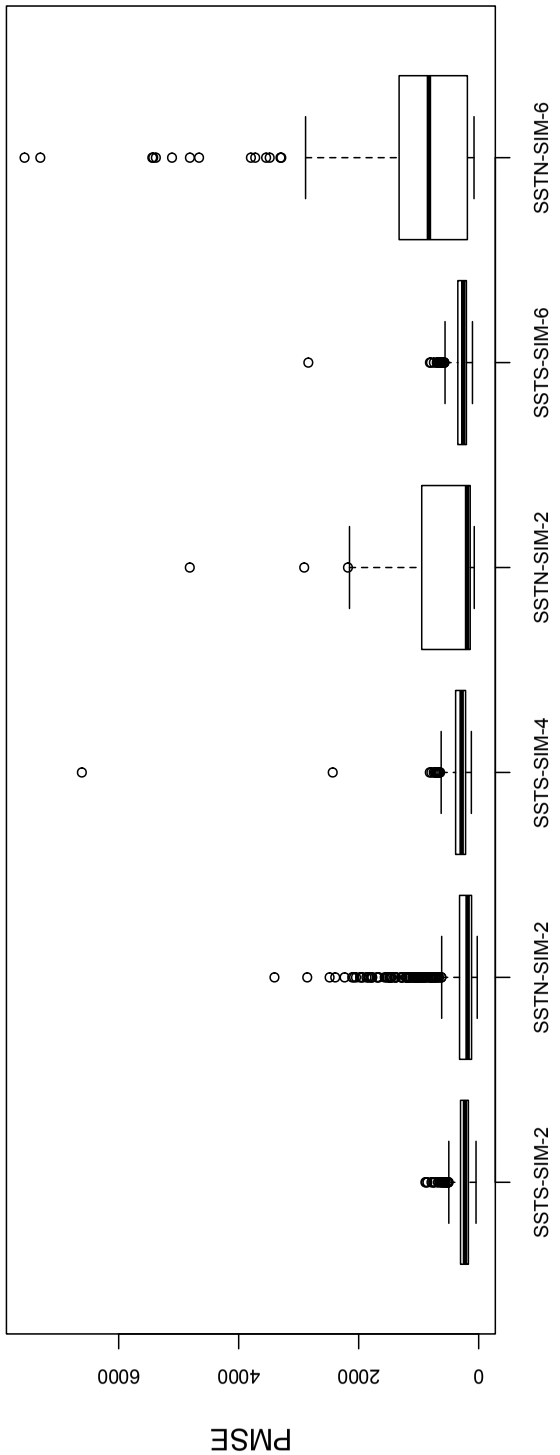


Figure 4.11: Boxplots of Prediction Mean Square Error (PMSE) of the proposed two models; SSTS-SIM=semiparametric spatio-temporal single index additive model and SSTN-SIM= semiparametric spatio-temporal single index nonaddative model at different evaluation data set sizes ($n = 2, 4, 6$)

4.6 Summary

We have proposed two models to incorporate the spatially and temporally correlated random effects into the single index model. One is that the spatial random effects and time effects are separated additively from the unknown function and the relationship between the mean response and the random effects is linear. The other is that the correlated spatial random effects are included in the unknown function and the time effect are separated additively so that the spatial random effects can not be separated from the unknown function. The relationship between the spatial random effects and the mean response is unknown. We have proposed nonseparable model because nonparametric function $m(\cdot)$ is different from that of each spatial location and time point. To the best of our knowledge, there is no such nonparametric single index model, that is, it can not be separated with spatial random effects. We showed this nonseparable model provides not only accurate parameter estimation but also better prediction accuracy. The advantage of SSTN-SIM does not need to have a restriction on the single index coefficients parameters which enables us to estimate all the parameters and the unknown function without additional constrains.

For each model we proposed an algorithm to simultaneously estimate the unknown function, the single index coefficients parameters, the variance of the spatial effects, the spatial random effects, and time effect based on Markov Chain Expectation Maximization algorithm.

The two models (SSTS-SIM, SSTN-SIM) have been applied to the South Korea data. It is found that Busan city has highest non accident mortality when we fit SSTS-SIM or SSTN-SIM and the city which has lowest mortality in both models is Daejeon city. The other cities in between have the same rank in mortality level in both models. Also, we found that the shape of mortality functions of the 6 cities are the same in SSTS-SIM, but they are different in SSTN-SIM.

To evaluate which model fits the data better among SIM, SSS-SIM, SSTSIM, SSTN-SIM, and SSTSN-SIM, we calculated R^2, LogLE and MSE. We found that SSTN-SIM is the most appropriate model for monthly South Korea mortality data. This means including spatial effect and time effect into single index model improved the model and including spatial effect in a nonseparable format is much better than separated it additively.

To address the prediction and estimation performance of SSTS-SIM and SSTN-SIM, we divided the data into two parts: training data and testing data. We calculated several estimation and prediction criteria for both models. It is found that SSTN-SIM outperforms SSTS-SIM when we have enough training data.

Chapter 5

General Conclusions and Future Research

Major conclusions and contributions of this dissertation are summarized in this chapter and possible future research areas are introduced.

5.1 Conclusions

Environmental health studies are of great interest in human research to evaluate the relationship between daily/weekly mortality and temperature. In our real data from South Korea, non accident mortality and other covariates, such as mean temperature, mean humidity, mean pressure, time, has been recorded daily during the period from January, 2000 to December, 2007 for six major cities: Seoul, Incheon, Daejeon, Daegu, Gwangju, and Busan. In total we have 2922 observations for each city.

In Chapter 2, Simulation result for investigating the performance of our proposed model in

detecting the change point(s) and compared it to the performance of the generalized linear model and additive generalized linear model suggests that the SICM is better than the other two models in terms of Type I error and power. For our real data, we found that the proposed single index change point model which simultaneously estimates the nonlinear relationship and detect change points works well to estimate the unknown mortality function and detect the change points in temperature. Using the permutation test procedure, two change points were detected in mortality one at temperature degree 15.8 °C and the other is at 23.1 °C, while the previous studies found only one change point. It is found that mortality before 15.8 °C is decreasing and after that temperature is sharp decreasing and once reach 23.1 °C, it starts to be sharp increasing.

In Chapter 3, the performance of two proposed models to incorporate the correlated spatial random effects into single index model, SSS-SIM and SSN-SIM, has been investigated and found from the simulation that they works well. SSN-SIM has two advantages over SSS-SIM; (1) in terms of estimating the spatially correlated random effects, it performs better than SSS-SIM and (2) it does not need a restriction on the single index coefficients parameters which enable us to estimate all the parameters and the unknown function without additional constraints on identifiability.

Additional simulation study was done to make sure the two models have the same performance when the unknown function is the identity function. It is found that they give almost the same estimates. The two models have been applied to the South Korea data, SSS-SIM and SSN-SIM. It is found that Busan city has highest non accident mortality when we fit SSS-SIM or SSN-SIM, however the city which has lowest mortality is different in both models: Seoul city has the lowest mortality using SSN-SIM and Daejeon using SSS-SIM. The other cities in between have the same rank in mortality level in both models. Also, we found that the shape of mortality functions of the 6 cities are the same in SSS-SIM, but

they are different in SSN-SIM. To see which model fits the data better, SSS-SIM, SSN-SIM, and SIM has been compared. We found that SSN-SIM is the most suitable model for our mortality data. This means including spatial effect in single index model improved the model and including it in nonseparable format is much better for both the prediction and fitting performance.

In Chapter 4, The two proposed models (SSTS-SIM, SSTN-SIM) have been applied to the South Korea data. It is found that Busan city has highest non accident mortality when we fit SSTS-SIM or SSTN-SIM and the city which has lowest mortality in both models is Daejeon city. The other cities in between have the same order in mortality level in both models. Also, we found that the shape of mortality functions of the 6 cities are the same in SSTS-SIM, but they are different in SSTN-SIM.

To evaluate which model fits the data better among SIM, SSS-SIM, SSN-SIM, SSTS-SIM, and SSTN-SIM, we calculated R^2, LogLE and MSE. We found that SSTN-SIM is the most appropriate model for monthly South Korea mortality data. This means including spatial effect and time effect into single index model improved the model and including spatial effect in a nonseparable format is much better than separated it additively.

To address the prediction and fitting performance for SSTS-SIM and SSTN-SIM models, We calculated several estimation and prediction criteria for both models. It is found that SSTN-SIM outperforms SSTS-SIM when we have enough training data.

5.2 Contributions

Three semiparametric research problems have been studied. In Chapter 2, we proposed a change point single index model as a development of single index model by incorporating

the change point. This model is more flexible than the common models have been used, generalized linear model and generalized additive model. The estimation method we used enables us to detect the change points and estimate the unknown function simultaneously. A permutation test procedure has been adopted to test the significance of change points. The asymptotic properties of the estimator of number of change points using the proposed permutation test has been introduced using an approximation of the unknown function and found that it is consistent. The proposed model is applied to the South Korea data set and interesting results have been found.

In Chapter 3, another development of single index model has been introduced by incorporating spatial correlated random effects into the model. We propose two models. One is to extend generalized single index additive model which can be separated with spatially correlated random effects with the single index function, $m(\cdot)$, SSS-SIM. However the other model we propose in this chapter, SSN-SIM, is the nonparametric single index model which can not be separated with spatially correlated random effect with the single index function, $m(\cdot)$. To the best of our knowledge, there is no such nonparametric single index model which can not be separated with spatial random effects. Our two models, separable and nonseparable models, were estimated via adapted Markov Chain Expectation Maximization (MCEM) algorithm. The advantage of SSN-SIM is that it does not need to have a restriction on the single index coefficients parameters which enable us to estimate all the parameters and the unknown function without additional constrains. After estimating the nonparametric function, we also provide the prediction accuracy to predict "unobserved" mortality at location s.

In Chapter 4, we have proposed two models to incorporate the spatially and temporally correlated random effects into the single index model. One is that the spatial random effects and time effects are separated additively from the unknown function and the relationship

between the mean response and the random effects is linear. The other is that the correlated spatial random effects are included in the unknown function and the time effect are separated additively so that the spatial random effects can not be separated from the unknown function. The relationship between the spatial random effects and the mean response is unknown. Also to the best of our knowledge, there is no such nonparametric single index model, that is, it can not be separated with spatial random effects with separated time effects. We showed this nonseparable model provides not only accurate parameter estimation but also better prediction accuracy. The advantage of SSTN-SIM over SSTS-SIM is that it does not need to have a restriction on the single index coefficients parameters which enable us to estimate all the parameters and the unknown function without additional constrains. We have applied the proposed models to the South Korea data and found interesting results.

5.3 Future Work

Our work in this dissertation can be extended in different ways:

1. Single index change point model is estimated using Ichimura's method and to fix the identifiability problem the first component of the single index coefficients vector is set to be 1, $\alpha_1 = 1$, but other estimation methods for the single index coefficient parameters vector, $\boldsymbol{\alpha}$, can be applied and another approach for fixing the identifiability problem can be used, such as $||\boldsymbol{\alpha}|| = 1$. That is possible to be applied for SICM, SSS-SIM, and SSTS-SIM. However, if $||\boldsymbol{\alpha}|| = 1$ is used, maybe two of the proposed models which have spatial effect not separated from the unknown function, SSN-SIM and SSTN-SIM, will not be possible to be estimated.

2. We included the change points in our proposed model, SICM, and applied this model

to only Seoul city data set but the change points can be included in the other proposed models, SSS-SIM, SSN-SIM, SSTS-SIM, and SSTN-SIM and apply to the data of all the 6 South Korea cities to see if they have a common change points or different change points in temperature. Figure 5.1 and Figure 5.2 show that there are change points in mortality for each city.

3. In SSTS-SIM and SSTN-SIM, we included the spatial and time effects but we did not include the interaction effect between time and spatial effects. Figure 5.3 shows that there is an interaction between year and location: the mortality functions of the 6 cities have different patterns along with year. Future research could focus on including interaction in these models.

4. In SSTS-SIM and SSTN-SIM, modified RW(1) and Gaussian process are used in this dissertation to model the time effect but other time series models can be used. In addition, separable covariance functions for spatial and time effects are used but non-separable function also can be used in future work.

5. In all the proposed models, SICM, SSS-SIM, SSN-SIM, SSTS-SIM and SSTN-SIM, the number of observations is greater than the number of variables. Studying those proposed models in this case is an interesting future point of research.

6. We introduced estimation methods for the proposed models, SICM, SSS-SIM, SSN-SIM, SSTS-SIM and SSTN-SIM, and it is found that the estimation methods using simulated and real data applications work well but the estimators of the parameters need further study to find the asymptotic properties of the estimators especially for the models have nonseparable components, SSN-SIM and SSTN-SIM.

7. In our dissertation, we consider a small number of spatial random effects where our data set has only 6 different locations. As a result, the covariance matrix has a small

rank but our model can be extended to a large number of spatial effects.

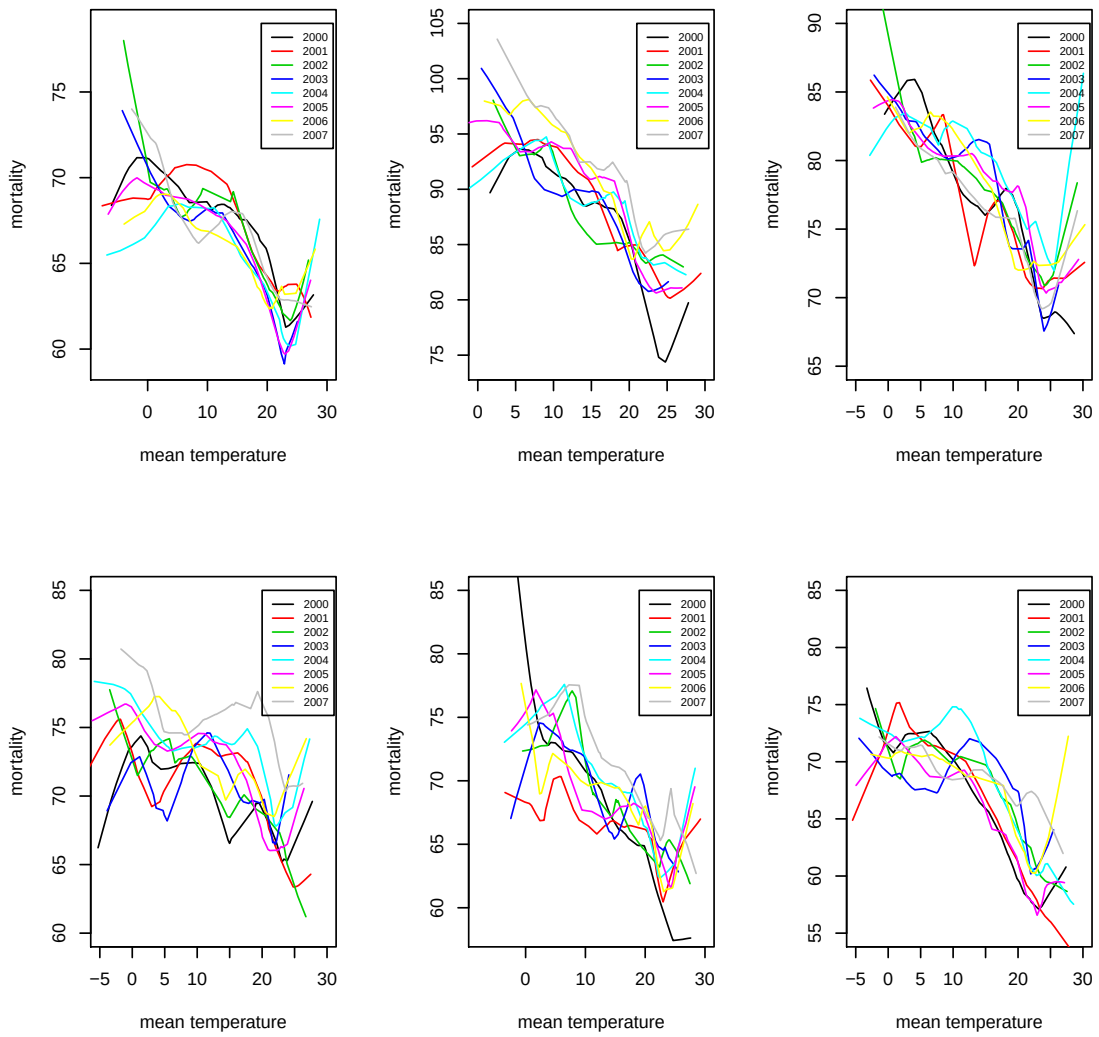


Figure 5.1: Smoothed mortality function with temperature at different years (2000-2007) of each city

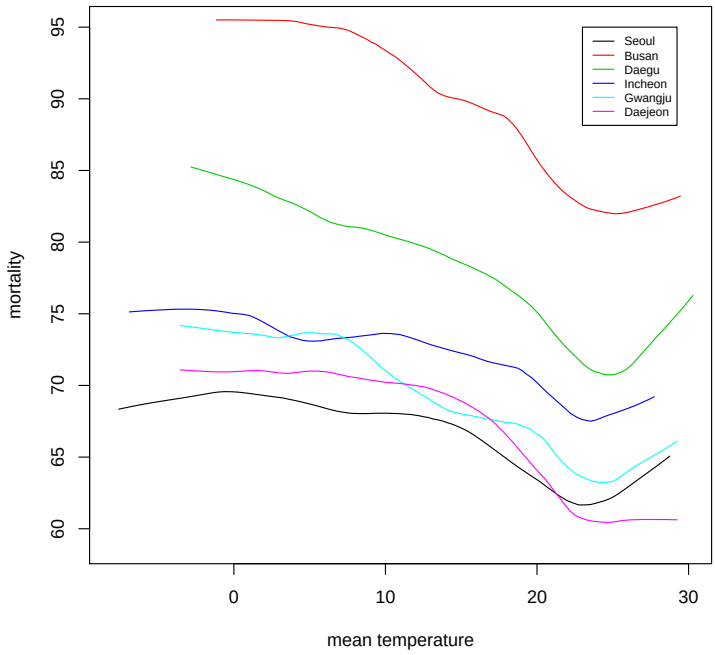


Figure 5.2: Smoothed mortality function for each city during the period from 2000 to 2007

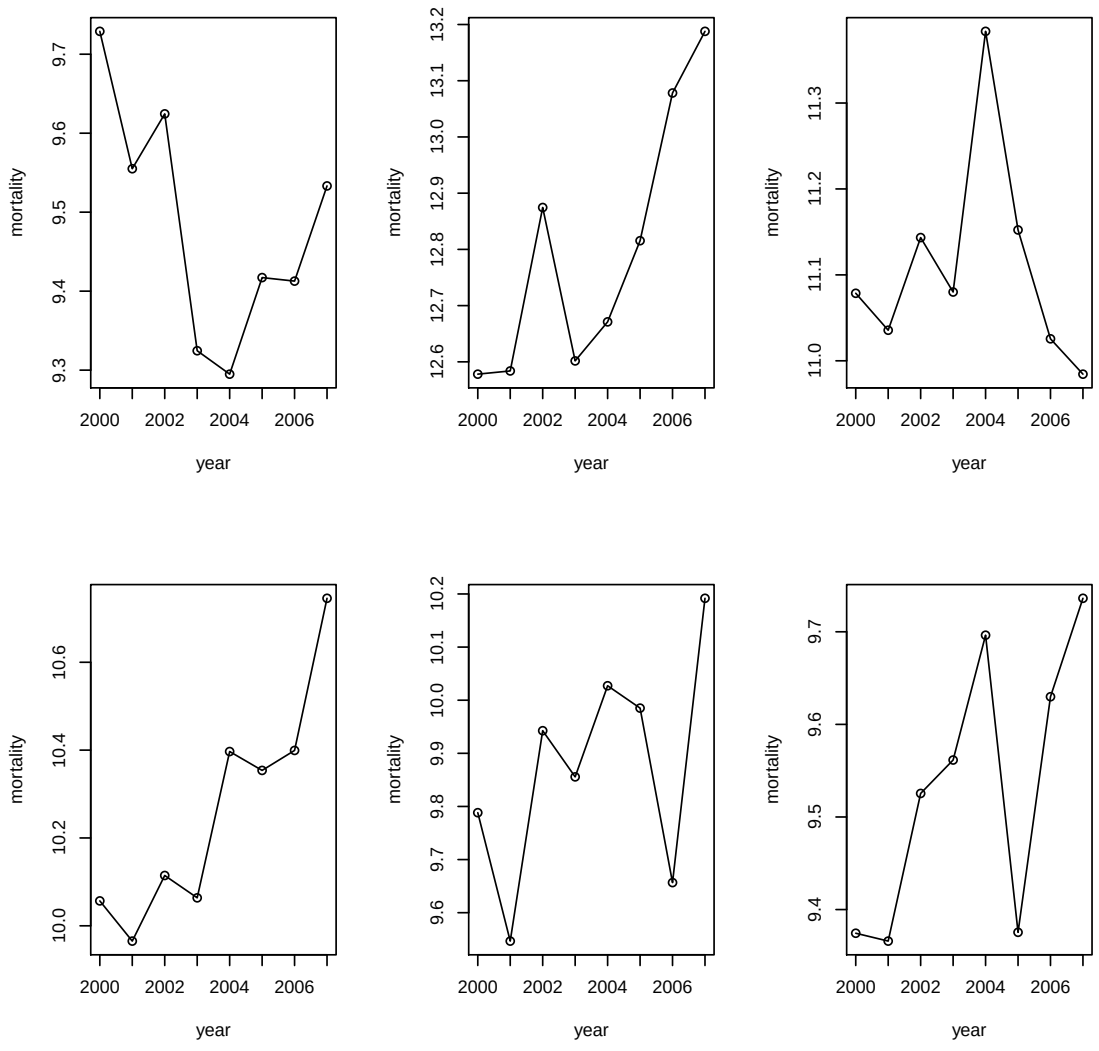


Figure 5.3: Yearly non accident mortality for each city

Bibliography

Agarwal, D., Chen, B. and Elango, P. (2009). Spatio-temporal models for estimating click-through rate. *18th International World Web Wide Conference*, 21-29.

An, X., and Bentler, P. M. (2012). Efficient direct sampling MCEM algorithm for latent variable models with binary responses. *Computational Statistics & Data Analysis*, **56**(2), 231-244.

Anselin, L. and Florax, R. J. G. M. (1995). *New Directions in Spatial Econometrics*. Berlin: Springer.

Arcuti, S., Calculli, C., Pollice, A., D'Onghia, G., Maiorano, P. and Tursi, A. (2013). Spatio-temporal modelling of zero-inflated deep-sea shrimp data by Tweedie generalized additive. *Statistica*, **73**, 103-122.

Armstrong, B. (2006). Models of the Relationship between ambient temperature and daily mortality. *Epidemiology*, **17**, 624-631.

Basu, R. and Samet, J. M. (2002). Relationship between elevated ambient temperature and mortality: A review of the Epidemiologic evidence. *Epidemiologic Review*, **24**, 190 -202.

Bilonick, R. A. (1983). Risk qualified maps of hydrogen ion concentration for the New York State area for 1966-1978. *Atmospheric Environment*, **17**, 2513-2524.

BIBLIOGRAPHY

Bilonick, R. A. (1985). The space-time distribution of sulfate deposition in the Northeast United States. *Atmospheric Environment*, **19**, 1829-1845.

Bilonick, R. A. (1985). Monthly hydrogen ion deposition maps for the Northeastern U.S. from July 1982 to September 1984. *Atmospheric Environment*, **22**, 1909-1924.

Bilonick, R. A. and Nichols, D. G. (1983). Temporal variations in acid precipitation over New York State − What the 1965-1979 USGS data reveal. *Atmospheric Environment*, **17**, 1063-1072.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo em algorithm. *Journal of the Royal Statistical Society: Series B*, **61**, 265285.

Banerjee, S., Carlin, C. P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial*. London: Chapman and Hall.

Caffo, B. S., Jank, W., and Jones, G. L. (2005). Ascent-based Monte Carlo expectation maximization. *Journal of the Royal Statistical Society: Series B*, **67**, 235251.

Chang, Z. Q., Xue, L. G., and Zhu, L. X. (2010). On asymptotically more efficient estimation of the single-index model. *Journal of Multivariate Analysis*, **101**, 1898-1901.

Chiles, J. P., and Delfiner, P. (1999). *Geostatistics*. New York: Wiley.

Chung, J., Honda, Y., Hong, Y., Pan, X., Guo, Y., and Kim, H. (2009). Ambient temperature and mortality: An international study in four capital cities East Asia. *Science of the Total Environment*, **408**, 390-396.

Clayton, D. (1996). *Generalized linear mixed models*. London: Chapman & Hall.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. New York: Wiley.

BIBLIOGRAPHY

Cressie N. A. C. (1993). *Statistics for Spatial Data, revised edition*. New York: Wiley.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: Wiley.

Cressie, N, and Hawkins, D. M. (1980). Robust estimation of the variogram, I, *Journal of the International Association for Mathematical Geology*, **12**, 115-125.

Cressie, N, and Huang, H. C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330-1340.

Cressie, N and Huang, H. C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330-1340.

Deschenes, O. and Moretti, E. (2009). Extreme weather events, mortality and migration. *Review of Economic and Statistics*, **91**(4), 659 -681.

Diggle, P. J., and Ribeiro, P. J., Jr. (2007). *Model-based Geostatistics*. New York: Springer.

El-Zein, A., Tewtel-Salem, M, and Nehme, G. (2004). A time-series analysis of mortality and air temperature in Greater Beirut. *Science of the Total Environment*, **330**, 71-80.

Genton, M. G., Butry, D. T., Gumpertz, M. L. and Prestemon, J. P. (2006). Spatio-temporal analysis of wildfire ignitions in the St Johns River Water Management District, Florida. *International Journal of Wildland Fire*, **15**, 87-97.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, **7**(4), 473-483.

Gilks, W. R., Richardsons, S., and Spoegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. New York: Chapman and Hall/CRC.

Gneiting, T. (2002). Nonseparable stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590-600.

BIBLIOGRAPHY

Guyon, X. (1995). *Random Fields on a Network*. New York: Springer.

Gu, C. and Ma, P. Optimal smoothing in nonparametric mixed-effect models. *The Annals of Statistics*, **33**, 1357-1379.

Hardle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single index models. *The Annals of Statistics*, **21**, 157-178.

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

Hashizume, M., Wagatsuma, Y., Hayashi, T., Saha, S., K., Streatfield, and Yunus, M. (2009). The effect of temperature on mortality in rural Bangladesh a population-based time series study. *International Journal of Epidemiology*, **38**, 1689-1697.

Hayn, M., Beirle, S., Hamprecht, F. A., Platt, U., Menze, B. H. and Wagner, T. (2009). Analysing spatio-temporal patterns of the global NO2-distribution retrieved from GOME satellite observations using a generalized additive model. *Atmospheric Chemistry and Physics*, **9**, 6459-6477.

Horowitz, J. L. and Hardle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, **91**, 1623-1629.

Hridtache, M., Juditski, A., and Spokoiny, V. (2001). Direct estimation of the single coefficients in a single-index model. *The Annals of Statistics*, **29**, 595-623.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. New York: Springer.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71-120.

BIBLIOGRAPHY

Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford: Oxford University Press.

Jones, R. H. and Zhang, Y. (1997). Models for continuous stationary space-time processes. In *Modeling Longitudinal and Spatially Correlated Data,* eds. T. Gregorire, D. Brillinger. P. Diggle, *et al.* Springer, New York, 289-298.

Kan, H., London, S., J., Chen, H., Song, G., Chen, G., Jiang, L., Zhao, N., Zhang, Y., and Chen, B. (2007). Diurnal temperature range and daily mortality in Shanghai, China. *Environmental Research*, **103**, 424-431.

Kanevski, M. and Maignan, M. (2004). *Analysis and Modeling of Spatial Environmental Data.*. Switzerland: EPFL Press.

Ke, C. and Wng, Y. (2001). Semiparametric nonlinear mixed effects models and their applications (with discussion). *Journal of the American Statistical Association*, **96**(456), 1272-1298.

Kim, H-J., Fay, M., Feuer, E. J., and Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, **19** , 335-351.

Kim, H-J., Yu, B., and Feuer, E. J. (2009). Selecting the number of change-points in segmented line regression. *Statistica Sinica*, **19**, 597-609 .

Klein, R. L. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, **61**, 387-421

Knorr-Held, L. (2000). Bayesian modeling of inseparable space-time variation in disease risk. *Statistics in medicine*, **19**(17-18): 2555-2567.

Landagan, E. B., and Barrios, O. Z. (2007). An estimation procedure for a spatial-temporal model, *Statistics and Probability Letters*, **77**, 401-406.

BIBLIOGRAPHY

Lekdee, K. and Ingsrisawang, L. (2013). Generalized linear mixed models with spatial random effects for spatio-temporal data: an application to dengue fever mapping. *Journal of Mathematics and Statistics*, **9**, 137-143.

Lerman, P. M.. (1980). Fitting segmented regression models by grid search. *Applied Statistics*, **29**, 77-89.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-342.

Li, B., Genton, M. G. and Sherman, M. (2007). A nonparametric assessment of properties of space-time covariance functions. *Journal of the American Statistical Association*, **102**, 736-744.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **24**(3), 673-687.

Lin, W. and Kulasekera, K. B. (2007). Identifiability of single index models and additive index models. *Biometrika*, **94**, 496-501.

Liu, H., Davidson, R. A. and Apanasovich, T. V. (2008). Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms. *Reliability Engineering and System Safety*, **93**, 875-890.

Lui, J., Wu, S., and Zidek, J. V. (1997). On segmented multivariate regression. *Statistica Sinica*, **7**, 497-525.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**, 330-335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162-170.

BIBLIOGRAPHY

McCulloch, C. E. (2003). *Generalized Linear Mixed Models. NSF-CBMS Regional Conference Series in Probability and Statistics* **7**. Ohio: IMS, Beachwood.

Nadaraya, E. A. (1964). On estimating regression. *Theory of probability and its applications*, **9**, 141-142.

Naik, P. and Tsai, C. L. (2000). Partial least squares estimator for single-index. *Journal of Royal Statistical Society, series B*, **62**, 763-771.

Nelson, T. A., Duffus, D., Robertson, C., Laberfee, K. and Feyrer, L. J. (2009). Spatial-temporal analysis of marine wildlife. *Journal of Coastal Research*, **56**, 1537-1541.

Pang, Z. and Xue, L. (2012). Estimation of the single-index models with random effects. *Computational Statistics & Data Analysis*, **56**, 1837-1853.

Possolo, A. (1991). *Spatial Statistics and Imaging*. California: IMS, Hayward.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403-1430.

Ripley, B. (1981). *Spatial Statistics*. New York: Wiley.

Robinson, P. M. (2009). Inference on nonparametrically trending time series with fractional errors. *Economic Theory*, **25**(6), 1716-1733.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge Press.

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data*. New York: Wiley.

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data*. New York: Wiley.

BIBLIOGRAPHY

Sherman, R.P. (1994). U-process in analysis of a generalized semi-parametric regression estimator. *Economic Theory*, **10**, 372-395.

Son, J., Lee, J., Anderson, G., B., and Bell, M., L. (2011). Vulnerability to temperature-related mortality in Seoul, Korea. *Environmental Research Letters*. **6**, 1-8.

Stein, M. L. (1999). *Interpolation of Spatial Data*. New York: Springer.

Stein, M. L5 (2002). Space-time covariance functions. *Journal of the American Statistical Association*, **100**, 310-321.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, **54**, 1461-1481

Stute, W. and Zhu, L. X. (2005). Nonparametric checks for single-index models. *The Annals of Statistics*, **33**, 1048-1083.

Tan, M., Tian, G-L., and Fang, H-B. (2007). An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data. *Journal of Statistical Computation and Simulation*, **77**(11), 929-943.

Tobler W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**(2), 234-240.

Tobler, W. (1979). "Cellular Geography." In S. Gale and G. Olsson, Philosophy in Geography, pp. 379-386. Dordrecht: Reidel. Tukey, J.W., 1977. Exploratory Data Analysis. Reading: Addison-Wesley.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Wang, J. L., Xue, L. G., Zhu, L. X., and Chong, Y. S. (2010). Extension for a partial-linear single-index model. *The Annals of Statistics*, **38**, 246-274.

BIBLIOGRAPHY

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: Series A*, **26**, 359-372.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, **22**, 1112-1137.

Xia, Y., Li, W. K., Tong, H., and Zhang, D. (2004). A goodness-of-fit for single index models. *Statistica Sinica*, **14**, 1-39.

Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, **94**, 1275-1285.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society, series B*, **64**, 363-410.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, **58**, 129-136

Zhu, L. X. and Ng, K. W. (1995). Asymptotics for sliced inverse regression. *Statistica Sinica*, **5**, 727-736.

Zhu, L. X., and Xue, L. G. (2006). Empirical likelihood confidence regions in a partially linear single-index-model. *Journal of the Royal Statistical Society: Series B*, **68**, 549-570.

Zhu, L. P. and Zhu, L. X. (2009a). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *Journal of Multivariate Analysis*, **100**, 862-875.

Zhu, L. P. and Zhu, L. X. (2009b). On distribution weighted partial least squares with diverging number of highly correlated predictors. *Journal of the Royal Statistical Society: Series B*, **71**, 525 -548.

Appendix A

Technical Report

To evaluate the asymptotic properties of the permutation test in single index case, we consider the local linear approximation of the function $g(X\boldsymbol{\alpha})$. Since $g(X\boldsymbol{\alpha})$ is a smooth function, it can be approximated and written as follows:

$$y \approx g(X\boldsymbol{\alpha}_0) + \acute{g}(X\boldsymbol{\alpha}_0)[X\boldsymbol{\alpha} - X\boldsymbol{\alpha}_0] + O\|X\boldsymbol{\alpha} - X\boldsymbol{\alpha}_0\|^2 + \boldsymbol{\epsilon}$$

where $\acute{g}$ is the estimated derivative of the function g and $\boldsymbol{\alpha}_0$.

When $n \longrightarrow \infty$ and $\boldsymbol{\alpha} \longrightarrow \boldsymbol{\alpha}_0$,

$$\begin{aligned} y &\approx c + \acute{g}(X\boldsymbol{\alpha}_0)X\boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ &= c + A\boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ &= H\boldsymbol{\alpha} + \boldsymbol{\epsilon} \end{aligned} \tag{1}$$

where c is a vector of constant, A is $n \times n$ matrix, and H is a $n \times (p + K)$ matrix. K is number of change points and p is number of coefficients parameters. Hence our model matrix

is H. For each t=1,2,.....,n, we have $y_t = H_t\boldsymbol{\alpha} + \boldsymbol{\epsilon_t}$.