AN ABSTRACT OF THE THESIS OF

George W. Weaver for the degree of Doctor of Philosophy in Statistics presented on October 2, 1996. Title: Model Based Estimation of Parameters of Spatial Populations from Probability Samples.

Redacted for Privacy

Abstract Approved: _____

W. Scott Overton

Many ecological populations can be interpreted as response surfaces; the spatial patterns of the population vary in response to changes in the spatial patterns of environmental explanatory variables. Collection of a probability sample from the population provides unbiased estimates of the population parameters using design based estimation. When information is available for the environmental explanatory variables, model based procedures are available that provide more precise estimates of population parameters in some cases. In practice, not all of these environmental explanatory variables will be known. When the spatial coordinates of the population units are available, a spatial model can be used as a surrogate for the unknown, spatially patterned explanatory variables. Design based and model based procedures will be compared for estimating parameters of the population of Acid Neutralizing Capacity (ANC) of lakes in the Adirondack Mountains in New York. Results from the analysis of this population will be used to elucidate some general principles for model based estimation of parameters of spatial populations. Results indicate that using model based estimates of population parameters provide more precise estimates than design based estimates in some cases. In addition, including spatial information as a surrogate for spatially patterned missing covariates improves the precision of the estimates in some cases, the degree to which depends upon the model chosen to represent the spatial pattern.

When the probability sample is selected from the spatial population is a stratified sample, differences in stratum variances need to be accounted for when residual spatial covariance estimation is desired for the entire population. This can be accomplished by scaling residuals by their estimated stratum standard deviation functions, and calculating the residual covariance using these scaled residuals. Results here demonstrate that the form of scaling influences the estimated strength of the residual correlation and the estimated correlation range.

# Model Based Estimation of Parameters of Spatial Populations from Probability Samples

by

George W. Weaver

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented October 2, 1996
Commencement June 1997

Doctor of Philosophy thesis of George W. Weaver presented on October 2, 1996

APPROVED:

Redacted for Privacy

Major Professor, representing Statistics

Redacted for Privacy

Chair of Department of Statistics

Redacted for Privacy

Dean of Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Redacted for Privacy

George W. Weaver, Author

ACKNOWLEDGEMENTS

I am indebted to Dr. W. Scott Overton for his encouragement, commitment and support during my graduate studies. This research was a continuation of work by Dr. W. Scott Overton and Henrietta Jager, and I am grateful to them for posing the specific questions to be addressed and identifying possible solutions. I thank Henrietta Jager for providing the data set, Dr. David Birkes for his GAUSS program for spatial prediction, and Dr. W. Scott Overton, Dr. David Birkes and Dr. David Thomas for many helpful discussions and suggestions.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

# LIST OF TABLES

For my wife Carol, without whose prayers, support
and sacrifice this would not have been possible.

"...the writing of many books is endless, and excessive devotion to books is wearying to the body. The conclusion, when all has been heard is: fear God and keep His commandments, because this applies to every person."
Ecclesiastes 12:12-13.

# Model Based Estimation of Parameters of Spatial Populations from Probability Samples.

## 1. Introduction.

### 1.1 Overview.

The main objective of this document is model-based estimation of the parameters of spatial populations. We assume that our data consists of a probability sample taken from a finite spatial population where there are covariates known for the entire population. A general protocol will be developed for the estimation of the distribution function (abbreviated as the cumulative distribution function, CDF) and other parameters from a probability sample using model based methodology. The protocol consists of four steps: (1) estimation of parameters of a statistical model using the sample data and known covariate information, (2) prediction of population values for the unsampled population units, (3) estimation of the population CDF using the sample data and predicted values for the unsampled population units, and (4) use the CDF from (3) to calculate estimates of other population parameters, such as the mean and standard deviation. Several issues will be addressed: the effect of a missing covariate on a model based analysis, the effect of stratified sampling on spatial covariance estimation and using spatial information to estimate the finite population CDF and other parameters. These issues will be investigated using simulated data as well as actual survey data. The hypothetical populations are patterned after the descriptions of the population sampled by the actual data. The remainder of this introduction describes the perspective of spatial populations to be used here and gives a brief overview of probability sampling, design based estimation and some basic regression analysis results that will be used, particularly relating to prediction. Chapter 2 begins with a discussion of

regression analysis of stratified sample data. It then gives an overview of spatial linear models, especially for the purpose of spatial prediction, and concludes with an investigation of the effect of stratification on spatial covariance estimation. Chapter 3 begins with a review of a model based CDF estimator to be considered here, and shows how it can be applied to spatial populations, and then its performance is assessed with some simulated populations. The chapter concludes with a discussion of the uses of spatial information for estimating parameters of finite, spatial populations. Chapter 4 applies the protocols assessed in chapters 2 and 3 to an actual probability sample from the Eastern Lake Survey (USEPA). Chapter 5 contains an overall summary of these research results.

## 1.2 Spatial Populations.

A population Y is defined as a function on a Universe U = {u:$\cdot$u$\epsilon$U} such that Y = {y(u):u$\epsilon$U}. A **spatial** population is then defined as any population that is 'geo-referenced', i.e. for each individual unit in the universe, the spatial location of the unit is known in some spatial coordinate system. In this document, the coordinate system will be based on Latitude and Longitude, which will be labeled as $l_1$ and $l_2$ respectively. When choosing a coordinate system for geo-referencing a spatial population, it is important to take into account the influence of the coordinate system on the representation of geographic distances between the population units. For latitude and longitude, there is a distortion of distances for different directional orientations, i.e. one degree of longitude doesn't correspond to the same geographic distance as one degree of latitude (unless the data are taken at the earth's equator). In particular, sample data combined by distance classes to estimate the covariance function will be incorrect. When latitude and longitude are used, the distance classes will be grouped incorrectly depending upon the directional orientation of the pair. Therefore, $l_1$ and $l_2$ will be transformed to a Cartesian coordinate system so that calculated distances and plotted maps will accurately represent geographic distances. We will label these transformed coordinates $i_u$ and $j_u$, representing the spatial location of population

unit $u$.

Additionally, spatial populations are interpreted here as varying in response to spatially patterned explanatory variables which, if known, could be used to predict the spatial pattern of the population. These variables are never completely known in practice, but some may be available from ecological surveys, geographic information system (GIS) data layers or from meteorological data. Because the spatial coordinates of the members of the universe are known, a spatial model can be used as a surrogate for the unknown spatially patterned explanatory variables. The perspective taken here can be referred to as a response surface perspective (Jager and Overton, 1991); the spatial patterns of the population vary in response to spatially patterned causal variables. Suppose the causal variables $x_u$, $z_u$ are known and the spatial population of interest is $y_u$. The model that we will use to express this response surface is:

$$\begin{aligned} y_u &= \mu + x_u\beta + z_u\alpha + \epsilon_u, \\ &\quad E(\epsilon_u) = 0, \\ &\quad Var(\epsilon_u) = \sigma^2, \\ &\quad Cov(\epsilon_u , \epsilon_{u'}) = 0, \quad u \neq u'. \end{aligned} \tag{1}$$

For some populations, we may also want to allow the residuals to be heteroscedastic: $Var(\epsilon_u) = g(E(y_u);\gamma)\sigma^2$. In almost all cases, some of the covariates will be unknown. This perspective differs from thinking about spatial populations as spatially autocorrelated, where there is assumed to be some spatial connectivity or association between the population units that causes their response values to be similar. These two perspectives are quite different and are reflective of the way that researchers are thinking about the nature of the phenomenon under study. If the population units are functionally linked in some fashion, then we are using the spatial autocorrelation perspective. Examples of populations that fit into this framework are the distribution of water temperature in a lake, water chemistry samples from an aquifer or moisture content samples from a soil plot. The model that we will use to express this perspective is:

$$\begin{aligned} y_u &= \mu + x_u\beta + i_u\phi_1 + j_u\phi_2 + \epsilon_u, \\ &\quad E(\epsilon_u) = 0, \\ &\quad Var(\epsilon_u) = \sigma^2, \\ &\quad Cov(\epsilon_u , \epsilon_{u'}) = \sigma(d_{uu'}; \theta), \quad u \neq u'. \end{aligned} \tag{2}$$

$$d_{uu'}=\sqrt{(i_u-i_{u'})^2+(j_u-j_{u'})^2}$$

In general, any function of the spatial coordinates could be included in the regression equation, but only linear functions will be considered here. Notice that this model still allows $y_u$ to depend upon a known covariate $x_u$. Also, for some populations we may want to allow the variances of the residuals to be heteroscedastic; $\mathrm{Var}(\epsilon_u) = g(E(y_u);\gamma)\sigma^2$. If one of the covariates, say $z_u$ is unknown but is spatially patterned, we will use the spatial information in model (2) as a surrogate for the spatially patterned, missing covariate.

## 1.3 Missing Covariate Effect.

To illustrate the effect of a missing covariate on residual spatial covariance, suppose a spatial population follows the linear model (1) but with some additional assumptions about the covariates:

$$\begin{aligned} y_u = \mu + x_u\beta + z_u\alpha + \epsilon_u, \\ & E(\epsilon_u) = 0, \\ & \mathrm{Var}(\epsilon_u) = \sigma^2, \\ & \mathrm{Cov}(\epsilon_u\ ,\ \epsilon_{u'}) = 0,\ u \neq u'. \\ & \qquad (3) \\ & E(z_u) = \delta, \\ & \mathrm{Var}(z_u) = \tau^2, \\ & \mathrm{Cov}(z_u\ ,\ z_{u'}) = \tau(d_{uu'};\ \theta),\ u \neq u'. \end{aligned}$$

And assume that z and $\epsilon$ are uncorrelated. The stated assumptions about $z_u$ are for mathematical convenience, the only necessary assumption is that $z_u$ is spatially patterned and that the model expression for the variance and covariance of $z_u$ provides an adequate approximation. Now suppose that $z_u$ is unknown, then we are actually analyzing the following population:

$$y_u = \mu^* + x_u\beta + \epsilon_u^*, \qquad (4)$$

$$\mu^* = \mu + \alpha\delta,$$
$$\mathrm{E}(\epsilon_u^*) = 0,$$
$$\mathrm{Var}(\epsilon_u^*) = \alpha^2\tau^2+\sigma^2,$$
$$\mathrm{Cov}(\epsilon_u^*,\epsilon_{u'}^*) = \alpha^2\tau(\mathrm{d}_{uu'}\,;\,\theta),\ u \neq u'.$$

To see this, note that:

$\mathrm{Var}(\epsilon_u^*) = \mathrm{Var}(\alpha z_u + \epsilon_u) = \mathrm{Var}(\alpha z_u) + \mathrm{Var}(\epsilon_u) = \alpha^2\tau^2+\sigma^2$, and

$$\begin{aligned}
\mathrm{Cov}(\epsilon_u^*,\epsilon_{u'}^*) &= \mathrm{Cov}(z_u\alpha+ \epsilon_u\ ,\ z_{u'}\alpha+\epsilon_{u'}) \\
&= \mathrm{Cov}(z_u\alpha,\ z_{u'}\alpha) + \mathrm{Cov}(z_u\alpha,\ \epsilon_{u'}) \\
&+ \mathrm{Cov}(\epsilon_u,\ z_{u'}\alpha) + \mathrm{Cov}(\epsilon_u,\ \epsilon_{u'}) \\
&= \alpha^2\mathrm{Cov}(z_u,\ z_{u'}) + 0 + 0 + 0 \\
&= \alpha^2\tau(\mathrm{d}_{uu'};\ \theta).
\end{aligned}$$

Notice that the missing covariate induces spatial covariance and increased variance in the residuals. If the missing covariates are not spatially patterned, they still induce increased variance in the residuals. In this latter case, the model that we will use to analyze the population is:

$$\mathrm{y}_u = \mu + \mathrm{x}_u\beta + \epsilon_u, \qquad (5)$$
$$\mathrm{E}(\epsilon_u) = 0,$$
$$\mathrm{Var}(\epsilon_u) = \sigma^2,$$
$$\mathrm{Cov}(\epsilon_u\ ,\epsilon_{u'}\ ) = 0,\ u \neq u'.$$

Similarly to model (2) we will allow the variances to be heteroscedastic; $\mathrm{Var}(\epsilon_u) = \mathrm{g}(\mathrm{E}(\mathrm{y}_u);\gamma)\sigma^2$. For the spatial populations to be analyzed here, model (1) will be the model that motivates our thinking about the populations under study. Models (2) and (5) will be used as tools to analyze samples from these populations. In all the cases considered here, isotropic spatial covariance models will be used (i.e. the form of spatial covariance does not depend on the directional orientation of the distance between the pair). The procedures generalize to include anisotropic covariance models, but these will not be considered here explicitly.

## 1.4 Probability Samples and Design Based Estimation.

In addition to having a statistical model (1) that we are using to motivate our thinking about the populations under study, we assume that the data we have collected from the population is a probability sample, defined as follows:
Definition:

A probability sample, where $\pi_u$ is the inclusion probability of element $u$ of the universe, will be defined according to the weak definition of Overton (1991) :

"A probability sample is a subset of the universe
selected by an explicit protocol so that $\pi_u$ is:
a) Known for each element in the realized sample
b) Positive for each element in the universe
c) $\pi_{uu'}$ is known on the sample "

Given a probability sample S, design based estimation proceeds according to the theorem of Horvitz and Thompson (1952):

If $S \subset U$ is selected such that $\pi_u > 0$ for all $u \epsilon U$,
then $\widehat{T}_y = \sum_S y_u/\pi_u$ is unbiased for $T_y = \sum_U y_u$, and

$$V(\widehat{T}_y) = \sum_U y_u^2 \left(\frac{1-\pi_u}{\pi_u}\right) + \sum_U \sum_{U-u} \frac{y_u}{\pi_u} \frac{y_{u'}}{\pi_{u'}} (\pi_{uu'} - \pi_u \pi_{u'})$$

Here $\pi_{uu'}$ is the pairwise inclusion probability, the probability that elements $u$ and $u'$ are both in the sample. To obtain a design based estimator of the finite population CDF, replace $I(y_u \leq t)$ for $y_u$ in the HT theorem to obtain:

$$\widehat{F}(t) = \frac{1}{N} \sum_S \frac{I(y_u \leq t)}{\pi_u} \qquad t \epsilon T.$$

T is the set of population values over which the CDF is to be evaluated. Design based estimation will not be the focus of this paper, but it must be emphasized that given a probability sample, rigorous design based estimation methodology for population parameters is available and appropriate.

## 1.5 Regression Parameter Estimation.

Suppose we are analyzing our data according to the missing covariate perspective, and our missing covariates are assumed or known not to be spatially patterned. Then we will use model (5) as our analytical tool for the sample data. Writing X as the matrix whose rows are the vectors of known covariates $x_u$ (plus a column of ones for the intercept term), D=Diag[$g(E(y_u);\gamma)\sigma^2$] as the diagonal matrix of variance functions for the vector of response values Y, the best linear unbiased estimate of the vector of regression parameters $\beta$ is $\widetilde{\beta} = (X'D^{-1}X)^{-1}X'D^{-1}Y$. It is important to note that the ordinary least squares (OLS) estimator $\widehat{\beta} = (X'X)^{-1}X'Y$ is also unbiased for $\beta$, though less efficient than the former estimator when the matrix D is known. When D is unknown, the variance function $g(E(y_u);\gamma)\sigma^2$ is usually estimated by using a transformation of the OLS residuals and minimizing some goodness of fit criterion to obtain an estimate of $\gamma$. In particular, regressing the logarithm of the squared residuals on the logarithm of $g(E(y_u);\gamma)\sigma^2$ is a commonly used method for the case where $g(E(y_u);\gamma) = E(y_u)^{\gamma}$ (Carroll and Ruppert 1988, Davidian and Carroll 1987):

$$\ln(\widehat{\epsilon}_u^2) = a + b \ln(\widehat{y}_u)+c,$$

$\widehat{\gamma}$ is then estimated using the slope of this line, $\widehat{b}$. The implementation of this method involves estimation of $\beta$, so an iterative method is required. This method will be briefly explored in the case study in chapter 4. Estimation of variance functions is a large literature unto itself (see also Welsh et al, 1994). Work done elsewhere has demonstrated the difficulty in estimating $\gamma$. In practice, it seems adequate to choose between $\gamma$=0, 1, 1.5, 2 in the model $g(E(y_u);\gamma) = E(y_u)^{\gamma}$ and choose the value which is optimal according to some fitting criterion. The gains in efficiency in using $\widetilde{\beta}$ will be lessened when it is necessary to estimate the variance function $g(E(y_u);\gamma)$. Therefore, for cases considered here the easier to compute $\widehat{\beta}$ will be used. However, the form of the variance function is an

important component of the model based CDF estimator to be considered here and the topic will be raised again in that context.

## 1.6 Regression Prediction

Given an unbiased estimate of $\beta$ and the matrix $X_0$ whose rows are the covariate values for the unsampled population units, the vector of best linear unbiased predictions of the response under population model (5) with heteroscedastic variances is $\widetilde{Y}_0=X_0\widetilde{\beta}$. Note that $\widehat{Y}_0=X_0\widehat{\beta}$ is also a linear unbiased predictor of $Y_0$, though is less efficient than $\widetilde{Y}_0$ when D is known. Under population model (5) for large samples, there is little difference in the result between using the unweighted least squares prediction $\widehat{Y}_0$ and the weighted least squares prediction in any particular instance (Carrol and Ruppert 1988, p. 52). When D must be estimated, the advantages of using weighted least squares estimates of $\beta$ for making predictions are lessened and so weighting has even less value, especially for the large sample sizes to be considered here. Therefore, for obtaining predictions, the unweighted least squares predictions will be adequate as well as simpler to compute. It is important to point out that the predictions are biased with respect to the response surface model (1):

$$
\begin{aligned}
E(\widetilde{Y}_0) &= E(X_0(X'D^{-1}X)^{-1}X'D^{-1}Y) \\
&= X_0(X'D^{-1}X)^{-1}X'D^{-1}E(Y) \\
&= X_0(X'D^{-1}X)^{-1}X'D^{-1}(X\beta + Z\alpha) \\
&= X_0(X'D^{-1}X)^{-1}X'D^{-1}X\beta + X_0(X'D^{-1}X)^{-1}X'D^{-1}Z\alpha \\
&= X_0\beta + X_0(X'D^{-1}X)^{-1}X'D^{-1}Z\alpha \\
&= E(Y_0) + X_0(X'D^{-1}X)^{-1}X'D^{-1}Z\alpha.
\end{aligned}
$$

Adding the spatial coordinates to the regression equation and spatial covariance structure to the residuals will be used to account for this bias by modeling the pattern of Z with the spatial information known for the entire population.

## 1.7 Summary.

Given a probability sample, design based estimation of the parameters of the finite population are available using the Horvitz Thompson theorem and other design based methodology. In addition, as covariate and spatial information is made available for the finite population under study, statistical models can be used. Some of these procedures will be investigated in the subsequent chapters here. The primary tools for model based estimation of the population parameters will be obtaining predicted values of the response variable for the nonsample units, and a model based methodology to estimate the finite population distribution function. In terms of obtaining predicted values for the nonsample units, we are primarily interested in obtaining predictions that are unbiased with respect to the chosen model (used to analyze the data) and in simplicity of computation. In the missing covariate perspective, these predictions will be biased (for the population values) because of the model misspecification but adding spatial information to the model may reduce the bias in cases where the missing covariates are spatially patterned. In addition, using the known covariate and spatial information for the entire population may provide gains in efficiency in estimation of the parameters of the finite population relative to the design based estimates.

# 2. Estimation of Residual Spatial Covariance for Stratified Sample Data.

## 2.1 Regression Analysis of Stratified Sample Data.

Under simple random sampling in each stratum with adequate sample sizes, it is sufficient to fit the regression model separately for each stratum when it is assumed that the regression parameters $\beta$ differ between strata. It is intuitive to fit strata separately, since differences in the population Y and the covariate population X are the usual motivation for stratification. Design effects of stratification have been examined by several authors and cases do exist where this is not the analysis of interest (DuMouchel and Duncan 1983, Jewell 1985, Quesenberry and Jewell 1986), but these issues are restricted to the cases where estimation of a single $\beta$ for the entire spatial population is desired, a restriction not necessary here for estimation of the regression parameters. For spatial covariance estimation of a single covariance model for all strata, the influence of stratification on the estimates does need to be considered, and this issue will be addressed here.

## 2.2 Model Selection for Residual Covariance.

As discussed earlier, the perspective in this document is that these spatial populations of concern vary in response to environmental causal variables. When some of these covariates are unknown but spatially patterned, we can attempt to represent these patterns of the spatial population through the spatial components of the linear model (2). The covariance is assumed to depend upon the population units only through a vector of parameters $\theta$ and the geographic distances between the population units $d_{uu'}=\sqrt{(i_u-i_{u'})^2+(j_u-j_{u'})^2}$, where i,j are the spatial coordinates of the units. The entire set of explanatory variables will seldom, if

ever be known in practice, so the spatial model (2) , which includes spatial coordinates in the regression equation and residual covariance can be used as a surrogate for spatially patterned missing covariates. The spatial covariance is modelled by a function that describes the covariance as a function of the distances between the population units. Many covariance models are available, three of the most commonly used are:

$$\sigma(d;\theta) = \theta_0 + \theta_1 \qquad d=0$$
$$\sigma(d;\theta) = \theta_1 * \exp(-\theta_2 d) \qquad d>0$$

$$\sigma(d;\theta) = \theta_0 + \theta_1 \qquad d=0$$
$$\sigma(d;\theta) = \theta_1 * \exp(-\theta_2 d^2) \qquad d>0$$

$$\sigma(d;\theta) = \theta_0 + \theta_1 \qquad d=0$$
$$\sigma(d;\theta) = \theta_1 \left(1-\frac{3}{2}\left(\frac{d}{\theta_2}\right) + \left(\frac{d^3}{2*\theta_2^3}\right)\right. \qquad d>0$$
$$\sigma(d;\theta) = 0 \qquad d>\theta_3$$

These are often named the exponential model, the gaussian model and the spherical model respectively. The values of the parameters $\theta_0$ and $\theta_1$ determine the behavior of the covariance function near the origin and the parameter $\theta_2$ influences the distance at which the population units are correlated. In some situations the form of the covariance model can be chosen according to the nature of the phenomenon under study (see Matérn, 1986). For this work the exponential model was selected because, 1) there is no theoretical form of the covariance function for the populations under study, 2) the exponential model does not have the multimodality problems in likelihood estimation encountered when using the spherical covariance model (Mardia and Watkins, 1989, Warnes and Ripley, 1987). A large literature exists regarding choosing and fitting the best covariance model and the consequences of model misspecification. The main result of interest here is that of Stein (1988) which states that predictions using a misspecified covariance model are asymptotically efficient (meaning as the sample size approaches the population size, or infill asymptotics in the geostatistical terminology) relative to predictions using the actual covariance function if the two covariance models are 'compatible'. Compatible essentially means that the two covariance functions have the same shape for small d.

## 2.3 Estimation of the Model Parameters.

For estimation of the regression parameters in the spatial linear model with a general covariance matrix, the best linear unbiased estimator (BLUE) of $\beta$ is $\widetilde{\beta}$ $= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$ when $\Sigma$ is known, here $\Sigma=\Sigma(d_{uu'};\ \theta)$. When $\Sigma$ is unknown, in practice the BLUE is modified by using $\widehat{\Sigma}=\Sigma(d_{uu'};\widehat{\theta})$ in place of $\Sigma$ and fitting involves iterating between estimates of $\beta$ and $\Sigma$ until some stopping rule is achieved. When using an estimated $\Sigma$, the gains in efficiency may be lost depending upon how well $\Sigma$ is estimated, and therefore the OLS estimate $\widehat{\beta}$ will be considered adequate.

In practice, the vector of spatial covariance parameters $\theta$ are unknown and must be estimated from the data. There are three main issues in estimating the parameters of the covariance function: 1) obtaining an estimate of the empirical covariance, 2) choosing a covariance model (discussed in the previous section), and 3) choosing an algorithm for estimating the parameters $\theta$. Many estimators of $\theta$ are available which consist of taking the residuals from an OLS fit for the regression parameters, "smoothing" the observed covariances and fitting a parametric covariance model. One particular estimator for the empirical covariance function is as follows: for a pair of residuals from the sample data, their observed covariance is $\widehat{\epsilon}_u * \widehat{\epsilon}_{u'}$ and separation distance $d_{uu'} = \sqrt{(i_u-i_{u'})^2 + (j_u-j_{u'})^2}$, the method of moments estimator of the residual covariance is:

$$\widehat{C}(d) = \frac{1}{N_d} \sum_{\{u,u':d_{uu'}=d\}} \widehat{\epsilon}_u\widehat{\epsilon}_{u'} \ , \qquad (6)$$

where the sum is over all pairs $N_d$ that are d units apart. In practice, the $d_{uu'}$ are all unique and are aggregated into groups of approximately the same d. Of concern is the potential for bias when using this estimator with variable probability samples. We will return to this point in section 2.7. Using $\widehat{C}(d)$ (which depends only on d) a parametric covariance function can be fit using a variety of methods, such as M-estimators or likelihood based estimators. Alternately, a model is fit to the semi-variogram: $\frac{1}{2}\gamma(d)= C(0) - C(d)$.

Zimmerman and Zimmerman (1991) compared two types of M-estimators and normal theory maximum likelihood (ML) and restricted (or residual) maximum likelihood (REML) and found that for normally distributed data and an exponential semi-variogram model, ML and REML performed slightly better in terms of the variances of the estimated $\theta$ values over repeated samples.

## 2.4 Spatial Prediction (Kriging).

Under model (2) the predictions for the nonsample will be $\widetilde{Y}_0 = X_0\widetilde{\beta} + \Sigma_0\Sigma^{-1}(Y-X\widetilde{\beta})$, here $X_0$ is the matrix of known covariates and spatial coordinates for the nonsample (with a column of ones for the intercept term), $\Sigma_0'=\Sigma_0(d_{uu'};\theta)$ is the transpose of the spatial covariance matrix between the sample units and nonsample population units, $\Sigma$ is the (spatial) covariance matrix on the sample. $\Sigma$ is a matrix whose elements are $\sigma(d_{uu'};\theta)$ and $\sigma$ is a covariance function that insures that $\Sigma$ is positive definite. Prediction with estimated covariance parameters is accomplished by plugging in the estimate $\Sigma(d_{uu'};\widehat{\theta})$ for $\Sigma$ in the prediction equation. When $\theta$ is estimated, the predictions are unbiased with respect to the model if the distribution of $\epsilon$ is symmetric and the estimator of $\theta$ is even and translation invariant (Zimmerman and Harville, 1989, Christensen 1991, Zimmerman and Cressie, 1992), which is the case with most commonly used estimators. The precision of the predictions does depend upon the estimation of $\theta$. If these parameters are estimated poorly, the precision of predictions based on them may be worse than if just OLS predictions were used (Cordy and Griffith, 1993).

## 2.5 Residual Spatial Covariance for Stratified Sample Data.

Before proceeding with spatial covariance modelling and spatial prediction using data from a stratified sample, two issues need to be considered; the effect of stratum differences on the analysis, and the potential for bias. Suppose we are using the missing covariate model for a stratified population, and each stratum is assumed to follow the model:

$$y_{uh} = \mu_h + x_{uh}\beta_h + z_{uh}\alpha_h + \epsilon_{uh} \ ,$$

$$E(\epsilon_{uh}) = 0,$$
$$Var(\epsilon_{uh}) = \sigma_h^2,$$
$$Cov(\epsilon_{uh},\epsilon_{u'h'}) = 0, \ u \neq u'.$$

and the (unknown) $z_{uh}$'s can be approximated with the following second order properties:

$$E(z_{uh}) = \delta$$
$$Var(z_{uh}) = \tau^2$$
$$Cov(z_{uh},z_{u'h'}) = \tau(d_{uu'}; \theta) \quad u \neq u'.$$

Note that the properties of the $z_{uh}$'s are independent of the strata, but that the $z_{uh}\alpha_h$ are dependent upon the strata. Using the same notation as in section 1.3, we are analyzing the residuals $\epsilon^*_{uh}$ for spatial covariance.

$$\epsilon^*_{uh} = z_{uh}\alpha_h + \epsilon_{uh} - \delta\alpha_h,$$

$$E(\epsilon^*_{uh}) = 0,$$

$$Var(\epsilon^*_{uh}) = \alpha_h^2 \tau^2 + \sigma_h^2,$$

$$Cov(\epsilon^*_{uh} , \epsilon^*_{u'h'}) = Cov((z_u\alpha_h + \epsilon_{uh})(z_{u'}\alpha_{h'}+\epsilon_{u'h'})),$$

$$= Cov(z_u\alpha_h, z_{u'}\alpha_{h'}) + Cov(z_u\alpha_h, \epsilon_{u'h'})$$

$$+ Cov(\epsilon_{uh}, z_{u'}\alpha_{h'}) + Cov(\epsilon_{uh}, \epsilon_{u'h'}),$$

$$= \alpha_h \alpha_{h'} \mathrm{Cov}[z_u, z_{u'}] + 0 + 0 + 0,$$

$$= \alpha_h \, \alpha_{h'} \tau(d_{uu'}; \theta).$$

If the residuals are in the same stratum, then this reduces to:

$$= \alpha_h^2 \tau(d_{uu'}; \theta).$$

When using the method of moments estimator of the residual covariance and fitting a parametric model to this estimator, the standard operating assumptions are that the residuals have a constant mean (this can be relaxed to be a function of the spatial coordinates), constant variance and a spatial covariance whose value depends on the data only through the geographic distance between the population units. This is not to be expected for stratified samples, simply because anticipated differences in mean and variance for subsets of the population are one of the reasons for stratification. Stratum differences in mean are already accounted for by fitting the regression models separately for each stratum. In the missing covariate perspective, we are also assuming that the distribution of the spatially patterned missing covariate is independent of stratum. The stratum differences in variance can be accounted for before the spatial analysis by scaling the stratum specific residuals by their respective standard deviation functions: $\epsilon^*_{uh} / \sqrt{\alpha_h^2 \tau^2 + \sigma_h^2}$

Then $\mathrm{Var}(\epsilon^*_{uh} / \sqrt{\alpha_h^2 \tau^2 + \sigma_h^2}) = 1$, and:

$$\mathrm{Cov}\Big(\epsilon^*_{uh} / \sqrt{\alpha_h{}^2 \tau^2 + \sigma_h^2}, \; \epsilon^*_{u'h'} / \sqrt{\alpha_{h'}^2 \tau^2 + \sigma_{h'}^2}\Big)$$

$$= (1/\sqrt{\alpha_h^2 \tau^2 + \sigma_h^2}) \; (1/\sqrt{\alpha_{h'}^2 \tau^2 + \sigma_{h'}^2}) \; \mathrm{Cov}(\epsilon^*_u, \epsilon^*_{u'}),$$

$$= (\alpha_h / \sqrt{\alpha_h^2 \tau^2 + \sigma_h^2}) \; (\alpha_{h'} / \sqrt{\alpha_{h'}^2 \tau^2 + \sigma_{h'}^2}) \; \tau(d_{uu'}; \theta).$$

If the residuals are from the same stratum, this reduces to:

$$= (\alpha_h^2/\alpha_h^2\tau^2 + \sigma_h^2)\ \tau(\mathrm{d}_{uu'};\ \theta).$$

Note that if $\alpha^2\tau^2 >> \sigma^2$ then the covariance is approximately equal to:

$$\simeq \frac{\alpha_h\alpha_{h'}\tau(\mathrm{d}_{uu'};\ \theta)}{\sqrt{\alpha_h^2\tau^2}\sqrt{\alpha_{h'}^2\tau^2}},$$

$$= \frac{\alpha_h\alpha_{h'}\tau(\mathrm{d}_{uu'};\ \theta)}{\alpha_h\tau\ \alpha_{h'}\tau},$$

$$= \rho(\mathrm{d}_{uu'};\ \theta),$$

which is the spatial correlation function of the missing covariate.

If $\alpha^2\tau^2 << \sigma^2$

$$\simeq \frac{\alpha_h\alpha_{h'}\tau(\theta;h_{uk})}{\sigma_{h'}\sigma_h}$$

$$= \frac{\tau^2\alpha_h\alpha_{h'}\tau(\mathrm{d}_{uu'};\ \theta)}{\tau^2\sigma_{h'}\sigma_h}$$

$$= \rho(\mathrm{d}_{uu'};\ \theta)*\frac{\tau^2\alpha_h\alpha_{h'}}{\sigma_{h'}\sigma_h} << \rho(\mathrm{d}_{uu'};\ \theta).$$

For residuals from the same stratum, this reduces to:

$$= \rho(\mathrm{d}_{uu'};\ \theta)*\frac{\tau^2\alpha_h^2}{\sigma_h^2}.$$

In all cases, the spatial correlation of the scaled residuals is the spatial correlation of the missing covariate multiplied by a number that is less than one. Note that if $\alpha_h$=$\alpha_{h'}$= $\alpha$ (i.e. the relationship of y with the missing covariate is independent of stratum) then these results will be the same for pairs within the same stratum or from different strata. When the unexplained variability in the residuals is dominated by the missing covariate ($\alpha^2\tau^2 >> \sigma^2$) the covariance function of the scaled residuals will approximate the correlation function that describes the spatial pattern of the missing covariate. In the opposite case ($\alpha^2\tau^2 << \sigma^2$), the

estimated spatial correlation will be less than the actual spatial correlation, possibly by a considerable amount. In this case spatial covariance estimation will be of less interest because the missing covariate constitutes a minor part of the unexplained population variability, and population model (5) is adequate.

## 2.6 Residual Scaling.

The results in the previous section were derived using the actual residuals $\epsilon_u^*$ and variance functions $\alpha_h^2\tau^2 + \sigma_h^2$ for each of the strata. In an actual data analysis setting, we will only have estimated residuals $\widehat{\epsilon}_u^*$ and a hypothesized variance function, which we will write in general as $\mathrm{Var}(\epsilon_u^*) = \mathrm{g}(\mathrm{E}(\mathrm{y}_u);\gamma)$. The primary case where residual scaling is of interest is in stratified sampling where the variance functions differ between strata and it is desired to estimate residual spatial covariance for the entire population of residuals. When analyzing nonstratified populations or stratified populations where each stratum is analyzed separately, scaling may not be needed. Three types of residuals are identified for spatial covariance estimation:

1) $\widehat{\epsilon}_u^*$ unscaled,

2) $\frac{\widehat{\epsilon}_u^*}{\widehat{\sigma}}$ homogeneously scaled,

3) $\frac{\widehat{\epsilon}_u^*}{\sqrt{\widehat{\mathrm{g}}(\mathrm{E}(\mathrm{y}_u;\gamma)}\ \widehat{\sigma}}$ heterogeneously scaled.

Case (1) analyzes unscaled residuals from an ordinary least squares regression model. Case (2) scales residuals by their estimated standard deviation, estimated separately for each stratum for stratified populations. In case (3), the scaling takes place in two steps, first the residuals are scaled by the square root of the variance function;

$$\widetilde{\epsilon} = \frac{\widehat{\epsilon}_u^*}{\sqrt{\widehat{g}(E(y_u;\gamma))}}.$$

If a combined spatial covariance analysis is desired for a stratified population, then these residuals are additionally scaled by their stratum standard deviation;

$$\widetilde{\epsilon}' = \frac{\widetilde{\epsilon}}{\sqrt{\mathrm{Var}(\widetilde{\epsilon})}}.$$

We wish to see if residual scaling followed by covariance estimation gives evidence of spatial correlation in the residuals that would otherwise have been undetected had scaling not been used. In practice, estimated residuals and estimated standard deviation functions must be used for this analysis, and so the utility of the concept of scaling may be diminished. The influence on residual spatial covariance estimation of scaling will be investigated in the simulation work to come.

## 2.7 The Design Effect.

The design effect is defined as the bias of any parameter estimate when analyzing the sample as an independent random sample when it is actually taken by a structured design. Accounting for the design effect in stratified samples is done in two ways; incorporation of the inclusion probabilities into the model based analysis (Little 1991, Skinner and Holt, 1989), or adding additional parameters to the model to account for the design effect (Scott and Holt, 1982, Christensen 1987). Failing to account for the design effect can lead to inference about the wrong population (Overton, 1991). The design effect of stratified samples on spatial covariance estimation is only of concern when the data analysis uses combined strata. Because the shape of the covariance function at the shortest distances has the greatest affect on predictions (Stein, 1988), minimization of the design effect at these short distances will be the most important. If the strata are highly fragmented or patchy, there will be many pairs of residuals separated by short distances that cross stratum boundaries. This is the case when the design

effect would be expected to be greatest. Under stratified random sampling with proportional allocation, there is essentially no bias because the pairwise inclusion probabilities will be approximately the same as if the entire sample were just SRS.

The magnitude of this effect will depend upon the difference in the first order inclusion probabilities among strata. In cases where allocation does not deviate greatly from proportional allocation and the strata are large and convex in shape, this design effect on the method of moments estimator $\widehat{C}(d)$ for small d is expected to be small. If we select a simple random sample of size $n_h$ out of $N_h$ from stratum $h$ and a simple random sample of size $n_{h'}$ out of $N_{h'}$ from stratum $h'$ their inclusion probabilities respectively are $n_h/N_h$ and $n_{h'}/N_{h'}$. For pairs of points in the same stratum, their pairwise inclusion probability is $(n_h(n_h-1))/(N_h(N_h-1))$ . The pairwise inclusion probability of a pair of points from different strata will be $n_h n_{h'}/N_h N_{h'}$. Under proportional allocation, $n_h/N_h = n_{h'}/N_{h'}$ and the inclusion probabilities for all pairs will be approximately the same. When the allocation is not proportional, they are not. A modified estimation procedure analogous to the method of moments estimator is:
The pairwise inclusion probability for units in the same stratum, say stratum H, will be:

$$\pi_{uu'} = \frac{n_h(n_h-1)}{N_h(N_h-1)}$$

and for pairs in separate strata, say h and h', these will be:

$$\pi_{uu'} = \frac{n_h n_{h'}}{N_h N_{h'}}.$$

So, setting $w_{uu'} = 1/\pi_{uu'}$, an estimator of the residual covariance function will be:

$$\frac{1}{\sum w_{uu'}} \sum w_{uu'} \hat{\epsilon}_u \hat{\epsilon}_{u'} \qquad (7)$$

where both summations are over the set: $\{u,u'$: $d_{uu'} = d\}$. In practice, the weights will differ greatly in only the most extreme cases of nonproportional allocation. Additionally, since this empirical function is then smoothed with a fitted covariance model, the effect is further reduced. However, since the pairwise inclusion probabilities are known for probability samples, this design effect should

be investigated. A recommended procedure would be to calculate the method of moments estimator (6) and the weighted estimator (7) and fit the covariance model to each. If it is judged that the fitted models differ considerably, then the weighted estimator should be used. If not then the unweighted estimator will be adequate.

## 2.8 Simulation Study.

Several populations were generated to assess the influence of heteroscedasticity and residual scaling on spatial covariance function estimation. Comparative results were obtained for analysis of simple random samples and stratified random samples with proportional and nonproportional allocation. For both designs, populations with homogeneous error variance and heteroscedastic errors were analyzed. The populations were simulated using the fitted regression equations from the original Eastern Lake Survey data analyzed in chapter 4. Three nonstratified populations were created using the fitted regression equation for the first stratum of the original eastern lake survey, each population with increasing levels of residual variance and heteroscedasticity. Five stratified populations were created, four with stratum specific regression coefficients for the missing covariate, one where the regression coefficient for the missing covariate is the same for all strata. The simulated populations are as follows:

Nonstratified Populations:

Population N1 (Size=600):

$$y_u = -4900 - 0.49 * x_{1u} + 1270 * x_{2u} + e_u, \quad e_u \sim N(0,1)$$

Population N2 (Size=600):

$$y_u = -4900 - 0.49 * x_{1u} + 1270 * x_{2u} + e_u, \quad e_u \sim N(0, 25 * E(y_u))$$

Population N3 (Size=600):

$$y_u = -4900 - 0.49 * x_{1u} + 1270 * x_{2u} + e_u, \quad e_u \sim N(0, 100 * E(y_u))$$

Stratified Populations:

Population S1 (Size=600, 200 in each stratum): Separate models for each stratum:

$$y_{1u} = -4900 - 0.49*x_{1u} + 1270*x_{2u} + e_{1u}, \quad e_{1u} \sim N(0,1)$$
$$y_{2u} = -7535 + 0.564*x_{1u} + 1783*x_{2u} + e_{2u} \quad e_{2u} \sim N(0,1)$$
$$y_{3u} = 17730 - 1.57*x_{1u} - 3833*x_{2u} + e_{3u}, \quad e_{3u} \sim N(0,1)$$

Population S2 (Size=600, 200 in each stratum): Separate models for each stratum:

$$y_{1u} = -4900 - 0.49*x_{1u} + 1270*x_{2u} + e_{1u}, \quad e_{1u} \sim N(0,E(y_{1u}))$$
$$y_{2u} = -7535 + 0.564*x_{1u} + 1783*x_{2u} + e_{2u}, \quad e_{2u} \sim N(0,E(y_{2u}))$$
$$y_{3u} = 17730 - 1.57*x_{1u} - 3833*x_{2u} + e_{3u}, \quad e_{3u} \sim N(0,E(y_{3u}))$$

Population S3 (Size=600, 200 in each stratum): Separate models for each stratum:

$$y_{1u} = 400 - 0.49*x_{1u} + 50*x_{2u} + e_{1u}, \quad e_{1u} \sim N(0,E(y_{1u}))$$
$$y_{2u} = 200 + 0.564*x_{1u} + 50*x_2 + \epsilon_{2u} \quad e_{2u} \sim N(0,E(y_{2u}))$$
$$y_{3u} = 2000 - 1.57*x_{1u} + 50*x_2 + e_{3u}, \quad e_{3u} \sim N(0,E(y_{3u}))$$

Population S4 (Size=600, 200 in each stratum): Separate models for each stratum.

$$y_{1u} = -4900 - 0.49*x_{1u} + 1270*x_{2u} + e_{1u}, \quad e_{1u} \sim N(0,4*E(y_{1u}))$$
$$y_{2u} = -7535 + 0.564*x_{1u} + 1783*x_{2u} + e_{2u}, \quad e_{2u} \sim N(0,4*E(y_{2u}))$$
$$y_{3u} = 17730 - 1.57*x_{1u} - 3833*x_{2u} + e_{3u}, \quad e_{3u} \sim N(0,4*E(y_{3u}))$$

Population S5 (Size=600, 200 in each stratum): Separate models for each stratum.

$$y_{1u} = -4900 - 0.49*x_{1u} + 1270*x_{2u} + e_{1u}, \quad e_{1u} \sim Unif(0.5,1.2)$$
$$y_{2u} = -7535 + 0.564*x_{1u} + 1783*x_{2u} + e_{2u} \quad e_{2u} \sim Unif(0.5,1.2)$$
$$y_{3u} = 17730 - 1.57*x_{1u} - 3833*x_{2u} + e_{3u}, \quad e_{3u} \sim Unif(0.5,1.2)$$

For all populations, $x_{1u}$ is the elevation of the unit, $x_{2u}$ and is the pH of the unit and $E(y_u)>0$. The spatially patterned pH variable was treated as unknown in order to induce spatial pattern in the residuals and the elevations and spatial coordinates were assumed known for the entire population. The nonstratified populations were sampled with size n=75, and residual covariance is estimated using no scaling, homogeneous scaling of residuals before spatial covariance estimation, and heterogeneous scaling of residuals before spatial covariance

estimation. For the stratified populations, all were sampled with proportional allocation and n=25 in each stratum, populations . To illustrate the influence of variable probabilities of selection on estimated covariance parameters, a subset of the stratified populations were sampled via stratified sampling with simple random sampling in each stratum but without proportional allocation. Only homogeneous scaling was used for these samples. For populations S1, S2, the stratum sample sizes are $n_1$=10, $n_2$=25, $n_3$=40. For populations S4 and S5, the sample sizes are: $n_1$=15, $n_2$=15, $n_3$=40.

The details of the parameter estimation protocol are as follows: Given the simple random sample, the regression model of $y_u$ on $x_{1u}$ is fit by OLS. In the case of the stratified populations, this is fit separately for each stratum. The residuals $\widehat{\epsilon}_u$ from this model are scaled by their estimated standard deviation, or estimated heteroscedastic standard deviation function. Again, for the stratified populations the residuals from each stratum are scaled separately:

$$\widetilde{\epsilon}_{uh} = \widehat{\epsilon}_{uh}/\widehat{\sigma}_h$$

or

$$\widetilde{\epsilon}_{uh} = \widehat{\epsilon}_{uh}/\sqrt{\widehat{y}_{uh}}\ \widehat{\sigma}_h$$

Heterogeneous scaling and homogeneous scaling will be used for these simulated populations to determine the influence of the form of scaling on covariance estimation. In addition to the results using scaled residuals, residual covariance for the nonstratified populations is estimated using unscaled residuals. Using these residuals, the parameters of the following model are estimated using the exponential covariance function:

$$\begin{aligned}\widetilde{\epsilon}_{uh} &= \eta + i_u\phi_1 + j_u\phi_2 + \varepsilon_{uh},\\ &\quad E(\varepsilon_u) = 0,\\ &\quad \mathrm{Var}(\varepsilon_u) = \sigma^2,\\ &\quad \mathrm{Cov}(\varepsilon_u\ ,\ \varepsilon_{u'}) = \sigma(d_{uu'};\ \theta),\ \ u \neq u'.\end{aligned}$$

## 2.9 Results.

The results for the estimated covariance parameters are summarized in several ways: the mean and standard deviation of the 100 estimates are calculated, the mean and median of a statistic estimating the distance at which the correlation between two population units is 0.05:

$$d_5 = \ln[\ \frac{1}{0.05} * (\frac{\theta_1}{\theta_0 + \theta_1})\ ]\ /\theta_2,$$

this is often called an estimate of the correlation range. The mean and median of a statistic that estimates the correlation of residuals that are close together ($d \approx 0$) is also calculated:

$$d_o = \frac{\theta_1}{\theta_0 + \theta_1}\ .$$

The results for the mean and median of $d_5$ and $d_o$ are shown in table 2.1, means and standard deviations for the estimates of $\theta$ are given in table 2.2 for each population and for the different forms of scaling. In addition, a column labeled 'reps' indicates the number of samples for which residual spatial covariance was estimated to be 0. All 100 estimates whether they were 0 or not are used in the calculation of the summary statistics. For the stratified populations, there are no results using no scaling because the REML algorithm did not converge to a single set of estimates for the $\theta$'s. For these populations, it was necessary to use some form of scaling in order to obtain estimates of residual spatial covariance. Examination of $d_o$ in table 2.1 reveals that scaling the residuals before covariance estimation decreased the correlation of the residuals for small d. In addition, mean and median values of $d_o$ using homogeneous scaling were smaller than when when using heterogeneous scaling except for three cases. This indicates that using homogeneously scaled residuals treats less of the unexplained residual variation as spatial correlation and treats more as random variation. Using heterogeneously scaled residuals attributes more of the unexplained residual variation as spatial correlation and less as random variation.

Cressie, (1991 p. 127-134) discusses the influence of this behavior at the origin in terms of the variogram. He mentions that the magnitude of the

difference is a measure of how much residual variability is random variation and how much is due to spatial correlation. The greater the discontinuity, the greater the amount of variation that is considered random. The simulated results here indicate that using homogeneous scaling decreases the estimated residual spatial correlation for small d. The three exceptions to this result are for populations N3, S3, and S4. These are the cases where the random variation in the residual is much larger than the variation due to the missing covariate, the estimated correlation for small d is less than 0.5, and the difference in $d_o$ estimates between the different forms of scaling is small.

The influence of scaling on the estimate of $d_5$ is less clear. For the nonstratified populations, N1 and N2 have approximately the same median value of $d_5$ when no scaling is used (98.4 Km and 110.9Km respectively), but $d_5$ is increased when scaling is used for N1 and decreased when scaling is used for N2. Residual covariance was estimated to be 0 when no scaling was used for population N3, but $d_5$ was about the same for both types of scaling (23 km and 30 km). This is likely due to the residual variance being much greater than the variance due to the missing covariate. There was not a consistent affect of scaling on the $d_5$ values for the stratified populations. For population S2, heterogeneous scaling resulted in a larger median $d_5$ than for homogeneous scaling( 47.98 km and 6.3 km respectively). The opposite was true for populations S3, S4, S5 though the median $d_5$ values were not drastically different.

Table 2.1 Means and Medians of $d_5$ and $d_O$.

| | $d_5$ | | $d_o$ | | |
|---|---|---|---|---|---|
| Pop/scale | mean | median | mean | median | reps |
| N1/none | 169.9 | 98.4 | 0.96 | 0.999 | 7 |
| N1/hom | 3716.4 | 254.6 | 0.491 | 0.409 | 98 |
| N1/het | 2906.6 | 544.3 | 0.804 | .811 | 55 |
| N2/none | 3898.1 | 110.9 | 0.521 | 0.511 | 39 |
| N2/hom | 4071.5 | 34.12 | 0.412 | 0.345 | 98 |
| N2/het | 2301.2 | 47.8 | 0.472 | 0.380 | 97 |
| N3/hom | 5298.6 | 22.99 | 0.472 | 0.312 | 33 |
| N3/het | 6727.3 | 30.7 | 0.381 | 0.232 | 38 |
| S1/hom | 37.63 | 11.17 | 0.73 | 0.400 | 78 |
| S2/hom | 13.85 | 6.31 | 0.549 | 0.475 | 60 |
| S2/het | 138.1 | 47.7 | 0.814 | 0.84 | 90 |
| S3/hom | 14.12 | 8.49 | 0.372 | 0.032 | 22 |
| S3/het | 27.6 | 6.59 | 0.334 | 0.00 | 29 |
| S4/hom | 5422.6 | 16.32 | 0.324 | 0.273 | 21 |
| S4/het | 16.96 | 11.08 | 0.308 | 0.102 | 29 |
| S5/hom | 361.3 | 158.5 | 0.926 | 0.979 | 85 |
| S5/het | 334.7 | 143.3 | 0.931 | 0.998 | 93 |

In terms of the covariance parameter estimates, the form of residual scaling prior to fitting influences the estimate of $\theta_2$. The means and standard deviations of the $\theta_0$, $\theta_1$ are changed little between the two forms of scaling, but the mean of the estimated $\theta_2$ increases when using heterogeneous scaling relative to homogeneous scaling (see table 2.1 and compare the solid and dashed pairs of lines for N1 and N2 in figure 2.1(a) and for S3, S4 in figure 2.1(b)). This result doesn't hold for population N3 which represents a case where $\sigma^2 >> \alpha^2\tau^2$ and the form of scaling has little affect on the estimated covariance function (see table 2.1 and figure 2.1) and for population S2, where the form of scaling has little effect on the parameter estimates.

Table 2.2 Statistics for Estimated Covariance Parameters.
Values are: mean (sd) for the 100 samples.

| Pop/scale | $\theta_0$ | $\theta_1$ | $\theta_2$ |
|---|---|---|---|
| N1/none | 5976(2888) | 165.8(504.7) | 128.4(754.7) |
| N1/hom | .6622(.4302) | .2282(.7518) | .0145(.0645) |
| N1/het | .3265(.2944) | 2.902(5.3592) | .8546(1.5192) |
| N2/none | 11346(4902) | 43199(12559) | 9.1633(66.52) |
| N2/hom | .8400(.2670) | 3.820(11.145) | .0276(.0644) |
| N2/het | .8725(.2295) | 5.322(14.296) | .0404(.0727) |
| N3/hom | .8382(.2647) | 23.94(141.462) | .0207(.0639) |
| N3/het | .8690(.2379) | 22.46(141.245) | .0277(.0672) |
| S1/hom | .1447(.0185) | 2.6508(.3977) | 3.783(1.3184) |
| S2/hom | .2846(.02461) | 2.6465(.9815) | 4.0377(1.4238) |
| S2/het | .2702(.2462) | 1.972(3.598) | 3.796(14.859) |
| S3/hom | .8823(.2581) | .1187(.2619) | .0424(.1306) |
| S3/het | .8760(.2762) | .1217(.2738) | .0645(.1910) |
| S4/hom | .9231(.1785) | 2.643(11.05) | .0242(.0705) |
| S4/het | .8961(.2511) | 1.821(9.281) | .0441(.0948) |
| S5/hom | .1350(.1668) | 2.4703(3.93) | 3.299(12.20) |
| S5/het | .1248(.1634) | 2.433(3.76) | 3.799(12.45) |

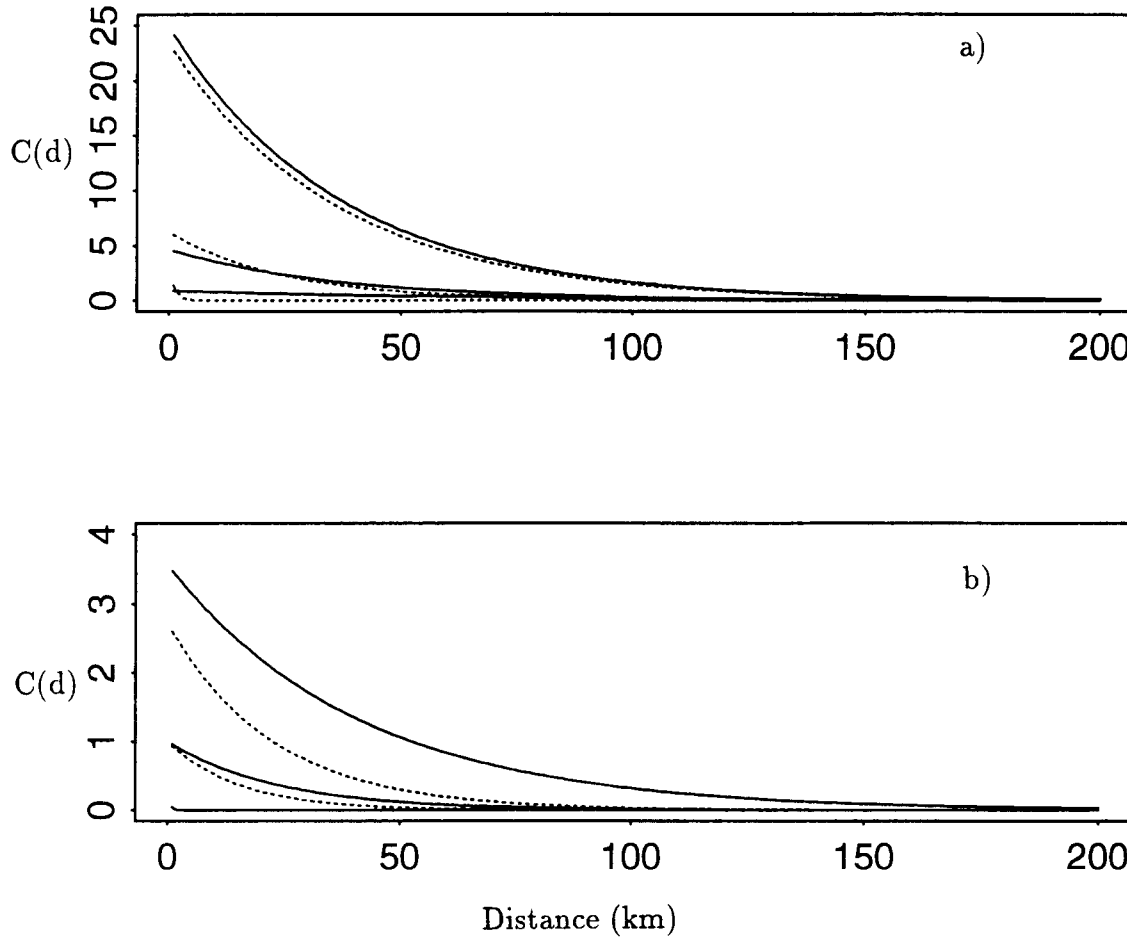


Figure 2.1 Estimated Residual Covariance Functions. a) Nonstratified Populations. Functions are plotted using the mean estimates of $\theta$, solid lines are from heterogeneous scaling, dashed are from homogeneous scaling. Top pair is for population N3, middle pair for N2, bottom pair for N1. b) Stratified Populations. Top pair is population S4, middle pair is population S3, the bottom curve is the superposition of populations S1 and S2.

Results for the stratified samples without proportional allocation are given in table 2.3. Results for the mean of the parameter estimates are also included from table 2.2. The results for proportional allocation are included for comparison (these are the same as the results in table 2.1). The populations were analyzed with homogeneous scaling. For populations S1, S2, the stratum sample sizes are $n_1$=10, $n_2$=25, $n_3$=40. For populations S4 and S5, the sample sizes are: $n_1$=15, $n_2$=15, $n_3$=40. Examination of the estimates indicates that the type of allocation primarily influences the estimates of $\theta_2$, decreasing the average estimated value for populations S1, S2, (equivalent to stronger estimated spatial dependence) and increasing it for populations S4 and S5 (equivalent to decreasing the estimated spatial dependence). These results serve to illustrate the potential for the design affect, the results using the weighted estimator (7) are not included and still need to be investigated to see if the differences are accounted for by weighting.

Table 2.3. Means of Estimated Covariance Parameters with Nonproportional Allocation.

| Pop | $\theta_0$ | $\theta_1$ | $\theta_2$ |
|---|---|---|---|
| S1/prop | .1447 | 2.6508 | 3.783 |
| S1/non | .1770 | .9722 | 1.833 |
| S2/prop | .2846 | 2.6465 | 4.0377 |
| S2/non | .2089 | 1.2601 | 1.6207 |
| S4/prop | .9231 | 2.643 | .0242 |
| S4/non | .5880 | 2.420 | .5035 |
| S5/prop | .1350 | 2.4703 | 3.299 |
| S5/non | .1713 | 2.053 | 26.165 |

## 2.10 Conclusions.

Residual scaling before spatial covariance estimation influenced the estimated residual correlation at small distances and the correlation range. Using homogeneously scaled residuals for covariance estimation yielded a weaker estimated correlation than using heterogeneously scaled residuals. The strength of the estimated spatial correlation is related to the amount of residual variation that is considered random. The form of scaling sets aside a portion of the variability that is considered to be purely random spatially. Heterogeneous scaling attributes part of the residual spatial pattern to the spatial pattern of the $E(y_u)$, with large residuals being associated with large values of $E(y_u)$ for the variance functions considered here. This puts more of the residual variation into the spatial correlation model than homogeneous scaling, which treats more of the residual variation as spatially unpatterned. Using no scaling for the nonstratified populations gave the strongest estimated correlation of the residual spatial pattern and gave approximately the same representation of the patterns for both populations N1 and N2 (i.e. similar median values for $d_0$ and $d_5$, recall that the missing covariate and its regression coefficient is the same for these two populations, so their estimated residual spatial patterns should be the same). This is evidence that using no scaling is the preferred option in these cases.

Work with the stratified populations considered here has shown that it was not possible to estimate residual spatial covariance without some form of scaling when the covariance is estimated using combined strata. Again, the estimated correlation was greater when using heterogeneous scaling. The estimated correlation range ($d_5$) of the residuals for the different forms of scaling is less clear. In some cases it is greater for homogeneous scaling and in some cases it is less. In terms of prediction we expect spatial predictions to differ the most from OLS predictions when the spatial correlation is strong and the correlation range is large. We expect little difference when the spatial correlation is weak and the correlation range is small. Populations S3, S4 are the latter case, the median estimated correlations were less than 0.032 and 0.273 respectively for both forms of scaling and the median estimated ranges were less than 9 Km and 16 Km. We would expect little difference between OLS prediction and spatial predictions in

these cases. Populations N1, N2, S1, S2, S5 have higher correlations and in some cases large ranges so we expect spatial predictions to differ the most from OLS predictions for these populations. Nonproportional allocation affected the estimation of the $\theta$'s, the increase or decrease in the average estimated value depending upon the nature of the nonproportional allocation and the population under study. For populations S1 and S2, the average estimated $\theta_2$ decreased under the nonproportional allocation, for populations S4 and S5, it increased. The investigation of residual scaling examined here assumed the form of the variance function is known and only the components $\sigma$ and $y_u$ needed to be estimated from the data. In some cases, this may be possible based upon understanding of the variance structure of the population under study. In other cases, even the form of the variance function to be used for scaling must be estimated from the data. Based upon this work it seems the primary issue in choosing the type of scaling (when it is needed) is to choose between a standard deviation or a heterogeneous variance function for scaling.

# 3. Model Based Estimation of Distribution Functions and other Parameters.

## 3.1 The Chambers and Dunstan Estimator

Discussion of model based CDF estimation begins with the design based estimator. For equal probability sampling it is:

$$\widehat{F}_d(t) = \frac{1}{n} \sum_{u \epsilon S} I(y_u \leq t) \qquad (8)$$

The first modification that can be made in the direction we are headed is to obtain model based predictions for the nonsample units and add an additional component to (8) based upon these:

$$\widehat{F}_m(t) = \frac{1}{N}[ \sum_{u \epsilon S} I(y_u \leq t) \quad + \sum_{u \epsilon U\text{-}S} I(\hat{y}_u \leq t) ] \qquad (9)$$

We will call this the naïve model based estimator because it makes no distributional assumptions about the residuals and does not account for it in the estimated CDF. It is necessary to put the regression residuals back into the CDF; one estimator that does this is modified from Chambers and Dunstan (1986):

$$\widetilde{F}(t) = \frac{1}{N} [ \sum_{u \epsilon S} I(y_u \leq t) + \sum_{u \epsilon U\text{-}S} \Phi^{-1}((t\text{-}\hat{y}_u)/g(\hat{y}_u;\hat{\gamma})\hat{\sigma}) ] \qquad (10)$$

The second summation is an estimate of the sum of $P(y_u \leq t)$ for the nonsample population units. By noting that $y_u = \hat{y}_u + \epsilon_u$ it is apparant that (9) does not account for the component of the F(t) due to the distribution of the residuals. Suppose the residuals have distribution $N(0,g(E(y_u);\gamma)$. Then

$$\begin{aligned} \sum_{U\text{-}S} P(y_u \leq t) \quad &= \sum_{u \epsilon U\text{-}S} P(\hat{y}_u + \epsilon_u \leq t), \\ &= \sum_{u \epsilon U\text{-}S} P(\epsilon_u \leq t\text{-}\hat{y}_u), \end{aligned}$$

$$= \sum_{u\epsilon \text{U-S}} P\left(\frac{\epsilon_u}{g(E(y_u);\gamma)\ \sigma} \leq \frac{t\text{-}\widehat{y}_u}{g(E(y_u);\gamma)\ \sigma}\right),$$

$$= \sum_{u\epsilon \text{U-S}} \Phi^{-1}\left(\frac{t\text{-}\widehat{y}_u}{g(E(y_u);\gamma)\ \sigma}\right).$$

This is estimated by using model based estimates of the unknown quantities:

$$\sum_{u\epsilon \text{U-S}} \widehat{P}(y_u \leq t) \quad = \sum_{u\epsilon \text{U-S}} \Phi^{-1}\left(\frac{t\text{-}\widehat{y}_u}{g(E(\widehat{y}_u);\widehat{\gamma})\ \widehat{\sigma}}\right).$$

Therefore the Chambers and Dunstan estimator can be represented by the naïve with a bias correction term over the nonsample $B(t) = \sum \widehat{P}(\widehat{y}_u + \epsilon_u \leq t)$ - $I(\widehat{y}_u \leq t)$ to account for the missing component due to the distribution of the residuals:

$$\widetilde{F}(t) = \widehat{F}_m(t) + \widehat{B}(t),$$

$$\widetilde{F}(t) = \widehat{F}_m(t) + \frac{1}{N}\left[\sum_{u\epsilon \text{U-S}} \Phi^{-1}\left(\frac{t\text{-}\widehat{y}_u}{g(E(\widehat{y}_u);\widehat{\gamma})\ \widehat{\sigma}}\right) - \sum_{u\epsilon \text{U-S}} I(\widehat{y}_u \leq t)\right].$$

The original Chambers and Dunstan estimator has been evaluated by several authors as an estimator of F(t). When the model is specified correctly, Chambers and Dunstan (1986) showed that their model based estimator has greater precision than the design based estimator when model (5) holds, at the cost of slight design bias. Rao, Kovar and Mantel (1990) showed that under model misspecification, the CD estimator can have considerable bias and can be less efficient than other estimators, especially for large samples. Chambers, et al (1992) showed that it is possible, for large n, for other estimators to be more efficient than CD even when the model is correct. Dorfman (1993) reiterated that other model based estimators (and sometimes even the standard design based estimator) can have smaller bias and greater precision than CD in the presence of model misspecification. He also showed that if the model is specified correctly, considerable gains in precision are possible with the CD estimator.

Specification of the residual variance structure $g(E(y_u;\gamma)\sigma)$ is an important component of the CD model based estimator. The form of the function g influences how the residual distribution is put back into the estimated distribution. Figure 3.1 illustrates this influence, the solid curve is $F(t) = \sum I(y_u \leq t)$ for a random sample from a N(10,5) distribution. The three dashed lines represent the curves $\widetilde{F}(t) = \sum I(y_u \leq t) + \sum \Phi^{-1}(\frac{t-y_u}{g(E(y_u);\gamma)\sigma})$ for different combinations of g() and $\sigma$. As g changes from 1 to $y_u$, we see that the lower tail of the CDF is increasingly shortened while the upper tail flattens out. As the heteroscedasticity specified by g increases, more probability is put in the tails of the distribution.

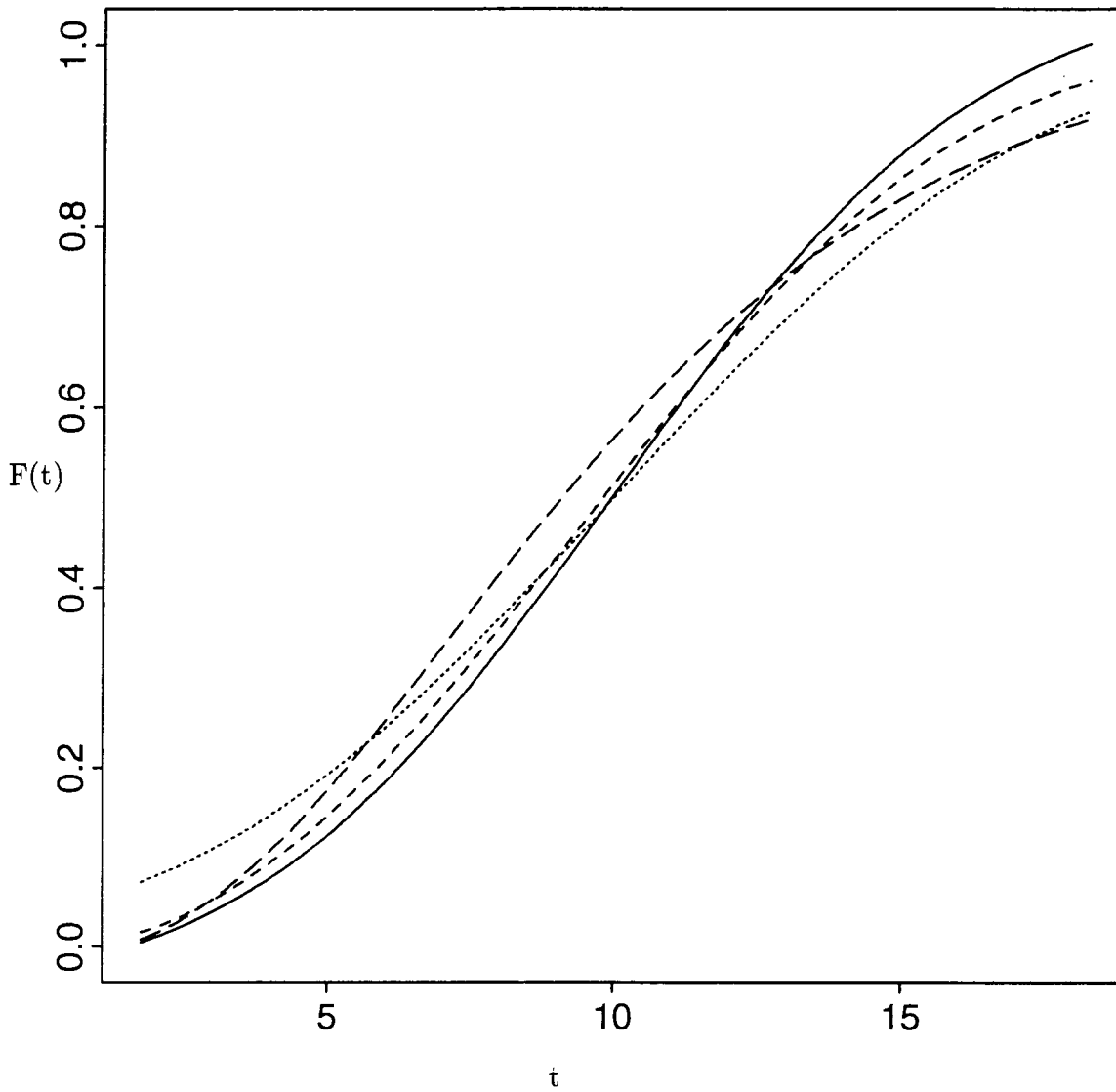


Figure 3.1 Illustration of the influence of the chosen Variance Function on the Estimated CDF. $\hat{F}(t)$ for different combinations of g() and $\sigma$. For the short dashed curve, g() = 1 and $\sigma = \sqrt{\mathrm{Var}(y)}$, for the medium dashed curve, $g() = y_u^{0.5}$ and $\sigma = \sqrt{\mathrm{Var}(y_u^{0.5})}$, and for the large dashed curve g() = $y_u$ and $\sigma = 1$.

## 3.2 The Chambers and Dunstan Estimator for Spatial Populations.

We will consider here the Chambers and Dunstan estimator with several different model based predictions for the nonsample population units to be used in the bias correction term. The two simplest will be predictions from an OLS regression model, and 'spatial' predictions from a kriging model:

$$\widehat{y}_u = x_u\widehat{\beta}\ ,\ \ \widehat{\beta} = (X'X)^{-1}X'Y$$

$$\widetilde{y}_u = x_u\widetilde{\beta} + \Sigma_0'\Sigma^{-1}(Y - X\widetilde{\beta}\ ),\ \ \widetilde{\beta}\ = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$
$$\forall\ u\ \varepsilon\ \text{U-S}$$

$x_u$ is the vector of covariates (with a 1 for the intercept) for a nonsample population unit, X is the matrix of covariates (and a column of ones) for the sample units, Y is the vector of $y_u$ values for the sample units, $\Sigma_0$ is the spatial covariance matrix between the sample units and the nonsample units and $\Sigma$ is the spatial covariance among the sample units. A third set of predictions is constructed to exploit the results from using scaled residuals to estimate spatial residual covariance as discussed in the previous chapter. Recall that given a set of OLS residuals $\widehat{\epsilon}_u$, these are scaled by the square root of the variance function $g()\sigma^2$ to obtain the scaled set $\widetilde{\epsilon}$ , before estimating the covariance function. Given this estimated residual spatial covariance function, predicted residuals are obtained for the nonsample population units:

$$\overset{*}{\epsilon} = J_0(J'\widehat{\Sigma}^{-1}J)^{-1}J'\widehat{\Sigma}^{-1}\widetilde{\epsilon}\ \ + \widehat{\Sigma}_0'\widehat{\Sigma}^{-1}(I - J(J'\widehat{\Sigma}^{-1}J)^{-1}J'\widehat{\Sigma}^{-1})\widetilde{\epsilon}$$

Where $J_0$ is the (N-n) $\times$ 3 matrix of spatial locations (with a column of ones for the intercept term) of the nonsampled population units, J is the matrix for the sampled population units. The covariance matrices $\Sigma$ $\Sigma_0$ are estimated using the exponential covariance model and REML (Christensen, 1991). These predicted residuals on the nonsample are then rescaled accoi ding to the same variance function used for scaling the $\widehat{\epsilon}_u$, and added to the OLS predictions.

$$y'_u = x_u\widehat{\beta} + \widehat{g}\ \widehat{\sigma}\overset{*}{\epsilon}_u$$

As stated in the previous chapter, residual scaling is of interest primarily when estimation of residual spatial covariance from a stratified sample for the entire population of residuals is required. For nonstratified populations, or when analyzing strata separately, scaling is not necessarily needed. In the previous chapter residual scaling was shown to influence the strength of the estimated correlation of the residual spatial pattern and the estimated range over which the residuals are spatially correlated. In this chapter we seek to determine if the inclusion of spatial information as a surrogate for a spatially patterned missing covariate improves the prediction of the nonsample values and model based estimates of the finite population CDF and other parameters.

## 3.3 Simulation Study.

Several nonstratified and stratified populations were generated to assess the performance of the Chambers and Dunstan estimator using the various model based predictions discussed in the previous section. The populations were simulated using the fitted regression equations from the original Eastern Lake Survey data to be analyzed in chapter 4. The nonstratified populations were created using the fitted regression equation for the first stratum of this original survey. The first two stratified populations have stratum specific regression coefficients for the missing covariate. A third stratified population was constructed where the regression coefficient for the missing covariate was the same for all strata. The simulated populations are as follows:

Nonstratified Populations:

Population N1 (Size=600):

$$y_u = -4900 - 0.49 * x_{1u} + 1270 * x_{2u} + e_u, \quad e_u \sim N(0,1)$$

Population N2 (Size=600):

$$y_u = -4900 - 0.49 * x_{1u} + 1270 * x_{2u} + e_u, \quad e_u \sim N(0, 25 * E(y_u))$$

Stratified Populations:

Population S1 (Size=600, 200 in each stratum): Separate models for each stratum:

$$y_{1u} = -4900 - 0.49{*}x_{1u} + 1270{*}x_{2u} + e_{1u}, \qquad e_{1u} \sim N(0,1)$$
$$y_{2u} = -7535 + 0.564{*}x_{1u} + 1783{*}x_{2u} + e_{2u} \qquad e_{2u} \sim N(0,1)$$
$$y_{3u} = 17730 - 1.57{*}x_{1u} - 3833{*}x_{2u} + e_{3u}, \qquad e_{3u} \sim N(0,1)$$

Population S2 (Size=600, 200 in each stratum): Separate models for each stratum:

$$y_{1u} = -4900 - 0.49{*}x_{1u} + 1270{*}x_{2u} + e_{1u}, \qquad e_{1u} \sim N(0,E(y_{1u}))$$
$$y_{2u} = -7535 + 0.564{*}x_{1u} + 1783{*}x_{2u} + e_{2u}, \qquad e_{2u} \sim N(0,E(y_{2u}))$$
$$y_{3u} = 17730 - 1.57{*}x_{1u} - 3833{*}x_{2u} + e_{3u}, \qquad e_{3u} \sim N(0,E(y_{3u}))$$

Population S3 (Size=600, 200 in each stratum): Separate models for each stratum:

$$y_{1u} = 400 - 0.49{*}x_{1u} + 50{*}x_{2u} + e_{1u}, \qquad e_{1u} \sim N(0,E(y_{1u}))$$
$$y_{2u} = 200 + 0.564{*}x_{1u} + 50{*}x_{2} + \epsilon_{2u} \qquad e_{2u} \sim N(0,E(y_{2u}))$$
$$y_{3u} = 2000 - 1.57{*}x_{1u} + 50{*}x_{2} + e_{3u}, \qquad e_{3u} \sim N(0,E(y_{3u}))$$

For all populations, $x_{1u}$ is the elevation of the unit, $x_{2u}$ and is the pH of the unit. The pH variable was treated as unknown in the data analysis and CDF estimation procedure for these populations in order to induce the missing covariate effect. The pH variable has a simple spatial pattern and induces spatial correlation in the residuals in addition to increased residual variance. For each design the sample consists of $y_u$, elevation, latitude, longitude (transformed to accurately represent geographic distances) known on the sample units, and elevation, latitude, longitude known on the nonsample units. For all populations, four models were used to analyze the sample data and for making predictions for the nonsample units:

REG1: $y_u = \beta_0 + \beta_1 \text{elevation}_u + \epsilon_u$

$$\text{Var}(\epsilon_u) = \sigma^2$$
$$\text{Cov}(\epsilon_u , \epsilon_{u'}) = 0.$$

REG2: $y_u = \beta_0 + \beta_1 \text{elevation}_u + \beta_2\, i_u + \beta_3\, j_u + \epsilon_u$

$$\text{Var}(\epsilon_u) = \sigma^2$$
$$\text{Cov}(\epsilon_u , \epsilon_{u'}) = 0.$$

UK1: $y_u = \beta_o + \beta_1 \text{elevation}_u + \beta_2\, i_u + \beta_3\, j_u + \epsilon_u$

$$\text{Var}(\epsilon_u) = \sigma^2$$

$$\text{Cov}(\epsilon_u\ ,\ \epsilon_{u'}) = \sigma(d_{uu'};\ \theta)$$

UK2: This model is fit in two stages, $y_u = \beta_o + \beta_1 \text{elevation}_u + \epsilon_u$,

$$\text{Var}(\epsilon_u) = \sigma^2\ ,\ (\text{or Var}(\epsilon_u) = \sigma^2 y_u),$$

$$\text{Cov}(\epsilon_u\ ,\ \epsilon_{u'}) = 0.$$

Fitted by OLS , and

$$E(\epsilon_u \mid i_u, j_u) = \mu_0 + \alpha_1 i_u + \alpha_2\, j_u$$

$$\text{Var}(\epsilon_u) = \sigma^2,$$

$$\text{Cov}(\epsilon_u\ ,\ \epsilon_{u'}) = \sigma(d_{uu'};\ \theta)$$

Fitted by REML.

Predictions from this model are the sum of the predicted values from the two stages. The residuals from the first stage are scaled by a standard deviation function (either heterogeneous or homogeneous) before estimating $\mu_0$, $\alpha_1$, $\alpha_2$ and $\theta$.

The nonstratified populations were sampled with simple random samples of size n=75. All four models (REG 1, REG 2, UK 1, UK 2) were used to analyze the samples. The regression parameters for REG 1 and REG 2 are unweighted least squares estimates (OLS). The model parameters for UK 1 are estimated using REML. For UK2, $\beta_0$ and $\beta_1$ are estimated using OLS and the spatial parameters $\mu$, $\alpha_1$ and $\alpha_2$ are estimated using REML. Homogeneous residual scaling was used between the two stages of estimation in model UK 2 for all populations and heterogeneous scaling was used in addition for population S3. For population S1 an additional model was fit:

UK 3: $y_u = \beta_o + \beta_1 \text{elevation}_u + \epsilon_u$

$$\text{Var}(\epsilon_u) = \sigma^2$$

$$\text{Cov}(\epsilon_u\ ,\ \epsilon_{u'}) = \sigma(d_{uu'};\ \theta)$$

This differs from UK 1 because the spatial coordinates are not included in the regression equation. We wish to investigate the difference in results between

including the spatial information via the spatial coordinates, spatial covariance or both.

The predicted values for the nonsample are the sum of the OLS predictions from stage 1 and the predicted residuals from stage two. For each set of model based predictions, the Chambers and Dunstan estimator was used with both variance function g=1 and g=E($y_u$).

The stratified populations were sampled with a simple random sample of n=25 in each of the three strata. Each stratum CDF was estimated using the within stratum OLS predictions for the the regression model REG 1 and REG 2 as described above for a simple random sample. In addition, estimation using a spatial analysis of the entire set of residuals across strata using model UK 2 was used with homogeneous scaling and heterogeneous scaling of the OLS residuals between the two stages. The predicted values for the nonsample units are then the sum of the OLS predictions from the stratum specific regression equations and the spatially predicted residuals. For each set of model based predictions, the Chambers and Dunstan estimator was used with both variance functions g=1 and g=E($y_u$).

For comparison, results for the design based estimator (8) and the model based estimator (9) are included. An estimate of the population mean and standard deviation are computed for each estimated CDF by first computing estimates of the first two population moments. $\widehat{\mu}_1$ is the estimated population mean and the estimated population standard deviation is $\widehat{\sigma}$. The values of $p_i$ are obtained from the estimated CDF by taking the step height that leads up to the value $y_i$ (figure 3.2).

$$\widehat{\mu}_1 = \sum_{i=1}^{N} p_i y_i \qquad \widehat{\mu}_2 = \sum_{i=1}^{N} p_i y_i^2 \qquad (12)$$

$$\widehat{\sigma} = \sqrt{\widehat{\mu}_2 - \widehat{\mu}_1^2}\,. \qquad (13)$$

In addition to being of interest in their own right, estimating the population mean and standard deviation in this manner provide a measure of the overall goodness of fit of the estimated CDF to the actual finite population CDF. Given a set of estimated CDFs for a set of samples, the average bias and root mean square error for the estimated mean and standard deviation are also calculated and the average bias and root mean square error of $\widehat{F}(t)$ where the values of t are the true population values of $F^{-1}(p)$ for $p \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$. The means and standard deviations for the simulation study are estimated using $F^{-1}(p)$ for these 11 lattice points. In this context, the mean and sd are also a measure of how well the estimated CDF fits the true CDF at these population points.

$$\text{BIAS}=\frac{1}{M}\sum(\widehat{\mu_1}-\mu)$$

$$\text{RMSE}=\sqrt{\frac{1}{M}\sum(\widehat{\mu}_1-\mu)^2}$$

$$\text{BIAS}=\frac{1}{M}\sum(\widehat{\sigma}-\sigma)$$

$$\text{RMSE}=\sqrt{\frac{1}{M}\sum(\widehat{\sigma}-\sigma)^2}$$

$$\text{BIAS}=\frac{1}{M}\sum(\widehat{F}(t)-F(t))$$

$$\text{RMSE}=\sqrt{\frac{1}{M}\sum(\widehat{F}(t)-F(t))^2}$$

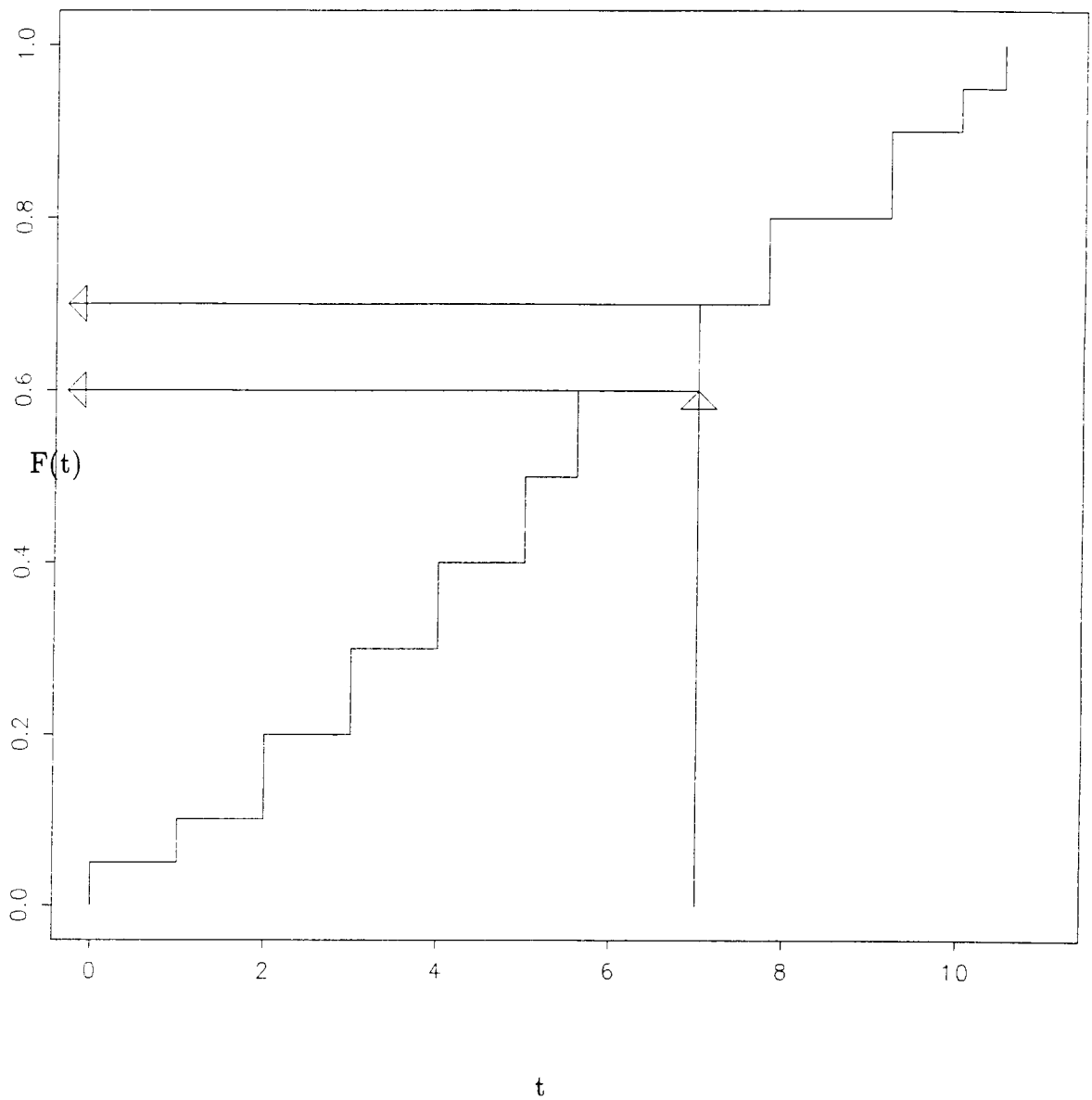


Figure 3.2 Example: Calculation of an Individual Probability. The probability associated with each $y_i$ is the height of the step in the estimated CDF.

M is the number of replications of the sampling/estimation strategy and the summation is over all M replications. 100 samples were drawn from each of the populations, and the Chambers and Dunstan estimator is used with the various model based predictions and the two standard deviation functions described above. The results for the bias and rmse of the estimated means, standard deviations and CDFs are summarized below.

## 3.4 Results: Estimation of Means and Standard Deviations.

Tables showing the bias and root mean square error for the estimated population means and standard deviations are shown in the tables in Appendix 1. The design based estimators (labeled hte for Horvitz Thompson Estimation) are unbiased for the population parameters according to the theorem in Chapter 1, therefore the bias in hte is indicative of the representativeness of the 100 samples used in the simulations. 100 replications is a very low number and more will be used in future investigations of this work. The greatest anticipated gains from using model based estimators will be increased accuracy relative to hte or to each other. Though both the naïve and the CD estimator are biased we expect the bias of CD to be less because it puts more of the distribution of the residuals back into the estimate, which the naive estimator does not. The model based means and standard deviations were estimated as in equations (12) and (13) using the estimated CDFs.

For the nonstratified population N1, the naïve estimator using predictions from REG2 had better precision than the design based estimator for the mean. The CD estimators gave additional improvements in bias and precision over the naïve estimator for the mean. For the population standard deviation we see the same pattern of improvements when predictions from UK 1 are used. In Chapter 2 we saw that only seven samples from N1 had significant spatial covariance, though the spatial correlation at short distances was strong (median value of 0.99) and the correlation range was large (median value of 98.5 km). Therefore this improvement in estimation of the standard deviation is due to these samples.

For population N2 the naïve estimator using REG 2 predictions did not improve precision over the design based estimator but the CD estimator did improve the bias for both the estimated population mean and standard deviation and improved the precision for the mean. For the estimated population standard deviation, the CD estimators have slight improvements over the design based estimator but not large. Using predictions from UK1 did not improve over REG 1 prediction in the model based estimators. Chapter 2 found that the correlation for small distances was moderate (median value of 0.511) and the range was large (median value of 110.9 km). The lack of additional gains when including the residual covariance seems to be due to the lack of strong residual correlation. Recall that the only difference between populations N1 and N2 is the variance structure of the residual distribution, the increased variance of N2 reduced the estimated residual correlation from a median value of 0.99 to 0.511. Put another way, the increased residual variation decreased the missing covariate effect, because the missing covariate was a smaller part of the unexplained population variability for N2.

For the stratified populations, the naïve estimator using REG 2 predictions improved the precision over the design based estimator for Stratum 1 of S1. The CD estimators reduced the bias and also improved the precision slightly for the mean, improved the bias for the stratum standard deviation but worsened the precision slightly. Adding residual covariance through UK 2 predictions improved the precision and bias for the mean, but gave similar results for the stratum standard deviation. From Chapter 2 the short range correlation was moderate to weak (mean value of 0.73, median value of 0.4) and the range was short (mean value of 37.63 km, median value of 11.7 km), therefore the residual covariance was not strong or spatially extensive when the spatial coordinates are already included in the model. For Stratum 2, of S1 both the naïve and CD estimators have better precision than the design based estimator. The CD estimator gives essentially the same answer as the naïve with a slight worsening in bias and improvement in precision (remember that the residual variance is small, therefore the residuals are a small component of the population variability here) using UK 2 predictions reduced bias and improved precision slightly for both the stratum mean and standard deviation. For Stratum 3, the naïve estimator using REG 2 predictions improves the precision over the design based estimator. The CD estimators are worse than the naïve estimators for both the mean and standard deviation.

For population S2, the naïve estimator had better precision than the design based estimator for the stratum mean but not for the stratum standard deviation. The CD estimators improved the bias for the mean but worsened the precision. For the standard deviation, the CD estimators worsened the bias and the precision relative to the naïve estimator. For Stratum 2, the naïve estimator improved the precision for the mean and standard deviation over the design based estimator. CD estimators improved precision and worsened bias for the mean, and improved both for the standard deviation. For stratum 3, the naïve again improved the precision over the design based estimator, the CD estimators worsened the bias and the precision for the stratum mean, but improved bias and precision for the stratum standard deviation. For this population the correlation at short distances was weak (mean value of 0.549, median value of 0.475) and the range was short (mean value of 13.85 km, median value of 6.31 km) and so the model based estimators using UK 2 predictions did not have great improvements over those using REG 2 predictions.

For population S3, the naïve estimator improved the precision over the design based estimator for all strata, the CD estimators improved precision and bias for strata 1 and 2 and all model based estimators are the same for stratum 3. Short range correlation was weak (median value of 0.032) and the correlation range was short (median value of 8.49 km), therefore adding residual covariance did not improve the results when the spatial coordinates were already in the regression equation.

In general the anticipated behavior was realized, for a given set of model based predictions, the naïve estimator had better precision than the design based based estimator but was considerably biased in some cases. The CD estimator using the same model based predictions generally reduced the bias and in some cases improved the precision as well. There were some exceptions to this general result. For the populations where the residual variance was largest, the naïve estimator did not improve the precision over the design based estimator. For the populations where the residual variance was small, the CD estimator did not always improve the bias of the naïve estimator for estimating the population mean. This makes sense because the model based predictions used in the estimators are predicting E[Y|X] and are optimal for the mean, regardless of the magnitude of the residual variation. The CD estimators did improve the results

for the estimated standard deviation in these cases, because putting the residual distribution back in reflects the true population variability.

Using spatial information as a surrogate for the spatially patterned missing covariate improved the estimation of the means and standard deviations. For population N1 including the spatial coordinates in the regression model improved the precision of estimates of the mean and standard deviation. Adding residual spatial covariance to the model improved the precision of the estimated standard deviation only. Using residual covariance in the model but not the spatial coordinates (model UK 3) improved the precision over REG 1 (which uses no spatial information at all) but the improvement is not as great as when using spatial coordinates. For population N2 adding the spatial coordinates did not improve the precision of the parameter estimates when using the naïve estimator, but did improve the precision when the CD estimator was used. Adding spatial covariance to these actually worsened the precision.

For the stratified populations S1 and S2, including the spatial coordinates in the regression equation improved the precision of the estimated mean and standard deviation. Additionally, adding residual spatial covariance improved the precision slightly more. For population S1 the parameter estimates obtained by using residual spatial covariance without the spatial coordinates in the regression equation (model UK 3) gave less precise estimates than using no spatial information at all. For this model the short range correlation had a median value of 1 over the samples and the correlation range had a median value of 29.8 km which is an increase in strength and spatial extent over the model with the spatial coordinates ( which had median values 0.4 and 11.7 respectively). With this strong residual spatial covariance the lack of improvement is surprising, but seems to be due to the fact that the planar missing covariate is more accurately modelled by the spatial coordinates than by a residual spatial covariance function alone. In addition, the residual spatial covariance for the stratified populations is estimated using scaled residuals where the scaling function is estimated from the sample. The increased variability due to estimating the scaling function seems to explain why the the model using residual covariance only improves the results over using no spatial information for the nonstratified population but doesn't for the stratified one. For population S3, adding spatial information gave essentially the same results as including only the known covariates. The spatially patterned missing covariate was a small part of the unexplained population variability.

## 3.5 Results: CDF Estimation.

Figures 3.3 - 3.16 show the superimposed estimated distribution functions for the 100 replications for each of the populations using the design based estimator and the naïve and Chambers and Dunstan estimators using predicted values from REG 2. In addition, results for the bias and root mean square error for the estimated percentiles are shown in Appendix 2.

Figure 3.3 shows the estimated distribution functions for the nonstratified population N1, the improvement in precision from the model based estimators is highly visible. Examination of the tables in Appendix 2 show that UK 1 predictions with a homogeneous variance model and REG 2 predictions with a heterogeneous variance model had similar precision for the estimated percentiles, with a slight edge to UK 1. Thus, including the spatial coordinates in the regression model improved the estimation of the CDF. Including the estimated residual covariance in the prediction improved the results slightly more. As noted earlier, this population had only 7 samples with significant spatial covariance, so the gains were only made for those samples. For population N2, figure 3.4 shows the improvement in precision when the model based estimates are used. The bias correction of the CD can be seen relative to the naïve estimator. The naïve estimator puts more of the probability in the middle of the distribution giving it an 'S' shape, the CD estimators give the distribution a straighter look, more like the design based estimators. Using kriging predictions from UK 1 led to gains in precision for the estimated percentiles, except for in the upper tail of the distribution.

For the stratified populations S1-S3, the distribution function estimates are estimated separately for each stratum. The most obvious result is the improvement in precision of the model based estimators over the design based estimators (figures 3.5, 3.7, 3.9, 3.11-3.16). The figures shown use REG 2 predicted values. The distribution function estimates using UK 3 predictions are shown for the three strata of population S1 in figures 3.6, 3.8, and 3.10. Improvement in precision over the design based estimator can be seen in all cases, but the improvement is not as great as when predictions from REG 2 are used. This shows that the spatial coordinates in the regression equation are a better

model for the missing covariate than the residual spatial covariance function alone. Looking at the estimated percentiles in Appendix 2, the CD estimators with REG2 predictions led to improvements in precision over REG 1 for some percentiles. Adding residual covariance using UK2 did not lead to further improvements. For S3 CD estimators with REG 1 predictions had the best precision.

## 3.6 Conclusions.

Greatest gains in precision over the design based estimates of population parameters from model based estimators were realized by including the known explanatory variable elevation and the spatial coordinates in the regression model. Including residual spatial covariance to the prediction model lead to small additional gains in some cases, especially for estimating the population (or stratum) standard deviation. The spatial pattern of the missing covariate influences the best way that spatial information should be included. Even though residual spatial correlation was stronger when spatial coordinates were not included in the regression equation the estimation of population parameters was poorer than when spatial coordinates were used in the regression equation. The missing covariate in these simulated data is a planar surface and so it was anticipated that the coordinates alone would be adequate, however for other populations where the missing covariates are highly spatially patterned, the spatial covariance component is likely to lead to larger gains than realized here. Therefore the planar component in the model is a better surrogate than a spatial residual covariance function. Using both spatial components together was a better surrogate. In general, the nature of the spatially patterned missing covariates will determine the best way to include spatial information.

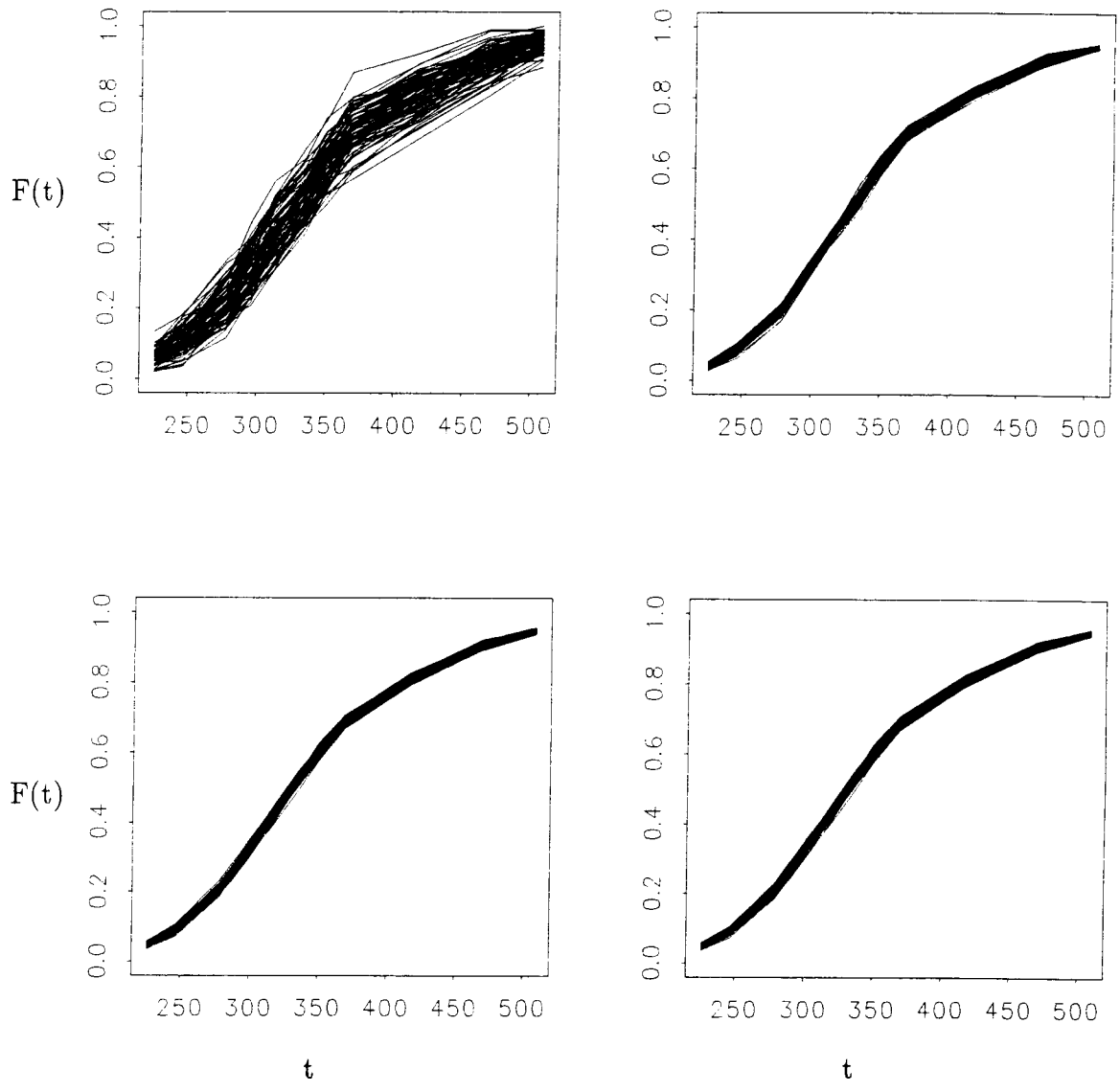


Figure 3.3. Estimated CDFs for Population N1 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

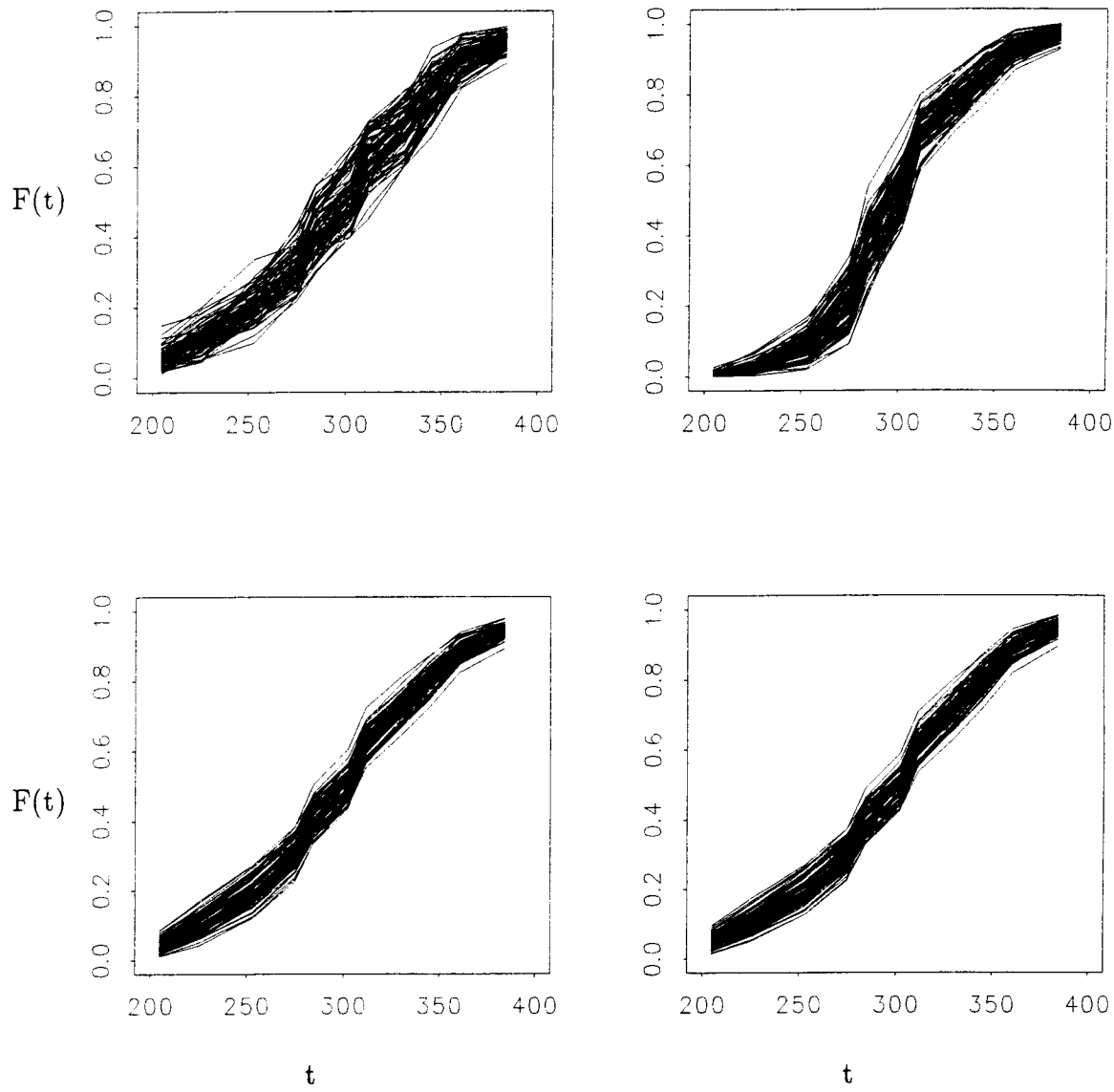


Figure 3.4. Estimated CDFs for Population N2 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

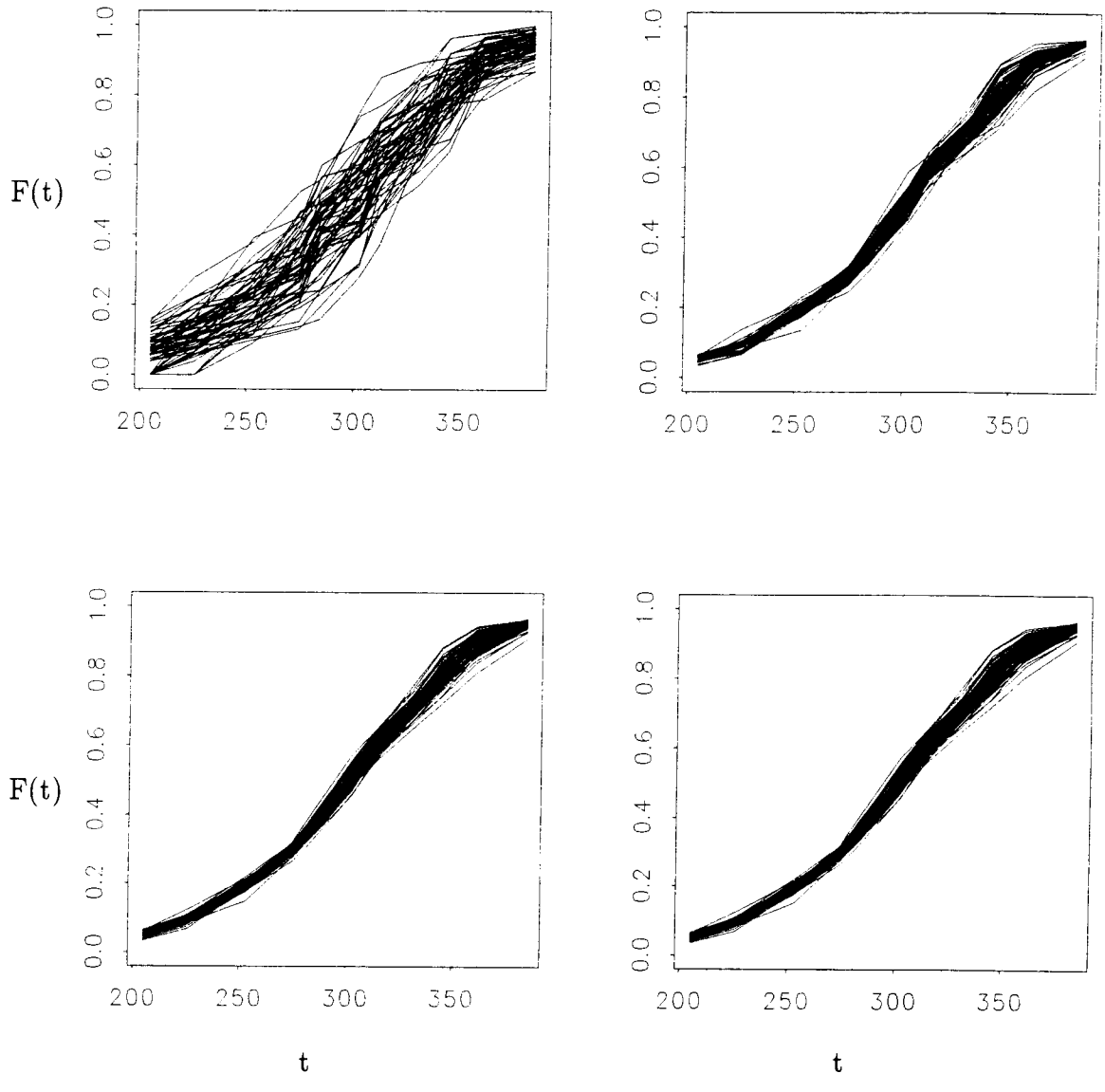


Figure 3.5. Estimated CDFs for Population S1 Stratum 1 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

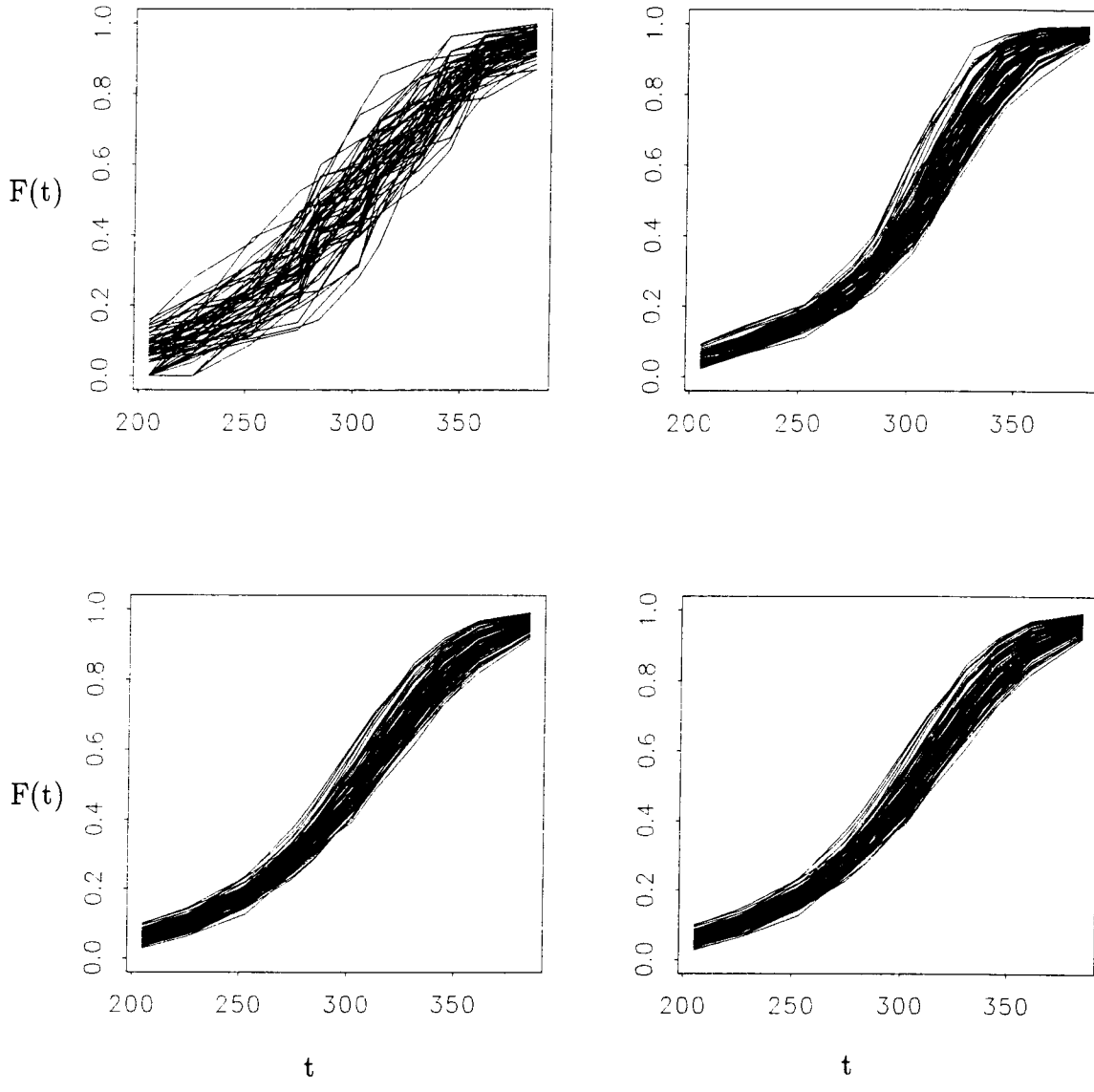


Figure 3.6. Estimated CDFs for Population S1 Stratum 1 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model UK 3.

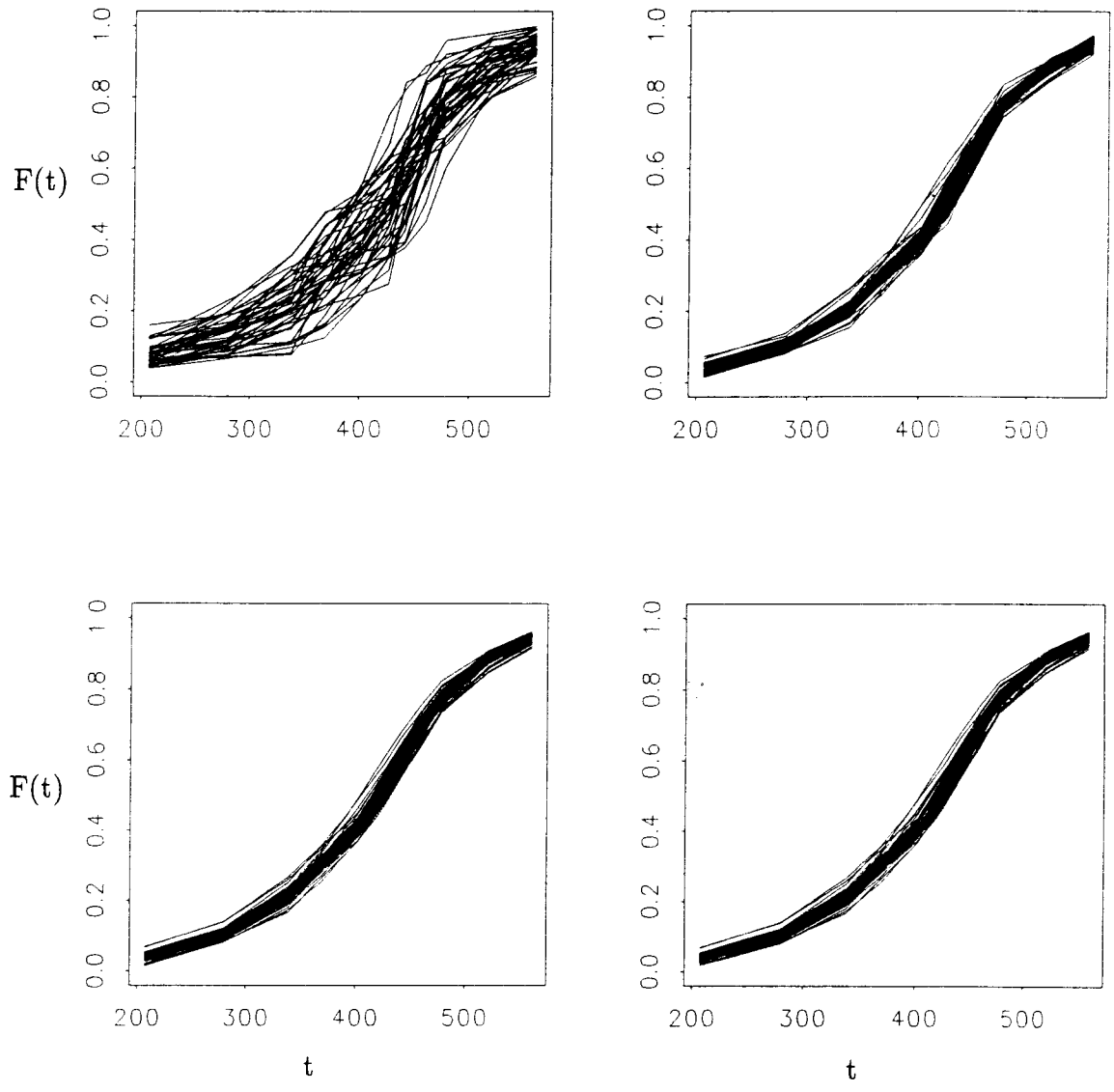


Figure 3.7. Estimated CDFs for Population S1 Stratum 2 for the 100 replications. Design based (upper left), naive (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

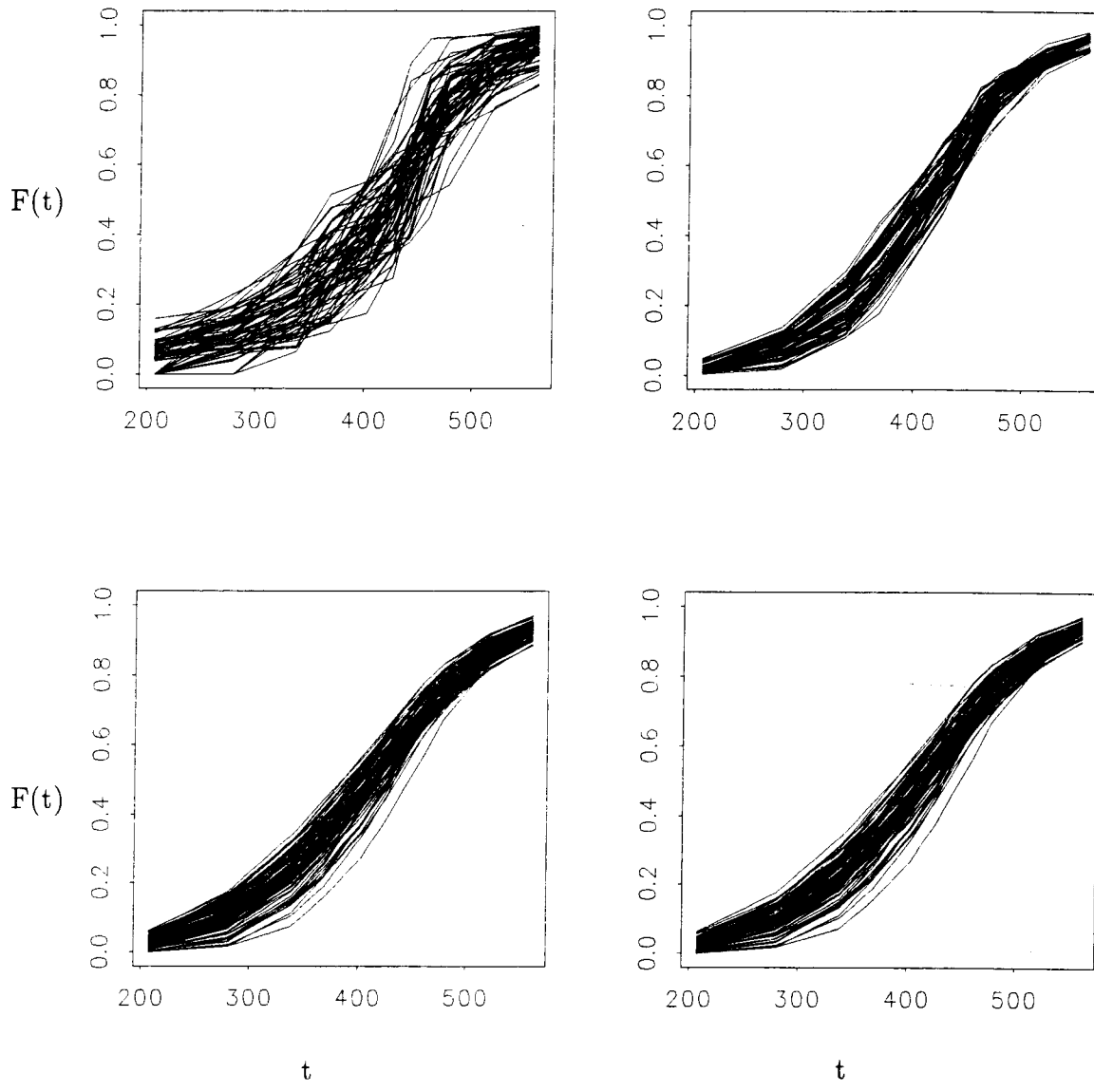


Figure 3.8. Estimated CDFs for Population S1 Stratum 2 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model UK 3.

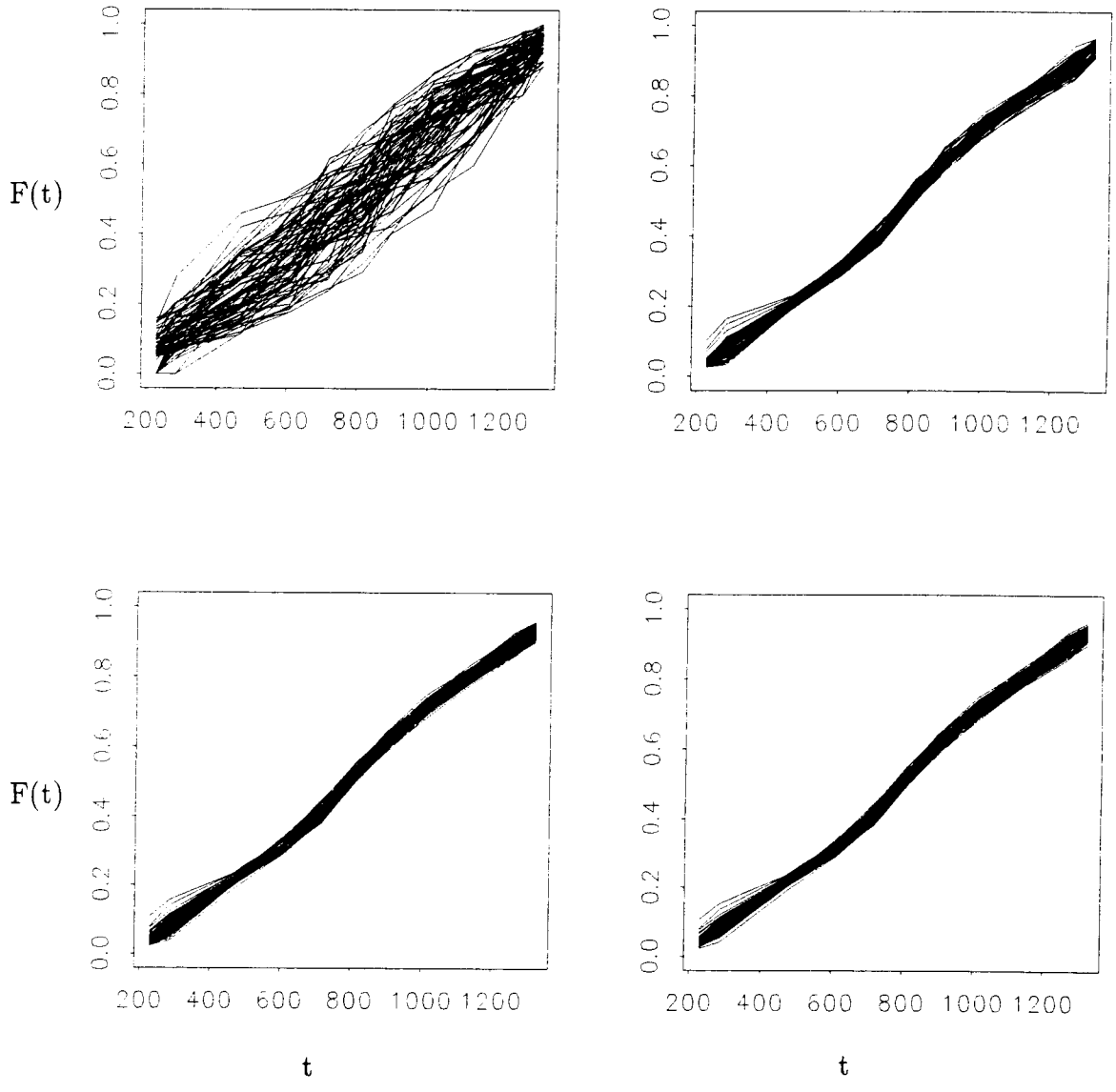


Figure 3.9. Estimated CDFs for Population S1 Stratum 3 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

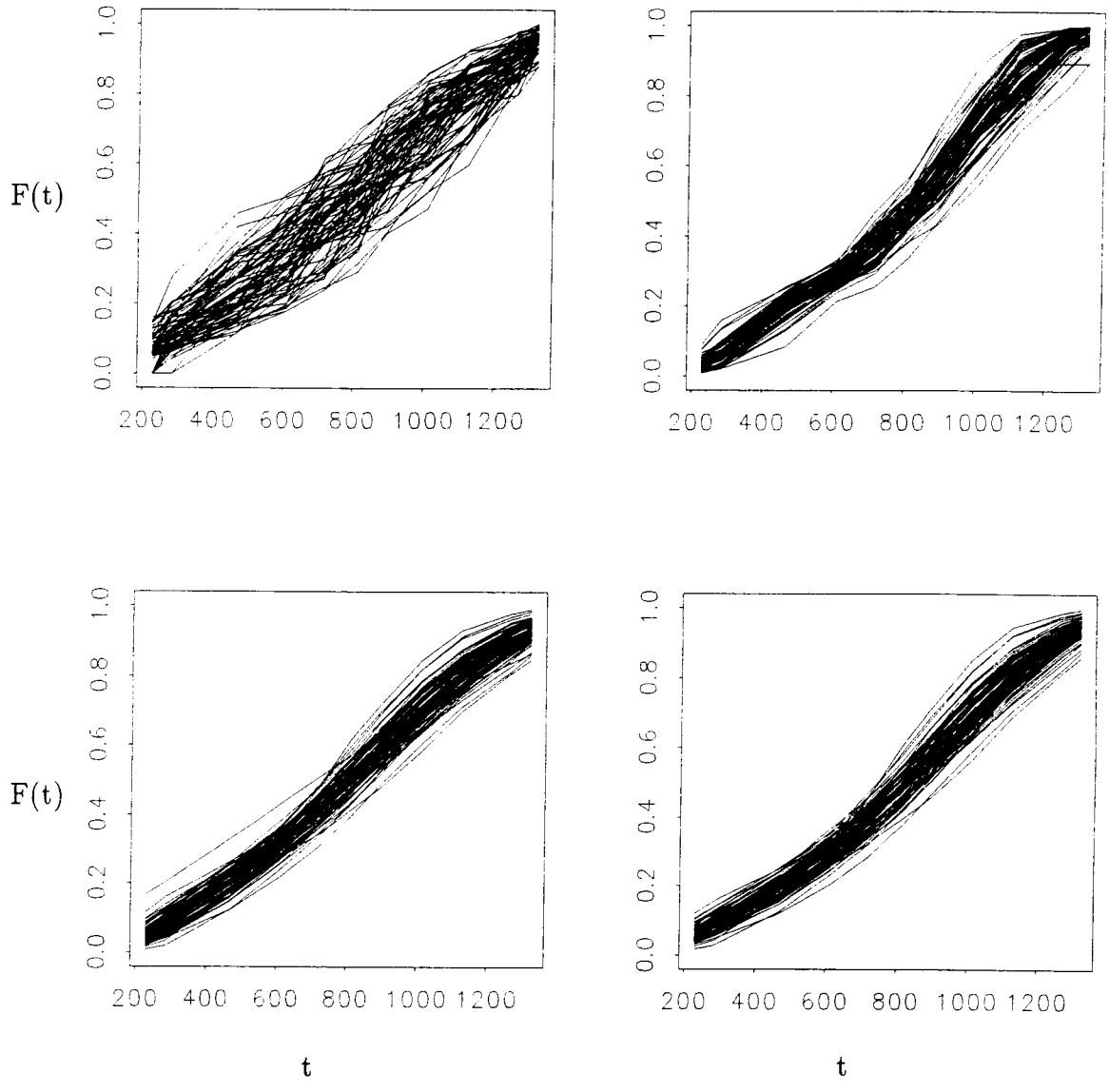


Figure 3.10. Estimated CDFs for Population S1 Stratum 3 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model UK 3.

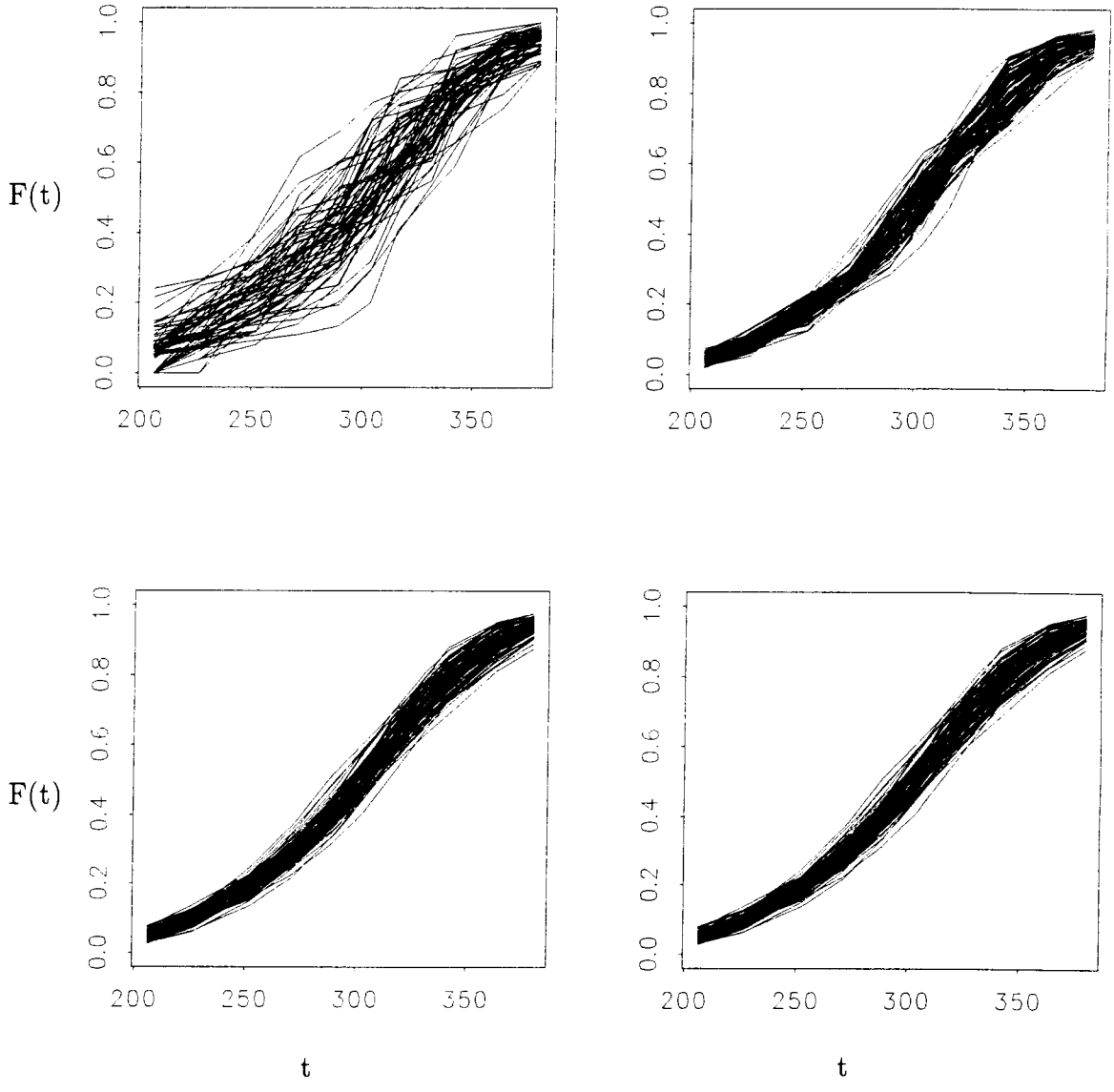


Figure 3.11. Estimated CDFs for Population S2 Stratum 1 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

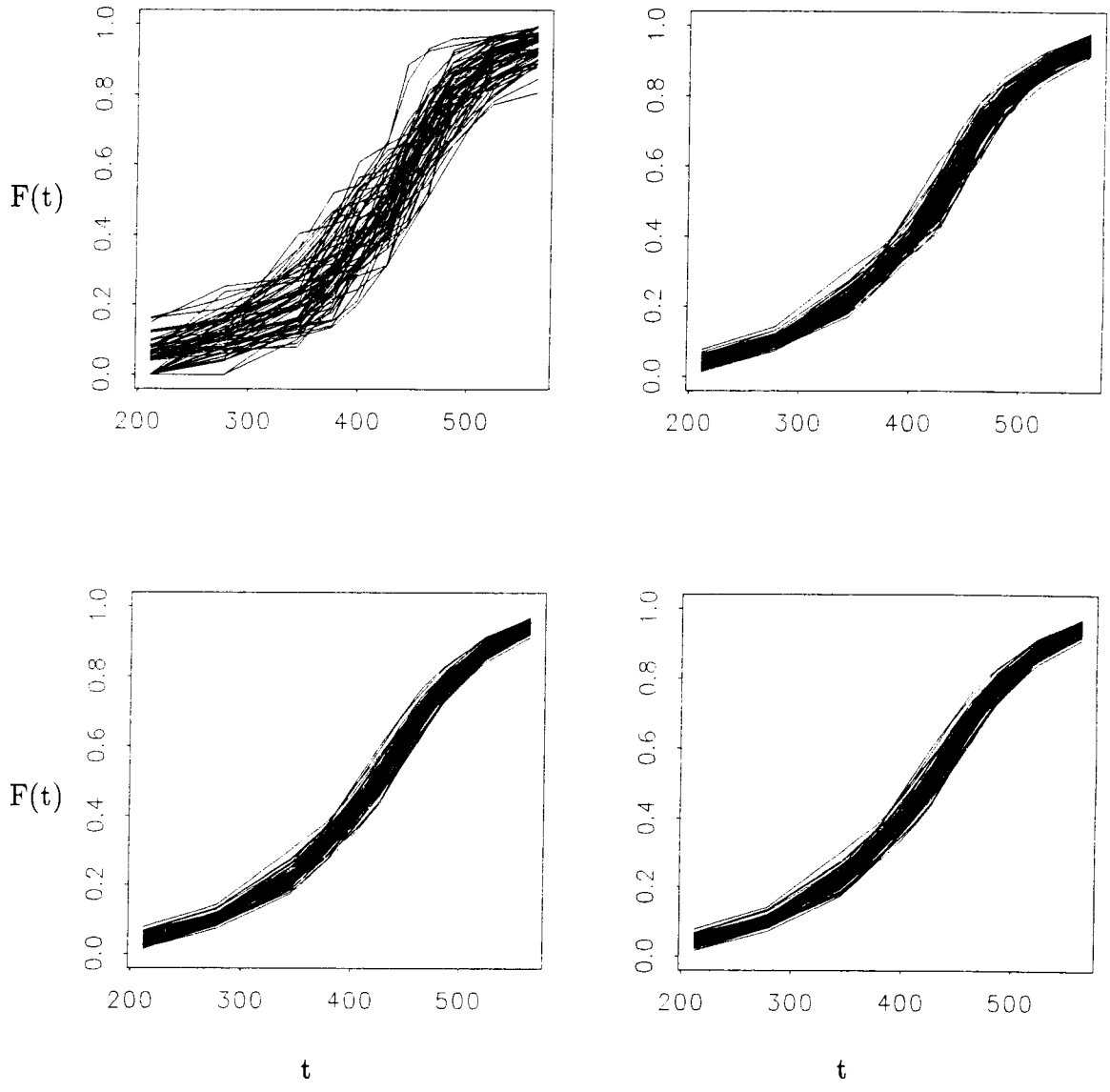


Figure 3.12. Estimated CDFs for Population S2 Stratum 2 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

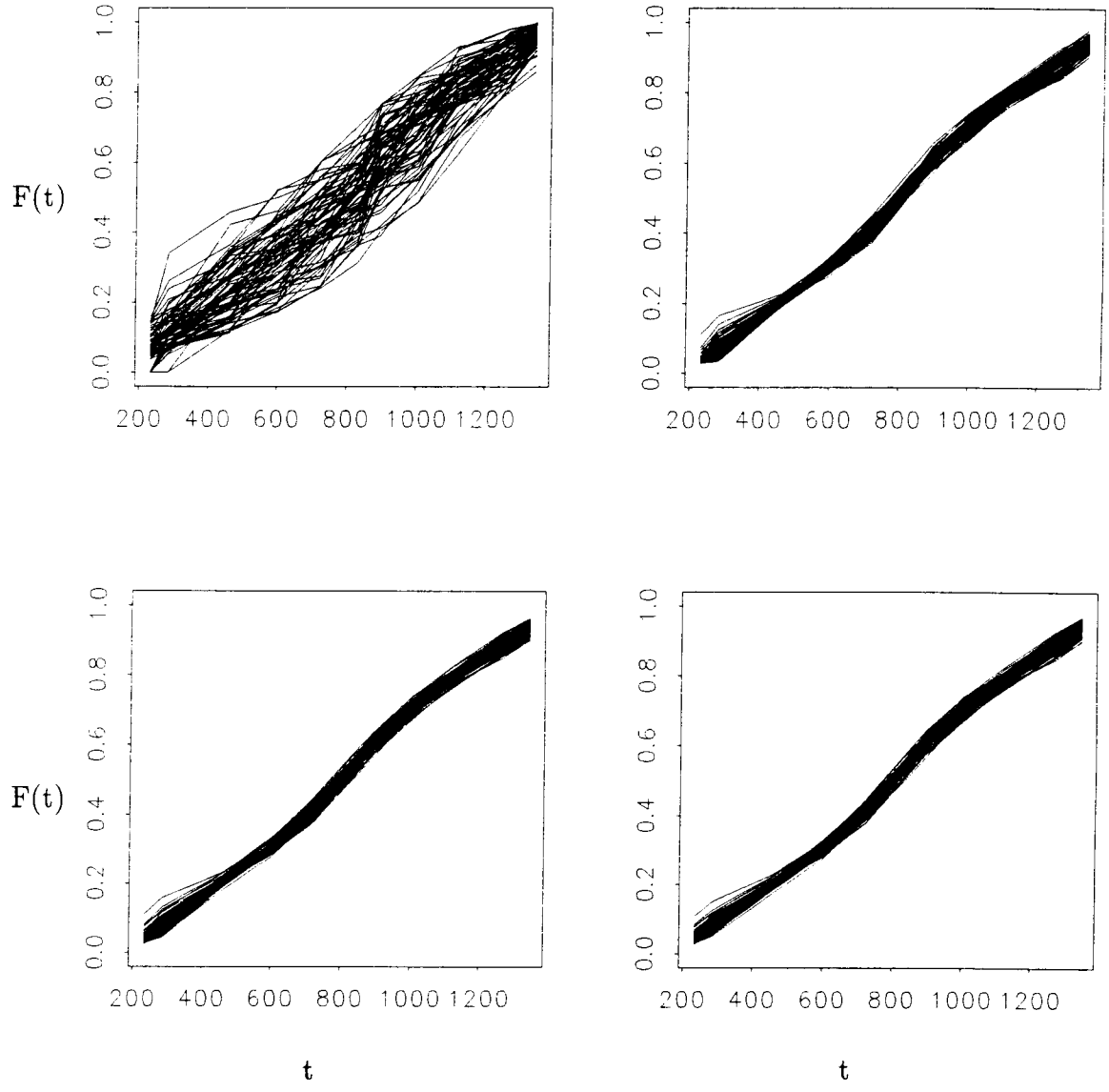


Figure 3.13. Estimated CDFs for Population S2 Stratum 3 for the 100 replications. Design based (upper left), naive (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

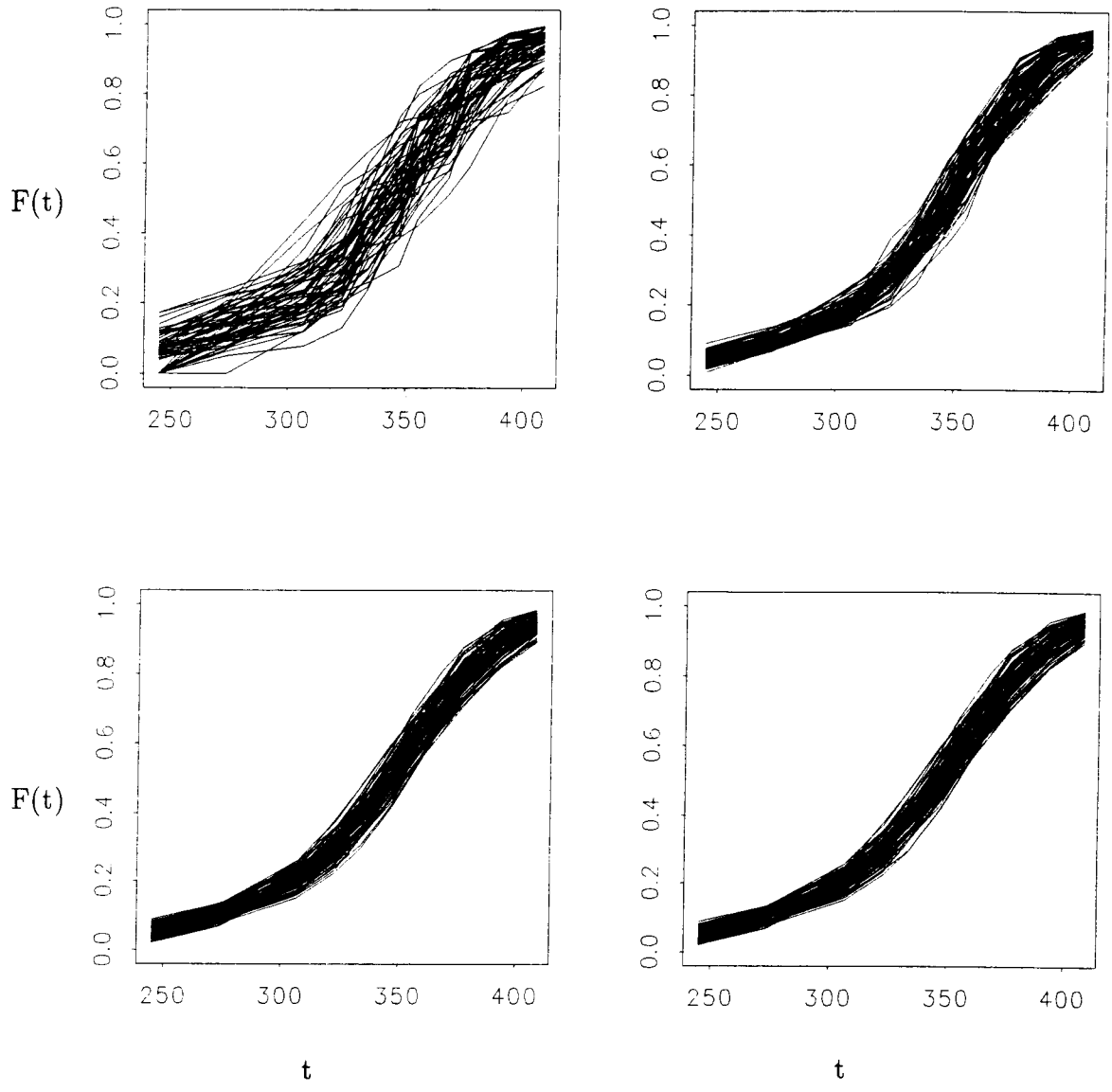


Figure 3.14. Estimated CDFs for Population S3 Stratum 1 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

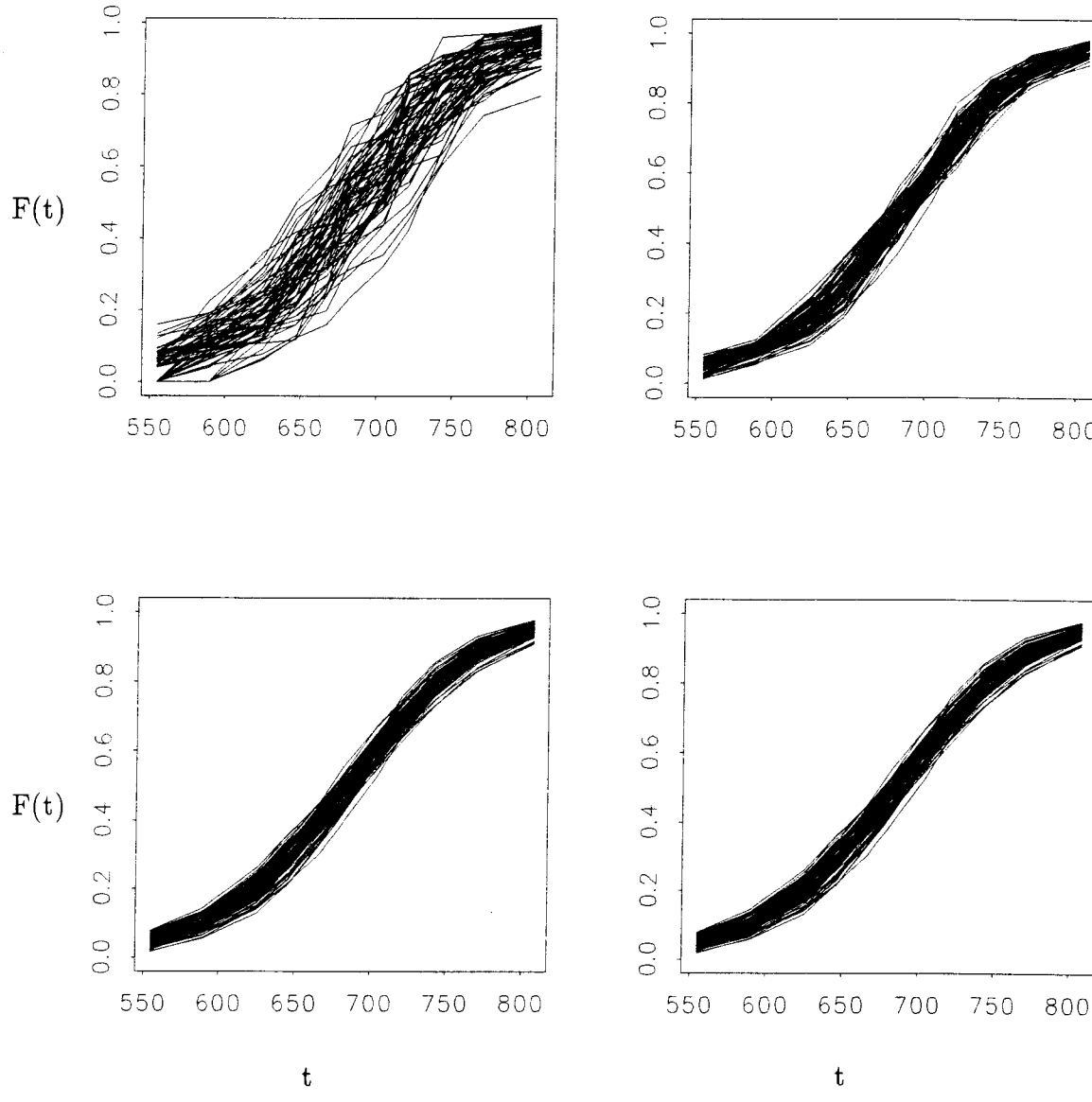


Figure 3.15. Estimated CDFs for Population S3 Stratum 2 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

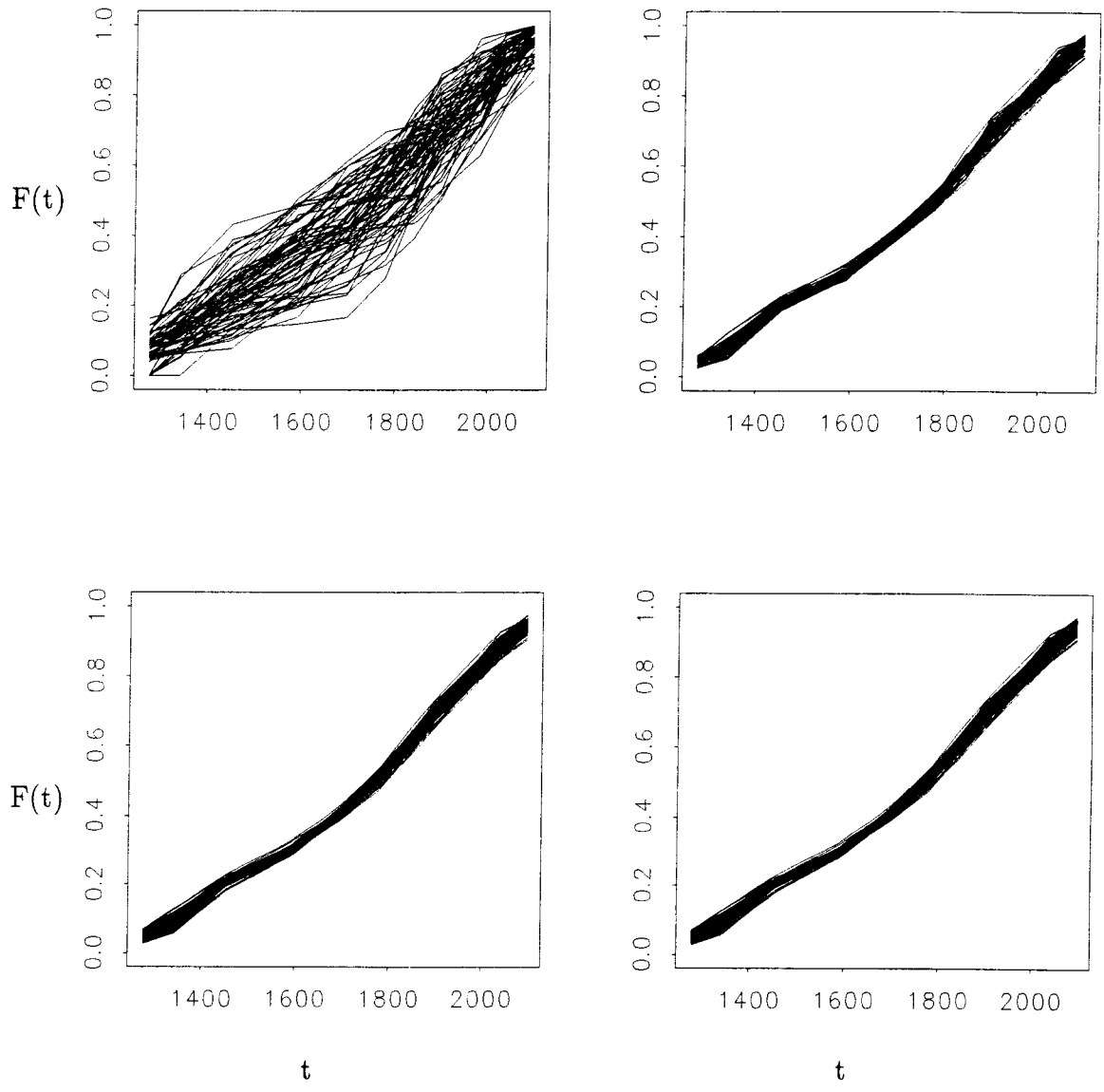


Figure 3.16. Estimated CDFs for Population S3 Stratum 3 for the 100 replications. Design based (upper left), naïve (upper right), CD with a homogeneous variance function (lower left) and CD with a heterogeneous variance function (lower right). Model based estimators use predictions from model REG 2.

# 4. Model Based Estimation of Population Parameters: Case Study from the Eastern Lake Survey.

## 4.1 Introduction.

The estimator of Chambers and Dunstan (1986) is used for model based estimation of the cumulative distribution function (CDF) and other parameters for acid neutralizing capacity (ANC) for the universe of lakes in region 1A (Adirondack Mountains) of the Eastern Lake Survey (USEPA). The data consist of a simple random sample from each of three strata (alkalinity classes) with different sampling probabilities among strata. Design based methods provide one estimate of the CDF using the inclusion probabilities. With the existence of covariate information on the entire population (sample and nonsample) model based estimation of parameters of the population is also possible. This example has been investigated in Jager and Overton (1991); this document represents an extension and refinement of their results, and additional work on the spatial analysis of the residuals from regression models under stratified sampling. The lakes data set consists of a sample of 155 lakes (simple random samples of 57, 51 47 from each of the three strata respectively) taken from a frame population of 1293 lakes (548, 430 and 315 in each stratum). Acid Neutralizing Capacity (ANC) was measured on the sample lakes from field sampling and elevation and rainfall pH were obtained for all lakes in the frame using Geographic Information System (GIS) data. The pH variable is a planar interpolated surface from rain gauge data. The basic data set was provided by Henrietta Jager; the explanatory variables were not includeed in the original EPA data set.

## 4.2 Model Based Estimation.

Interest is in estimation of the finite population distribution function (CDF) and associated parameters (population mean, population standard deviation) for the ANC of the lakes: $F(t) = \frac{1}{N}[\sum_{U} I(y_u \leq t)]$. Where I(B) is the indicator function taking on the value 1 if B is true, 0 if B is false. Since we are interested in the finite universe of lakes, the CDF will be estimated for a finite set of units in the population, where the number of units, N, is the known lake population size.

The model based estimator that will be used here is modified from Chambers and Dunstan (1986):

$$F(t) = \frac{1}{N}\left[\sum_{u\epsilon S} I(y_u \leq t) + \sum_{u\epsilon U-S} \Phi^{-1}((t-\hat{y}_u)/\hat{y}_u\hat{\sigma}) \right. \qquad (1)$$

I(B) is the indicator function of the set B, U-S is the set of unit labels in the nonsample and $\hat{y}_u$ is the ordinary least squares (OLS) prediction for unit $u$. This estimator has been shown to provide considerable gains in efficiency and reduction in bias over the design based estimator when the model is fit correctly (Chambers and Dunstan 1986, Dorfman 1993). This estimator is a model based estimator due to the use of predicted values $\hat{y}_u$, an estimated standard deviation $\hat{\sigma}$ from a statistical model for the data, and a n assumed distribution for the residuals, in this case the normal distribution. We will consider two statistical models for the sample data. The following regression model for the ANC values in each stratum:

$$y_u = \mu + x_u\beta + \epsilon_u, \qquad (2)$$
$$E(\epsilon_u) = 0,$$
$$Var(\epsilon_u) = E(y_u)\ \sigma^2,$$
$$Cov(\epsilon_u,\epsilon_{u'}) = 0,\ \ u \neq u',$$

and a linear model of which (2) is a special case:

$$y_u = \mu + x_u\beta + \epsilon_u,$$
$$E(\epsilon_u) = 0,$$
$$Var(\epsilon_u) = E(y_u)\ \sigma^2 \qquad (3)$$
$$Cov(\epsilon_u,\epsilon_{u'}) = \sigma(d_{uu'};\ \theta)\ \ u \neq u'.$$

where $x_u$ is a matrix of covariates and spatial coordinates, $\sigma(.\ ;\ .)$ is a spatial covariance function of a vector of parameters $\theta$ and the distances between the

spatial locations of the population units $d_{uu'}$. The residual covariance in (3) can be interpreted as a real part of the process that generated the realization of the finite population (interpreting the data as a realization of a random field) , or can be used as a surrogate for response surface explanatory variables that were not measured in the survey (Jager and Overton 1991). It is this second interpretation that will be used here adding this additional structure to the model may change the CDF estimate by treating missing covariates as heterogeneous variance structure, induced residual spatial correlation, or both. This will be investigated in this case study.

For the analysis of each stratum separately, the Chambers and Dunstan (CD) estimator will be the same as in equation (1) but the predictions will be OLS predictions for population model (2) $\widehat{y}_u = x_0(x'x)^{-1}x'y$ and 'spatial' predictions using an estimated spatial covariance matrix under population model (3) $\widehat{y}_u = x_0(x'\widehat{\Sigma}^{-1}x)^{-1}x'\widehat{\Sigma}^{-1}y + \widehat{\Sigma}_0\widehat{\Sigma}^{-1}(y - x(x'\widehat{\Sigma}^{-1}x)^{-1}x'\widehat{\Sigma}^{-1}y )$, $\widehat{\Sigma}_0$ is the transpose of the estimated residual spatial covariance between the sample and nonsample locations, $\widehat{\Sigma}$ is the estimated residual spatial covariance between the sample locations and $x_0$ is the matrix of covariates and the spatial coordinates for the nonsample units. Because the design is stratified, a further modification of the analysis of the $\epsilon$'s is to analyze the residuals from all three strata at the same time. This is particularly of interest when the unexplained spatial pattern of the population is independent of strata. Combining the strata for this purpose increases the sample size used for spatial covariance estimation by allowing spatially adjacent residuals that are on opposite sides of a stratum boundary to be included in a common spatial analysis.

Stratum differences in the residuals need to be accounted for before proceeding. To do this, the response model will be fit separately for each stratum and the set of residuals will be adjusted for their different stratum variance so that a single spatial linear model can be estimated for the entire population. There are three options, depending upon which model is the most feasible for the stratum variances: (1) No scaling, this is equivalent to assuming that the variance functions are all equal, or that the stratification is only correlated with the first moments of Y, (2) Scaling by a separate $\widehat{\sigma}$ for each stratum. This assumes that the residuals are homoscedastic, and that the strata have different residual variances, (3) scaling by a function of the fitted value $g(\widehat{y})$ and (4)

Scaling by a separate, heterogeneous standard deviation function $\hat{\sigma}g(\hat{y})$ for each stratum, this assumes that the variances are heteroscedastic and different for each stratum. Heterogeneous scaling (4) was chosen for these data. This decision was based upon examination of residual plots, which indicated two things; (1) large residuals are associated with large values of ANC and (2) the variances of the three strata are different, especially for stratum 3. From results in chapter 2, this form of scaling sets aside less of the residual variability than form (2) and for these highly patterned residuals this is desired. Each residual is scaled by its estimated stratum specific standard deviation function giving a set of homogeneous residuals for the entire sample:

$$\tilde{\epsilon}_{uh} = \hat{\epsilon}_{uh}/\hat{\sigma}_h g(\hat{y}_{uh}) \quad (4)$$

This set of scaled residuals is then analyzed for spatial covariance (actually spatial correlation because they have been scaled by their standard deviations). If significant spatial covariance is found, the spatial covariance matrix is used to make predicted residuals for the nonsample units using an ordinary kriging predictor:

$$\overset{*}{\epsilon}_{uh} = \hat{\Sigma}_0 \hat{\Sigma}^{-1} \tilde{\epsilon}_{uh} \quad (5)$$

Where $\hat{\Sigma}_0$ is the transpose of the estimated residual spatial covariance between the sample and nonsample locations and $\hat{\Sigma}$ is the estimated residual spatial covariance between the sample locations.

These predicted residuals on the nonsample are then rescaled according to the estimated standard deviation functions used in (4), and added to the stratum specific OLS predictions. These predictions are then used for the nonsample units in the model based CDF estimator (1).

$$\epsilon'_{uh} = \hat{\sigma}_h \sqrt{\hat{y}_{uh}}\, \epsilon^*_{uh}$$

$$\tilde{y}_{uh} = \hat{y}_{uh} + \epsilon'_{uh}.$$

Given a set of residuals and predictions for the sample units, the parameter $\gamma$, in the variance function $g(E(y_u);\gamma) = \hat{y}_u^{\gamma}$ is estimated using the methodology introduced in the introduction:

$$\ln(\widehat{\epsilon}_u^2) = \text{a} + \text{b}\ \ln(\widehat{\text{y}}_u) + \text{c},$$

$\widehat{\gamma}$ is then estimated using the slope of this line, $\widehat{\text{b}}$.

In particular, for the data discussed in the introduction, two regression models were fit to the sample values in each stratum:

(1) ANC=$\beta_0$ + $\beta_1$elevation + $\beta_2$pH + $\epsilon$

(2) ANC=$\beta_0$ + $\beta_1$elevation + $\beta_2$pH + $\beta_3$lat + $\beta_4$long + $\epsilon$

Here the latitude and longitude variables have been transformed to accurately represent geographic distance. Strata were analyzed separately with each of the two models, and each model was fit using two structures for the residuals, the spatially independent version in (2) above and the spatially correlated version in (3). Predictions with spatially independent residuals will be termed ols predictions, predictions using spatially correlated residuals will be called uk predictions (based on the commonly used term 'universal kriging' for this model). The CDFs were estimated using the Chambers and Dunstan estimator with both a homogeneous and heterogeneous standard deviation function as in equation (1).

A further analysis was made by fitting regression model (A) for each stratum, and then estimating spatial covariance for the residuals from all of the strata together. For this modification of the spatial analysis, the residuals from all three strata were scaled by their standard deviation functions before spatial covariance estimation. Predicted residuals were made for the nonsample lakes in each stratum and these predicted residuals (5) were added to the OLS predictions after rescaling by the same standard deviation function used in (4). The CDFs were estimated using predicted values $\widehat{y}_u$ and estimated standard deviations $\widehat{\sigma}$ in the CD estimator.

## 4.3 Results: Fitted Models.

For stratum 1 the fitted equations are:

**Model 1**

ANC = −5157 −0.49∗elevation + 1270∗pH

$$\widehat{\gamma} = 0.694$$

ANC = −2675 −0.561∗elevation + 731∗pH

$$\begin{aligned} \mathrm{Cov}(\epsilon_u,\epsilon_{u'}) &= 2043.9 + 4591.5 \qquad d_{uu'} = 0 \\ &= 4591.5{*}\exp(0.0194{*}\ d_{uu'}\ ) \quad d_{uu'} > 0 \end{aligned}$$

$$\widehat{\gamma} = 0.338$$

**Model 2**

ANC = −2017 −0.43∗elevation + 251∗pH +0.899∗lat +0.296∗long

spatial covariance was estimated to be 0.

$$\widehat{\gamma} = 0.699$$

For stratum 2 the fitted equations are:

**Model 1**

ANC = −7335.84 −0.654∗elevation +1782.9∗pH

$$\widehat{\gamma} = .8409$$

ANC = −5936.5 −0.538∗elevation +14591.1∗pH

$$\begin{aligned} \mathrm{Cov}(\epsilon_u,\epsilon_{u'}) &= 6727.5 + 7744.8 \qquad d_{uu'} = 0 \\ &= 7744.8{*}\exp(0.0199{*}\ d_{uu'}\ ) \quad d_{uu'} > 0 \end{aligned}$$

$$\widehat{\gamma} = 1.05$$

**Model 2**

ANC = −1215.96 −0.5614∗elevation + 65.32∗pH +1.247∗lat +0.3194∗long

$$\widehat{\gamma} = 1.06$$

ANC = −1270 −0.548∗elevation +13.84∗pH +1.22∗lat +0.377∗long

$$\begin{aligned} \mathrm{Cov}(\epsilon_u,\epsilon_{u'}) &= 8814 + 7888.3 \qquad d_{uu'} = 0 \\ &= 7888.3{*}\exp(0.004{*}\ d_{uu'}\ ) \quad d_{uu'} > 0 \end{aligned}$$

$$\widehat{\gamma} = .979$$

For stratum 3 the fitted equations are:

**Model 1**

$$\text{ANC} = 17630 - 1.57*\text{elevation} - 3833*\text{pH}$$

$$\widehat{\gamma} = 1.06$$

$$\text{ANC} = 11993 - 1.58*\text{elevation} - 2467.7*\text{pH}$$

$$\begin{aligned} \text{Cov}(\epsilon_u, \epsilon_{u'}) &= 52158.9 + 2488080 \qquad d_{uu'} = 0 \\ &= 2488080*\exp(0.005* d_{uu'}) \quad d_{uu'} > 0 \end{aligned}$$

$$\widehat{\gamma} = 1.45$$

**Model 2**

$$\text{ANC} = 15650 - 1.71*\text{elevation} - 940.4*\text{pH} - 1.91*\text{lat} - 2.36*\text{long}$$

$$\widehat{\gamma} = 1.09$$

$$\text{ANC} = 39681 - 1.61*\text{elevation} - 355*\text{pH} - 0.48*\text{lat} - 8.13*\text{long}$$

$$\begin{aligned} \text{Cov}(\epsilon_u, \epsilon_{u'}) &= 49592 + 2186442 \qquad d_{uu'} = 0 \\ &= 2186442*\exp(0.006* d_{uu'}) \quad d_{uu'} > 0 \end{aligned}$$

$$\widehat{\gamma} = 1.24$$

Additionally, residuals from the stratum specific regression models (for model 1) were analyzed as a set for spatial covariance using the exponential covariance model. This model was fit via REML for the three types of scaling discussed above. Table 4.1 gives the estimated parameters.

Table 4.1 Estimated Covariance Parameters for the Different Forms of Scaling.

| | $\theta_o$ | $\theta_1$ | $\theta_2$ |
|---|---|---|---|
| No Scaling | NO CONVERGENCE | | |
| Homogeneous Scaling | 0.0062 | 0.9661 | 0.5527 |
| Heterogeneo· s Scaling | 0.0 | 1.0042 | 0.3227 |

We see that the form of scaling primarily affects $\theta_2$, which describes the rate at which the covariance decreases with spatial distance. For heterogeneous scaling, this rate is more gradual than for homogeneous scaling, indicating an increase in correlation range between the heterogeneous scaled residuals relative to the homogeneously scaled ones.

The estimates of $\gamma$ range between 0.3 - 1.5, which indicates that the standard deviation function $\sqrt{g(E(y_u);\gamma)}$ ranges between 0.15 and 0.75. Stratum 1 has the smallest estimated values for $\gamma$, and stratum 3 the largest.

## 4.4 Results: Residual Analysis.

The residual plots for the individual strata reveal differences in the variance-covariance between strata. Figures 4.1-4.3 show, plots for model (1) in the left column and model (2) on the right. The top plot is of the residuals versus the fitted values and the bottom plot is the semi-variogram. The semi-variogram, $\gamma(d)$, is related to the covariance as $\frac{1}{2}\gamma(d)=C(0) - C(d)$ and is a useful residual diagnostic tool. The empirical semi-variogram is estimated as:

$$\frac{1}{2N_d} \sum_{\{uu':\, d_{uu'}=d\}} (\epsilon_u - \epsilon_{u'})^2 \qquad (9)$$

Where $N_d$ is the number of pairs of residuals separated by a distance of d units (kilometers for the data analyzed here) and the sum is over the set of all pairs of residuals that are d units apart. Notable features of the semi-variogram are the value of the semi-variance at the point where the fitted curve flattens out (called the sill in geostatistical terminology). This is a measure of the variance of the residuals. The separation distance at which the fitted semi-variogram flattens out (often called the range) is a measure of the distance beyond which the residuals are uncorrelated. Examination of the residual plots indicates that strata 1 and 2 appear to have variances increasing with the fitted values. Stratum 3 is less clear but appears to have two large positive residuals and variance increasing with the fitted value. The semi-variograms reveal that the residual variances are different between strata, with sills of approximately 3500, 9500 and 450000 for strata 1, 2, 3

respectively. The correlation ranges are similar for strata 1 and 2, about 50 Km, but for stratum 3 is appears to be much greater, about 200 Km. These plots motivated the fitting of the spatial versions of the two models for each of the strata and using the specific form of the Chambers and Dunstan estimator (8).

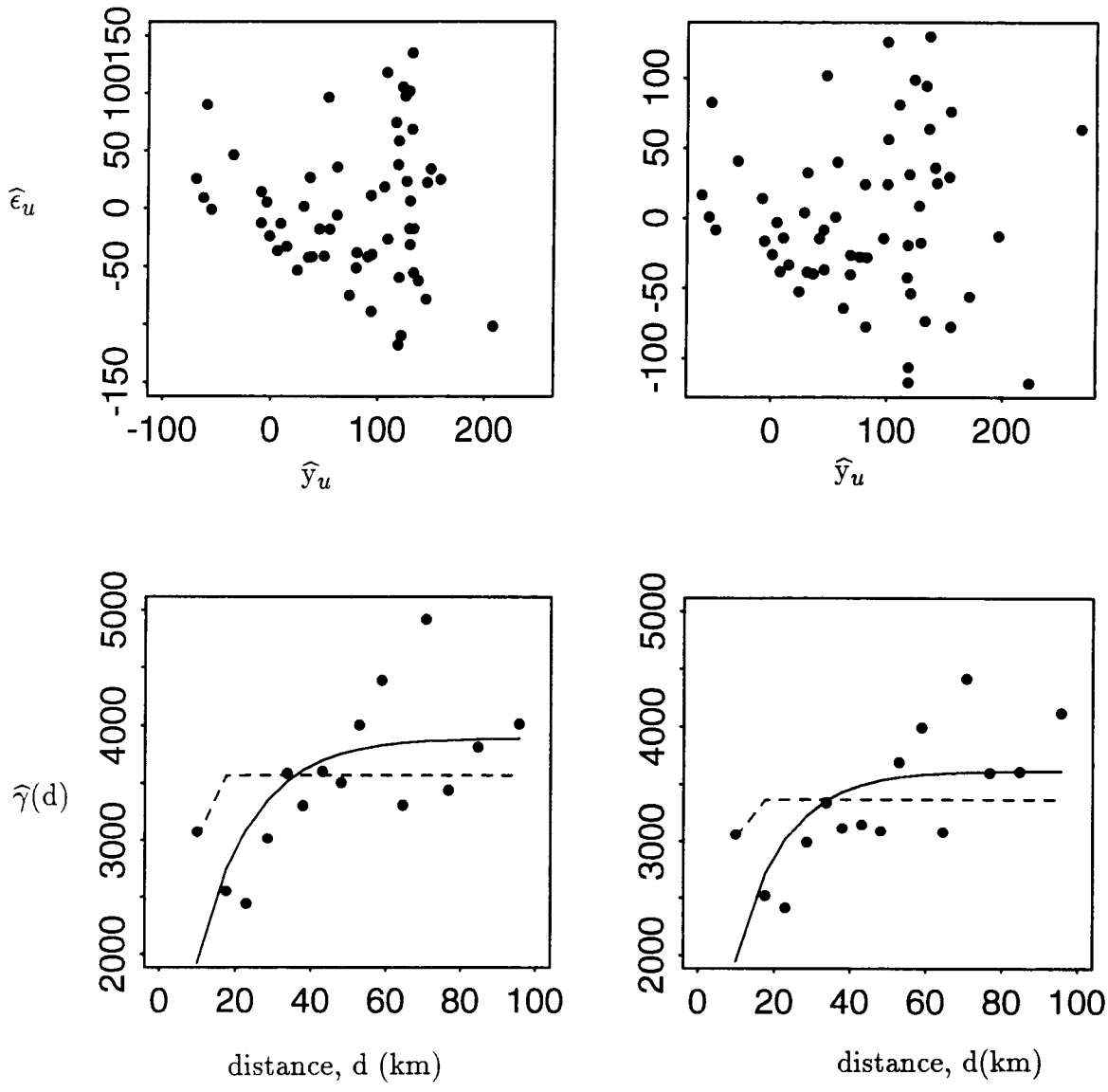


Figure 4.1 Residual Plots for Stratum 1. Model 1 (left) and model 2 (right). Estimated residuals versus fitted values (top) and semi-variograms with two fitted models
(bottom).

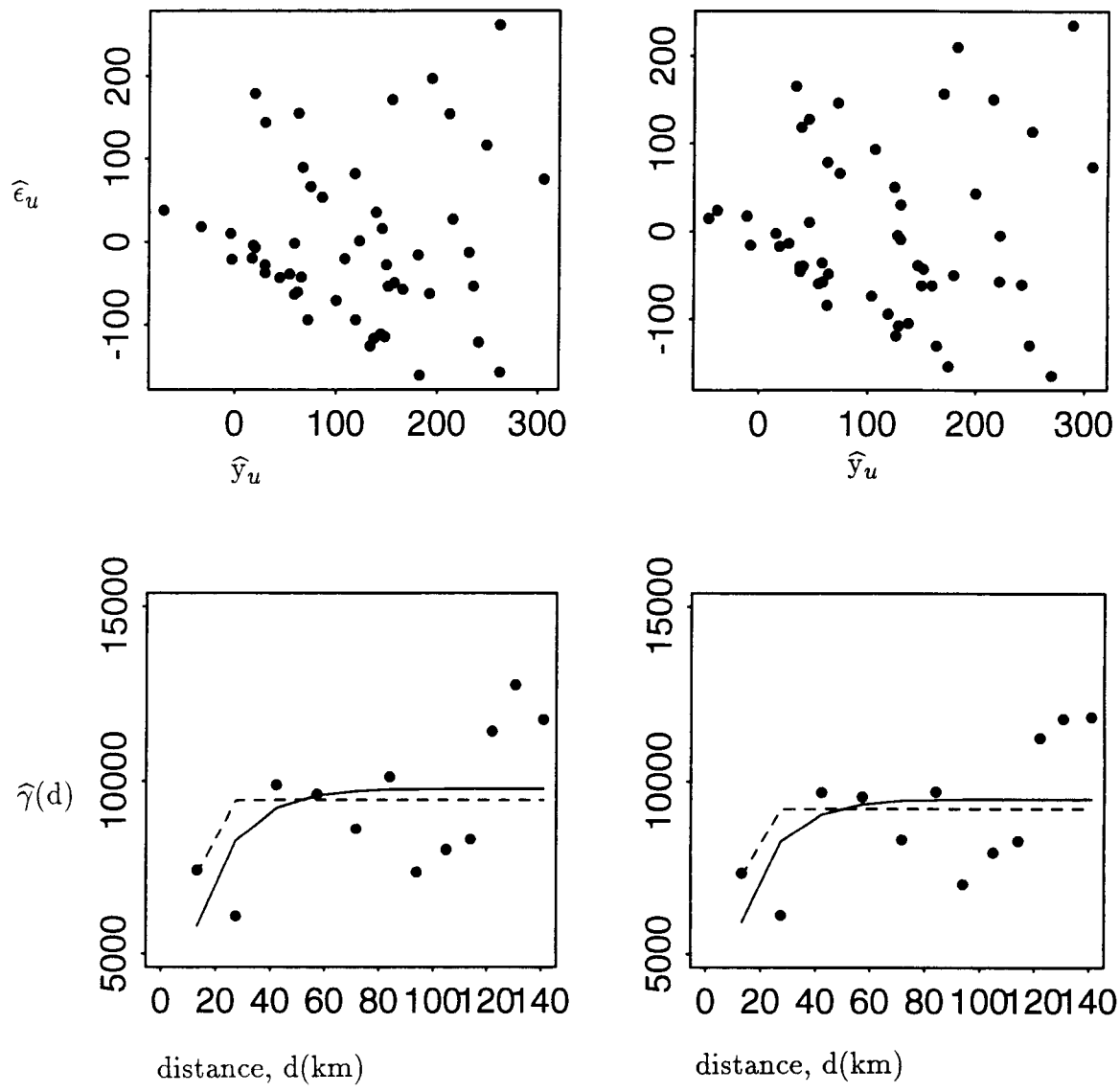


Figure 4.2 Residual Plots for Stratum 2. Model 1 (left) and model 2 (right). Estimated residuals versus fitted values (top) and semi-variograms with two fitted models
(bottom).

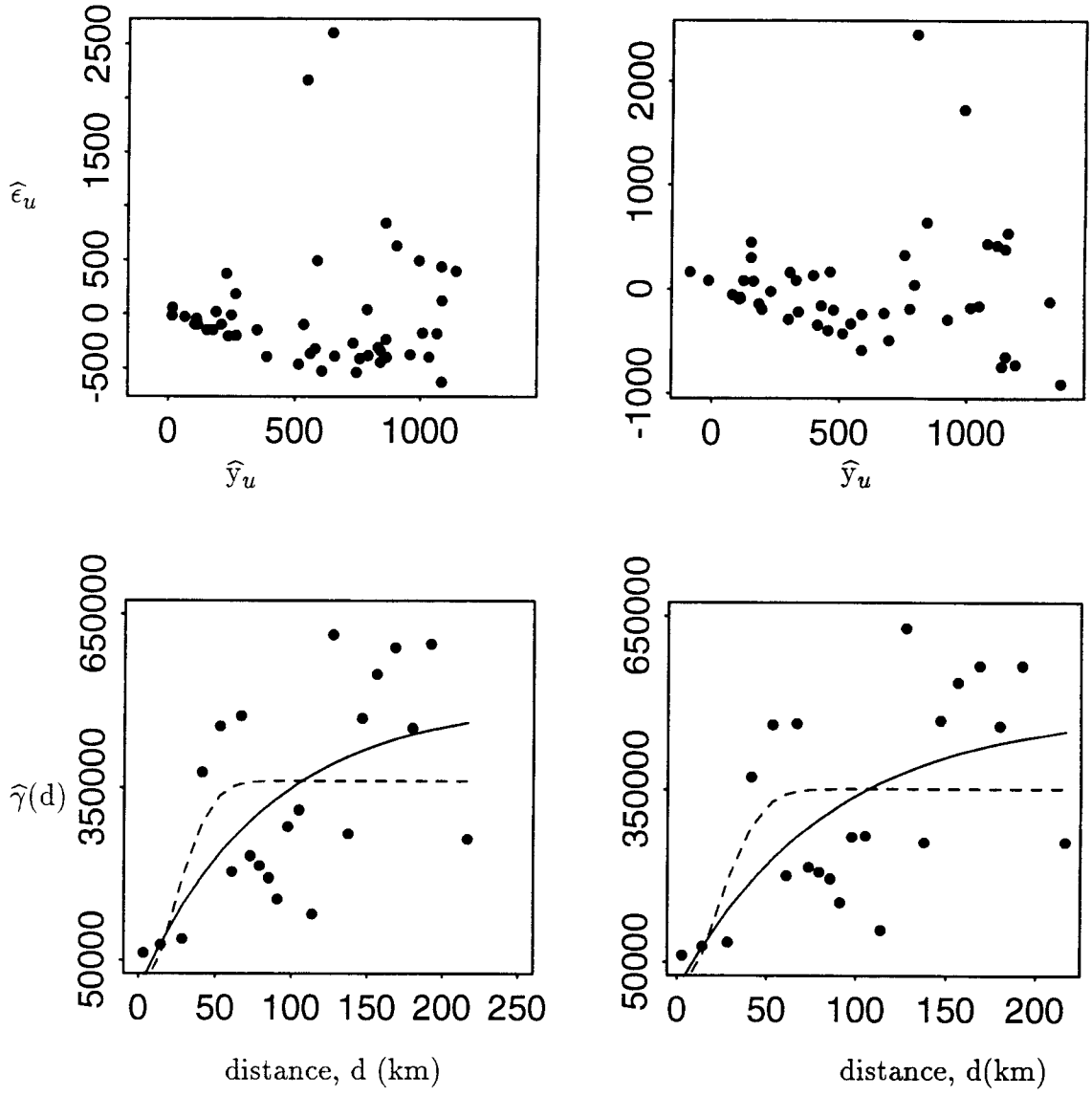


Figure 4.3 Residual Plots for Stratum 3. Model 1 (left) and model 2 (right). Estimated residuals versus fitted values (top) and semi-variograms with two fitted models
(bottom).

Figure 4.4 shows plots of the model 1 residuals from all three strata combined into one plot. The left column shows the residuals plotted versus the fitted values (left), residuals divided by the estimated stratum standard deviation $\widehat{\sigma}_u$ (middle), and divided by the estimated stratum standard deviation function $\widehat{\sigma}_u\sqrt{\widehat{y}_u}$ as in (6) (bottom). The right column shows the empirical semi-variogram and two fitted semi-variogram models. The unscaled residuals have some evidence of heterogeneous variance, but no evidence of spatial correlation. Scaling by the stratum standard deviation drastically changes the shape of the residual plot and the semi-variogram. There is evidence of spatial correlation with a correlation range of 45 or 100 Km, depending upon which variogram model is used. Scaling by the heterogeneous variance function markedly reduces the correlation range of the residual semi-variograms, to between 5 and 20 Km depending upon which model is used. Figure 4.5 is a similar plot for model 2. The residual assessment is similar to that of model 1.

## 4.5 Design Effect.

Another issue that warrants investigation is the influence of the variable probabilities of selection on the covariance function (or alternately, the variogram) estimate (9) for the entire set of residuals. The sample set of lakes is not an SRS from the population of lakes due to the differing probabilities of selection for each of the strata. The pairwise inclusion probabilities will not be the same for all pairs of lakes in the same distance class set {ij: $h_{ij}$=h}. In particular, the pairwise inclusion probabilities for a pair in the same stratum, say stratum h will be:

$\pi_{uu'} = \dfrac{n_h(n_h-1)}{N_h(N_h-1)}$, and for pairs in separate strata:

$\pi_{uu'} = \dfrac{n_h n_{h'}}{N_h N_{h'}}$.

Setting $w_{uu'} = 1/\pi_{uu'}$, an estimator of the empirical covariance function will be:

$$\frac{1}{\sum w_{uu'}} \sum w_{uu'}\, \epsilon_u^* \, \epsilon_{u'} \qquad (10)$$

where both summations are over the set: $\{u,u'\colon d_{uu'} = d\}$. The exponential and the gaussian semi-variogram models are fit to the estimator (9) and the weighted estimator (10). The models are: $\sigma(d;\theta) = \theta_0 * \exp(-\theta_1 d)$ and $\sigma(d;\theta) = \theta_0 * \exp(-\theta_1 d^2)$. The estimated coefficients are given in table 4.2.

Table 4.2 Estimated Semi-variogram Parameters using the two Empirical Estimators.

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| | $\theta_0$ | $\theta_1$ | $\theta_0$ | $\theta_1$ |
| (9) | 1.29 | .0196 | 1.02 | .00228 |
| (10) | 1.31 | .0189 | 1.02 | .00224 |

In this case, adjusting for the variable probabilities of selection made little difference in the fitted functions and this would have little affect on the predictions, and the variable probabilities are ignored here. This may not be the case in general and if the inclusion probabilities for a data set are known then this comparison should be made. In general, when the inclusion probabilities are known they should be used in the covariance estimator simply because the additional information about the sample may lead to better model based estimates of finite population parameters. In this case the differences were very subtle, because the sampling rates were not distinctly different in the three strata.

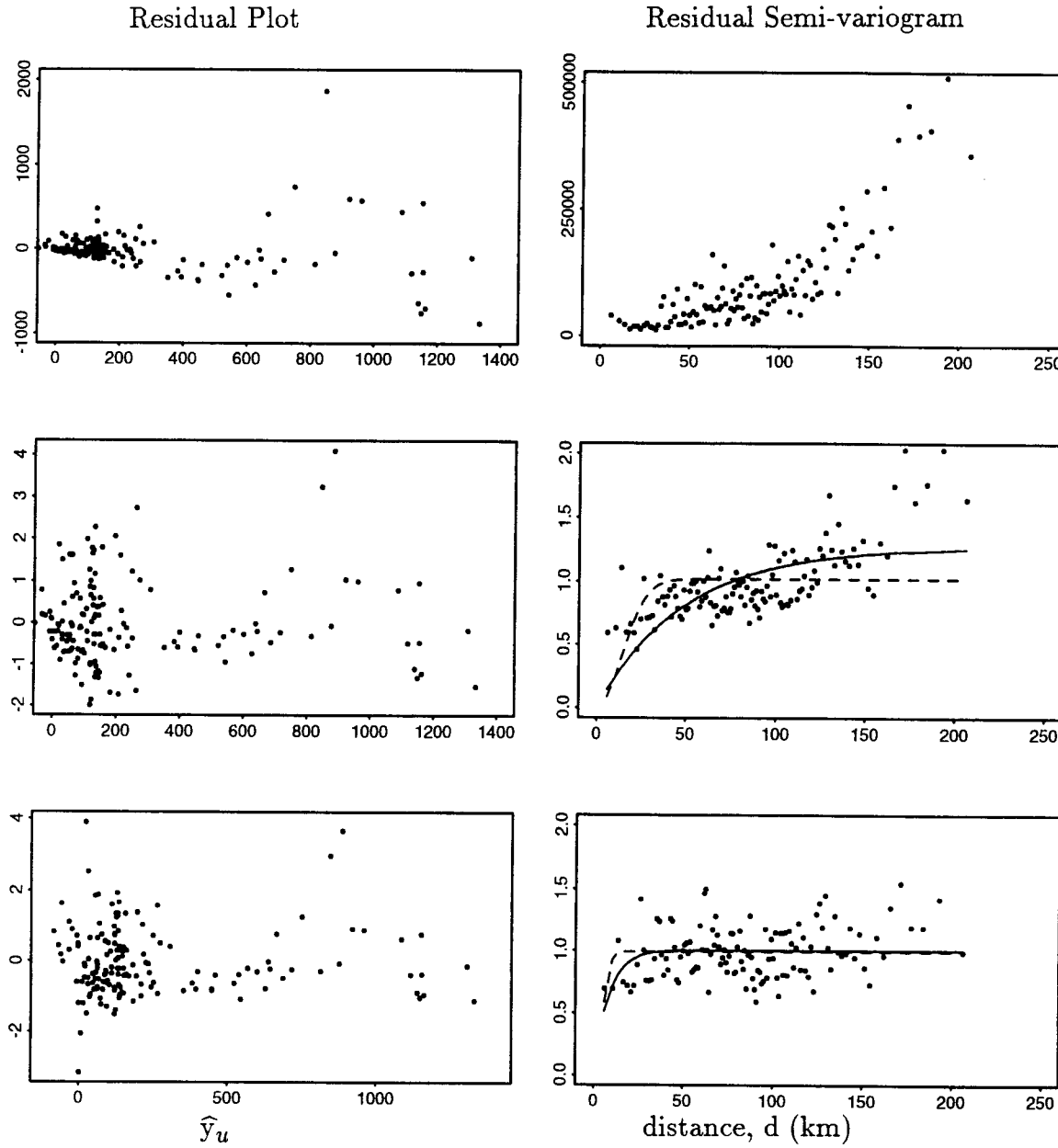


Figure 4.4 Residual Plots for Model 1: All Strata Combined. The left column shows estimated residuals versus fitted values and the right column is the semi-variogram with two estimated models. Rows in the plots using residuals with no scaling (top), homogeneous scaling (middle), and heterogeneous scaling (bottom).

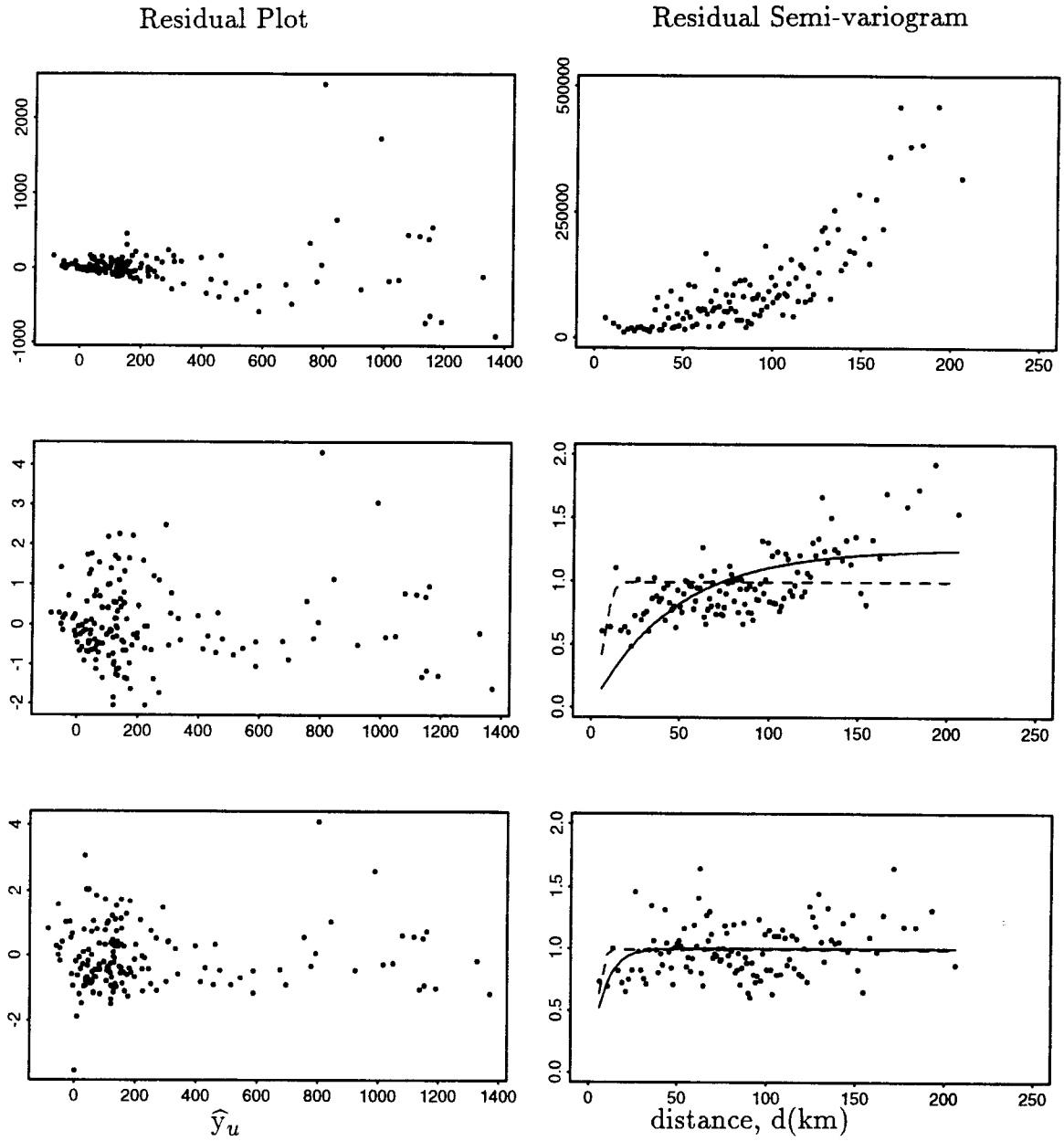


Figure 4.5 Residual Plots for Model 2: All Strata Combined. The left column shows estimated residuals versus fitted values and the right column is the semi-variogram with two estimated models. Rows in the plots using residuals with no scaling (top), homogeneous scaling (middle), and heterogeneous scaling (bottom).

## 4.6 Results: Parameter Estimation.

The Chambers and Dunstan estimator was used to estimate the stratum CDFs using models 1 and 2, spatial and nonspatial versions, and the spatial version of model 1 where the entire population of residuals was used to make predicted residuals for each stratum. Figures 4.6 - 4.7 show estimates using model based predictions from model 1. The figures contain two plots apiece, each plot shows four estimates of the CDF, the design based estimator (step function solid line), the naïve model based estimator (smooth solid line) and two Chambers and Dunstan estimators, one with a homogeneous variance function and the other heterogeneous (dashed lines). All three model based estimators in a plot use the same model based predictions, $\widehat{y}_{uh}$. For the top plot, the model based predictions are from regression model (1). For the bottom plot, the model based predictions are from the version of (1) with spatially correlated residuals with the covariance estimated separately for each stratum. In addition, estimates of the stratum means and standard deviations from each estimator are given in table 4.3. Figure 4.6 shows the estimated CDFs for stratum 1. The bias correction of the CD estimators over the naïve estimator is apparent, especially in the upper quartile. The naïve model based estimators between the two plots are distinct, reflecting the influence of the 3 different prediction models for the nonsample units. The CD estimator using uk predictions where the residual covariance is estimated within the stratum is the most distinct. This is reflected in the estimated mean for the stratum, which is largest for this model (table 4.3). For stratum 2 the bias correction of the CD estimators relative to the naïve model based estimator are most noticeable in lower half of the distribution (figure 4.7). The relative performance of the CD estimators for the mean and standard deviation are the same as for stratum 1. For stratum 3, the estimator using uk predictions with spatial covariance estimated within the stratum has heavier tails (figure 4.8). This is reflected in the large estimate of the population standard deviation for this model.

Figures 4.9-4.11 show the CD estimates using models 1 and 2 for Strata 1, 2 and 3 respectively. The design based estimator is included for comparison. For stratum 1, the model based estimates are all different from the design based

estimate. The model based estimates that use only the stratum specific analysis are parallel to each other (recall that the spatial and nonspatial versions of model 2 are identical here, because the spatial covariance estimated to be 0), with the biggest differences around the upper quartile.

For stratum 2, the model 2 estimates (both spatial and non spatial) were very similar to the design based estimate except in the lower tail (figure 4.9). The 2 estimators based upon model 1 using only stratum 2 information are quite different from these, and different from each other except in the upper tail.

For stratum 3, all model based estimators differ appropriately from the design based one. This is the only stratum where the design based estimator was steeper and to the left of the model based estimates. The estimators using within stratum analysis were parallel to each other until the 70th percentile, where the kriging versions of models 1 and 2 estimated longer flatter upper tails.

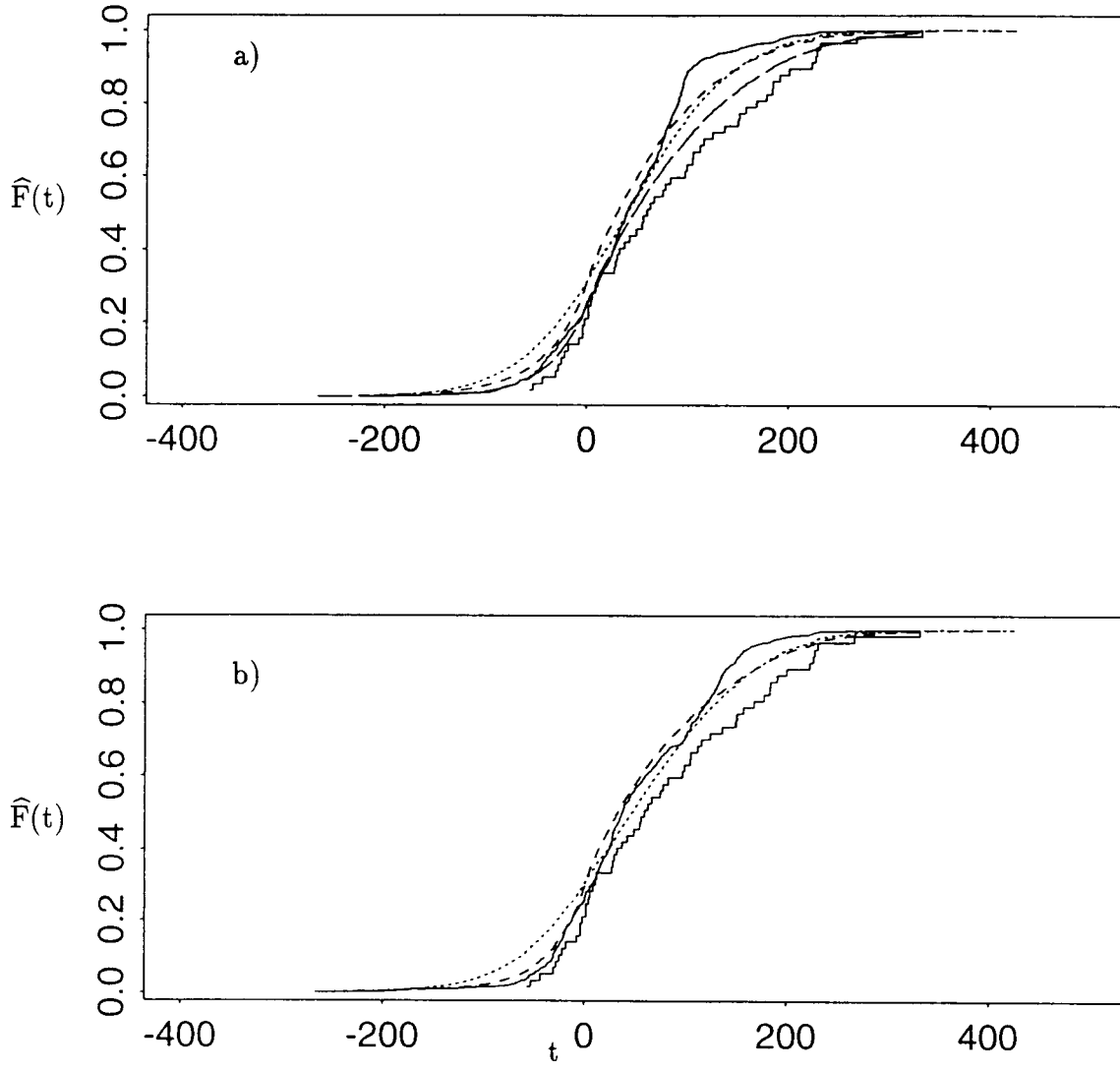


Figure 4.6 Estimated CDFs for Stratum 1. a) uses regression predictions from model 1, the elevation and pH model. b) uses kriging predictions from this model where the residual spatial covariance is estimated separately for each stratum. Lines are design based (step function solid line), naive estimator (smooth solid), and CD with homogeneous variance function (short dashed) and a heterogeneous variance function (long dashes).

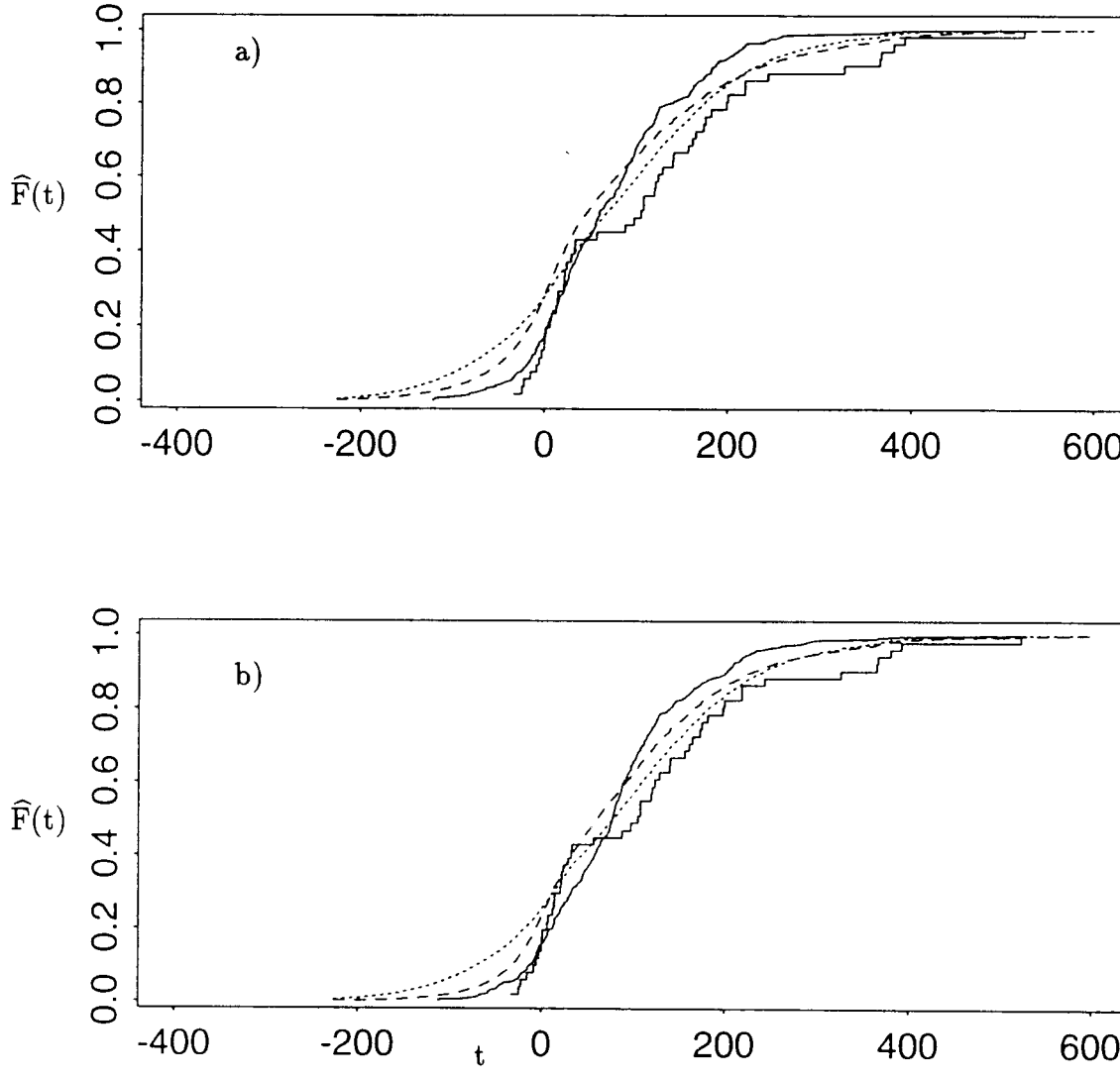


Figure 4.7 Estimated CDFs for Stratum 2. a) uses regression predictions from model 1, the elevatio ' and pH model. b) uses kriging predictions from model (1) where the residual spatial covariance is estimated separately for each stratum. Lines are design based (step function solid line), naïve estimator (smooth solid), and CD with homogeneous variance function (short dashes) and heterogeneous variance function (long dashes).

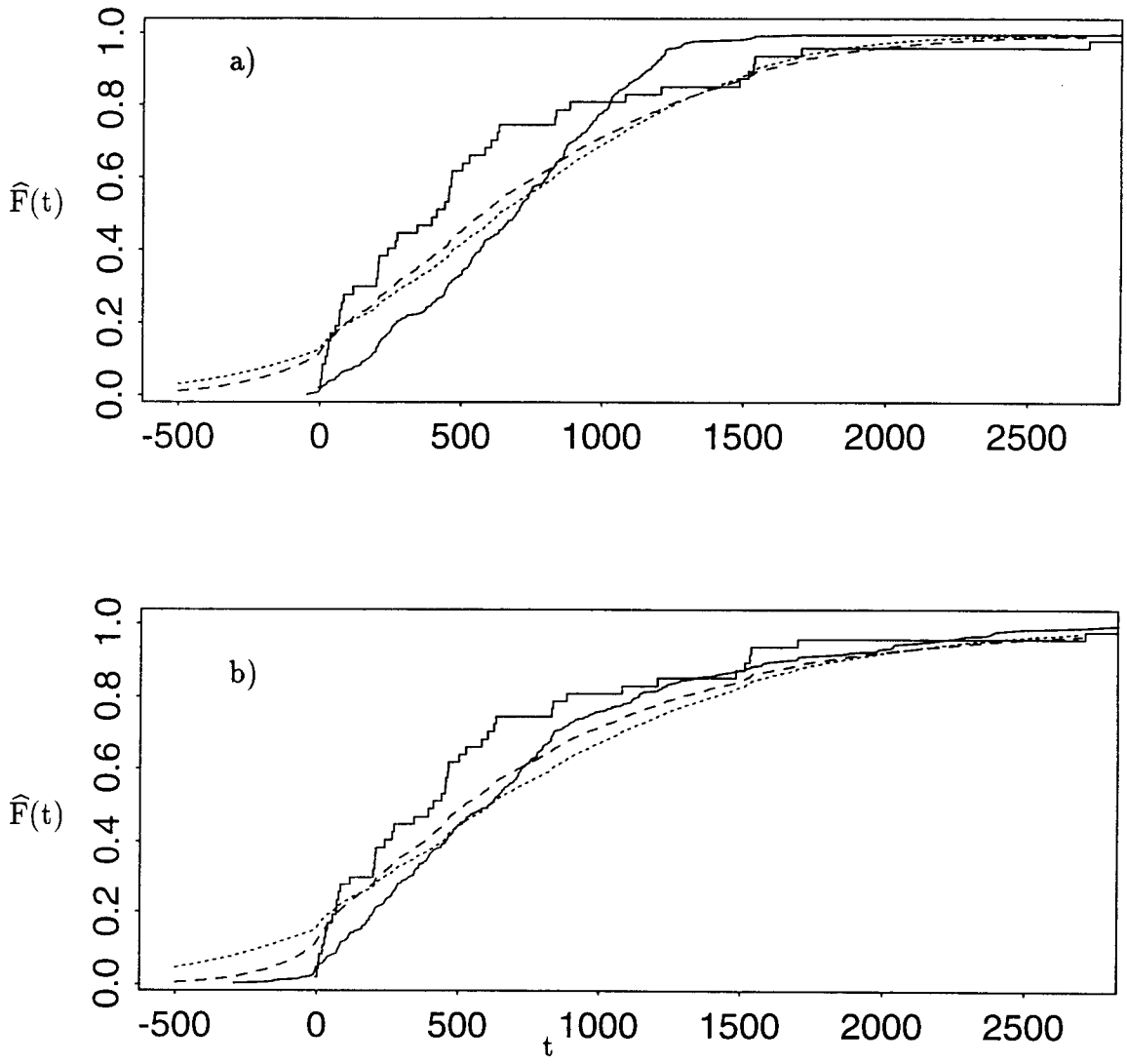


Figure 4.8 Estimated CDFs for Stratum 3. a) uses regression predictions from model 1, the elevation and pH model. b) uses kriging predictions from model (1) where the residual spatial covariance is estimated separately for each stratum. Lines are design based (step function solid line), naïve estimator (smooth solid), and CD with homogeneous variance function (short dashes) and heterogeneous variance function (long dashes).

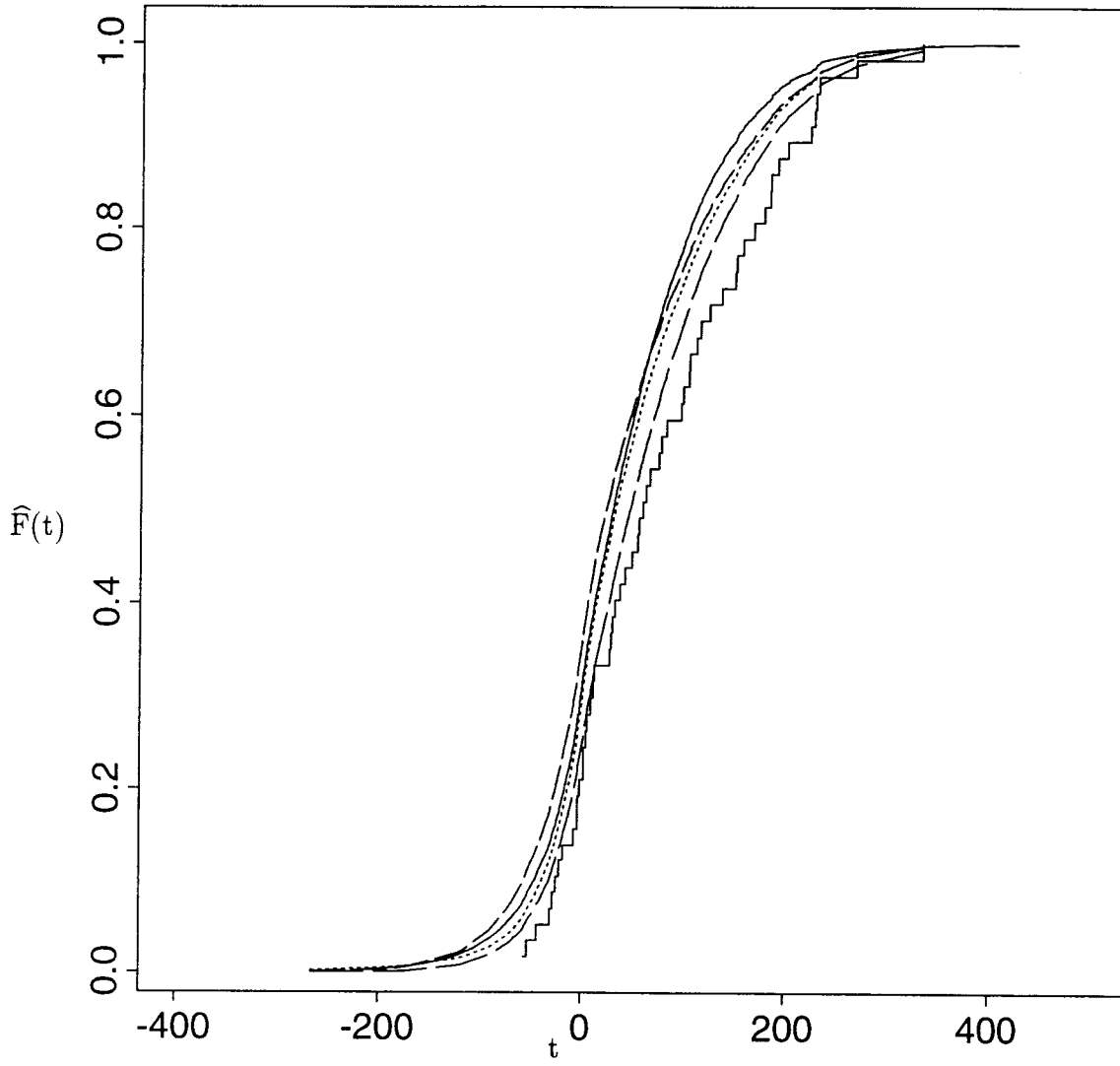


Figure 4.9 Estimated CDFs for Stratum 1. The solid line is the design based estimate. All other lines are CD estimates with heterogeneous variance functions but using predicted values from different models. The two short dashed lines use ols predictions and uk predictions from model 1 respectively, the medium dashed lines use ols and uk predictions from model 2, and the mixed dashed line uses uk predictions from model 1, but estimates combined stratum residual covariance.

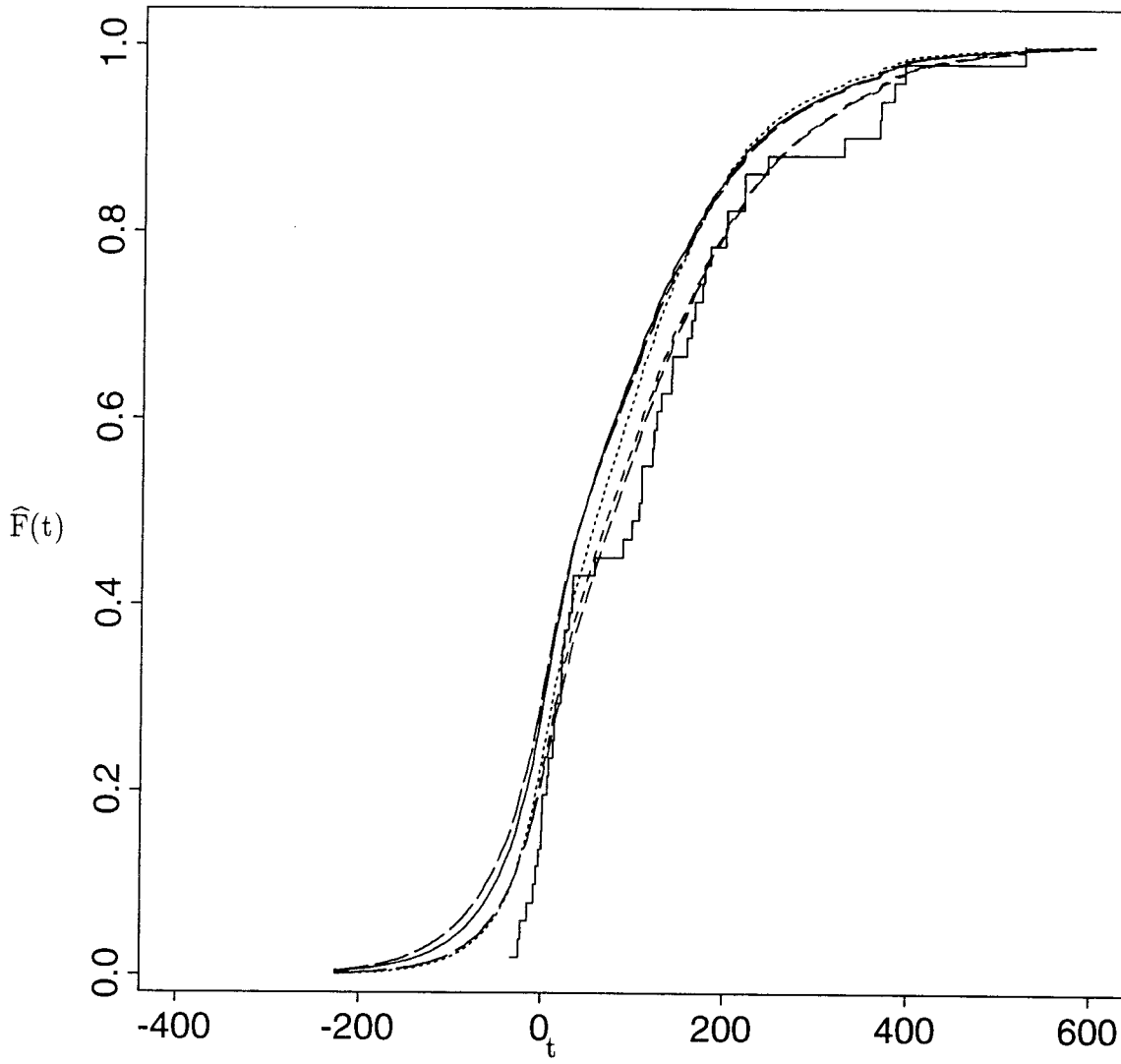


Figure 4.10 Estimated CDFs for Stratum 2. The solid line is the design based estimate. All other lines are CD estimates with heterogeneous variance functions but using predicted values from different models. The two short dashed lines use ols predictions and uk predictions from model 1 respectively, the medium dashed lines use ols and uk predictions from model 2, and the mixed dashed line uses uk predictions from model 1, but estimates combined stratum residual covariance.

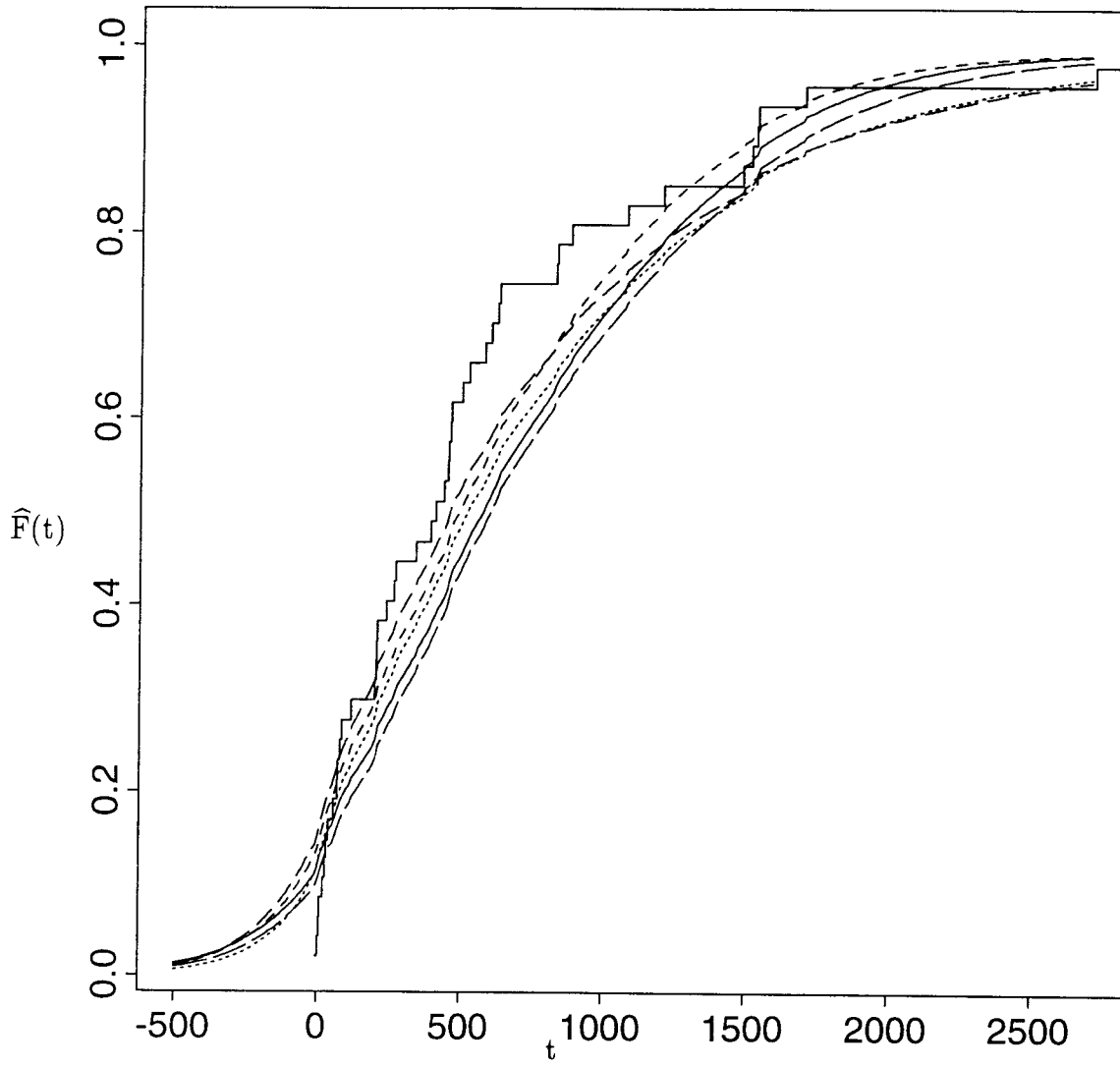


Figure 4.11 Estimated CDFs for Stratum 3. The solid line is the design based estimate. All other lines are CD estimates with heterogeneous variance functions but using predicted values from different models. The two short dashed lines use ols predictions and uk predictions from model 1 respectively, the medium dashed lines use ols and uk predictions from model 2, and the mixed dashed line uses uk predictions from model 1, but estimates combined stratum residual covariance.

Table 4.3 Estimates of Means and Standard Deviations for the three Strata.

| Stratum 1 | Mean | Standard Deviation | |
|---|---|---|---|
| design based | 78.9381 | 90.9386 | |
| naïve | 41.9305 | 63.9296 | |
| CD - het sd function | 41.7447 | 89.0370 | ols predictions |
| CD - hom sd function | 41.7206 | 92.3414 | |
| naïve | 50.0982 | 71.3797 | |
| CD - het sd function | 51.1911 | 84.6666 | uk within stratum |
| CD - hom sd function | 51.1236 | 91.3446 | |
| naïve | 41.2143 | 73.4937 | |
| CD - het sd function | 41.7447 | 89.0370 | uk combined strata |
| CD - hom sd function | 41.7206 | 92.6633 | |
| **Stratum 2** | **Mean** | **Standard Deviation** | |
| design based | 116.000 | 128.5978 | |
| naïve | 73.0358 | 82.8924 | |
| CD - het sd function | 74.3278 | 119.3598 | ols predictions |
| CD - hom sd function | 75.4274 | 122.6496 | |
| naïve | 82.7436 | 82.2808 | |
| CD - het sd function | 83.4756 | 111.7280 | uk within stratum |
| CD - hom sd function | 84.7303 | 123.0652 | |
| naïve | 71.8833 | 94.2002 | |
| CD - het sd function | 73.2386 | 123.1837 | uk combined strata |
| CD - hom sd function | 75.1058 | 130.1062 | |
| **Stratum 3** | **Mean** | **Standard Deviation** | |
| design based | 587.900 | 698.5401 | |
| naïve | 684.6976 | 399.0496 | |
| CD - het sd function | 668.0643 | 616.6783 | ols predictions |
| CD - hom sd function | 691.6280 | 611.8690 | |
| naïve | 734.8775 | 686.8112 | |
| CD - het sd function | 631.5018 | 648.7975 | uk within stratum |
| CD - hom sd functioı | 692.5509 | 651.1833 | |
| naïve | 739.3184 | 478.7105 | |
| CD - het sd function | 702.0567 | 638.4802 | uk combined strata |
| CD - hom sd function | 737.5866 | 640.5866 | |

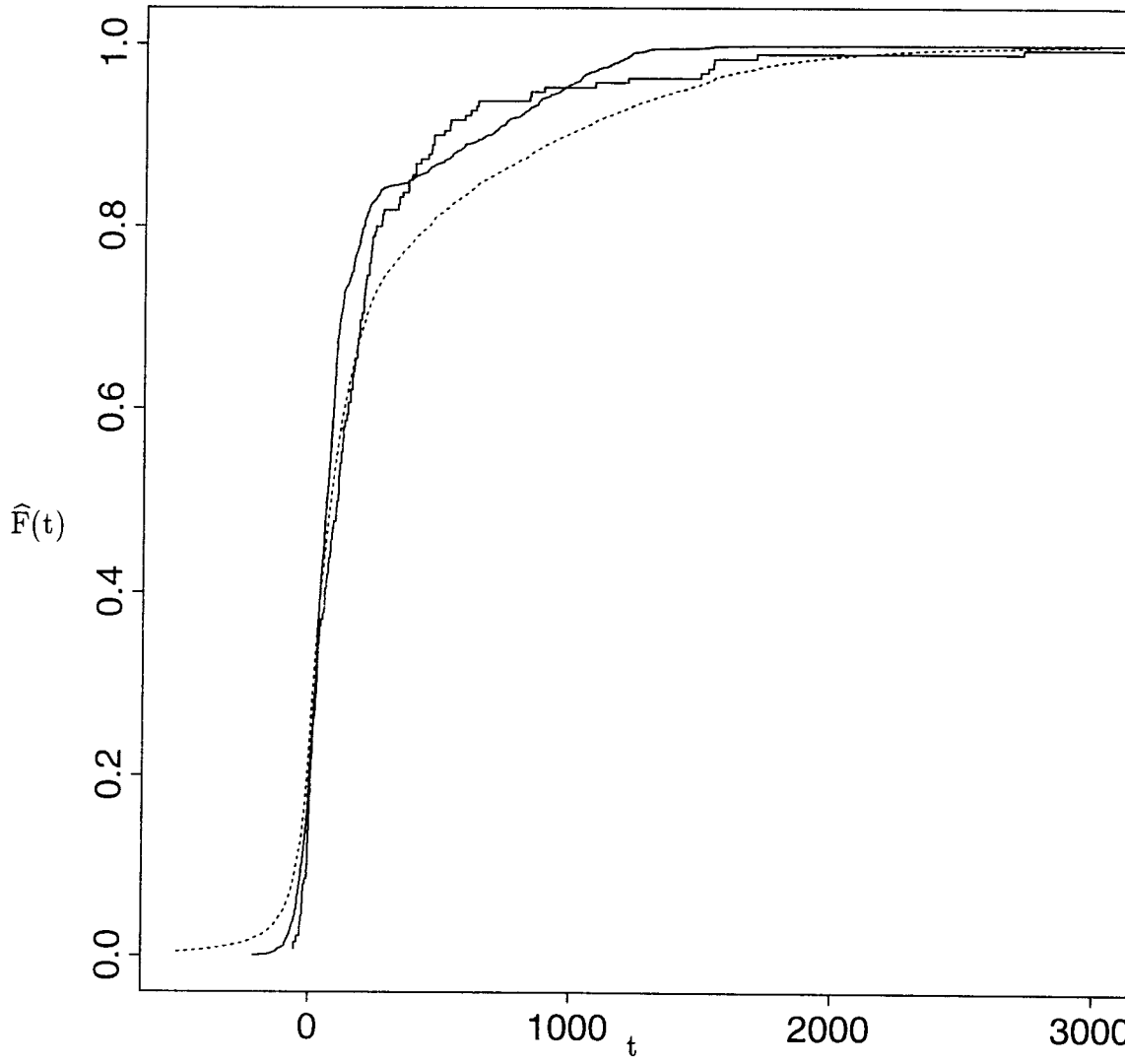


Figure 4.12 Estimated CDFs for the Entire Population. The solid step function is the design based estimate. The model based estimators use ols predictions. The solid line is the naïve estimator, the dashed line is the CD estimator with a heterogeneous variance function.

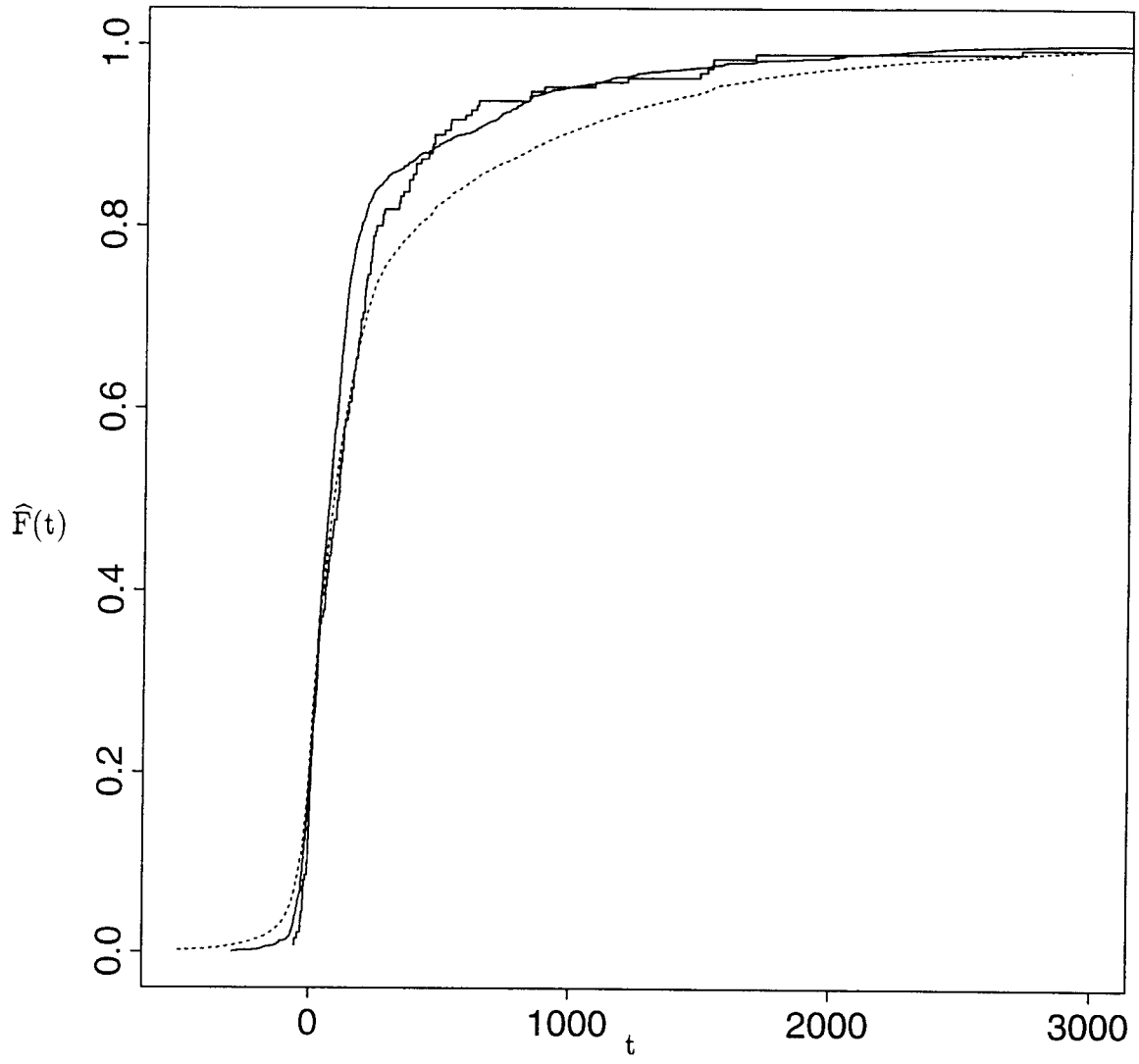


Figure 4.13 Estimated CDF for the Entire Population. The solid step function is the design based estimate. The model based estimators use kriging predictions. The covariance function is estimated separately for each stratum. The solid line is the naïve estimator, the dashed line is the CD estimator with a heterogeneous variance function.

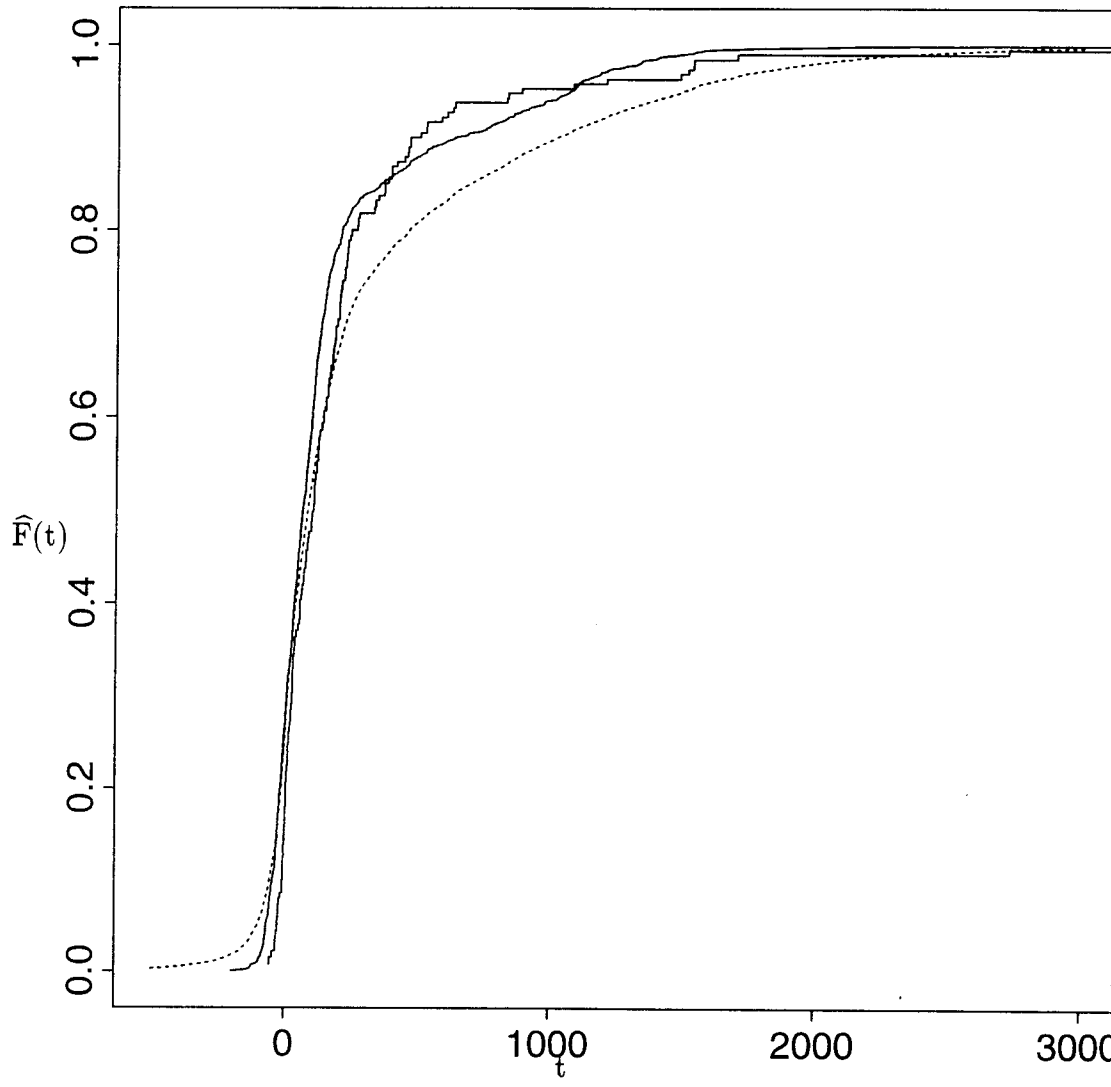


Figure 4.14 Estimated CDF for the Entire Population. The solid step function is the design based estimate. The model based estimators use kriging predictions. The covariance function is estimated using all of the residuals after being scaled by a heterogeneous variance function. The solid line is the naïve estimator, the dashed line is the CD estimator with a heterogeneous variance function.

## 4.7 Discussion.

For stratum 1, estimating residual correlation changes the distribution function estimates little. Adding the latitude and longitude coordinates to the regression model seems to account for most of the spatial pattern in residuals. Additionally, adding information from residuals from surrounding strata changed the estimate only slightly. This is probably due to stratum 1 being a single, convex shaped area of the geographic area. For strata 2 and 3, there is more of a difference between the estimates using residual covariance information within the strata versus using the residual covariance from all strata. This is due to the fragmented nature of these strata. Many of the sampled lakes that are geographically close to lakes in strata 2 and 3 are in different strata, using a residual analysis of all the stratum residuals combined has a greater affect on the shape of the estimated distribution.

# 5. Conclusions.

This research has focussed upon model based methods for estimating the parameters of spatial populations from probability samples. Design based methods provide unbiased, or at least consistent, estimates of population parameters using the properties of the sampling design. The model based procedures are used in cases where there is additional information that is known for the entire population, such as covariate information. The spatial populations considered here are interpreted to have spatial patterns in response to patterns in the values of causal variables. In practice, causal variables may be unknown, and known explanatory variables, though not directly causal themselves, can be used as surrogates for these unknown causal variables. Additionally, when the spatial coordinates are known for the entire population, a spatial model can be used as a surrogate for unknown spatially patterned causal variables. The explanatory variables are used in a regression model that estimates the relationship between the population values and the variables. The spatial information can be incorporated into the regression model in two ways, by including the spatial coordinates in the regression equation or by invoking a spatial covariance function for the residuals from the regression model.

When estimating residual spatial covariance for probability samples which are stratified, and the strata have different residual variance structures, it is sometimes necessary to scale the residuals before doing the analysis. The use of residual scaling in residual spatial covariance estimation of spatial populations sampled with stratified designs was investigated. The form of residual scaling influenced the estimated correlation of the residual spatial pattern at short distances and the estimated distance at which the populations units are uncorrelated. The estimated correlation at short distances was always weaker when scaling by a residual standard deviation than when scaling by a heterogeneous variance function $E[y_u]^{\gamma}\sigma$. Heterogeneous scaling leaves more of the spatial variability of the residuals in the spatial pattern model, whereas homogeneous scaling treats more of the residual variability as purely random. Which form of scaling to use needs to be determined from residual diagnostics and any intuitive understanding about the missing covariates that are influencing the spatial pattern of the population. Residual scaling is only necessary when it is of

interest to analyze residuals from mixed strata for spatial covariance when there is evidence that the strata have different variances. For the nonstratified populations, the results for the estimated CDF and other parameters using predicted values from an estimated spatial covariance model with no residual scaling were generally better, and the implications are that using no residual scaling for spatial covariance estimation for stratified populations would be better as well in cases where the stratum variances are similar.

Using spatial information for the estimation of finite population parameters (CDF, mean, standard deviation) yielded improved precision of the estimates over the design based estimates in some cases for the populations studied here. The improvements were largest for the cases where the spatial coordinates were included in the regression model, given that the known explanatory variables were already included in the model. Accounting for the residual spatial covariance led to additional improvements in estimation, especially for the population standard deviation, though this additional improvement was not always large. In general, using residual spatial covariance as part of a modeling strategy to obtain predictions to use in the Chambers and Dunstan estimator did not lead to large improvements in the estimates. In most cases, the simpler method of including the spatial coordinates (or some function of them) in the regression model was adequate for estimation of the population parameters for the populations studied here. These results are dependent upon the strength of the spatial pattern of the missing covariate, the explanatory power of covariates available for the analysis, and which spatial model is a better approximation of the spatial pattern of the missing covariate (using spatial coordinates, residual covariance or both). For the simulation studies conducted here, the missing covariate had a planar spatial pattern and therefore the spatial coordinates were a better approximation than a residual spatial covariance function. The spatial scale of the population is also important. The populations considered here cover a broad scale, with distinct population units separated by distances on the order of kilometers. This result may not hold for other populations. The nature of the population and the scale of the observations may lead to the opposite situation, residual spatial covariance is an important and meaningful component of the population.

# BIBLIOGRAPHY

Carroll, R.J. and D. Ruppert. 1988. Transformation and Weighting in Regression. Chapman and Hall. New York, NY.

Chambers, R.L. and R. Dunstan. 1986. Estimating Distribution Functions from Survey Data. Biometrika. 79:577-582.

Christensen, R. 1991. Linear Models for Multivariate, Time Series and Spatial Data. Springer-Verlag. New York, NY.

Cordy, C.B. and D.A. Griffith. 1993. Efficiency of Least Squares Estimators in the Presence of Spatial Autocorrelation. Commun. Stat.- simulat. 22:1161-1179

Davidian, M. and R.J. Carroll. 1987. Variance Function Estimation. JASA 82:1079-1091

Dorfman, A.H. 1993. A Comparison of Design Based and Model Based Estimators of the Finite Population Distribution Function. Aust.J.Statist. 35:29-41.

Jager, H.I. and W.S. Overton. 1991. Explanatory Models for Ecological Response Surfaces. in Environmental Modeling with GIS. Edited by Goodchild, M.F., B.O. Parks and L.T. Steyaert. Oxford University Press.

Jewell, N.P. 1985. Least Squares Regression with Data arising from Stratified Samples of the Dependent Variable. Biometrika. 72:11-21.

Mardia, K.V. and A.J. Watkins. 1989. On the Multimodality of the Likelihood in the Spatial Linear Model. Biometrika 76:289-295

Mardia, K.V. and R.J. Marshall. 1984. Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression. Biometrika. 71:135-146.

Overton, W.S. 1991. Probability Sampling and Population Inference in Monitoring Programs. in Environmental Modeling with GIS. Edited by Goodchild, M.F., B.O. Parks and L.T. Steyaert. Oxford University Press.

Quesenberry, C.P. and N.P. Jewell 1986. Regression Analysis of Stratified Samples. Biometrika. 73:605-614.

Rao, J.N.K., J.G. Kovar and H.J. Mantel. 1990. On Estimati ıg Distribution Functions and Quantiles from Survey Data using Auxiliary Information. Biometrika. 77:365-375.

Stein, M.L. 1988. Asymptotically Efficient Prediction of a Random Field with a Misspecified Covariance Function. The Annals of Statistics. 16:55-63.

Warnes, J.J. and B.D. Ripley. 1987. Problems with Likelihood Estimation of Covariance Functions of Spatial Gaussian Processes. Biometrika. 74:640-642

Welsh, A.H. R.J. Carroll and D. Ruppert. 1994. Fitting Heteroscedastic Regression Models. JASA. 89:2100-116

Zimmerman, D.L. and N. Cressie. 1992. Mean Squared Prediction Error in the Spatial Linear Model with Estimated Covariance Parameters. Ann. Inst. Statist. Math. 44:27-43.

Zimmerman, D.L. and M.B. Zimmerman. 1991 A Comparison of Semi-Variogram estimators and Corresponding Ordinary Kriging Predictors. Technometrics 33:77-99.

# APPENDICES

Appendix 1. Results for estimated means and standard deviations.
Notation: Each model takes on several forms, which are distinct based upon the predictions used in the estimator, and the form of the standard deviation model used in the Chambers and Dunstan estimator.
hte - design based estimator
olp - estimate using regression predictions and the CDF estimator (11).
ol1 - CD using regression model predictions and $g(y_u;\gamma) = \sqrt{y_u}$.
ol2 - CD using regression model predictions and $g(y_u;\gamma) = 1$.
ukp - estimate using kriging predictions and the CDF estimator (11).
uk1 - CD using kriging model predictions and $g(y_u;\gamma) = \sqrt{y_u}$.
uk2 - CD using kriging model predictions and $g(y_u;\gamma) = 1$.

Results for Population N1, models REG 2 and UK 1.

| | Population Mean | | Population Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | -1.3611 | 10.2336 | -0.3001 | 7.1283 |
| olp | -0.8936 | 2.92886 | -4.5245 | 5.9914 |
| ol1 | -0.4705 | 2.69392 | -3.0630 | 4.5739 |
| ol2 | -0.7598 | 2.69167 | -2.4220 | 4.0919 |
| ukp | -0.8453 | 2.90731 | -4.3355 | 5.8525 |
| uk1 | -1.2179 | 3.02515 | -1.5142 | 3.9009 |
| uk2 | -1.0866 | 2.81813 | -1.4864 | 3.6789 |

Results for Population N1, REG 1 and UK 2 with heterogeneous scaling between stages.

| | Population Mean | | Population Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | 22.1141 | 23.1451 | -23.1612 | 25.1846 |
| ol1 | 5.40281 | 8.41538 | -1.2689 | 7.90459 |
| ol2 | 5.79221 | 8.72716 | 1.55462 | 8.15991 |
| ukp | 22.1929 | 23.4753 | -22.683 | 25.2815 |
| uk1 | 3.56706 | 8.05986 | 1.03546 | 8.45375 |
| uk2 | 3.86471 | 8.28695 | 4.08096 | 9.47838 |

Results for Population N2, models REG 2 and UK1.

| | Population Mean | | Population Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | -2.2055 | 14.6205 | -0.5700 | 11.8863 |
| olp | 18.5206 | 26.9438 | -39.565 | 45.0352 |
| ol1 | -3.5138 | 12.2823 | -2.4261 | 10.7063 |
| ol2 | -0.5061 | 11.7106 | 0.09879 | 11.2312 |
| ukp | 18.6222 | 27.6784 | -38.983 | 44.5835 |
| uk1 | -5.8732 | 16.1312 | 0.14826 | 12.8310 |
| uk2 | -0.9214 | 11.9757 | 0.96207 | 11.4683 |

Results for Population N2, REG 1 and UK 2 with heterogeneous scaling between stages.

| | Population Mean | | Population Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | 28.6632 | 34.6943 | -53.3169 | 57.8088 |
| ol1 | -0.9124 | 11.9954 | -1.0220 | 11.1796 |
| ol2 | 2.26291 | 12.0246 | 1.11565 | 12.1264 |
| ukp | 27.7011 | 34.3539 | -51.1381 | 56.0224 |
| uk1 | -2.3469 | 12.7121 | 0.07128 | 11.3712 |
| uk2 | 0.81761 | 12.4875 | 2.39192 | 12.6129 |

Results for Population S1 stratum 1, REG1

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | -52.8 | 92.754 | 35.687 | 43.82 |
| ol1 | -24.287 | 30.664 | 28.783 | 32.2003 |
| ol2 | -24.077 | 30.7276 | 28.915 | 32.394 |

Results for Population S1 stratum 1, REG 2 and UK 2 with homogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | -1.676 | 10.0664 | -1.9085 | 9.0089 |
| olp | 2.7351 | 3.77562 | -2.2593 | 4.6581 |
| ol1 | 1.4808 | 3.25150 | -0.7763 | 5.0061 |
| ol2 | 1.6577 | 3.37334 | -0.8618 | 5.1079 |
| ukp | 2.0666 | 3.54885 | -2.1560 | 4.7536 |
| uk1 | 0.8444 | 3.09817 | -0.7113 | 5.0977 |
| uk2 | 1.0292 | 3.20021 | -0.8099 | 5.2052 |

Results for Population S1 stratum2, REG1.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | -26.953 | 94.730 | 1.0524 | 66.673 |
| ol1 | -6.2961 | 17.6615 | -3.638 | 15.2701 |
| ol2 | -1.9188 | 15.439 | -5.804 | 15.377 |

Results for Population S1 stratum 2, REG 2 and UK2 with homogeneous scaling between the stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | 3.5204 | 19.097 | -2.5197 | 17.351 |
| olp | 1.9820 | 8.9818 | -2.2999 | 10.283 |
| ol1 | 1.0934 | 7.0943 | -1.1258 | 7.7442 |
| ol2 | 1.5202 | 7.1791 | -1.3988 | 7.6881 |
| ukp | 1.9069 | 8.8582 | -2.4684 | 10.018 |
| uk1 | 1.1375 | 7.0062 | -1.2999 | 7.6746 |
| uk2 | 1.5518 | 7.0807 | -1.5387 | 7.6386 |

Results for Population S1 stratum 3, REG 1

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | 39.504 | 80.919 | -34.92 | 83.448 |
| ol1 | -28.629 | 42.66 | 1.1215 | 28.26 |
| ol2 | -6.380 | 23.787 | 0.8577 | 28.049 |

Results for Population S1 stratum 3, REG 2 and UK 2 with homogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rsme | Bias | Rmse |
| hte | -3.72767 | 62.6714 | 4.13466 | 33.5027 |
| olp | -6.58126 | 20.5937 | -12.956 | 18.7958 |
| ol1 | -22.3422 | 29.0904 | -5.2547 | 14.0857 |
| ol2 | -19.2440 | 26.3185 | -3.7557 | 13.3597 |
| ukp | -8.15566 | 20.7042 | -13.106 | 18.5007 |
| uk1 | -22.8870 | 29.4822 | -6.1084 | 14.0821 |
| uk2 | -19.8526 | 26.8945 | -4.7613 | 13.3533 |

Results for Population S2 stratum 1, REG 1

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | 11.239 | 13.692 | -13.289 | 20.088 |
| ol1 | -1.2407 | 9.3749 | 2.1411 | 13.515 |
| ol2 | 0.2934 | 9.0492 | 1.197 | 13.236 |

Results for Population S2 stratum 1, REG 2 and UK 2 with homogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | -1.9990 | 10.6255 | -2.1618 | 8.36893 |
| olp | 2.48131 | 6.4159 | -3.8108 | 9.62236 |
| ol1 | -3.2658 | 8.1063 | 2.01239 | 10.0109 |
| ol2 | -2.8128 | 7.7176 | 1.98828 | 9.8693 |
| ukp | 2.12423 | 6.4030 | -3.7722 | 9.8398 |
| uk1 | -3.6419 | 8.3828 | 2.07467 | 10.0851 |
| uk2 | -3.2066 | 8.0047 | 2.06943 | 9.9596 |

Results for Population S2 stratum 2, REG 1.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | -25.5767 | 93.771 | -2.030 | 65.1289 |
| ol1 | -7.953 | 18.757 | -2.954 | 16.149 |
| ol2 | -3.653 | 16.1389 | -4.939 | 16.102 |

Results for Population S2 stratum 2, REG 2 and UK 2 with homogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | 3.9476 | 19.5241 | -2.7894 | 17.6021 |
| olp | 3.1536 | 11.5156 | -4.8394 | 14.3084 |
| ol1 | -0.5267 | 8.40087 | -0.6196 | 9.7063 |
| ol2 | 0.3637 | 8.40280 | -0.9942 | 9.6456 |
| ukp | 3.3189 | 11.4313 | -5.0656 | 14.3050 |
| uk1 | -0.4561 | 8.3714 | -0.7304 | 9.6812 |
| uk2 | 0.4171 | 8.3790 | -1.0702 | 9.6365 |

Results for Population S2 stratum 3, REG 1.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | 51.806 | 77.586 | -38.403 | 79.705 |
| ol1 | -19.113 | 32.204 | 3.052 | 28.719 |
| ol2 | 2.343 | 22.519 | 2.782 | 28.861 |

Results for Population S2 stratum 3, REG 2 and UK 2 with homogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | -3.7145 | 64.2147 | 3.74833 | 34.2504 |
| olp | 2.2967 | 19.5241 | -11.6321 | 19.1103 |
| ol1 | -14.018 | 24.4566 | -3.18476 | 14.8481 |
| ol2 | -10.6790 | 22.4625 | -1.13038 | 14.3690 |
| ukp | 0.8880 | 19.5112 | -11.6735 | 18.9060 |
| uk1 | -14.5362 | 24.4516 | -3.5364 | 14.5451 |
| uk2 | -11.3054 | 22.4859 | -1.5129 | 14.0306 |

Results for Population S3 stratum 1 REG 1.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | 8.9396 | 11.7079 | -10.578 | 16.449 |
| ol1 | 0.1227 | 8.707 | -0.391 | 12.112 |
| ol2 | 1.238 | 8.504 | -1.365 | 12.069 |

Results for Population S3 stratum 1, REG 2 and UK 2 with homogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | -1.2892 | 8.7067 | -1.5053 | 9.2073 |
| olp | 8.2401 | 11.672 | −10.5251 | 17.3648 |
| ol1 | −0.1518 | 9.6691 | −0.8658 | 13.3007 |
| ol2 | 0.8595 | 9.4139 | −1.7207 | 13.2266 |
| ukp | 8.1214 | 11.6265 | −10.3537 | 17.3098 |
| uk1 | −0.1853 | 9.6874 | −0.8227 | 13.3077 |
| uk2 | 0.8283 | 9.4303 | −1.6801 | 13.2341 |

Results for Population S3 stratum 1, UK 2 with heterogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| ukp | 8.1960 | 11.6401 | -10.4041 | 17.2675 |
| uk1 | -0.1968 | 9.6385 | -0.7769 | 13.2453 |
| uk2 | 0.8242 | 9.3807 | -1.6401 | 13.1694 |

Results for Population S3 stratum 2, REG 1.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | 19.832 | 25.696 | -28.103 | 37.458 |
| ol1 | 4.941 | 16.3098 | -6.699 | 22.090 |
| ol2 | 5.3277 | 16.3711 | -6.6707 | 21.993 |

Results for Population S3 stratum 2, REG 2 and UK 2 with homogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | 3.3439 | 14.5086 | -2.1245 | 12.4388 |
| olp | 16.5247 | 24.0517 | -22.2775 | 34.0796 |
| ol1 | 4.2148 | 16.7129 | -5.1537 | 21.9008 |
| ol2 | 4.6287 | 16.7465 | -5.2166 | 21.8178 |
| ukp | 16.4872 | 24.0704 | -22.2538 | 34.0598 |
| uk1 | 4.1844 | 16.6916 | -5.1219 | 21.8932 |
| uk2 | 4.5974 | 16.7284 | -5.1847 | 21.8137 |

Results for Population S3, stratum 2, UK 2 with heterogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| ukp | 16.6009 | 24.0714 | -22.4220 | 34.0637 |
| uk1 | 4.3185 | 16.6901 | -5.3276 | 21.9261 |
| uk2 | 4.7263 | 16.7244 | -5.3833 | 21.8385 |

Results for Population S3 stratum 3, REG 1.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| olp | 21.356 | 32.8108 | -28.024 | 43.669 |
| ol1 | 1.955 | 33.6008 | -8.7609 | 42.7122 |
| ol2 | 3.4297 | 34.023 | -9.627 | 43.574 |

Results for Population S3 stratum 3, REG 2 and UK 2 with homogeneous scaling between stages.

| | Stratum Mean | | Stratum Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| hte | -7.5038 | 50.0281 | 3.0564 | 24.6336 |
| olp | 15.2559 | 34.5588 | -21.5084 | 45.0492 |
| ol1 | -0.6993 | 35.0168 | -6.1526 | 43.3065 |
| ol2 | 0.74038 | 35.2981 | -7.0552 | 44.0129 |
| ukp | 14.7465 | 34.2216 | -20.9146 | 44.6146 |
| uk1 | -0.8149 | 34.7619 | -6.0357 | 42.9624 |
| uk2 | 0.62360 | 35.0391 | -6.9346 | 43.6602 |

Results for Population S3 stratum 3, UK 2 with heterogeneous scaling between stages.

| | Stratum | | Stratum | |
|---|---|---|---|---|

| | Mean | | Standard Deviation | |
|---|---|---|---|---|
| | Bias | Rmse | Bias | Rmse |
| ukp | 14.4657 | 34.1535 | -20.3206 | 44.2832 |
| uk1 | -1.3862 | 34.8488 | -5.1718 | 42.7793 |
| uk2 | 0.0434 | 35.1156 | -6.0716 | 43.4844 |

# APPENDIX 2.

Tables for Bias and RMSE of Estimated Percentiles.

Notation: Each model takes on several forms, which are distinct based upon the predictions used in the estimator, and the form of the standard deviation model used in the Chambers and Dunstan estimator. The rows in the tables are as follows:

hte - design based estimator
olp - estimate using regression predictions and the CDF estimator (11).
ol1 - CD using regression model predictions and $g(y_u;\gamma) = \sqrt{y_u}$.
ol2 - CD using regression model predictions and $g(y_u;\gamma) = 1$.
ukp - estimate using kriging predictions and the CDF estimator (11).
uk1 - CD using kriging model predictions and $g(y_u;\gamma) = \sqrt{y_u}$.
uk2 - CD using kriging model predictions and $g(y_u;\gamma) = 1$.

Results for Population N1, models REG 2 and UK 1.

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.92 | 0.78 | 1.47 | 1.49 | 1.22 | 1.74 | 1.14 | 0.61 | 0.65 | 0.68 | 0.67 |
| -0.91 | -1.47 | -0.04 | -0.10 | 0.50 | 2.29 | 1.62 | 0.22 | 1.49 | 0.97 | -0.32 |
| -0.73 | -1.22 | 0.36 | 0.11 | 0.04 | 2.40 | 0.68 | -0.82 | 1.15 | 0.71 | -0.32 |
| -0.59 | -1.15 | 0.43 | -0.00 | -0.12 | 2.27 | 0.54 | -0.92 | 1.11 | 0.73 | -0.33 |
| -0.88 | -1.42 | -0.05 | -0.10 | 0.45 | 2.18 | 1.53 | 0.24 | 1.44 | 0.91 | -0.31 |
| -0.46 | -0.90 | 0.79 | 0.25 | 0.06 | 2.30 | 0.31 | -1.23 | 1.04 | 0.54 | -0.37 |
| -0.42 | -0.96 | 0.64 | 0.03 | -0.18 | 2.15 | 0.28 | -1.16 | 1.03 | 0.63 | -0.34 |

Tabled values are rmse*100.

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 18.99 | 7.74 | 14.52 | 14.74 | 12.08 | 17.20 | 11.24 | 6.00 | 6.47 | 6.76 | 6.60 |
| 9.03 | 14.59 | 0.37 | 1.01 | 4.96 | 22.66 | 16.07 | 2.21 | 14.78 | 9.61 | 3.15 |
| 7.26 | 12.09 | 3.60 | 1.05 | 0.38 | 23.71 | 6.73 | 8.15 | 11.36 | 7.08 | 3.15 |
| 5.89 | 11.39 | 4.25 | 0.05 | 1.24 | 22.45 | 5.30 | 9.09 | 11.03 | 7.27 | 3.28 |
| 8.66 | 14.03 | 0.45 | 1.01 | 4.50 | 21.61 | 15.16 | 2.33 | 14.23 | 8.99 | 3.02 |
| 4.56 | 8.93 | 7.83 | 2.51 | 0.61 | 22.72 | 3.05 | 12.14 | 10.33 | 5.36 | 3.67 |
| 4.15 | 9.45 | 6.31 | 0.26 | 1.77 | 21.30 | 2.81 | 11.49 | 10.23 | 6.27 | 3.41 |

Results for Population N1, using models REG 1 and UK 2 with heterogeneous scaling between the two stages.

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| -1.81 | -4.27 | -5.15 | -6.49 | -0.98 | 4.64 | 1.96 | -1.12 | 1.23 | 2.84 | 3.22 |
| 0.45 | -0.04 | 1.17 | -0.29 | -1.06 | 0.58 | -2.03 | -3.61 | 0.62 | 1.13 | 1.01 |
| 1.18 | 0.71 | 1.51 | -0.39 | -1.62 | -0.39 | -3.10 | -4.62 | 0.14 | 1.17 | 1.27 |
| -1.84 | -4.20 | -4.99 | -6.28 | -0.84 | 4.93 | 1.77 | -1.52 | 0.99 | 2.40 | 3.15 |
| 0.70 | 0.30 | 1.53 | -0.03 | -0.96 | 0.45 | -2.26 | -3.94 | 0.25 | 0.77 | 0.71 |
| 1.48 | 1.08 | 1.87 | -0.15 | -1.54 | -0.55 | -3.38 | -5.00 | -0.28 | 0.79 | 0.96 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 17.99 | 42.49 | 51.29 | 64.59 | 9.76 | 46.16 | 19.49 | 11.11 | 12.22 | 28.27 | 32.03 |
| 4.46 | 0.45 | 11.63 | 2.93 | 10.59 | 5.77 | 20.16 | 35.95 | 6.17 | 11.29 | 10.08 |
| 11.77 | 7.02 | 14.98 | 3.89 | 16.09 | 3.91 | 30.82 | 46.00 | 1.35 | 11.65 | 12.61 |
| 18.28 | 41.75 | 49.67 | 62.44 | 8.36 | 49.05 | 17.58 | 15.15 | 9.88 | 23.89 | 31.33 |
| 6.92 | 2.97 | 15.25 | 0.32 | 9.60 | 4.52 | 22.53 | 39.23 | 2.49 | 7.67 | 7.06 |
| 14.77 | 10.78 | 18.56 | 1.46 | 15.35 | 5.50 | 33.62 | 49.78 | 2.75 | 7.82 | 9.54 |

Results for Population N2, models REG 2 and UK1.
Tabled Values are BIAS*100 (top block) and RMSE*100 (bottom block).

| Estimated Percentile | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
| -3.19 | -7.47 | -11.8 | -9.29 | -2.50 | 3.45 | 9.52 | 8.64 | 6.03 | 3.50 | 2.53 |
| -0.95 | 0.17 | -0.19 | 0.84 | 0.99 | 0.63 | 2.49 | 2.27 | 0.08 | -1.35 | -0.67 |
| -0.18 | 0.74 | -0.22 | 0.11 | -0.21 | -0.80 | 1.11 | 1.23 | -0.51 | -1.30 | -0.22 |
| -3.12 | -7.26 | -11.38 | -9.08 | -2.82 | 2.87 | 9.22 | 8.56 | 5.95 | 3.41 | 2.53 |
| -0.48 | 0.72 | 0.29 | 1.12 | 1.05 | 0.50 | 2.15 | 1.81 | -0.41 | -1.79 | -0.99 |
| -0.06 | 0.88 | -0.13 | 0.13 | -0.26 | -0.91 | 0.93 | 1.03 | -0.70 | -1.44 | -0.30 |
| rmse*100 | | | | | | | | | | |
| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
| 16.81 | 9.97 | 10.17 | 13.22 | 11.02 | 16.33 | 20.25 | 12.62 | 12.67 | 7.76 | 4.36 |
| 31.37 | 73.56 | 115.8 | 91.54 | 24.65 | 33.97 | 93.80 | 85.09 | 59.38 | 34.45 | 24.91 |
| 9.32 | 1.63 | 1.87 | 8.27 | 9.72 | 6.24 | 24.55 | 22.34 | 0.78 | 13.34 | 6.58 |
| 1.82 | 7.34 | 2.16 | 1.12 | 2.05 | 7.84 | 10.89 | 12.08 | 5.04 | 12.76 | 2.14 |
| 30.74 | 71.54 | 112.1 | 89.39 | 27.76 | 28.28 | 90.83 | 84.31 | 58.62 | 33.58 | 24.89 |
| 4.71 | 7.12 | 2.85 | 11.00 | 10.33 | 4.91 | 21.15 | 17.85 | 4.00 | 17.62 | 9.73 |
| 0.60 | 8.67 | 1.23 | 1.30 | 2.59 | 9.00 | 9.15 | 10.14 | 6.89 | 14.16 | 2.94 |

Results for Population N2, using models REG 1 and UK2 with hetereogeneous scaling
between the two stages.
Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block).

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.74 | 1.02 | 0.96 | 1.28 | 1.08 | 1.49 | 2.01 | 1.28 | 1.24 | 0.80 | 0.47 |
| -3.11 | -7.73 | -13.7 | -14.5 | -6.34 | 4.29 | 9.47 | 9.42 | 7.71 | 6.33 | 4.32 |
| -0.38 | 0.71 | -0.05 | 0.45 | 0.29 | -0.26 | 1.59 | 1.61 | -0.25 | -1.17 | -0.18 |
| 0.30 | 1.23 | -0.03 | -0.08 | -0.73 | -1.54 | 0.26 | 0.55 | -0.87 | -1.11 | 0.30 |
| -3.04 | -7.67 | -13.4 | -13.7 | -5.52 | 4.31 | 9.25 | 9.20 | 7.28 | 5.84 | 4.17 |
| -0.26 | 0.91 | 0.19 | 0.73 | 0.47 | -0.14 | 1.59 | 1.51 | -0.42 | -1.39 | -0.37 |
| 0.46 | 1.45 | 0.21 | 0.12 | -0.58 | -1.47 | 0.22 | 0.41 | -1.08 | -1.36 | 0.10 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 17.3 | 10.1 | 9.5 | 12.6 | 10.6 | 14.8 | 19.9 | 12.7 | 12.3 | 7.9 | 4.6 |
| 30.8 | 76.5 | 135.8 | 144.0 | 62.7 | 42.5 | 93.8 | 93.3 | 76.3 | 62.6 | 42.8 |
| 3.81 | 6.99 | 0.54 | 4.87 | 2.85 | 2.54 | 15.71 | 15.95 | 2.47 | 11.63 | 1.75 |
| 2.94 | 12.15 | 0.31 | 0.83 | 7.24 | 15.25 | 2.55 | 5.49 | 8.65 | 11.01 | 2.99 |
| 30.1 | 76.0 | 132.3 | 135.5 | 54.6 | 42.7 | 91.6 | 91.0 | 72.1 | 57.9 | 41.3 |
| 2.55 | 9.03 | 1.93 | 7.21 | 4.69 | 1.40 | 15.7 | 14.98 | 4.12 | 13.77 | 3.69 |
| 4.52 | 14.34 | 2.07 | 1.23 | 5.78 | 14.53 | 2.14 | 4.11 | 10.69 | 13.43 | 0.96 |

Results for Population S1 by stratum, models REG 2, UK2 with homogeneous scaling between the two stages, and REG 1.
Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block).
STRATUM 1.

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.48 | 1.95 | 2.03 | 2.44 | 1.86 | 2.92 | 2.13 | 3.85 | 2.94 | -7.38 | -33.2 |
| 0.25 | -0.68 | 0.11 | -1.16 | -3.07 | 0.82 | 0.37 | 0.33 | 0.86 | 1.11 | 0.80 |
| 0.02 | -0.68 | -0.38 | -0.40 | -2.97 | 1.17 | -1.02 | 0.65 | 0.70 | -0.22 | 0.20 |
| 0.02 | -0.62 | -0.44 | -0.38 | -3.01 | 1.07 | -1.12 | 0.53 | 0.66 | -0.19 | 0.23 |
| 0.28 | -0.76 | -0.04 | -1.06 | -2.80 | 2.10 | 0.86 | 0.73 | 1.49 | 1.10 | 0.74 |
| 0.02 | -0.73 | -0.43 | -0.19 | -2.68 | 1.94 | -0.40 | 1.07 | 1.05 | -0.20 | 0.12 |
| 0.02 | -0.66 | -0.48 | -0.17 | -2.72 | 1.84 | -0.51 | 0.95 | 1.01 | -0.16 | 0.16 |

REG 1 results BIAS*100

| 13.85 | 0.03 | -4.63 | -7.05 | -10.9 | -11.9 | -13.5 | -11.5 | -12.7 | -13.0 | -9.27 |
|---|---|---|---|---|---|---|---|---|---|---|
| -1.08 | -2.69 | -5.38 | -7.05 | -10.3 | -11.1 | -13.5 | -12.8 | -13.3 | -14.1 | -10.93 |
| -1.04 | -2.65 | -5.45 | -7.21 | -10.5 | -11.4 | -13.7 | -13.0 | -13.5 | -14.3 | -10.9 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 14.77 | 19.48 | 20.26 | 24.41 | 18.64 | 29.15 | 21.28 | 38.48 | 29.42 | 73.77 | 331.66 |
| 2.49 | 6.75 | 1.14 | 11.59 | 30.66 | 8.24 | 3.68 | 3.28 | 8.61 | 11.12 | 7.99 |
| 0.17 | 6.84 | 3.83 | 3.98 | 29.68 | 11.68 | 10.19 | 6.48 | 7.02 | 2.24 | 1.95 |
| 0.24 | 6.16 | 4.39 | 3.80 | 30.09 | 10.67 | 11.23 | 5.28 | 6.57 | 1.86 | 2.34 |
| 2.77 | 7.59 | 0.43 | 10.59 | 28.00 | 20.98 | 8.61 | 7.28 | 14.95 | 11.00 | 7.37 |
| 0.19 | 7.32 | 4.31 | 1.91 | 26.79 | 19.40 | 4.03 | 10.74 | 10.51 | 1.96 | 1.24 |
| 0.22 | 6.63 | 4.82 | 1.74 | 27.17 | 18.40 | 5.14 | 9.52 | 10.13 | 1.58 | 1.63 |

REG 1 results rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 138 5 | 0.34 | 46.31 | 70.52 | 109.3 | 118.8 | 135.3 | 114.9 | 127.1 | 130.1 | 92.66 |
| 10.79 | 26.89 | 53.85 | 70.5 | 102.6 | 111.2 | 134.9 | 127.6 | 132.8 | 141.4 | 109.30 |
| 10.35 | 26.54 | 54.53 | 72.09 | 104.5 | 113.7 | 137.6 | 130.1 | 134.9 | 142.8 | 109.09 |

STRATUM 2

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.36 | 1.40 | 0.88 | 0.32 | 0.37 | 0.27 | 0.11 | 0.66 | 0.50 | -5.47 | -18.52 |
| -0.75 | 0.64 | 0.40 | 0.64 | 0.08 | 1.50 | -0.42 | -0.06 | -1.99 | -1.35 | -0.11 |
| -0.99 | 0.55 | 0.77 | 0.25 | 0.93 | 1.55 | -0.33 | -0.70 | -3.07 | -1.71 | -0.49 |
| -1.00 | 0.56 | 0.74 | 0.15 | 0.77 | 1.37 | -0.45 | -0.70 | -3.02 | -1.67 | -0.44 |
| -0.85 | 0.66 | 0.56 | 0.88 | 0.27 | 1.41 | -0.51 | -0.02 | -2.11 | -1.43 | -0.16 |
| -1.08 | 0.54 | 0.87 | 0.44 | 1.05 | 1.51 | -0.41 | -0.79 | -3.16 | -1.81 | -0.53 |
| -1.07 | 0.54 | 0.84 | 0.34 | 0.90 | 1.34 | -0.53 | -0.78 | -3.11 | -1.77 | -0.47 |

REG 1 results, BIAS*100

| 12.09 | -2.51 | -1.40 | -0.56 | 4.68 | 5.89 | 3.75 | 3.57 | 1.88 | 1.10 | 1.68 |
|---|---|---|---|---|---|---|---|---|---|---|
| -2.52 | -0.07 | 3.25 | 3.65 | 6.45 | 6.07 | 2.45 | -0.23 | -3.92 | -3.13 | -2.00 |
| -2.35 | -0.33 | 2.40 | 2.61 | 5.48 | 5.37 | 1.98 | -0.37 | -3.76 | -2.50 | -1.30 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.57 | 14.04 | 8.83 | 3.21 | 3.68 | 2.65 | 1.08 | 6.57 | 4.95 | 54.74 | 185.23 |
| 7.47 | 6.43 | 3.95 | 6.43 | 0.75 | 14.95 | 4.19 | 0.63 | 19.95 | 13.55 | 1.14 |
| 9.91 | 5.54 | 7.67 | 2.54 | 9.29 | 15.47 | 3.32 | 7.04 | 30.73 | 17.07 | 4.93 |
| 9.97 | 5.61 | 7.43 | 1.52 | 7.70 | 13.73 | 4.52 | 6.99 | 30.22 | 16.69 | 4.36 |
| 8.47 | 6.59 | 5.65 | 8.80 | 2.68 | 14.10 | 5.11 | 0.22 | 21.09 | 14.28 | 1.62 |
| 10.78 | 5.39 | 8.71 | 4.37 | 10.54 | 15.13 | 4.12 | 7.87 | 31.63 | 18.09 | 5.29 |
| 10.75 | 5.44 | 8.43 | 3.36 | 8.96 | 13.38 | 5.35 | 7.81 | 31.08 | 17.73 | 4.69 |

REG 1 results, rmse*100

| 120.8 | 25.15 | 14.04 | 5.62 | 46.84 | 58.90 | 37.53 | 35.74 | 18.76 | 11.03 | 16.77 |
|---|---|---|---|---|---|---|---|---|---|---|
| 25.22 | 0.68 | 32.47 | 36.52 | 64.52 | 60.69 | 24.46 | 2.30 | 39.18 | 31.32 | 20.05 |
| 23.52 | 3.27 | 24.03 | 26.12 | 54.85 | 53.68 | 19.78 | 3.68 | 37.61 | 24.99 | 12.96 |

STRATUM 3

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.94 | 11.64 | 45.15 | 47.06 | 41.13 | 33.98 | 27.23 | 20.64 | 13.79 | 6.99 | 3.42 |
| -1.30 | -2.25 | 0.80 | 0.72 | 0.64 | 2.70 | 1.43 | 2.03 | 0.52 | -0.07 | -0.38 |
| -1.07 | -1.89 | 0.95 | 0.80 | 1.10 | 2.31 | 1.16 | 1.58 | 0.10 | -0.43 | -1.56 |
| -0.69 | -1.64 | 0.89 | 0.60 | 0.85 | 2.18 | 1.07 | 1.56 | 0.03 | -0.42 | -1.28 |
| -1.32 | -2.51 | 0.73 | 0.59 | 0.65 | 2.82 | 1.55 | 2.21 | 0.40 | -0.06 | -0.53 |
| -1.15 | -2.20 | 0.81 | 0.65 | 1.09 | 2.41 | 1.36 | 1.72 | 0.05 | -0.51 | -1.66 |
| -0.81 | -1.94 | 0.71 | 0.46 | 0.84 | 2.28 | 1.28 | 1.71 | -0.05 | -0.49 | -1.39 |

REG 1 results, BIAS*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.26 | -3.34 | -0.02 | -1.94 | -2.81 | -3.05 | -3.51 | 2.18 | 5.20 | 6.36 | 3.38 |
| 0.23 | -1.59 | 0.52 | 1.61 | 1.31 | 0.66 | -0.77 | 0.78 | 0.87 | 0.04 | -1.92 |
| 1.06 | -1.13 | -0.60 | -0.12 | -0.54 | -1.09 | -2.19 | 0.22 | 1.34 | 1.30 | -0.57 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 29.4 | 116.4 | 451.5 | 470.6 | 411.3 | 339.8 | 272.3 | 206.4 | 137.9 | 69.9 | 34.2 |
| 13.03 | 22.53 | 7.99 | 7.17 | 6.42 | 27.01 | 14.27 | 20.27 | 5.17 | 0.75 | 3.81 |
| 10.75 | 18.92 | 9.54 | 7.95 | 10.97 | 23.09 | 11.65 | 15.77 | 1.02 | 4.31 | 15.64 |
| 6.95 | 16.37 | 8.86 | 6.03 | 8.53 | 21.82 | 10.75 | 15.55 | 0.25 | 4.25 | 12.78 |
| 13.17 | 25.14 | 7.30 | 5.94 | 6.48 | 28.21 | 15.50 | 22.07 | 4.05 | 0.61 | 5.33 |
| 11.47 | 21.97 | 8.05 | 6.48 | 10.94 | 24.13 | 13.58 | 17.18 | 0.48 | 5.06 | 16.65 |
| 8.10 | 19.36 | 7.15 | 4.61 | 8.44 | 22.83 | 12.77 | 17.07 | 0.47 | 4.93 | 13.94 |

REG 1 results, rmse*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 32.59 | 33.45 | 0.16 | 19.37 | 28.09 | 30.45 | 35.15 | 21.78 | 51.96 | 63.60 | 33.81 |
| 2.28 | 15.94 | 5.19 | 16.10 | 13.13 | 6.60 | 7.68 | 7.81 | 8.73 | 0.37 | 19.23 |
| 10.56 | 11.32 | 6.04 | 1.24 | 5.38 | 10.85 | 21.91 | 2.18 | 1[illegible].35 | 12.98 | 5.71 |

Results for Population S2 by stratum, models REG 2 and UK 2 with homogeneous scaling between the two stages, and REG 1.

STRATUM 1

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.52 | 2.43 | 2.75 | 3.12 | 2.70 | 3.22 | 3.52 | 3.31 | 1.19 | -9.03 | -30.62 |
| -0.05 | -0.94 | -0.83 | -3.40 | -0.44 | 1.68 | 2.49 | 2.10 | 0.07 | 2.19 | 0.65 |
| 0.08 | -0.18 | -0.75 | -1.54 | 0.46 | 1.67 | 1.37 | 2.01 | -0.91 | -0.24 | -0.82 |
| 0.17 | -0.10 | -0.84 | -1.67 | 0.24 | 1.39 | 1.09 | 1.80 | -0.99 | -0.09 | -0.68 |
| -0.12 | -0.91 | -0.88 | -3.25 | -0.45 | 2.14 | 2.87 | 2.27 | 0.09 | 2.01 | 0.55 |
| 0.05 | -0.20 | -0.75 | -1.45 | 0.68 | 1.96 | 1.65 | 2.18 | -0.82 | -0.29 | -0.90 |
| 0.14 | -0.13 | -0.84 | -1.58 | 0.46 | 1.69 | 1.37 | 1.97 | -0.91 | -0.15 | -0.76 |

REG 1 results, BIAS*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.57 | -0.93 | -4.27 | -7.10 | -6.45 | -3.12 | 1.62 | 6.63 | 5.29 | 5.49 | 3.13 |
| 1.08 | 0.14 | -1.86 | -2.98 | -1.26 | 0.06 | 0.40 | 2.04 | -0.45 | 0.41 | 0.11 |
| 1.13 | 0.10 | -2.08 | -3.39 | -1.83 | -0.50 | -0.01 | 1.92 | -0.36 | 0.78 | 0.48 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 15.1 | 24.2 | 27.3 | 31.0 | 26.8 | 32.0 | 35.1 | 32.9 | 11.8 | 89.8 | 304.7 |
| 0.55 | 9.33 | 8.29 | 33.82 | 4.35 | 16.73 | 24.78 | 20.90 | 0.66 | 21.81 | 6.44 |
| 0.76 | 1.83 | 7.45 | 15.36 | 4.55 | 16.62 | 13.67 | 19.97 | 9.02 | 2.40 | 8.15 |
| 1.67 | 1.04 | 8.36 | 16.62 | 2.39 | 13.87 | 10.88 | 17.93 | 9.83 | 0.92 | 6.79 |
| 1.19 | 9.08 | 8.76 | 32.39 | 4.52 | 21.27 | 28.55 | 22.56 | 0.92 | 20.00 | 5.46 |
| 0.47 | 2.03 | 7.50 | 14.46 | 6.75 | 19.53 | 16.41 | 21.67 | 8.20 | 2.90 | 8.92 |
| 1.39 | 1.26 | 8.37 | 15.69 | 4.62 | 16.80 | 13.64 | 19.62 | 9.05 | 1.48 | 7.60 |

REG 1 results, rmse*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5.68 | 9.30 | 42.74 | 70.96 | 64.46 | 31.19 | 16.23 | 66.29 | 52.89 | 54.92 | 31.33 |
| 10.81 | 1.40 | 18.64 | 29.78 | 12.60 | 0.58 | 4.04 | 20.44 | 4.54 | 4.11 | 1.13 |
| 11.33 | 1.03 | 20.84 | 33.95 | 18.35 | 5.02 | 0.11 | 19.23 | 3.60 | 7.76 | 4.77 |

STRATUM 2

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.69 | 8.31 | 20.23 | 18.03 | 12.41 | 6.27 | 0.03 | -6.31 | -12.3 | -17.9 | -20.23 |
| -0.75 | 0.06 | 1.56 | 1.98 | -1.15 | -0.02 | -0.15 | 0.30 | -0.48 | -1.09 | 0.20 |
| -0.88 | 0.24 | 2.39 | 1.82 | -0.02 | 0.98 | -0.10 | -1.01 | -2.01 | -1.95 | -0.58 |
| -0.83 | 0.25 | 2.30 | 1.61 | -0.30 | 0.69 | -0.31 | -1.07 | -1.96 | -1.88 | -0.43 |
| -0.82 | 0.08 | 1.69 | 2.04 | -1.12 | -0.05 | -0.17 | 0.25 | -0.56 | -1.17 | 0.20 |
| -0.93 | 0.24 | 2.45 | 1.89 | 0.01 | 0.96 | -0.15 | -1.08 | -2.08 | -1.99 | -0.59 |
| -0.87 | 0.26 | 2.37 | 1.68 | -0.27 | 0.66 | -0.36 | -1.14 | -2.03 | -1.91 | -0.45 |

REG 1 results, BIAS*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 11.18 | -2.15 | -0.36 | 1.21 | 2.25 | 4.85 | 4.29 | 4.18 | 3.47 | 1.33 | 1.65 |
| -2.30 | -0.52 | 4.83 | 5.57 | 4.65 | 5.16 | 2.58 | -0.13 | -2.64 | -3.32 | -2.06 |
| -2.12 | -0.73 | 3.94 | 4.51 | 3.63 | 4.40 | 2.09 | -0.30 | -2.44 | -2.71 | -1.37 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 26.8 | 82.6 | 201.3 | 179.4 | 123.4 | 62.4 | 0.30 | 62.8 | 122.4 | 177.8 | 201.3 |
| 7.49 | 0.62 | 15.55 | 19.72 | 11.43 | 0.19 | 1.49 | 2.94 | 4.80 | 10.87 | 2.04 |
| 8.74 | 2.37 | 23.74 | 18.12 | 0.20 | 9.80 | 0.95 | 10.07 | 20.00 | 19.43 | 5.76 |
| 8.31 | 2.52 | 22.89 | 16.02 | 3.03 | 6.83 | 3.05 | 10.68 | 19.51 | 18.71 | 4.32 |
| 8.18 | 0.77 | 16.85 | 20.34 | 11.12 | 0.46 | 1.66 | 2.44 | 5.56 | 11.62 | 1.98 |
| 9.21 | 2.44 | 24.40 | 18.82 | 0.13 | 9.55 | 1.47 | 10.73 | 20.66 | 19.77 | 5.90 |
| 8.70 | 2.55 | 23.55 | 16.71 | 2.69 | 6.56 | 3.58 | 11.37 | 20.21 | 19.02 | 4.46 |

REG 1 results, rmse*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 111.8 | 21.49 | 3.64 | 12.12 | 22.50 | 48.53 | 42.87 | 41.79 | 34.66 | 13.28 | 16.52 |
| 23.03 | 5.19 | 48.29 | 55.69 | 46.49 | 51.59 | 25.83 | 1.32 | 26.40 | 33.15 | 20.64 |
| 21.21 | 7.30 | 39.36 | 45.07 | 36.32 | 44.00 | 20.94 | 2.99 | 24.41 | 27.12 | 13.67 |

STRATUM 3

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 4.09 | 10.84 | 43.55 | 46.74 | 40.90 | 34.12 | 27.25 | 20.32 | 13.73 | 6.88 | 3.52 |
| -1.23 | -2.50 | 0.29 | 0.21 | 0.79 | 4.14 | 1.17 | 1.01 | -0.38 | -0.82 | -0.02 |
| -0.80 | -2.02 | 0.63 | 0.06 | 1.46 | 3.67 | 0.85 | 0.58 | -0.81 | -0.91 | -1.14 |
| -0.32 | -1.70 | 0.53 | -0.14 | 1.20 | 3.51 | 0.74 | 0.52 | -0.90 | -0.94 | -0.83 |
| -1.23 | -2.65 | 0.32 | 0.11 | 0.87 | 4.31 | 1.27 | 1.11 | -0.44 | -0.77 | -0.12 |
| -0.84 | -2.14 | 0.58 | -0.00 | 1.48 | 3.76 | 0.95 | 0.66 | -0.82 | -0.95 | -1.20 |
| -0.37 | -1.83 | 0.47 | -0.20 | 1.22 | 3.60 | 0.85 | 0.61 | -0.92 | -0.98 | -0.90 |

REG 1 results, BIAS*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.44 | -4.40 | -0.58 | -2.58 | -2.81 | -1.95 | -4.05 | 0.78 | 3.72 | 6.18 | 3.69 |
| 0.42 | -1.70 | 0.02 | 0.64 | 1.41 | 1.78 | -1.21 | -0.38 | -0.20 | -0.25 | -1.46 |
| 1.30 | -1.17 | -1.01 | -1.04 | -0.43 | 0.07 | -2.67 | -1.09 | 0.08 | 0.92 | -0.18 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 40.7 | 107.9 | 433.3 | 465.1 | 406.9 | 339.5 | 271.2 | 202.2 | 136.6 | 68.4 | 35.1 |
| 12.24 | 24.90 | 2.88 | 2.08 | 7.87 | 41.20 | 11.67 | 10.02 | 3.82 | 8.15 | 0.20 |
| 7.94 | 20.07 | 6.32 | 0.60 | 14.55 | 36.53 | 8.49 | 5.76 | 8.11 | 9.04 | 11.35 |
| 3.17 | 16.96 | 5.29 | 1.41 | 11.96 | 34.88 | 7.40 | 5.17 | 8.97 | 9.39 | 8.28 |
| 12.26 | 26.38 | 3.14 | 1.07 | 8.67 | 42.85 | 12.65 | 11.06 | 4.42 | 7.63 | 1.24 |
| 8.34 | 21.28 | 5.79 | 0.05 | 14.78 | 37.44 | 9.50 | 6.59 | 8.20 | 9.44 | 11.91 |
| 3.71 | 18.17 | 4.65 | 2.01 | 12.16 | 35.81 | 8.45 | 6.08 | 9.16 | 9.74 | 8.96 |

REG1 results, rmse*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 24.38 | 44.01 | 5.81 | 25.83 | 28.10 | 19.52 | 40.50 | 7.76 | 37.20 | 61.80 | 36.87 |
| 4.21 | 17.00 | 0.24 | 6.40 | 14.14 | 17.82 | 12.09 | 3.80 | 1.97 | 2.49 | 14.65 |
| 12.98 | 11.68 | 10.10 | 10.41 | 4.28 | 0.69 | 26.71 | 10.87 | 0.78 | 9.22 | 1.80 |

Results for Population S3 by stratum, models REG 2 and UK 2 with homogeneous scaling between the two stages and REG 1.

STRATUM 1

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 23.57 | -2.88 | -12.5 | -19.4 | -25.9 | -32.2 | -39.1 | -45.6 | -52.5 | -59.2 | -62.7 |
| -0.06 | -0.37 | -0.69 | -1.83 | -3.25 | -0.58 | -0.08 | 2.77 | 0.59 | 2.34 | 1.92 |
| -0.11 | 0.07 | 0.73 | 0.14 | -1.05 | 0.92 | -0.92 | 1.17 | -1.03 | -0.72 | -0.16 |
| -0.11 | 0.06 | 0.63 | -0.06 | -1.34 | 0.63 | -1.17 | 1.06 | -1.03 | -0.53 | 0.06 |
| -0.04 | -0.36 | -0.73 | -1.81 | -3.23 | -0.54 | -0.08 | 2.76 | 0.57 | 2.34 | 1.91 |
| -0.11 | 0.07 | 0.72 | 0.15 | -1.04 | 0.93 | -0.91 | 1.16 | -1.05 | -0.73 | -0.17 |
| -0.11 | 0.06 | 0.63 | -0.06 | -1.32 | 0.63 | -1.17 | 1.04 | -1.05 | -0.54 | 0.05 |

REG 1 results, BIAS*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.15 | -0.35 | -1.10 | -2.00 | -4.68 | -1.74 | -0.91 | 3.22 | 1.48 | 2.83 | 2.18 |
| 0.03 | 0.17 | 0.59 | -0.26 | -1.66 | 0.27 | -1.44 | 1.00 | -1.00 | -0.58 | -0.05 |
| 0.03 | 0.16 | 0.48 | -0.49 | -1.97 | -0.05 | -1.71 | 0.88 | -0.98 | -0.38 | 0.19 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 234.5 | 28.7 | 124.6 | 192.9 | 257.7 | 320.6 | 388.7 | 454.0 | 522.3 | 588.9 | 624.0 |
| 0.62 | 3.69 | 6.91 | 18.19 | 32.31 | 5.77 | 0.75 | 27.58 | 5.87 | 23.24 | 19.14 |
| 1.12 | 0.66 | 7.23 | 1.42 | 10.49 | 9.17 | 9.12 | 11.68 | 10.27 | 7.16 | 1.62 |
| 1.11 | 0.63 | 6.30 | 0.63 | 13.30 | 6.23 | 11.62 | 10.51 | 10.27 | 5.26 | 0.58 |
| 0.40 | 3.55 | 7.26 | 17.99 | 32.13 | 5.39 | 0.85 | 27.45 | 5.66 | 23.26 | 19.00 |
| 1.09 | 0.68 | 7.16 | 1.49 | 10.36 | 9.26 | 9.10 | 11.56 | 10.45 | 7.30 | 1.69 |
| 1.08 | 0.64 | 6.25 | 0.55 | 13.17 | 6.31 | 11.61 | 10.38 | 10.45 | 5.41 | 0.51 |

REG 1 results, rmse*100

| 1.46 | 3.53 | 11.03 | 20.04 | 46.78 | 17.45 | 9.11 | 32.20 | 14.75 | 28.32 | 21.84 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.33 | 1.68 | 5.92 | 2.62 | 16.08 | 2.70 | 14.44 | 9.97 | 9.96 | 5.83 | 0.52 |
| 0.28 | 1.64 | 4.82 | 4.91 | 19.65 | 0.50 | 17.13 | 8.85 | 9.77 | 3.76 | 1.89 |

STRATUM 2

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 30.30 | 26.99 | 20.06 | 13.28 | 6.75 | 0.02 | -6.57 | -13.0 | -19.7 | -26.4 | -29.68 |
| -0.66 | -0.38 | -1.61 | -1.89 | -0.39 | -1.69 | -0.00 | 0.33 | 1.98 | -0.40 | 1.46 |
| -0.41 | 0.10 | -0.16 | -0.65 | -0.24 | -1.61 | 0.29 | -0.60 | -0.66 | -1.60 | 0.05 |
| -0.35 | 0.13 | -0.16 | -0.74 | -0.41 | -1.81 | 0.11 | -0.72 | -0.69 | -1.58 | 0.13 |
| -0.66 | -0.38 | -1.62 | -1.88 | -0.35 | -1.68 | 0.02 | 0.33 | 1.98 | -0.38 | 1.46 |
| -0.41 | 0.10 | -0.15 | -0.63 | -0.23 | -1.60 | 0.28 | -0.61 | -0.67 | -1.60 | 0.05 |
| -0.35 | 0.13 | -0.14 | -0.72 | -0.40 | -1.80 | 0.10 | -0.73 | -0.70 | -1.58 | 0.13 |

REG 1 results, BIAS*100

| -1.08 | -0.26 | -1.21 | -2.32 | -0.84 | -1.53 | -0.23 | 0.74 | 2.45 | -0.14 | 1.61 |
|---|---|---|---|---|---|---|---|---|---|---|
| -0.60 | 0.09 | 0.07 | -0.50 | -0.17 | -1.52 | 0.45 | -0.43 | -0.56 | -1.56 | 0.04 |
| -0.53 | 0.12 | 0.05 | -0.61 | -0.35 | -1.73 | 0.26 | -0.56 | -0.59 | -1.54 | 0.12 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 301.5 | 268.5 | 199.6 | 132.2 | 67.2 | 0.2 | 65.4 | 129.1 | 195.9 | 263.2 | 295.3 |
| 6.61 | 3.80 | 15.98 | 18.84 | 3.87 | 16.85 | 0.04 | 3.29 | 19.75 | 3.94 | 14.53 |
| 4.07 | 0.96 | 1.64 | 6.44 | 2.39 | 15.99 | 2.85 | 5.98 | 6.58 | 15.89 | 0.52 |
| 3.50 | 1.26 | 1.58 | 7.36 | 4.11 | 17.97 | 1.06 | 7.17 | 6.91 | 15.76 | 1.28 |
| 6.59 | 3.77 | 16.10 | 18.75 | 3.52 | 16.75 | 0.16 | 3.31 | 19.69 | 3.80 | 14.52 |
| 4.04 | 1.03 | 1.48 | 6.25 | 2.25 | 15.92 | 2.80 | 6.06 | 6.64 | 15.90 | 0.51 |
| 3.48 | 1.33 | 1.42 | 7.18 | 3.97 | 17.90 | 1.02 | 7.25 | 6.96 | 15.76 | 1.27 |

REG 1 results, rmse*100

| 10.85 | 2.62 | 12.06 | 23.22 | 8.45 | 15.33 | 2.26 | 7.44 | 24.51 | 1.44 | 16.07 |
|---|---|---|---|---|---|---|---|---|---|---|
| 5.99 | 0.94 | 0.67 | 5.03 | 1.73 | 15.24 | 4.49 | 4.34 | 5.64 | 15.58 | 0.37 |
| 5.25 | 1.21 | 0.52 | 6.11 | 3.54 | 17.34 | 2.57 | 5.59 | 5.94 | 15.37 | 1.21 |

STRATUM 3

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 64.18 | 60.80 | 54.14 | 48.06 | 41.32 | 34.53 | 27.32 | 20.41 | 13.93 | 6.87 | 3.42 |
| -0.29 | -1.76 | 0.83 | -0.50 | 1.42 | 0.69 | 0.22 | -0.11 | 0.84 | -0.84 | 0.55 |
| -0.53 | -1.21 | 0.35 | -0.28 | 0.86 | 0.92 | 0.10 | -0.75 | 0.78 | -1.40 | -0.42 |
| -0.49 | -1.09 | 0.25 | -0.29 | 0.81 | 0.86 | 0.04 | -0.78 | 0.73 | -1.38 | -0.34 |
| -0.29 | -1.78 | 0.83 | -0.50 | 1.42 | 0.70 | 0.23 | -0.12 | 0.82 | -0.86 | 0.52 |
| -0.53 | -1.22 | 0.34 | -0.28 | 0.87 | 0.93 | 0.11 | -0.75 | 0.78 | -1.41 | -0.43 |
| -0.50 | -1.11 | 0.25 | -0.29 | 0.82 | 0.86 | 0.04 | -0.78 | 0.73 | -1.39 | -0.35 |

REG 1 results, BIAS*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| -0.28 | -1.73 | 0.79 | -0.60 | 1.63 | 0.78 | 0.64 | 0.13 | 0.91 | -0.58 | 0.89 |
| -0.52 | -1.15 | 0.35 | -0.33 | 0.93 | 1.03 | 0.20 | -0.71 | 0.85 | -1.31 | -0.27 |
| -0.48 | -1.03 | 0.24 | -0.33 | 0.88 | 0.96 | 0.13 | -0.75 | 0.79 | -1.29 | -0.18 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 638.6 | 605.0 | 538.6 | 478.2 | 411.1 | 343.5 | 271.9 | 203.0 | 138.6 | 68.4 | 34.0 |
| 2.87 | 17.51 | 8.26 | 5.00 | 14.17 | 6.84 | 2.23 | 1.10 | 8.34 | 8.37 | 5.48 |
| 5.23 | 12.00 | 3.47 | 2.81 | 8.57 | 9.18 | 1.00 | 7.43 | 7.79 | 13.98 | 4.21 |
| 4.89 | 10.89 | 2.52 | 2.87 | 8.09 | 8.51 | 0.37 | 7.74 | 7.27 | 13.75 | 3.39 |
| 2.88 | 17.66 | 8.23 | 4.98 | 14.10 | 6.93 | 2.30 | 1.23 | 8.19 | 8.56 | 5.20 |
| 5.28 | 12.15 | 3.41 | 2.78 | 8.61 | 9.26 | 1.06 | 7.42 | 7.74 | 14.08 | 4.31 |
| 4.94 | 11.05 | 2.46 | 2.85 | 8.13 | 8.59 | 0.43 | 7.73 | 7.22 | 13.85 | 3.49 |

REG 1 results, rmse*100

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.80 | 17.35 | 7.92 | 6.02 | 16.34 | 7.84 | 6.36 | 1.34 | 9.11 | 5.82 | 8.88 |
| 5.22 | 11.53 | 3.45 | 3.26 | 9.34 | 10.31 | 1.96 | 7.14 | 8.48 | 13.10 | 2.66 |
| 4.80 | 10.35 | 2.42 | 3.29 | 8.82 | 9.59 | 1.31 | 7.50 | 7.90 | 12.85 | 1.78 |

Results for Population S3 by stratum, model UK 2 with heterogeneous scaling between the two stages.

STRATUM 1

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| -0.05 | -0.37 | -0.75 | -1.83 | -3.26 | -0.61 | -0.14 | 2.73 | 0.56 | 2.34 | 1.91 |
| -0.12 | 0.06 | 0.70 | 0.13 | -1.07 | 0.89 | -0.95 | 1.13 | -1.07 | -0.75 | -0.18 |
| -0.12 | 0.06 | 0.61 | -0.07 | -1.36 | 0.59 | -1.20 | 1.01 | -1.07 | -0.56 | 0.04 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.51 | 3.75 | 7.48 | 18.27 | 32.59 | 6.14 | 1.44 | 27.29 | 5.64 | 23.38 | 19.13 |
| 1.18 | 0.63 | 7.00 | 1.33 | 10.72 | 8.91 | 9.48 | 11.31 | 10.68 | 7.55 | 1.85 |
| 1.17 | 0.59 | 6.07 | 0.74 | 13.56 | 5.94 | 12.01 | 10.12 | 10.68 | 5.64 | 0.39 |

STRATUM 2

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| -0.67 | -0.38 | -1.62 | -1.88 | -0.34 | -1.65 | 0.02 | 0.36 | 2.00 | -0.38 | 1.47 |
| -0.42 | 0.09 | -0.16 | -0.63 | -0.20 | -1.58 | 0.30 | -0.59 | -0.64 | -1.58 | 0.06 |
| -0.36 | 0.12 | -0.15 | -0.72 | -0.38 | -1.78 | 0.12 | -0.71 | -0.68 | -1.57 | 0.14 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 6.68 | 3.81 | 16.22 | 18.78 | 3.37 | 16.55 | 0.24 | 3.59 | 19.97 | 3.77 | 14.72 |
| 4.16 | 0.94 | 1.58 | 6.26 | 2.03 | 15.83 | 2.99 | 5.91 | 6.44 | 15.81 | 0.64 |
| 3.59 | 1.23 | 1.52 | 7.19 | 3.75 | 17.82 | 1.19 | 7.10 | 6.75 | 15.69 | 1.39 |

STRATUM 3

Tabled values are BIAS*100 (top block) and RMSE*100 (bottom block)

Estimated Percentile

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| -0.28 | -1.76 | 0.83 | -0.50 | 1.42 | 0.70 | 0.23 | -0.13 | 0.81 | -0.86 | 0.52 |
| -0.52 | -1.21 | 0.35 | -0.28 | 0.87 | 0.93 | 0.10 | -0.75 | 0.78 | -1.43 | -0.44 |
| -0.49 | -1.10 | 0.25 | -0.29 | 0.82 | 0.86 | 0.04 | -0.78 | 0.73 | -1.40 | -0.36 |

rmse*100

| 0.05 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.85 | 17.62 | 8.31 | 5.00 | 14.20 | 7.01 | 2.29 | 1.29 | 8.15 | 8.60 | 5.19 |
| 5.24 | 12.09 | 3.46 | 2.82 | 8.66 | 9.31 | 1.02 | 7.53 | 7.78 | 14.26 | 4.43 |
| 4.90 | 10.98 | 2.52 | 2.88 | 8.18 | 8.63 | 0.40 | 7.84 | 7.25 | 14.03 | 3.61 |