Extensions of small area models with applications to the National Resources Inventory

by

Pushpal Mukhopadhyay

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Tapabrata Maiti, Major Professor
Wayne A. Fuller
Sarah M. Nusser
Soumendra N. Lahiri
Leslie Miller

Iowa State University

Ames, Iowa

2006

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Pushpal Mukhopadhyay

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

To my parents

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER 1. Introduction 1

1.1 Small Area Estimation 1

1.1.1 Small Area Models 2

1.1.2 Small Area Predictions Using EBULP 4

1.1.3 MSE of EBLUP and an Estimator of the MSE 5

1.2 Kernel Regression and Local Polynomial Regression 7

1.2.1 Nonparametric Fixed Effects Model 8

1.2.2 Nadaraya-Watson Estimator 8

1.2.3 Local Polynomial Estimator 10

1.2.4 Bandwidth Selection for Local Estimators 12

1.3 Dissertation Organization 15

CHAPTER 2. Small Area Estimation For A Nonlinear Transformation 16

2.1 Introduction 16

2.2 The NRI Survey 17

2.3 Variables of Interest for Wind Erosion 18

2.4 Exploratory Analysis 19

2.5 A Regression Based Calibrated Small Area Estimator 27

2.6 Conclusions 38

CHAPTER 3. Small Area Estimation: A Nonparametric Approach . 40

3.1 Introduction 40

3.2 Kernel-Based Approach 41

3.2.1 Approximation of Mean Squared Error 44

3.3 Simulation for the Nadaraya-Watson Estimator 45

3.3.1 Simulation Results 47

3.4 Application to the NRI 48

3.4.1 Estimates for Wind Erosion 52

3.5 Conclusions 53

CHAPTER 4. Local Polynomial Regression 56

4.1 Introduction 56

4.2 Framework for Local Polynomial Estimators 57

4.3 Theory for Local Polynomial Estimators 61

4.3.1 Exact Bias and Variance 62

4.3.2 Asymptotics for Local Polynomial Estimators 63

4.4 Approximation of Mean Squared Error 65

4.4.1 Bandwidth Selection for Local Polynomial Estimators 66

4.5 Conclusions 67

CHAPTER 5. Small Area Estimation Using Imputed Values 68

5.1 Introduction 68

5.2 The NRI Survey 69

5.2.1 Variables of Interest for Soil Erosion 70

5.2.2 Imputation Procedure for the C Factor 71

5.3 Estimator of the Mean C Factor 71

5.3.1 Multivariate Small Area Model 72

5.3.2 Estimator of the Covariance 74
5.4 Estimates for the C Factor 2002 77
5.4.1 Imputation Model 77
5.4.2 Small Area Model and County Level Estimates 87
5.5 Conclusions 92
CHAPTER 6. Summary 95
APPENDIX A. Proofs of Chapter 3 103
APPENDIX B. Proofs of Chapter 4 108
APPENDIX C. Proof of Chapter 5 119
BIBLIOGRAPHY 122
ACKNOWLEDGEMENTS 126

LIST OF TABLES

Table 2.1 Summary Statistics for Survey Weighted Mean WEQ02 and the Design Standard Error for Iowa Counties 20

Table 2.2 Parameter Estimates for the Small Area Models 22

Table 2.3 Summary Statistics for Estimated County Means 32

Table 2.4 County Estimates for WEQ02 34

Table 2.5 Summary Statistics for Estimated RMSEP 36

Table 2.6 Predicted Means and RMSEP for Eight Selected Counties 39

Table 3.1 Predictions for Linear Populations 48

Table 3.2 Predictions for Cubic Populations 49

Table 3.3 Predictions for Exponential Populations 49

Table 3.4 Predictions for Mixed-Exponential Populations 50

Table 3.5 Summary Statistics for Observed Counties 51

Table 3.6 Summary Statistics for County Means and the Estimated MSE 53

Table 5.1 Summary Statistics for C Factor 2002 78

Table 5.2 Parameter Estimates and Standard Errors for Imputation Models 79

Table 5.3 Summary Statistics for Fitted Values and the NRI Imputed Values 79

Table 5.4 Estimated Variances for 99 Iowa Counties 80

Table 5.5 Summary Statistics for the Correlations of County Means 86

Table 5.6 Summary Statistics for Square Root of Estimated Design Variances and Estimated Coefficient of Variation of County Means . 87

Table 5.7 Three Counties with High Values for the Variance Ratio for Two Phase Estimator 87

Table 5.8 Estimates for Regression Parameters and the Between Area Variance Parameter 89

Table 5.9 Summary Statistics for the Predicted County Means 91

Table 5.10 Summary Statistics for the Root Mean Square Error of Prediction 92

LIST OF FIGURES

Figure 2.1 Supplemented Panel Design for the NRI. 18

Figure 2.2 Scatter Plot and Residual Plot from Model I. 22

Figure 2.3 Scatter Plot and Residual Plot from Model II. 23

Figure 2.4 Scatter Plot and Residual Plot from Model III. 23

Figure 2.5 Estimated Design Variances for Transformed Estimates. 25

Figure 2.6 Normal Quantiles Plot for Model III. 26

Figure 2.7 Predicted Means from Three Small Area Models. 33

Figure 2.8 Ratio of the Standard Error of Design Weighted Mean to the Root Mean Square Error of Prediction. 37

Figure 3.1 Scatter Plot for Simulated Populations. 46

Figure 3.2 Scatter Plot of WEQ 2003 and Erodibility Index. 51

Figure 3.3 Direct County Means for WEQ 2003. 54

Figure 3.4 Estimates for County Means using Fay-Herriot Model. 54

Figure 3.5 Estimates for County Means using Non-Parametric Model. . . . 55

Figure 5.1 Ratio of the Estimated Two Phase Variance to the Estimated Variance using Observed Data. 88

Figure 5.2 Plot of Predicted C Factors. 90

Figure 5.3 Root Mean Square Error of Prediction. 91

Figure 5.4 Scatter Plot of Missing Rate and the Ratio of RMSEP. 93

CHAPTER 1. Introduction

1.1 Small Area Estimation

A sample survey is a cost effective way to draw inferences from a target population. Surveys are used in practice to provide estimates not only for the total population but also for a variety of subpopulations. The term "Small Area" refers to a subpopulation, or domain, where the domain sample size is not large enough to support direct sample based estimates with adequate precision. Sometimes a sampling fraction that is larger than the average sampling fraction is used in some domains in order to increase the precision of domain estimates. Frequently, domains are not defined during the design stage and, hence, oversampling is not possible. Even when the domains of interest are known beforehand, it is not always possible to have a large enough, overall sample size to support reliable direct estimates.

Domain estimators (or direct domain estimators) rely solely on domain specific sample data and may use known auxiliary information. On the other hand, small area estimators "borrow strength" by using information across similar domains. Small area estimators may use known auxiliary information as well. Small area estimation techniques are divided into two major types: traditional indirect estimators and model based estimators. Traditional indirect estimators use implicit linking models. Traditional indirect estimators are generally design biased and their design variances are usually small compared to the design variances of the direct domain estimators. For a finite population $\mathcal{F}$ the design bias and the design variance of an estimator $\hat{\theta}$ of θ are defined by $E[\hat{\theta}-\theta|\mathcal{F}]$

and $V[\hat{\theta}|\mathcal{F}]$ respectively. The design bias of indirect estimators does not decreases as the overall sample size increases (Rao, 2003). Examples of indirect estimators include synthetic and composite estimators. See Chapter 4, Rao (2003) for detailed descriptions of indirect estimators and their properties.

Model based small area estimation techniques are generally accepted in the small area literature. Model based estimators use explicit linking models. The model based mean squared errors (the mean squared error under the model assumptions) are usually smaller than the design variances of the direct domain estimators. One major advantage of the model based technique is the ability to validate the explicit model from the sample data. In this work, we focus on explicit linking models and model based techniques for small area estimation.

1.1.1 Small Area Models

Small area models provide explicit linking of related small areas through supplementary data. Most commonly used small area models have a sampling error component and an area specific random error component. Thus small area models can be thought of as generalized linear mixed effects models. Depending on the availability of the auxiliary information, the small area models can be divided into two basic types, area level models and unit level models. Area level models relate area level mean responses to area level auxiliary information and unit level models relate the unit values of the study variable to unit level auxiliary information.

The area level small area model was first used in the survey setting by Fay and Herriot (1979) in the context of estimating per capita income for small places in the US. Let $\bar{y}_{i.}$ be the survey weighted mean for the small area i, where $i = 1, 2, ..., m$. The area level small area model using area level covariates $\mathbf{x}_i$ can be written as

$$\bar{y}_{i.} = \theta_i + \epsilon_i, \text{ and} \tag{1.1}$$

$$\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + u_i, \tag{1.2}$$

where θ_i's are the "true" means, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a set of area level parameters, ϵ_i's are sampling errors and u_i's are area specific random errors. The sampling errors ϵ_i are assumed to be independent with $E[\epsilon_i|\mathcal{F}] = 0$ and $\text{Var}[\epsilon_i|\mathcal{F}] = \psi_i$. The area specific random effects are assumed to be independent and identically distributed with $E[u_i] = 0$ and $V[u_i] = \sigma_u^2$. We denote these assumptions as $\epsilon_i \overset{ind}{\sim} (0, \psi_i)$ and $u_i \overset{iid}{\sim} (0, \sigma_u^2)$ respectively. We also assume that ϵ_i and u_i are independent. It is common to assume the normality of both error components. It is also customary to assume that ψ_i's are known.

Battese et al. (1988) used a unit level small area model to estimate county crop areas using survey and satellite data. Unit level models assume that unit specific auxiliary information $\mathbf{x}_{ij}$'s are available for each population unit j in each small area i. A unit level small area model using unit level covariates $\mathbf{x}_{ij}$ can be written as a nested error regression model of the form

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + \epsilon_{ij}, \tag{1.3}$$

where y_{ij} is the response variable, $j = 1, 2, ..., N_i$, $i = 1, 2, ..., m$, u_i's are area specific random effects, $\boldsymbol{\beta}$ is a set of fixed parameters, N_i is the number of population units in the area i and m is the number of small areas. Let $\bar{\mathbf{x}}_{i.}$ be the mean $\mathbf{x}_{ij}$'s for the area i. The small area means can be written as

$$\theta_i = \bar{\mathbf{x}}_{i.}^T\boldsymbol{\beta} + u_i, \tag{1.4}$$

provided the population size N_i is large for every i. Typically, ϵ_{ij} and u_i are mutually independent with $\epsilon_{ij} \overset{iid}{\sim} (0, \sigma_e^2)$ and $u_i \overset{iid}{\sim} (0, \sigma_u^2)$. The normality of the error components are also commonly assumed. We further assume that a sample of size n_i is drawn from N_i units in area i using simple random sampling or using the auxiliary information $\mathbf{x}_{ij}$.

The restriction in the sample selection strategy guarantees that the sampled values also follow model (1.3). See Rao (2003).

Linear mixed effects models can be written as

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\epsilon}, \tag{1.5}$$

where $\mathbf{y}$ is a vector of responses, X and Z are known design matrices, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a set of fixed effects parameters, $\mathbf{u}$ is a set of random effects and $\boldsymbol{\epsilon}$ is a set of random errors. The $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are assumed to be independently distributed with $\mathbf{u} \sim (0, G)$ and $\boldsymbol{\epsilon} \sim (0, R)$, where G and R are positive definite covariance matrices. See Searle et al. (1992) for further details. Model (1.2) is a special case of the linear mixed model (1.5) with $\mathbf{y} = (y_1, ..., y_m)^T$, $X = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_m^T)^T$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$, $Z = I_m$, $\mathbf{u} = (u_1, ..., u_m)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_m)^T$, $R = \text{diag}\{\psi_i\}$ and $G = \sigma_u^2 I_m$ where I_m is the identity matrix of dimension m and $\text{diag}\{a_i\}$ is a diagonal matrix with a_i as the i^{th} diagonal element. Model (1.4) is also a special case of model (1.5) with $\mathbf{y} = (y_{11}, ..., y_{1N_1}, ..., y_{m1}, ..., y_{mN_m})$, $Z = \text{blockdiag}(\mathbf{1}_{N_i})$, $\mathbf{u} = (u_1, ..., u_m)^T$, $\boldsymbol{\epsilon} = (\epsilon_{11}, ..., \epsilon_{mN_m})^T$, $G = \sigma_u^2 I_m$ and $R = \sigma_e^2 I_N$ where $\mathbf{1}_{N_i}$ is a column vector of ones of length N_i, $N = \sum_{i=1}^m N_i$ and blockdiag denotes a block diagonal matrix. In this work we focus on the area level small area models.

1.1.2 Small Area Predictions Using EBULP

Let $\mu = \mathbf{l}^T\boldsymbol{\beta} + \mathbf{b}^T\mathbf{u}$ be any linear combination of the regression parameter $\boldsymbol{\beta}$ and the realization of the random component $\mathbf{u}$. For known covariance matrices G and R , Henderson (1950) proposed the best linear unbiased predictor (BLUP) of μ as

$$\tilde{\mu} = \mathbf{l}^T\tilde{\boldsymbol{\beta}} + \mathbf{b}^T\tilde{\mathbf{u}}, \tag{1.6}$$

where

$$\tilde{\boldsymbol{\beta}} = (X^T V^{-1} X)^{-1} X^T V^{-1} \mathbf{y}, \tag{1.7}$$

$$\tilde{\mathbf{u}} = GZ^T V^{-1}(\mathbf{y} - X\boldsymbol{\beta}), \tag{1.8}$$

and $V = \text{Var}[\mathbf{y}] = R + G$. If we assume normality of the error components $\boldsymbol{\epsilon}$ and $\mathbf{u}$ then an estimator for $\boldsymbol{\beta}$ and $\mathbf{u}$ can be obtained by solving the Henderson equations

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} \mathbf{y} \\ Z^T R^{-1} \mathbf{y} \end{bmatrix}. \tag{1.9}$$

The solution of the Henderson equations (1.9) is identical to the components of the BLUP (1.6) of μ. Since G and R are not known, we replace G and R in (1.6) with their estimated values. The empirical best linear unbiased predictor (EBLUP) of μ is given by

$$\hat{\mu} = \mathbf{l}^T (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} \mathbf{y} + \mathbf{b}^T \hat{G} Z^T \hat{V}^{-1} (\mathbf{y} - X\hat{\boldsymbol{\beta}}), \tag{1.10}$$

where $\hat{V}$ and $\hat{G}$ are estimated covariance matrices and $\hat{\boldsymbol{\beta}} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} \mathbf{y}$. In particular, for the area level small area model (1.2), the EBLUP of the "true" mean θ_i is given by

$$\hat{\theta}_i = \hat{\gamma}_i \bar{y}_{i\cdot} + (1 - \hat{\gamma}_i) \mathbf{x}_i^T \hat{\boldsymbol{\beta}}, \tag{1.11}$$

where $\gamma_i = (\sigma_u^2 + \psi_i)^{-1} \sigma_u^2$, $\hat{\gamma}_i = (\hat{\sigma}_u^2 + \psi_i)^{-1} \hat{\sigma}_u^2$ and $\mathbf{x}_i$ is the covariate information available for the area i.

1.1.3 MSE of EBLUP and an Estimator of the MSE

The mean squared error (MSE) of $\hat{\mu}$ is $E(\hat{\mu} - \mu)^2$. Assuming normality of the error components $\mathbf{u}$ and $\boldsymbol{\epsilon}$, Prasad and Rao (1990) proposed a second order approximation for the MSE of $\hat{\theta}$. For the area level model (1.2), Prasad and Rao (1990) proposed

$$\text{MSE}(\hat{\theta}_i) = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) + O(m^{-2}), \tag{1.12}$$

where

$$g_{1i}(\sigma_u^2) = (\sigma_u^2 + \psi_i)^{-1} \psi_i \sigma_u^2, \tag{1.13}$$

$$g_{2i}(\sigma_u^2) = (1 - \gamma_i)^2 \mathbf{x}_i^T (\sum \mathbf{x}_i^T (\sigma_u^2 + \psi_i)^{-1} \mathbf{x}_i)^{-1} \mathbf{x}_i, \tag{1.14}$$

$$g_{3i}(\sigma_u^2) = (\sigma_u^2 + \psi_i)^{-3}\psi_i^2 V(\hat{\sigma}_u^2), \tag{1.15}$$

$V(\hat{\sigma}_u^2)$ is the variance of $\hat{\sigma}_u^2$ and γ_i is defined in (1.11). Let $\{a_n\}$ be a sequence of real numbers and $\{b_n\}$ be a sequence of positive real numbers. We write $a_n = O(b_n)$ if there exists a positive real number M such that $b_n^{-1}|a_n| < M$ for all n. Assuming the regularity conditions

1) ψ_i's are bounded, and

2) $\sup_i \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i = O(m^{-1})$, the first term $g_{1i}(\sigma_u^2) = O(1)$ and the second term $g_{2i}(\sigma_u^2) = O(m^{-1})$. Assuming the normality of the error components $\mathbf{u}$ and $\boldsymbol{\epsilon}$, the third term $g_{3i}(\sigma_u^2) = O(m^{-1})$ and the neglected terms in (1.12) are of order $O(m^{-2})$.

The between area variance parameter σ_u^2 can be estimated using the method of moments equation, the maximum likelihood (ML) equation, or the residual maximum likelihood equation. A simple method of moments estimator of σ_u^2 is

$$\hat{\sigma}_{u,MM}^2 = \max\{(m-p)^{-1}[(\bar{y}_{i\cdot} - \mathbf{x}_i^T\boldsymbol{\beta})^2 - \sum_{i=1}^{m}\psi_i(1 - \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i)], 0\}, \tag{1.16}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_2, \ldots, \beta_p)^T$, and $X = (\mathbf{x}_1^T, \ldots, \mathbf{x}_m^T)^T$. Assuming normality of $\mathbf{u}$ and $\boldsymbol{\epsilon}$ the ML estimator of σ_u^2 can be obtained by solving

$$\sigma_u^{2(a+1)} = \sigma_u^{2(a)} + [\mathcal{I}(\sigma_u^{2(a)})]^{-1}h(\tilde{\boldsymbol{\beta}}, \sigma_u^{2(a)}) \tag{1.17}$$

iteratively where

$$\mathcal{I}(\sigma_u^2) = \sum_{i=1}^{m}\{2(\psi_i + \sigma_u^2)^2\}^{-1}, \tag{1.18}$$

$$h(\tilde{\boldsymbol{\beta}}, \sigma_u^{2(a)}) = -\sum_{i=1}^{m}\{2(\psi_i + \sigma_u^2)\}^{-1} + 2^{-1}\sum_{i=1}^{m}\{(\psi_i + \sigma_u^2)^2\}^{-2}(\bar{y}_{i\cdot} - \mathbf{x}_i^T\tilde{\beta})^2, \tag{1.19}$$

and $\sigma_u^{2(a)}$ denotes the value of σ_u^2 after the a^{th} iteration. Similarly the REML estimator of σ_u^2 can be obtained from

$$\sigma_u^{2(a+1)} = \sigma_u^{2(a)} + [\mathcal{I}_R(\sigma_u^{2(a)})]^{-1}h_R(\sigma_u^{2(a)}), \tag{1.20}$$

iteratively where

$$\mathcal{I}_R = 2^{-1}\text{tr}[P^2], \tag{1.21}$$

$$S_R(\sigma_u^2) = -2^{-1}\text{tr}[P] + 2^{-1}[\mathbf{y}^T P^2 \mathbf{y}], \tag{1.22}$$

$P = V^{-1} - V^{-1}X[X^T V^{-1} X]^{-1} X^T V^{-1}$, $P^2 = PP$, tr$[A]$ is the trace of a square matrix A and a is the iteration number.

For the area level model (1.2), Prasad and Rao (1990) proposed an estimator of the MSE (1.12) by

$$\text{mse}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2), \tag{1.23}$$

where $\hat{\sigma}_u^2$ is the method of moment (1.16) estimator or the REML (1.20) estimator of σ_u^2. If the ML estimator (1.17) of σ_u^2 is used then an estimator of the MSE is

$$\text{mse}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2) - b(\hat{\sigma}_u^2)\nabla g_{1i}(\hat{\sigma}_u^2), \tag{1.24}$$

where $\nabla g_{1i}(\sigma_u^2) = (1-\gamma_i)^2$,

$$b(\sigma_u^2) = -[2\mathcal{I}(\sigma_u^2)]^{-1}\text{tr}[\{\sum_{i=1}^{m} \mathbf{x}_i^T(\psi_i - \sigma_u^2)^{-1}\mathbf{x}_i\}^{-1}\{\mathbf{x}_i^T(\psi_i - \sigma_u^2)^{-2}\mathbf{x}_i\}], \tag{1.25}$$

and $\mathcal{I}(\sigma_u^2)$ is defined in (1.18). See Chapter 7 Rao (2003) and references cited there.

1.2 Kernel Regression and Local Polynomial Regression

A nonparametric approach has significant advantages over its parametric counterpart. Erroneous specification of the parametric model can result in a biased estimator. Despite several advantages of the nonparametric model, there is no substantial use of this technique in small area estimation. This is largely due to the difficulties in incorporating nonparametric mixed effects models into the estimation tools used by survey statisticians.

There are several ways to use nonparametric smoothing. Splines, orthogonal series expansion, and local modeling are the most common nonparametric smoothing techniques. For any smoothers we assume a smooth mean curve. Splines allow possible

discontinuities of the derivative of the mean curve. The locations of the discontinuity points are called knots. The orthogonal series expansion method expands the mean curve by using orthogonal basis decompositions. A few useful subsets of basis functions are then chosen to approximate the mean curve. Local modeling selects a local neighborhood for any given point and fits a polynomial using data near that point. In particular, for local linear modeling, we solve many linear regression problems. The size of the local neighborhood is called the bandwidth. Bandwidth can be chosen by minimizing an objective function, or using the data. We have considered only local modeling as smoothers.

1.2.1 Nonparametric Fixed Effects Model

Let (x_i, y_i), $i = 1, 2, ..., n$, be a set of observations. Assume the model

$$y_i = m(x_i) + \epsilon_i \tag{1.26}$$

where $m(x)$ is a smooth function of the covariate x and is called the mean function and ϵ_i's are random errors. Homoscedastic nonparametric models assume $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$ and heteroscedastic nonparametric models assume $\epsilon_i \overset{ind}{\sim} (0, v(x))$ where σ^2 (> 0) is a constant parameter and $v(x)$ is a smooth function of x. The function $v(x)$ is assumed to be positive for every x and is called the variance function. It is common to assume that ϵ_i's are normally distributed. x_i's can be fixed or random. In general, it is assumed that x_i's are realizations of random variables X_i where $X_i \overset{iid}{\sim} f(\cdot)$ and $f(\cdot)$ is a probability density function.

1.2.2 Nadaraya-Watson Estimator

Let K be a real valued function that satisfies the following conditions:

i) $K(\cdot)$ is symmetric,

ii) $K(\cdot)$ is bounded and continuous on the range of X, say $\mathcal{X}$,

iii) $\int_{\mathcal{X}} K(a)da = 1$.

The function K is called a kernel function. The Nadaraya-Watson estimator of $m(x_0)$ at a given point x_0 is

$$\hat{m}_h(x_0) = \left\{ \sum_{i=1}^{n} K_h(x_i - x_0) \right\}^{-1} \sum_{i=1}^{n} K_h(x_i - x_0) y_i, \tag{1.27}$$

where h is a nonnegative number known as the bandwidth and $K_h(u) = h^{-1}K(u/h)$. See Nadaraya (1964) and Watson (1964).

Commonly used kernel functions include the Gaussian kernel

$$K(u) = (\sqrt{2\pi})^{-1} exp(-u^2/2), \tag{1.28}$$

and the symmetric Beta family

$$K(u) = \{\text{Beta}(1/2, t+1)\}^{-1} \{(1-u^2)_+\}^t, \tag{1.29}$$

where $\text{Beta}(p,q) = \int_0^1 x^p (1-x)^q dx$, $t = 0, 1, 2, ...$ and

$$(1-u^2)_+ = \begin{cases} 1-u^2, & u^2 \leq 1 \\ 0, & u^2 > 1. \end{cases} \tag{1.30}$$

The choices $t = 0$, 1, 2, and 3 lead to the uniform, the Epanechnikov, the biweight, and the triweight kernel respectively. Note that the constant factors in (1.28) and (1.29) are normalization constants and are not used in $\hat{m}(x_0)$.

An important question to consider is the selection of bandwidth h. One may use local bandwidths where h is a function of x_0 or a global bandwidth where h does not depend on x_0. Section 1.2.4 gives two common approaches of bandwidth selection.

Let $\{a_n\}$ be a sequence of real numbers and $\{b_n\}$ be a sequence of positive real numbers. We write $a_n = o(b_n)$ if $b_n^{-1} a_n \to 0$ as $n \to \infty$. Under certain regularity conditions the bias and the variance of the Nadararya-Watson estimator are given by

$$E[\hat{m}(x_0) - m(x_0)] = h^2 \left\{ m^{(2)}(x_0) + 2f^{-1}(x_0) m^{(1)}(x_0) f^{(1)}(x_0) \right\} \int_{-\infty}^{\infty} 2^{-1} u^2 K(u) du + o(h^2), \tag{1.31}$$

and

$$V[\hat{m}(x_0)] = (nh)^{-1}\left\{f^{-1}(x_0)\sigma^2\int_{-\infty}^{\infty}K^2(u)du\right\} + o\left((nh)^{-1}\right), \tag{1.32}$$

where $m^{(r)}(x_0)$ denotes the r^{th} derivative of $m(x)$ evaluated at $x = x_0$, $f(x)$ is the distribution function of x, and x_0 is an interior point of the X space $\mathcal{X}$. See Härdle (2002) for a discussion on the Nadaraya-Watson estimation method.

Although the Nadaraya-Watson estimator has several attractive features, it has some limitations. The bias (1.31) of the estimator depends on the derivative of the distribution of x and could be large when the ratio $f^{(1)}(x)/f(x)$ is large. The Nadaraya-Watson estimator fits a local constant in the sense that it minimizes a local least squares $\sum_{i=1}^{m}\{y_i - m(x)\}^2 w_i$ to obtain $\hat{m}(x) = (\sum_{i=1}^{m} w_i)^{-1}\sum_{i=1}^{m} w_i y_i$ where $w_i = h^{-1}K\{h^{-1}(X_i - x\}$ and $K(\cdot)$ is a kernel function. The Nadaraya-Watson estimator has a higher order bias at a boundary point than at an interior point (Fan and Gijbels, 1996). Local linear fit overcomes some limitations of a local constant fit. See Fan and Gijbels (1996), Hastie and Loader (1993), and Chu and Marron (1991) for a comparison between local constant and local linear fits.

1.2.3 Local Polynomial Estimator

Local polynomial estimators fit a polynomial model locally. This suggests minimizing the objective function

$$\sum_{i=1}^{n}\{y_i - \sum_{j=0}^{p}\beta_j(X_j - x)^j\}^2 K_h(X_i - x) \tag{1.33}$$

with respect to the parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ where $K_h(u) = h^{-1}K(h^{-1}u)$, $K(\cdot)$ is a kernel function, and p is the degree of the local polynomial. A local polynomial estimator of the function $m^{(\nu)}(x)$ at the point x_0 is given by

$$m^{(\nu)}(x_0) = \nu!\mathbf{e}_{\nu+1}^T[X_{x_0}^T W_{x_0} X_{x_0}]^{-1}[X_{x_0}^T W_{x_0}\mathbf{y}], \tag{1.34}$$

where $m^{(\nu)}(x_0) = \frac{d^\nu}{dx^\nu} m(x)|_{x=x_0}$,

$$X_{x_0} = \begin{bmatrix} 1 & X_1 - x_0 & \dots & (X_1 - x_0)^p \\ 1 & X_2 - x_0 & \dots & (X_2 - x_0)^p \\ \dots & \dots & \dots & \dots \\ 1 & X_n - x_0 & \dots & (X_n - x_0)^p \end{bmatrix}, \tag{1.35}$$

$W_{x_0} = \text{diag}\{K_h(X_i - x_0)\}_{i=1}^n$, $\mathbf{e}_\nu$ is the identity vector with the ν^{th} element as one, and diag $\{a_1, \dots, a_n\}$ denotes a diagonal matrix with diagonal elements $a_1, \dots, a_n$. Note that the Nadaraya-Watson estimator (1.27) is a special case of the local polynomial estimator (1.34) with $p = 0$.

The residual variance σ^2 in model (1.26) can be estimated by

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{m} \{y_i - \hat{m}(x_i)\}^2. \tag{1.36}$$

Smooth estimators for σ^2 are also available in the literature. Ruppert et al. (1997) proposed to smooth the observed squared residuals $\{y_i - \hat{m}(x_i)\}^2$ using a local polynomial of order p_2 and a bandwidth h_2.

The bias and the variance of the local polynomial estimator (1.34) are given in Theorem 1.2.1.

Theorem 1.2.1 *Let $K(\cdot)$ be a kernel function. Let $\mu_j = \int_{\mathcal{X}} u^j K(u)du$, $\nu_j = \int_{\mathcal{X}} u^j K^2(u)du$, $S = [\mu_{j+l}]_{0 \le j,l \le p}$, $S^* = [\nu_{j+l}]_{0 \le j,l \le p}$, $\mathbf{c}_p = (\mu_{p+1}, \dots, \mu_{2p+1})^T$, and $\tilde{\mathbf{c}}_p = (\mu_{p+2}, \dots, \mu_{2p+2})^T$. Let $\hat{m}^{(\nu)}(x_0)$ be the local polynomial estimator (1.34) of $m^{(\nu)}(x_0)$. Further let $X_i \overset{iid}{\sim} f(x)$. Assume the following conditions:*

(A1) $f(x_0) > 0$.

(A2) $f(x)$, $m^{(p+1)}(x)$ and $\sigma^2(x)$ are continuous in a neighborhood $N_\delta(x_0)$ of x_0.

(A3) $h \to 0$ and $nh \to \infty$.

Then the following results are true:

(R1) The conditional bias for estimating $m^{(\nu)}(x_0)$ when $p-\nu$ odd is given by

$$E[\hat{m}^{(\nu)}(x_0) - m^{(\nu)}(x_0)|\mathbb{X}] = \mathbf{e}_{\nu+1}^T S^{-1}\mathbf{c}_p \frac{\nu!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-\nu} + o_p(h^{p+1-\nu}), \quad (1.37)$$

where $\mathbb{X}$ is the sigma algebra generated by X_1, X_2, ..., X_n.

(R2) The conditional variance of $\hat{m}^{(\nu)}(x_0)$ is given by

$$V[\hat{m}^{(\nu)}(x_0)|\mathbb{X}] = \mathbf{e}_{\nu+1}^T S^{-1} S^* S^{-1} \mathbf{e}_{\nu+1} \frac{\nu!^2 \sigma^2}{f(x_0) n h^{1+2\nu}} + o_p\left((nh^{1+2\nu})^{-1}\right). \quad (1.38)$$

Further assume the following conditions:

(A4) $f^{(1)}(x)$ and $m^{(p+2)}(x)$ are continuous in a neighborhood of $N_{\delta_2}(x_0)$ of x_0.

(A5) $nh^3 \to \infty$.

Then the next result is true:

(R3) The conditional bias for estimating $m^{(\nu)}(x_0)$ when $p-\nu$ is even is given by

$$\begin{aligned} & E[\hat{m}^{(\nu)}(x_0) - m^{(\nu)}(x_0)|\mathbb{X}] \quad (1.39) \\ &= \mathbf{e}_{\nu+1}^T S^{-1} \tilde{c}_p \frac{\nu!}{(p+2)!} \{m^{(p+2)}(x_0) + (p+2) m^{(p+1)}(x_0) \frac{f^{(1)}(x_0)}{f(x_0)}\} h^{p+2-\nu} \\ &\quad + o_p(h^{p+2-\nu}). \end{aligned}$$

See Fan and Gijbels (1996) for the proof of Theorem 1.2.1.

1.2.4 Bandwidth Selection for Local Estimators

Performances of the Nadaraya-Watson estimator (1.27) and the local polynomial estimator (1.34) largely depend on the selection of bandwidth h. The bandwidth parameter h controls both the bias and the variance of a local smoother. The Nadaraya-Watson estimators are local polynomial estimators with $p = 0$, where p is the degree of the local polynomial. In this section we discuss the selection of h for the local polynomial estimators. The results follow directly for the Nadaraya-Watson estimators. The bias (1.37) for estimating the mean using a local polynomial regression estimator vanishes as h tend to zero. On the other hand, the variance (1.38) of a local polynomial regression

estimator increases as h decreases. Several methods are available in the literature to select a "good" bandwidth for local smoothers. Methods for selecting bandwidths are based on minimizing the mean squared error (MSE)

$$\text{MSE}[\hat{m}^{(\nu)}(X_i)|\mathbb{X}] = E[\hat{m}^{(\nu)}(X_i) - m^{(\nu)}(X_i)|\mathbb{X}]^2, \tag{1.40}$$

where $m(X_i)$ is the mean function evaluated at X_i, $m^{(\nu)}(x_i) = \frac{d^\nu}{dx^\nu} m(x)|_{x=x_i}$, $\hat{m}^{(\nu)}(X_i)$ is an estimator of $m^{(\nu)}(X_i)$, and $\nu = 1, 2, \ldots, p$. Local optimal bandwidths and global optimal bandwidths are the two most commonly used bandwidths in practice. Local optimal bandwidths minimizes the MSE locally at a given point x_0. Thus the local optimal bandwidths at x_0 are obtained by minimizing

$$\begin{aligned} \text{MSE}[\hat{m}^{(\nu)}(x_0)|\mathbb{X}] &= \left\{ \mathbf{e}_{\nu+1}^T S^{-1} \mathbf{c}_p \frac{\nu!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-\nu} \right\}^2 \\ &\quad + \mathbf{e}_{\nu+1}^T S^{-1} S^* S^{-1} \mathbf{e}_{\nu+1} \frac{\nu!^2 \sigma^2(x_0)}{f(x_0) n h^{1+2\nu}} \end{aligned} \tag{1.41}$$

as a function of h, where $\mu_j = \int_{\mathcal{X}} u^j K(u) du$, $\nu_j = \int_{\mathcal{X}} u^j K^2(u) du$, $S = [\mu_{j+l}]_{0 \le j,l \le p}$, $S^* = [\nu_{j+l}]_{0 \le j,l \le p}$, $\mathbf{c}_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$ and $\sigma^2(x_0) = \text{Var}[y|X = x_0]$. Terms of order $o_p(a_n)$, where $a_n = \max\{h^{p+1-\nu}, (nh^{1+2\nu})^{-1}\}$, are ignored in (1.41). A minimizer of (1.41) is given by,

$$h_{\text{opt}} = C_{\nu,p}(K) \left[\{m^{p+1}(x_0)\}^{-2} \{f(x_0)\}^{-1} \sigma^2(x_0) \right]^{1/(2p+3)} n^{-1/(2p+3)}, \tag{1.42}$$

where

$$C_{\nu,p}(K) = \{2(\mathbf{e}_{\nu+1}^T S^{-1} \mathbf{c}_p)(\mathbf{e}_{\nu+1}^T S^{-1} \mathbf{c}_p)^T (p+1-\nu)\}^{-1} [\{\mathbf{e}_{\nu+1}^T S^{-1} S^* S^{-1} \mathbf{e}_{\nu+1}\} \{(p+1)!\}^2 (2\nu+1)]. \tag{1.43}$$

Global optimal bandwidth minimizes the overall MSE

$$\int_{\mathcal{X}} [\text{MSE}(\hat{m}(x)|\mathbb{X})] \, w(x) dx \tag{1.44}$$

as a function of h, where $w(x) \ge 0$ is a weight function and $\mathcal{X}$ is the entire X space. The minimizer of (1.44) is

$$h_{opt} = C_{\nu,p}(K) \left[\left\{ \int_{\mathcal{X}} \{m^{p+1}(x)\}^2 w(x) dx \right\}^{-1} \right. \tag{1.45}$$

$$\left\{\int_{\mathcal{X}} \sigma^2(x)w(x)\{f(x)\}^{-1}dx\right\}\Bigg]^{1/(2p+3)} n^{-1/(2p+3)}. \tag{1.46}$$

One advantage of the local bandwidths relative to a global bandwidth is the local bandwidths could be small at high peaked regions of the mean curve and large at the flat regions. Thus, local bandwidths allow reduction of the bias at peaked regions and the variance at flat regions of the mean curve (Fan and Gijbels, 1996).

Optimal bandwidths (1.42) and (1.46) involve unknown parameters $\sigma^2(\cdot)$ and $m^{(p+1)}(\cdot)$. The term $C_{\nu,p}$ depends on the kernel function $K(\cdot)$ and the degree of the local polynomial p but the other terms in (1.42) and (1.46) are unknown. The two most commonly used methods are the plug-in method and the cross validation method. The plug-in method replaces the unknown values of $\sigma^2(\cdot)$ and $m^{(p+1)}(\cdot)$ with their estimated values. $m^{p+1}(x)$ can be estimated by fitting a local polynomial of degree $p+3$ using an initial bandwidth h_0. $\sigma^2(x_0)$ can be estimated from the residual sum of squares by fitting a local polynomial of degree $p+3$ with bandwidth h_0. To avoid the dependence of the optimal bandwidth on h_0, an iterative procedure is proposed. For examples see Gasser et al. (1991) and Ruppert et al. (1995). Starting with a large h_0, h_{i+1} is given by

$$h_{i+1} = C_{\nu,p}(K)\{\hat{m}_{h_i}^{p+1}(x_0)\}^{-2}\{f(x_0)\}^{-1}\hat{\sigma}_{h_i}^2(x_0) \tag{1.47}$$

for a local optimal bandwidth and by

$$h_{i+1} = C_{\nu,p}(K)\left\{\int_{\mathcal{X}}\{\hat{m}_{h_i}^{p+1}(x)\}^{-2}w(x)dx\right\}^{-1}\left\{\int_{\mathcal{X}}\hat{\sigma}_{h_i}^2(x)w(x)\{f(x_0)\}^{-1}dx\right\} \tag{1.48}$$

for a global bandwidth.

The cross validation method for selecting a global optimal bandwidth minimizes the weighted least squares

$$\sum_{i=1}^{n}\{y_i - \hat{m}_{h,-i}(x_i)\}^2 w(x_i), \tag{1.49}$$

where $\hat{m}_{h,-i}(x_i)$ is the local polynomial estimate of $m(x_i)$ using bandwidth h and deleting the i^{th} data element. See Chapter 3 in Fan and Gijbels (1996).

1.3 Dissertation Organization

The National Resources Inventory (NRI) Survey is a longitudinal survey of non-federal lands in the US and its territories. The data obtained from the NRI survey is used to estimate wind erosion for counties. A transformed Fay-Herriot model using observed county means is covered in Chapter 2. A soil erodibility index is available from the administrative records for each county and is used as the predictor. A county level estimator that is adjusted for the transformation bias and is calibrated to the state level is proposed.

Chapters 3 and 4 introduce a nonparametric area level model for small area estimation. Chapter 3 covers a Nadaraya-Watson estimator for small area means. Assuming the normality of the error components $\mathbf{u}$ and $\boldsymbol{\epsilon}$ and assuming equal design variances for county means, we propose a second order approximation of the MSE of the proposed estimate. The results from a simulation study and an application of the proposed method to estimate soil erosion due to wind are also discussed in Chapter 3.

The results of the Nadaraya-Watson estimator of county means are generalized in Chapter 4 using local polynomial regression. A class of estimators based on local polynomial regression is proposed in Chapter 4. Both the small area mean function and the between area variance function are modeled as smooth functions of the area level covariates. An approximation for the MSE of the proposed estimator based on a Taylor linearization was developed and its asymptotic properties are studied.

Small area estimation in the presence of missing values is considered in Chapter 5. An estimation technique is developed for the cover and crop management factor that can be used in small area estimation for the counties. We propose an estimator for the sampling error covariance matrix adjusted for the imputed values. A multivariate area-level model is proposed that uses the estimated covariance matrix. Finally, discussions are given in Chapter 6.

CHAPTER 2. Small Area Estimation For A Nonlinear Transformation

2.1 Introduction

Wind erosion is a severe problem in some mid-western states in the US where soil loss can result in large decreases in soil productivity. Although some work has been done by the Wind Erosion Research Unit (WERU) to estimate wind erosion at the national level, there has been little work to estimate wind erosion at the county level. Our main objective is to build a small area model for counties to estimate wind erosion. The weighted sum of the predicted county means are not always same as the state direct estimate. Our second objective is to propose a methodology so that the small area predictions are calibrated up to a higher level. We propose an approach to estimate wind erosion at the county level using the National Resources Inventory (NRI) data set. A transformed Fay-Herriot model is used to predict county means. We include the design weight in our proposed model in such a way that the final estimates are calibrated with the state estimates. The proposed model fits the data well and produces a mean squared error of prediction that is approximately one half of the design standard error.

Section 2.2 gives an introduction to the NRI survey. The problem of wind erosion and some previous results are discussed in Section 2.3. In Section 2.4, we present results from an exploratory study and in Section 2.5, calibration for estimated county means is proposed. Finally, Section 2.6 discusses the advantages and disadvantages of the proposed method.

2.2 The NRI Survey

The NRI is a longitudinal survey conducted by the US Department of Agriculture's (USDA) Natural Resources Conservation Service (NRCS) in cooperation with the Center for Survey Statistics and Methodology (CSSM). The survey is designed to assess conditions and trends for land cover, soil, water, and related natural resources on non-federal lands in the US. The NRI was conducted every 5 years during 1982-1997. The basic design of NRI surveys is a stratified, two-stage area sample. The land area of most states in the US is divided according to the Public Land Survey (PLS) system, which has provided a convenient structure for developing NRI sample selection procedures and for locating primary sampling units (PSUs) in the field. Three sample points are selected within each PSU according to a restricted randomization procedure, see Nusser and Goebel (1997). The 1997 NRI contains approximately 300,000 PSU and over 800,000 sample points. Sampling rates across the US generally range from 2% to 6% of the land area, though rates occasionally fall outside this 6% range. The sampling rate within a county is increased when larger sample sizes are needed for special studies or when heterogeneous patterns exist for irrigation soil types, land uses, major land resource areas, or hydrologic regions (Nusser and Goebel, 1997).

Since 2000, the full panel structure of the NRI has been replaced by a two-phase supplemented panel sampling design in which the 1997 NRI segments serve as a first phase and each year a partially overlapping panel is selected through a stratified sampling design as a second phase. The annual second phase sample includes approximately 42,000 "core" segments that are to be observed every year (Fig. 2.1). An additional 30,000 segments are selected from the remaining 268,000 PSU each year to form a supplemental sample. All points in every selected segment are part of the annual sample (Fuller, 2003). Data are collected in two levels. Urban land, water, etc. are collected at the PSU level whereas soil properties, land use etc. are collected at the point level.

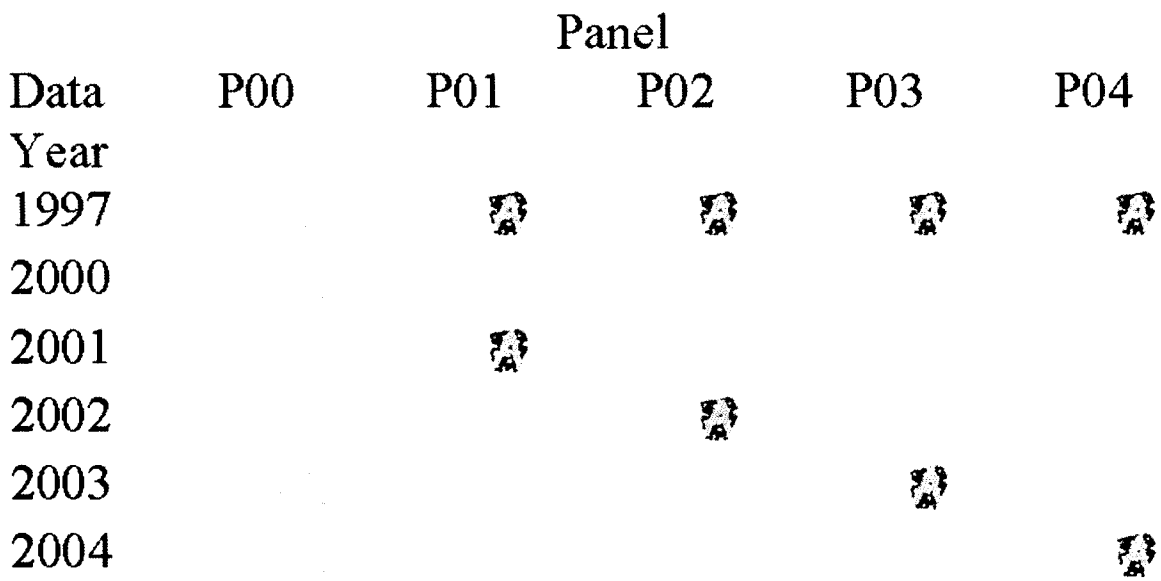

Figure 2.1 Supplemented Panel Design for the NRI.

2.3 Variables of Interest for Wind Erosion

Wind erosion is a serious problem in many parts of the world including arid and semiarid regions where it is known to be considerably worse. The areas most susceptible to wind erosion on agricultural lands include much of North Africa and the Near East; parts of southern, central, and eastern Asia; the Siberian Plains; Australia; northwest China; southern South America; and North America (Wind Erosion Research Unit (WERU), *http://www.weru.ksu.edu/problem.html.*

An extensive dry spell during the 1930's ended in dust storms and severe soil damage of catastrophic size. Wind erosion continues to threaten the sustainability of Americas natural resources even seventy years after the Dust Bowl ended. Still today, as early as 1997, wind erosion severely damaged agricultural land throughout the Great Plains (WERU). On average, wind erosion is responsible for about 40 percent of the total soil loss in the US (Hagen, 1994), and can increase drastically during drought years (Hagen and Woodruff, 1973). In the US, wind erosion is a dominant problem on approximately 73.6 million acres and moderately to severely damage approximately 4.9 million acres

annually (USDA, 1965). According to the 1992 National Resources Inventory (NRI), the estimated annual soil loss from wind erosion on non-federal rural land in the US was 2.5 tons per year (SCS-USDA, 1994).

It is of much interest to the local governments to estimate the wind erosion pertinent to their area. Although several studies address the issue of estimating soil loss at a national level, no effort has been made to provide precise estimation of soil loss at a lower level (e.g., county or city). Our objective is to produce estimations of wind erosion for counties in various eastern, and mid-western (highly susceptible area for wind erosion) states in the US. Data was used from the NRI survey to estimate wind erosion in 2002 for counties in Iowa. In this work, we used the soil erodibility index (IFact) as the predictor variable and wind equation for 2002 (WEQ02) as the response variable. The use of IFact as a predictor has several advantages. IFact is directly related with wind erosion. Higher IFact values indicate greater susceptibility to wind erosion. IFact can be obtained from the Natural Resources Conservation Service (NRCS) soil survey database available through the NRCS Soil Data Mart (SDM) for each county in the US. Since IFact is a soil characteristic, it doesn't change much over time. For this study, we used IFact values from the 1997 NRI sampled points. WEQ02 is not directly observed in the field, rather it is calculated as a function of several factors. Soil erodibility, climate, slope, and land cover are just a few of those factors (Bell et al., 2003). WEQ02 is measured in tonnes/ha and is used as observed wind erosion in this study.

2.4 Exploratory Analysis

Table 2.1 gives summary statistics for the 2002 [1] survey weighted county mean of wind erosion, denoted by WEQ02, for the counties in Iowa. The range is the the difference between the highest and the lowest weighted means. Data from the core panel and

[1]The 2002 NRI data set has not yet been released for public use. All values are strictly for research purposes.

Table 2.1 Summary Statistics for Survey Weighted Mean WEQ02 and the Design Standard Error for Iowa Counties

	First Quartile	Median	Mean	Third Quartile	Range
Weighted Mean	0.2037	0.514	0.696	0.963	3.474
Standard Error	0.0694	0.131	0.164	0.217	0.648

the 2002 supplemental panel are used in Table 2.1. Many counties have a high standard error of the mean relative to the direct mean. We propose a small area model to produce estimates of county means. Let y_i be the survey weighted mean WEQ02 for county i, and let x_i be the mean IFact for the same county, where $i = 1, 2, ..., m$. Following Fay and Herriot (1979), we propose the small area model

Model I (Untransformed Model):

$$y_i = \beta_0 + \beta_1 x_i + u_i + e_i \tag{2.1}$$

where β_0 are β_1 are fixed parameters, e_i is the sampling error and u_i is the area specific random effect. We assume $e_i \overset{ind}{\sim} N(0, D_i)$, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ and assume the e_i's and u_i's are independent. D_i's are estimated from the unit level information using the surveymeans procedure in SAS and assumed to be known. The empirical best linear unbiased predictor (EBLUP) of $\mu_i = \mathbf{x}_i^T\boldsymbol{\beta} + u_i$ is given by

$$\begin{aligned} \hat{\mu}_i &= \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\mathbf{x}_i^T\hat{\boldsymbol{\beta}} && (2.2) \\ &= \mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}) && (2.3) \end{aligned}$$

where

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}\mathbf{y} \\ &= \left[\sum_{i=1}^{m} \mathbf{x}_i\mathbf{x}_i^T(\hat{\sigma}_u^2 + D_i)^{-1}\right]^{-1}\left[\sum_{i=1}^{m} \mathbf{x}_i y_i(\hat{\sigma}_u^2 + D_i)^{-1}\right] && (2.4) \end{aligned}$$

is the empirical generalized least square (EGLS) estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\gamma_i = (\sigma_u^2 + D_i)^{-1}\sigma_u^2$,

$$\hat{\gamma}_i = (\hat{\sigma}_u^2 + D_i)^{-1}\hat{\sigma}_u^2, \tag{2.5}$$

$\mathbf{y} = (y_1, \ldots, y_m)^T$, $\mathbf{x}_i = (1, x_i)^T$, $X = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_m^T)^T$, $V = \text{diag}\{\sigma_u^2 + D_i\}_{i=1}^m$, $\hat{V} = \text{diag}\{\hat{\sigma}_u^2 + D_i\}_{i=1}^m$, and $\hat{\sigma}_u^2$ is an estimator of σ_u^2. For a set of scalars $\{a_1, a_2, \ldots, a_m\}$, $\text{diag}\{a_i\}_{i=1}^m$ denotes a diagonal matrix with elements $a_1, a_2, \ldots, a_m$ and A^T denotes the transpose of a matrix A. Prasad and Rao (1990) expressed the mean squared error (MSE) of the EBLUP (2.3) as

$$E[\hat{\mu}_i - \mu_i]^2 = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2), \tag{2.6}$$

where

$$g_{1i}(\sigma_u^2) = (\sigma_u^2 + D_i)^{-1}\sigma_u^2 D_i \tag{2.7}$$

is the MSE when $\boldsymbol{\beta}$ and σ_u^2 are known,

$$g_{2i}(\sigma_u^2) = (1 - \gamma_i)^2 \mathbf{x}_i^T [\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T (\sigma_u^2 + D_i)^{-1}]^{-1} \mathbf{x}_i \tag{2.8}$$

is the effect of estimating $\hat{\beta}$, and

$$g_{3i}(\sigma_u^2) = D_i^2 (\sigma_u^2 + D_i)^{-3} \bar{V}(\hat{\sigma}^2) \tag{2.9}$$

is the effect of estimating σ_u^2 where $\bar{V}(\hat{\sigma}^2)$ is the asymptotic variance of $\hat{\sigma}_u^2$. The asymptotic variance of the residual maximum likelihood (REML) estimator of σ_u^2 is (Rao, 2003)

$$\bar{V}(\hat{\sigma}^2) = 2[\sum_{i=1}^m (\sigma_u^2 + D_i)^{-2}]^{-1}. \tag{2.10}$$

The MSE in (2.6) can be estimated by

$$\text{mse}[\hat{\mu}_i] = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2), \tag{2.11}$$

where $\hat{\sigma}_u^2$ is the REML estimator of σ_u^2 (Prasad and Rao, 1990).

Table 2.2 Parameter Estimates for the Small Area Models

	β_0	β_1	β_2	σ_u^2
Model I	-1.532	0.0365	-	0.1102
	(0.312)	(0.0054)	-	(0.0283)
Model II(b)	-8.374	0.1900	-0.00099	0.1639
	(2.991)	(0.0993)	(0.00807)	(0.0887)
Model III	-0.293	0.0182	-	0.0204
	(0.145)	(0.0025)	-	(0.0062)

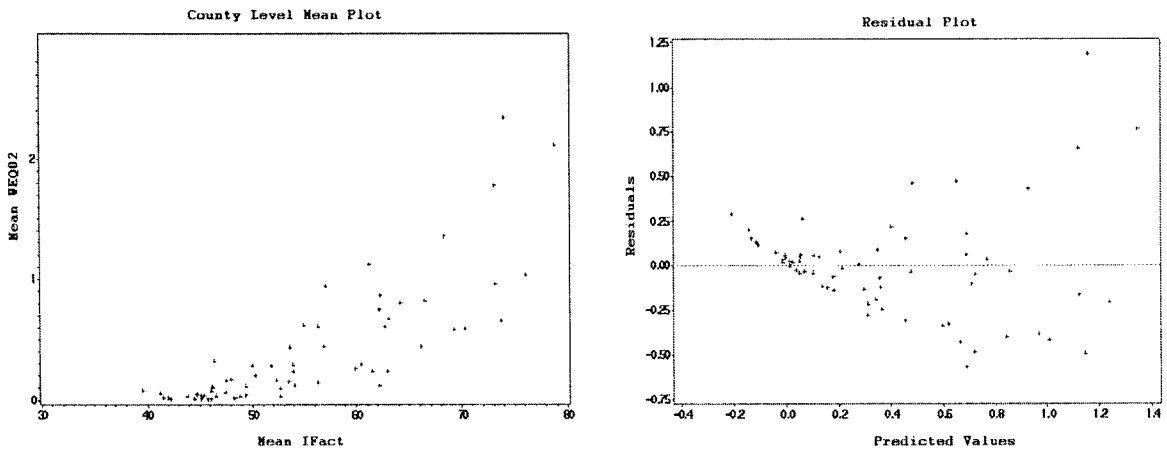

Figure 2.2 Scatter Plot and Residual Plot from Model I.

The regression parameter of model (2.1) was estimated using EGLS and the between area variance parameter was estimated using the REML. The PROC MIXED procedure in SAS is used to estimate the parameters. The parameter estimates and their standard errors are given in Table 2.2. The standard errors are given in parentheses. The first part of Figure 2.2 is a scatter plot of the mean WEQ02 against the mean IFact. The second part of Figure 2.2 is the residual plot from regressing mean WEQ02 on mean IFact.

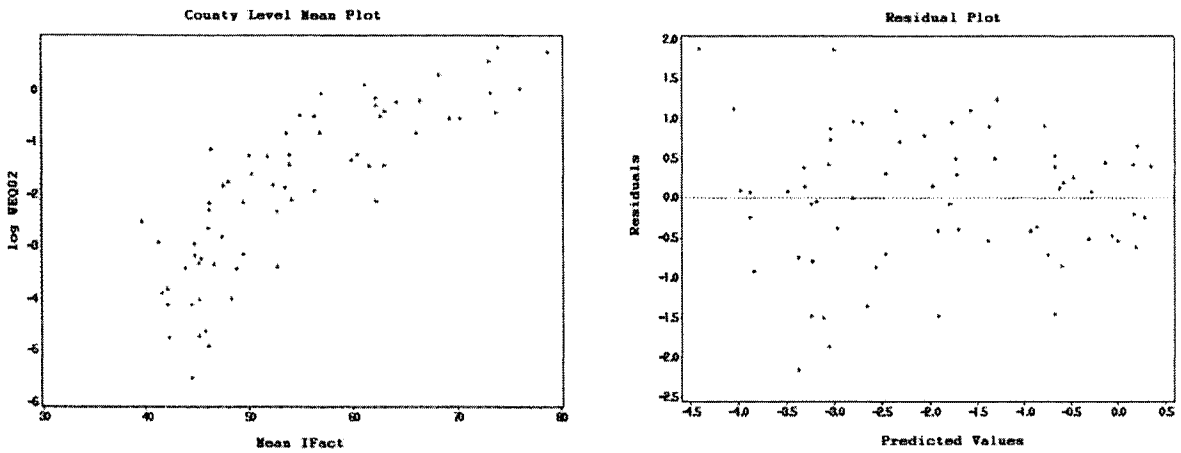

Figure 2.3 Scatter Plot and Residual Plot from Model II.

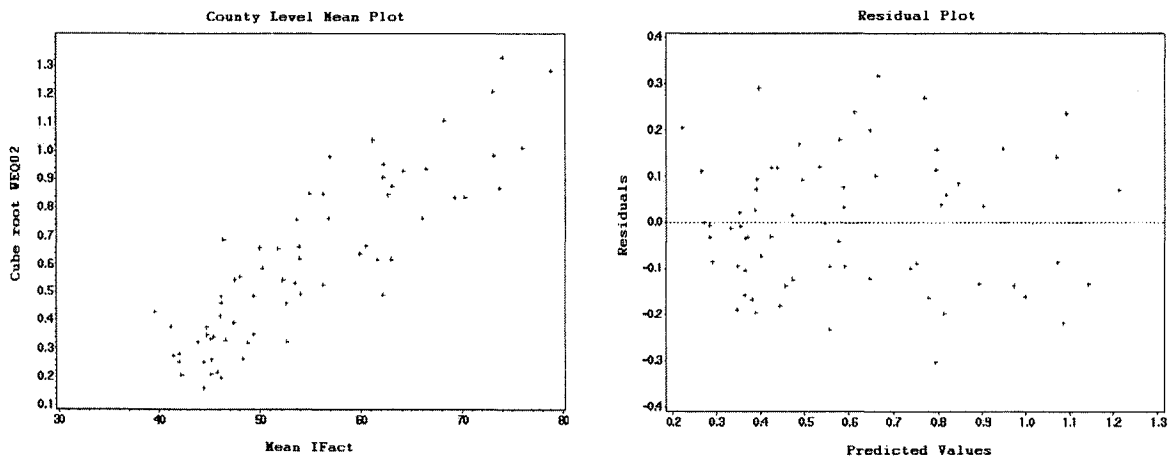

Figure 2.4 Scatter Plot and Residual Plot from Model III.

Nonlinearity and unequal residual variance are clear from the plot which leads us to consider a transformations of y_i. Three models are considered,

Model II (Log Transformation):

$$log(y_i) = \beta_0 + \beta_1 x_i + u_i + e_i, \tag{2.12}$$

where $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $e_i \overset{ind}{\sim} N(0, D_i^{**})$, u_i's and e_i's are independent, $D_i^{**} = y_i^{-2} D_i$ and D_i's are known.

Model II(b) (Log Transformation with IFact Squared):

$$log(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i + e_i, \tag{2.13}$$

where $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $e_i \overset{ind}{\sim} N(0, D_i^{**})$, u_i's and e_i's are independent, $D_i^{**} = y_i^{-2} D_i$ and D_i's are known.

Model III (Cube Root Transformation):

$$(y_i)^{1/3} = \beta_0 + \beta_1 x_i + u_i + e_i, \tag{2.14}$$

where $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $e_i \overset{ind}{\sim} N(0, D_i^{*})$, u_i's and e_i's are independent, $D_i^{*} = (9 y_i^{4/3})^{-1} D_i$ and D_i's are known.

Figure 2.5 is a plot for D_i, D_i^*, and D_i^{**} against n_i^{-1}, where n_i's are county sizes. The D_i^*'s are more nearly constant with respect to n_i^{-1}. The regression parameters of the models were estimated using the EGLS (2.4) and the between area variance parameter was estimated using the REML. Parameter estimates and their standard errors as obtained from PROC MIXED in SAS are given in Table 2.2. Figure 2.3 contains the scatter plot of log of WEQ02 and IFact and the residual plot for model II(b). Similar plots for model III are shown in Figure 2.4. The scatter plot in Figure 2.3 shows a quadratic pattern and accordingly a second order term of IFact is also included in the model to create model II(b). The scatter plot in Figure 2.4 has a linear trend and the residual plot in Figure 2.4 supports a linear model. Model III (Figure

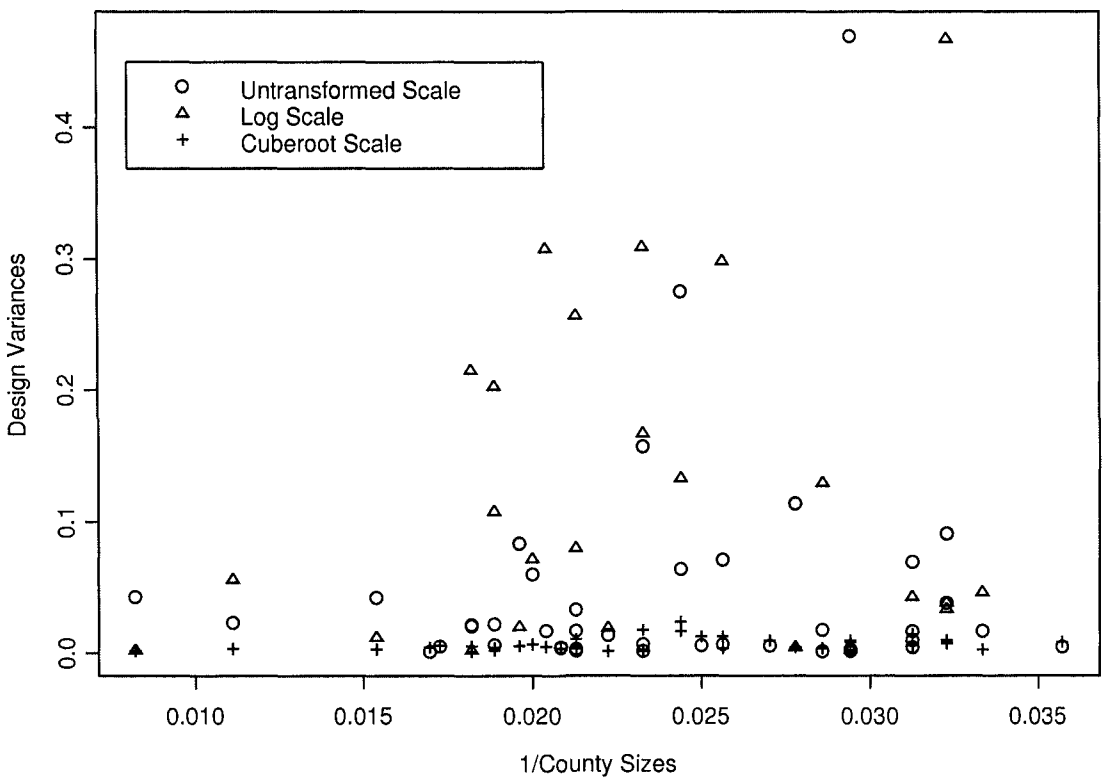

Figure 2.5 Estimated Design Variances for Transformed Estimates.

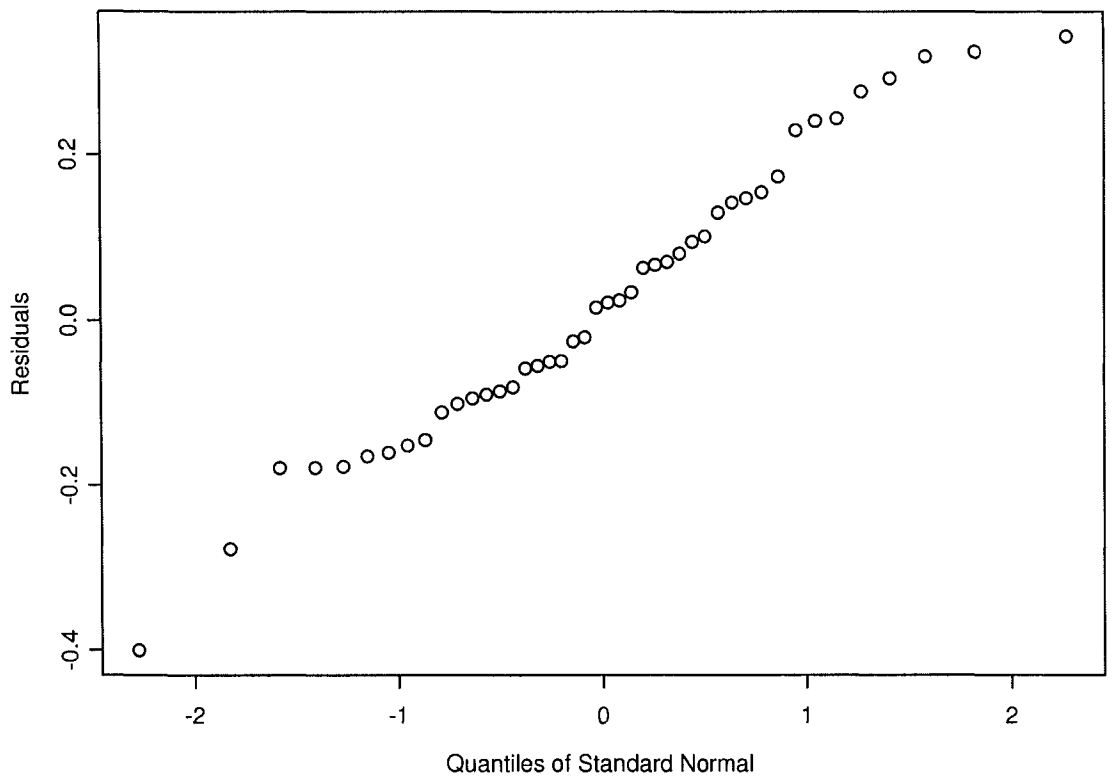

Figure 2.6 Normal Quantiles Plot for Model III.

2.4) is a better fit with an adjusted R-Squared of 0.63. The normal quantiles plot for the observed marginal residuals from model III is given in Figure 2.6. There is no evidence from the plot that the distributions of the observed residuals deviate from the normality assumption. We dropped model II and II(b) from further analysis. Since the analysis suggests a transformation on y_i, the standard small area predictions (2.3) and the estimation of MSE (2.11) need to be adjusted. Following Slud and Maiti (2006) we used

$$\hat{\mu}_i = \{\hat{\gamma}_i(y_i)^{1/3} + (1-\hat{\gamma}_i)x_i^T\hat{\beta}\}^3\{3\hat{\sigma}_u^2\hat{\gamma}_i + (x_i^T\hat{\beta})^2\}^{-1}\{3\hat{\sigma}_u^2 + (x_i^T\hat{\beta})^2\} \tag{2.15}$$

for the model III, where $\hat{\boldsymbol{\beta}}$ is the EGLS estimate of $\boldsymbol{\beta}$, $\hat{\gamma}_i = (\hat{\sigma}_u^2 + D_i^*)^{-1}\hat{\sigma}_u^2$, and $\hat{\sigma}_u^2$ is the REML estimator of σ_u^2 for model III. Summary statistics for the predicted means from models I and III are given in Table 2.3.

2.5 A Regression Based Calibrated Small Area Estimator

The direct domain estimators are frequently used for populations and subpopulations with large sample sizes. If a subpopulation with acceptable direct estimates is divided into a number of small areas, then it is desirable that the weighted sum of the small area predicted means is the same as the direct subpopulation mean. For the NRI survey, it is desired that the design weighted predicted county means be close to the state direct mean. By calibrated small area estimation, we mean the weighted sum of the estimated county means is equal to the direct estimate of the state mean. The direct estimate of the state mean (SDE) SDE $= \sum_{i=1}^{m}(w_i y_i)/\sum_{i=1}^{m}(w_i)$, and the state level model based estimate (SME) SME $= \sum_{i=1}^{m}(w_i\hat{\mu}_i)/\sum(w_i)$, where the w_i's are survey weights associated with county means y_i and $\hat{\mu}_i$'s are predictions from the small area model. We define the relative absolute calibration (RAC) error by RAC = |SME - SDE|/SDE. The RAC are 0.13 and 0.12 for the model I and the model III, respectively.

The calibration at the state level is not necessarily achieved through the proposed model based estimators. For a closer look, consider model I when σ_u^2 is known. The normal equations for estimating $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ for model (2.1) are

$$X^T V^{-1} X\tilde{\boldsymbol{\beta}} = X^T V^{-1}\mathbf{y}, \tag{2.16}$$

where X is the design matrix for model I, $V = \text{diag}\{(\sigma_u^2 + D_i)^{-1}\}_{i=1}^{m}$, and $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$. One of the normal equations for $\boldsymbol{\beta}$ can be written as

$$\sum_{i=1}^{m} \frac{y_i - x_i^T\tilde{\boldsymbol{\beta}}}{\sigma_u^2 + D_i} = 0. \tag{2.17}$$

But for $\hat{\mu}_i$ from (2.3), $\sum_{i=1}^{m} w_i\hat{\mu}_i$ can be written as,

$$\sum_{i=1}^{m} w_i\hat{\mu}_i = \sum_{i=1}^{m} w_i y_i - \sum_{i=1}^{m} w_i \frac{D_i}{\sigma_u^2 + D_i}(y_i - x_i^T\tilde{\boldsymbol{\beta}}). \tag{2.18}$$

If $w_i \propto D_i^{-1}$ the second term in (2.18) is zero by (2.17) and the calibration condition is achieved. Several approaches have been taken to obtain a fully calibrated estimator. Following Wang and Fuller (2002), we propose to include the design weight in the small area model as a covariate in such a way that the score function for the normal equations remains zero. The proposed estimator will have the properties of EBLUP under the model and will be fully calibrated. For model I we include $x_{2i} = D_i w_i$ as a covariate.

Model I(b):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + u_i + e_i. \tag{2.19}$$

where β_2 is a fixed parameter and β_0, β_1, u_i and e_i are defined in model I. We assume $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $e_i \overset{ind}{\sim} N(0, D_i)$ and u_i's and e_i's are mutually independent. Small area means, $\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$, are estimated by

$$\begin{aligned} \hat{\mu}_i &= \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}) && (2.20) \\ &= y_i - (1 - \hat{\gamma}_i)(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}), && (2.21) \end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}_i$ are given in (2.4) and (2.5) respectively. From one of the normal equations for $\boldsymbol{\beta}$ when σ_u^2 is known

$$\sum_{i=1}^{m} \frac{1}{\sigma_u^2 + D_i}\{D_i w_i y_i - D_i w_i \tilde{\beta}_0 - D_i w_i x_i \tilde{\beta}_1 - (D_i w_i)^2 \tilde{\beta}_2\} = 0, \tag{2.22}$$

and using the predicted mean from model I(b)

$$\begin{aligned} &\sum_{i=1}^{m} w_i\hat{\mu}_i \\ &= \sum_{i=1}^{m} w_i y_i - \sum_{i=1}^{m} \frac{1}{\sigma_u^2 + D_i}\{D_i w_i y_i - D_i w_i \tilde{\beta}_0 - D_i w_i x_i \tilde{\beta}_1 - (D_i w_i)^2 \tilde{\beta}_2\}. \end{aligned} \tag{2.23}$$

The second term on the R.H.S of (2.23) vanishes by (2.22). Therefore we achieved calibration by including x_{2i} in model I. We can test the significance of calibration by conducting a significance test on β_2. If model I is correct then by including x_{2i} we have included an unnecessary variable.

In model I, y_i is assumed to be linearly related with the predictor x_i. Proposition 2.5.1 shows how to obtain calibrated predictors when a smooth function of y_i is linearly related with x_i. Let $\mathcal{F}$ be the finite population. The procedure requires iteration and we let $(r) = (1), (2), \ldots, (r^*)$ denote the iteration number.

Proposition 2.5.1 *Let the small area model be*

$$\begin{aligned} y_i &= h(z_i^{(1)}), \\ z_i^{(r)} &= \beta_0^{(r)} + \beta_1^{(r)} x_{1i} + \beta_2^{(r)} \xi_i^{(r-1)} + u_i^{(r)} + e_i, \qquad (2.24) \\ r &= 1, 2, \ldots, r^*, \qquad (2.25) \end{aligned}$$

where $\beta_0^{(r)}$, $\beta_1^{(r)}$ and $\beta_2^{(r)}$ are fixed regression parameters, $u_i^{(r)}$'s are area specific random effects, the e_i's are random errors, $h(\cdot)$ is a smooth function and $i = 1, 2, ..., m$ denote the m small areas. Assume $e_i \overset{ind}{\sim} N(0, D_i)$, $u_i^{(r)} \overset{iid}{\sim} N(0, \sigma_u^{2(r)})$, e_i and $u_i^{(r)}$ are independent, $D_i = V[h^{-1}(y_i)|\mathcal{F}]$ and D_i are known. Let $\gamma_i^{(r)} = (\sigma_u^{2(r)} + D_i)^{-1}\sigma_u^{2(r)}$, $\hat{\gamma}_i^{(r)} = (\hat{\sigma}_u^{2(r)} + D_i)^{-1}\hat{\sigma}_u^{2(r)}$, $\hat{\sigma}_u^{2(r)}$ be the REML estimate of $\sigma_u^{2(r)}$, and w_i be the survey weights for county i. Let

$$\hat{z}_i^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_{1i} + \hat{\beta}_2^{(r)} \xi_i^{(r-1)}, \qquad i = 1, 2, \ldots, m, \qquad (2.26)$$

where $\hat{\boldsymbol{\beta}}^{(r)} = (\hat{\beta}_0^{(r)}, \hat{\beta}_1^{(r)}, \hat{\beta}_2^{(r)})^T = [\{X^{(r)}\}^T \{\hat{V}^{(r)}\}^{-1} X^{(r)}]^{-1} \{X^{(r)}\}^T \{\hat{V}^{(r)}\}^{-1} \mathbf{y}$,
$\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$, $X^{(r)} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_m^T)^T$, $\hat{V}^{(r)} = diag\{(\hat{\sigma}_u^{2(r)} + D_i)^{-1}\}_{i=1}^m$, $\mathbf{x}_i = (1, x_{1i}, \xi_i^{(r-1)})^T$, *and*

$$\xi_i^{(r)} = \begin{cases} (1 - \hat{\gamma}_i^{(r)}) w_i h'(\hat{z}_i^{(r)}) & , \quad r = 1, 2, \ldots, r^* \\ 0 & , \quad r = 0. \end{cases} \qquad (2.27)$$

Let

$$\hat{q}_i^{(r)} = \hat{\alpha}_0^{(r)} + \hat{\alpha}_1^{(r)} x_{1i} + \hat{\alpha}_2^{(r)} \xi_i^{(r)}, \qquad i = 1, 2, \ldots, m, \tag{2.28}$$

where

$$\hat{\boldsymbol{\alpha}}^{(r)} = (\hat{\alpha}_0^{(r)}, \hat{\alpha}_1^{(r)}, \hat{\alpha}_2^{(r)})^T = \left[\{A^{(r)}\}^T \{A^{(r)}\}\right]^{-1} \{A^{(r)}\}^T \mathbf{q}^{(\mathbf{r})}, \tag{2.29}$$

$$\mathbf{q}^{(\mathbf{r})} = (q_1^{(r)}, q_2^{(r)}, \ldots, q_m^{(r)})^T, \quad q_i^{(r)} = \{h'(\hat{z}_i^{(r)})\}^{-1}\{y_i - h(\hat{z}_i^{(r)})\}, \tag{2.30}$$

$A^{(r)} = (\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_m^T)^T$, $\mathbf{a}_i = (1, x_{1i}, \xi_i^{(r)})^T$ *and* $h'(\hat{z}_i)$ *is the derivative of* $h(z_i)$ *evaluated at* $\hat{z}_i$. *Then*

$$\sum_{i=1}^{m} w_i y_i = \sum_{i=1}^{m} w_i \hat{\mu}_i, \tag{2.31}$$

where

$$\hat{\mu}_i = \hat{\gamma}_i^{(r^*)} y_i + (1 - \hat{\gamma}_i^{(r^*)}) \hat{y}_i^{(r^*)}, \tag{2.32}$$

and $\hat{y}_i^{(r^*)} = \hat{\gamma}_i^{(r^*)} y_i + (1 - \hat{\gamma}_i^{(r^*)})\{h(\hat{z}_i^{(r^*)}) + h'(\hat{z}_i^{(r^*)}) \hat{q}_i^{(r^*)}\}$.

Proof of Proposition 2.5.1 By (2.27), (2.29), and (2.30)

$$\sum_{i=1}^{m} \xi_i^{(r*)} (\hat{\alpha}_0^{(r*)} + x_i \hat{\alpha}_1^{(r*)} + \hat{\alpha}_2^{(r*)} \xi_i^{(r*)} - q_i^{(r*)}) = 0. \tag{2.33}$$

By (2.30), $y_i = h(\hat{z}_i^{(r)}) + h'(\hat{z}_i^{(r)}) q_i^{(r)}$, the calibration constraint is

$$\begin{aligned}
&\textstyle\sum_{i=1}^m w_i y_i - \sum_{i=1}^m w_i \hat{\mu}_i \\
&\quad = \textstyle\sum_{i=1}^m w_i (1 - \hat{\gamma}_i^{(r*)}) y_i - \sum_{i=1}^m w_i (1 - \hat{\gamma}_i^{(r*)}) \{h(\hat{z}_i^{(r*)}) + h'(\hat{z}_i^{(r*)}) \hat{q}_i^{(r*)}\} \\
&\quad = \textstyle\sum_{i=1}^m w_i (1 - \hat{\gamma}_i^{(r*)}) \{h(\hat{z}_i^{(r*)}) + h'(\hat{z}_i^{(r*)}) q_i^{(r*)}\} \\
&\qquad - \textstyle\sum_{i=1}^m w_i (1 - \hat{\gamma}_i^{(r*)}) \{h(\hat{z}_i^{(r*)}) + h'(\hat{z}_i^{(r*)}) \hat{q}_i^{(r*)}\} \\
&\quad = \textstyle\sum_{i=1}^m w_i (1 - \hat{\gamma}_i^{(r*)}) h'(\hat{z}_i^{(r*)}) (q_i^{(r*)} - \hat{q}_i^{(r*)}) \\
&\quad = \textstyle\sum_{i=1}^m w_i (1 - \hat{\gamma}_i^{(r*)}) h'(\hat{z}_i^{(r*)}) (q_i^{(r*)} - \hat{\alpha}_0^{(r*)} - \hat{\alpha}_1^{(r*)} x_i - \hat{\alpha}_2^{(r*)} \xi_i^{(r*)}).
\end{aligned}$$

Therefore, with $\xi_i^{(r*)} = w_i (1 - \gamma_i^{(r*)}) h'(\hat{z}_i^{(r*)})$ the right hand side of the last equation is exactly zero by (2.33). □

Proposition 2.5.1 gives an iterative approach to achieve calibration in a generalized linear model. We apply Proposition 2.5.1 with $r^* = 2$ to get calibrated estimators of the small area means. Model III can be extended as,

Model III(b):

$$\begin{aligned} y_i^{1/3} &= z_i^{(1)}, \\ z_i^{(r)} &= \beta_0^{(r)} + \beta_1^{(r)} x_{1i} + \beta_2^{(r)} \xi_i^{(r-1)} + u_i^{(r)} + e_i, \end{aligned} \tag{2.34}$$

$$\tag{2.35}$$

where

$$\xi_i^{(r)} = \begin{cases} (1 - \hat{\gamma}_i^{(r)}) w_i h'(\hat{z}_i^{(r)}) & , \quad r = 1, 2 \\ 0 & , \quad r = 0, \end{cases} \tag{2.36}$$

$$\hat{z}_i^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_{1i} + \hat{\beta}_2^{(r)} \xi_i^{(r-1)}, \tag{2.37}$$

and $\hat{\boldsymbol{\beta}}^{(r)} = (\hat{\beta}_0^{(r)}, \hat{\beta}_1^{(r)}, \hat{\beta}_2^{(r)})^T$ is defined in Proposition 2.5.1. Assume $u_i^{(r)} \overset{iid}{\sim} N(0, \sigma_u^{2(r)})$, $e_i \overset{ind}{\sim} N(0, D_i^*)$, u_i and e_i are independent and D_i^* is defined in model III. Let

$$\hat{q}_i^{(r)} = \hat{\alpha}_0^{(r)} + \hat{\alpha}_1^{(r)} x_i + \hat{\alpha}_2^{(r)} w_i (1 - \hat{\gamma}_i^{(r)}) \hat{z}_i^{2(r)}, \tag{2.38}$$

where $\hat{\boldsymbol{\alpha}}^{(r)} = (\hat{\alpha}_0^{(r)}, \hat{\alpha}_1^{(r)}, \hat{\alpha}_2^{(r)})^T$ is defined in (2.29), and

$$q_i^{(r)} = \{\hat{z}_i^{(r)}\}^{-2} \left[y_i - \{\hat{z}_i^{(r)}\}^3 \right]. \tag{2.39}$$

The predicted means are given by

$$\hat{\mu}_i = \hat{\gamma}_i^{(2)} y_i + (1 - \hat{\gamma}_i^{(2)}) \left[\{\hat{z}_i^{(2)}\}^3 + \hat{q}_i^{(2)} \{\hat{z}_i^{(2)}\}^2 \right]. \tag{2.40}$$

The RAC from model I(b) and model III(b) are zero. Thus the weighted sum of the small area predicted means is the weighted sum of the direct estimates. Figure 2.7 shows a plot of survey weighted county means and predicted means from small area models. The dotted line is the overall state mean and the solid line is the 45^o line. Predictions from model I and model III are between the overall state mean and the observed mean. Predictions from model III(b) follow similar trends except for a few selected counties where the predicted values are further from the overall mean than the observed mean. There are two counties in Figure 2.7 where the predictions from model III(b) are not

Table 2.3 Summary Statistics for Estimated County Means

Model	First Quartile	Median	Mean	Third Quartile	Range
Direct Estimates	0.202	0.514	0.696	0.963	3.48
Model I	0.206	0.540	0.620	0.901	1.80
Model III	0.232	0.489	0.596	0.812	2.00
Model III(b)	0.248	0.536	0.706	0.933	3.14

close to the predictions from model III. County 141 has the highest direct mean (3.53) and the highest IFact (81.91). Predicted means for county 141 from model III is 2.06 and model III(b) is 3.14. Higher values for IFact and WEQ02 are probably the reason for a higher relative difference for county 141. County 149 has the second highest value for the IFact (80.54), the second highest value for the county weight (3877), and the fourth highest value for the direct mean (1.42). Thus, county 149 has the highest value for $\xi_i^{(2)} = (D_i^* + \sigma_u^{2(2)})^{-1} w_i \hat{z}_i^{2(2)}$ and the highest value for $\hat{q}_i^{(2)}$ (1.105). Individual scatter plots of $q_i^{(2)}$ and IFact, $q_i^{(2)}$ and $\xi_i^{(2)}$, and $q_i^{(2)} - \hat{q}_i^{(2)}$ and $\hat{q}_i^{(2)}$ (not shown) suggest that county 149 is highly influential.

The MSE of prediction for model I is obtained from (2.11). For transformed models we used the delta method to estimate the MSE in the original scale. If μ_i is the true small area mean and μ_i^* is the small area mean in the transformed scale so that $\mu_i = h(\mu_i^*)$ then

$$\text{mse}(\hat{\mu}_i) = \{h'(\hat{\mu}_i^*)\}^2 \text{mse}(\hat{\mu}_i^*) \tag{2.41}$$

is the estimated MSE in the untransformed scale where $h(\cdot)$ is a smooth function and $\text{mse}(\hat{\mu}_i^*)$ is given in (2.11). In particular for model III, $\mu_i^* = \beta_0 + \beta_1 x_i + u_i$ and $\mu_i = (\mu_i^*)^3$. Therefore, $\text{mse}(\hat{\mu}_i) = 9(\hat{\mu}_i^*)^4 \text{mse}(\hat{\mu}_i^*)$. Table 2.4 contains county means for the IFact, direct estimates for WEQ in 2002, standard errors of direct estimates, county totals of survey weights, predicted values from model III(b), and root mean square error of prediction (RMSEP).

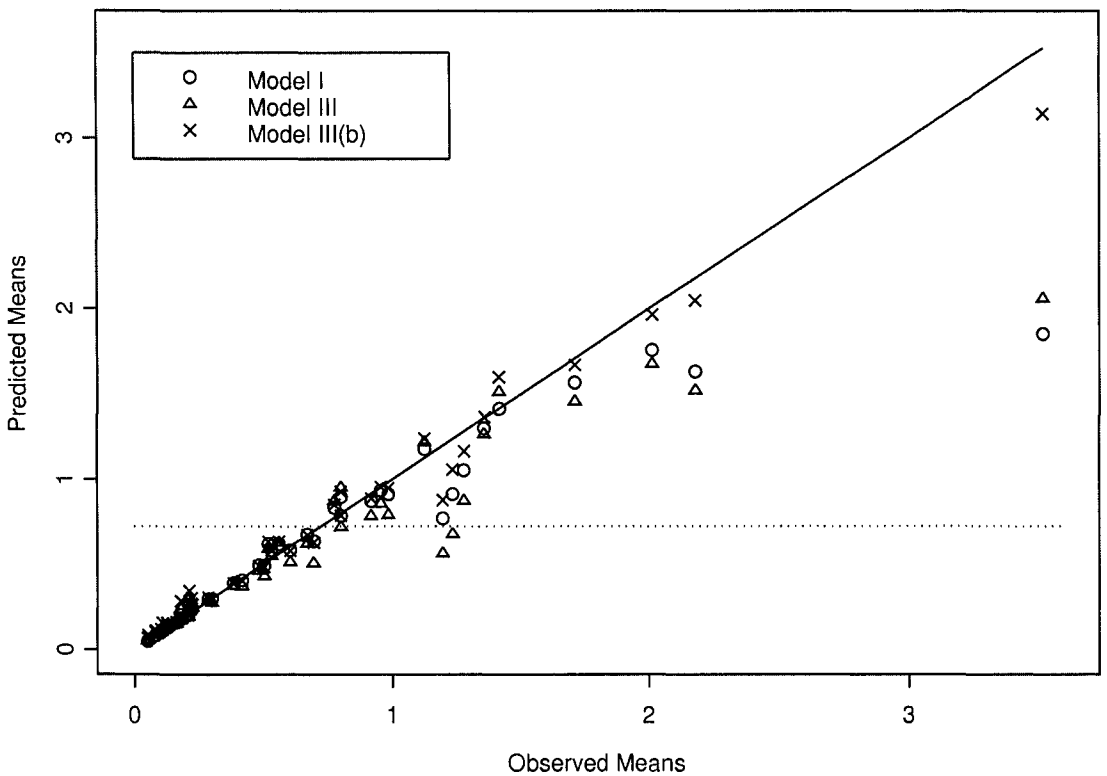

Figure 2.7 Predicted Means from Three Small Area Models.

Table 2.4: County Estimates for WEQ02

ID	n_i	Mean IFact	Direct Mean	Standard Error	Predicted Mean Model III(b)	RMSEP	Weight
3	37	46.7	0.130	0.074	0.150	0.035	1387
15	48	58.6	0.225	0.062	0.263	0.038	2462
21	43	65.6	0.984	0.396	0.949	0.266	2265
27	53	47.7	0.419	0.079	0.403	0.051	2479
33	47	52.8	0.507	0.130	0.484	0.079	2318
35	36	72.1	2.178	0.337	2.047	0.242	1748
41	48	59.1	0.221	0.067	0.270	0.041	2186
47	59	49.8	0.081	0.038	0.111	0.017	3048
59	32	53.7	0.286	0.099	0.305	0.057	1261
63	41	63.6	0.522	0.253	0.633	0.151	1822
67	32	44.9	0.163	0.067	0.164	0.033	1597
71	39	56.8	0.699	0.267	0.630	0.125	1345
73	35	54.2	0.607	0.133	0.580	0.085	1795
75	43	41.0	0.210	0.041	0.204	0.023	2369
77	40	48.6	0.110	0.078	0.157	0.037	2562
79	30	75.0	0.777	0.131	0.852	0.112	1899
83	49	57.5	0.486	0.131	0.491	0.083	2486
85	55	66.7	0.564	0.147	0.634	0.107	2241
91	45	56.1	0.918	0.118	0.889	0.092	2066
93	31	61.8	0.534	0.195	0.584	0.116	1385

continued on next page

Table 2.4: County Estimates for WEQ02 (Continued)

ID	n_i	Mean IFact	Direct Mean	Standard Error	Predicted Mean Model III(b)	RMSEP	Weight
109	55	64.3	1.708	0.144	1.670	0.138	2752
119	90	61.6	0.801	0.153	0.789	0.109	1753
129	32	58.7	0.213	0.129	0.346	0.074	1270
131	28	48.7	0.128	0.070	0.159	0.034	1232
133	65	73.1	1.360	0.206	1.364	0.177	2943
135	34	45.4	0.052	0.038	0.083	0.015	1190
141	34	81.9	3.528	0.685	3.143	0.392	1567
143	31	56.2	1.238	0.301	1.058	0.145	1511
145	47	40.9	0.050	0.042	0.067	0.016	1772
147	53	60.8	0.674	0.149	0.672	0.103	2716
149	51	80.5	1.416	0.289	1.599	0.316	3877
151	31	63.2	1.281	0.301	1.167	0.174	1823
153	39	48.5	0.300	0.083	0.293	0.047	1580
155	58	62.3	0.179	0.073	0.286	0.045	4405
157	34	44.5	0.079	0.053	0.101	0.023	2121
161	50	66.6	0.955	0.246	0.956	0.167	2423
165	35	47.5	0.104	0.039	0.118	0.018	2327
167	122	72.3	2.013	0.207	1.965	0.193	3180
169	43	54.8	0.386	0.083	0.389	0.054	1862
187	41	58.4	1.199	0.525	0.879	0.376	3011
189	32	76.3	1.127	0.264	1.238	0.208	1644

continued on next page

Table 2.4: County Estimates for WEQ02 (Continued)

ID	n_i	Mean IFact	Direct Mean	Standard Error	Predicted Mean Model III(b)	RMSEP	Weight
193	47	75.1	0.801	0.182	0.927	0.152	2319
195	47	46.5	0.183	0.056	0.185	0.030	1290
197	34	71.4	0.221	0.050	0.302	0.032	1754

A plot of the ratio of the standard error of the survey weighted mean to the estimated RMSEP from model III(b) is shown in Figure 2.8. Model III has the lowest RMSEP except for one county where the RMSEP from model I is better. The RMSEP from model III(b) are similar to the RMSEP from model III. Summary statistics for the estimated RMSEP are given in Table 2.5. Overall, model III has the lowest RMSEP with model III(b) a close competitor.

The coefficient of variation (CV) of an unbiased estimator $\hat{\mu}$ is defined as $\hat{\mu}^{-1}\{\text{Var}(\hat{\mu})\}^{0.5}$. The predicted means, RMSEP's and estimated CV's of 8 selected counties are given in Table 2.6. The estimated CV of the survey weighted mean for county 145 is 84%. The estimated CV from model I is 86% but the estimated CV from model III and model III(b) are 28% and 24% respectively. Hence, the prediction using model III or model III(b) is very effective for county 145, whereas for county 167, the estimated CV of

Table 2.5 Summary Statistics for Estimated RMSEP

Model	First Quartile	Median	Mean	Third Quartile	Range
M1	0.068	0.123	0.131	0.186	0.29
M3	0.036	0.086	0.103	0.150	0.37
M3(b)	0.037	0.084	0.110	0.151	0.38

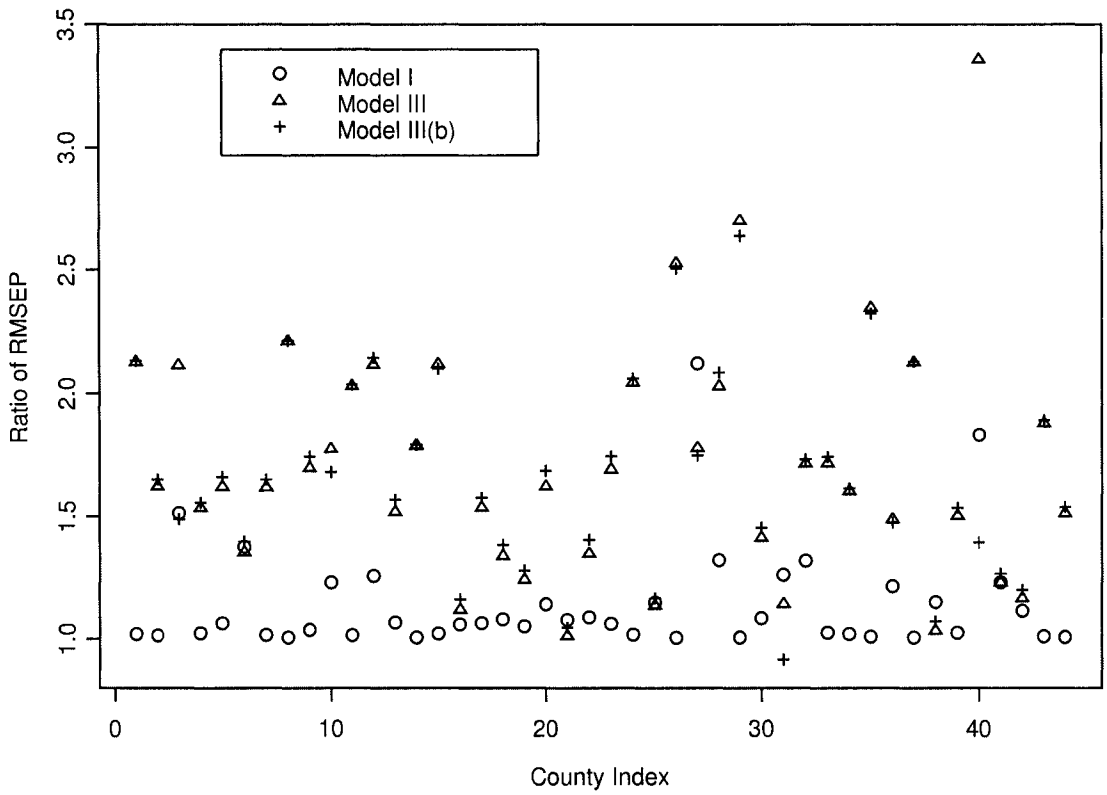

Figure 2.8 Ratio of the Standard Error of Design Weighted Mean to the Root Mean Square Error of Prediction.

small area predicted means using models are similar to the estimated CV of the survey weighted mean. Overall, the average estimated CV using model III is 19% and the average estimated CV using model III(b) is 18% which is a big improvement over the average estimated CV using the direct estimates (33%). Model III(b) RMSEP is sometimes higher than the RMSEP using model III. The mean of the differences of RMSEP between model III and model III(b) is 0.004 with a maximum difference of 0.117. So, the use of model III(b) instead of model III gives us a greater advantage of calibration with very little sacrifice of RMSEP.

2.6 Conclusions

We used the NRI data to estimate wind erosion for counties for the year 2002. County sample sizes vary from 28 to 122. Due to small sample sizes in many counties, the direct survey estimates have high standard errors. The coefficient of variation ranges from 8.4% to 84.0% with an average of 33.1%. We fit a Fay-Herriot model to estimate the small area means. A cube root transformation of the response is found to be linear with the covariate. We proposed weight adjusted small area means which calibrate to the direct state level estimates. A general approach of calibration for any smooth transformation of the response is discussed. In the final estimate, the average county CV is 17.1% with a maximum CV of 42.3%.

Table 2.6 Predicted Means and RMSEP for Eight Selected Counties

County	n_i		DE	MI	MIII	MIII(b)
		Mean	0.534	0.583	0.554	0.584
93	31	RMSEP	0.195	0.171	0.120	0.116
		CV	0.365	0.293	0.217	0.198
		Mean	0.221	0.240	0.267	0.302
197	34	RMSEP	0.050	0.049	0.033	0.032
		CV	0.225	0.205	0.123	0.107
		Mean	0.607	0.585	0.517	0.580
73	35	RMSEP	0.133	0.125	0.088	0.085
		CV	0.219	0.213	0.169	0.146
		Mean	0.210	0.206	0.199	0.204
75	43	RMSEP	0.041	0.041	0.023	0.023
		CV	0.195	0.197	0.115	0.112
		Mean	0.050	0.049	0.056	0.067
145	47	RMSEP	0.042	0.042	0.016	0.016
		CV	0.840	0.859	0.277	0.239
		Mean	0.955	0.935	0.860	0.956
161	50	RMSEP	0.246	0.202	0.165	0.167
		CV	0.257	0.216	0.192	0.174
		Mean	1.708	1.566	1.459	1.670
109	55	RMSEP	0.144	0.134	0.142	0.138
		CV	0.084	0.0853	0.0972	0.082
		Mean	2.012	1.758	1.682	1.965
167	122	RMSEP	0.207	0.180	0.200	0.193
		CV	0.103	0.102	0.119	0.098

CHAPTER 3. Small Area Estimation: A Nonparametric Approach

3.1 Introduction

Small area estimators commonly "borrow strength" from other related areas. The indirect estimators use models (explicit or implicit) that relate the small area means to the supplementary data. Various unit-level and area-level small area models are proposed in the literature (Rao, 2003). Small area models use parametric estimation procedures to relate covariates and unobserved small area means. We propose a non-parametric smoothing approach to predict the unobserved small area means. An approximation of the mean squared error (MSE) of the proposed predictor is developed and an estimator of the MSE is proposed. A limited simulation study shows that if the linearity breaks, the predictions from the non-parametric model are better compared to the predictions from a linear model. Even when the linear relationship is true, the non-parametric prediction is 'as good as' the linear prediction.

Section 3.2 introduces the kernel based non-parametric approach for small area estimations. A simulation study is conducted to check the performance of the proposed estimator. The description of the simulation study is given in Section 3.3. In Section 3.4, an application of the proposed method to estimate soil erosion due to wind is discussed. We conclude with a brief discussion in Section 3.5 and proof of the theorems are given in the appendix.

3.2 Kernel-Based Approach

Small area means are usually modeled using a mixed linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{y}$ is the vector of mean responses, X and Z are design matrices, $\mathbf{u}$ is a random vector commonly known as small area effects, and $\boldsymbol{\epsilon}$ is a vector of sampling errors. In particular, a basic area level model with one covariate can be written as

$$\begin{aligned} y_i &= \theta_i + \epsilon_i, \\ \theta_i &= \beta_0 + \beta_1 x_i + u_i, \end{aligned} \tag{3.2}$$

where x_i's are area specific covariates, θ_i's are the unobserved means, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is a vector of regression parameters, ϵ_i's are sampling errors, u_i's are area specific random effects, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $\epsilon_i \overset{ind}{\sim} N(0, D_i)$, and u_i and ϵ_i are independent (Fay and Herriot, 1979). The empirical best linear unbiased predictor of θ_i is given by

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\mathbf{x}_i^T \hat{\boldsymbol{\beta}}, \tag{3.3}$$

where $\gamma_i = (\sigma_u^2 + D_i)^{-1}\sigma_u^2$, $\hat{\gamma}_i = (\hat{\sigma}_u^2 + D_i)^{-1}\hat{\sigma}_u^2$ and $\mathbf{x}_i = (1, x_i)^T$. Assume D_i's are known. Prasad and Rao (1990) proposed an estimator of the MSE of the best linear unbiased predictor for model (3.2).

$$\text{mse}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2), \tag{3.4}$$

where

$$g_{1i}(\sigma_u^2) = (\sigma_u^2 + D_i)^{-1} D_i \sigma_u^2, \tag{3.5}$$

$$g_{2i}(\sigma_u^2) = (1 - \gamma_i)^2 \mathbf{x}_i^T \left[\sum \mathbf{x}_i^T (\sigma_u^2 + D_i)^{-1} \mathbf{x}_i\right]^{-1} \mathbf{x}_i, \tag{3.6}$$

$$g_{3i}(\sigma_u^2) = (\sigma_u^2 + D_i)^{-3} D_i^2 V(\sigma_u^2), \tag{3.7}$$

where $\hat{\sigma}_u^2$ is the REML or the method of moments estimator of σ_u^2 and $V(\hat{\sigma}_u^2)$ is the variance of $\hat{\sigma}_u^2$.

In almost every application of small area estimation, linear mixed effects models are assumed. However, the estimates could be sensitive to the linearity assumption. If the assumption of linearity between the small area mean and the supplementary information fails, borrowing strength from other areas using a linear model may not be appropriate. We propose a nonparametric model of the form

$$y_i = \theta_i + \epsilon_i, \tag{3.8}$$

$$\theta_i = m(x_i) + u_i, \tag{3.9}$$

where $i = 1, 2, ..., m$ denotes the number of small areas and $m(\cdot)$ is a smooth function. Assume $u_i \overset{iid}{\sim} (0, \sigma_u^2)$, $\epsilon_i \overset{ind}{\sim} (0, D_i)$, u_i and ϵ_i are independent and D_i's are known.

To estimate $m(x)$ we propose a Nadaraya-Watson estimator

$$\hat{m}_h(x) = \frac{\sum_i K_h(x - x_i) y_i}{\sum_i K_h(x - x_i)}, \tag{3.10}$$

where $K_h(\cdot)$ is a kernel function with bandwidth h and is of the form $K_h(u) = \frac{1}{h} K(u/h)$ with $K(\cdot)$ satisfying:

i) $K(\cdot)$ is symmetric,

ii) $K(\cdot)$ is bounded and continuous on the range of x and,

iii) $\int_{\chi} K(a)da = 1$, where $\mathcal{X}$ is the range of x.

The Nadaraya-Watson estimator (3.10) is linear in y_i and can be rewritten as:

$$\hat{m}_h(x) = \frac{1}{m} \sum_{i=1}^{m} W_{hi}(x) y_i, \tag{3.11}$$

where $W_{hi}(x) = \frac{K_h(x - x_i)}{1/n \sum_i K_h(x - x_i)}$. It is easy to show that the best predictor for small area means θ_i can be written as

$$E(\theta_i | y_i) = \tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i) \hat{m}_h(x_i), \tag{3.12}$$

where $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + D_i}$ and we assume σ_u^2 is known. In the second stage, we estimate

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\hat{m}_h(x_i), \tag{3.13}$$

where $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + D_i}$ and $\hat{\sigma}_u^2$ is a consistent estimator of σ_u^2. From the theory of kernel regression it can be shown that $\hat{m}_h(x)$ is a consistent estimator for $m(x)$ at every point of continuity $m(\cdot)$. Assuming $x_i \overset{iid}{\sim} f(x_i)$, we can prove Theorem 3.2.1.

Theorem 3.2.1 *Assume the response variable y is related to a one-dimensional predictor variable x through (3.9) and:*

(A1) $\int |K(a)|da < \infty$.

(A2) $\lim_{|a|\to\infty} aK(a) = 0$.

(A3) $Ey_i^2 < \infty$ *for all* i *and* $f(x_i) \neq 0$.

(A4) $m \to \infty$, $mh \to \infty$.

Then, at every point of continuity for $m(x)$,

$$m^{-1} \sum_{i=1}^{m} \frac{K_h(x - x_i)y_i}{\sum_i K_h(x - x_i)} \xrightarrow{p} m(x), \tag{3.14}$$

where the notation $Z_n \xrightarrow{p} Z$ *indicates that the sequence of random variables* Z_n *converges in probability to* Z.

Proof of Theorem 3.2.1 is given in Appendix A. For certain bound conditions of x_i and $K(\cdot)$ the MSE for estimating $m(\cdot)$ by $\hat{m}(\cdot)$ is obtained. An approximation of the MSE is given in Theorem 3.2.2

Theorem 3.2.2 *Assume the non-parametric model (3.9) with a one-dimensional predictor x. Define* $c_k = \int K^2(a)da$, $d_k = \int a^2 K(a)da$ *and assume the following conditions:*

(A5) $m(\cdot)$ *is continuous.*

(A6) $max_{1\leq i\leq m}|x_i - x_{i-1}| = O(m^{-1})$.

(A7) $D_i = D$ *for all* $i = 1, 2, \ldots, m$ *and* D *is finite.*

(A8) $m \to \infty$, $mh \to \infty$.

Then,

$$E[\hat{m}_h(x_i) - m(x_i)]^2 \approx (mh)^{-1}\sigma^2 c_k f^{-1}(x_i) + h^4 d_k^2 \{m^{(2)}(x_i) + 2f^{-1}(x_i)f^{(1)}(x_i)m^{(1)}(x_i)\}^2/4, \tag{3.15}$$

where $\sigma^2 = \sigma_u^2 + D$, *and for any smooth function* $g(x)$ *we write* $g^{(k)}(x_i) = \frac{\partial^k}{\partial x^k} g(x)|_{x=x_i}$.

Proof of Theorem 3.2.2 is given in Appendix A. In Theorem 3.2.2, we ignore the terms which are higher than the order of m^{-1}. From expression (3.15), the MSE of $\hat{m}(x)$ has two parts. The first part comes from the bias and the second part is related to the variance. A suitable selection of bandwidth h can compromise between the bias and the variance. A global fixed bandwidth $h \propto m^{-1/5}$ is used (Härdle, 2002) to estimate the mean function. For a more detailed discussion on bandwidth selection, see Härdle (2002). To estimate θ_i by $\hat{\theta}_i$ we propose an estimator for σ_u^2. The between area variance is estimated using the adjusted residuals. A method of moments type estimator is given by

$$\hat{\sigma}_u^2(x) = \max\{0, \frac{1}{m-1}\sum_{i=1}^{m} W_{hi}(x)\{y_i - \hat{m}(x_i)\}^2 - D\}. \tag{3.16}$$

3.2.1 Approximation of Mean Squared Error

We provide an approximation for the MSE of $\hat{\theta}_i$ and propose an estimator of the approximated MSE. The square difference of $\hat{\theta}_i$ and θ_i is divided into three terms. The first term $E(\theta_i^* - \theta_i)^2$ is due to the form of θ_i^*, where $\theta_i^* = \gamma_i y_i + (1 - \gamma_i)m(x_i)$. The second term $E(\tilde{\theta}_i - \theta_i^*)^2$ is due to the estimation of $m(x_i)$ and the third term $E(\hat{\theta}_i - \tilde{\theta}_i)^2$ is due to the estimation of σ_u^2. The result is stated in Theorem 3.2.3

Theorem 3.2.3 *Assume (A1) to (A8) are true. Further assume the following conditions:*

(A9) ϵ_i *and* u_i *are independently normally distributed.*

(A10) The bias for estimating $\hat{m}(x_i)$ can be ignored.

Then

$$MSE(\hat{\theta}_i) \approx \frac{D\sigma_u^2}{\sigma_u^2 + D} + (1-\gamma)^2 MSE[\hat{m}_h(x_i)] + D^2(\sigma_u^2 + D)^{-4} E[(y_i - m(x_i))(\hat{\sigma}_u^2 - \sigma_u^2)]^2, \tag{3.17}$$

where $MSE(\hat{m}_h(x_i))$ is given by Theorem 3.2.2.

Proof of Theorem 3.2.3 is given in Appendix A. Expanding the product term in the right side of (3.17) by a first order Taylor and then plugging the parameter estimates, an estimator of the MSE can be obtained as,

$$\text{mse}(\hat{\theta}_i) = \frac{D\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + D} + (1-\hat{\gamma})^2 \text{mse}[\hat{m}_h(x_i)] + 2D^2(\hat{\sigma}_u^2 + D)^{-3} \text{mse}(\hat{\sigma}_u^2). \tag{3.18}$$

The unobserved small area mean is estimated using a non-parametric model. An approximation of the MSE and an estimator for the approximated MSE is proposed. Moreover, if $\hat{m}(x_i) = x_i^T \hat{\beta}$ we get the same form of linear mixed effects estimators as in Rao (2003).

3.3 Simulation for the Nadaraya-Watson Estimator

The performance of the kernel based estimator (3.13) are compared to the parametric estimator (1.11) through a simulation study. A wide range of smooth functions are considered as the true mean function. Three different ratios of small area variances and error variances are considered for each mean function. The following four mean functions are used:

i) Linear: $m_1(x) = 50 + 2x$.

ii) Cubic: $m_2(x) = .01 + .2x - .005x^3$.

iii) Exponential: $m_3(x) = exp(.5x)$.

iv) Mixed Exponential: $m_4(x) = \{1 - x + exp((x-5)^2)\}10^{-6}$.

x_i's are generated from uniform (0,10) distribution, $i = 1, 2, ..., 100$. Area specific random effects are generated from $N(0, .25)$ and D_i's are .1 for the first 33 areas, .25 for

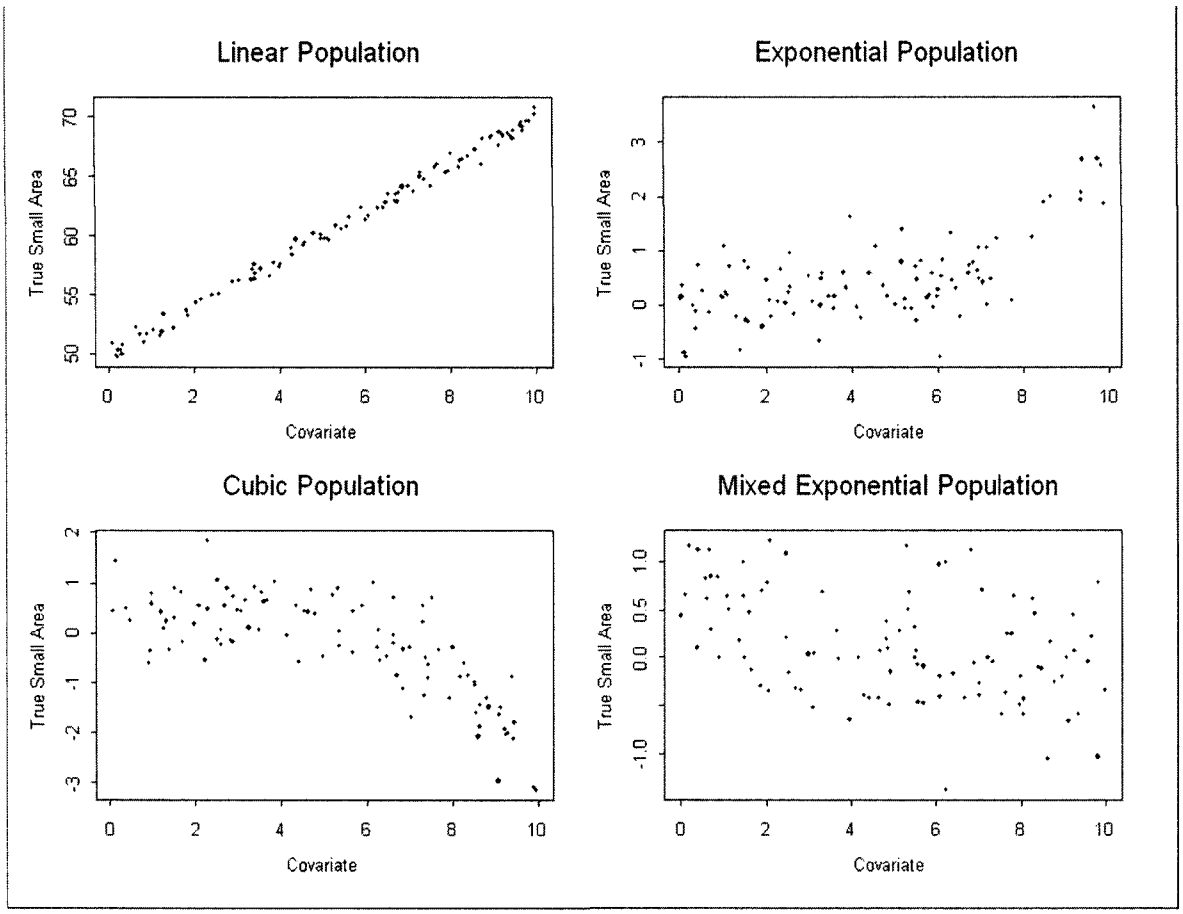

Figure 3.1 Scatter Plot for Simulated Populations.

the second 33 areas and .5 for the rest of the 34 areas. The parameters of the linear model (3.2) and the mean function of the non-parametric model (3.9) are estimated for each population. Small area means and estimates of the MSE are computed using both models. We generate the populations R times and estimate the following quantities:

i) Relative bias (RB):

$$\mathrm{RB}(\hat{\theta}_i) = \frac{1}{R}\sum_{r=1}^{R}\{\hat{\theta}_i^{(r)} - \theta_i^{(r)}\}, \tag{3.19}$$

where $\theta_i^{(r)}$ are the true means from the r^{th} population, and $\hat{\theta}_i^{(r)}$ are the estimated values of $\theta_i^{(r)}$, and $i = 1, 2, \ldots, 100$.

ii) True MSE of the estimated mean:

$$\text{MSE}(\hat{\theta}_i) = \frac{1}{R}\sum_{r=1}^{R}\{\hat{\theta}_i - \theta_i\}^2 \tag{3.20}$$

iii) Relative bias of estimated MSE:

$$\text{RB}\{\text{mse}(\hat{\theta}_i)\} = [\text{MSE}(\hat{\theta}_i)]^{-1}\frac{1}{R}\sum_{r=1}^{R}\{\text{mse}(\hat{\theta}_i)^{(r)} - \text{MSE}(\hat{\theta}_i)\}, \tag{3.21}$$

where $mse(\hat{\theta}_i)$, and $MSE(\hat{\theta}_i)$ are the estimated MSE and the true MSE for the i-th area.

iv) Coefficient of variation (CV) of the estimated MSE:

$$\text{CV}\{\text{mse}(\hat{\theta}_i)\} = [\text{MSE}(\hat{\theta}_i)]^{-1}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\{\text{mse}(\hat{\theta}_i)^{(r)} - \text{MSE}(\hat{\theta}_i)\}^2} \tag{3.22}$$

3.3.1 Simulation Results

Summary statistics of RB_i, MSE_i, $\text{RB}(\text{mse}_i)$, and $\text{CV}(\text{mse}_i)$ for 100 simulated small areas are presented in Table 3.1 through Table 3.4. Predictions from the Fay-Herriot model are denoted as FH and predictions from the non-parametric mixed effects model are denoted as NPME. Mean, square root of the estimated variance, and the first and third quartiles for 100 small area predictions are presented. Square root of the estimated variance is denoted by $\hat{V}^{1/2}$ and the first and the third quartiles are denoted by 1-st Quartile and 3-rd Quartile respectively. For the linear population in Table 3.1, predictions from the NPME model are 'as good as' the predictions from the FH model. For all other populations considered in the simulation, the NPME predictions have a smaller RB as compared to the FH predictions. True MSE using the FH model and using the NPME model are similar. The estimated MSE from the NPME model can be reduced by changing the bandwidth. However, reduction of the estimated MSE by enlarging the bandwidth will increase the relative bias of the NPME predictors. The mean for the RB(mse) using the NPME model is smaller as compared to the mean for

Table 3.1 Predictions for Linear Populations

	Model	Mean	$\hat{V}^{1/2}$	1-st Quartile	3-rd Quartile
RB	FH	0.00034	0.0021	-0.0009	0.0012
	NPME	0.00072	0.0024	-0.0011	0.0019
MSE	FH	0.132	0.079	0.080	0.163
	NPME	0.178	0.121	0.103	0.227
RB(mse)	FH	0.373	0.309	0.193	0.422
	NPME	0.257	0.168	0.144	0.366
CV(mse)	FH	7.19	6.16	3.77	8.21
	NPME	8.14	9.56	3.85	8.76

the RB(mse) using the FH model. Except for the linear population, the mean for the CV(mse) using the NPME model is smaller relative to the mean for the CV(mse) using the FH model. Therefore, for the linear population, the NPME predictions are similar to the FH predictions. However, for the populations with a nonlinear trend, the NPME predictors have a smaller bias as compared to the FH predictors.

3.4 Application to the NRI

The National Resource Inventory (NRI) is a nation-wide survey of the US non federal land. The NRI is designed to assess conditions and trends of land cover, soil properties, and related environmental resources on a yearly basis. The data were collected using a two-stage, two-phase, supplemented panel, longitudinal area sample design at the national level (Nusser and Goebel, 1997; Fuller, 2003). In some Midwestern states, soil erosion due to wind is a severe problem. It may be beneficial for the local and state governments to estimate soil loss due to wind at the local level. Due to the national level design of the NRI, sample size within one county could be as low as 5, but there might be some similarities between the adjacent counties. We should "borrow strength"

Table 3.2 Predictions for Cubic Populations

	Model	Mean	$\hat{V}^{1/2}$	1-st Quartile	3-rd Quartile
RB	FH	0.373	12.05	-0.642	0.379
	NPME	0.214	9.94	-0.481	0.144
MSE	FH	0.182	0.117	0.102	0.233
	NPME	0.142	0.091	0.068	0.180
RB(mse)	FH	6.16	5.27	3.23	7.02
	NPME	-3.91	4.59	-4.21	-1.85
CV(mse)	FH	10.64	9.11	5.58	12.14
	NPME	6.83	8.02	3.23	7.35

Table 3.3 Predictions for Exponential Populations

	Model	Mean	$\hat{V}^{1/2}$	1-st Quartile	3-rd Quartile
RB	FH	0.472	5.952	-0.550	0.483
	NPME	0.244	6.228	-0.507	0.265
MSE	FH	0.156	0.104	0.091	0.201
	NPME	0.141	0.094	0.079	0.183
RB(mse)	FH	0.982	0.842	0.519	1.123
	NPME	-1.074	1.260	-1.150	-0.507
CV(mse)	FH	5.033	4.311	2.635	5.741
	NPME	4.067	4.767	1.922	4.382

Table 3.4 Predictions for Mixed-Exponential Populations

	Model	Mean	$\hat{V}^{1/2}$	1-st Quartile	3-rd Quartile
RB	FH	0.422	9.203	-0.618	0.437
	NPME	0.216	7.442	-0.502	0.246
MSE	FH	0.129	0.092	0.077	0.170
	NPME	0.152	0.099	0.075	0.201
RB(mse)	FH	3.520	3.838	2.894	4.468
	NPME	-2.477	3.310	-2.112	-0.988
CV(mse)	FH	7.052	6.983	6.014	8.816
	NPME	5.554	6.374	2.851	6.925

from other counties with similar trends (soil properties, landscape, weather, etc.) in order to increase precision of our estimation. In this application, the soil erodibility index (IFact) is used as auxiliary information and soil loss due to wind for the year 2003[1] (WEQ03) is used as the response variable. WEQ03 is not directly observed in the field, rather it is calculated as a function of several factors. Soil erodibility, climate, slope, and land cover are just a few of those factors (Bell et al., 2003). WEQ03 is measured in ton/ha and is used as observed wind erosion in this study. The use of IFact as a predictor has several advantages. IFact is directly related with wind erosion. Higher IFact values indicate greater susceptibility to wind erosion. IFact can be obtained from the Natural Resources Conservation Service (NRCS) soil survey database available through the NRCS Soil Data Mart (SDM) for each county in the US. Since IFact is a soil characteristic, it doesn't change much over time. Table 3.5 shows a summary of observations in each county. There are 152 counties with less than 20 observations. Figure 3.2 is a scatter plot for WEQ03 and IFact. The county level mean plot for soil loss due to wind suggests a non-linear relationship among WEQ03 and IFact. This motivates

[1]The 2003 NRI data set has not yet been released for public use. All values are strictly for research purposes.

Table 3.5 Summary Statistics for Observed Counties

Total Observations	:	75573
Number of States	:	3
Number of Counties	:	276
County with Size 0	:	57
County with Size < 10	:	114
County with Size < 20	:	152

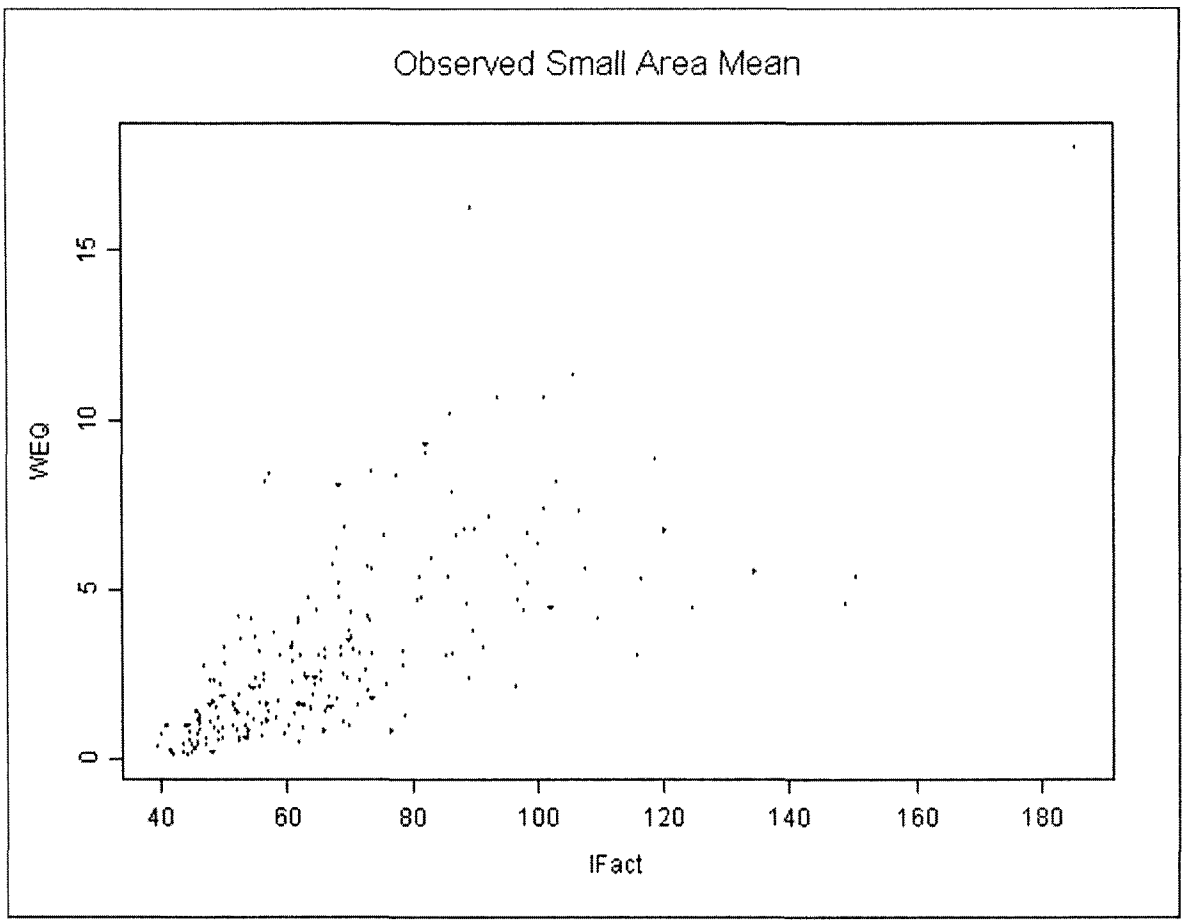

Figure 3.2 Scatter Plot of WEQ 2003 and Erodibility Index.

us to use a non-parametric small area model

$$y_i = m(x_i) + u_i + \epsilon_i, \tag{3.23}$$

where y_i's are the observed county means for WEQ03 in 2003, x_i's are means for IFact, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $\epsilon_i \overset{ind}{\sim} N(0, D_i)$, and u_i's and ϵ_i's are independent. The mean function $m(x_i)$ from model (3.23) is estimated using the Nadaraya-Watson kernel estimator. The estimates are discussed in the next subsection.

3.4.1 Estimates for Wind Erosion

Table 3.6 presents summary statistics for estimated means and estimated MSE for the counties. For presentation purposes, counties are divided into five size categories. For each size category, the number of counties in that category is given within parenthesis. The direct survey weighted estimates (DE), the small area estimates using the Fay-Herriot model (FH), and the small area estimates using the non-parametric model (NP) are given in Table 3.6. The interquartile range for each estimated value is given within parenthesis. The relative differences of the observed means and the NPME means are smaller than the relative differences of the observed means and the FH means. The estimated MSE for the NP model are smaller relative to the other two methods when county sizes are less than 20. For counties with sample sizes over 50, the estimated MSE using all three methods are similar. When county sample sizes are small, predictions from the FH model have smaller estimated MSE relative to the estimated design variance of the DE. Predictions form the NP model have smaller estimated MSE relative to the predictions from the FH model. This is not surprising as the data plot suggests a deviation from linearity. A plot for estimated county means (for the three states under study) using direct estimates is given in Figure 3.3. Plots for predicted means using the FH model and the NP model are shown in Figure 3.4, and Figure 3.5 respectively. Dark

Table 3.6 Summary Statistics for County Means and the Estimated MSE

	County Size = 1 (8)			County Size = 2-10 (52)	
	Mean	mse		Mean	mse
DE	0.93	-	DE	3.04	3.53
	(0.52, 1.19)			(0.44, 4.48)	(0.03, 2.60)
FH	3.03	0.24	FH	2.94	0.98
	(1.36, 3.77)	(0.23, 0.26)		(0.83, 5.12)	(0.07, 1.12)
NP	1.50	0.22	NP	2.99	0.77
	(1.31, 1.89)	(0.19, 0.22)		(0.49, 4.62)	(0.14, 0.99)
	County Size = 11-20 (38)			County Size > 50 (52)	
	Mean	mse		Mean	mse
DE	2.02	0.65	DE	4.02	0.82
	(0.68, 2.53)	(0.02, 0.53)		(2.08, 3.36)	(0.23, 0.95)
FH	2.48	0.18	FH	3.16	0.74
	(1.35, 2.81)	(0.10, 0.17)		(2.55, 3.66)	(0.15, 0.90)
NP	1.97	0.11	NP	3.74	0.85
	(0.73, 2.41)	(0.09, 0.14)		(2.07, 4.61)	(0.24, 1.01)

values of red imply high values of estimated means. Figure 3.3, 3.4 and 3.5 suggest that both the FH and NP predictions make the plots smoother relative to the DE.

3.5 Conclusions

We propose a non-parametric regression estimator for small area estimation. A two stage estimation technique is developed that uses the Nadaraya-Watson estimator. An approximation for the MSE and an estimator of the approximated MSE is proposed. A simulation study demonstrates the efficiency of the proposed estimator relative to its linear counterpart. The proposed estimator is applied to the NRI data set to estimate soil loss due to wind for the counties in three mid-western states in the US. Predicted values using the proposed non-parametric model have lower estimated MSE than the predicted values from a linear model.

All theorems are stated under the assumption that the sampling variances are the

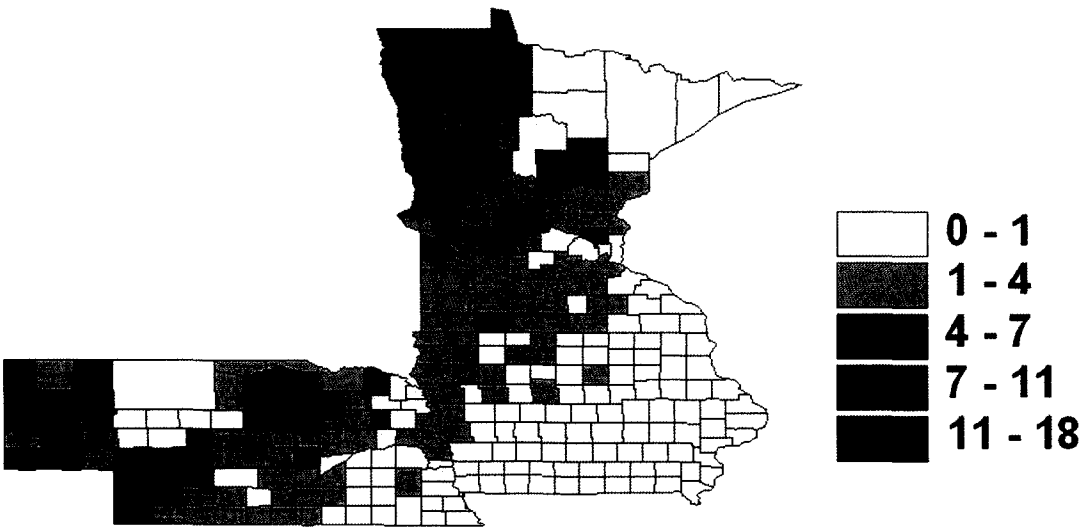

Figure 3.3 Direct County Means for WEQ 2003.

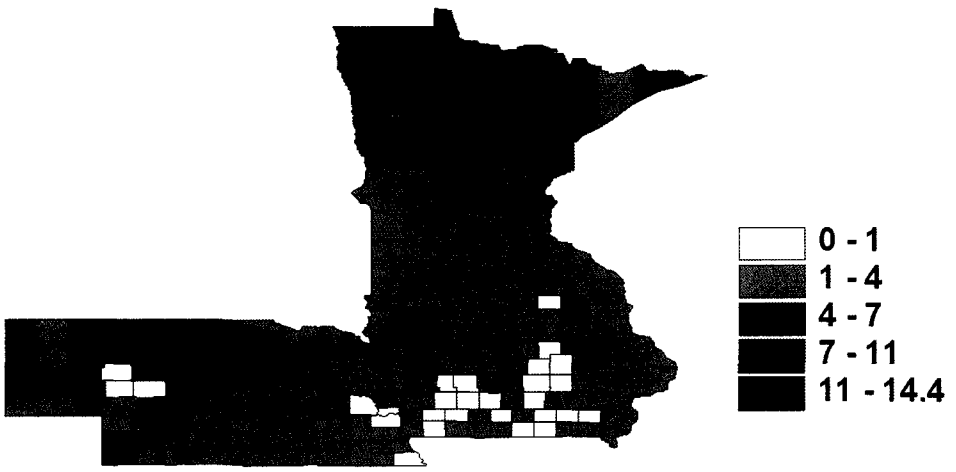

Figure 3.4 Estimates for County Means using Fay-Herriot Model.

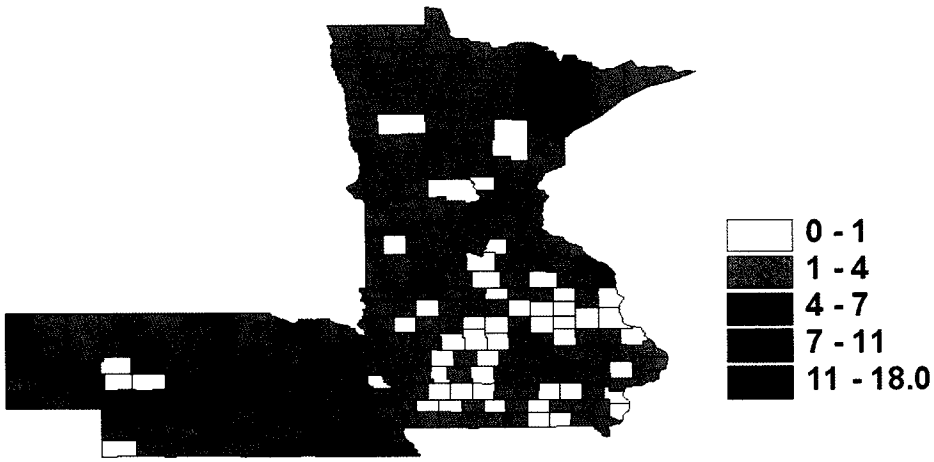

Figure 3.5 Estimates for County Means using Non-Parametric Model.

same. Work needs to be done to incorporate unequal sampling variance.

CHAPTER 4. Local Polynomial Regression

4.1 Introduction

Survey statisticians frequently provide estimates for small domains within the overall population of interest. Depending on the overall survey sample size, design-based inference methods may be inappropriate for all or some of these small domains. Survey practitioners have often resorted to model-based estimators in this case. The term "small area estimation" is used to denote this kind of estimation setting. Rao (2003) gives extensive review of the most commonly used estimators, including synthetic and composite estimators, empirical best linear unbiased predictors, empirical Bayes, and hierarchical Bayesian approaches. To date, all the approaches in use for small area estimations have relied on parametric, most often linear, modeling techniques. In this chapter, we propose a small area estimator that relies on a nonparametric model formulation.

A nonparametric approach has significant advantages over its parametric counterpart. Erroneous specification of the parametric model can result in a biased estimator. Despite several advantages of the nonparametric model, there is no substantial use of this technique in small area estimation. This is largely because of the difficulties in incorporating nonparametric mixed effect models into the estimation tools used by the survey statisticians.

Our main theoretical contributions are results on the bias and the variance of the estimated mean and the estimated variance functions for a nonparametric mixed effects model. We develop predictors of small area mean function. The theoretical properties

of the proposed estimators are studied. An optimal bandwidth selection method based on the estimated mean squared error of small area means is discussed. Our framework is expandable to the empirical Bayes estimation under a hierarchical nonparametric model assumption.

In Section 4.2, the construction of a local polynomial regression estimator for small area means is covered. The theoretical properties of the proposed estimators are discussed in Section 4.3. Section 4.4 investigates the mean squared error for the proposed small area estimator and proposes a technique for bandwidth selection. Conclusions are given in Section 4.5.

4.2 Framework for Local Polynomial Estimators

For each area $i = 1, 2, ..., n$, assume that y_i is the Horvitz-Thompson estimator (Särndal et al., 1991) of the true mean θ_i with design variance D_i. Let $\mathbf{x}_i$ be a vector of area level covariates. The basic area level small area model can be written as a special case of a linear mixed effect model,

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i, \qquad i = 1, 2, ..., n, \tag{4.1}$$

where $\boldsymbol{\beta}$ is a vector of regression parameters, u_i's are random effects and e_i's are sampling errors. Note that the design-induced error e_i accounts for within area variation and the model-induced error u_i accounts for between area variation. We also assume $e_i \overset{ind}{\sim} (0, D_i)$, $u_i \overset{iid}{\sim} (0, \sigma_u^2)$ and that they are independent. For estimation purposes, D_i is usually assumed to be known, see Rao (2003).

The linearity assumption and the assumption of homoscedastic between area variance are restrictive in many applications. A limited simulation study indicates that a violation of the linear relationship may reduce the efficiency of the current method (see Chapter 3). We consider an extension of model (4.1). We assume that y_i and $\mathbf{x}_i$ are related

through a smooth function $m(\cdot)$ and the between area variance component is also a smooth function of $\mathbf{x}_i$. Let $\mathbf{X}$ be the random vector of predictors. Thus

$$y_i = m(\mathbf{x}_i) + u_i + e_i, \qquad i = 1, 2, ..., n, \tag{4.2}$$

where $u_i|\mathbf{X} \overset{ind}{\sim} (0, v(\mathbf{x}_i))$, $e_i|\mathbf{X} \overset{ind}{\sim} (0, D_i)$, and u_i and e_i are conditionally independent. We call m the mean function and v the between area variance function. The small area mean functions

$$\theta_i(\mathbf{x}_i) = m(\mathbf{x}_i) + u_i \tag{4.3}$$

are linear combinations of the mean $m(x_i)$ and the random effects u_i. We propose an estimator of the mean function using a linear smoother. By this, we mean $\hat{\mathbf{m}} = P_1\mathbf{y}$ for some $n \times n$ matrix P_1, often referred to as the smoother matrix, and $\mathbf{y}$ and $\mathbf{m}$ denote the column vectors with elements of y_i and $m(x_i)$ respectively. Examples of linear smoothers include smoothing splines, regression splines, and local polynomial regression (Hastie and Tibshirani, 1990).

We concentrate on local polynomial regression estimators of $\mathbf{m}$ and $\mathbf{v}$; see Fan and Gijbels (1996), or Wand and Jones (1995) for an introduction. With one dimensional covariate x, we estimate $m(x)$ by fitting a p_1th-degree polynomial to the data using weighted least squares. As commonly used in the literature, we will use the weight

$$K_{h_1}(X_i - x) = h_1^{-1} K\left(h_1^{-1}(X_i - x)\right), \tag{4.4}$$

where K is a probability density function known as the kernel function and h_1 is a bandwidth (see Chapter 1) parameter. The weighted least squares estimators of m(x) is

$$\hat{m}(x) = \mathbf{e}_1^T [X_{p_1}(x)^T W_{p_1}(x) X_{p_1}(x)]^{-1} X_{p_1}(x)^T W_{p_1}(x)\mathbf{y}, \tag{4.5}$$

where

$$X_{p_1}(x) = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^{p_1} \\ 1 & X_2 - x & \cdots & (X_2 - x)^{p_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & ... & (X_n - x)^{p_1} \end{bmatrix},$$

$W_{p_1}(x) = \text{diag}_{1 \le i \le n}\{K_1(X_i - x)\}$, $\mathbf{e}_i$ denotes the unit vector of appropriate order with 1 in the ith-position, and $\text{diag}_{1 \le i \le n}\{a_i\}$ denotes the diagonal matrix with $a_1, a_2, ..., a_n$ on the diagonal. The (i, j) entry of the pth-degree local polynomial smoother for the mean function is

$$[P_1]_{ij} = \mathbf{e}_1^T [X_{p_1}(x_i)^T W_{p_1}(x_i) X_{p_1}(x_i)]^{-1} X_{p_1}(x_i)^T W_{p_1}(x_i) \mathbf{e}_j. \tag{4.6}$$

To the best of our knowledge we are the first to consider a nonparametric variance components model of the form (4.2) where one part of the variance is known from the survey design and the other part of the variance is assumed to be a smooth function of covariate x. We estimate the between area variance function by smoothing the adjusted observed residuals using a p_2th-degree polynomial

$$\hat{\mathbf{v}} = \frac{P_2(\mathbf{r}^2 - \boldsymbol{\Delta}_2)}{\mathbf{1} + P_2 \boldsymbol{\Delta}_1}, \tag{4.7}$$

where P_2 is a smoother matrix similar to P_1 except it uses a p_2th-degree polynomial and a different bandwidth h_2, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, $\mathbf{r} = \mathbf{y} - P_1\mathbf{y}$ is the observed residual,

$$\boldsymbol{\Delta}_1 = \text{diag}\{P_1 P_1^T - 2P_1\}, \tag{4.8}$$

$$\boldsymbol{\Delta}_2 = \text{diag}\{D + P_1 D P_1^T - 2P_1 D\}, \tag{4.9}$$

$D = \text{diag}_{1 \le i \le n} D_i$, $\mathbf{1}$s is a column vector of ones, $\text{diag}\{A\}$ is the column vector containing the diagonal elements of any square matrix A, and vector multiplications and divisions are elementwise.

We define a composite estimator of small area means by taking a convex combination of the survey weighted mean and the mean function $m(\cdot)$ from model (4.2),

$$\theta_i^* = \gamma_i y_i + (1 - \gamma_i) m_i, \tag{4.10}$$

where $m_i = m(x_i)$ and $\theta_i = \theta_i(x_i) = m_i + u_i$. The ratio γ_i is obtained by minimizing the mean squared error of θ_i^*. The mean squared error for θ_i^* can be written as

$$\begin{aligned} E[\theta_i^* - \theta_i]^2 &= E[\gamma_i y_i + (1-\gamma_i)m_i + m_i - u_i]^2 \\ &= E[\gamma_i(y_i - m_i) - u_i]^2 \qquad (4.11) \\ &= E[\gamma_i(e_i + u_i) - u_i]^2 \qquad (4.12) \\ &= E[\gamma_i e_i + (1-\gamma_i)u_i]^2 \qquad (4.13) \\ &= \gamma_i^2 D_i + (1-\gamma_i)^2 v_i \qquad (4.14) \\ &= \gamma_i^2(v_i + D_i) - 2\gamma_i v_i + v_i, \qquad (4.15) \end{aligned}$$

since $E[e_i u_i | \mathbf{X}] = 0$. Therefore $\gamma_i = (v_i + D_i)^{-1} v_i$ minimizes the mean squared error of θ_i^*. Assume D_i's are known. Estimate γ_i by

$$\hat{\gamma}_i = (\hat{v}_i + D_i)^{-1} \hat{v}_i. \qquad (4.16)$$

Thus, a two stage estimator for θ_i is given by,

$$\begin{aligned} \hat{\theta}_i &= \hat{\gamma}_i y_i + (1-\hat{\gamma}_i)\hat{m}_i \qquad (4.17) \\ &= \hat{m}_i + \hat{\gamma}_i(y_i - \hat{m}_i), \qquad (4.18) \end{aligned}$$

where

$$\hat{m}_i = \mathbf{e}_i^T P_1 \mathbf{y}, \qquad (4.19)$$

$$\hat{v}_i = \mathbf{e}_i^T \frac{P_2(\mathbf{r}^2 - \boldsymbol{\Delta}_2)}{\mathbf{1} + P_2 \boldsymbol{\Delta}_1} \qquad (4.20)$$

and $\hat{\gamma}_i$ is given in (4.16). For the linear mixed effects model (4.1), the plug-in estimator of γ_i gives an empirical best linear unbiased predictor for θ_i (Rao, 2003).

Remark 1. Ruppert et al. (1997) proposed similar estimators of $\hat{\mathbf{m}}$ and $\hat{\mathbf{v}}$ for the model $y_i = m(x_i) + e_i$, where $e_i \overset{ind}{\sim} (0, v(x_i))$ and $v(\cdot)$ is a smooth function. The nonparametric model we considered is different in the sense that it accounts for separate

within area variabilities and between area variability. The effect of estimating the within area sampling variance D_i is generally ignored in small area estimation (Rao, 2003).

Remark 2. If the between area variance function is assumed to be the same for all x then one should simply replace the smoother matrix P_2 by $n^{-1}\mathbf{1}\mathbf{1}^T$.

Remark 3. For theoretical convenience we consider only one covariate X; however the results can be extended to a vector of covariates.

Remark 4. We assume the random matrix $X_{p_1}(x_0)^T W_{p_1}(x_0) X_{p_1}(x_0)$ is invertible. In other words, the P_X probability that $X_{p_1}(x_0)^T W_{p_1}(x_0) X_{p_1}(x_0)$ is singular is zero. This assumption is not new in sample survey (Breidt and Opsomer, 2000) and is meaningful in small area estimation.

Remark 5. The nature of the marginal mean function $m(\cdot)$ and the variance function $v(\cdot)$ are quite different. We should be able to model high spikes on $m(\cdot)$ but usually in practice the variance function $v(\cdot)$ is more smooth. Hence, a lower degree polynomial fit is often used for $v(\cdot)$. Ruppert et al. (1997) recommended the use of $p_1 = 2$ and $p_2 = 1$ in most situations.

4.3 Theory for Local Polynomial Estimators

Estimators of the following quantities are proposed: (i) the mean function $m(\cdot)$, (ii) the between area variance function $v(\cdot)$, and (iii) the small area mean function $\theta_i(\cdot)$. In this section, the exact matrix algebraic expressions for the bias and the variance of the proposed estimators are obtained. Asymptotic approximations for the bias and the variance of the proposed estimators are also obtained under certain regularity conditions. Asymptotic approximations are useful for choosing bandwidths or evaluating the performances of the proposed estimators. Proofs are given in Appendix B.

In practice, X_i can either be fixed or random. For theoretical convenience, we assume X_is are random and $X_i \overset{iid}{\sim} f_X(\cdot)$. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ be the random vector of

predictors. Results for exact bias and covariance are conditional on $\mathbf{X}$ and therefore do not depend on a particular form of the distribution of $\mathbf{X}$. Results are derived for an interior point x_0 and for odd integers p_1 and p_2. Similar to local polynomial estimators for a fixed effects model (Fan and Gijbels, 1996), our results can easily be derived for boundary points and for even integers.

4.3.1 Exact Bias and Variance

Theorem 4.3.1 *The expectation and variance of* $\hat{m}(x_0)$ *are given by*

$$E[\hat{m}(x_0)|\mathbf{X}] = m(x_0) + \mathbf{e}_1[[X_{p_1}(x_0)^T W_{p_1}(x_0) X_{p_1}(x_0)]]^{-1} X_{p_1}(x_0)^T W_{p_1}(x_0)\mathbf{t}_{x_0}, \quad (4.21)$$

and

$$\begin{aligned} V[\hat{m}(x_0)|\mathbf{X}] &= \mathbf{e}_1[X_{p_1}(x_0)^T W_{p_1}(x_0) X_{p_1}(x_0)]^{-1} \\ & X_{p_1}(x_0)^T\{\Sigma^{(1)}_{1,x_0} + \Sigma^{(2)}_{1,x_0}\} X_{p_1}(x_0)[X_{p_1}(x_0)^T W_{p_1}(x_0) X_{p_1}(x_0)]^{-1}\mathbf{e}_1^T \end{aligned} \quad (4.22)$$

where $\mathbf{t}_{x_0} = \mathbf{m} - X_{p_1}(x_0)\boldsymbol{\beta}(x_0)$ *is the remainder from the Taylor series expansion of* $\mathbf{m}$ *around* $m(x_0)$, $\Sigma^{(1)}_{1,x_0} = diag\{K^2_{h_1}(X_i - x_0)v(X_i)\}$ *and* $\Sigma^{(2)}_{1,x_0} = diag\{K^2_{h_1}(X_i - x_0)D_i\}$ *are the weighted variance components.*

Proposition 4.3.2 *For a homoscedastic model, the expected value of the squared residuals after fitting the conditional mean function is given by*

$$E[\mathbf{r}^2|\mathbf{X}] = \{E[P_1\mathbf{y} - \mathbf{m}]|\mathbf{X}\}^2 + \sigma^2(\mathbf{1} + \boldsymbol{\Delta}_1) + \boldsymbol{\Delta}_2, \quad (4.23)$$

where $v(x_i) = \sigma^2$, $D_i = \psi$ *for all* i, *and* $\boldsymbol{\Delta}_1$ *and* $\boldsymbol{\Delta}_2$ *are defined in (4.8) and (4.9).*

Theorem 4.3.3 *Let* $G_u = diag\{Eu_i^3\}$, $T_u = diag\{Eu_i^4\}$, $G_e = diag\{Ee_i^3\}$, $T_e = diag\{Ee_i^4\}$, $P_2\mathbf{1} = P_2$, $D = diag_{1\le i\le n}\{D_i\}$, $\boldsymbol{\psi} = (\psi_1, ..., \psi_n)^T$ *and* $\Sigma = diag_{1\le i\le n}\{v(x_i)\}$. *The expectation and variance of* $\hat{\mathbf{v}}$ *are given by*

$$\begin{aligned} & E[(\hat{\mathbf{v}} - \mathbf{v})|\mathbf{X}] = \\ & \left[(P_2 - I)\mathbf{v} + P_2\{\mathbf{b}^2 + diag\{P_1\Sigma P_1^T - 2P_1\Sigma\}\} - P_2\boldsymbol{\Delta}_1\mathbf{v}\right]/\{\mathbf{1} + P_2\boldsymbol{\Delta}_1\}, \end{aligned} \quad (4.24)$$

and

$$
\begin{aligned}
Cov[\hat{\mathbf{v}}|\mathbf{X}] = {} & P_2[\{(P_1 - I) \odot (P_1 - I)\}(T - 3V^2)\{(P_1 - I) \odot (P_1 - I)\}^T \\
& +2diag\{\mathbf{b}\}(P_1 - I)G\{(P_1 - I) \odot (P_1 - I)\}^T \\
& +2\{(P_1 - I) \odot (P_1 - I)\}G(P_1 - I)diag\{\mathbf{b}\} \\
& +2\{(P_1 - I)V(P_1 - I)^T\} \odot \{(P_1 - I)V(P_1 - I)^T\} \\
& +4\{(P_1 - I)V(P_1 - I)^T\} \odot (\mathbf{b}\mathbf{b}^T)]P_2^T/\{(\mathbf{1} + P_2\boldsymbol{\Delta}_1)(\mathbf{1} + P_2\boldsymbol{\Delta}_1)^T\}
\end{aligned}
\tag{4.25}
$$

where $\mathbf{b} = E[\hat{\mathbf{m}} - \mathbf{m}|\mathbf{X}]$ *is the bias due to the estimation of the mean,* $G = G_u + G_e$, $T = T_u + T_e$ *and* $\odot$ *denotes element wise matrix multiplication.*

4.3.2 Asymptotics for Local Polynomial Estimators

The exact bias and variance expressions involve unknown quantities. Approximations of the bias and the variance are required for most applications. Asymptotic approximations are derived under certain regularity assumptions about the nature of $f_X(\cdot)$, $m(\cdot)$, $v(\cdot)$ and D_i. Most of these assumptions are standard for local polynomial regression (Fan and Gijbels, 1996).

Theorem 4.3.4 *Assume the following:*

(A1) x_0 *is an interior point in the* $\mathbf{X}$ *space.*

(A2) $f_X(x_0) \geq 0$.

(A3) There exists a $\delta_1 \geq 0$ *such that* $f_X(\cdot)$, $m^{(p_1+1)}(\cdot)$, *and* $v(\cdot)$ *are continuous and* $m^{(p_1+2)}(\cdot)$ *is bounded on* $N_{\delta_1}(x_0)$.

(A4) $n^{-1}\sum_i D_i$ *and* $n^{-1}\sum_i D_i^2$ *are bounded.*

(A5) p_1 *is an odd integer.*

(A6) $h_1 \to 0$ *and* $nh_1 \to \infty$ *as* $n \to \infty$.

The asymptotic bias and variance of $\hat{m}(x_0)$ *are given by*

$$
Bias[\hat{m}(x_0)|\mathbf{X}] = \mathbf{e}_1^T S_1^{-1}\mathbf{c}_{p_1}\frac{1}{(p_1+1)!}m^{(p_1+1)}(x_0)h_1^{p_1+1} + o_P(h_1^{p_1+1}), \tag{4.26}
$$

and

$$Var[\hat{m}(x_0)|\mathbf{X}] = \mathbf{e}_1^T S_1^{-1} S_1^* S_1^{-1} \mathbf{e}_1 \frac{v(x_0) + n^{-1}\sum_i D_i}{f_X(x_0) n h_1} + o_P(n^{-1}h_1^{-1}) \tag{4.27}$$

where $\mathbf{c}_{p_1} = (\mu_{p_1}, ..., \mu_{2p_1+1})^T$, $S_1 = [\mu_{j+l}]_{0\le j+l\le p_1}$, $S_1^* = [\nu_{j+l}]_{0\le j+l\le p_1}$, $\mu_j = \int u^j k(u)du$, *and* $\nu_j = \int u^j K^2(u)du$.

Theorem 4.3.5 *Assume that (A1) - (A6) hold. In addition, assume the following:*

(A7) There exists a $\delta_2 > 0$ *such that* $f_X(\cdot)$ *and* $v^{(p_2+1)}(\cdot)$ *are continuous and* $v^{(p_2+2)}(\cdot)$ *is bounded on* $N_{\delta_2}(x_0)$.

(A8) $h_2 \to 0$, $nh_2 \to \infty$, *as* $n \to \infty$.

(A9) p_2 *is an odd integer.*

(A10) $h_1^{2(p_1+1)} + (nh_1)^{-1} = o(h_2^{p_2+1})$.

(A11) $Ee_i^4 = \kappa_i$ *where* $n^{-1}\sum \kappa_i = O(1)$, $n^{-1}\sum \kappa_i^2 = O(1)$ *and* k_i *are known from the sampling design.*

(A12) Skewness in both error components can be ignored. i.e., $Ee_i^3 = 0$ *and* $Eu_i^3 = 0$.

The asymptotic bias and variance of $\hat{v}(x_0)$ *are given by*

$$Bias[\hat{v}(x_0)|\mathbf{X}] = \mathbf{e}_1^T S_2^{-1} c_{p_2} \frac{1}{(p_2+1)!} v^{(p_2+1)}(x_0) h_2^{p_2+1} + o_p(h_2^{p_2+1}), \tag{4.28}$$

and

$$\begin{aligned} &Var[\hat{v}(x_0)|\mathbf{X}] = \\ &\mathbf{e}_1^T S_2^{-1} S_2^* S_2^{-1} \mathbf{e}_1 [\{f_X(x_0)\}^{-1}\{\eta(x_0) + \bar{\eta}_\pi - 2v(x_0)\bar{\psi}\}](nh_2)^{-1} + o_p(nh_2)^{-1} \end{aligned} \tag{4.29}$$

where $\bar{\eta}_\pi = \bar{\kappa} - (\bar{\psi})^2$, $\mathbf{c}_{p_2} = (\mu_{p_2}, ..., \mu_{2p_2+1})^T$, $S_2 = [\mu_{j+l}]_{0\le j+l\le p_2}$, $S_2^* = [\nu_{j+l}]_{0\le j+l\le p_2}$, $\mu_j = \int u^j k(u)du$, *and* $\nu_j = \int u^j K^2(u)du$, *and* $\eta(x_i) = Eu_i^4 - (Eu_i^2)^2$.

Remark 6. Assumption (A10) is satisfied if $p_1 = p_2$ and for any optimal selection of bandwidth. Ruppert et al. (1997) used the same assumption for variance function estimators of nonparametric fixed effect models.

Remark 7. D_i's are variances of county means. For many survey designs, the assumptions about D_i and e_i are valid (Fuller, 2006).

Remark 8. We are smoothing the observed residuals, not the true residuals. Asymptotically $\hat{v}(x_0)$ behaves like a local polynomial smooth of the true residuals (Theorem 4.3.5). There is no loss in asymptotic efficiency of $\hat{\mathbf{v}}$ due to the estimation of $\hat{\mathbf{m}}$.

4.4 Approximation of Mean Squared Error

We provide an approximation for the MSE of $\hat{\theta}_i$ using the formulas derived in Section 4.3.2. Let

$$\tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i)\hat{m}_i \tag{4.30}$$

and

$$\theta_i^* = \gamma_i y_i + (1 - \gamma_i)m_i. \tag{4.31}$$

We write

$$E(\hat{\theta}_i - \theta_i)^2 = E(\theta_i^* - \theta_i)^2 + E(\tilde{\theta}_i - \theta_i^*)^2 + E(\hat{\theta}_i - \tilde{\theta}_i)^2 + E(\hat{\theta}_i - \tilde{\theta}_i)(\tilde{\theta}_i - \theta_i^*), \tag{4.32}$$

since the expected values for the other product terms vanish (see Appendix B). The first two terms on the right side of (4.32) have similar expressions to those in Prasad and Rao (1990). However, the last two terms are not tractable in general. We approximate the last two terms by the Taylor series expansions. More formally, Theorem 4.4.1 can be shown.

Theorem 4.4.1 *Assume that (A1) - (A12) hold. Assume* $u_i \overset{ind}{\sim} N(0, v_i)$ *and* $e_i \overset{ind}{\sim} N(0, D_i)$. *Then*

$$E[\hat{\theta}_i - \theta_i|\mathbf{X}]^2 = g_{1i}(v_i) + g_{2i}(v_i, m_i) + g_{3i}(v_i) + g_{4i}(v_i) + O_P(a_{nh}), \tag{4.33}$$

where

$$g_{1i}(v_i) = (v_i + D_i)^{-1} v_i D_i, \tag{4.34}$$

$$g_{2i}(v_i, m_i) = (1-\gamma_i)^2 MSE(\hat{m}_i), \tag{4.35}$$

$$g_{3i}(v_i) = \{b_i^2 + v_i(1+\mathbf{\Delta}_{1i}) + \mathbf{\Delta}_{2i}\}(v_i + D_i)^{-4} D_i^2 MSE(\hat{v}_i), \tag{4.36}$$

$$g_{4i}(v_i) = (D_i + v_i)^{-3} D_i^2 \{b_i^2 + v_i(1+\mathbf{\Delta}_{1i}) + \mathbf{\Delta}_{2i}\} Bias(\hat{v}_i), \tag{4.37}$$

and $a_{nh} = max\{(nh_2)^{-3/2}, h_2^{2p_2+2}\}$. *Asymptotic expressions for* $g_{2i}(v_i)$, $g_{3i}(v_i)$, *and* $g_{4i}(v_i)$ *are given by*

$$\begin{aligned} & g_{2i}(v_i, m_i) \\ &= (1-\gamma_i)^2 \left[\mathbf{e}_1^T S_1^{-1} \mathbf{c}_{p_1} \mathbf{c}_{p_1} S_1^{-1} \mathbf{e}_1 \{(p_1+1)!\}^{-2} \{m^{(p_1+1)}\}^2(x_i) h_1^{2p_1+2} \right. \qquad (4.38) \\ &\quad \left. + \mathbf{e}_1^T S_1^{-1} S_1^* S_1^{-1} \mathbf{e}_1 \{v(x_0) + n^{-1} \sum_i D_i\} \{f_X(x_0) nh_1\}^{-1} \right] + o_p(b_{nh}), \qquad (4.39) \end{aligned}$$

$$\begin{aligned} & g_{3i}(v_i) \\ &= (v_i + D_i)^{-3} D_i^2 \left[\mathbf{e}_1^T S_2^{-1} c_{p_2} c_{p_2}^T S_2^{-1} \mathbf{e}_1^T \{(p_2+1)!\}^{-2} \{v^{(p_2+1)}(x_0)\}^2 h_2^{2(p_2+1)} \right. \qquad (4.40) \\ &\quad \left. + \mathbf{e}_1^T S_2^{-1} S_2^* S_2^{-1} \mathbf{e}_1 [\{f_X(x_0)\}^{-1} \{\eta(x_0) + \bar{\eta}_\pi - 2v(x_0)\bar{\psi}\}](nh_2)^{-1}] + o_p(c_{nh}), \end{aligned}$$

$$g_{4i}(v_i) = (v_i + D_i)^{-2} D_i^2 \mathbf{e}_1^T S_2^{-1} c_{p_2} \{(p_2+1)!\}^{-1} v^{(p_2+1)}(x_0) h_2^{(p_2+1)} + o_P(h_2^{p_2+1}), \tag{4.41}$$

where $b_{nh} = max\{h_1^{2p_1+2}, (nh_1)^{-1}\}$, *and* $c_{nh} = max\{h_2^{2p_2+2}, (nh_2)^{-1}\}$.

4.4.1 Bandwidth Selection for Local Polynomial Estimators

An important issue is the choice of bandwidth parameters, h_1, and h_2. Local optimal bandwidths and global optimal bandwidths are common in practice (Fan and Gijbels, 1996). We provide a methodology for local optimal bandwidths selection. Ideally, the local optimal bandwidths should minimize the MSE of the small area predicted means. Thus,

$$(h_1^{opt}, h_2^{opt})_i = \text{argmin}_{h_1,h_2} \text{MSE}(\hat{\theta}_i | \mathbf{X}), \tag{4.42}$$

subject to

$$h_1^{2p_1+2} + (nh_1)^{-1} = o_p(h_2^{p_2+1}), \tag{4.43}$$

where $MSE[\hat{\theta}_i|\mathbf{X}]$ is given in (4.33). Finding (h_1^{opt}, h_2^{opt}) by minimizing (4.42) is difficult to do in practice because the effects of h_1 on the MSE of $\hat{\theta}_i$ are of second order. Using Theorem 4.3.4 and Theorem 4.3.5, we propose the following strategy:

1. Select an asymptotically optimal bandwidth, h_1^{opt}, for the estimation of the mean function $m(x_i)$ by minimizing $\text{MSE}[\hat{m}(x_i)|\mathbf{X}]$. One can use any bandwidth selection strategy described in Section 1.2.4.

2. Find the residuals

$$\hat{y}_i - \hat{m}_{h_1^{opt}}(x_i) \tag{4.44}$$

using the asymptotic optimal bandwidth h_1^{opt}.

3. Assume the observed residuals are true residuals. Apply the same bandwidth selector as in Step 1 to the observed squared residuals from Step 2 and obtain h_2^{opt}. By Remark 8, the proposed bandwidth selector will produce asymptotically optimal bandwidths (Ruppert et al., 1997).

4.5 Conclusions

A nonparametric mixed effects model is considered for small area estimation. Local polynomial estimators for both the mean function and the between area variance function are proposed. The between area variance function is estimated by smoothing the observed residuals. Theoretical properties of the proposed estimators are studied. A shrinkage estimator using the direct mean and the local polynomial estimator is proposed for the small area mean function. Asymptotic approximation for the MSE of the proposed estimator of the small area mean is derived. An asymptotic optimal bandwidth selection technique is discussed.

CHAPTER 5. Small Area Estimation Using Imputed Values

5.1 Introduction

Survey statisticians frequently encounter small area estimation (SAE) problems. In small area estimation problems, estimates are sought for a domain with a "small" or "moderate"' sample size. Because of this small sample size, direct domain estimates have low precision. Estimation approaches that "borrow strength" from similar areas using explicit or implicit models are described in Rao (2003). In almost all large surveys, some form of imputation is used. Several approaches have been taken to produce a valid estimate and its variance when imputed data are present. For example see Rao and Shao (1992) and Särndal (1992). Not much work has been done to consider the effects of imputation on SAE.

The National Resources Inventory (NRI) survey collects annual data on US non-federal land. Among other variables, the C factor (a variable highly related with soil erosion) is recorded for each selected sample point for which erosion is to be calculated. In this work, we will estimate the average C factor for each county in Iowa for the year 2002. In practice, a number of variables including erosion would be estimated, but here we study only the C factor. The NRI collects data through a supplemented panel design, where a fixed panel (core) is observed every year. Although the entire core panel is usually observed in each year, only a random sample of the core was observed in the year 2002. The unobserved part of the core is imputed using a hot-deck type single imputation procedure.

If we consider a small area model for the C factor, then the design variances of county means are required and these variances depend on the imputation procedure. The sampling errors from two different counties are not independent because of imputation. We will (i) estimate the sampling error covariance matrix adjusted for imputation, (ii) fit a multivariate area-level model using the estimated sampling error covariance, and (iii) contrast the predicted means from the data analysis using imputed values with the predicted means from an analysis of the first phase sample. To estimate the sampling error covariance matrix we fit a regression model within each imputation cell which closely matches the imputation procedure used in the NRI.

In Section 5.2 we describe the design of the NRI survey and the current imputation procedure. In Section 5.3 we describe small area models for county level means and propose a method for estimating the sampling error covariance matrix. Results and findings are given in Section 5.4 and conclusions are in Section 5.5.

5.2 The NRI Survey

The NRI is a longitudinal survey conducted by the US Department of Agriculture's (USDA) Natural Resources Conservation Service (NRCS) in cooperation with the Iowa State University, Center for Survey Statistics and Methodology (CSSM). The survey was designed to assess conditions and trends for land cover, soil, water, and related natural resources on non-federal lands in the United States. The NRI sample is a stratified two-stage area sample. The primary sampling units (PSU) are the divisions of the US land defined by the Public Land Survey System (PLS). Three sample points are selected within most PSU's according to a restricted randomization procedure, see Nusser and Goebel (1997) for details. Since 2000, the full panel structure of the NRI has been replaced by a two-phase supplemented panel sampling design in which the 1997 NRI segments serve as a first phase, and each year a partially overlapping panel

is selected through a stratified sampling design as a second phase. The annual sample includes approximately 42,000 "core" segments that are to be observed every year. An additional 30,000 segments are selected from the remaining 258,000 PSUs each year to form a supplemental sample.

5.2.1 Variables of Interest for Soil Erosion

Data collection to estimate soil loss is one of the major focuses for the NRI. Soil loss is estimated using the universal soil loss equation (USLE). The USLE is not collected directly, rather it is calculated using several factors related to soil properties and farming practices. The cover and crop management factor (C factor) is one important factor in the USLE. Other factors are the soil support factor (P factor), the rainfall factor (R factor), the soil erodibility factor (K factor), slope length, and slope percent. R factor, P factor, and K factor can be obtained from administrative records (NRCS, soil science data base). Slope percent and slope length are directly observed in the field. In this chapter, we focus only on the C factor, as it is observed for each selected sample point and unobserved values are imputed for all points that require USLE. The C factor in the USLE measures the combined effect of all the interrelated cover and crop management variables. It is defined as the ratio of soil loss from land maintained under specified conditions to the corresponding loss from continuous bare land. The value of C is usually expressed as an annual value for a particular cover and crop management system but is calculated from the soil loss ratios for short periods of time within which cover and management effects are relatively uniform. The soil loss ratios are combined in proportion to the applicable percentages of erosion index (EI) to derive annual C values. Broad use of the land and land cover use are related to the C factor. Also slope percent, irrigation practice, rotation of crop and Cowardin classification of wetland systems are used to determine the value for the C factor (see Rosewell (1993)).

5.2.2 Imputation Procedure for the C Factor

The C factor was only observed on half of the core (P00.1) in 2002. The observed set is a systematic subsample where the original sample was ordered geographically. The missing values were imputed with a single imputed value for each missing value. Imputation cells were created using broad use, land cover use, slope percent, irrigation type, and Cowardin wetland classification. Then a donor is chosen from the same imputation cell as the recipient. Finally, the missing value is imputed using a ratio adjusted donor value based on values for the years 2001 and 2003. Let $C_{2002,R}$ and $C_{2002,D}$ denote the C factor in the year 2002 for the recipient point and for the donor point respectively. Then, except for some special cases (mainly based on land cover use, see Bell et al. (2003) for details) the missing value $C_{2002,R}$ is calculated as

$$\begin{aligned} C_{2002,R} &= 0.5\{(\alpha_1 C_{2001,R} + \alpha_2 C_{2003,R}) - (\alpha_1 C_{2001,D} + \alpha_2 C_{2003,D}) + C_{2002,D}\} + \\ &\quad 0.5\{(\alpha_1 C_{2001,R} + \alpha_2 C_{2003,R})(\alpha_1 C_{2001,D} + \alpha_2 C_{2003,D})^{-1} C_{2002,D}\} \end{aligned} \tag{5.1}$$

where $\alpha_1 = \alpha_2 = 1/2$.

By definition, imputation cells and small areas (counties) are not the same. For a missing observation in county i the donor can come from a different county. Hence, estimated county means are not independent. If we assume that observed values in two distinct counties are independent then the correlation between two county estimates is due to the imputed values.

5.3 Estimator of the Mean C Factor

Let the finite population $\mathcal{F}$ with index set $U = \{1, 2, ..., N\}$ be divided into m subdivisions (counties) $\{U_i\}_{i=1}^m$. Let A_1 be a set of indexes for a sample of size n from the population, let A_r be a set of indexes for the r observed values in A and let A_m

be the set of indexes of the $n - r$ unobserved values. Assume that A_1 can be divided into G poststrata (imputation cells) such that $A_1 = U_{g=1}^{G} A_{1g}$, $A_r = U_{g=1}^{G} A_{rg}$, and $A_m = U_{g=1}^{G} A_{mg}$. Further, let y_{igk} be the k^{th} C factor in county i and imputation cell g, and let π_{igk} be the probability that y_{igk} is selected for A_1. An estimator of the mean C factor for county i, denoted by $\bar{y}_{i..}$, is

$$\bar{y}_{i..} = N_{i+}^{-1}\{\sum_{g=1}^{G}\sum_{k\in A_{rig}} w_{igk}y_{igk} + \sum_{g=1}^{G}\sum_{k\in A_{mig}} w_{igk}z_{igk}\}, \tag{5.2}$$

where $N_{i+} = \sum_g N_{ig}$ is the population size of county i, $w_{igk} = \pi_{igk}^{-1}$ and z_{igk} are imputed values.

5.3.1 Multivariate Small Area Model

If we have reasonable county level covariates then the county covariate mean and estimated means, $\bar{y}_{i..}$, can be used to estimate the parameters of a small area model. Let μ_i be the true unobserved mean C factor for county i. Then with $\mathbf{y} = (\bar{y}_{1..}, ..., \bar{y}_{m..})^T$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)^T$, $\mathbf{u} = (u_1, ..., u_m)^T$, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ we write,

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\mu} + \mathbf{e}, \\ \boldsymbol{\mu} &= X\boldsymbol{\beta} + \mathbf{u} \end{aligned} \tag{5.3}$$

where X is a $m \times p$ matrix of covariates, $\mathbf{e}$ is the sampling error, and $\mathbf{u}$ is a vector of area-specific random variables. We assume that

$$(\mathbf{u}^T, \mathbf{e}^T) \sim N(\mathbf{0}, \begin{bmatrix} \sigma^2 I & \mathbf{0} \\ \mathbf{0} & \Sigma_{ee} \end{bmatrix}). \tag{5.4}$$

Then the dispersion matrix for $\mathbf{y}$ is $\Sigma_{zz} = \sigma^2 I + \Sigma_{ee}$. Assuming σ^2 and Σ_{zz} are known, the best linear unbiased predictor (BLUP) of $\boldsymbol{\mu}$ is,

$$\tilde{\boldsymbol{\mu}} = X\tilde{\boldsymbol{\beta}} + \sigma^2\Sigma_{zz}^{-1}(\mathbf{y} - X\tilde{\boldsymbol{\beta}}), \tag{5.5}$$

where

$$\tilde{\boldsymbol{\beta}} = (X^T \Sigma_{zz}^{-1} X)^{-1} X^T \Sigma_{zz}^{-1} \mathbf{y}. \tag{5.6}$$

Since σ^2 is unknown in practice, the empirical BLUP (EBLUP) estimator of $\boldsymbol{\mu}$ can be obtained by substituting the estimated σ^2 in (5.5) and (5.6). The explicit form is

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} + \hat{\sigma}^2 \tilde{\Sigma}_{zz}^{-1} (\mathbf{y} - X\hat{\boldsymbol{\beta}}), \tag{5.7}$$

where

$$\hat{\boldsymbol{\beta}} = (X^T \tilde{\Sigma}_{zz}^{-1} X)^{-1} X^T \tilde{\Sigma}_{zz}^{-1} \mathbf{y}, \tag{5.8}$$

$\tilde{\Sigma}_{zz} = \hat{\sigma}^2 I + \Sigma_{ee}$ and Σ_{ee} is assumed to be known.

The hot deck imputation procedure used in the NRI will change the covariance structure of the small area model (5.3). In section (5.3.2) a methodology to estimate the covariance matrix is proposed, given a data set with imputed values.

Datta et al. (1992) obtained a second-order approximation for the covariance matrix $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ as,

$$\text{MSE}(\hat{\boldsymbol{\mu}}) \approx \mathbf{G}_1(\sigma^2) + \mathbf{G}_2(\sigma^2) + \mathbf{G}_3(\sigma^2) \tag{5.9}$$

where

$$\mathbf{G}_1(\sigma^2) = \Sigma_{ee} - \Sigma_{ee} \Sigma_{zz}^{-1} \Sigma_{ee}, \tag{5.10}$$

$$\mathbf{G}_2(\sigma^2) = \Sigma_{ee} \Sigma_{zz}^{-1} X (X^T \Sigma_{zz}^{-1} X)^{-1} X^T \Sigma_{zz}^{-1} \Sigma_{ee}, \tag{5.11}$$

and

$$\mathbf{G}_3(\sigma^2) = \Sigma_{ee} K^3 \Sigma_{ee} V(\hat{\sigma}^2) \tag{5.12}$$

with $K = \Sigma_{zz}^{-1} - \Sigma_{zz}^{-1} X (X^T \Sigma_{zz}^{-1} X)^{-1} X^T \Sigma_{zz}^{-1}$ and $V(\hat{\sigma}_u^2)$ is the variance of $\hat{\sigma}_u^2$. The first term in expression (5.9) is the prediction covariance matrix if all parameters are known. The second term is due to the uncertainty of estimating β and the third term is due to the uncertainty of estimating σ^2.

A second-order approximation to the estimator of the MSE($\hat{\boldsymbol{\mu}}$) can be obtained by replacing σ^2 by its estimator $\hat{\sigma}^2$ and by accounting for the bias associated with $\mathbf{G}_1(\hat{\sigma}^2)$. If $\hat{\sigma}^2$ is an unbiased estimator of σ^2, then the MSE($\hat{\boldsymbol{\mu}}$) can be estimated by

$$\text{mse}(\hat{\mu}) = \mathbf{G}_1(\hat{\sigma}^2) + \mathbf{G}_2(\hat{\sigma}^2) + 2\mathbf{G}_3(\hat{\sigma}^2). \tag{5.13}$$

5.3.2 Estimator of the Covariance

An estimator of $V(\bar{y}_{i..}|\mathcal{F})$ that ignores the fact that some values are imputed may seriously under estimate the true variance, see Särndal (1992). To estimate the variance for an imputed data set, we must consider the response mechanism, the survey design, and the imputation model. The response mechanism defines the nature of the response given the sample (which is defined by the supplemented panel design in the NRI), the survey design defines how the sample is chosen, and the imputation model (implicit or explicit) defines how the missing values are imputed.

For the NRI survey, missing values are imputed using imputation cells. We define an explicit imputation model (ξ) which closely matches the model used in the NRI. The randomness due to the imputation mechanism used in the NRI is approximated using the explicit cell-model ξ. We define cell-model ξ by

$$\xi : y_{igk} = \mathbf{q}_{igk}^T \boldsymbol{\gamma}_g + \epsilon_{igk}, \tag{5.14}$$

where $\mathbf{q}_{igk}$ are unit level covariates, $\boldsymbol{\gamma}_g$ is a vector of regression parameters in cell g, ϵ_{igk}'s are random errors and $\epsilon_{igk} \overset{ind.}{\sim} (0, \sigma_g^2)$. Assume the conditional expectations of the predicted values from the imputation model ξ closely match the conditional expectations of the original imputed values. The randomness associated with choosing a residual from the imputation model (5.14) will approximate the randomness in the current NRI hot deck imputation procedure. Define

$$\hat{y}_{igk} = \begin{cases} y_{igk}, & k \in A_{rig} \\ \mathbf{q}_{igk}^T \hat{\boldsymbol{\gamma}}_g, & k \in A_{mig} \end{cases}, \tag{5.15}$$

and

$$y^*_{igk} = \begin{cases} y_{igk}, & k \in A_{rig} \\ z_{igk}, & k \in A_{mig} \end{cases}, \tag{5.16}$$

where $\hat{\gamma}_g$ are ordinary least squares estimators of γ_g, z_{igk} are original imputed values, and A_{rig}, and A_{mig} are the sets of indexes for the observed units and the missing units, respectively, in county i and imputation cell g. The estimated standard deviations from model ξ are used to estimate the design variance of imputed county means in Proposition 5.3.1.

Proposition 5.3.1 *Let n_{1ig} be the number of sampled elements for county i and imputation cell g, let n_{rig} be the number of observed elements, and let $n_{mig} = n_{1ig} - n_{rig}$. Let $n_{1i+} = \sum_{g=1}^{G} n_{1ig}$, $n_{ri+} = \sum_{g=1}^{G} n_{rig}$, $n_{mi+} = \sum_{g=1}^{G} n_{mig}$, $n_{1+g} = \sum_{i=1}^{m} n_{1ig}$, $n_{r+g} = \sum_{i=1}^{m} n_{rig}$, and $n_{m+g} = \sum_{i=1}^{m} n_{mig}$. Let $p_g = n_{1+g}^{-1} n_{r+g}$ be the subsampling rate in the imputation cell g. Assume*

(A1) Common nonresponse probability within imputation cell g.

(A2) If there are no missing values, the county estimates are independent.

(A3) Missing values are imputed through a single hot-deck imputation using imputation cells such that

$$E[\hat{z}_{igk}|A_r, A_1, \mathcal{F}] = \{\sum_{k \in A_{rig}} w_{igk}\}^{-1} \sum_{k \in A_{rig}} w_{igk} y_{igk}, \tag{5.17}$$

and

$$V[\hat{z}_{igk}|A_r, A_1, \mathcal{F}] = (1 - p_g)^2 N_{ig}^2 (1 - n_{r+g}^{-1} n_{m+g})(n_{m+g})^{-1} \sigma_g^2, \tag{5.18}$$

where z_{igk} are imputed values using hot deck imputation, σ_g^2 are residual variances from the imputation model ξ.

(A4) A donor is not used twice.

(A5) The number of donors from county i used to impute missing values in county i' is known and is denoted by $\tau_{ii'}$.

Then

(R1) $E[\bar{y}_{i..}|\mathcal{F}] = \theta_i + O_p(n_{1ig}^{-1})$, *where* $\bar{y}_{i..}$ *is defined in (5.2) and* θ_i *is the population mean for county* i.

(R2) The variance of the estimated county mean $\bar{y}_{i..}$ *is*

$$\begin{aligned} V([\bar{y}_{i..}|\mathcal{F}]) &= V[\tilde{y}_{1i..}|\mathcal{F}] + N_{i+}^{-2}\sum_g E[p_g(1-p_g)\sum_{k\in A_{1ig}}\{w_{igk}y_{igk} - R_{A_{1ig}}w_{igk}\}^2|\mathcal{F}] \\ &\quad + n_{mi+}n_{1i+}^{-2}(n_{ri+}-1)n_{ri+}^{-1}(\sum_g n_{1i+}^{-1}n_{mig}\sigma_g^2) + O_p(n_{1i+}^{-3/2}), \end{aligned} \tag{5.19}$$

where $R_{A_{1ig}} = (\sum_{k\in A_{1ig}} w_{igk})^{-1}\sum_{k\in A_{1ig}} w_{igk}y_{igk}$, *and* $\tilde{y}_{i..}$ *is the sample mean of* y_{igk} *for county* i *if there were no missing values.*

Assuming a simple random nonreplacement sample (SRSWOR) within county, a consistent estimator of the variance in (5.19) is

$$\begin{aligned} \hat{V}[\bar{y}_{i..}|\mathcal{F}] &= n_{1i+}^{-1}S_{\xi i}^2 + n_{ri+}^{-1}(\sum_g n_{1i+}^{-1}n_{1ig}\hat{\sigma}_g^2) \\ &\quad + n_{mi+}n_{1i+}^{-2}(n_{ri+}-1)n_{ri+}^{-1}(\sum_g n_{1i+}^{-1}n_{1ig}\hat{\sigma}_g^2) \end{aligned} \tag{5.20}$$

where

$$S_{\xi i}^2 = (n_{i1+}-1)^{-1}\sum_g\sum_k(\hat{y}_{igk} - \bar{\hat{y}}_{i..})^2, \tag{5.21}$$

$\bar{\hat{y}}_{i..}$ *is the mean of* $\hat{y}_{igk}$ *in county* i *and* $\hat{\sigma}_g^2$ *is the estimated* σ_g^2 *from the imputation model (5.14).*

(R3) Assuming SRSWOR within county, the covariance of the imputed county means $\bar{y}_{i..}$ *and* $\bar{y}_{i\prime..}$ *is given by*

$$Cov\{(\bar{y}_{i..}, \bar{y}_{i\prime..})|\mathcal{F}\} = (n_{1i+}n_{1i'+})^{-1}\sum_{g=1}^{G}\{\sigma_g^2(\tau_{ii'g} + \tau_{i'ig})\} \tag{5.22}$$

and an estimator of the covariance in (5.22) is

$$cov\{(\bar{y}_{i..}, \bar{y}_{i\prime..})|\mathcal{F}\} = (n_{1i+}n_{1i'+})^{-1}(\tau_{ii'} + \tau_{i'i})G^{-1}\sum_{g=1}^{G}\{\hat{\sigma}_g^2\} \tag{5.23}$$

where $\hat{\sigma}_g^2$ *is the estimated residual variance from model (5.14).*

Using equations (5.20) and (5.23), the form of Σ_{ee} is known and we can apply methods discussed in Section (5.3.1) to estimate the county mean for the C factor. Assumptions (A1) and (A2) are standard and are justified for our setup. We will justify assumption (A3) in the next section. For the NRI surveys, we know the number of points from county i that are used as donors for recipient points in county i'. Given this information, (A5) is known. Although (A4) is not exactly true, the number of times the same donor is used twice is small. See the appendix.

5.4 Estimates for the C Factor 2002

We are interested in estimating the mean C factor for the counties in Iowa for the year 2002 [1]. There are a total of 99 counties for which small area estimates are required. Point level data for the survey years 1997, 2001, and 2002 are available. There are a total of 8340 sample points in 2002 but only 4557 required USLE and C factor for 2002. Although there are a total of 12 broad use categories, C factor and USLE are required only for four categories; viz., cultivated cropland, non-cultivated cropland, pasture land, and Conservation Reserve land. Among the 4557 sampled units, the C factor is observed for 3255 sampled points and 1302 units have imputed values. We consider two types of analysis. In one we use only the 3255 observed units (M1). The analysis based on 3225 units is called the observed unit analysis. The data analysis (M2) uses all 4557 sampled units with imputed values for the 1302 units with missing data.

5.4.1 Imputation Model

In the NRI survey, imputation cells were created based on broad use, land cover use, slope percent, irrigation type, and Cowardin wetland classification. Most of the cells created by full cross classification have either no observations or very few observations

[1]The 2002 NRI data set has not yet been released for public use. All values are strictly for research purposes.

Table 5.1 Summary Statistics for C Factor 2002

Cell	# Obs.	# Miss.	Mean	100 × Estimated Standard Deviation	$100 \times (\hat{\sigma}_g^2)^{1/2}$
1	2165	873	0.2142	5.78	3.65
5	23	11	0.1574	6.96	4.53
9	82	30	0.0824	5.70	4.58
21	100	41	0.0111	1.79	1.79
33	405	141	0.0130	0.98	0.98
46	264	125	0.0040	6.20E-3	6.20E-3

in Iowa. We merged small imputation cells to obtain six cells with a reasonable number of sample points. Cells 21, 33, and 46 are composed of points that have the broad use non-cultivated crop, pasture and CRP, respectively, during 2002. Cell 1 contains cropland segments with a land cover of horticulture or row crops. Cell 5 contains cropland segments with a land cover of close grown crop, such as wheat, oats or barley. Cell 9 contains cropland segments with a land cover of hay or pasture. Table 5.1 shows the six imputation cells with the number of observed points, number of missing points, design estimated mean and estimated S^2 for the C factor 2002. Cell 1 and Cell 5 have a high average C factor and high estimated standard deviation whereas Cell 46 has a low average C factor and low estimated standard deviation.

We considered the C factor for the year 1997 (C97) and the C factor for the year 2001 (C01) as covariates for imputation models. Slope percent is a continuous variable but we treated it as a factor with four levels to create imputation cells. We also considered slope percent for the year 2002 (SP02) as a possible continuous covariate for the imputation model. Design weights (Weight) vary across sample points and are therefore also considered as a possible covariate for the imputation model. Given these possible covariates, we searched for the best parsimonious imputation model using adjusted R^2. Models were estimated from observed data for each imputation cell. We fit an overall

Table 5.2 Parameter Estimates and Standard Errors for Imputation Models

Cell	γ_0	γ_1	γ_2	$10^4 \times \gamma_3$	$10^4 \times \gamma_4$
1	0.0699	0.6345	0.0435	-8.69	0.45
	(0.0036)	(0.0141)	(0.0114)	(2.18)	(0.33)
5	0.0094	0.3334	-0.1142	-0.39	0.25
	(0.0010)	(0.0272)	(0.0183)	(0.57)	(0.16)
9	0.0531	0.4807	-0.0571	0.90	-1.98
	(0.0157)	(0.0742)	(0.0607)	(1.16)	(0.98)

Table 5.3 Summary Statistics for Fitted Values and the NRI Imputed Values

	Min.	First Quartile	Median	Mean	Third Quartile	Max.
Fitted	0.004	0.013	0.195	0.155	0.233	0.360
Imputed	0.003	0.026	0.190	0.157	0.240	0.410

mean in cells 21, 33, and 46; and a model of the form

$$y_{igk} = \gamma_0 + \gamma_1 C01_{igk} + \gamma_2 C97_{igk} + \gamma_3 SP02_{igk} + \gamma_4 Weight_{igk} + \epsilon_{igk} \tag{5.24}$$

in cells 1, 5, and 9 where γ_0, γ_1, γ_2, γ_3, γ_4 are regression parameters and $\epsilon_{igk} \sim (0, \sigma_g^2)$. The regression parameters from model (5.24) are estimated by ordinary least squares and the residual variance σ_g^2 is estimated by

$$\hat{\sigma}_g^2 = (n-5)^{-1} \sum_{i=1}^{m} \sum_{k \in A_{rig}} (y_{igk} - \hat{y}_{igk})^2 \tag{5.25}$$

where $\hat{y}_{igk}$'s are the predicted values from model (5.24) or from the simple mean model. Table 5.2 shows parameter estimates and standard errors for model (5.24). The standard errors are given in parenthesis. $(\hat{\sigma}_g^2)^{1/2}$ from the imputation model are given in Table 5.1.

The models were developed to approximate the nearest neighbor procedure actually used in the NRI. The predicted values from the fitted imputation model are compared to the original imputed values for panel P00.2 in Table 5.3. This table suggests that the

distributions of the predictions from the explicit imputation model closely match the distribution of the imputed values.

Correlations of county means are estimated from (5.23) using the estimated residual standard deviation from the imputation model. There are only 393 positive correlations out of a total of 4851 (= 99 × 49) possible correlations. Summary statistics for (i) all 4851 correlations, (ii) the 393 positive correlations, and (iii) number of donors from a different county are given in Table 5.5. The average positive correlation is 0.16% with a maximum correlation of 1.32%. The low correlation is reasonable because only a few donors are from a different county.

Table 5.4: Estimated Variances for 99 Iowa Counties

ID	n_{1i}	n_{ri}	$\bar{y}_{Ava}$	$\bar{y}_{Imp}$	$10^3\times \hat{V}(\bar{y}_{1i.})$	$10^3\times \hat{V}(\bar{y}_{ri.})$	$10^3\times \hat{V}(\bar{y})_{2phase}$	$10^3\times \hat{V}(\bar{y})_{Imp}$
1	49	37	0.119	0.124	0.192	0.254	0.197	0.201
3	34	23	0.119	0.143	0.390	0.576	0.399	0.405
5	46	30	0.070	0.078	0.172	0.264	0.184	0.191
7	54	35	0.090	0.085	0.201	0.310	0.209	0.214
9	46	24	0.169	0.148	0.193	0.370	0.214	0.224
11	42	36	0.177	0.191	0.316	0.368	0.320	0.324
13	44	35	0.166	0.179	0.199	0.250	0.206	0.211
15	46	33	0.232	0.216	0.127	0.177	0.137	0.145
17	23	19	0.235	0.224	0.260	0.314	0.271	0.279
19	27	16	0.154	0.183	0.245	0.413	0.275	0.292
21	38	30	0.196	0.194	0.218	0.276	0.227	0.233
23	40	28	0.185	0.180	0.158	0.226	0.172	0.181
25	45	31	0.196	0.196	0.071	0.102	0.084	0.093

continued on next page

Table 5.4: Estimated Variances for 99 Iowa Counties (Continued)

ID	n_{1i}	n_{ri}	$\bar{y}_{Ava}$	$\bar{y}_{Imp}$	$10^3\times \hat{V}(\bar{y}_{1i.})$	$10^3\times \hat{V}(\bar{y}_{ri.})$	$10^3\times \hat{V}(\bar{y})_{2phase}$	$10^3\times \hat{V}(\bar{y})_{Imp}$
27	52	37	0.207	0.216	0.105	0.147	0.114	0.121
29	47	34	0.081	0.082	0.128	0.177	0.135	0.14
31	39	36	0.183	0.180	0.270	0.293	0.272	0.275
33	45	33	0.199	0.208	0.155	0.211	0.165	0.171
35	33	26	0.264	0.231	0.154	0.196	0.164	0.172
37	27	21	0.245	0.244	0.198	0.255	0.213	0.223
39	33	25	0.031	0.025	0.130	0.172	0.132	0.133
41	46	29	0.164	0.165	0.216	0.342	0.229	0.237
43	71	55	0.134	0.143	0.144	0.186	0.148	0.152
45	74	56	0.159	0.163	0.138	0.182	0.142	0.146
47	47	31	0.153	0.146	0.151	0.229	0.164	0.172
49	44	30	0.180	0.162	0.201	0.295	0.212	0.219
51	30	19	0.042	0.060	0.253	0.399	0.267	0.276
53	46	32	0.078	0.058	0.272	0.391	0.276	0.279
55	44	28	0.183	0.180	0.188	0.295	0.204	0.214
57	23	19	0.184	0.184	0.548	0.664	0.558	0.565
59	31	21	0.214	0.203	0.292	0.431	0.309	0.320
61	58	44	0.102	0.109	0.125	0.164	0.130	0.134
63	39	32	0.204	0.195	0.236	0.287	0.242	0.246
65	39	28	0.193	0.180	0.274	0.381	0.286	0.294
67	31	23	0.219	0.223	0.466	0.628	0.478	0.486

continued on next page

Table 5.4: Estimated Variances for 99 Iowa Counties (Continued)

ID	n_{1i}	n_{ri}	$\bar{y}_{Ava}$	$\bar{y}_{Imp}$	$10^3\times \hat{V}(\bar{y}_{1i.})$	$10^3\times \hat{V}(\bar{y}_{ri.})$	$10^3\times \hat{V}(\bar{y})_{2phase}$	$10^3\times \hat{V}(\bar{y})_{Imp}$
69	51	42	0.147	0.154	0.106	0.129	0.112	0.116
71	38	30	0.204	0.212	0.333	0.422	0.341	0.348
73	33	23	0.209	0.214	0.139	0.199	0.156	0.167
75	39	24	0.221	0.224	0.067	0.108	0.087	0.100
77	40	29	0.107	0.123	0.256	0.354	0.264	0.270
79	30	27	0.247	0.246	0.106	0.118	0.111	0.115
81	28	21	0.201	0.204	0.057	0.076	0.073	0.084
83	46	28	0.189	0.218	0.206	0.339	0.223	0.233
85	52	39	0.202	0.185	0.157	0.210	0.165	0.171
87	45	31	0.128	0.149	0.242	0.352	0.251	0.258
89	31	21	0.170	0.186	0.360	0.532	0.378	0.389
91	43	26	0.247	0.248	0.066	0.110	0.086	0.097
93	29	20	0.231	0.223	0.108	0.156	0.129	0.143
95	35	32	0.132	0.129	0.368	0.402	0.370	0.371
97	63	43	0.105	0.089	0.144	0.212	0.150	0.154
99	42	31	0.151	0.145	0.199	0.270	0.209	0.215
101	50	35	0.171	0.161	0.246	0.351	0.253	0.258
103	117	77	0.134	0.152	0.068	0.103	0.073	0.076
105	39	30	0.168	0.174	0.206	0.268	0.215	0.222
107	47	29	0.146	0.149	0.251	0.407	0.262	0.268
109	51	40	0.264	0.259	0.110	0.140	0.117	0.122

continued on next page

Table 5.4: Estimated Variances for 99 Iowa Counties (Continued)

ID	n_{1i}	n_{ri}	$\bar{y}_{Ava}$	$\bar{y}_{Imp}$	$10^3\times \hat{V}(\bar{y}_{1i.})$	$10^3\times \hat{V}(\bar{y}_{ri.})$	$10^3\times \hat{V}(\bar{y})_{2phase}$	$10^3\times \hat{V}(\bar{y})_{Imp}$
111	35	23	0.170	0.153	0.392	0.597	0.408	0.417
113	53	34	0.146	0.144	0.137	0.213	0.149	0.157
115	28	19	0.196	0.188	0.276	0.407	0.296	0.309
117	24	18	0.025	0.030	0.222	0.296	0.228	0.232
119	84	65	0.185	0.193	0.074	0.096	0.078	0.082
121	63	35	0.091	0.121	0.183	0.330	0.193	0.198
123	43	26	0.125	0.130	0.201	0.333	0.216	0.225
125	42	36	0.081	0.080	0.173	0.201	0.175	0.178
127	47	33	0.120	0.137	0.108	0.153	0.118	0.125
129	29	23	0.211	0.206	0.416	0.525	0.426	0.433
131	27	15	0.157	0.192	0.280	0.505	0.314	0.332
133	60	39	0.211	0.216	0.141	0.217	0.152	0.159
135	34	20	0.111	0.092	0.346	0.588	0.358	0.365
137	30	22	0.122	0.143	0.189	0.258	0.201	0.210
139	33	26	0.195	0.184	0.236	0.300	0.247	0.255
141	32	26	0.236	0.243	0.170	0.209	0.179	0.186
143	29	20	0.250	0.246	0.047	0.068	0.068	0.081
145	42	31	0.163	0.159	0.209	0.283	0.218	0.224
147	50	29	0.243	0.243	0.105	0.182	0.124	0.134
149	43	34	0.166	0.167	0.194	0.245	0.201	0.206
151	27	19	0.250	0.253	0.050	0.071	0.071	0.085

continued on next page

Table 5.4: Estimated Variances for 99 Iowa Counties (Continued)

ID	n_{1i}	n_{ri}	$\bar{y}_{Ava}$	$\bar{y}_{Imp}$	$10^3\times \hat{V}(\bar{y}_{1i.})$	$10^3\times \hat{V}(\bar{y}_{ri.})$	$10^3\times \hat{V}(\bar{y})_{2phase}$	$10^3\times \hat{V}(\bar{y})_{Imp}$
153	34	24	0.225	0.216	0.257	0.364	0.271	0.281
155	51	35	0.114	0.138	0.166	0.242	0.176	0.182
157	32	22	0.185	0.168	0.166	0.241	0.183	0.195
159	79	44	0.033	0.055	0.055	0.099	0.059	0.061
161	44	35	0.231	0.223	0.125	0.157	0.132	0.138
163	36	26	0.122	0.119	0.232	0.322	0.244	0.252
165	34	31	0.186	0.187	0.241	0.264	0.244	0.247
167	118	82	0.215	0.219	0.054	0.078	0.059	0.062
169	41	33	0.251	0.245	0.224	0.279	0.232	0.237
171	39	32	0.116	0.115	0.217	0.265	0.223	0.227
173	51	34	0.076	0.090	0.101	0.152	0.108	0.113
175	36	27	0.101	0.086	0.268	0.358	0.274	0.278
177	29	17	0.055	0.079	0.364	0.621	0.384	0.395
179	38	24	0.140	0.160	0.440	0.697	0.452	0.459
181	44	32	0.062	0.077	0.198	0.272	0.203	0.207
183	30	28	0.124	0.122	0.316	0.339	0.319	0.321
185	56	43	0.079	0.082	0.121	0.158	0.125	0.127
187	37	17	0.220	0.211	0.131	0.284	0.170	0.187
189	29	18	0.224	0.253	0.190	0.306	0.216	0.231
191	47	37	0.119	0.124	0.251	0.319	0.257	0.260
193	44	36	0.167	0.176	0.299	0.366	0.305	0.309

continued on next page

Table 5.4: Estimated Variances for 99 Iowa Counties (Continued)

ID	n_{1i}	n_{ri}	$\bar{y}_{Ava}$	$\bar{y}_{Imp}$	$10^3\times \hat{V}(\bar{y}_{1i.})$	$10^3\times \hat{V}(\bar{y}_{ri.})$	$10^3\times \hat{V}(\bar{y})_{2phase}$	$10^3\times \hat{V}(\bar{y})_{Imp}$
195	42	31	0.260	0.263	0.333	0.452	0.343	0.349
197	32	24	0.212	0.219	0.164	0.219	0.177	0.186

County variances are estimated by Proposition 5.3.1. For comparison purposes the following variances for county means are estimated:

1. The design variance using the complete first phase sample

$$\hat{V}(\bar{y}_{1i.}) = n_{1i+}^{-1}\{S_{\xi i}^2 + \sum_g n_{1i+}^{-1} n_{1gi}\hat{\sigma}_g^2\}, \tag{5.26}$$

where $S_{\xi i}^2$ is defined in Proposition 5.3.1, and the $\hat{\sigma}_g^2$ are the estimated σ_g^2 from imputation model (5.14).

2. The design variance using the observed units

$$\hat{V}(\bar{y}_{ri.}) = \hat{V}(\bar{y}_{Ava}) = n_{ri+}^{-1}\{S_{\xi i}^2 + \sum_g n_{1i+}^{-1} n_{1gi}\hat{\sigma}_g^2\}. \tag{5.27}$$

3. The design variance for a two phase estimator of the county mean using observed units

$$\hat{V}(\bar{y})_{2phase} = n_{1i+}^{-1} S_{\xi i}^2 + n_{ri+}^{-1} \sum_g n_{1i+}^{-1} n_{1gi}\hat{\sigma}_g^2. \tag{5.28}$$

4. The design variance using available values of the observed units and imputed values for missing units

$$\hat{V}(\bar{y})_{Imp} = \hat{V}(\bar{y})_{2phase} + n_{mi+} n_{1i+}^{-2}(n_{ri+} - 1) n_{ri+}^{-1}(\sum_g n_{1i+}^{-1} n_{1ig}\hat{\sigma}_g^2). \tag{5.29}$$

Table 5.4 contains the number of sampled units n_{1i+} , number of observed units n_{ri+}, county means using available data $\bar{y}_{Ava} = \bar{y}_{ri..}$, county means using all data $\bar{y}_{Imp} =$

Table 5.5 Summary Statistics for the Correlations of County Means

	Min.	First Quartile	Med.	Mean	Third Quartile	Max.
Positive Corr. ×100	0.0387	0.0751	0.1051	0.1642	0.1882	1.3231
All Corr. ×100	0.0000	0.0000	0.0000	0.0132	0.00	1.3231
Number Donors	0	0	0	0.1528	0	12

$\bar{y}_{i..}$, estimated design variances using the first phase sample $\hat{V}(\bar{y}_{1i.})$, estimated design variances using the available data $\hat{V}(\bar{y}_{Ava})$, estimated design variances for the two phase estimator $\hat{V}(\bar{y})_{2phase}$ and estimated design variances using the imputed values $\hat{V}(\bar{y}_{Imp})$.

Summary statistics for the square root of the estimated design variance for the 99 county means, using available data and imputed data, are given in Table 5.6. "Variance ratio" is the ratio of the estimated variance for the estimated mean using all data to the estimated variance for the estimated mean using available data. "Subsampling rate" is the ratio of number of observed units to the number of sampled units. Summary statistics for variance ratios and subsampling rates are also given in Table 5.6. The coefficient of variation (CV) for the direct survey estimator $\bar{y}_{i..}$ is defined by $\text{CV}(\bar{y}_{i..}) = [abs(\bar{y}_{i..})]^{-1}\{\text{Var}(\bar{y}_{i..})\}^{0.5}$ where $abs(x)$ denotes the absolute value of x. Summary statistics for CV using all data and available data are also given in Table 5.6. The average CV is 12% using available data and 11% using all data. There are approximately 23 counties where the CV using available data is greater than 15%. On average, the estimated variance using all data is approximately 80% of the estimated variance using available data.

"Variance ratio two phase" is the ratio of the estimated variance for the two phase estimator to the estimated variance for the estimated mean using available data. Figure 5.1 is a scatter plot of variance ratio two phase and subsampling fraction. The variance ratio two phase is almost one in counties 81, 143 and 151. Table 5.7 presents $S^2_{\xi i}$, and

Table 5.6 Summary Statistics for Square Root of Estimated Design Variances and Estimated Coefficient of Variation of County Means

	Min.	First Quartile	Med.	Mean	Third Quartile	Max.
Available Data $\times 10^2$	0.823	1.357	1.644	1.644	1.886	2.641
Imputed Data $\times 10^2$	0.779	1.219	1.463	1.456	1.659	2.376
Variance Ratio	0.600	0.735	0.800	0.800	0.850	1.189
Subsampling Rate	0.459	0.659	0.718	0.718	0.781	0.933
CV (M1)	0.033	0.076	0.101	0.123	0.145	0.689
CV (M2)	0.035	0.064	0.092	0.106	0.123	0.450

Table 5.7 Three Counties with High Values for the Variance Ratio for Two Phase Estimator

County	Cell	n_{1ig}	n_{rig}	$10^4 \times S^2_{\xi i}$	$10^4 \times \sum_g n_{1i}^{-1} n_{1ig} \hat{\sigma}^2_g$	$\hat{V}(\bar{y})_{2phase}/\hat{V}(\bar{y})_{Ava}$
81	1	28	21	2.705	0.1329	0.958
143	1	29	20	0.389	0.1329	0.991
151	1	27	19	0.220	0.1329	0.995

$\sum_g n_{1i}^{-1} n_{1ig} \hat{\sigma}^2_g$ for the three counties. A closer inspection shows that the three counties consist of cultivated crop land (cell ID 1) with very little variation for the C factor. Although the same $\hat{\sigma}^2_g$ is useful for most counties, it is not beneficial for these three particular counties. A separate imputation cell for the counties with low variation for the C factor could be useful.

5.4.2 Small Area Model and County Level Estimates

Several soil properties, such as the soil erodibility index, the soil support factor, soil texture, erosion index, and slope percent are potential covariates for the small area model. Soil information for each county can be obtained from the USDA soil science database and can be treated as known. Given a set of county level covariates, we searched for the most parsimonious small area model using all data. A small area model of the

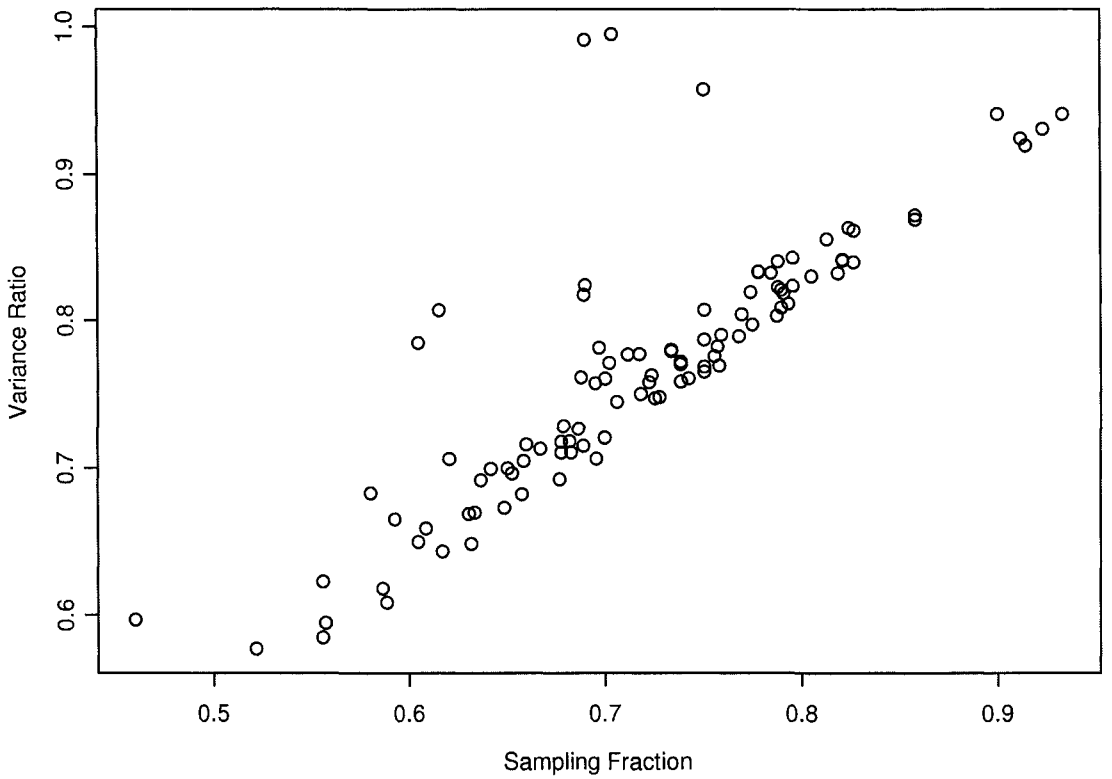

Figure 5.1 Ratio of the Estimated Two Phase Variance to the Estimated Variance using Observed Data.

Table 5.8 Estimates for Regression Parameters and the Between Area Variance Parameter

Procedure	β_0	$10^2 \times \beta_1$	$10^3 \times \sigma_u^2$
M1	0.2532 (0.0101)	-1.491 (0.154)	1.550 (0.258)
M2	0.2556 (0.0091)	-1.497 (0.139)	1.278 (0.211)

form

$$\bar{\mathbf{y}}_{..} = X\boldsymbol{\beta} + \mathbf{u} + \mathbf{e} \tag{5.30}$$

where $\bar{\mathbf{y}}_{..} = (\bar{y}_{1..}, \bar{y}_{2..}, \ldots, \bar{y}_{m..})^T$, $\mathbf{x}_i = (1, \text{mean (Slope)}_i)^T$, $X = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_m^T)$ $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, and $\mathbf{u}$ and $\mathbf{e}$ are defined in (5.4) was selected. Model (5.30) is fitted using the estimated mean based on (i) observed data (M1), and the estimated mean based on (ii) observed data and imputed values (M2). The regression parameters are estimated using (5.8) and the between area variance parameter σ_u^2 is estimated using residual maximum likelihood (Rao, 2003). Marginal and conditional residual plots (not shown) suggested an adequate fit of the model. Table 5.8 shows parameter estimates along with their standard errors. Mean slope percent and mean C factor are negatively correlated. This is because higher precautions are taken to prevent soil loss in the field if slope percent is high.

Small area means are predicted from (5.7). Summary statistics for survey weighted county means and small area predictions are given in Table 5.9. The overall mean for data set M1 is similar to the overall mean for data set M2. The interquartile ranges for the predicted means using small area models are always smaller than the interquartile ranges for the direct means. Plots of predicted values are in Figure 5.2. Predicted values using available data are shown in the top plot and predicted values using all data are shown in the bottom plot. Both plots have a random scatter around the 45° line.

Mean square error of predictions are estimated using (5.13). Summary statistics for the root mean square error of prediction (RMSEP) using available data and using all

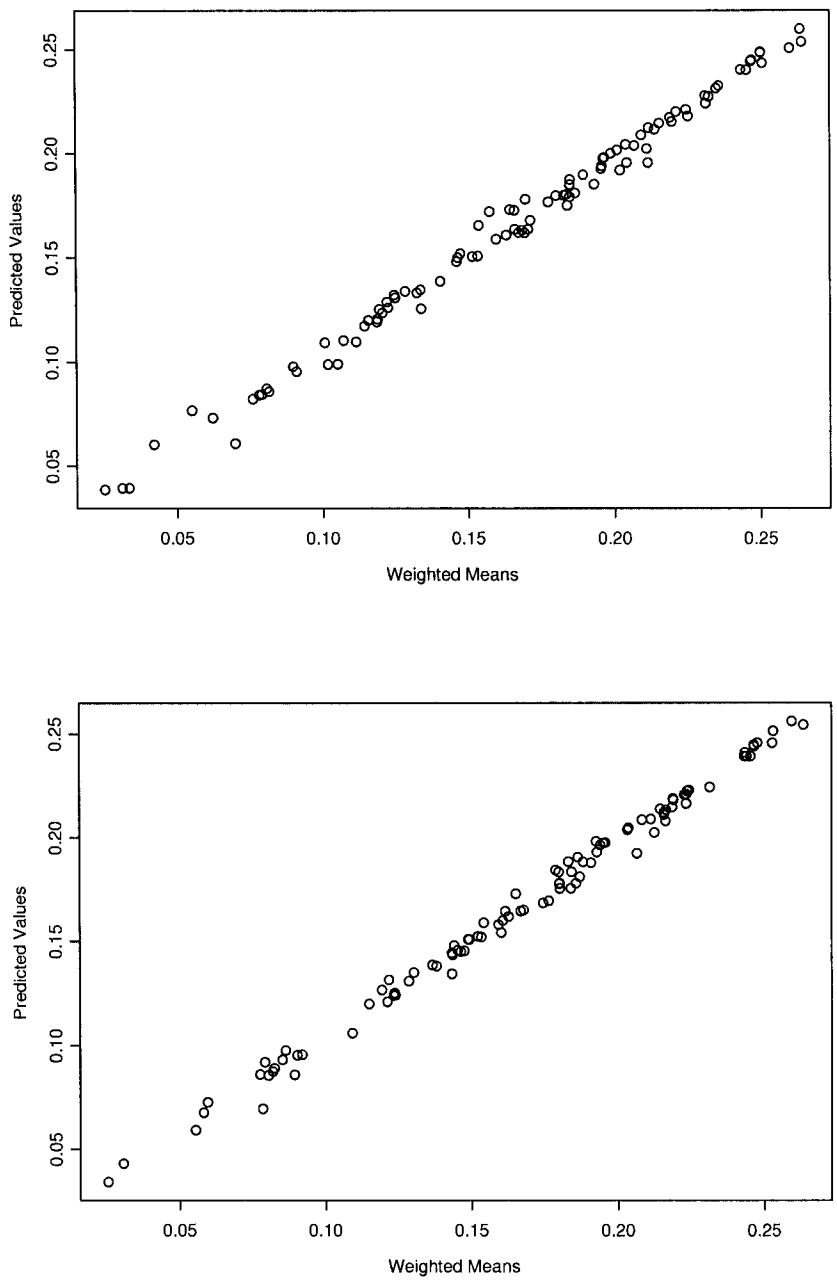

Figure 5.2 Plot of Predicted C Factors.

Table 5.9 Summary Statistics for the Predicted County Means

Procedure	Min.	First Quartile	Median	Mean	Third Quartile	Max.
Weighted (M1)	0.0250	0.121	0.170	0.164	0.210	0.264
Weighted (M2)	0.0253	0.129	0.176	0.167	0.213	0.263
Predicted (M1)	0.0386	0.125	0.173	0.165	0.204	0.260
Predicted (M2)	0.0341	0.133	0.173	0.166	0.209	0.256

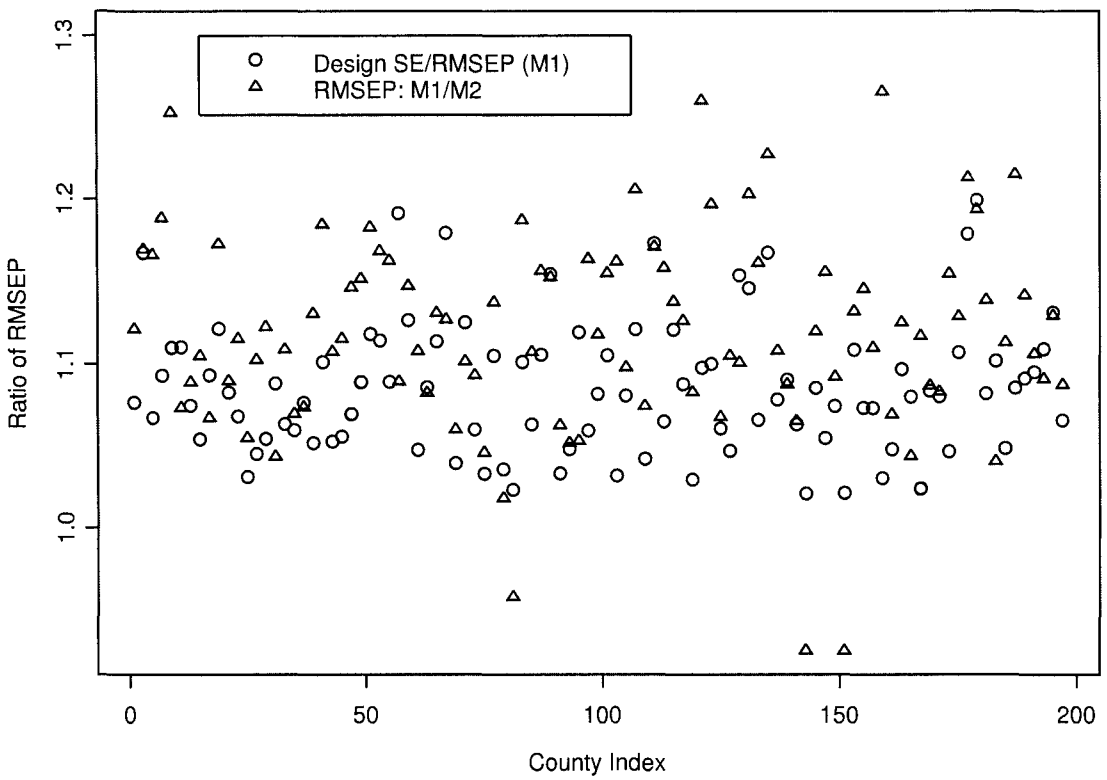

Figure 5.3 Root Mean Square Error of Prediction.

Table 5.10 Summary Statistics for the Root Mean Square Error of Prediction

	100× Min.	100× First Quartile	100× Med.	100× Mean	100× Third Quartile	100× Max.
M1	0.810	1.287	1.522	1.504	1.705	2.202
M2	0.762	1.158	1.356	1.339	1.508	1.984

data are given in Table 5.10. The average RMSEP is 0.01504 using available data and 0.01339 when using all data. Thus, there is a 11.0% relative improvement using all data. RMSEP using all data is smaller than using available data except for the three counties (81, 143, and 151). Table 5.7 presents a detailed description of the three counties. The ratio of design standard error to RMSEP using available data and the ratio of RMSEP using available data to RMSEP using all data is shown in Figure 5.3. A value greater than 1 indicates smaller estimated MSEP. From Table 5.10 and Figure 5.3 it is clear that small area prediction using M1 is preferable to direct survey means. Small area prediction using M2 produces smaller RMSEP in most counties than using M1. Figure 5.4 is a scatter plot of the missing rate and the ratio of RMSEP using all data and available data in each county. The smooth line in Figure 5.4 is obtained through scatter plot smoothing. An increasing trend suggests that as the missing rates increase the efficiency of M2 relative to M1 increases.

5.5 Conclusions

We considered the effect of imputation on the estimates constructed for a small area model. The missing values were imputed through hot deck imputation via regression. A unit level imputation model was built which closely matches the imputation procedure used in the NRI. Since the imputation cells cross county boundaries, the county estimates are correlated. A method of estimating the correlations using the fitted imputation

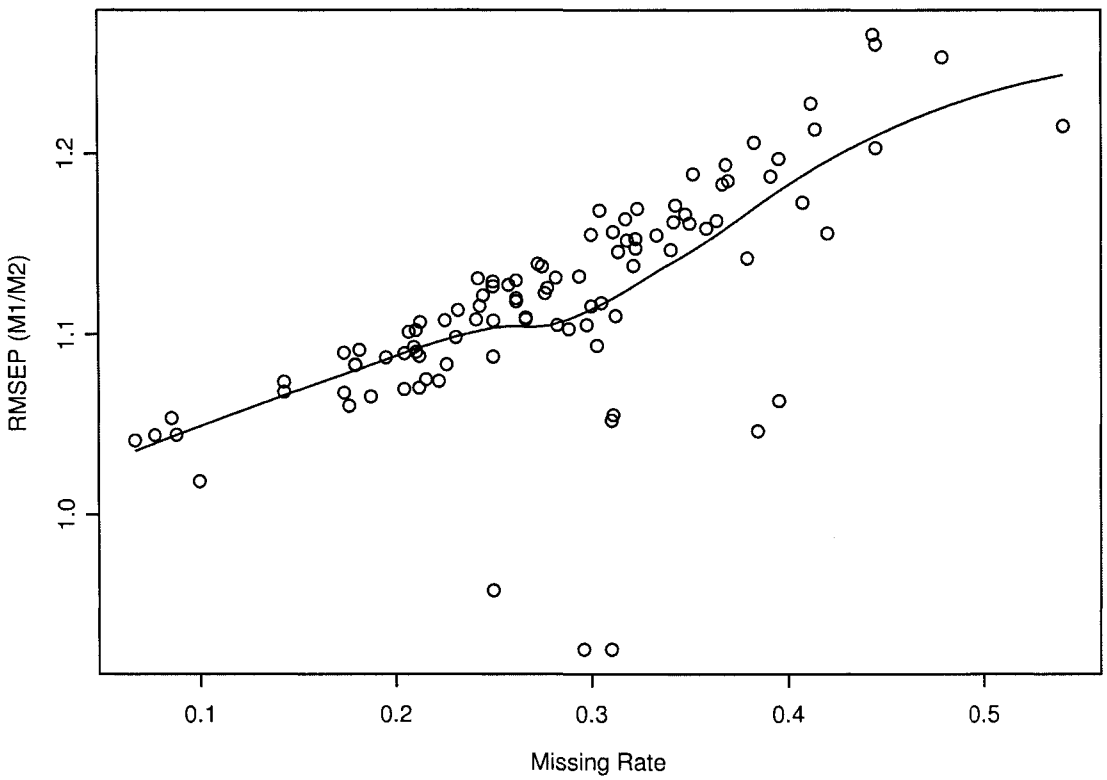

Figure 5.4 Scatter Plot of Missing Rate and the Ratio of RMSEP.

model was proposed. Variances within imputation cells were used to estimate the extra variability due to the imputation. A multivariate small area model was then fitted to the county level data, assuming the estimated design variance was known. The EBLUP estimator from the fitted small area model and estimated MSEP were used to produce county estimates. It was shown that the randomness due to the imputation mechanism should be considered for small area estimation and the proposed methodology was adjusted to account for the randomness due to the imputation mechanism.

CHAPTER 6. Summary

In the first section, a methodology is developed to obtain calibrated small area estimators when a smooth nonlinear function of direct means are linearly related with the covariates. The methodology is applied to estimate wind erosion for the counties in Iowa. The second section proposes a nonparametric small area model. Small area means are estimated using Nadaraya-Watson estimators and local polynomial regression estimators. Theoretical properties of the proposed estimators are studied under mild assumptions. A limited simulation study is conducted to verify the performances of the proposed estimators. The Nadaraya-Watson estimator is used to estimate wind erosion within three states in the US. In the final section, the effects of imputed values for small area estimations are studied. A methodology is proposed to estimate the design variance of the mean C factors using imputed values. The estimated design variance is used to fit a multivariate Fay Herriot model. Predictions from the multivariate Fay Herriot model have low standard errors relative to the small area predictions using available data. This chapter summarizes the statistical methods developed in this thesis.

The area level model is

$$\bar{y}_{i.} = \theta_i + \epsilon_i, \text{ and} \tag{6.1}$$

$$\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \tag{6.2}$$

where $\bar{y}_{i.}$'s are design weighted means, $\mathbf{x}_i$'s are area level covariates, $\boldsymbol{\beta}$ is a set of regression parameters, u_i's are random effects, ϵ_i's are sampling errors, and $i = 1, 2, \ldots, m$ are small areas. True small area means $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i$ are linear combinations of regres-

sion parameters $\boldsymbol{\beta}$ and area specific random effects $u_1, u_2, \ldots, u_m$. Assume $u_i \overset{iid}{\sim} (0, \sigma_u^2)$, $\epsilon_i \overset{ind}{\sim} (0, \psi_i)$ and u_i's and ϵ_i's are mutually independent. The BLUP for θ_i is

$$\tilde{\theta}_i = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} + \gamma_i (y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}), \tag{6.3}$$

where

$$\tilde{\boldsymbol{\beta}} = [X^T V^{-1} X]^{-1} X^T V^{-1} \mathbf{y}, \tag{6.4}$$

$$\gamma_i = (\sigma_u^2 + \psi_i)^{-1} \sigma_u^2, \tag{6.5}$$

$X = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_m^T)^T$, $V = \sigma_u^2 I_m + \text{diag}\{\psi_i\}_{i=1}^m$, and $\mathbf{y} = (y_1, \ldots, y_m)^T$. Further assume $\hat{\sigma}_u^2$ is an estimator of σ_u^2. Thus, the EBLUP for θ_i is

$$\hat{\theta}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_i (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}), \tag{6.6}$$

where

$$\hat{\boldsymbol{\beta}} = [X^T \hat{V}^{-1} X]^{-1} X^T \hat{V}^{-1} \mathbf{y}, \tag{6.7}$$

and $\hat{\gamma}_i = (\hat{\sigma}_u^2 + \psi_i)^{-1} \hat{\sigma}_u^2$. Further, if $u_i \overset{ind}{\sim} N(0, \sigma_u^2)$, $e_i \overset{iid}{\sim} N(0, \psi_i)$ and the ψ_i's are known, then the mean squared error (MSE) for $\hat{\theta}_i$ is given by,

$$E\{\hat{\theta}_i - \theta_i\}^2 = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) + O(m^{-2}), \tag{6.8}$$

where

$$g_{1i}(\sigma_u^2) = (\sigma_u^2 + \psi_i)^{-1} \psi_i \sigma_u^2, \tag{6.9}$$

$$g_{2i}(\sigma_u^2) = (1 - \gamma_i)^2 \mathbf{x}_i^T \left[\sum \mathbf{x}_i^T (\sigma_u^2 + \psi_i \right]^{-1} \mathbf{x}_i)^{-1} \mathbf{x}_i, \tag{6.10}$$

$$g_{3i}(\sigma_u^2) = (\sigma_u^2 + \psi_i)^{-3} \psi_i^2 V(\hat{\sigma}_u^2), \tag{6.11}$$

$V(\hat{\sigma}_u^2)$ is the variance of $\hat{\sigma}_u^2$ and γ_i is defined in (6.5). The order of approximation in (6.8) is valid under certain regularity conditions (Prasad and Rao, 1990). The approximated MSE (6.8) can be estimated by

$$\text{mse}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2), \tag{6.12}$$

where $\hat{\sigma}_u^2$ is the REML or the MM estimator of σ_u^2 (Prasad and Rao, 1990; Rao, 2003). In the first part of this dissertation, the NRI data is used to estimate county level wind erosion for 2002. The WEQ is used as the response variable. The coefficient of variation for direct estimates ranges from 8.4% to 84% with an average of 31.1%. A soil erodibility index is used to fit a small area model for the WEQ. A cube root transformation of the response is found to be linear with the covariate. The small area model for the WEQ is

$$(\bar{y}_i)^{1/3} = \beta_0 + \beta_1 x_i + u_i + e_i, \tag{6.13}$$

where $\bar{y}_i$'s are the mean WEQ's, x_i's are the mean IFact, β_0 and β_1 are fixed parameters, u_i's are area specific random components, and e_i's are sampling errors. Assume $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $e_i \overset{ind}{\sim} N(0, D_i^*)$ and u_i's and e_i's are mutually independent, where $D_i^* = E[\bar{y}_i^{1/3}|\mathcal{F}]$ and $\mathcal{F}$ is the finite population. The small area means, $\mu_i = \beta_0 + \beta_1 x_i + u_i$, are predicted by

$$\hat{\mu}_i = \hat{\gamma}_i \bar{y}_i + (1 - \hat{\gamma}_i)(\hat{z}_i^3 + \hat{q}_i \hat{z}_i^2), \tag{6.14}$$

where $\hat{z}_i = \{\hat{\bar{y}}_{i.}\}^{1/3} = \hat{\gamma}_i \bar{y}_{i.}^{1/3} + (1 - \hat{\gamma}_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i)$, $\hat{\gamma}_i = (\hat{\sigma}_u^2 + D_i^*)^{-1}\hat{\sigma}_u^2$, $\hat{q}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i + \hat{\alpha}_2 w_i(1 + \hat{\gamma}_i)\hat{z}_i^2$, $\hat{\alpha} = (Z^T Z)^{-1} Z^T \tilde{\mathbf{q}}$, $\tilde{\mathbf{q}} = (q_1, \ldots, q_m)^T$, $Z = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m)^T$, $\mathbf{a}_i = (1, x_i, \xi_i)^T$, $\xi_i = w_i(1 + \hat{\gamma}_i)\hat{z}_i^2$ and w_i's are the survey weights for area i. The weighted sum of the predicted means is the state direct estimate. A methodology to obtain calibrated estimators for any smooth transformation of the response is discussed. The average coefficient of variation for the final estimates is 17.6% and the maximum coefficient of variation is 37.6%.

In the second part of this dissertation, we propose a non-parametric mixed effects model of the form

$$y_i = \theta_i + \epsilon_i, \text{ and} \tag{6.15}$$

$$\theta_i = m(x_i) + u_i, \qquad i = 1, 2, \ldots, n, \tag{6.16}$$

where x_i's are area level covariates, y_i's are direct estimators of mean responses for the area i, $m(\cdot)$ is a smooth mean function, θ_i's are the unobserved small area means, u_i's are area specific random effects, and e_i's are sampling errors. Assume $u_i \overset{iid}{\sim} (0, \sigma_u^2)$, $\epsilon_i \overset{ind}{\sim} (0, D_i)$, and D_i's are known constants.

The mean function $m(x)$ is estimated using the Nadaraya-Watson estimator

$$\hat{m}_h(x) = \left\{ \sum_i K_h(x - x_i) \right\}^{-1} \sum_i K_h(x - x_i) y_i, \tag{6.17}$$

where $K_h(\cdot)$ is a kernel function with bandwidth h. Assuming $D_i = D$ for all $i = 1, 2, \ldots, n$, the between area variance function is estimated by

$$\hat{\sigma}_u^2(x) = \min\{0, (n-1)^{-1} \sum_{i=1}^{n} W_{hi}(x)[y_i - \hat{m}_h(x_i)]^2 - D\}, \tag{6.18}$$

where $W_{hi}(x) = \{\sum_i K_h(x - x_i)\}^{-1} K_h(x - x_i)$.

Assuming u_i's and ϵ_i's are normally distributed, it is shown that the best predictor of θ_i is

$$E(\theta_i | y_i) = \tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i)\hat{m}_h(x_i), \tag{6.19}$$

where $\gamma_i = (\sigma_u^2 + D_i)^{-1}\sigma_u^2$. A two stage estimator of θ_i is obtained by

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\hat{m}_h(x_i), \tag{6.20}$$

where $\hat{\gamma}_i = (\hat{\sigma}_u^2 + D_i)^{-1}\hat{\sigma}_u^2$.

A Taylor series approximation of the MSE of the proposed estimator is obtained. An estimator of the approximated MSE of θ_i is proposed by

$$\text{mse}(\hat{\theta}_i) = \frac{D\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + D} + (1 - \hat{\gamma})^2 \text{mse}[\hat{m}_h(x_i)] + 2D^2(\hat{\sigma}_u^2 + D)^{-3}\text{mse}(\hat{\sigma}_u^2), \tag{6.21}$$

where $\text{mse}[\hat{m}_h(x_i)]$ and $\text{mse}(\hat{\sigma}_u^2)$ are the estimated MSE for $\hat{m}(x_i)$ and $\hat{\sigma}_u^2$.

A limited simulation study shows that the proposed estimator performs similarly to the Fay-Herriot estimator (6.6) when the mean function is linear. The proposed estimator performs better than the Fay-Herriot estimator when the mean function is

nonlinear. The methodology is applied to estimate wind erosion for the counties within the three states in the US. The final results are encouraging.

Theoretical properties of the Nadaraya-Watson estimator were studied using local polynomial regression estimators. Model (6.16) is extended to

$$y_i = \theta_i + \epsilon_i, \text{ and} \tag{6.22}$$

$$\theta_i = m(x_i) + u_i, \tag{6.23}$$

where x_i's are area level covariates, y_i's are direct estimators of mean responses for the area i, $m(\cdot)$ is a smooth mean function, θ_i's are the unobserved small area means, u_i's are area specific random effects, and e_i's are sampling errors. Assume $u_i \overset{iid}{\sim} (0, \sigma^2(x_i))$, $\epsilon_i \overset{ind}{\sim} (0, D_i)$, and D_i's are known constants. Assume $\sigma^2(x_i) > 0$ is a smooth function of x_i. The mean function and the between area variance function are estimated using local polynomial regression. An estimator of the ν^{th} derivative $m^{(\nu)}(\cdot)$ of the mean function $m(\cdot)$ is given by

$$\hat{m}^{(\nu)}(x_0) = \nu! \mathbf{e}_{\nu+1}^T \hat{\boldsymbol{\beta}}(x_0), \qquad \nu = 0, 1, \ldots, p_1, \tag{6.24}$$

where

$$\hat{\boldsymbol{\beta}}(x_0) = [X_{p_1}(x_0)^T W_{p_1}(x_0) X_{p_1}(x_0)]^{-1} X_{p_1}(x_0)^T W_{p_1}(x_0) \mathbf{y}, \tag{6.25}$$

$\mathbf{e}_\nu$ is a unit vector of correct dimension with the ν^{th} element as one, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is an n-vector of survey weighted means, $X_1(x_o) = [(X_i - x_0)^j]_{1 \le i \le n, 0 \le j \le p_1}$ is a $n \times p_1$ random design matrix, and $W_1(x_0) = diag\{K_{h_1}(X_i - x_0)\}$ is a $n \times n$ diagonal matrix of kernel weights.

The between area variance function $\sigma^2(x_i)$ is estimated by using the observed residuals. Assume $P_2 \mathbf{1} = P_2$, where P_2 is a smoother matrix (Section 4.2) for a p_2 degree polynomial with a bandwidth of h_2, and $\mathbf{1}$ is a vector of ones. We propose an estimator of $\boldsymbol{\sigma}^2 = (\sigma^2(X_1), ..., \sigma^2(X_n))^T$ by

$$\hat{\boldsymbol{\sigma}^2} = \frac{P_2(\mathbf{r}^2 - \boldsymbol{\Delta}_2)}{\mathbf{1} + P_2 \boldsymbol{\Delta}_1}, \tag{6.26}$$

where $\mathbf{r} = \mathbf{y} - P_1\mathbf{y}$, $\mathbf{\Delta}_1 = diag\{P_1P_1^T - 2P_1\}$, $\mathbf{\Delta}_2 = diag\{D + P_1DP_1^T - 2P_1D\}$, $D = \text{diag}\{D_i\}_{i=1}^n$, and P_1 is a smoother matrix for a p_1 degree polynomial with a bandwidth h_1. Asymptotic properties of the proposed estimators are discussed and it is shown that there is no loss in asymptotic efficiency of $\hat{\boldsymbol{\sigma}}^2$ due to the estimation of the mean function m.

Small area means $\theta_i = m(x_i) + u_i$ are estimated by

$$\tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i)\hat{m}(x_i), \tag{6.27}$$

where $\gamma_i = (\sigma_u^2(X_i) + D_i)^{-1}\sigma_u^2(X_i)$ is the minimizer of the conditional mean squared error $E[\theta_i^* - \theta_i|\mathbb{X}]^2$ and

$$\theta_i^* = \gamma_i y_i + (1 - \gamma_i)m_i, \tag{6.28}$$

where $m_i = m(x_i)$ A two stage estimator for θ_i is obtained by replacing γ_i with its estimated value

$$\hat{\gamma}_i = [\hat{\sigma}_u^2(X_i) + D_i]^{-1}\hat{\sigma}_u^2(X_i). \tag{6.29}$$

Thus,

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\hat{m}(x_i) \tag{6.30}$$

The mean squared error for the proposed estimator is

$$\begin{aligned} & E[\{\hat{\theta}_i - \theta_i\}^2|\mathbb{X}] \qquad (6.31) \\ &= g_{1i}(\sigma^2(X_i)) + g_{2i}(\sigma^2(X_i), m_i) + g_{3i}(\sigma^2(X_i)) + g_{4i}(\sigma^2(X_i)) + o_p(r_{nh}), \end{aligned}$$

where r_{nh}, $g_{1i}(\sigma_i^2)$, $g_{2i}(\sigma_i^2, m_i)$, $g_{3i}(\sigma_i^2)$, and $g_{4i}(\sigma_i^2)$ are given in Theorem 4.33.

In the third and the final part of this dissertation, the effect of imputed values on small area estimation is considered. The NRI data is used to estimate the C factor for the counties in Iowa in the year 2002. Imputation cells are created using broad use and land cover use. Cell regression models are fitted within each imputation cell. Estimated residual variances from the cell models are used to approximate a part of the variance

due to the original hot deck imputation used for the NRI. The design variance for the direct mean $\bar{y}_{i.}$ is estimated by

$$\begin{aligned}\hat{V}[\bar{y}_{i..}|\mathcal{F}] &= n_{1i+}^{-1}S_{\xi i}^2 + n_{ri+}^{-1}\left(\sum_g n_{1i+}^{-1}n_{mig}\hat{\sigma}_g^2\right) \\ &\quad + n_{mi+}n_{1i+}^{-2}(n_{ri+}-1)n_{ri+}^{-1}\left(\sum_g n_{1i+}^{-1}n_{mig}\hat{\sigma}_g^2\right), \end{aligned} \tag{6.32}$$

where

$$S_{\xi i}^2 = (n_{1i+}-1)^{-1}\sum_g\sum_k(\hat{y}_{igk}-\bar{\hat{y}}_{i..})^2, \tag{6.33}$$

$\bar{\hat{y}}_{i..}$ is the mean of $\hat{y}_{igk}$ in county i, $\hat{\sigma}_g^2$ is the estimated σ_g^2 from imputation model (5.14), $\hat{y}_{igk}$'s are predictions from imputation model (5.24), and n_{1i+}, n_{ri+}, n_{mi+}, n_{rig} and n_{mig} are defined in Proposition 5.3.1. The estimated design variances are used to fit a multivariate small area model

$$\begin{aligned}\mathbf{y} &= \boldsymbol{\mu} + \mathbf{e}, \\ \boldsymbol{\mu} &= X\boldsymbol{\beta} + \mathbf{u}\end{aligned} \tag{6.34}$$

where $\mathbf{y} = (\bar{y}_{1..}, \bar{y}_{2..}, \ldots, \bar{y}_{m..})^T$, X is a $m \times p+1$ matrix of covariates, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is a vector of regression parameters, $\mathbf{e}$ is the sampling error, and $\mathbf{u}$ is an area specific random quantity. Further assume

$$(\mathbf{u}^T, \mathbf{e}^T) \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2 I & \mathbf{0} \\ \mathbf{0} & \Sigma_{ee} \end{bmatrix}\right). \tag{6.35}$$

The small area means are predicted by

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} + \hat{\sigma}^2\hat{\Sigma}_{zz}^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}}), \tag{6.36}$$

where

$$\hat{\boldsymbol{\beta}} = (X^T\hat{\Sigma}_{zz}^{-1}X)^{-1}X^T\hat{\Sigma}_{zz}^{-1}\mathbf{y}, \tag{6.37}$$

$\hat{\Sigma}_{zz} = \hat{\sigma}^2 I + \Sigma_{ee}$ and Σ_{ee} is assumed to be known. The MSE of $\hat{\boldsymbol{\mu}}$ and an estimator of the MSE are given in (5.9) and (5.13) respectively. Small area predictions are obtained

using all data analysis and available data analysis. The all data analysis uses observed C factors for observed units and imputed C factors for missing units. The available data analysis uses only observed units. It is shown that predictions from the proposed methodology has smaller estimated RMSEP relative to the estimated RMSEP by using the available data.

APPENDIX A. Proofs of Chapter 3

Proof of Theorem 3.2.1

Consider the numerator and denominator separately. We want to show that:

1. The numerator,

$$\hat{r}_h(x) = m^{-1} \sum_i K_h(x - x_i) y_i \xrightarrow{p} m(x) f(x), \tag{A.1}$$

where x is a point of continuity of $m(x)$.

2. The denominator,

$$\hat{f}_h(x) = m^{-1} \sum_i K_h(x - x_i) \xrightarrow{p} f(x). \tag{A.2}$$

If $f(x) > 0$, then by the Slutzky's theorem (Casella and Berger, 2002)

$$\hat{m}(x) = \{\hat{f}_h(x)\}^{-1} \hat{r}_h(x) \xrightarrow{p} m(x). \tag{A.3}$$

The result (A.2) is standard for kernel estimators. For examples, see Härdle (2002).

$$E[\hat{r}_h(x)] = \int \int m^{-1} K_h(x - u) y_i f(u, y_i) du dy_i, \tag{A.4}$$

where $f(u, y_i)$ are the joint density of (x_i, y_i). Using conditional expectations,

$$\begin{aligned} & E[\hat{r}_h(x)] \\ &= m^{-1} \int K_h(x - u) \{ \int y_i f(u, y_i) dy_i \} du \\ &= m^{-1} \int K_h(x - u) \{ m(u) \int f(u, y_i) dy_i \} du \\ &= m^{-1} \int K_h(x - u) \{ m(u) f(u) du. \end{aligned} \tag{A.5}$$

Let $r(x) = m(x)f(x)$. Following lemma 3.1.1 Härdle (2002),

$$\begin{aligned} & E[|\hat{r}_h(x) - m(x)f(x)|] \\ & \leq \sup_{|t|\leq\delta} |r(x-t) - r(t)| \int |K(t)|dt + \delta^{-1} \sup_{|t|\geq h^{-1}\delta} |tK(t)| \int |r(t)|dt \\ & \quad + |r(x)| \int_{|t|\geq h^{-1}\delta} |K(t)|dt. \end{aligned} \tag{A.6}$$

Since x is a point of continuity of $r(x)$,

$$\sup_{|t|\leq\delta} |r(x-t) - r(t)| \to 0 \text{ as } \delta \to 0. \tag{A.7}$$

The second and the third term in (A.6) vanish by (A1) and (A2) and therefore

$$E[|\hat{r}_h(x) - m(x)f(x)|] \to 0 \text{ as } h \to 0. \tag{A.8}$$

Let $E[y_i^2] = s^2(x) + D_i$. Note that,

$$\begin{aligned} m^{-1}V[K_h(x-x_i)y_i] &= \left\{ \int K_h^2(x-u)s^2(u)f(u)du - \left(\int K_h(x-u)r(u)du \right)^2 \right. \\ &\quad \left. + D_i \int K_h^2(x-u)f(u)du \right\}. \end{aligned} \tag{A.9}$$

Writing $u = x + th$,

$$\int K_h^2(x-u)s^2(u)f(u)du = h^{-1} \int K^2(t)s^2(x+th)f(x+th)dt \tag{A.10}$$

$$= h^{-1}s^2(x)f(x) \int K^2(t)dt + o(h^{-1}). \tag{A.11}$$

Similar expressions for the last two terms in (A.9) can be obtained. Therefore if D_i's are bounded then

$$V[\hat{r}_h(x)] = m^{-2} \sum_{i=1}^{m} V[K_h(x-x_i)y_i] \tag{A.12}$$

$$= m^{-1}h^{-1}\{s^2(x) + \bar{D}\}f(x) \int K^2(u)du + o\left((mh)^{-1}\right) \tag{A.13}$$

where $\bar{D} = m^{-1} \sum_{i=1}^{m} D_i$.

Therefore from (A.8) and (A.13) and using Chebyshev's inequality (Casella and Berger, 2002) $\hat{r}_h(x) \xrightarrow{p} m(x)f(x)$.

Proof of Theorem 3.2.2

The MSE can be written as

$$E[\hat{m}_h(x_i) - m(x_i)]^2 = V[\hat{m}_h(x_i)] + E^2[\hat{m}_h(x_i) - m(x_i)]. \tag{A.14}$$

The expression for the bias is

$$E[\hat{m}_h(x_i) - m(x_i)] = h^2 d_k\{m^{(2)}(x_i) + f^{-1}(x_i)f^{(1)}(x_i)m^{(1)}(x_i) + o(h^2)\}. \tag{A.15}$$

See Fan and Gijbels (1996) for the proof. The variance of y_i is $V[y_i] = \sigma_u^2(x) + D_i = E[y_i^2] - E^2[y_i]$. Therefore $\sigma_u^2(x_i) = s^2(x) - m(x_i)^2$. Also $\hat{f}_h(x_i) = f(x_i) + O_p(a_n)$ where $a_n \to 0$. Therefore, after simplification, the variance for $\hat{m}_h(x_i)$ can be obtained from (A.13) as

$$V[\hat{m}_h(x_i)] = (mh)^{-1} f^{-1}(x_i)\{\sigma_u^2 + D_i\}c_k + o\left((mh)^{-1}\right). \tag{A.16}$$

Proof of Theorem 3.2.3

Expand the squares of $\hat{\theta}_i - \theta_i$ into $\tilde{\theta}_i - \theta_i$ and $\hat{\theta}_i - \tilde{\theta}_i$.

$$\begin{aligned}
\tilde{\theta}_i - \theta_i &= \gamma_i y_i + (1-\gamma_i)\hat{m}_i - \theta_i \\
&= \gamma_i y_i + (1-\gamma_i)m_i - \theta_i + (1-\gamma_i)(\hat{m}_i - m_i) \\
&= \gamma_i(m_i + u_i + \epsilon_i) + (1-\gamma_i)m_i - (m_i + u_i) + (1-\gamma_i)(\hat{m}_i - m_i) \\
&= \{\gamma_i\epsilon_i - (1-\gamma_i)u_i\} + \{(1-\gamma_i)(\hat{m}_i - m_i)\} \qquad \text{(A.17)}
\end{aligned}$$

Any linear estimator of m_i is of the form $\sum_j a_{ij}y_j$ where a_{ij} are constants. Therefore for any linear estimator of m_i the covariance term in (A.17) can be written as,

$$\begin{aligned}
\text{Cov}[\hat{m}_i, \gamma_i\epsilon_i - (1-\gamma_i)u_i] &= \text{Cov}[\sum_j a_{ij}y_j, \gamma_i\epsilon_i - (1-\gamma_i)u_i] \\
&= \gamma_i a_{ii}\psi_i - (1-\gamma_i)a_{ii}\sigma_u^2 \\
&= a_{ii}\{\gamma_i\psi_i - (1-\gamma_i)\sigma_u^2\} \\
&= 0. \qquad \text{(A.18)}
\end{aligned}$$

Accordingly

$$\begin{aligned} & E[(\tilde{\theta}_i - \theta_i)^2] \\ &= E\left[\{\gamma_i \epsilon_i - (1-\gamma_i)u_i\}^2\right] + E\left[\{(1-\gamma_i)(m_i - \hat{m}_i)\}^2\right] \\ &= g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) \end{aligned} \tag{A.19}$$

where $g_{1i}(\sigma_u^2) = E\left[\{\gamma_i \epsilon_i - (1-\gamma_i)u_i\}^2\right]$ is the mean squared error if all the parameters are known, and $g_{2i}(\sigma_u^2) = E\left[\{(1-\gamma_i)(m_i - \hat{m}_i)\}^2\right]$ is the mean squared error due of the estimation of the mean function m_i.

$$\begin{aligned} g_{1i}(\sigma_u^2) &= E\left[\{\gamma_i \epsilon_i - (1-\gamma_i)u_i\}^2\right] \\ &= (\sigma_u^2 + D_i)^{-1}\sigma_u^2 D_i \end{aligned} \tag{A.20}$$

and

$$\begin{aligned} g_{2i}(\sigma_u^2) &= E\left[\{(1-\gamma_i)(m_i - \hat{m}_i)\}^2\right] \\ &= (1-\gamma_i)^2 \mathrm{MSE}(\hat{m}_i) \end{aligned} \tag{A.21}$$

Now write

$$\hat{\theta}_i - \theta_i = \hat{\theta}_i - \tilde{\theta}_i + \tilde{\theta}_i - \theta_i. \tag{A.22}$$

Now,

$$\begin{aligned} & E[(\hat{\theta}_i - \tilde{\theta}_i)(\tilde{\theta}_i - \theta_i)] \\ &= E[(\hat{\gamma}_1 - \gamma_i)(y_i - \hat{m}_i)(1-\gamma_i)(\hat{m}_i - m_i)]. \end{aligned} \tag{A.23}$$

Therefore using the normality of the error components u_i and e_i, and ignoring the conditional bias $E[(\hat{m}_i - m_i)|(y_i - \hat{m}_i)]$, the product term in (A.23) vanish. Hence using (A.18),

$$E[\hat{\theta}_i - \theta]^2 = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2), \tag{A.24}$$

where $g_{3i}(\sigma_u^2) = E[\hat{\theta}_i - \tilde{\theta}_i]^2$. We approximate the expectation in g_{3i} by a Taylor series expansion. Using Fuller (1996) (see appendix of Chapter 4 for a detailed proof), by a two step Taylor approximation,

$$E[\hat{\theta}_i - \tilde{\theta}_i]^2 = E[(\hat{\gamma}_1 - \gamma_i)^2(y_i - \hat{m}_i)^2] \tag{A.25}$$

$$= D^2(\sigma_u^2 - D)^{-4}E[(y_i - m_i)^2(\hat{\sigma}_u^2 - \sigma_u^2)^2] + O(a_m^3), \tag{A.26}$$

where $a_m = \max\{(nh)^{-1/2}, h^{2/3}\}$ and $a_m \to 0$ as $m \to \infty$.

APPENDIX B. Proofs of Chapter 4

Proof of Theorem 4.3.1

The expectation and the variance for

$$\hat{\boldsymbol{\beta}}(x_0) = [[X_1(x_0)^T W_1(x_0) X_1(x_0)]]^{-1} X_1(x_0)^T W_1(x_0)\mathbf{y}, \tag{B.1}$$

are

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}(x_0)|\mathbf{X}] &= [[X_1(x_0)^T W_1(x_0) X_1(x_0)]]^{-1} X_1(x_0)^T W_1(x_0) E[\mathbf{y}|\mathbf{X}] && \text{(B.2)} \\ &= [[X_1(x_0)^T W_1(x_0) X_1(x_0)]]^{-1} X_1(x_0)^T W_1(x_0)\mathbf{m} && \text{(B.3)} \\ &= \boldsymbol{\beta}(x_0) + [[X_1(x_0)^T W_1(x_0) X_1(x_0)]]^{-1} X_1(x_0)^T W_1(x_0)\mathbf{t}_1(x_0), && \text{(B.4)} \end{aligned}$$

and

$$\begin{aligned} & V[\hat{\boldsymbol{\beta}}(x_0)|\mathbf{X}] \\ &= [X_1(x_0)^T W_1(x_0) X_1(x_0)]^{-1} X_1(x_0)^T W_1(x_0) V[\mathbf{y}|\mathbf{X}][X_1(x_0)^T W_1(x_0)]^T \\ &\quad [X_1(x_0)^T W_1(x_0) X_1(x_0)]^{-1} \\ &= [X_1(x_0)^T W_1(x_0) X_1(x_0)]^{-1} X_1(x_0)^T W_1(x_0) \Sigma_1(x_0) [X_1(x_0)^T W_1(x_0)]^T \\ &\quad [X_1(x_0)^T W_1(x_0) X_1(x_0)]^{-1} \\ &= [X_1(x_0)^T W_1(x_0) X_1(x_0)]^{-1} X_1(x_0)^T \{\Sigma_1^{(1)}(x_0) + \Sigma_1^{(2)}(x_0)\} X_1(x_0) \\ &\quad [X_1(x_0)^T W_1(x_0) X_1(x_0)]^{-1}, \end{aligned} \tag{B.5}$$

where $\mathbf{t}_1(x_0) = \mathbf{m} - X_1(x_0)\boldsymbol{\beta}(x_0)$ is the vector remainders for expanding $m(x_i)$ around x_0. Writing $\hat{m}(x_0) = \mathbf{e}_1^T \hat{\boldsymbol{\beta}}(x_0)$, the result follows directly from (B.4) and (B.5).

Proof of Proposition 4.3.2

Note that,

$$\begin{aligned} E[r_i^2|\mathbf{X}] &= E[\hat{m}_i - m_i|\mathbf{X}]^2 \\ &= V[\sum_j p_{1,ij}y_j - y_i|\mathbf{X}] + E^2[\sum_j p_{1,ij}y_j - y_i|\mathbf{X}], \end{aligned} \tag{B.6}$$

where $p_{1,ij}$ are the (i,j)-th elements of P_1. Using matrix notations,

$$\begin{aligned} E[\mathbf{r}^2|\mathbf{X}] &= \{EP_1\mathbf{y}|\mathbf{X} - \mathbf{m}\}^2 + \sigma^2[1 + \text{diag}P_1P_1^T - 2\text{diag}P_1] \\ &\quad + [D + \text{diag}P_1DP_1^T - 2\text{daig}P_1D]. \end{aligned} \tag{B.7}$$

(B.8)

Proof of Theorem 4.3.3

The expression for the bias term is the same as in Theorem 4.3.1. We give proof for the variance of $\hat{\mathbf{v}}$. The following lemma is from McCullagh (1987):

Lemma B.0.1 *Let* $\mathbf{y}$ *be a random vector having all entries independent and* $\mathbf{m} = E(\mathbf{y})$, $V = diag\{E(\mathbf{y} - \mathbf{m})^2\}$, $G = diag\{E(\mathbf{y} - \mathbf{m})^3\}$, $T = diag\{E(\mathbf{y} - \mathbf{m})^4\}$ *then for any square matrix of constants,* A, *with the same number of rows as* $\mathbf{y}$

$$\begin{aligned} &Cov\{(A\mathbf{y})^2\} \\ &= Cov(diag\{A\mathbf{y}\mathbf{y}^TA\}) \\ &= (A \odot A)(T - 3V^2)(A \odot A)^T + 2diag\{(A\mathbf{m})AG(A \odot A)^T \\ &\quad + (A \odot A)GA^Tdiag\{A\mathbf{m}\}\} + 2(AVA^T) \odot (AVA^T) \\ &\quad + 4(AVA^T) \odot \{(A\mathbf{m})(A\mathbf{m})^T\} \end{aligned} \tag{B.9}$$

where $\odot$ *is the element-wise matrix multiplication.*

Note that,

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\sigma}_i^2|\mathbf{X}) &= \mathrm{Cov}\{\frac{P_2(\mathbf{r}^2-\boldsymbol{\Delta}_2)}{\mathbf{1}+P_2\boldsymbol{\Delta}_1}|\mathbf{X}\} && \text{(B.10)} \\ &= \frac{P_2\mathrm{Cov}\left[\{(P_1-I)\mathbf{y}\}^2|\mathbf{X}\right]P_2^T}{(\mathbf{1}+P_2\boldsymbol{\Delta}_1)(\mathbf{1}+P_2\boldsymbol{\Delta}_1)^T}. && \text{(B.11)} \end{aligned}$$

The variance in (4.25) is obtained from Lemma B.0.1 with $A = P_1 - I$ and $(P_1 - I)\mathbf{m} = \mathbf{b}$.

Proof of Theorem 4.3.4

The main steps for getting Theorem 4.3.4 from Theorem 4.3.1 are the following lemmas.

Lemma B.0.2 *Let* $S_{1,n,j} = \sum_{i=1}^n (X_i - x_0)^j k_{h_1}(X_i - x_0)$, $\mu_j = \int u^j k(u)du$, $S_1 = [\mu_{j+l}]_{0\le j+l\le p_1}$, $H_1 = diag\{1, h_1, ..., h_1^{p_1}\}$ *then*

$$n^{-1}S_{1,n,j} = h_1^j f_X(x_0)\mu_j\{1+o_P(1)\}, \qquad \textit{and} \tag{B.12}$$

$$n^{-1}S_{1,n} = f_X(x_0)H_1S_1H_1\{1+o_P(1)\}. \tag{B.13}$$

Lemma B.0.3 *Let* $S_{1,n,j}^{*(1)} = \sum_{j=1}^n (X_i - x_0)^j k_{h_1}^2(X_i - x_0)\sigma^2(X_i)$, $S_{1,n}^{*(1)} = [S_{1,n,j+l}^{*(1)}]_{0\le j+l\le p_1}$, $\nu_j = \int u^j k^2(u)du$, *and* $S_1^* = [\nu_{j+l}]_{0\le j+l\le p_1}$ *then*

$$\begin{aligned} n^{-1}S_{1,n,j}^{*(1)} &= h_1^{j-1}f_X(x_0)\sigma^2(x_0)\nu_j\{1+o_P(1)\}, \qquad \textit{and} && \text{(B.14)} \\ n^{-1}S_{n,j}^{*(1)} &= h_1^{-1}f_X(x_0)\sigma^2(x_0)H_1S_1^*H_1\{1+o_P(1)\}. && \text{(B.15)} \end{aligned}$$

Lemma B.0.4 *Let* $S_{1,n,j}^{*(2)} = \sum_{j=1}^n (X_i - x_0)^j k_{h_1}^2(X_i - x_0)D_i$, *and* $S_{1,n}^{*(2)} = [S_{1,n,j+l}^{*(2)}]_{0\le j+l\le p_1}$. *If* $n^{-1}\sum_i D_i$ *and* $n^{-1}\sum_i D_i^2$ *are bounded then*

$$\begin{aligned} n^{-1}S_{1,n,j}^{*(2)} &= \bar{D}h_1^{j-1}f_X(x_0)\nu_j\{1+o_P(1)\}, \qquad \textit{and} && \text{(B.16)} \\ n^{-1}S_{1,n}^{*(2)} &= h_1^{-1}f_X(x_0)\bar{D}H_1S_1^*H_1\{1+o_P(1)\}. && \text{(B.17)} \end{aligned}$$

Lemma B.0.5 *If* $m^{(p_1+2)}(\cdot)$ *and* $\mathbf{c}_{p_1} = (\mu_{p_1}, ..., \mu_{2p_1+1})^T$ *is bounded then*

$$n^{-1}X_{p_1}(x_0)^T W_{p_1}(x_0)\mathbf{t}_{1,x_0} = f_X(x_0)H_1\mathbf{c}_{p_1}h_1^{p_1+1}\beta_{p_1+1}\{1+o_P(1)\}. \tag{B.18}$$

Proof. The proof of Lemma (B.0.2), (B.0.3) and (B.0.5) follow directly from Fan and Gijbels (1996). We give a proof of Lemma (B.0.4). Note that,

$$\begin{aligned} E[K_{h_1}^2(X_i-x_0)(X_i-x_0)^j D_i] &= D_i h_1^{j-2}\int K^2(u)u^j f_x(x_0+uh_1)h_1 du \quad \text{(B.19)} \\ &= D_i h_1^{j-1}\{f_x(x_0)\nu_j + O(h_1)\}. \quad \text{(B.20)} \end{aligned}$$

Therefore,

$$E[n^{-1}S_{1,n,j}^{*(2)}] = \bar{D}h_1^{j-1}f_x(x_0)\nu_j\{1+O(h)\}. \tag{B.21}$$

Similarly,

$$E[n^{-1}\sum_i\{K_{h_1}^2(X_i-x_0)(X_i-x_0)^j\}^2] = \bar{D}^2 O(h_1^{2j-3}). \tag{B.22}$$

By Chebyshev's inequality,

$$\begin{aligned} n^{-1}S_{1,n,j}^{*(2)} &= E[n^{-1}S_{1,n,j}^{*(2)}] + O_p\left(V[n^{-1}S_{1,n,j}^{*(2)}]^{1/2}\right) \quad \text{(B.23)} \\ &= \bar{D}h_1^{j-1}f_X(x_0)\nu_j\{1+o_p(1)\}. \quad \text{(B.24)} \end{aligned}$$

By simple matrix multiplications, equation (B.18) is obtained.

Proof of Result (4.27): The conditional variance for $\hat{\boldsymbol{\beta}}(x_0)$ is

$$\begin{aligned} V[\hat{\boldsymbol{\beta}}(x_0)|\mathbf{X}] &= S_{1,n}^{-1}X_1(x_0)^T W_1(x_0)V[\mathbf{y}|\mathbf{X}]W_1(x_0)X_1(x_0)S_{1,n}^{-1} \quad \text{(B.25)} \\ &= S_{1,n}^{-1}S_{1,n}^{*(1)}S_{1,n}^{-1} + S_{1,n}^{-1}S_{1,n}^{*(2)}S_{1,n}^{-1} \quad \text{(B.26)} \\ &= \{nh_1 f_X(x_0)\}^{-1}\{\sigma^2(x_0)+\bar{D}\}H_1^{-1}S_1^{-1}S_1^*S_1^{-1}H_1^{-1}\{1+o_P(1)\}. \end{aligned}$$

Since

$$\hat{m}(x_0) = \mathbf{e}_1^T\hat{\boldsymbol{\beta}}(x_0), \tag{B.27}$$

result (4.27) is proven.

Proof of Result (4.26): Apply Lemma (B.0.5) to obtain an expression for the bias of $\hat{\boldsymbol{\beta}}(x_0)$.

$$\begin{aligned} & E[\{\hat{\boldsymbol{\beta}}(x_0) - \boldsymbol{\beta}(x_0)\}|\mathbf{X}] \\ &= S_{1,n}^{-1} X_1(x_0)^T W_1(x_0)\mathbf{t}(x_0) && \text{(B.28)} \\ &= [nf_X(x_0)H_1S_1H_1]^{-1} nh_1^{p_1+1} f_X(x_0)\beta_{p_1+1}(x_0)H\mathbf{c}_p\{1+o_p(1)\} && \text{(B.29)} \\ &= H_1^{-1}S_1^{-1}h_1^{p_1+1}\beta_{p_1+1}(x_0)\mathbf{c}_p\{1+o_p(1)\}, && \text{(B.30)} \end{aligned}$$

where $\beta_{p_1+1}(x_0) = \{(p_1+1)!\}^{-1} m^{(p_1+1)}(x_0)$. The result (4.26) is proven by (B.27) and (B.30).

Proof of Theorem 4.3.5

The following lemmas are applied to Theorem 4.3.3 to obtain the asymptotic expressions for the variance function.

Lemma B.0.6 *If $\sigma^{2(p_2+1)}(\cdot)$ is bounded in $N_\delta(x_0)$ then*

$$n^{-1}X_{2,x_0}^T W_{2,x_0}\mathbf{t}_{2,x_0} = f_X(x_0)H_2\mathbf{c}_{p_2}h_2^{p_2+1}\alpha_{p_2+1}\{1+o_P(1)\} \qquad \text{(B.31)}$$

Lemma B.0.7 *Suppose g has $p+2$ continuous derivatives and f is differentiable. Then*

$$\begin{aligned} diag(P_1) &= O_P\left((nh_1)^{-1}\right), && \text{(B.32)} \\ diag(P_2) &= O_P\left((nh_2)^{-1}\right), && \text{(B.33)} \\ P_1 diag_{1\le i\le n}\{g(x_i)\}P_1^T &= O_P\left((nh_1)^{-1}\right), \quad \textit{and} && \text{(B.34)} \\ P_2 diag_{1\le i\le n}\{g(x_i)\}P_2^T &= O_P\left((nh_2)^{-1}\right). && \text{(B.35)} \end{aligned}$$

Lemma B.0.8 *Let the conditions of Theorem 4.3.4 are satisfied, then*

$$(\frac{1}{\mathbf{1}+\mathbf{\Delta}_1})(\frac{1}{\mathbf{1}+\mathbf{\Delta}_1})^T = \mathbf{1} + O_P(nh_1)^{-1}. \tag{B.36}$$

Proof of Lemma B.0.6 is similar to the proof of Lemma B.0.5. Lemma B.0.7 is a direct consequence of Lemma B.0.2, Lemma B.0.3 and Lemma B.0.4. Lemma B.0.8 is obtained by repeated applications of Lemma B.0.7.

Proof of Result (4.28): By repeated applications of Lemma (B.0.7) and by assuming (A10), the dominating term in (4.24) is $(P_2 - I)\mathbf{v}$. Since P_2 is a smoother matrix similar to P_1, equation (4.28) is obtained from Lemma B.0.6.

Proof of Result (4.29): By (A12), $Ee_i^3 = 0$ and $Eu_i^3 = 0$. Therefore,

$$\begin{aligned}\mathrm{Cov}[\hat{\mathbf{v}}|\mathbf{X}] = {} & P_2[\{(P_1 - I) \odot (P_1 - I)\}(T - 3V^2)\{(P_1 - I) \odot (P_1 - I)\}^T \\ & + 2\{(P_1 - I)V(P_1 - I)^T\} \odot \{(P_1 - I)V(P_1 - I)^T\} \\ & + 4\{(P_1 - I)V(P_1 - I)^T\} \odot (\mathbf{b}\mathbf{b}^T)]P_2^T/\{(\mathbf{1} + P_2\mathbf{\Delta}_1)(\mathbf{1} + P_2\mathbf{\Delta}_1)^T\}\end{aligned} \tag{B.37}$$

Applying Lemma B.0.8 and Lemma B.0.7 repetitively and using assumption (A10), the leading term in the numerator of (B.37) is $P_2\{T - 3V^2 + 2V^2\}$. By assumption (A11), and by arguments similar to Lemma B.0.3 and Lemma B.0.4 result (4.29) is obtained.

Proof of Theorem 4.4.1

Note that,

$$\begin{aligned}\tilde{\theta}_i - \theta_i &= \gamma_i y_i + (1 - \gamma_i)\hat{m}_i - \theta_i \\ &= \gamma_i y_i + (1 - \gamma_i)m_i - \theta_i + (1 - \gamma_i)(\hat{m}_i - m_i) \\ &= \gamma_i(m_i + u_i + e_i) + (1 - \gamma_i)m_i - (m_i + u_i) + (1 - \gamma_i)(\hat{m}_i - m_i) \\ &= \{\gamma_i e_i - (1 - \gamma_i)u_i\} + \{(1 - \gamma_i)(\hat{m}_i - m_i)\}\end{aligned} \tag{B.38}$$

Any linear estimator of m_i is of the form $\sum_j a_{ij}y_j$ where a_{ij} are constants. Therefore the covariance term in (B.38) can be written as,

$$\begin{aligned}\mathrm{Cov}[\hat{m}_i, \gamma_i e_i - (1-\gamma_i)u_i] &= \mathrm{Cov}[\sum_j a_{ij}y_j, \gamma_i e_i - (1-\gamma_i)u_i] \\ &= \gamma_i a_{ii} D_i - (1-\gamma_i)a_{ii}\sigma_u^2 \\ &= a_{ii}\{\gamma_i D_i - (1-\gamma_i)\sigma_u^2\} \\ &= 0. \end{aligned} \tag{B.39}$$

In particular, for a local polynomial regression estimator of m_i,

$$a_{ij} = \mathbf{e}_1^T[[X_{p_1}(x_i)^T W_{p_1}(x_i) X_{p_1}(x_i)]]^{-1}\mathbf{x}_{j,x_i} w_j, \tag{B.40}$$

where $\mathbf{x}_{j,x_i} = (1, X_j - x_i, ..., (X_j - x_i)^{p_1})^T$ and w_j is the j^{th} diagonal element of $W_{p_1}(x_i)$. Accordingly,

$$\begin{aligned} E[(\tilde{\theta}_i - \theta_i)^2|\mathbf{X}] &= E\left[\{\gamma_i e_i - (1-\gamma_i)u_i\}^2|\mathbf{X}\right] \\ &\quad + E\left[\{(1-\gamma_i)(m_i - \hat{m}_i)\}^2|\mathbf{X}\right] \\ &= g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2), \end{aligned} \tag{B.41}$$

where $g_{1i}(\sigma_u^2) = E\left[\{\gamma_i e_i - (1-\gamma_i)u_i\}^2|\mathbf{X}\right]$ is the mean squared error if all the parameters are known, and $g_{2i}(\sigma_u^2) = E\left[\{(1-\gamma_i)(m_i - \hat{m}_i)\}^2|\mathbf{X}\right]$ is the mean squared error due to the estimation of the mean function m_i.

$$g_{1i}(\sigma_u^2) = E\left[\{\gamma_i e_i - (1-\gamma_i)u_i\}^2|\mathbf{X}\right] = (\sigma_u^2 + D_i)^{-1}\sigma_u^2 D_i, \tag{B.42}$$

and

$$g_{2i}(\sigma_u^2) = E\left[\{(1-\gamma_i)(m_i - \hat{m}_i)\}^2|\mathbf{X}\right] = (1-\gamma_i)^2 \mathrm{MSE}(\hat{m}_i). \tag{B.43}$$

We express

$$E[(\hat{\theta}_i - \theta_i)|\mathbf{X}]^2 = E[(\theta_i^* - \theta_i)|\mathbf{X}]^2 + E[(\tilde{\theta}_i - \theta_i^*)|\mathbf{X}]^2 + E[(\hat{\theta}_i - \tilde{\theta}_i)|\mathbf{X}]^2 + E[(\hat{\theta}_i - \tilde{\theta}_i)(\tilde{\theta}_i - \theta_i^*)|\mathbf{X}], \tag{B.44}$$

and provide a Taylor series approximation of the last two terms on the right side of (B.44).

Approximation of the $E(\hat{\theta}_i - \tilde{\theta}_i)^2$:

Our results are conditional on $\mathbf{X}$ and for notational convenience we write $E[Z|\mathbf{X}]$ as $E[Z]$, where Z is any random vector. By Proposition 4.3.2,

$$Er_i^2 = b_i^2 + \sigma_i^2(1 + \Delta_{1i}) + \Delta_{2i}, \tag{B.45}$$

where $\mathbf{r} = (r_1, r_2, \ldots, r_n)^T$, Δ_{1i} is the ith element of $\text{diag}\{P_1 P_1^T - 2P_1\}$ and Δ_{2i} is the ith element of $\text{diag}\{D + P_1 D P_1^T - 2P_1\}$. Let $(y_i - \hat{m}_i)^2 = r_i^2 = \hat{\kappa}_i$ and $Er_i^2 = \kappa_i$ and write

$$g(\hat{\kappa}_i, \hat{\sigma}_i^2) = (y_i - \hat{m}_i)^2(\hat{\gamma} - \gamma_i)^2 = \hat{\kappa}(\hat{\gamma}_i - \gamma_i)^2. \tag{B.46}$$

Note that,

$$\frac{\partial \hat{\sigma}_i^2}{\partial \hat{\kappa}_i} = \mathbf{e}_i^T \frac{P_2}{\mathbf{1} + P_2 \boldsymbol{\Delta}_1} \mathbf{e}_i \equiv t_{1i}, \tag{B.47}$$

$$\frac{\partial \hat{\gamma}_i}{\partial \hat{\kappa}_i} = \frac{\partial \hat{\gamma}_i}{\partial \hat{\sigma}_i^2} \frac{\partial \hat{\sigma}_i^2}{\partial \hat{\kappa}_i} = \frac{D_i}{(\hat{\sigma}_i^2 + D_i)^2} t_{1i}, \qquad \text{and} \tag{B.48}$$

$$\frac{\partial \hat{\kappa}_i}{\partial \hat{\sigma} \mathbf{i}} = \frac{\partial}{\partial \hat{\sigma}_i^2} \{\mathbf{e}_i^1 P_2^{-1} \mathbf{e}_i (\mathbf{1} + P_2 \boldsymbol{\Delta}_1) \hat{\boldsymbol{\sigma}}^2 + P_2^{-1} \boldsymbol{\Delta}_2\} = \mathbf{e}_i^T P_2^{-1} (\mathbf{1} + P_2 \boldsymbol{\Delta}_1) \mathbf{e}_i \equiv t_{2i} \tag{B.49}$$

Partial derivatives of $g(\hat{\kappa}_i, \hat{\sigma}_i^2)$ with respect to $\hat{\sigma}_i^2$ are obtained using the chain rule. Thus,

$$\frac{\partial}{\partial \hat{\sigma}_i^2} g(\hat{\kappa}_i, \hat{\sigma}_i^2) = \frac{\partial}{\partial \hat{\kappa}_i} g(\hat{\kappa}_i, \hat{\sigma}_i^2) \frac{\partial \hat{\kappa}_i}{\partial \hat{\sigma}_i^2} = t_{2i} \frac{\partial}{\partial \hat{\kappa}_i} g(\hat{\kappa}_i, \hat{\sigma}_i^2), \tag{B.50}$$

and

$$\frac{\partial}{\partial \hat{\kappa}_i} g(\hat{\kappa}_i, \hat{\sigma}_i^2) = \frac{\partial}{\partial \hat{\kappa}_i} \left\{\hat{\kappa}_i (\hat{\gamma}_i - \gamma_i)^2\right\} = (\hat{\gamma}_i - \gamma_i)^2 + 2\hat{\kappa}_i (\hat{\gamma}_i - \gamma_i) \frac{D_i}{(\hat{\sigma}_i^2 + D_i)^2} t_{1i}. \tag{B.51}$$

Similarly, the higher order derivatives are also obtained using the chain rule. Note that the 8-th moments of the random components are finite by the normality assumption. Hence a two step Taylor series expansion of $g(\hat{\kappa}_i, \hat{\sigma}_i^2)$ around κ_i, σ_i is given by,

$$E[(y_i - \hat{m}_i)^2 (\hat{\gamma}_i - \gamma_i)^2]$$

$$= t_{1i}^2\kappa_i\frac{D_i^2}{(\sigma_i^2+D_i)^4}E(y_i-\hat{m}_i)^4 + 2t_{1i}\kappa_i\frac{D_i^2}{(\sigma_i^2+D_i)^4}E\{(y_i-\hat{m}_i)^2(\hat{\sigma}_i^2-\sigma_i^2)\}$$
$$+\kappa_i\frac{D_i^2}{(\sigma_i^2+D_i)^4}E(\hat{\sigma}_i^2-\sigma_i^2)^2 + O_p(a_{nh}), \tag{B.52}$$

where $a_{nh} = \max\{(nh_2)^{-3/2}, h_2^{3(p_2+1)}\}$. By arguments similar to the asymptotic bias of $\hat{\boldsymbol{\sigma}}^2$,

$$[\frac{P_2}{\mathbf{1}+P_2\boldsymbol{\Delta}_1}][\mathbf{b}^2+\boldsymbol{\sigma}^2(\mathbf{1}+\boldsymbol{\Delta}_1)+\boldsymbol{\Delta}] = O_p(h_2^{p_2+1}). \tag{B.53}$$

Therefore,

$$t_{1i}^2\kappa_i(\sigma_i^2+D_i)^{-4}D_i^2E(y_i-\hat{m}_i)^4 = o_p(h_2^{p_2+1}). \tag{B.54}$$

Consider the product term in (B.52),

$$E\{(y_i-\hat{m}_i)^2(\hat{\sigma}_i^2-\sigma_i^2)\} = \hat{\kappa}(\hat{\gamma}_i-\gamma_i). \tag{B.55}$$

By a one step Taylor series approximation,

$$E\{(y_i-\hat{m}_i)^2(\hat{\gamma}-\gamma_i)\}$$
$$= t_{1i}\kappa_i\frac{D_i}{(\sigma_i^2+D_i)^2}E(y_i-\hat{m}_i)^2 + \kappa_i\frac{D_i}{(\hat{\sigma}_i^2+D_i)^2}E(\hat{\sigma}_i^2-\sigma_i^2) + O_p(\lambda_{nh})$$
$$= t_{1i}\kappa_i^2\frac{D_i}{(\sigma_i^2+D_i)^2} + \kappa_i\frac{D_i}{(\hat{\sigma}_i^2+D_i)^2}\ \ Bias\ (\hat{\sigma}_i^2) + O_p(\lambda_{nh}), \tag{B.56}$$

where $\lambda_{nh} = \max\{(nh_2)^{-1}, h_2^{p_2+1}\}$. Therefore,

$$t_{1i}\kappa_i\frac{D_i^2}{(\sigma_i^2+D_i)^4}E\{(y_i-\hat{m}_i)^2(\hat{\sigma}_i^2-sigma_i^2)^2\}$$
$$= t_{1i}\kappa_i\frac{D_i^2}{(\sigma_i^2+D_i)^4}\left\{t_{1i}\kappa_i^2\frac{D_i}{(\sigma_i^2+D_i)^2} + \kappa_i\frac{D_i}{(\hat{\sigma}_i^2+D_i)^2}\ \ Bias\ (\hat{\sigma}_i^2) + O_p(\lambda_{nh})\right\}$$
$$= o_p(\lambda_{nh}), \tag{B.57}$$

since,

$$\mathbf{t}_1 = \frac{P_2}{\mathbf{1}+P_2\boldsymbol{\Delta}_1}\ Bias\ (\hat{\sigma}_i^2) = o_p(h^{p_2+1}). \tag{B.58}$$

Therefore, by (B.52), (B.54) and (B.57) and by using $\kappa_i = b_i^2+\sigma_i^2(1+\Delta_{1i})+\Delta_{2i}$

$$E[(\hat{\theta}_i-\tilde{\theta}_i)|\mathbf{X}]^2$$

$$= g_{3i}(\sigma_i^2) + O_p(a_{nh})$$

$$= \{b_i^2 + \sigma_i^2(1+\Delta_{1i}) + \Delta_{2i}\}(\sigma_i^2 + D_i)^{-4} D_i^2 E[\hat{\sigma}_i^2 - \sigma_i^2 | \mathbf{X}] + O_p(a_{nh}). \quad \text{(B.59)}$$

Approximation of the $E[(\hat{\theta}_i - \tilde{\theta}_i)(\tilde{\theta}_i - \theta_i^*)|\mathbf{X}]$:

Note that,

$$(\hat{\theta}_i - \tilde{\theta}_i)(\tilde{\theta}_i - \theta_i) = (y_i - \hat{m}_i)(\gamma_i - \hat{\gamma}_i)\{(y_i - \hat{m}_i)(\gamma_i - 1) + e_i\}. \quad \text{(B.60)}$$

Since $E[e_i|\mathcal{F}, \mathbf{X}] = 0$, using the conditional expectation,

$$E[(\hat{\theta}_i - \tilde{\theta}_i)(\tilde{\theta}_i - \theta_i)] = (1-\gamma_i)E[(y_i - \hat{m}_i)^2(\hat{\gamma}_i - \gamma_i)]. \quad \text{(B.61)}$$

Since the error components are independently and normally distributed and u_i's and e_i's are independent, we derive the following covariances.

$$\text{Cov}[(u_i + e_i)^2, u_k u_l] = \begin{cases} 2\sigma_i^4, & i = k = l, \\ 0, & ow. \end{cases} \quad \text{(B.62)}$$

$$\text{Cov}[(u_i + e_i)^2, e_k u_l] = \begin{cases} 2D_i\sigma_i^2, & i = k = l, \\ 0, & ow. \end{cases} \quad \text{(B.63)}$$

and

$$\text{Cov}[(u_i + e_i)^2, e_k e_l] = \begin{cases} 2D_i^2, & i = k = l, \\ 0, & ow. \end{cases} \quad \text{(B.64)}$$

From (B.62), (B.63) and (B.64),

$$\text{Cov}[(u_i + e_i)^2, y_k y_l] = \begin{cases} 2(\sigma_i^2 + D_i)^2, & i = k = l \\ 0, & ow. \end{cases} \quad \text{(B.65)}$$

Write $\hat{\sigma}_i^2 = \sum_{j=1}^n \delta_{ij}(y_j - \hat{m}_j)^2$, $\hat{m}_i = \sum_{j=1}^n a_{ij} y_j$, where a_{ij} is the ij-th element of P_1 and δ_{ij} is the ij-th element of $(\mathbf{1} + P_2\mathbf{\Delta}_1)^{-1}P_2$. Therefore,

$$\text{Cov}[(y_i - m_i)^2, (\hat{\sigma}_i^2 - \sigma_i^2)]$$

$$= \text{Cov}\left[(u_i + e_i)^2, \sum_j \delta_{ij} y_j^2 + \sum_j \delta_{ij} \sum_k \sum_l a_{jk} a_{kl} y_k y_l - 2\sum_j \delta_{ij} y_j \sum_k a_{jk} y_k\right]$$

$$= 2(\sigma_i^2 + D_i)^2 \delta_{ii}(1 - a_{ii})^2. \quad \text{(B.66)}$$

By a Taylor series expansion of $\hat{\gamma}_i$ around σ_i^2,

$$E[\hat{\gamma}_i - \gamma_i] = (D_i + \sigma_i^2)^{-1} D_i E[\hat{\sigma}_i^2 - \sigma_i^2] + O_P(a_{nh}), \tag{B.67}$$

and using (B.66) and assumption (A10),

$$\mathrm{Cov}[(y_i - \hat{m}_i)^2, (\hat{\gamma}_i - \gamma_i)] = 2D_i\delta_{ii}(1 - a_{ii})^2 + O_P(a_{nh}) = O_P(a_{nh}), \tag{B.68}$$

where $a_{nh} = \max\{h_2^{2p_2+2}, (nh_2)^{-2}\}$. Further,

$$E[y_i - \hat{m}_i]^2 = b_i^2 + \sigma_i^2(1 + \Delta_{1i}) + \Delta_{2i}, \tag{B.69}$$

where b_i is the bias for estimating the mean function m_i, and Δ_{1i} and Δ_{2i} are defined in Chapter 4. Hence,

$$\begin{aligned} & E[(y_i - \hat{m}_i)^2(\hat{\gamma}_i - \gamma_i)] \\ &= \mathrm{Cov}[(y_i - \hat{m}_i)^2, (\hat{\gamma}_i - \gamma_i)] + E[(y_i - \hat{m}_i)^2]E[\hat{\gamma}_1 - \gamma_i] \\ &= 2D_i\delta_{ii}(1 - a_{ii})^2 + (D_i + \sigma_i^2)^{-2} D_i\{b_i^2 + \sigma_i^2(1 + \Delta_{1i}) + \Delta_{2i}\}\mathrm{Bias}(\hat{\sigma}_i^2) + O_P(a_{nh}). \end{aligned} \tag{B.70}$$

Accordingly,

$$\begin{aligned} & E[(\hat{\theta}_i - \tilde{\theta}_i)(\tilde{\theta}_i - \theta_i)] \\ &= (D_i + \sigma_i^2)^{-3} D_i^2\{b_i^2 + \sigma_i^2(1 + \Delta_{1i}) + \Delta_{2i}\}\mathrm{Bias}(\hat{\sigma}_i^2) + O_P(a_{nh}). \end{aligned} \tag{B.71}$$

Result (4.33) follows directly from (B.42), (B.43), (B.59), (B.71).

Results (4.39), (4.41), and (4.41) follow from (4.26), (4.27), (4.36), (4.37), (4.28), and (4.29).

APPENDIX C. Proof of Chapter 5

Proof of proposition 5.3.1

Proof of (R1) :

$$\begin{aligned} E[\bar{y}_{i..}|\mathcal{F}] &= E[E(\bar{y}_{i..}|A_1,\mathcal{F})|\mathcal{F}] \\ &= E[E\{E(\bar{y}_{i..}|A_r,A_1,\mathcal{F})|A_1,\mathcal{F}\}|\mathcal{F}]. \end{aligned} \tag{C.1}$$

By (A3),

$$E(\bar{y}_{i..}|A_r,A_1,\mathcal{F}) = N_{i+}^{-1}\sum_{g=1}^{G}\{\sum_{k\in A_{rig}} w_{igk}y_{igk}\}\{(\sum_{k\in A_{rig}} w_{igk})^{-1}\sum_{k\in A_{1ig}} w_{igk}\} \tag{C.2}$$

is a two phase estimator of the population mean θ_i in county i, hence using (A1) and following Cochran (1977)

$$E[E(\bar{y}_{i..}|A_r,A_1,\mathcal{F})|A_1,\mathcal{F}] = N_{i+}^{-1}\sum_{g=1}^{G}\sum_{k\in A_{1ig}} w_{igk}y_{igk} + O_p(n_{1ig}^{-1}). \tag{C.3}$$

The result follows from (C.1) and (C.3) and since $E[N_{i+}^{-1}\sum_{g=1}^{G}\sum_{k\in A_{1ig}} w_{igk}y_{igk}|\mathcal{F}] = \theta_i$.

Proof of (R2) :

$$V[\bar{y}_{i..}|\mathcal{F}] = V[E(\bar{y}_{i..}|A_r,A_1,\mathcal{F})|\mathcal{F}] + E[V(\bar{y}_{i..}|A_r,A_1,\mathcal{F})|\mathcal{F}]. \tag{C.4}$$

The first term in (C.4),

$$V[E(\bar{y}_{i..}|A_r, A_1, \mathcal{F})|\mathcal{F}] = V[E\{E(\bar{y}_{i..}|A_r, A_1, \mathcal{F})|A_1, \mathcal{F}\}|\mathcal{F}] + E[V\{E(\bar{y}_{i..}|A_r, A_1, \mathcal{F})|A_1, \mathcal{F}\}|\mathcal{F}] \quad (C.5)$$

is the variance from a two phase estimator of the mean. Following Cochran (1977) and from (C.3) and (C.5),

$$V[E(\bar{y}_{i..}|A_r, A_1, \mathcal{F})|\mathcal{F}] \quad (C.6)$$

$$= V[\sum_g \sum_{k\in A_{1ig}} w_{igk} y_{igk}|\mathcal{F}] + N_{i+}^{-2} \sum_g E[p_g(1-p_g) \sum_{k\in A_{1ig}} \{w_{igk}y_{igk} - R_{A_{1ig}} w_{igk}\}^2|\mathcal{F}] + O_p(n_{1i+}^{-3/2}). \quad (C.7)$$

By (A3),

$$E[V(\bar{y}_{i..}|A_r, A_1, \mathcal{F})|\mathcal{F}] = n_{mi+} n_{1i+}^{-2} (n_{ri+} - 1) n_{ri+}^{-1} (\sum_g n_{1i+}^{-1} n_{mig} \sigma_g^2). \quad (C.8)$$

Combining (C.4), (C.7) and (C.8) we get (5.19). (5.20) follows from (5.19) since $w_{gik} = N_{i+}^{-1} n_{1i}$ for SRSWOR and $p_g = n_{1+g}^{-1} n_{r+g}$. The consistency follows by applying Slutzky's theorem if $\hat{\sigma}_g^2$ is a consistent estimator for σ_g^2.

Proof of (R3) :

The covariance of estimated means of two counties i and j can be written as

$$\mathrm{Cov}[(\bar{y}_{i..}, \bar{y}_{j..})|\mathcal{F}] = \mathrm{Cov}[\{E(\bar{y}_{i..}|A_r, A_1, \mathcal{F}), E(\bar{y}_{j..}|A_r, A_1, \mathcal{F})\}|\mathcal{F}] + E[\mathrm{Cov}\{(\bar{y}_{i..}, \bar{y}_{j..})|A_r, A_1, \mathcal{F}\}|\mathcal{F}]. \quad (C.9)$$

The first term in equation (C.9) is zero by (A2). By (A4),

$$\begin{aligned}
&\mathrm{Cov}[(\bar{y}_{i..}, \bar{y}_{j..})|A_r, A_1, \mathcal{F}] \\
&\quad = \mathrm{Cov}\Big[(N_{i+}^{-1}\{\textstyle\sum_{g=1}^{G}\sum_{k\in A_{rig}} w_{igk}y_{igk} + \sum_{g=1}^{G}\sum_{k\in A_{mig}} w_{igk}z_{igk}\}, \\
&\qquad N_{j+}^{-1}\{\textstyle\sum_{g=1}^{G}\sum_{k\in A_{rjg}} w_{jgk}y_{jgk} + \sum_{g=1}^{G}\sum_{k\in A_{mjg}} w_{jgk}z_{jgk}\}|\mathcal{F}\Big] \\
&\quad = (N_{i+}^{-1}N_{j+}^{-1})\Big[\mathrm{Cov}(\textstyle\sum_g\sum_{k\in A_{rig}} w_{igk}y_{igk}, \sum_g\sum_{k\in A_{mjg}} w_{jgk}z_{jgk})|\mathcal{F} \\
&\qquad +\mathrm{Cov}(\textstyle\sum_g\sum_{k\in A_{mig}} w_{igk}z_{igk}, \sum_g\sum_{k\in A_{rjg}} w_{jgk}y_{jgk})|\mathcal{F}\Big].
\end{aligned}$$

Noting that the covariance in the last equation exists only through the common observations between the set of respondent points and missing points and assuming SRSWOR within each county, the use of a little algebra can show that

$$E[\mathrm{Cov}\{(\bar{y}_{i..}, \bar{y}_{j..})|A_r, A_1, \mathcal{F}\}|\mathcal{F}] = (n_{1i+}^{-1}n_{1j+}^{-1})\sum_g n_{1+g}\mathrm{V}_*(\bar{y}_{1.g.})(\tau_{ijg} + \tau_{jig}). \tag{C.10}$$

Finally, using a cell model as in (A3), and using (C.9) and (C.10) we have,

$$\mathrm{Cov}[(\bar{y}_{i..}, \bar{y}_{j..})|\mathcal{F}] = (n_{1i+}^{-1}n_{1j+}^{-1})\sum_g \sigma_g^2(\tau_{ijg} + \tau_{jig}).\square \tag{C.11}$$

BIBLIOGRAPHY

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satelite data. *Journal of the American Statistical Association*, 83:28–36.

Bell, A., Drignei, D., Dorsch, R. K., Fuller, W. A., Keinzler, J., Maiti, T., Nusser, S. M., Peterson, T. C., and Wolter, K. (2003). Estimation procedure for 2002 NRI. Technical report, Center for Survey Statistics and Methodology, National Resource Conservation Service.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1026–1053.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Duxbury, Pacific Grove, USA.

Chu, C. K. and Marron, J. S. (1991). Choosing a kernel regression estimator (with discussion). *Statistical Science*, 6:404–436.

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley & Sons, New York, NY.

Datta, G. S., Ghosh, M., Huang, E. T., Isaki, C. T., Shultz, L. K., and Tsay, J. H. (1992). Hierarchical and empirical bayes methods for adjustment of census undercount: The 1998 missouri dress rehearsal data. *Survey Methodology*, 18:95–100.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall: London.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74:269–277.

Fuller, W. A. (1996). *Introduction to Statistical Time Series.* John Wiley & Sons, Inc., New York, NY, second edition.

Fuller, W. A. (2003). Sample selection for the 2000 NRI-2004 NRI surveys. Unpublished Manuscript, Center for Survey Statistics and Methodology.

Fuller, W. A. (2006). Sampling statistics. Unpublished Manuscript, Center for Survey Statistics and Methodology.

Gasser, T., Kneip, A., and Kohler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association*, 86:643–652.

Hagen, L. J. (1994). Wind erosion in the united states. In *Proceedings of Wind Erosion Symposium*, pages 25–32. Poznan, Poland.

Hagen, L. J. and Woodruff, N. P. (1973). Air pollution from dust storms in the great plains. *Atmospheric Environment*, 7:323–332.

Härdle, W. (2002). *Applied non-parametric regression.* Cambridge University Press, Cambridge.

Hastie, T. J. and Loader, C. (1993). Local regression: Automatic kernel carpentry (with discussion). *Statistical Science*, 8:120–143.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman & Hall, London.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics*, 21:309–310.

McCullagh, P. (1987). *Tensor Methods in Statistics.* Chapman & Hall, London.

Nadaraya, E. A. (1964). On estimating regression. *Theory Prob. Appl.*, 9:141–142.

Nusser, S. M. and Goebel, J. (1997). The national resource inventory: a long-term multi-resource monitoring program. *Environmental and Ecological Statistics*, 4:181–204.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of the small area estimators. *Journal of the American statistical association*, 85:163–171.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley Series in Survey Methodology.

Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79:811–822.

Rosewell, C. (1993). A program to assist in the selection of management practices to reduce erosion. Technical report, Soil Conservation Service.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effictive bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90:1257–1270.

Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997). Local polynomial variance fuction estimation. *Technometrics*, 39(3):262–273.

Särndal, C. E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18(2):241–252.

Särndal, C. E., Swensson, B., and Wretman, J. (1991). *Model assisted survey sampling.* Springer, New York, NY.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components.* John Wiley & Sons Inc., New York, NY.

Slud, E. V. and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society Series B*, 68(2):239 – 258.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman & Hall/CRC, Boca Raton, FL.

Wang, J. and Fuller, W. A. (2002). Small area estimation under a restriction. In *Proceedings of the Joint Statistical Meetings*, volume CD-ROM.

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

ACKNOWLEDGEMENTS

This research was supported in part by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

The completion of this thesis would not have been possible without the support and generosity of several people.

I would first like to thank my adviser, Dr. Tapabrata Maiti, for giving me the opportunity to work in a very interesting area and for his support and guidance throughout the duration of this research. He has been generous both with his invaluable advice and encouragements. I thank Dr. Wayne A. Fuller for his guidance in many areas of this research. His insights in analyzing statistical problems have always inspired me. I thank Dr. Jean D. Opsomer for his suggestions on the development of the local polynomial estimators. I thank Dr. Sarah M. Nusser for her advice during the initial stages of my graduate career. I also thank Dr. Soumendra N. Lahiri and Dr. Leslie Miller for their time and effort.

I am especially grateful to Desireé Moffitt for her suggestions throughout the writing of this thesis. I would like to thank the NRI staff for their support during data analysis and the CSSM for the financial support to complete this dissertation. I would like to thank my friends and colleagues in Snedecor. I would also like to thank Jason C. Legg for his suggestions.

Finally, I want to thank my parents for their constant inspiration and encouragements.