DISSERTATION

OPTIMAL SAMPLING AND MODELING STRATEGIES FOR QUANTIFYING NATURAL RESOURCES OVER LARGE GEOGRAPHICAL REGIONS

Submitted by

Nantachai Pongpattananurak

Department of Forest, Rangeland, and Watershed Stewardship

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2008

COLORADO STATE UNIVERSITY

May 14, 2008

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY NANTACHAI PONGPATTANANURAK ENTITLED "OPTIMAL SAMPLING AND MODELING STRATEGIES FOR QUANTIFYING NATURAL RESOURCES OVER LARGE GEOGRAPHICAL REGIONS" BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Celedonio Aguirre-Bravo

Mohammed A. Kalkhan

Rajiv Khosla

Robin M. Reich, Adviser

Michael J. Manfredo, Interim Department Head

ABSTRACT OF DISSERTATION

OPTIMAL SAMPLING AND MODELING STRATEGIES FOR QUANTIFYING NATURAL RESOURCES OVER LARGE GEOGRAPHICAL REGIONS

This study addresses three important issues related to designing an inventory and monitoring program of natural resources in the State of Jalisco, Mexico; 1) selecting an appropriate modeling approach to describe the spatial variability of selected variables of interest; 2) selecting an appropriate sampling design; and 3) selecting an appropriate plot size and sample size.

Chapter 1 evaluates a new approach of modeling the spatial distribution of soil attributes over large geographical regions. A combination of three-stage least squares (3SLS) and multivariate regression trees (MRT) was used to model the spatial variability in soil texture. In 2006, 1427 soil samples were collected as part of a state-wide inventory and monitoring program (IMRENAT) implemented in the State of Jalisco, Mexico, located in the west central part of Mexico and covers an area approximately 78618 km^2. A two-way nested stratified design was used to allocate samples throughout the state based on the spectral variability of land cover and climatic conditions. Soil samples were collected from five subplots on a 30 m x 30 m primary sampling unit to form a composite surface soil sample (0 – 10 cm depth). The final set of models described 61% of the observed variability in soil pH, 62% of

the variability in sand and 56% for clay. Comparison with other interpolation techniques such as ordinary kriging, suggest that the approach used in this study is far superior in terms of the accuracy and precision.

Chapter 2 evaluates three sampling designs (i.e., simple random sampling, systematic sampling and two-way nested stratified design (IMRENAT)) for modeling the spatial variability in forest tree biomass in the State of Jalisco, Mexico. Normally distributed random errors were added to an existing spatial model of forest tree biomass and used as "truth" in this study. Monte Carlo simulations were used to implement the three sampling designs using samples of 500 and 1100 30 m x 30 m primary sampling units. Statistically, the two-way-nested stratified design outperformed the simple random and systematic sampling design. There was no significant difference between the simple random and systematic designs. The statistical performance of the two-way nested stratified design increased with increasing sample size.

Chapter 3 evaluates the statistical properties of plot size and sample intensities in estimating forest stand characteristics (i.e., tree basal area, tree density and total number of tree species) in seasonal dry evergreen forests in Huai Kha Khaeng Wildlife Sanctuary, Thailand. Monte Carlo simulations were used to evaluate plot sizes (5 m x 5 m, 10 m x 10 m, 20 m x 20 m, 25 m x 25 m and 50 m x 50 m) and sample intensities (0.5%, 1%, 2%, 5%, 10%, and 15%) on a 50 ha mapped dataset. All plot sizes and sampling intensities provided unbiased estimates of the population mean and variance for tree basal area and tree density. All plot sizes and sampling

intensities were biased with respect to estimating the total number of tree species on the 50 ha plot.

Nantachai Pongpattananurak
Department of Forest, Rangeland,
and Watershed Stewardship
Colorado State University
Fort Collins, CO 80523
Summer 2008

ACKNOWLEDGEMENTS

The learning process associated with earning a doctoral degree is not an easy task. Learning how to critically think about scientific questions and how to suggest appropriate solutions proved to be a daunting challenge. Also, writing this dissertation was not an easy process for me. There are a number of people who are involved in the succeed of my study and research. First, I am truly grateful to the Royal Thai Government for their full financial support. In addition, I must give special thanks to the Faculty of Forestry, Kasetsart University for providing me the opportunity to study at Colorado State University.

Second, I would like to thank my graduate advisor, Dr. Robin Reich, who for the past five years has been tremendously patient with me and has assisted through the tedious learning process to accomplish this research. Without his guidance, I would have never been able to complete my program of study. I would also like to give sincere thanks to my graduate committee, Dr. Celedonio Aguirre-Bravo, Dr. Mohammed Kalkhan and Dr. Rajiv Khosla for their guidance, advice and time to review the first draft of my dissertation.

I also would like to thank the support of the Governor of the Mexican state of Jalisco, the Secretary of Rural Development, FIPRODEFO (Fideicomiso para la Administración del Programa de Desarrollo Forestal del Estado de Jalisco), Jalisco's Agiculture Advisory Board, the USDA Forest Service, Rocky Mountain Research Station, Colorado State University, the Universidad de Colima, and the network of governmental institutions and nongovernmental organizations that are partners of the

Consortium for Advancing the Monitoring of Ecosystem Sustainability in the Americas (CAMESA). Thanks to them, Jalisco's Natural Resources Inventory and Monitoring Program (IMRENAT) generated and made available the data used in several sections of this dissertation.

I would alos like to give special thanks to Dr. Sarayudh Bunyavejchewin, my former employer and Dr. Soomboon Kiratiprayoon for their comments, suggestions and help in the field data collection used in Chapter 3. I would like to thank Rungsuriya Buasalee for his help in the field data collection and identifying tree species used in Chapter 3.

Special thanks is also extended to Dick Peterson, my first friend in Fort Collins, who helped me when I first came to Fort Collins. I also appreciate all the help, support and encouragement from my friends in Fort Collins and Thailand. I really appreciate what you have done for me for the past five year.

Lastly, tremendous thanks to my family in Thailand. They inspired me to overcome all the obstacles associated with living aboard for the past five years. Special thanks to my Mom and Dad for their support and helping me get through the good and bad times during my graduate studies. Thanks to all for your love and encouragement.

DEDICATION

I would like to dedicate my dissertation to the most influential women in my life. First, I would like to dedicate my dissertation to my beloved mother, Yupa, who has been waiting her whole life to see me fulfill my dream of completing my doctoral degree. Second, I owe a lot to my wife, Athisha who sacrificed her precious time, shared her life with me, and supported me. Last but not the least, to my daughter, Ananya, you truly become an inspiration in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: MODELING THE SPATIAL DISTRIBUTION OF SOIL ATTRIBUTES AT A REGIONAL LEVEL, A CASE STUDY IN THE STATE OF JALISCO

ABSTRACT

Information on the spatial variability of soil attributes, such as soil texture and pH, play a crucial role in the measurement of forest ecosystems and agricultural lands. Selecting an appropriate technique to spatially interpolate soils data is not straightforward especially when dealing with large geographical regions. In this study, a new approach using three-stage least squares (3SLS) and multivariate regression trees (MRT) was illustrated to model soil texture fractions. Additionally, the comparisons of modeling small scale variability based on 1) a stratified tree-based approach using regression trees (RT) and multivariate regression trees (MRT) and 2) a geostatistical approach using ordinary kriging (OK) are compared and evaluated. The soils data used in this study were obtained from a state-wide inventory implemented in the State of Jalisco, Mexico conducted in 2006 and included 1427 observations on soil texture and pH. The decisions to use three stage least squares and multivariate regressing trees were to ensure the prediction of soil texture fractions summed to 100 percent. Additionally, 3SLS allowed the use of highly correlated dependent variables as explanatory variables in some of the regression models, which violates the underlining assumption of ordinary least squares. The 3SLS models accounted for 30%, 43% and 39% of the variability observed in sand, clay and pH, respectively. The RT models explained an addition 31%, 19% and 6% of

the observed variability in the pH, sand and clay models, respectively. With respect to the 3SLS + RT models, the total observed variability explained for the soil pH, sand and clay models were 61%, 62% and 45%, respectively. The MRT models accounted for 19% and 17% of the observed variability in sand and clay, respectively while the final models (3SLS + MRT) accounted for 62% of the variability in sand and 56% for clay. Meanwhile, ordinary kriging explained only 9% and 17% of the observed variability in sand and clay, respectively. The results also suggest that only the stratified tree-based approach provided unbiased variance estimates for the mean response and new observations. The 3SLS + MRT model satisfied the constraint that the estimated values of the sum of sand, clay and silt summed to 100%, while 3SLS + RT had sums ranging from 82.19% to 121.60%. The stratified tree-based approach provided a more reliable model of soil attributes than ordinary kriging.

INTRODUCTION

Soil texture varies significantly within and across land cover types. In agricultural soils, for example, assessment of the spatial variability in soil texture is central to support a variety of management decision processes. Soil physical properties such as soil texture have a direct effect on water-holding capacity, cation-exchange capacity, crop yield, site productivity, and nitrogen loss, as well as other soil processes and conditions. Numerous statistical techniques have been advocated to describe and interpolate soil properties at the field level (McBratney et al. 2003, Scull et al. 2003).

Historically, modeling of soil attributes has relied primarily on ordinary least squares (i.e., multiple linear regressions) to explain the variability of soil attributes

(Troeh 1964, Walker et al. 1968, Moore et al. 1993, Skidmore et al. 1997). Multivariate techniques have also been used, particularly when dealing with a number of redundant independent variables, e.g., multispectral satellite imagery (McBratney et al. 2003) and/or topographic factors. Discriminant analysis has also been considered as a useful tool for the purposes of predicting soil attributes (Bell et al. 1994, Palvik and Hole 1997, Dobos et al. 2001). Spatial interpolation techniques have also been used to describe the spatial dependency in soil attributes. For example, Odeh et al. (1994, 1995), Knotters et al. (1995), De Gruijter et al. (1997) and Voltz et al. (1997) studied soil depth, and other soil properties using geostatistical techniques such as kriging and co-kriging. Additionally, universal kriging which, combines low order degree polynomials of geographical coordinates (i.e., trend surface analysis) and ordinary kriging have been evaluated for predicting a wide range of soil properties (Odeh et al. 1994, Meul and Van Meivernne 2003). Gotway-Crawford and Hergert (1997) and Meul and Meirvenne (2003) provide comprehensive examples of how to handle spatial soil attribute data without the assumption of stationarity. Generalized linear models have also been used to model and map soil attributes (McKenzie and Austin 1993, Gessler et al. 1995). More recently, classification and regression trees (CART) developed by Brieman et al. (1984) have been used by soil scientists as a predictive model to evaluate either continuous (Ryan et al. 2000, Henderson et al. 2005) or categorical (Bui and Moran 2001) soil attributes.

Most of the previous studies have concentrated on modeling soil attributes over small areas with a fine spatial resolution (Knotters et al. 1995, Odeh et al. 1994, Ryan et al. 2000). On the other hand, when modeling soil attributes over large geographical regions, soil scientists have focused primary on a coarse spatial resolution (Bui and

Moran 2001, Dobos et al. 2001, Henderson et al. 2005). Spatial models of soil attributes with a fine scale resolution have rarely been done over large geographical regions.

One approach of modeling spatial data is to decompose the data into two components, the large-scale and the small-scale variability. The large-scale variability in soil attribute may be influenced by such factors as elevation, slope, aspect, precipitation, and so on, while the small-scale variability is potentially influenced by differences in soil permeability, nutrient availability and so forth. The large-scale variability is generally modeled using multiple regression models, while the small-scale variability is modeled using geostatistical techniques such as kriging. Unfortunately, when trying to model soil attributes over large geographical regions, the data may not be spatially correlated, or weakly correlated making it almost impossible to model the small-scale variability in a set of data using geostatistical methods.

While the residuals from the regression models may not display any spatial dependency, they still contain important information useful in describing the spatial variability in a set of data. Reich and Aguirrie-Bravo (2008) introduced the concept of a tree-based stratified design capable of modeling the small-scale variability in a set of spatially independent data.

Another problem that arises when modeling soil attributes are implied constraints. For example, soil texture is probably one of the most common attributes modeled by soil scientists. Modeling efforts have concentrated on dealing with only one or two components of soil texture (e.g., sand, clay, or both sand and clay). If the third component is desired it is often obtained by subtraction. Recently, Van Meirvenne and

Van Cleemput (2006) employed compositional ordinary kriging to simultaneously model soil texture while constraining the three fractions of soil texture to sum to 100%.

In this paper, a new approach is presented for modeling soil texture fraction over large geographical regions based on a system of equations to ensure the three fractions of soil texture sum to a 100%. The method of three-stage least square (3SLS) is used to model the large-scale variability in soil texture (Zellner and Theil 1962). This approach allows one to statistically constrain the model such that the estimates of sand, silt and clay sum to 100%. In the case of modeling the small-scale variability, it is not clear what type of regression trees should be used. Thus, the main objective of this study was to evaluate the use of univariate and multivariate regression trees (De'ath 2002) in modeling the small scale variability in soil texture using the tree-based stratified approach advocated by Reich and Aguirrie-Bravo (2008). A secondary objective was to compare the use of ordinary kriging in modeling the small-scale variability with the use of the tree-based stratified approach. These methods are illustrated by modeling selected soil attributes in the Mexican State of Jalisco.

METHODS

Study Area

The State of Jalisco, Mexico is located in the west central part of Mexico, and covers an area of approximately 78618 km^2 (Figure 1). Four major ecological regions provide the natural resources and environmental conditions that make this region one of the most prosperous in Mexico. The eco-regions consist of: 1) the transversal neo-volcanic system, 2) the southern Sierra Madre, 3) the Southern and Western Pacific Coastal Plain and Hills and Canyons and 4) the Mexican High Plateau. Linked to these ecological regions are several important hydrological regions that drain to the Pacific Ocean (Lerma-Santiago, Huicicila, Ameca, Costa de Jalisco, Armeria-Coahuayana, Balsas, and El Salado). Elevations range from sea level to 4236 m.

Soil Data

In 2006, 1427 soil samples (Figure 1) were collected as part of a state-wide inventory and monitoring program (IMRENAT) implemented in the State of Jalisco, Mexico. A two-way nested stratified sampling design (Reich et al 2008) was used to allocate samples throughout the state based on the spectral variability of land cover and climatic conditions. Soil samples were collected from five subplots on a 30 m x 30 m primary sampling unit to form a composite surface soil sample (0 – 10 cm depth). Soil samples were analyzed to obtain some basic soil physical and chemical properties, including soil texture, soil depth and pH. Soil pH was estimated using a chemical measure of soil buffering, or the SMP buffer method (McLean 1982) while percent of

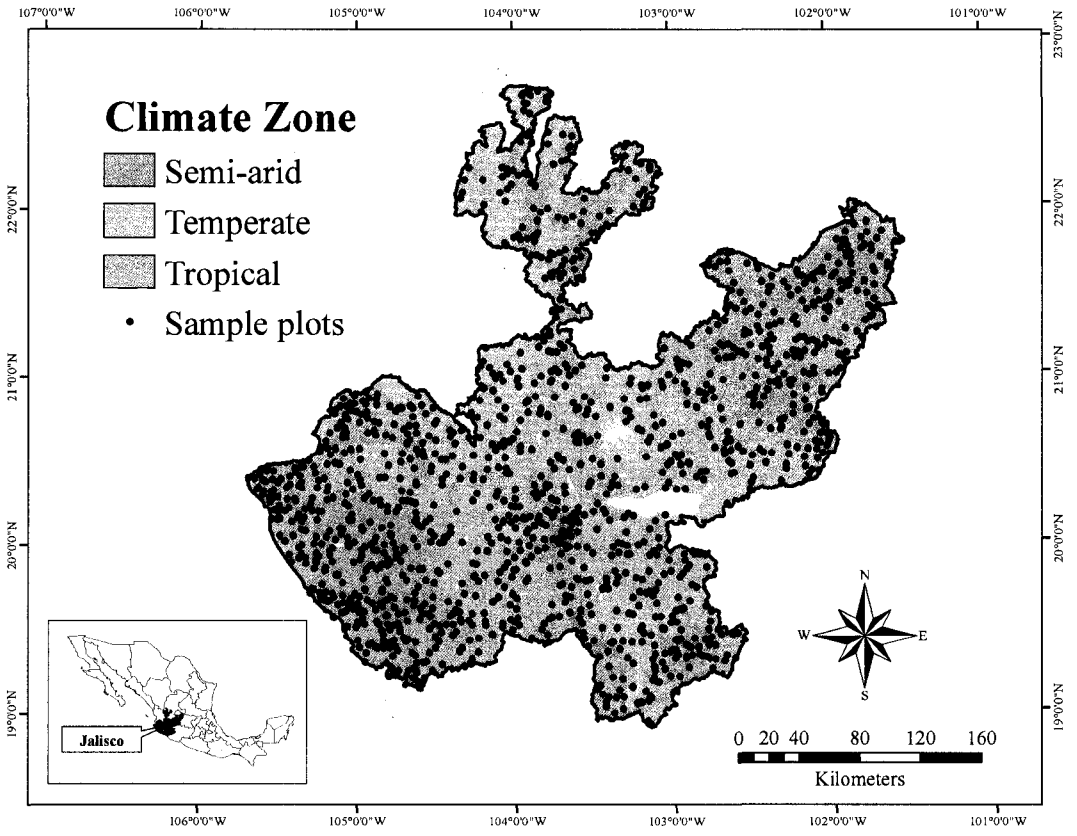


Figure 1. Locations of 1427 sample plots in the State of Jalisco, Mexico.

sand, clay and silt were determined using Bouyoucos hydrometer method (Bouyoucos 1936, Gee and Bauder 1979, Jones 2002). The soil textural classification and frequency of the observed soil data is given in Table 1.

GIS and Landsat TM Data

Ten cloud-free Landsat 7 ETM+ images obtained between January and March, 2004 were combined to create a seamless image using the *Mosaic* tool (ERDAS Inc. 1999). The thermal band 6L and 6H with a 57 m resolution and the panchromatic band 8 with a 14.25 m resolution were resampled to a 30 m resolution. A digital elevation model (DEM) with a 90 m spatial resolution, obtained from the U.S. Geological Survey (USGS) (Gesch et al. 2002, Rabus et al. 2003) was resampled to a 30 m spatial resolution using the *Resample* function with the *Bilinear* option (ARC/INFO, ESRI 1995) to correspond to the spatial resolution of the satellite imagery. The primary topographic attributes of elevation, aspect, and slope were derived from the DEM using *Spatial Analyst* tool (ARCGIS 9.1, ESRI 2005). In addition, a GRID layer of 12 climate zones (Reich et al. 2008) with a 30 m spatial resolution was incorporated as an additional covariate. All GIS analyses was included in ArcGIS 9.1 (ESRI 2005).

Modeling Soil Texture and pH

Large-Scale Variability

In the first step of the modeling process, ordinary least square (OLS) was used to identify the functional form of the regression equations for describing the large-scale variability in clay, sand, and pH. The *stepAIC* function, available in the *MASS* Package

Table 1. Soil textural classification of the soil samples (n = 1427) collected in the State of Jalisco Mexico.

Soil Textural Class	**Number of Samples**	**Percent**
Sand	16	1.12
Loamy Sand	123	8.62
Sandy Loam	767	53.75
Loam	147	10.30
Silt Loam	2	0.14
Sandy Clay Loam	252	17.66
Clay Loam	69	4.84
Silty Clay Loam	1	0.07
Sandy Clay	9	0.63
Clay	41	2.87
Total	1427	100

(Ripely 2008) in R (R Development Core Team 2006), was used to perform a backward stepwise selection procedure identifying significant predictors of each variable based on Akaike Information Criterion (AIC) (Akaike 1969). This process identified the following functional form of the models for sand, silt, clay and pH:

$$Sand = f(elev + slp + asp + czone + bands)$$

$$Clay = f(sand + elev + slp + asp + czone + bands)$$

$$Silt = f(100 - (sand + clay))$$

$$pH = f(clay + elev + slp + asp + czone + bands)$$

where, elev = elevation, slp = slope, asp = aspect, czone = climatic zone, bands = Landsat-7 ETM+ bands. An important characteristic of this system of equation is the presence of dependent variables on the right hand side of three of the four equations. This clearly violates the underlying assumption of the OLS model. To address this issue, the system of equations were fit using three-stage least square (3SLS).

The 3SLS approach combines two-stage least squares (2SLS) with seemingly unrelated regression (SUR). Two-stage least square is a method of using dependent variables as independent variables on the right-hand side of a regression model, while SUR is a technique for fitting a system of equations with cross-equation parameter restrictions and correlated error terms (Zellner and Theil 1962). The soil texture model contains three equations, which are seemingly dependent on one another. However, if the equations are using the same covariates, the errors obtained from OLS may be correlated across equations. Thus, rather than estimating the system equations individually by least squares, the method of SUR is applied.

To describe the 3SLS approach, let y_i represent a vector of sample values of the response variable *Y*. The system of equations can be described by

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & Z_j \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_j \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_j \end{bmatrix} \tag{1}$$

or in matrix notation

$$\underline{y} = Z\underline{\delta} + \underline{\eta}$$

where Z_j is a design matrix of predictive variables for the j^{th} linear model, including jointly dependent variables among the response variables (e.g. sand and clay), δ_j is a vector of estimated coefficients associated with the j^{th} linear model and η_j is the vector of independent residuals, or errors associate with the j^{th} linear model.

The estimated coefficients for 3SLS model (Zellner and Theil 1962, Greene 1990) is given by:

$$\hat{\delta}_{3SLS} = \left[\hat{Z}'(\Sigma^{-1} \otimes \mathrm{I})\hat{Z}\right]^{-1} \hat{Z}'(\Sigma^{-1} \otimes \mathrm{I})\underline{y} \tag{2}$$

where

$$\hat{Z} = \begin{bmatrix} X(X'X)^{-1}X'Z_1 & \cdots & 0 \\ \vdots & X(X'X)^{-1}X'Z_2 & \vdots \\ 0 & \cdots & X(X'X)^{-1}X'Z_j \end{bmatrix},$$

X is a design matrix of all independent variables for sand, clay, silt and pH excluding jointly dependent variables, Z_j is a matrix of predictive variables for the j^{th} linear model

including jointly dependent variables, I is an identity matrix, $\otimes$ [1] represents the Kronecker product, and Σ is the variance-covariance matrix among response variables obtained from 2SLS. Individual entries of the variance-covariance matrix are estimated as follows:

$$\hat{\sigma}_{ij} = \frac{(y_i - Z_i\hat{\delta}_{i,2SLS})'(y_j - Z_j\hat{\delta}_{j2SLS})}{(n_i - p_i - 1)(n_j - p_j - 1)} \tag{3}$$

where, y_i is the observed values of the i^{th} response, y_j is the observed value of the j^{th} response, $\hat{\delta}_{i,2SLS}$ is a vector of the estimated parameters from 2SLS for the i^{th} response, $\hat{\delta}_{j,2SLS}$ is a vector of estimated parameters from 2SLS for the j^{th} response, n_i is the sample size of i^{th} response, n_j is the sample size of j^{th} response, p_i is the number of parameters estimated for the i^{th} response and p_j is the number of parameter estimated for the j^{th} response.

The asymptotic variances-covariance matrix for the estimated regression coefficients is given by

$$Var[\hat{\delta}_{3SLS}] = \left[\hat{Z}'(\Sigma^{-1} \otimes \mathrm{I})\hat{Z}\right]^{-1}. \tag{4}$$

The estimated variance an estimate $\hat{y}_j$ is given by

$$Var(\hat{y}_j) = MSE_j\left(Z_j' Var[\hat{\delta}_j] Z_j\right) \tag{5}$$

---

[1] $a = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $a \otimes I = \begin{bmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & c & d \end{bmatrix}$

where MSE_j is the mean square error associated with the j^{th} covariates obtained from Eq. 3, $Var[\hat{\delta}_j]$ is the variance-covariance matrix of 3SLS coefficients associated with the j^{th} response, and Z_j is a design matrix of the j^{th} covariates including jointly dependent variables for a given response j.

The estimated variance associated with the prediction of a new observation is given by

$$Var(\hat{y}) = MSE_j\left(1 + S_j' Var[\hat{\delta}_j] S_j\right) \tag{6}$$

where $\hat{y}$ is the predicted value of the new observation and S_j is a design matrix of covariates including jointly dependent variables associated with the new observation.

Small-Scale Variability

Several studies (Odeh et al. 1995, Erxleben et al. 2002, Reich et al. 2008, and Reich et al. 2008) have shown that the small-scale variability in a set of data can be described by modeling the residuals obtained from a multiple linear regression model. Several approaches have been suggested to account for the small-scale variability in a set of data. For soil modeling, ordinary kriging (OK) is a well known technique to interpolate soil variables when the sample data has a strong degree of spatial dependency. However, Reich and Aguirrie-Bravo (2008) emphasized that data collected over large geographical regions generally lack spatial dependency because of large separation distances between sample points. Thus, geostatistical methods such as kriging may not be appropriate for describing the small-scale spatial variability in such data. Instead, the residuals obtained from modeling large-scale variability (i.e., a multiple linear regression model) can be modeled using binary regression trees. Therefore, three statistical

approaches including binary regression tree, multivariate regression tree and ordinary kriging were evaluated in modeling the small-scale variability associated with the residuals from the 3SLS model.

Tree-based approach

To describe the small-scale variability associated with the residuals obtained from the 3SLS model, both binary regression trees (RT) and multivariate regression trees (MRT) (De'ath 2002, 2007) were evaluated in this study. Binary regression trees is a nonparametric and non-linear regression procedure in which the data is repeatedly and successively split along a set of independent variables using a binary algorithm to maximize variances among groups of the dependent variable (Breiman, et al. 1984, Chamber and Hastie 1992, Venables and Ripley 1999, Crawley, 2002).

Reich and Aguirrie-Bravo (2008) suggested a new approach of modeling the small-scale variability using a tree-based stratified design. Let $y(s_i)$ represent a sample value of the variable Y at spatial location s_i. The sample data include a set of covariates X which are known for all locations in the population. Multiple regression is used to model the large-scale spatial variability in the sample data as a linear function in p known explanatory variables $x_j(s_i)$

$$y(s_i) = \beta_0 + \sum_{j=1}^{p} x_j(s_i)\beta_j + \eta(s_i) \qquad (7)$$

where $\beta_j, j = 0,\ldots p$ are p+1 unknown regression coefficients and $\eta(s_i)$ is an error process sometimes referred to as a random field, with $E[\eta(s_i)] = 0$ and covariance $C(x,y) = Cov(\eta(x_i), \eta(y_j))$. The error term in Eq. 7 is unknown because the true model is unknown. Once the model parameters have been estimated, the regression residuals are

defined as $\hat{\eta}(s_i) = y(s_i) - \hat{y}(s_i)$, where $\hat{y}(s_i)$ is the predicted value at spatial location s_i given the explanatory variables $x_j(s_i)$. The error process can be expressed as

$$\eta(s_i) = \hat{\eta}(s_i) + \mu(s_i) \tag{8}$$

with $E[\mu(s_i) | \hat{\eta}(s_i)] = 0$. Using the set of auxiliary variables, X as a basis of stratification assume

$$\hat{\eta}(s_i) = f(x(s_i)) + \delta(s_i) \tag{9}$$

with $E[\delta(s_i) | x(s_i)] = 0$, $f(x(s_i))$ is a deterministic function, and $\delta(s_i)$ is a zero-mean random term (Cocchi et al. 2002). Combining Eq. 8 and Eq. 9

$$\eta(s_i) = f(x(s_i)) + \varepsilon(s_i) \tag{10}$$

with $E[\eta(s_i) | f(x(s_i))] = f(x(s_i))$, $\varepsilon(s_i) = \mu(s_i) + \delta(s_i)$, and $E[\varepsilon(s_i) | x(s_i)] = 0$ provided that $\mu(s_i)$ and $\delta(s_i)$ are conditionally independent (Cocchi et al. 2002, Benedtti et al., 2005). The mean function $f(x(s_i))$ is estimated by $\hat{f}$ using the recursive partitioning method introduced by Brieman et al. (1984). Combining Eq. 7 and Eq. 10 the full model describing the spatial variability in the sample data is given by

$$y(s_i) = \beta_0 + \sum_{j=1}^{p} x(s_i)\beta_j + f(x(s_i)) + \varepsilon(s_i). \tag{11}$$

To implement the tree-based stratified design, RT is applied to the residuals from the 3SLS model. The algorithm for RT repeatedly partitions the residuals into strata to minimize the variability within strata (Breiman et al 1984). The recursive procedure determines a split starting from a single stratum containing all residuals, and ending once the sample data are split into new strata which minimize the variability within strata. Unlike RT, MRT simultaneously partitions the residuals from all the models in the 3SLS

model. This ensures that the residuals from the individual models will sum to zero using the same tree-based algorithm.

Defining the tree structure

To control the partitioning of the regression trees (RT and MRT) several parameter have to be defined. The parameter *minsplit* (or called *minsize* in the function *tree.control( )* in R) defines the number of observations (i.e., stratum size) at which the last split is attempted. The default value, *minsplit* = 5 means that the recursive partition keeps continuing to allocate observations into strata (i.e., terminal nodes) as long as there are at least five observations at a given node. Changing the parameter *minsplit* directly affects the maximum number of strata or terminal nodes and the path length of the tree (called tree size). The maximum number of strata has an upper bound ~*nobs*/*minsplit*, where *nobs* is the number of observations in the data set. The best optimal condition to minimize the cost complexity (see more details in the model evaluation session) was identified using different *minsplit* options (i.e., 5, 10 and 25). After obtaining the optimal number of terminal nodes based on given value for *minsplit*, the function *prune.rpart( )* is used to prune the tree by changing the argument *best* to the optimal number of terminal nodes or strata (i.e., strata size or tree size).

Ordinary kriging

Ordinary kriging is a common method used to interpolate spatially dependent data (Isaaks and Srivastava 1989, Fortin and Dale 2005). The ability to interpolate a set of data depends on the strength of the spatial dependency within the data. To evaluate the

feasibility of using kriging to interpolate the residuals from the 3SLS models, sample variaograms were constructed to describe how the variance changes with distance.

The sample variogram is defined by

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (y_i - y_j)^2 \tag{12}$$

where *N(h)* is the number of data pairs separated by distance h, y_i is the observed value at location s_i and y_j is the observed value at location j. The sample variogram models were fit to three theoretical variogram models: exponential, spherical, and Gaussian. Akaike Information Criterion (AIC) (Akaike 1969) was used to identify the best fitting variogram model. The spatial library for R created by Reich and Davis (2007a) was used to perform this operation.

Ordinary kriging (Isaaks and Srivastava 1989, Webster and Oliver 2001) can be used to estimate a value of interest at any location as a weighted combination of its neighbors:

$$\hat{y}_0 = \sum_{i=1}^{n} \lambda_i z_i \tag{13}$$

where $\hat{y}_0$ is the estimated value at a new location s_0, y_i is the observed value of the i^{th} neighbor and λ_i are the estimated weights, subject to the constraint $\sum_{i=1}^{n} \lambda_i = 1$.

The weights λ_i are calculated using the relationship

$$\underline{\lambda} = K^{-1}\underline{C} \tag{14}$$

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} C(x_1,x_i) & \cdots & C(x_1,x_n) & 1 \\ \vdots & \ddots & \vdots & 1 \\ C(x_n,x_1) & \cdots & C(x_n,x_n) & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} C(x_1,x_0) \\ C(x_2,x_0) \\ \vdots \\ C(x_n,x_0) \\ 1 \end{bmatrix}$$

where μ is a Lagrange multiplier, K is the covariance matrix among the sample data points, $\underline{C}$ is the vector of covariances between the point being estimated and its neighbors. The covariance were computed using the relationship $C(h) = \gamma(h) - \sigma^2$, where σ^2 is the sample variance of the sample data.

Variance Estimation

A common problem in using regression analysis to describe the variability in spatial data is overdispersion, in which the observed variability exceeds the variability predicted by the model. This leads to inaccurate estimates of standard errors and coverage rates that are not equal to the 0.95 nominal rate. To ensure that the variance estimates are consistent with the true errors, both RT and MRT were used to model the variability in the residuals from the 3SLS model.

The *rpart* Package (Ripley 2007) and the *mvpart* Package (De'ath, 2007), libraries in R, were utilized to perform univariate partitioning and multivariate partitioning of the 3SLS residuals, respectively. Using the function *rpart( )*, the splitting rule with the default *method* = "*anova*" is used to minimize the residual sum of squares associated with a terminal nodes or stratum.

Reich and Aguirrie-Bravo (2008) showed that the variance of the estimated mean response at a given location s_i, for a set of explanatory variables, $x(s_i)$ can be defined as

$$\mathrm{var}(y(s_i)) = \mathrm{var}(\hat{\eta}(s_i)) + \mathrm{var}(\hat{\delta}(s_i)), \tag{15}$$

where $\mathrm{var}(\hat{\eta}(s_i))$ represents the uncertainty in estimating the parameters of the 3SLS model and $\mathrm{var}(\hat{\delta}(s_i))$ reflects the uncertainty in estimating the error ($\hat{\eta}$) of the 3SLS model using RT or MRT. The variance $\mathrm{var}(\hat{\eta}(s_i))$ was computed using Eq. 5.

The variance associated with an estimate at a new location, s_0 is given by

$$\mathrm{var}(y(s_0)) = \mathrm{var}(\hat{\eta}(s_0)) + \mathrm{var}(z(s_0)) + \mathrm{var}(\hat{\delta}(s_0)), \tag{16}$$

where the additional term, $\mathrm{var}(z(s_0))$ reflects the random variation at a new location, s_0.

The uncertainty in estimating the residual of the 3SLS, $\mathrm{var}(\hat{\delta}(s_i))$, based on RT or MRT were determined by standard methods for a stratified random sample (Cochran, 1977):

$$\hat{\bar{\delta}}_k = \frac{1}{n_k}\sum_{l=1}^{n_{ki}} \hat{\delta}_{ki} \tag{17}$$

and

$$\hat{\sigma}_k^2 = \frac{\sum_{l=1}^{n_k}\left(\hat{\delta}_{ki} - \hat{\bar{\delta}}_k\right)^2}{n_k - 1} \tag{18}$$

where k denotes the stratum, $\hat{\delta}_{ki}$ denotes the mean residual error for observations assigned to the k^{th} stratum, n_k is the number of observations assigned to the k^{th} stratum, and $\hat{\sigma}_k^2$ is the within stratum variance for the k^{th} stratum. The use of the sample variance as a measure of the uncertainty in estimating the error of the regression model is justified by the fact that the mean square error is the best predictor of the variance given that the sample data belong to the stratum.

For ordinary kriging, an estimate of the prediction variance $\text{var}(y(s_i))$ at a given location s_i (Isaaks and Srivastava 1989, Reich and Davis 2007b) is given by:

$$\text{var}(y(s_i)) = \text{var}(\hat{\eta}(s_i)) + \hat{\sigma}^2_{i(OK)} \tag{19}$$

where

$$\hat{\sigma}^2_{i(OK)} = \sigma^2 - C' K^{-1} C + \mu. \tag{20}$$

Model Evaluation

A 10-fold cross validation (Efron and Tibshrani, 1993) was used to evaluate the predictive performance of the fitted models of soil attributes using 3SLS + RT, 3SLS + MRT and 3SLS + OK. The soils data were divided into 10 parts (K = 10), each of which consisted of approximately 140 sample plots. The predictive models were recursively fitted using nine parts (K-1) of the data as a training data set and the remaining data were treated as an independent dataset for estimating prediction errors. Repeating this procedure 10 times allowed each observation to be excluded from the model and independently predicted by the fitted models. Following this procedure, a set of statistics were calculated to evaluate the predictive performance of the models. Estimates of the prediction errors obtaining from the K-fold cross validation (Kravchenko and Bullock 1999, Schloeder et al. 2001, Reich et al. 2004) were compared to asses the effectiveness of three techniques for modeling the selected soil attributes. The prediction errors were inferred from the predicted minus actual values.

In this study, AIC was used to select the covariates used in the regression models of sand, clay and pH. The identified covariates were fixed while performing the 10-fold cross validation. For the tree-based approach, the effectiveness of the models using

different conditions of the function "*minsplit*" (i.e., to control the number of observations before the last split), and the function "*best*" (i.e., to control the number of terminal nodes) for different soil attributes were evaluated and compared via a 10-fold cross validation. For simplifications, *minsplit* is referred as *minsize* to avoid redundant terminology. Several statistics were calculated to evaluate the predictive performance of the models.

The effectiveness of the fitted models was evaluated using a goodness-of-prediction statistic (G-statistic) (Agterberg 1984, Kravchenka and Bullock 1999, Guisan and Zimmermann 2000, Schloeder et al. 2001):

$$\text{G-statisic} = 1 - \left(\frac{\sum_{i=1}^{n} \left[y(s_i) - \hat{y}(s_i) \right]^2}{\sum_{i=1}^{n} \left[y(s_i) - \bar{y}(s_i) \right]^2} \right). \tag{21}$$

The G-statistic is a measure of the effectiveness of a prediction relative to that which could have been derived using the sample mean. A G-statistic equal to one indicates perfect prediction, a positive value indicates a more reliable model than if one had used the sample mean, a negative value indicates a less reliable model than if one had used the sample mean, and a value of zero indicates that the sample mean should be used to estimate $y(s_i)$.

The mean absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y(s_i) - \hat{y}(s_i) \right| \tag{22}$$

and the root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[y(s_i) - \hat{y}(s_i) \right]^2} \tag{23}$$

were used to evaluate the accuracy of the predictions, where $y(s_i)$ is the observed value at a sample location s_i, $\hat{y}(s_i)$ is the estimated value at a sample location s_i obtained from the 10-fold cross validation, and n is the total number of samples used in the 10-fold cross validation.

To evaluate the ability of the three approaches in providing unbiased variance estimates, the standardized mean squared error (SMSE) (Reich et al. 2004) was calculated as follows:

$$SMSE = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varepsilon}^2(s_i)}{\mathrm{var}(\hat{y}(s_i))} \tag{24}$$

where $\hat{\varepsilon}(s_i) = (y(s_i) - \hat{y}(s_i))$, is the true error and $\mathrm{var}(\hat{y}(s_i))$ is the estimated variance obtained using Eq. 15 or Eq. 16 for 3SLS plus tree-based stratified design and from Eq. 20 for 3SLS + OK. The SMSE has a Chi-square distribution with n degree of freedom. A $1-\alpha$ confidence interval for SMSE under the null hypothesis of equal variances can be constructed using the Chi-square distribution as follows:

$$\Pr\left[\frac{\chi^2_{n,\frac{\alpha}{2}}}{n} \le SMSE \le \frac{\chi^2_{n,1-\frac{\alpha}{2}}}{n}\right] = 1-\alpha\,. \tag{25}$$

When n is large, SMSE can be approximated by a standard normal distribution with a mean of one and variance $2/n$ (SMSE $\sim N(1, 2/n)$). If the SMSE falls within the interval $1 \pm 1.96*(2/n)^{0.5}$, this would indicate that the true errors and estimation errors are consistent at the 0.95 level confidence (Hevesi et al 1992, Reich et al. 2004). Bonferroni joint confidence intervals were also constructed to make inference among the SMSE's associated with the system of equations: $1 \pm t(1-\alpha/2g)*(2/n)^{0.5}$, where g is the number of

confidence intervals being compared, t is the Student's t value, with a type I error equal to α.

Prediction coverage rates (CR) were computed assuming 95% confidence interval. The coverage rate indicates the proportion of individual intervals containing the true value. If the 95% prediction interval is unbiased, the coverage rate should not deviate from the nominal 0.95 rate. The 95% prediction interval was calculated assuming normality:

$$\hat{y}(s_0) \pm 1.96\sqrt{\text{var}(\hat{y}(s_0))}. \tag{26}$$

Coverage rates for the mean response were also computed assuming normality.

A decision rule (Reich and Aguirrie-Bravo 2008) to identify an optimal *minsize* and strata sizes was used to ensure that the error in estimating the variance of the mean response and the prediction variances were unbiased. The cost complexity criterion (CC) served as a decision rule is selecting *minsize* and strata sizes and is defined as

$$CC = \sqrt{(SMSEM - 1)^2 + (SMSEP - 1)^2} + \frac{MSEP}{df - n} \tag{27}$$

where SMSEM is the standardized mean square error of the variance for the mean response and SMSEP is the standardized mean square error for the prediction variance, MSEP is the mean squared error of prediction obtained from the 10-fold cross-validation, *df* is the degrees of freedom of the 3SLS model and n is the number of terminal nodes or strata sizes in the RT or MRT model and n is the number of neighbors used to estimate values at a given location using ordinary kriging.

Once the optimal criteria (i.e., *minsize* and strata sizes) to obtain unbiased estimates of the variances for RT and MRT were identified, subsets of the data were used to fit the system of equations using 3SLS. In spite of the 2-way nested stratified

sampling design (Reich et al. 2008) used to collect the data, the soil samples did not have a uniform distribution with respect to the soil textural classes. The soil data was sub-sampled to avoid overweighting the soil textural classes with the "sandy loam" and "sandy clay loam" classes which included 767 and 252 observations, respectively (Table 1). Each subset of the data contained the data from all soil textural classes except for the two dominant soil textural classes. To increase the efficiency of the regression models, 150 samples were randomly drawn from these two soil textural classes and added to each subset of the data. As a result, each subset of data included 713 samples and used to obtain parameter estimates for the 3SLS model. This process was repeated until the regression coefficients converged for all soil attributes. These averaged coefficients served as parameter estimates to generate the final surfaces describing the large-scale variability in sand, silt, clay and soil pH.

For model evaluation, the average 3SLS coefficients were applied to only one random subset of the data to generate a set of residuals. The 10-fold cross validation procedure was used to evaluate the performance of the tree-based approach (RT and MRT) using different conditions for the minimum stratum size (*minsize*) and number of strata (i.e., *best*). The optimal *minsize* and tree size were selected to minimize the bias associated with estimating the variance (SMSE) while minimizing the cost complexity criterion (CC). The identified optimal *minsize* and tree size were applied to 100 random subsets of the data (n = 713) to evaluate the variability among different random subsets of the data. Subsets of the data were drawn using sampling without replacement with fixed number of observations for each soil classes as mention previously. To make the

statistics comparable, all soil attributes used the same subset of the data by specifying a random seed.

Soil Attribute Classification and Mapping

The best fitting 3SLS + MRT, 3SLS + RT and 3SLS + OK models were used to generate the final surfaces of selected soil attributes. The grid surfaces of the small-scale variability of sand, silt, clay and soil pH were produced using the conditional statements obtained from MRT and RT using the raster calculator in ArcGIS 9.1. The3 SLS surfaces and the residual surfaces generating from RT, MRT and OK for the soil attributes were combined to form the final surfaces of sand, silt, clay and pH as a GRID layer in ArcGIS 9.1. Finally, the estimated surfaces of soil texture fractions were used to classify each pixel in the State of Jalisco into one of 12 soil textural classes (Soil Survey Division Staff 1993) and plotted on a soil triangle (Oom and Lemon 2005). The final surface of soil pH obtained from 3SLS + RT was classified into four different classes based on the degree of acidity and alkalinity (adapted from Jones 2002). Furthermore, the variance surfaces associated with the predicted soil attributes were developed as GRID layers.

RESULTS

The large-scale variability of soil texture and pH were modeled as a system of equations using 3SLS. Results of fitting the system of equations using 100 random subsets of the data indicated that the model for sand had the best fit in that the model accounted for 43% of the variability observed in the data, followed by clay (39%) and then soil pH (30%) (Table 2).

The residuals from the 3SLS model were modeled using a stratified tree-based approach. The influence of using different *minsize* and tree size on the fit statistics for the pH, sand and clay models are provided in Figures 2 through 4. The optimal conditions of *minsize* and tree size that minimized the cost complexity criteria for RT are given in Table 2. The sand model had an optimal *minsize* of 25 and tree size of 40, while the clay model had a *minsize* of 10 and tree size of 5. In constrast, the soil pH model had a *minsize* of 5 and tree size of 50. Applying the optimal *minsize* and tree size to the 100 random subsets of the data, the RT models accounted for an addition 31% of the observed variability in soil pH, 19% for the sand model and 6% for the clay model. The results also suggested that standardized mean square errors (SMSEM and SMSEP) were not significantly different from one (Table 2) indicating the variance estimates were consistent with the true errors. The prediction coverage rates for all three models were not significantly different from the nominal coverage rate of 0.95. The total observed variability explained by 3SLS + RT for the soil pH, sand and clay models were 61%, 62% and 45%, respectively (Table 2).

For the MRT models, optimal values for *minsize* and tree size for the sand and clay models are shown in Figures 5 and 6. The cost complexity criterion indicated that a *minsize* of 25 and the tree size of 41 provided the best estimates of the variances for sand,

Table 2. Fit statistics for the spatial statistical models of selected soil attributes based on 100 simulations of a random subset of the complete data set.

Model	Soil Attribute	Sample Size	*min-size*[ζ]	tree size[ζ]	R^2[ζ] (3SLS)		G-statistic[ζ]		SMSEM[ζ]		SMSEP[ζ]		MSEP[ζ]		CRM[ζ]		CRP[ζ]		CC[ζ]
					Mean	SD[ζ]	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
3SLS + RT	**Sand**	713	25	40	0.43	0.008	0.62	0.016	0.85	0.007	0.82	0.028	216.05	8.302	0.97*	0.004	0.96	0.005	0.56
	Clay	713	10	5	0.39	0.012	0.45	0.028	0.98	0.003	1.00	0.039	114.56	4.503	0.95	0.004	0.95	0.006	0.22
	pH	713	5	50	0.30	0.016	0.61	0.016	0.86	0.007	1.01	0.051	1.04	0.060	0.97*	0.005	0.95	0.007	0.14
3SLS + MRT	**Sand**	713	5	12	0.43	0.008	0.52	0.028	0.89	0.008	0.67*	0.027	191.48	7.951	0.96	0.004	0.98*	0.004	0.65
	Clay	713	5	12	0.39	0.012	0.46	0.027	0.97	0.003	1.04	0.058	114.78	4.195	0.96	0.004	0.95	0.006	0.23
	Sand	713	25	41	0.43	0.010	0.62	0.022	0.85	0.010	0.84	0.042	218.80	10.690	0.97*	0.005	0.96	0.006	0.58
	Clay	713	25	41	0.39	0.012	0.56	0.023	0.96	0.003	1.40*	0.085	130.85	6.268	0.96	0.005	0.91*	0.009	0.62
3SLS + Kriging	**Sand**	713	–	–	0.43	0.008	0.52	0.034	2.05*	0.192	0.68*	0.028	184.92	7.082	0.91*	0.010	0.98*	0.003	1.36
	Clay	713	–	–	0.39	0.012	0.56	0.014	2.49*	0.247	1.18*	0.057	109.62	3.997	0.89*	0.010	0.94	0.006	1.66

[ζ]*minsize* = a minimum number of observations in the terminal nodes of regression tree, tree size = a number of strata or terminal nodes in the tree, R^2 = coefficient of determination obtained from 3SLS, G-statistic = the total variability of the mean response accounted by the model, SMSEM = standardized mean square error of the mean response, SMSEP = standardized mean square error of the prediction, MSEP = mean square error of the prediction, CRM = coverage rate of the mean response, CRP = coverage rate of the prediction, CC = value of the cost complexity factor, SD = standard deviation.

*Significantly different from 1 for SMSEM and SMSEP and significantly different from 0.95 for CRM and CRP at the 0.05 level of significance.

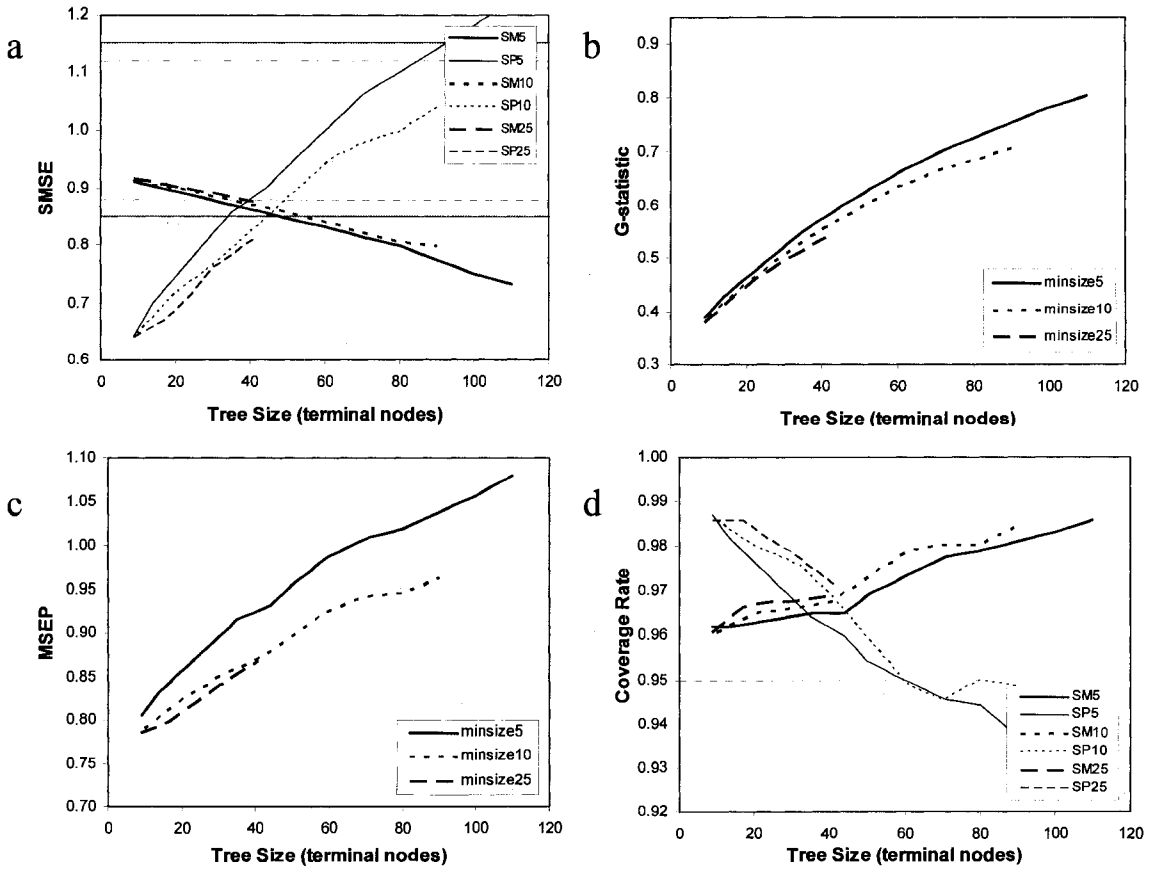


Figure 2. Influence of the minimum number of observations at a given terminal node (*minsize* = 5, 10 and 25) and tree size on a) the standardize mean square error (SMSE) for the mean response (SM) and predictions at new location (SP), b) the G-statistic, c) mean squared error of prediction (MSEP), and d) 0.95 coverage rates for the mean response (SM) and predictions at new locations (SP) for the soil pH model (3SLS + RT). In the upper left figure the two sets of straight lines represent the region in which the variance estimates are unbiased. The dotted lines represents the simultaneous confidence interval ($\alpha = 0.05$), while the solid lines represents the joint confident interval ($\alpha = 0.01$).

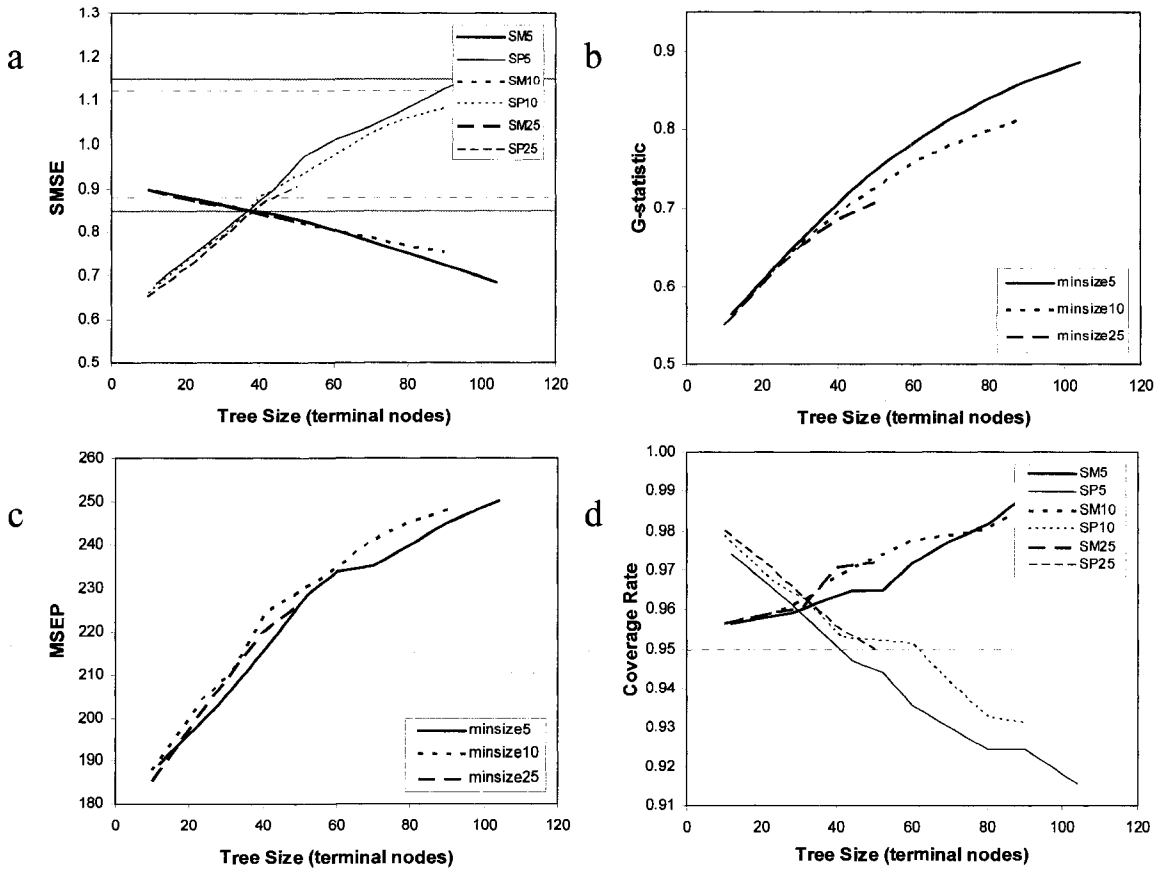


Figure 3. Influence of the minimum number of observations at a given terminal node (*minsize* = 5, 10 and 25) and tree size on a) the standardize mean square error (SMSE) for the mean response (SM) and predictions at new location (SP), b) the G-statistic, c) mean squared error of prediction (MSEP), and d) 0.95 coverage rates for the mean response (SM) and predictions at new locations (SP) for the sand model (3SLS + RT). In the upper left figure the two sets of straight lines represent the region in which the variance estimates are unbiased. The dotted lines represents the simultaneous confidence interval ($\alpha = 0.05$), while the solid lines represents the joint confident interval ($\alpha = 0.01$).

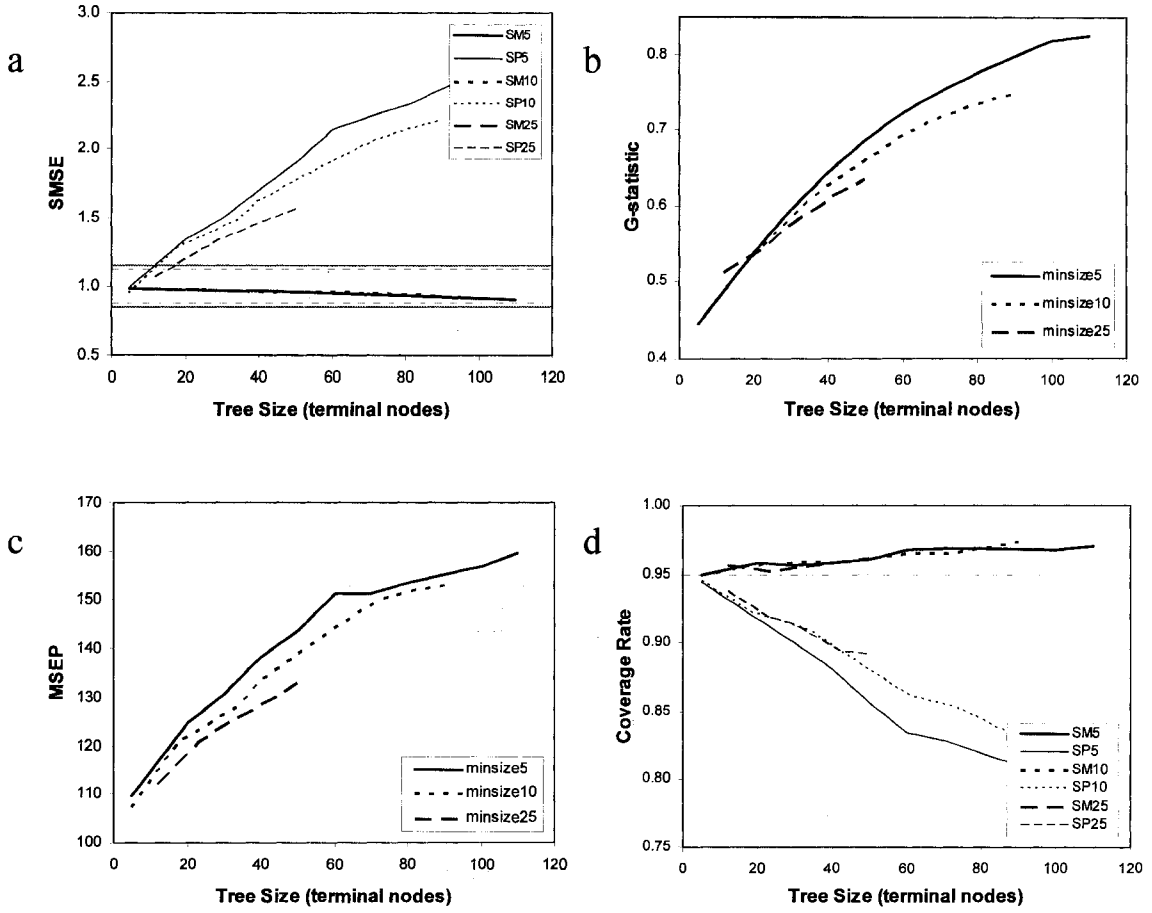


Figure 4. Influence of the minimum number of observations at a given terminal node (*minsize* = 5, 10 and 25) and tree size on a) the standardize mean square error (SMSE) for the mean response (SM) and predictions at new location (SP), b) the G-statistic, c) mean squared error of prediction (MSEP), and d) 0.95 coverage rates for the mean response (SM) and predictions at new locations (SP) for the clay model (3SLS + RT). In the upper left figure the two sets of straight lines represent the region in which the variance estimates are unbiased. The dotted lines represents the simultaneous confidence interval ($\alpha = 0.05$), while the solid lines represents the joint confident interval ($\alpha = 0.01$).

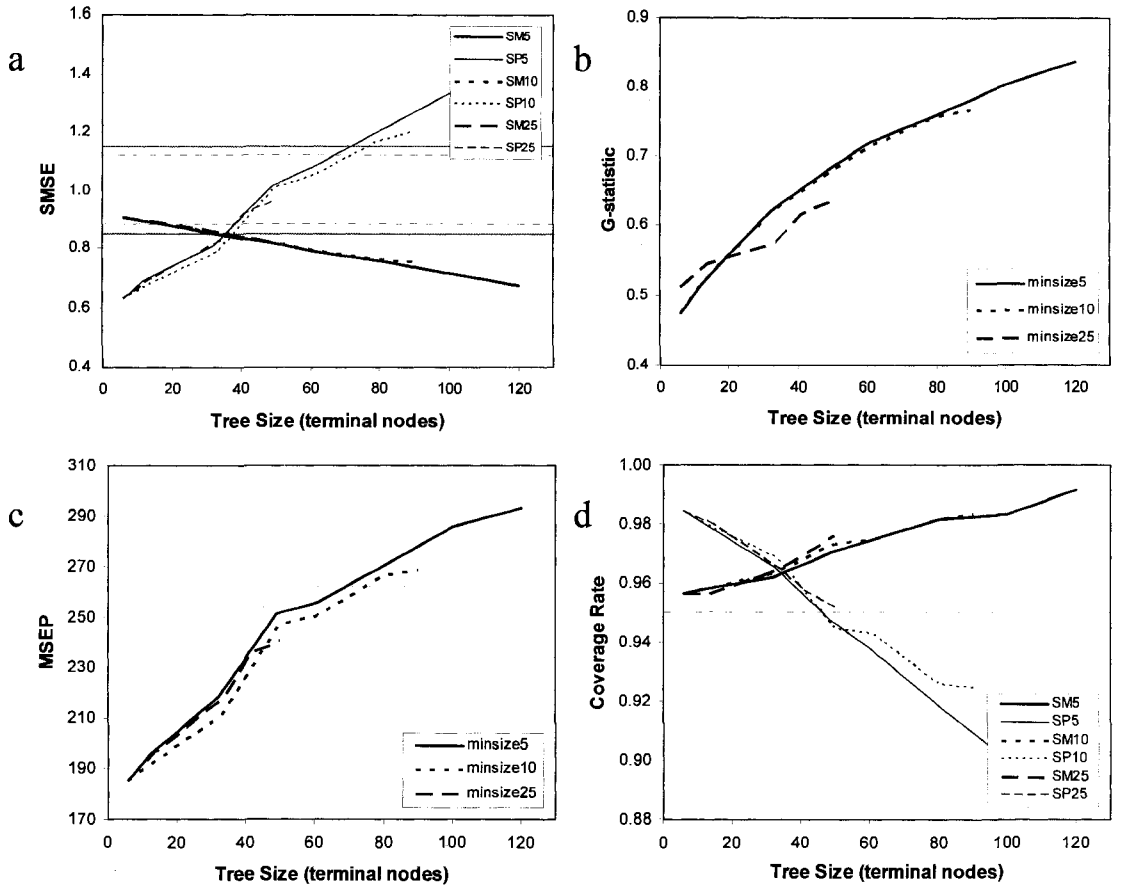


Figure 5. Influence of the minimum number of observations at a given terminal node (*minsize* = 5, 10 and 25) and tree size on a) the standardize mean square error (SMSE) for the mean response (SM) and predictions at new location (SP), b) the G-statistic, c) mean squared error of prediction (MSEP), and d) 0.95 coverage rates for the mean response (SM) and predictions at new locations (SP) for the sand model (3SLS + MRT). In the upper left figure the two sets of straight lines represent the region in which the variance estimates are unbiased. The dotted lines represents the simultaneous confidence interval ($\alpha = 0.05$), while the solid lines represents the joint confident interval ($\alpha = 0.01$).

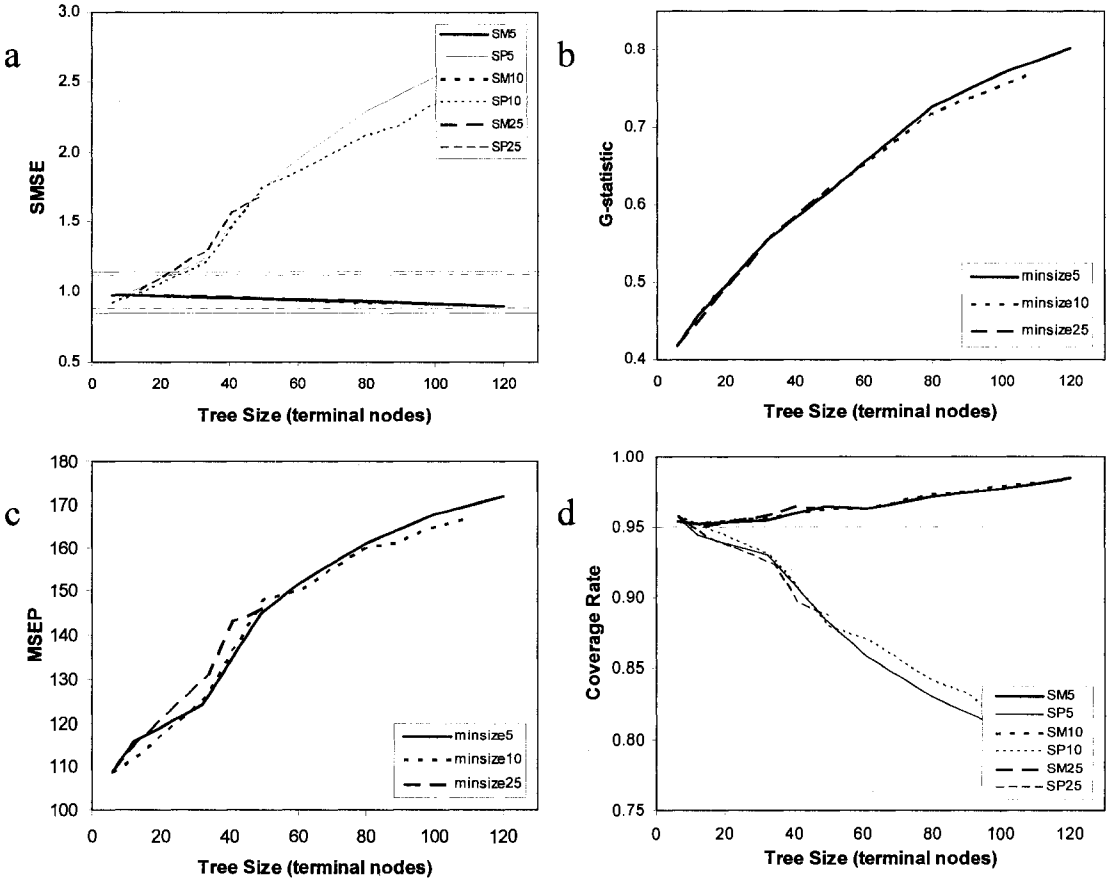


Figure 6. Influence of the minimum number of observations at a given terminal node (*minsize* = 5, 10 and 25) and tree size on a) the standardize mean square error (SMSE) for the mean response (SM) and predictions at new location (SP), b) the G-statistic, c) mean squared error of prediction (MSEP), and d) 0.95 coverage rates for the mean response (SM) and predictions at new locations (SP) for the clay model (3SLS + MRT). In the upper left figure the two sets of straight lines represent the region in which the variance estimates are unbiased. The dotted lines represents the simultaneous confidence interval ($\alpha = 0.05$), while the solid lines represents the joint confident interval ($\alpha = 0.01$).

while a *minsize* of 5 and the tree size of 12 were optimal for predicting clay. However, the MRT algorithm fits a regression tree to all of the variables simultaneously requiring only one set of parameters. Table 2 indicates that unbiased estimates of the variances for the mean response and new predictions of sand and clay were not achieved when the same *minsize* and tree size were used. Thus, the final models for estimating sand and clay were fitted based on a *minsize* of 25 and a tree size of 41, because of the better fit associated with these parameter. This resulted in a SMSEP of 1.40 for the clay model which suggests that variance estimates were underestimated by 40% compared to the true errors. Using the optimal *minsize* and tree size, the MRT models accounted for an additional 19% and 17% of the observed variability in sand and clay, respectively. The final models (3LSL + MRT) for sand and clay accounted for 62% and 56% of the observed variability, respectively (Table 2).

Sample variogram models were calculated for each of the soil attributes. The descriptive statistics of the soil data (n = 1427) revealed that the average distance among sample plots was 177 km and ranged from 0.35 km to 459 km (Table 3). The fitted sample variogram models for sand, clay, silt and pH had range parameters varying from 5.86 km to 10 km with small a nugget effect for both the raw data and the estimated residuals from the 3SLS models.

When applying ordinary kriging to the 3SLS residuals, the large values for the cost complexity functions for both sand and clay indicated that ordinary kriging did not perform well in terms of estimating the variances. The results suggest that the variance estimates for the mean response and new predictions were significantly different from one, indicating that the variance estimates were not consistent with the true errors.

Table 3. Descriptive statistics of the fitted variograms using the observed data and residuals obtained from 3SLS models.

Soil Attributes	Variogram Model	Observed Data			Residuals from 3SLS Model		
		Range (km)	Sill	Nugget	Range (km)	Sill	Nugget
Sand	Gaussian	5.86	117.90	0.00	5.48	112.35	0.00
Clay	Gaussian	5.93	66.46	0.00	5.59	62.96	0.00
Silt	Gaussian	6.11	44.04	0.00	5.64	41.56	0.00
pH	Exponential	10.00	0.68	0.09	10.00	0.58	0.09

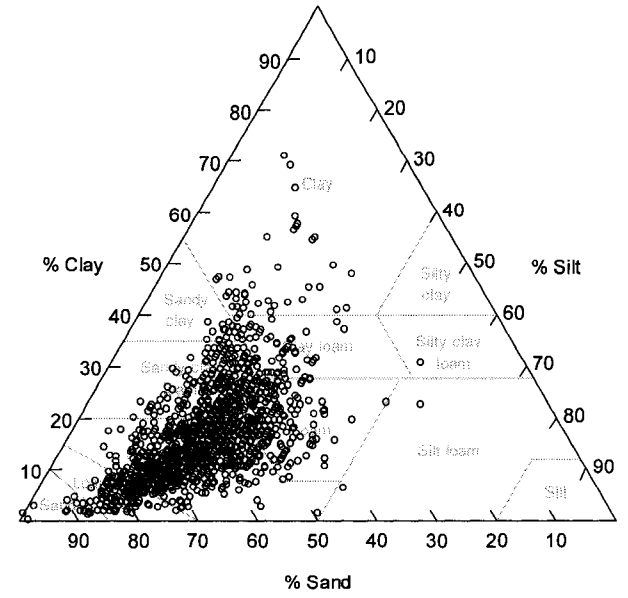
Ordinary kriging only accounted for an additional 9% and 17% of the observed variability in sand and clay, respectively. The final models (3SLS + OK) accounted for a total of 52% and 56% of the total variability observed in sand and clay, respectively. No discrepancies were detected between the frequency distribution of observed and predicted values for the sand, clay and silt model for the 3SLS + RT and 3SLS + MRT models (Table 4). The maximum value of predicted sand using the 3SLS + OK model was 126.64%, which is larger than the possible true value. The 3SLS + MRT model satisfied the constraint that the estimated values of the sum of sand, silt and clay equaled 100%, while the 3SLS + RT model had sums ranging from 82.19% to 121.60%.

The observed and predicted soil texture fractions were also compared in a soil triangle plot (Figure 7). The soil triangle plots indicated that the predicted values of soil textural fractions obtained from 3SLS + MRT had the same pattern as the observed data, but with less variability. Scatter plots of the predicted values and model residuals are displayed in Figure 8a for the 3SLS + RT model, Figure 9a for the 3SLS + MRT model and Figure 10a for the 3SLS + OK model. No systematic patterns were observed in the residual plots for models in which the residuals were fit using a stratified tree-based approach. This was not the case for the model that used ordinary kriging to describe the small-scale variability in the residuals from the 3SLS models (sand, $r = -0.46$; clay, $r = -0.39$) (Figure 10a). Scatter plots of the observed and predicted values of the soil attributes are provided in Figure 8b for the 3SLS + RT model, Figure 9b for the 3SLS + MRT model and Figure 10b for the 3SLS + OK model. The correlation between observed and predicted values for sand and clay from the 3SLS + OK model was 0.47 and 0.48, respectively, which were lower than those for the 3SLS + RT model (0.64 and 0.53) and 3SLS + MRT model (0.64 and 0.63), respectively.

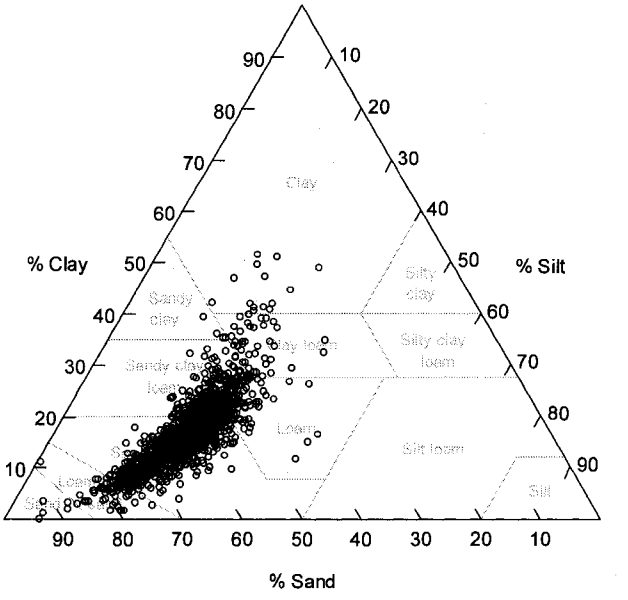
Table 4. Distribution of observed and predicted soil attributes based on the fitted models applied to the entire data set.

Model	Soil Attribute	Sample Size	*min-size*[1]	tree size[1]	Min	1st Quantile	Median	Mean	3rd Quantile	Max	SD[1]
Observed Data	**Sand**	1427	—	—	17.40	51.48	60.45	60.41	69.70	98.50	12.59
	Clay	1427	—	—	0.30	10.24	15.50	17.06	22.12	71.10	9.87
	Silt	1427	—	—	0.00	17.40	22.00	22.53	27.00	56.00	7.00
	ph	1427	—	—	4.03	5.61	6.20	6.27	6.85	10.72	0.92
3SLS + RT	**Sand**	1427	25	40	31.30	54.74	59.89	60.41	65.63	89.52	9.01
	Clay	1427	10	5	-3.85	13.20	17.02	17.06	20.74	50.69	6.48
	ph	1427	5	50	3.84	5.84	6.27	6.27	6.60	11.08	0.63
3SLS + MRT	**Sand**	1427	25	41	22.69	54.89	59.84	60.41	65.79	95.58	9.27
	Clay	1427	25	41	-3.40	12.56	16.59	17.06	20.46	51.48	6.98
	Sand	1427	5	12	22.69	54.90	59.56	60.41	64.92	95.58	8.58
	Clay	1427	5	12	-3.40	13.39	17.11	17.06	20.70	48.73	6.08
	Silt	1427	5	12	0.65	20.51	22.70	22.53	24.52	44.52	3.82
3SLS + Kriging	**Sand**	1427	—	—	-32.26	53.16	59.68	60.38	66.97	126.64	11.68
	Clay	1427	—	—	-8.79	11.39	16.44	16.98	21.98	71.73	8.41

[1]*minsize* = minimum number of observations required to split the data, tree size = number of strata or terminal nodes in the tree, SD = standard deviation.



Observed



Predicted

Figure 7. Distribution of observed and predicted soil textural classes.

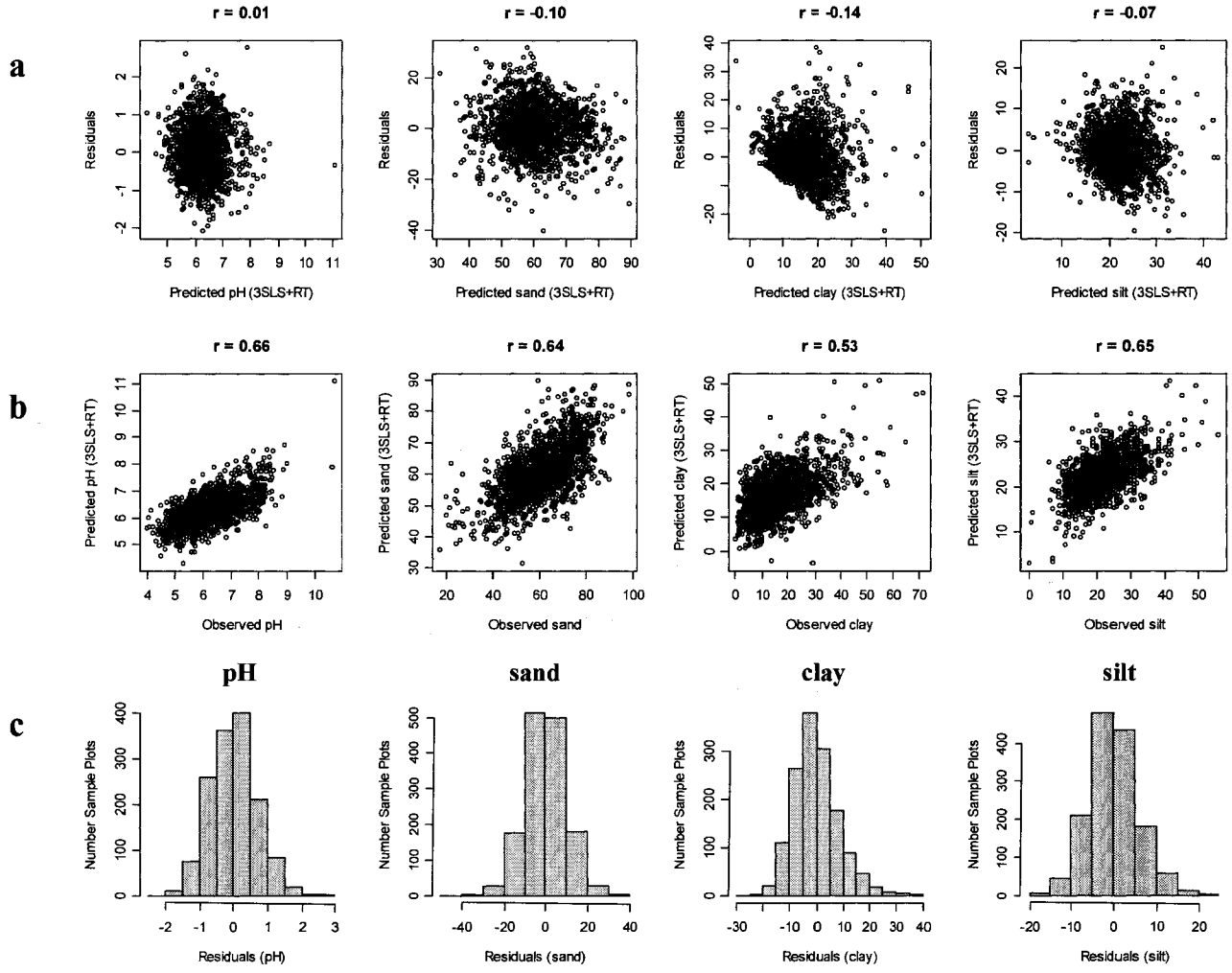


Figure 8. Scatter plots of a) predictions errors vs. predicted values, b) observed vs. prediction values and c) histograms of prediction errors from the three-stage least squares and regression tree models for soil pH, sand, clay and silt when applied to the entire data set.

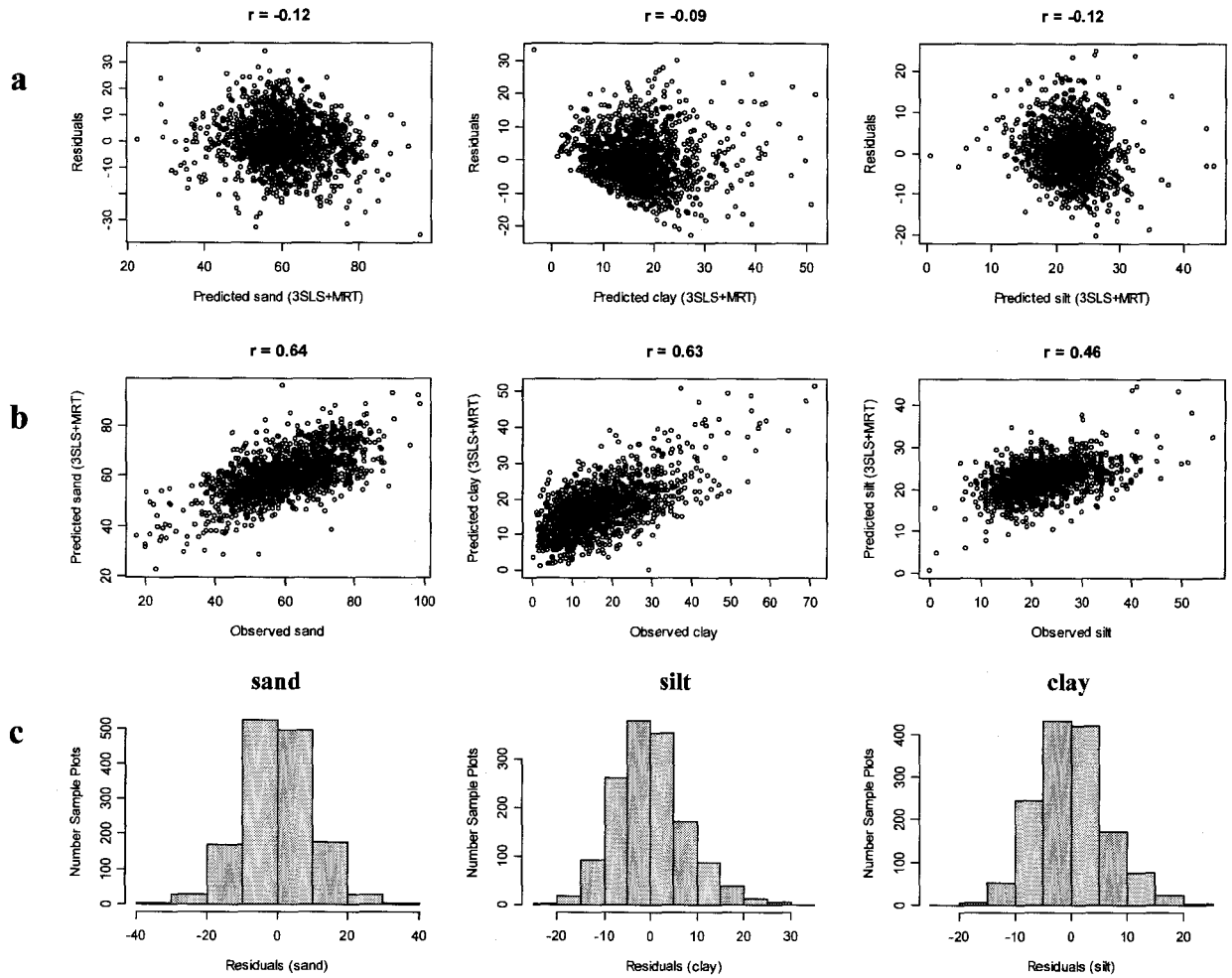


Figure 9. Scatter plots of a) predictions errors vs. predicted values, b) observed vs. prediction values and c) histograms of prediction errors from the three-stage least squares and multivariate regression tree models for soil pH, sand, clay and silt when applied to the entire data set.

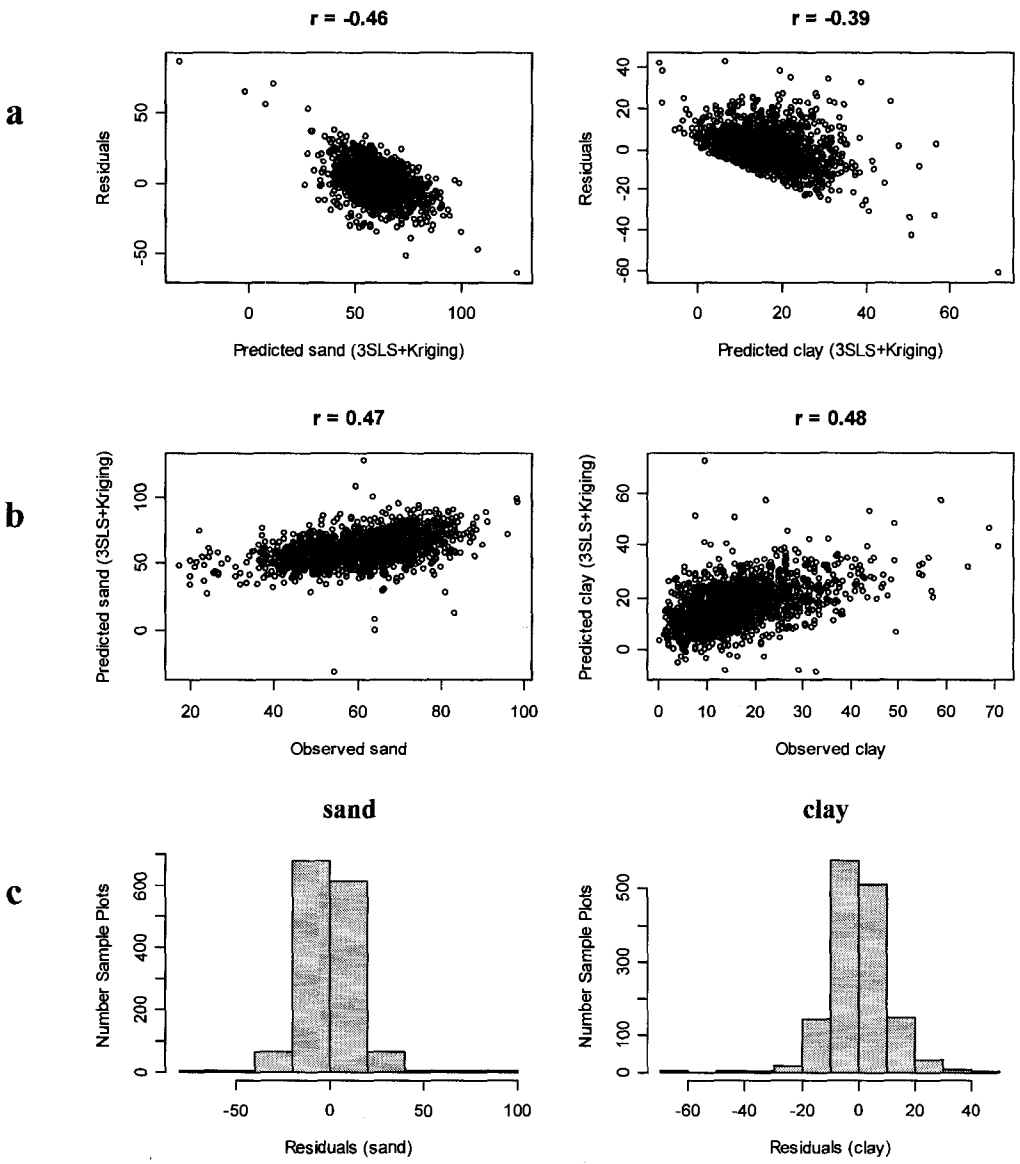


Figure 10. Scatter plots of a) predictions errors vs. predicted values, b) observed vs. prediction values and c) histograms of prediction errors from the three-stage least squares and ordinary kriging models for sand and clay when applied to the entire data set.

Also, histograms of the residuals from the 3SLS + RT (Figure 8c), the 3SLS + MRT (Figure 9c) and 3SLS + OK (Figure 10c) models were approximately normally distributed, suggest the models did not violate the underlying assumption (i.e., the residuals were independent and identically normally distributed).

Predictive surfaces of soil pH, sand, clay and silt were created as GIS layers and are displayed in Figure 11, while the standard deviation surfaces are displayed in Figure 12. The surface of soil pH was divided into four classes (Figure 11A) corresponding to acidity and alkalinity classification used in soil and crop management. Predictive surfaces of sand using 3SLS + OK were created and shown in Figure 13. A map representing the 12 soil textural classes were also generated (Figure 14). Figure 15 compares the distribution of observed and predicted soil textural classes for the sample data and the state as a whole. A Chi-square goodness of fit test indicated that the distribution of observed and the predicted soil textural classes were not significantly different (3SLS + RT, p-value = 0.13; 3SLS + MRT, p-value = 0.14). The distribution of predicted soil textural classes for the entire state (N = 94792466) was not significantly different from the observed data (p-value = 0.53).

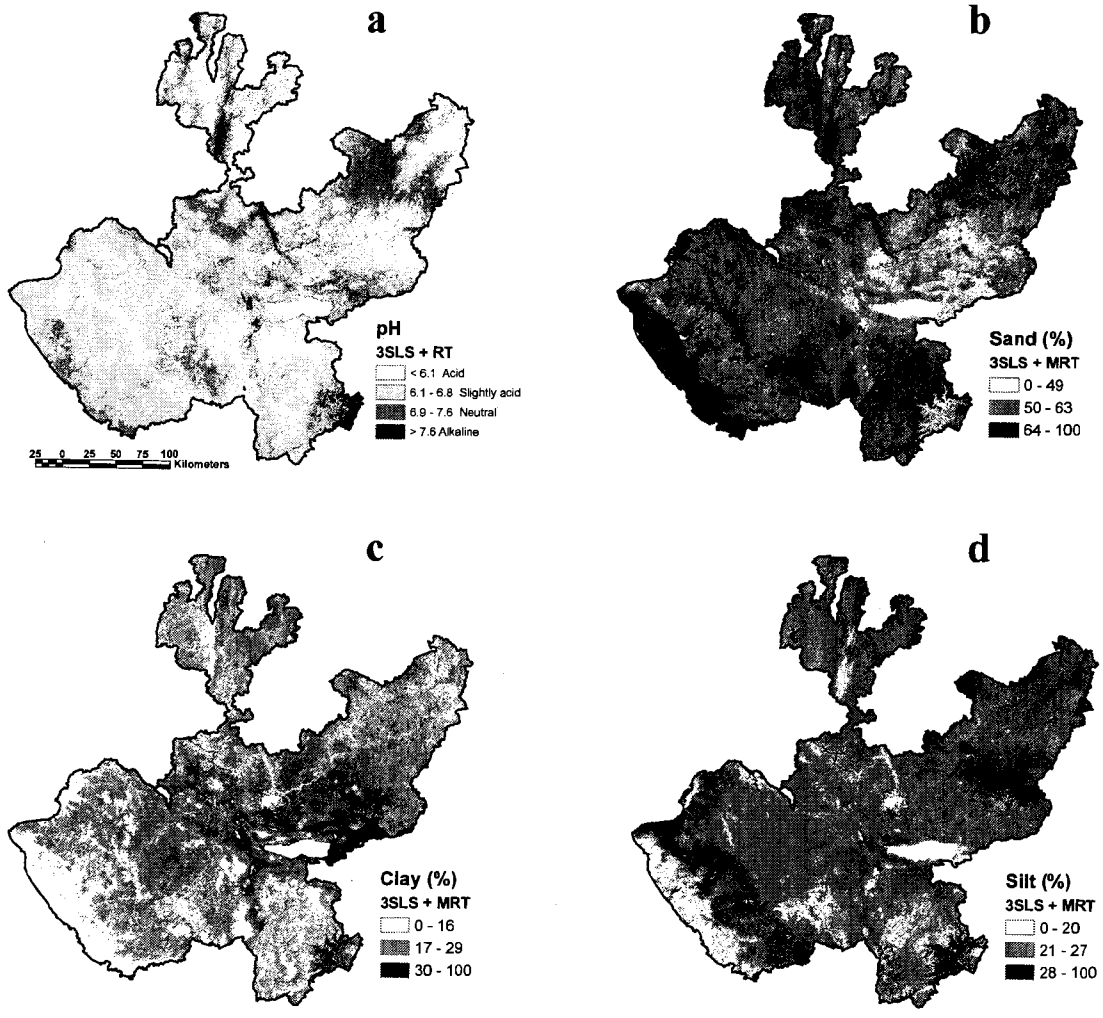


Figure 11. Spatial distribution of predicted a) soil pH, b) sand, c) clay and d) silt in the State of Jalisco, Mexico based on 3SLS + MRT model for soil texture and the 3SLS + RT model for the soil pH.

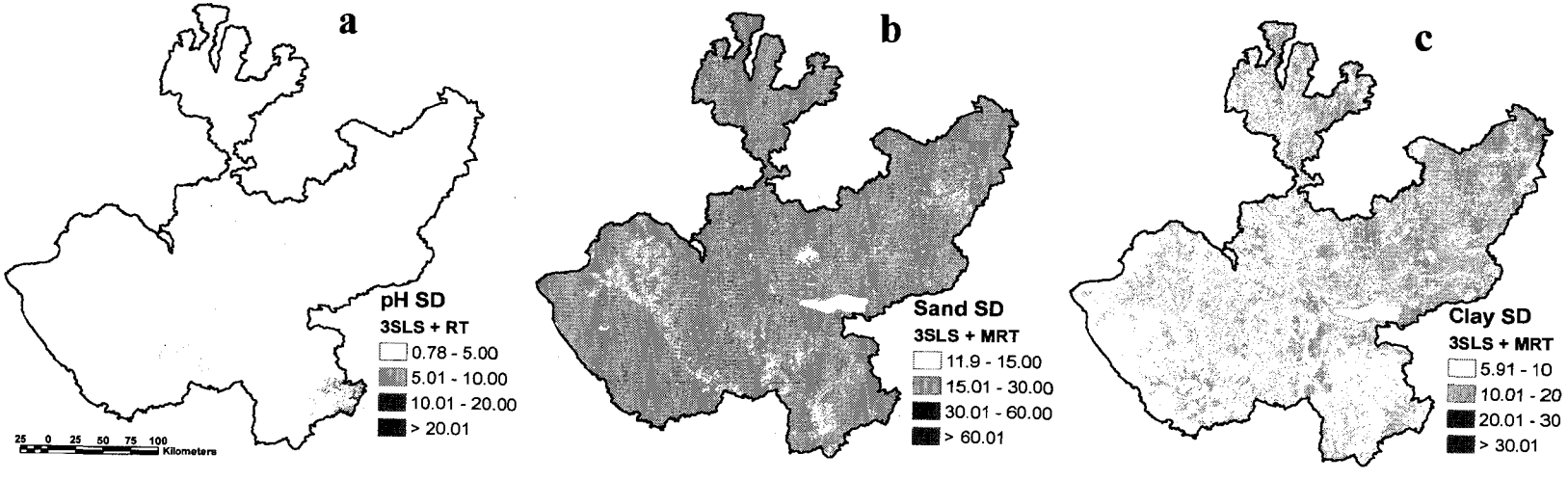


Figure 12. Spatial distribution of the prediction standard deviation (SD) for a) soil pH, b) sand, and c) clay in the State of Jalisco, Mexico.

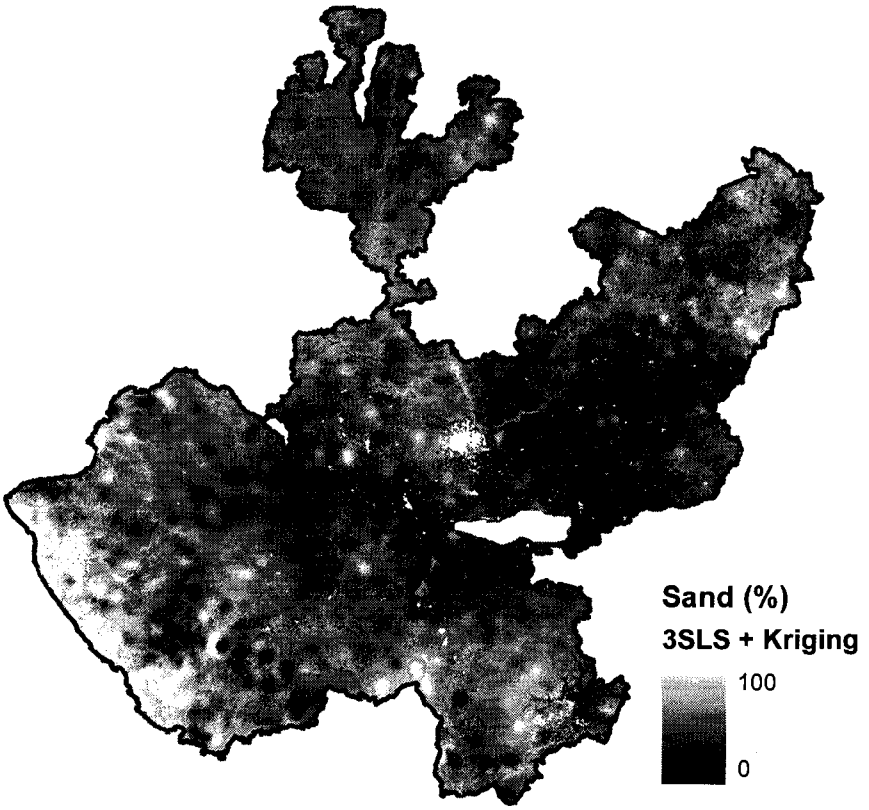


Figure 13. Spatial distribution of predicted sand in the State of Jalisco, Mexico based on three-staged least squares and ordinary kriging.

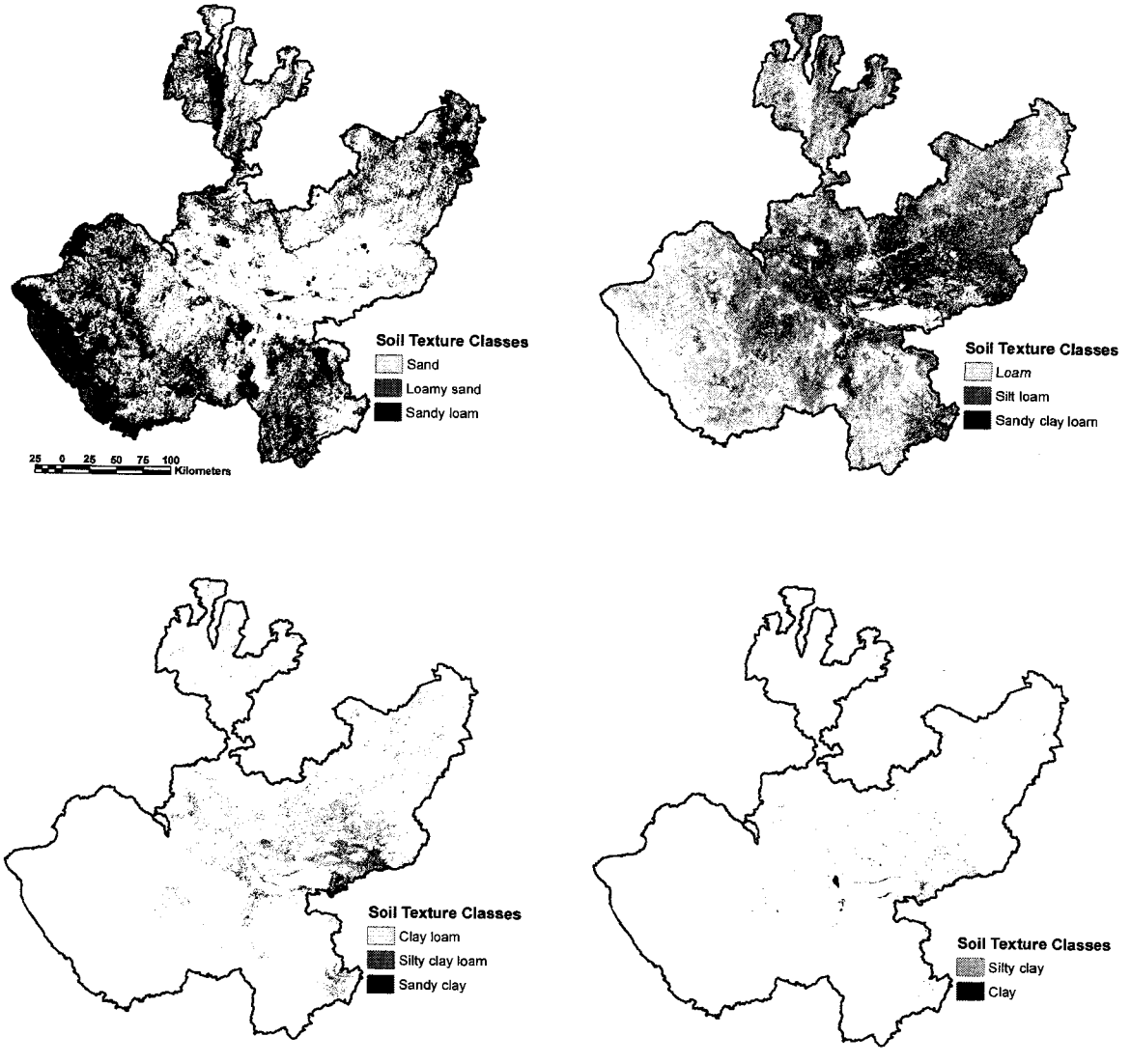


Figure 14. Spatial distribution of soil textural classes in the State of Jalisco, Mexico based on the 3SLS + MRT model.

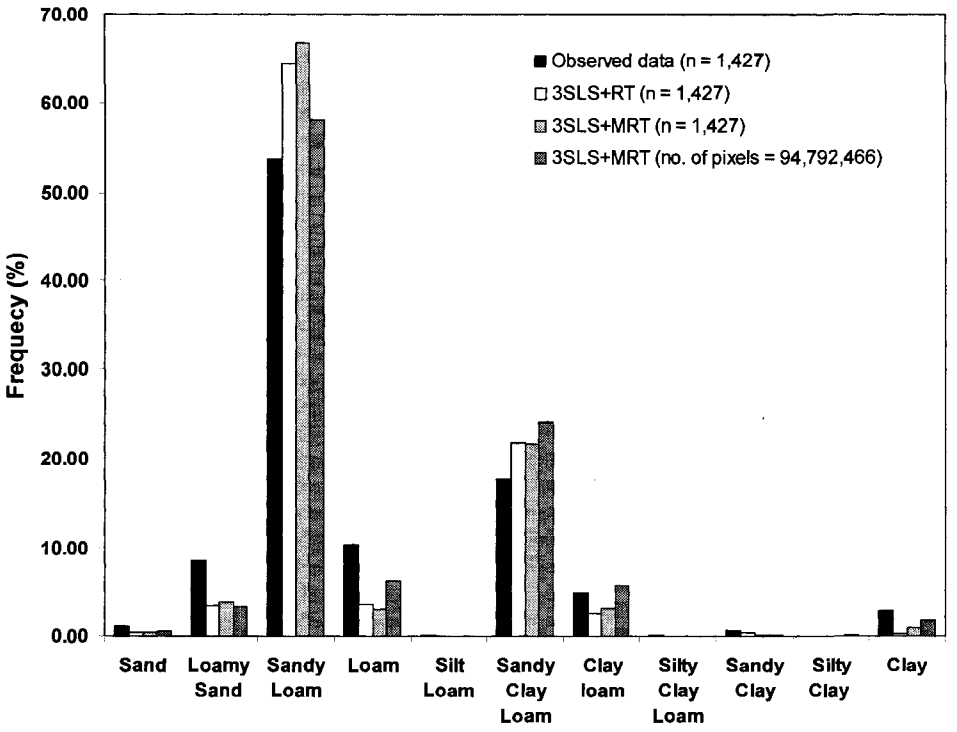


Figure 15. Frequency distribution of soil textural classes for the observed sample data, predicted values obtained form the three-stage least squares plus regression tree (3SLS + RT) and three-stage least square plus multivariate regression tree (3SLS + MRT) models.

DISCUSSION

The assessment of the spatial variability in soil texture is central to support a variety of management decision processes. Soil physical properties such as soil texture have influence the water-holding capacity, cation-exchange capacity, crop yields, site productivity, nitrogen loss, as well as other soil processes and conditions. Modeling the spatial distribution of soil attributes over large geographical regions at a fine spatial resolution presents some challenging obstacles that need to be addressed.

First, when modeling soil attributes such as soil texture, it is implied that the sum of the percent sand, silt and clay equal to 100%. To address this issue the method of three-stage least square was introduced to simultaneously model the fractions of soil texture to ensure they sum to 100%.

A second issue that is addressed in this paper deals with the problem of the lack of spatial dependency associated with the variables being modeled. Due to the lack of spatial dependency, geostatistical methods such as kriging may not be an appropriate technique for interpolating this type of data. For comparison purpose, ordinary kriging was used to interpolate the residuals from the models of sand and clay. Cross-validation indicated that while the estimates were unbiased, the estimated variances for the mean response and predictions at new locations were not adequate. The predicted surfaces obtained from this approach resulted in circular polygons surrounding the sample locations throughout the state. Such artifacts occur when the range parameter of the fitted variogram models are much smaller than the average distance separating the sample data. To address this issue, a stratified tree-based approach was used to model the residuals from the 3SLS models of the soil attributes.

This paper evaluated two approaches, the use of univariate regression trees applied to individual models, and multivariate regression trees which simultaneously modeled the residuals using a single tree structure. Cross-validation indicated that both approaches provided unbiased estimates of soil attributes as well as providing unbiased variance estimations of the mean response and predictions at unsampled locations. However, only the multivariate regression tree satisfied the constraint that the soil textural fractions sum to 100%. The final set of models accounted for 62% of the observed variability in sand, 56% for clay and 61% for the soil pH model.

Results of this study indicated that the State of Jalisco is dominated by slightly acidic soils, particulary at the lower elevations in the western and central portions of the state, which is dominated by tropical dry forests. Soil pH becomes more acidic with increasing elevation as the forests change to temperate pine-oak forests. More neutral soils are found in the eastern part of the state dominated by grasslands and vegetation characteristic of a semi-arid environment. The state is dominated by two major soil classes, sandy loam and sandy clay loam. The sandy clay loam soils are found primarily in the coastal region dominated by tropical dry forests and the deserts in the eastern part of the state. Sandy clay loam and clay loam soils occur primarily in the central portion of the state dominated by grasslands and agricultural lands.

Because of the nature of this study, it is difficult to compare the results of this study with previous studies conducted in the State of Jalisco or anywhere else in Mexico. Also, because of varying geographical scales and spatial resolutions used in other studies and the need to constrain estimates of soil fractions to sum to 100 percent, only two comparable studies were found in the literature. In one study, Henderson et al. (2005)

developed models for describing the spatial distribution of soil attributes using binary regression trees for the Australian continent (n = 135490) at a 250 m spatial resolution. The R^2 values for pH and clay were reported at 67% and 44%, respectively. These results are similar to the results presenting in this study. In a second study, Van Meirvenne and Van Cleemput (2006) used compositional ordinary kriging to simultaneously model soil texture while constraining the three soil fractions of soil texture to sum to 100 percent for an area covering 3000 km^2 in East Flanders, Belgium. The authors used a total of 4887 soil samples. The authors reported that the root mean square errors obtained from a 10-fold cross validation for sand and clay were 4.9 and 10.32 respectively, which were slightly lower than those obtained in this study.

Geostatistical techniques such kriging usually require tremendous sampling effort in order to capture the spatial variability in soil attributes. When dealing with large geographical areas, kriging might not be applicable because of the large sample size required to capture the spatial dependency among the soil attributes. Increasing sample size directly influences the cost of the survey making it prohibited in most cases. In this study 1427 soil samples were collected from an area covering 78618 km^2. While the sampling intensity is substantially lower than the two studies mentioned above, the method of 3SLS + MRT provided comparable results given the low sampling effort and extent of the study area.

CONCLUSION

Finding a suitable statistical approach for describing the spatial variability in a set of data is not considered an easy task. Based on a simulation study, the procedure of using either 3SLS + RT or 3SLS + MRT was able to provide reliable estimates of selected soil attributes over a large geographical region with the fine-scale spatial resolution. However, only the 3SLS + MRT approach was able to satisfy the constraint that estimates if sand, silt and clay to sum to 100%. The technique of combining the 3SLS and tree-based stratified approach also assures that the final models will be unbiased or at least optimal in terms of estimating the variances of the mean response and prediction variances. As a result, confidence intervals and prediction interval can be constructed for individual observations, using the prediction standard deviation surfaces developed in this study. Standard deviation surfaces not only convey meaningful information on the precision of the estimates but also provide information on where additional sampling is required to improve the precision of the predictive models of soil attributes.

It is well known that soil factors tremendously influence the productivity of forest and agricultural lands. Maps of soil attributes obtained from this study can serve as a useful surrogate explaining the spatial variability in soil attributes across large geographical region. Digital maps of soil attributes provided comprehensive information on the spatial variability of soil properties for the entire State of Jalisco at a fine spatial resolution (i.e., 30 m x 30 m). The GIS layers of soil attributes developed in this study could be used to support the applications of precision forestry and agriculture such as

managing soil fertility and crop production for site-specific management over both small and large geographical regions.

REFERENCES

Agterberg, F. P. 1984. Trend surface analysis. In: Spatial Statistics and models, G. L. Gaile and C. J. Willmott (eds.). Reidel, Dordrecht, The Netherlands, pp. 147-171.

Akaike, H. 1969. Fitting autoregressive models for prediction. Annals of the Institute for Statistics and Mathematics 21, 243-247.

Bell, J. C., R. L. Cunningham, and M.W. Havens. 1994. Soil drainage probability mapping using a soil -landscape model. Soil Science Society of America Journal 58, 464- 470.

Benedetti, R., G. Espa, and G. Lafratta. 2005. A tree-based approach to forming strata in multipurpose business surveys. Discussion Paper No. 5 2005, Dipartimento di Economia, Universita Degli Studi di Trento, Trento, Italy. 17 pp.

Bouyoucos, G. J. 1936. Directions for Making Mechanical Analysis of Soils by the Hydrometer Method. Soil Science. 42(3).

Breiman, L., J. H. Friedman, R.A. Olshen, and C. J. Stone. 1984. Classification and regression. Tress. Wadsworth International Group, Belmont, CA. 368 pp.

Bui, E. N., and C. J. Moran. 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103, 79-94.

Chambers, J. M. and T. J. Hastie. 1992. Statistical models in S. Wadsworth and Brooks/Cole. California, 608 pp.

Cocchi, D., E. Fabrizi, M. Raggi and C. Trivisano. 2002. Regression trees based stratification: An application to the analysis of the Italian post enumeration survey. Proceedings of the International Conference on Improving Surveys, August 25-28, 2002, Copenhagen, Denmark. [online] URL: http://www.icis.dk/ICIS-papers/B2_5_2.pdf.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. Wiley and Sons, New York. pp. 428.

Crawley M. J. 2002 Statistical Computing: An Introduction to Data Analysis using S-Plus. John Wiley & Sons. Chichester. 761 pp.

De Gruijter, J. J., D. J. J. Walvoort, and P. F. M. van Gaans, 1997. Continuous soil maps – a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma 77, 169-195.

De'ath, G. 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83, 1105-1117.

De'ath, G. 2007. An online help: the *mvpart* package in R. [online] URL: http://cran.r-project.org/doc/packages/mvpart.pdf

Dobos, E., L. Montanarella, T. Negre, E. Micheli, 2001. A regional scale soil mapping approach using integrated AVHRR and DEM data. International Journal of Applied Earth Observation and Geoinformation 3, 30-41.

Efron, B. and R. J. Tibshrani. 1993. An introduction to the bootstrap. Chapman and Hall. New York, 456 pp.

ERDAS, Inc. 1999. Earth Resource Data Analysis System Field Guide, Fifth Edition. ERDAS, Inc., Atlanta, Georgia, 672 pp.

Erxleben J., K. Elder, and R. Davis. 2002. Comparison of spatial interpolation methods for estimating snow distribution in the Colorado Rocky Mountains. Hydrological Process 16, 3627-3649.

ESRI. 1995. ARC/INFO Software and on-line help manual. Environmental Research Institute, Inc., Redlands, CA.

ESRI. 2005. Environmental Systems Research Institute, Inc., 380 New York St., Readlands, CA 97393. USA.

Fortin, M.-J. and M. Dale. 2005. Spatial analysis: a guide for ecologists. Cambridge University Press. Cambridge. 365 pp.

Gee, G. W. and J. W. Bauder. 1979. Particle size analysis by hydrometer: A simplified method for routine textural analysis and a sensitivity test of measurement parameters. Soil Sci. Soc. Am. J. 43:1004-1007.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, D. Tyler. 2002. The national elevation dataset. Photogrammetric Engineering & Remote Sensing 68, 5-11.

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. 1995. Soil-landscape modelling and spatial prediction of soil attributes. International Journal of Geographical Information Systems 9, 421-432.

Gotway-Crawford, C.A. and G.W. Hergert. 1997. Division S-8—Nutrient management and soil plant analysis. Incorporating spatial trends and anisotropy in geostatistical mapping of soil properties. Soil Sci. Soc. Am. J. 61, 298-309.

Greene, W. H. 1990. Econometric Analysis. Macmillan Publishing Company. New York. 783 pp.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 47-186.

Henderson, B.L., E.N. Bui, C.J. Moran, D.A.P. Simon. 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383-398.

Hevesi, J. A., J. D. Istok, and A. L. Flint. 1992. Precipitation estimation in mountainous terrain using multivariate geostatistics. Part I: structural analysis. Journal of Applied Meteorology 31, 661-676.

Isaaks, E. H., and R. M. Srivastava. 1989. An introduction to applied geostatistics. Oxford University Press, New York. 561 pp.

Jones, J. B. 2002. Agronomic Handbook: Management of Crops, Soils and Their Fertility. CRC Press, Florida. pp 352.

Knotters, M., D. J. Brus, J. H. Oude Voshaar. 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227- 246.

Kravchenko, A. and D. G. Bullock. 1999. A comparative study of interpolation methods for mapping soil properties. Agronomy Journal 91, 393-400.

McBratney, A.B., M.L. Mendonca Santos and B. Minasny. 2003. On digital soil mapping. Geoderma 117, 3-52.

McLean, E. O. 1982. Soil pH and lime requirement. *In* Page, A. L., R. H. Miller and D. R. Keeney (eds.) Methods of soil analysis. Part 2 - Chemical and microbiological properties. (2nd Ed.). Agronomy 9, 199-223.

McKenzie, N. J., M. P. Austin. 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57, 329-355.

Meul, M and M. Van Meirvenne. 2003. Kriging soil texture under different types of nonstationarity. Geoderma. 112, 217-233.

Moore, I. D., P. E. Gessler, G.A. Nielsen, and G.A. Peterson. 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57, 443-452.

Odeh, I. O. A., A. B. McBratney, and D.J. Chittleborough. 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197- 214.

Odeh, I. O. A., A. B. McBratney, and D.J. Chittleborough. 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67, 215- 225.

Oom, S. and Jim Lemon. 2005. An online help: the *plot.soiltexture* function in R. [online] URL: http://tolstoy.newcastle.edu.au/R/help/05/05/5433.html

Pavlik, H. F., F. D. Hole. 1977. Soilscape analysis of slightly contrasting terrains in southeastern Wisconsin. Soil Science Society of America Journal 41, 407- 413.

Rabus, B., M. Eineder, A. Roth, and R. Bamler. 2003. The shuttle radar topography mission—a new class of digital elevation models acquired by spaceborne radar. ISPRS Journal of Photogrammetry & Remote Sensing 57, 241- 262.

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. [online] URL http://www.R-project.org.

Reich, M. R., J. E. Lundquist, and V. A. Bravo. 2004. Spatial models for estimation fuel loads in the Black Hills, South Dakota, USA. International Journal of Wildland Fire 13, 119-129.

Reich, R. M., and R. A. Davis. 2007a. 'On-line spatial library for R package.' (Colorado State University: Fort Collins, CO). [online] URL: http://www.warnercnr.colostate.edu/~robin/)

Reich, R. M. and R. A. Davis. 2007b. Quantitative spatial analysis: course note for NR/ST523. Colorado State University, Fort Collins, CO, 481 pp.

Reich, R. M. and C. Aguirrie-Bravo. 2008. Empirical Evaluation of Confidence and Prediction Intervals for Spatial Models of Forest Structure in Jalisco, Mexico. Journal of Environmental Statistitics. (In review)

Reich, R. M., C. Aguirrie-Bravo and M. A. Mendoza Briseño. 2008. An innovative approach to inventory and monitoring of natural resources in the Mexican State of Jalisco. Journal Environmental Monitoring and Assessment. (In press)

Ripley, B. 2007. An online help: the *rpart* package. [online] URL: http://cran.r-project.org/doc/packages/rpart.pdf

Ripley, B. 2008. An online help: the *MASS* package. [online] URL: http://cran.r-project.org/web/packages/VR/VR.pdf

Ryan, P. J., N. J. McKenzie, D. O'Connell, A. N. Loughhead, P. M. Leppert, D. Jacquier, and L. Ashton. 2000. Integrating forest soils information across scales: spatial prediction of soil properties under Australian forests. Forest Ecology and Management 138, 139-157.

Scull P., J. Franklin, Q. A. Chadwick, and D. McArthur 2003. Predictive soil mapping: a review, Progress in Physical Geography 27, 171-197.

Schloeder, C. A., N. E. Zimmermann and M. J. Jacobs. 2001. Comparison of methods for interpolating soil properties using limited data. American Society of Soil Science Journal 65, 470-479.

Skidmore A. K., C. Varekamp, L. Wilson, E. Knowles and J. Delaney. 1997. Remote sensing of soils in a eucalypt forest environment. International Journal of Remote Sensing 18, 39-56.

Soil Survey Division Staff. 1993. Soil survey manual. Soil Conservation Service. U.S. Department of Agriculture Handbook 18. [online] URL: http://soils.usda.gov/technical/manual/print_version/complete.html

Troeh, F. R., 1964. Landform parameters correlated to soil drainage. Soil Science Society of America Proceedings 28, 808- 812.

Van Meirvenne, M and I. Van Cleemput. 2006. Pedometrical Techniques for soil texture mapping at different scales. In: Grunwald, S. (Ed.), Environmental Soil-Landscape Modeling - Geographic Information Technologies and Pedometrics. CRC Press, New York, pp. 323-341.

Venables, W. M. and B. D. Ripley. 1999. Modern Applied Statistics with S-PLUS. Springer-Verlag, New York. 501 pp.

Voltz, M., P. Lagacherie, and X. Louchart, 1997. Predicting soil properties over a region using sample information from a mapped reference area. European Journal of Soil Science 48, 19- 30.

Walker, P.H., G. F. Hall, and R. Protz. 1968. Relation between landform parameters and soil properties. Soil Science Society of America Proceedings 32, 101- 104.

Webster, R. and M. A. Oliver. 2001. Geostatistics for Environmental Scientists. John Wiley and Sons. New York. 271 pp.

Zellner, A. and H. Theil. 1962. Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. Econometrica 30, 54-78.

CHAPTER 2: EVALUATION OF THREE SAMPLING DESIGNS FOR DEVELOPING SPATIAL STATISTICAL MODELS

ABSTRACT

In natural resource studies, statistical models have been used extensively to account for the spatial variability in a set of data. The quality of a spatial model is closely related to the sampling design used to collect the sample data. Therefore, the reliability of estimates obtained from a spatial statistical model depends on how well the sample data represents the spatial variability in the population of interest. In many situations research scientists pay too much attention on selecting a suitable model but ignore how the data were collected. In this study, three sampling designs (simple random, systematic and stratified random sampling) were evaluated in modeling the spatial distribution of forest tree biomass in the State of Jalisco, Mexico. Results from a Monte Carlo simulation study suggested that stratifying the population based on the spectral properties of the vegetation provided a better fitting model compared to models based on simple random or systematic sampling which ignored the spectral variability in the population being modeled.

INTRODUCTION

The main objective of natural resource sampling is to make an inference about a population based on a representative sample selected from the population. Depending on the objectives of the survey, there are several methods one could use to select the sample units, the most common of which is design-based sampling. Designed-based sampling (Cochran 1977) uses probability sampling to select sample units in order to make inference about the population of interest. Examples of designed-based sampling include simple random, stratified random and cluster sampling, to name but a few. A second approach used to sample a population is model-based sampling. Many statistical models such as a linear regression models are developed using a model-based approach. The goal of the model-based approach is to find an appropriate predictive model to account for the behavior of the variable of interest in terms of a set of auxiliary variables (Warren 1998, Haining 2003, Lark and Cullis 2004, Stenvens 2006). In this situation, the model being fitted influences the way the sample data are collected. A good example of the model-based approach is developing an equation to predict the cubic volume of individual trees as function of tree diameter and tree height. Theory suggests that in order to minimize the variance of the mean response or the prediction variance, sample observations should be selected uniformly across the set of explanatory variables. In the case of the regression models, trees would be purposively selected such that the sample trees are uniformly distributed across all diameter-height classes. A disadvantage of this approach in selecting sample units is that the sample data can not be used to make an inference about the population of interest.

Recently, the model-assisted approach has emerged to fulfill the need of making inference about population (as a designed-based inference) while at the same time developing predictive models (Conquest 2003). The sample data are drawn from a population using a design-based approach while taking into consideration the requirement necessary for developing reliable models. Reich et al. (2008) developed a framework for designing multi-resource inventories based on the model-assisted design for inventorying and monitoring the natural resource in the State of Jalisco, Mexico.

Geostatistical techniques such as kriging has been influenced by both the model-based and sample-based approach. For example, in the field of geological exploration, samples can be located purposively to capture the spatial variability in the data while the fitted model can be used to make an inference about the population of interest. Theory also suggests that in order to minimize the prediction standard error associated with kriging, the sample data should be systematically located throughout the population of interest (Pettitt and McBratney 1993, Papritz and Webster 1995, Jardim and Ribeiro 2007). Brus and De Gruijter (1997) provide a comprehensive review of designed-based and model-based sampling strategies for describing soil properties using geostatistical approches. Recently, Stevens and Olsens (2003, 2004) introduced a sampling technique called "Spatial Balanced Sampling" which aims to spatially balance random samples throughout large geographical regions based on a systematic grid. The authors suggest that sample data with some degree of spatial regularity is more efficient than random samples, especially when surveying large geographical regions. Theobald et al. (2007) developed a new sampling technique called "Reversed Randomized Quadrant-Recursive Raster" based on a spatial balanced sampling design. In this approach, the inclusion

probability of limiting resources such as accessibility and cost are taken into consideration in spatially allocating samples across a landscape. However, a comparative study on the influence of different sampling strategies has rarely been done. Thus, the objective of this study was to evaluate the performance of developing a spatial statistical model of forest tree biomass using data obtained from two design-based approaches (simple random and systematic sampling) and a model-assisted approach (nested-stratified random sampling). The recommendation of a suitable sampling design will assist research scientists in developing model to describe the spatial variability is a set of data with confidence.

METHODS

Study Area

The State of Jalisco, Mexico is located in the west central part of Mexico, and covers an area of approximately 78618 km^2. The regions complex topography, geological substrata and climate are combined with the history of human influence to create an intricate mosaic of various vegetation types. Climatic variability in temperature, precipitation and evaporation define three broad climatic region (Rech et al. 2008). These zones coincide in general with those used to describe vegetation in Mexico (Rzedowski 1978). Furthermore, these zones define three broad ecological regions: 1) the first is the sub-humid tropical zone located along the Pacific coast and is characterized by high temperature, monsoon rain during the summer month (730-1200 mm) and an annual dry period that ranges from 5 to 9 months. Tropical dry forests dominate the region and occur on terrain with elevations from sea level to 2000 m. In the

northern part of this zone the forests are mesic, while in the south the forests are slightly drier, 2) at higher elevations the sub-humid temperate zone covers the greatest portions of the state. Pine, oak and mixed deciduous hardwood forests dominate this zone (1000-2500 m). Average annual rainfall ranges from 900-1500 mm. This zone gradually changes to 3) an arid and semi-arid zone that has a low annual precipitation of 400 mm or less and 8 to 12 dry months. Dominant vegetation includes mesquite-acacia and zerophitic shrubs.

GIS and Landsat-7 ETM+ Data

A digital elevation model (DEM) with a 90 m spatial resolution, obtained from the U.S. Geological Survey (USGS) (Gesch et al. 2002, Rabus et al. 2003) was resampled to a 30 m spatial resolution using the *Resample* function with the *Bilinear* option (ARC/INFO, ESRI 1995). The primary topographic attributes which included elevation, aspect and slope were derived from the DEM using *Spatial Analyst* tool (ARCGIS 9.2, ESRI 2006). In addition, a GRID layer of 12 climate zones (Reich et al. 2008) with a 30 m spatial resolution was included as an additional covariate. Ten cloud-free Landsat 7 ETM+ images obtained between January and March, 2004 were combined to create a seamless image using the *Mosaic* tool (ERDAS Inc. 1999). The thermal bands 6L and 6H with a 57 m resolution and the panchromatic band 8 with a 14.25 m resolution were resampled to a 30 m resolution. All GIS analyses were carried out using ArcGIS 9.2 (ESRI 2006).

Hypothetical Biomass Data

The GIS surface of forest tree biomass developed by Reich et al. (2008) was used to represent the population of biomass in the sate at a 30 m x 30 m spatial resolution.

Non-forested areas were not considered in this study to limit the analysis to only forested areas. To better represent field condition, a random error with a mean 0 and variance $4\sigma^2$ was added to the modeled values of biomass, where σ^2 is an estimate of the variability in forest tree biomass in the State of Jalisco. The "*create random raster*" tool available in the ArcGIS's toolbox (ArcGIS 9.2, ESRI 2006) was used to generate the error surface.

Sample Allocation and Sampling Designs

The State of Jalisco was initially stratified based on the climatic variability within the state. The three climate zones represent the tropical, temperate and semi-arid regions within the state. Each climatic zones was further stratified as to whether a sample unit represented a forested or non-forested region resulting in a total of six strata. Nested within each of the six strata, ten spectral classes were identified to represent the variability in land cover, resulting in a total of 60 strata. Initially samples were allocated to the forested regions within each climatic zones based on the economic importance of the region. A total of 600 sample plots were allocated to the temperate region, 400 to the tropical region and 300 to the semi-arid region. One-hundred sample plots were allocated to the non-forested areas in each of the tree climatic regions, resulting in a total sample size of 1600 plots. Because of difficulties in establish certain plots, only 1427 were established in the state.

In this study, two probability-based designs (simple random and systematic sampling) and one model-assisted sampling design (stratified random sampling) were evaluated to identify the best approach for modeling the spatial distribution of forest tree biomass in the State of Jalisco. Two sample sizes of 500 and 1100 were evaluated, where

the later reflects the sample size used in developing the original forest biomass surface. Only the 30 strata that were classified as being forested were considered in this study. For the three sampling designs, samples were allocated proportional to the size of the forested areas within each of the three bioclimatic zones. For the sample size of 500, the semi-arid, temperate and tropical zones were allocated 100, 300 and 100 sample plots, respectively, while for the sample size of 1100, the semi-arid, temperate and tropical zone were allocated 220, 660 and 220 sample plots, respectively. The three sampling designs were implemented independently within each of the climatic zones. For stratified sampling design, samples were allocated uniformly across the ten spectral classes within each climatic zone.

For a given sampling design, the location of the sample plots were overlaid on the various GIS grids to extract information on the elevation, slope, aspect, Landsat 7 ETM+ bands, forest type, bioclimatic zone, and estimated forest tree biomass.

Simple Random Sampling

Simple random sampling (SRS) is the most basic sampling design and usually serves as a basis for more complicated sampling designs. It is said to be a simple random sample when a sample of size n is drawn from a population of size N such that all samples of size n have the same chance of being selected (Cochran 1977).

Systematic Sampling

Systematic sampling is a cost-effective design in which the samples are uniformly distributed throughout the population (Cochran 1977, Scheaffer et al. 2006). Systematic sampling minimizes travel time, compared to SRS, making it more cost efficient in that it provides the same amount of information at minimal cost (Scheaffer et al. 2006, Stevens

and Olsens 2004). If the variable of interest is randomly distributed, estimates of the population mean and variance are identical to that of SRS.

A simple sequential inhibition process (SSI) (Reich and Davis 2007a) was used to simulate a systematic sample. The SSI process randomly locates sample plots within the population with the constraint that no two points can be within a given distance of one another. Since the state is irregular in shape, initially 2900 and 6200 samples were located with a minimum distance of 7200 m and 4800 m between sample points, respectively. Sample plots that did not fall within the state or forested areas were removed. The remaining sample plots were randomly thinned to a size of 500 and 1100 plots. A spatial library (Reich and Davis 2007b) for R version 2.4.0 (R Development Core Team 2006) was used to simulate the SSI process.

Stratified Random Sampling

A stratified random sample (ST) is obtained by dividing the population elements into non-overlapping groups, known as strata, and then selecting a simple random sample from each stratum (Johnson 2000, Scheaffer et al. 2006). The objective of this design is to create homogenous subgroups with minimum variance within stratum. If done correctly, a stratified random sample should be more precise than a simple random or systematic sample.

Simulation Study

A Monte Carlo simulation was used to compare the three sampling designs (Fishman 1995 and Manly 1998) in modeling the spatial distribution of forest tree biomass. For each Monte Carlo simulation, a sample of 500 or 1100 30 m x 30 m sample plots were randomly selected using SRS, SSI and ST to obtain a set of data for modeling

the spatial distribution of forest tree biomass. This process was repeated 50 times to evaluate the predictive performance of the three sampling designs. Example of the distribution of sample locations simulated using SRS, SSI and ST are displayed in Figure 16.

Modeling the Spatial Distribution of Biomass

The spatial variability in forest tree biomass was modeled using an approach developed by Reich and Aguirrie-Bravo (2008). Let $z(s_i)$ represent a sample value of the target variable Z at spatial location s_i. Also, assume the sample data contains a set of auxiliary variables (covariates) X, the values of which are known for all units in the population. Multiple regression is used to describe the large-scale spatial variability in the data as a linear function in p known explanatory variables $x_j(s_i)$

$$z(s_i) = \beta_0 + \sum_{j=1}^{p} x_j(s_i)\beta_j + \eta(s_i) \tag{28}$$

where $\beta_j, j = 0,\ldots p$ are p+1 unknown regression coefficients and $\eta(s_i)$ is an error process sometimes referred to as a random field, with $E[\eta(s_i)] = 0$ and covariance $C(x, y) = Cov(\eta(x_i), \eta(y_j))$. The error term in Eq. 28 is unknown because the true model is unknown. Once the model parameters have been estimated, the regression residuals are defined as $\hat{\eta}(s_i) = z(s_i) - \hat{z}(s_i)$, where $\hat{z}(s_i)$ is the predicted value at spatial location s_i given the explanatory variables $x_j(s_i)$. The error process can be expressed as

$$\eta(s_i) = \hat{\eta}(s_i) + \mu(s_i) \tag{29}$$

with $E[\mu(s_i) \mid \hat{\eta}(s_i)] = 0$. Using the set of auxiliary variables, X as a basis of stratification assume

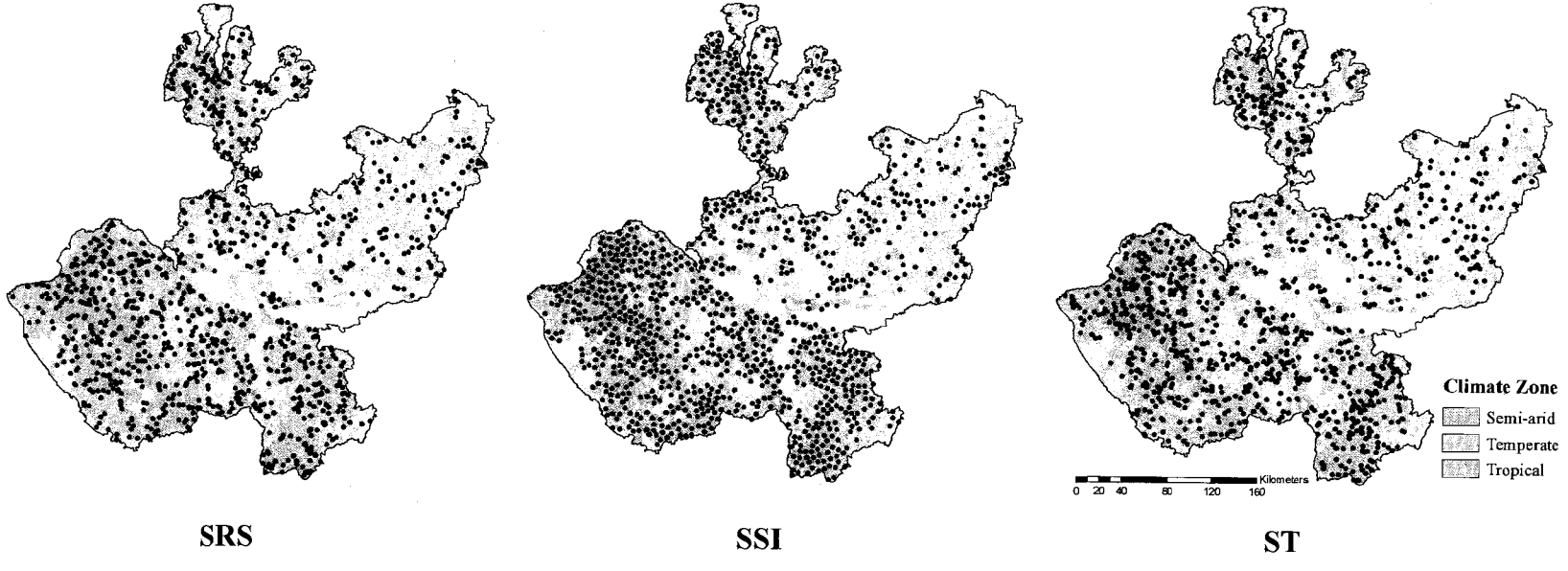


Figure 16. Spatial distribution of sample locations for simple random sampling (SRS), systematic sampling (SSI) and stratified random sampling (ST) with a sample size of 1100.

$$\hat{\eta}(s_i) = f(x(s_i)) + \delta(s_i) \tag{30}$$

with $E[\delta(s_i) \mid x(s_i)] = 0$, $f(x(s_i))$ is a deterministic function, and $\delta(s_i)$ is a zero-mean random term (Cocchi et al. 2002). Combining Eq. 29 and Eq. 30

$$\eta(s_i) = f(x(s_i)) + \varepsilon(s_i) \tag{31}$$

with $E[\eta(s_i) \mid f(x(s_i))] = f(x(s_i))$, $\varepsilon(s_i) = \mu(s_i) + \delta(s_i)$, and $E[\varepsilon(s_i) \mid x(s_i)] = 0$ provided that $\mu(s_i)$ and $\delta(s_i)$ are conditionally independent (Cocchi et al. 2002; Benedtti et al. 2005). The mean function $f(x(s_i))$ is estimated by $\hat{f}$ using the recursive partitioning method introduced by Brieman et al. (1984). Combining Eq. 28 and Eq. 29 the full model describing the spatial variability in the sample data is given by

$$z(s_i) = \beta_0 + \sum_{j=1}^{p} x(s_i)\beta_j + f(x(s_i)) + \varepsilon(s_i). \tag{32}$$

Variance Estimation

The variance of the estimated mean response at a given location s_i, for a set of explanatory variables, $x(s_i)$ is given by

$$\text{var}(z(s_i)) = \text{var}(\hat{\eta}(s_i)) + \text{var}\left(\hat{\delta}(s_i)\right) \tag{33}$$

where $\text{var}(\hat{\eta}(s_i))$ reflects the uncertainty in estimating the parameters of the regression model and $\text{var}\left(\hat{\delta}(s_i)\right)$ reflects the uncertainty in estimating the error ($\hat{\eta}$) of the regression model. The variance associated with an estimate at a new location, s_0, can be written as

$$\text{var}(z(s_0)) = \text{var}(\hat{\eta}(s_0)) + \text{var}(z(s_0)) + \text{var}\left(\hat{\delta}(s_0)\right) \tag{34}$$

where the additional term, $\text{var}(z(s_0))$ reflects the random variation at a new location, s_0.

The small-scale variability (i.e., estimated errors from the regression models) in biomass was modeled using a tree-based stratified design (Reich and Aguirrie-Bravo 2008). Independent variables considered in the stratification included elevation, slope, aspect, Landsat ETM+ bands and land cover type and predicted biomass obtained from a multiple regression model.

For implementing the tree-based approach, the option "*minsize*" in the *tree* function and the option "*best*" in the *prune.tree* function available in S-Plus (Insightful Crop. 2000) were fixed for a given set of simulations. The option "*minszie*" reflects the minimum number of observations required to split the data. The option "*prune.tree*" is used to prune the complete tree to have a desired number of partitions (i.e., tree size). Based on a preliminary study using only one simulated dataset of forest tree biomass, the cost complexity criterion (Reich and Aguirrie-Bravo 2008, Chapter 1) suggested that conditions set on the *minsize* and tree size for the two sample sizes were optimal in terms of variance estimation. In this study the options with *minsize* of 15 and tree size of 50 were used for the sample size of 500 and the options with *minsize* of 5 and tree size of 100 were applied to the sample size of 1100 to meet the requirement of unbiased estimates of the variacne.

A generalized linear model (McCullagh and Nelder, 1989, Chambers and Hastie 1992) was used to estimate the regression coefficients and variances associated with the large-scale variability of forest tree biomass. The *stepAIC* function (Venables and Ripley 1999), avaialbe in the *MASS* Package in S, was used to perform a backward stepwise selection procedure identifying significant predictors of a multiple regression model based on Akaike Information Criterion (AIC) (Akaike 1969). To stabilize the variance of

biomass, a square root transformation was applied to the sample data of biomass. The *tree* function in S-Plus platform was used to perform tree-based approach for modeling the small-scale variability in forest tree biomass.

Predicted values and associated standard errors (*se.fit* = *T*) of the fitted regression models were obtained using the function, *glm.pred <- predict.glm(object, predict="response", se.fit=T),* where *object* is a fitted glm model. The variance associated with estimating the regression coefficients were obtained by $\text{var}(\hat{\eta}(s_i))$= *glm.pred$se.fit* ^{2} while the uncertainty in estimating the error of the regression model, $\text{var}(\hat{\delta}(s_i))$ were calculated using standard methods for a stratified random sample (Cochran, 1977):

$$\hat{\bar{\delta}}_k = \frac{1}{n_k}\sum_{l=1}^{n_{ki}} \hat{\delta}_{kl} \tag{35}$$

and

$$\hat{\sigma}_k^2 = \frac{\sum_{l=1}^{n_k}\left(\hat{\delta}_{kl} - \hat{\bar{\delta}}_k\right)^2}{n_k - 1} \tag{36}$$

where k denotes the stratum, $\hat{\bar{\delta}}_{ki}$ denotes the mean residual error for observations assigned to the k^{th} stratum, n_k is the number of observations assigned to the k^{th} stratum, and $\hat{\sigma}_k^2$ is the within stratum variance for the k^{th} stratum. The use of the sample variance as a measure of the uncertainty in estimating the error of the regression model is justified by the fact that the mean square error is the best constant predictor given that the sample data belong to the stratum.

Jagger (2005) suggested that the function *predict.glm()* correctly calculates the variance for constructing confidence limits but those for the prediction intervals are not correct since the variance component due to the response is taken as the dispersion value. Assuming normality for the prediction interval, Jagger (2005) recommends calculating the variance used in constructing prediction intervals by *family()$var(fitted value) * dispersion + se(fitted value)^2*. Estimates of the variances used in constructing confidence and prediction intervals were computed by:

$$\text{var}(\hat{z}(s_i)) = se(fitted\ value)\text{^}2 + \hat{\sigma}_k^2(s_i) \quad (37)$$

$$\text{var}(\hat{z}(s_0)) = family()\$var(fitted\ value) * dispersion + se(fitted\ value)\text{^}2 + \hat{\sigma}_k^2(s_i) \quad (38),$$

respectively.

Model Evaluation

A 9-fold cross-validation (Efron and Tibshrani, 1993) was used to evaluate the predictive performance of the fitted biomass models. The sample data were divided into 9 parts (K = 9), each of which consisted of 55 and 122 observation for sample sizes of 500 and 1100, respectively. The fitted models were recursively fitted using eight parts (K-1) of the data as a training data set and the remaining data as an independent data set for estimating the prediction errors. Repeating this procedure nine times allow each observation to be excluded from the model and independently predicted by fitted models. Following this procedure, a set of statistics were calculated to evaluate the predictive performance of the models. Estimates of prediction errors obtaining from the K-fold cross validation (Kravchenko and Bullock 1999, Schloeder et al. 2001 and Reich et al. 2004) were compared to asses the effectiveness of the three sampling designs in modeling forest tree biomass.

The effectiveness of the fitted models was evaluated using a goodness-of-prediction statistic (G-statistic) (Agterberg, 1984; Kravchenka and Bullock, 1999; Guisan and Zimmermann, 2000; Schloeder et al., 2001).

$$\text{G-statistic} = 1 - \left(\frac{\sum_{i=1}^{n} [z(s_i) - \hat{z}(s_i)]^2}{\sum_{i=1}^{n} [z(s_i) - \bar{z}(s_i)]^2} \right). \tag{39}$$

The G-statistic is a measure of the effectiveness of a prediction relative to that which could have been derived using the sample mean. A G-statistic equal to one indicates perfect prediction, a positive value indicates a more reliable model than if one had used the sample mean, a negative value indicates a less reliable model than if one had used the sample mean, and a value of zero indicates that the sample mean should be used to estimate $z(s_i)$.

The mean absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |z(s_i) - \hat{z}(s_i)| \tag{40}$$

and the mean squared error of prediction (MSEP)

$$MSEP = \frac{1}{n} \sum_{i=1}^{n} [z(s_i) - \hat{z}(s_i)]^2 \tag{41}$$

were used to evaluate the accuracy of the predictions, where $z(s_i)$ is the actual value at a sample point i, $\hat{z}(s_i)$ is the estimated value at a sample location i obtained from the 9-fold cross validation, and n is the total number of samples used in the 9-fold cross validation.

The standardized mean squared error (SMSE) (Reich et al. 2004), was used to evaluate the reliability between the estimated variances and the true errors:

$$SMSE = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varepsilon}^2(s_i)}{\mathrm{var}(\hat{z}(s_i))} \tag{42}$$

where, $\hat{\varepsilon}(s_i) = (z(s_i) - \hat{z}(s_i))$, is the true error and $\mathrm{var}(\hat{z}(s_i))$ is the estimated variance of the mean response *i* obtained from Eq. 33 for the standardized mean square error of the variance of the mean response (SMSEM) and from Eq. 34 for the standardized mean square error of the prediction variance (SMSEP). SMSE has a Chi-square distribution with *n* degree of freedom which can be used to construct a confidence interval for SMSE under the null hypothesis of equal variances:

$$\Pr\left[\frac{\chi^2_{n,\frac{\alpha}{2}}}{n} \le SMSE \le \frac{\chi^2_{n,1-\frac{\alpha}{2}}}{n}\right] = 1-\alpha\,. \tag{43}$$

When *n* is large, SMSE can be approximated by a standard normal distribution with a mean of one and variance $2/n$ (SMSE ~ $N(1, 2/n)$). If the SMSE falls within the interval $1 \pm 1.96*(2/n)^{0.5}$, this would indicate that the true errors and estimation errors are consistent at the 0.95 level of confidence (Hevesi et al 1992, Reich et al. 2004). SMSE's were evaluated based on the same *minsize* and tree size for all three sampling designs.

These intervals were used to calculate coverage rates (CV) which are defined as the proportion of individual confidence intervals and prediction intervals containing the observed value. The 95% confidence and prediction intervals were calculated assuming normality

$$\hat{z}(s_0) \pm 1.96\sqrt{\mathrm{var}(\hat{z}(s_0))}\,. \tag{44}$$

Multi-Response Permutation Procedure

Multi-response permutation procedures (MRPP) were used to test for the significant differences among the three sampling designs and the two sample sizes. MRPP (Zimmerman et al. 1985, Mielke and Berry 2001) is a multivariate, nonparamentric statistics for testing for significant difference among groups of multivariate data. Unlike parametric statistics such as a *t*-test or *F*-test, Euclidean distances between all pairs of observations in multi-dimensional space are used to compute a test statistic. The test statistics does not rely on a standard normal distribution, but it is considered a distribution-free technique. Permutation procedures are used to develop a reference distribution under the null hypothesis for the purpose of testing for significant differences. The possible test statistics obtained from the permutation procedures under the null hypothesis of no difference are compared to the test statistic obtained from the observed data. A p-value is computed from the proportion of test statistics equal to or less than the observed statistic. Thus, a small p-value would indicate significant difference among groups. A detailed description of MRPP is provided by Mielke and Berry (2001).

In this study, MRPP was used to simultaneously test significant differences among the test statistics used to evaluate the predictive performance of the three sampling designs and two sample sizes. If a significant difference was detected all pair-wise combinations were evaluated to facilitate comparisons among the three sampling designs.

RESULTS

Average R^2 values associated with the regression methods and G-statistic for the combined models indicated that stratified random sampling for sample sizes of 500 and 1100 provided the best fit in terms of accounting for the total variability observed in the biomass data (Table 5 and 6). MRPP suggested that the R^2 for the stratified design was significantly larger compared to the other two sampling designs for the sample size of 500 (p-value < 0.001) and 1100 (p-value < 0.001) (Figure 17). MRPP also indicated that the R^2 values associated with the simple random and systematic designs were not significantly different from one another (n = 500, p-value = 0.533; n = 1100, p-value = 0.261) (Figure 17). Similarly, no significant differences were observed for the G-statistics for these two sampling designs (n = 500, p-value = 0.47; n = 1100, p-value = 0.79) (Figure 17).

All three sample designs provided unbiased variance estimate for the n = 500 sample size (Table 6). For the sample size of 1100 only the stratified design provided unbiased variance estimates due to the larger tree size used to model the small-scale variability in biomass. Noticeably, the averages for SMSEM and SMSEP associated with the stratified design for both sample sizes were closet to one. MRPP suggested that SMSEM's associated with the stratified design were significantly different from those associated with the simple random (p-value < 0.002) and systematic (p-value = 0.005) designs for the sample size of 500 (Figure 18), whereas, SMSEM's for the sample size of 1100 indicated no significant difference among the three sampling design (p-value = 0.22). Comparing the SMSEP's obtained from the three sampling designs, the pair-wise

Table 5. Summary statistics for comparing the spatial models of biomass (tones/ha) developed using simple random sampling (SRS), systematic sampling (SSI), and stratified random sampling (ST) for a sample size of 500.

Statistics[1]	Design	Mean	95% Lower[2]	95% Upper[2]	Median	SD[ζ]	Min	Max
$R^{2ζ}$	SRS	0.49	0.49	0.50	0.49	0.03	0.43	0.56
	SSI	0.50	0.49	0.51	0.50	0.03	0.42	0.56
	ST	0.52	0.51	0.53	0.52	0.03	0.46	0.58
G-statatic[ζ]	SRS	0.76	0.75	0.77	0.76	0.02	0.72	0.81
	SSI	0.76	0.75	0.77	0.76	0.02	0.70	0.81
	ST	0.77	0.76	0.78	0.77	0.02	0.73	0.80
SMSEM[ζ]	SRS	0.92	0.91	0.92	0.92	0.01	0.88	0.94
	SSI	0.92	0.91	0.92	0.92	0.01	0.89	0.94
	ST	0.91	0.91	0.91	0.91	0.01	0.88	0.94
CRM[ζ]	SRS	0.975	0.973	0.977	0.977	0.006	0.960	0.986
	SSI	0.974	0.972	0.975	0.974	0.006	0.956	0.986
	ST	0.975	0.973	0.976	0.976	0.006	0.964	0.986
SMSEP[ζ]	SRS	1.06	1.03	1.08	1.07	0.10	0.86	1.24
	SSI	1.05	1.02	1.08	1.04	0.11	0.85	1.31
	ST	0.96	0.94	0.99	0.97	0.10	0.77	1.21
CRP[ζ]	SRS	0.943	0.939	0.947	0.943	0.014	0.915	0.974
	SSI	0.944	0.940	0.948	0.944	0.015	0.899	0.972
	ST	0.951	0.947	0.954	0.954	0.012	0.915	0.974
MSEP[ζ]	SRS	16.68	16.31	17.05	16.64	1.29	14.35	19.36
	SSI	16.75	16.33	17.17	16.55	1.48	14.32	21.18
	ST	16.56	16.13	16.98	16.65	1.50	13.35	20.46
MAEP[ζ]	SRS	3.23	3.19	3.27	3.22	0.13	2.98	3.50
	SSI	3.24	3.20	3.28	3.22	0.15	2.93	3.64
	ST	3.17	3.13	3.22	3.18	0.14	2.84	3.47
MEAN (sample)	SRS	65.91	65.14	66.68	65.99	2.72	59.89	72.65
	SSI	67.18	66.24	68.11	66.88	3.29	58.24	73.87
	ST	66.17	65.19	67.15	66.28	3.44	58.95	74.00
MEAN (model)	SRS	60.96	60.19	61.72	60.63	2.69	54.86	67.65
	SSI	62.20	61.28	63.12	62.19	3.23	53.13	69.44
	ST	61.13	60.05	62.20	61.41	3.77	53.05	70.23

[1]The statistics are based on 50 simulations of each sampling design.
[2]The lower and upper confidence bound were constructed using a t-distribution, $t_{.025,\,49} = 2.31$.
[ζ]SD = standard deviation, R^2 = coefficient of determination obtained from 3SLS, G-statistic = the total variability of the mean response accounted by 3SLS + RT, SMSEM = standardized mean square error of the model, CRM = confidence coverage rates of the mean response, SMSEP = standardized mean square error of prediction, CRP = confidence coverage rates of prediction, MSEP = mean square error of the mean response, MAEP = mean absolute error of prediction.

Table 6. Summary statistics for comparing the spatial models of biomass (tones/ha) developed using simple random sampling (SRS), systematic sampling (SSI), and stratified random sampling (ST) for a sample size of 1100.

Statistics[1]	Designs	Mean	95% Lower[2]	95% Upper[2]	Median	SD[ζ]	Min	Max
$R^{2\zeta}$	SRS	0.48	0.47	0.49	0.48	0.02	0.42	0.52
	SSI	0.47	0.47	0.48	0.48	0.02	0.42	0.53
	ST	0.51	0.50	0.51	0.51	0.02	0.46	0.56
G-statatistic[ζ]	SRS	0.76	0.76	0.77	0.76	0.01	0.74	0.79
	SSI	0.76	0.76	0.76	0.76	0.01	0.73	0.78
	ST	0.78	0.78	0.78	0.78	0.01	0.75	0.80
SMSEM[ζ]	SRS	1.12	1.08	1.16	1.09	0.14	0.95	1.61
	SSI	1.15	1.10	1.19	1.10	0.15	0.95	1.54
	ST	1.09	1.05	1.12	1.06	0.12	0.94	1.45
CRM[ζ]	SRS	0.957	0.955	0.959	0.958	0.007	0.940	0.968
	SSI	0.956	0.954	0.958	0.955	0.006	0.943	0.965
	ST	0.959	0.957	0.961	0.959	0.007	0.945	0.972
SMSEP[ζ]	SRS	1.12	1.10	1.14	1.12	0.07	0.98	1.29
	SSI	1.15	1.12	1.17	1.14	0.08	0.93	1.35
	ST	1.04	1.02	1.06	1.04	0.07	0.89	1.23
CRP[ζ]	SRS	0.935	0.932	0.937	0.935	0.009	0.907	0.952
	SSI	0.932	0.929	0.935	0.932	0.010	0.908	0.961
	ST	0.942	0.939	0.944	0.942	0.009	0.922	0.960
MSEP[ζ]	SRS	16.67	16.44	16.90	16.65	0.81	15.04	18.62
	SSI	16.88	16.59	17.17	16.92	1.02	15.06	20.18
	ST	16.83	16.56	17.10	16.93	0.94	14.88	18.38
MAEP[ζ]	SRS	3.21	3.18	3.23	3.19	0.09	3.06	3.44
	SSI	3.22	3.19	3.25	3.23	0.10	3.05	3.47
	ST	3.17	3.15	3.20	3.17	0.10	2.94	3.35
MEAN (sample)	SRS	66.69	66.12	67.26	67.10	2.01	62.77	70.50
	SSI	66.55	65.96	67.15	66.68	2.09	62.83	70.72
	ST	65.93	65.38	66.48	66.18	1.93	60.90	70.94
MEAN (model)	SRS	61.79	61.20	62.37	62.38	2.07	57.98	65.24
	SSI	61.63	61.06	62.20	61.79	2.00	58.41	66.52
	ST	60.81	60.15	61.48	60.97	2.35	54.30	65.48

[1]The statistics are based on 50 simulations of each sampling design.
[2]The lower and upper confidence bound were constructed using a t-distribution, $t_{.025,\,49} = 2.31$.
[ζ]SD = standard deviation, R^2 = coefficient of determination obtained from 3SLS, G-statistic = the total variability of the mean response accounted by 3SLS + RT, SMSEM = standardized mean square error of the model, CRM = confidence coverage rates of the mean response, SMSEP = standardized mean square error of prediction, CRP = confidence coverage rates of prediction, MSEP = mean square error of the mean response, MAEP = mean absolute error of prediction.

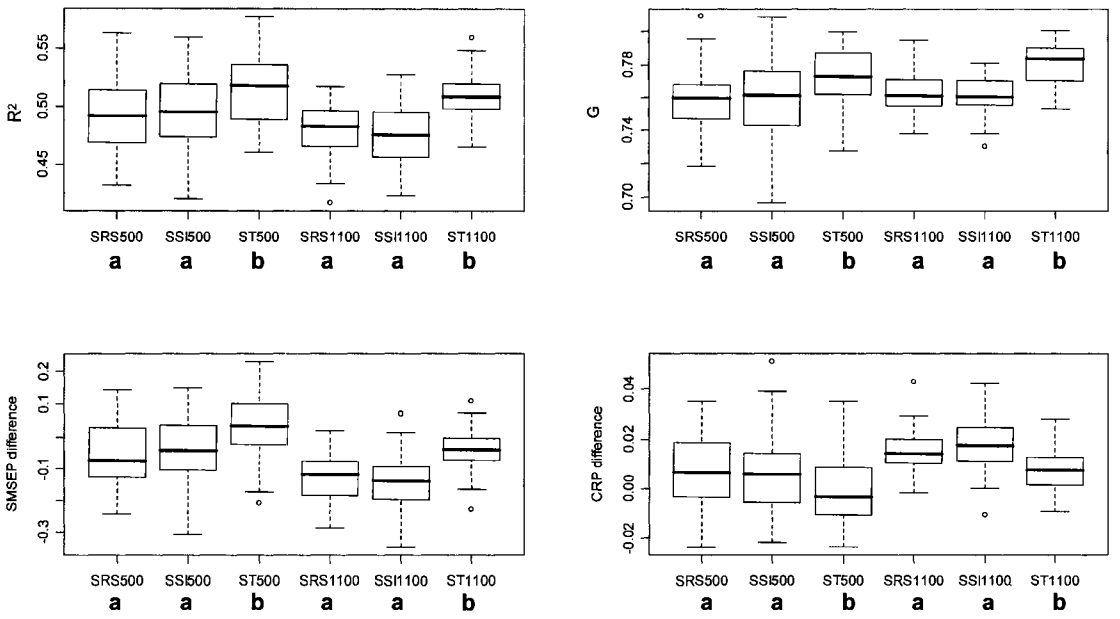


Figure 17. Box plots comparing model statistics for simple random sampling (SRS), systematic random sampling (SSI), and stratified random sampling (ST) using a sample size of 500 and 1100. The letter below the plots indicates a pair-wise comparison among the sampling designs using MRPP (R^2 = proportion of the observed variability accounted for by the ordinary least square model, G = proportion of the observed variability accounted for by the ordinary least square model plus the binary regression tree, SMSEP = standardized mean square error of the prediction, CRP = prediction coverage rate).

comparisons of MRPP indicated significant differences between the stratified and simple random designs (p-value < 0.001) and the stratified and systematic designs (p-value < 0.001) (Figure 17). MRPP also suggested that SRS and SSI were not significantly different for the two sample sizes (n = 500, p-value = 0.39; n= 1100, p-value = 0.21).

The average CRM's for the two sample sizes and the three sampling designs were significantly different from the nominal 0.95 level based on a two-tailed *t*-test (Table 5 and 6). However, the sample size of 1100 provided a coverage rate (CRM) closer to the nominal 0.95 rate. MRPP suggested that all sampling designs were not significantly different for the sample size of 500 (p-value = 0.36) and 1100 (p-value = 0.11) (Figure 18). Unlike CRM, the stratified design with a sample size of 500 resulted in an average CRP of 0.951 which was not significantly different from the nominal 0.95 rate (Table 5). Comparing the averaged CRP's obtained using a sample size of 1100, the stratified design yielded the closest value (0.942) to 0.95. MRPP also indicated that CRP's associated with the stratified design were significantly different from the simple random (p-value = 0.004) and systematic (p-value = 0.04) designs for the sample size of 500 (Figure 17). For a sample size of 1100, CRP's associated with the stratified design were significantly different from those for the simple random (p-value < 0.001) and systematic (p -value < 0.001) designs (Figure 17).

With respect to MSEP and MAEP, all three sampling designs provided similar results (Table 5 and 6) for both sample sizes. The stratified design provided a smaller MSEP and MAEP, especially with the smaller sample size. MRPP also confirmed that there was no significant difference among three sampling designs and sample sizes (Figure 18).

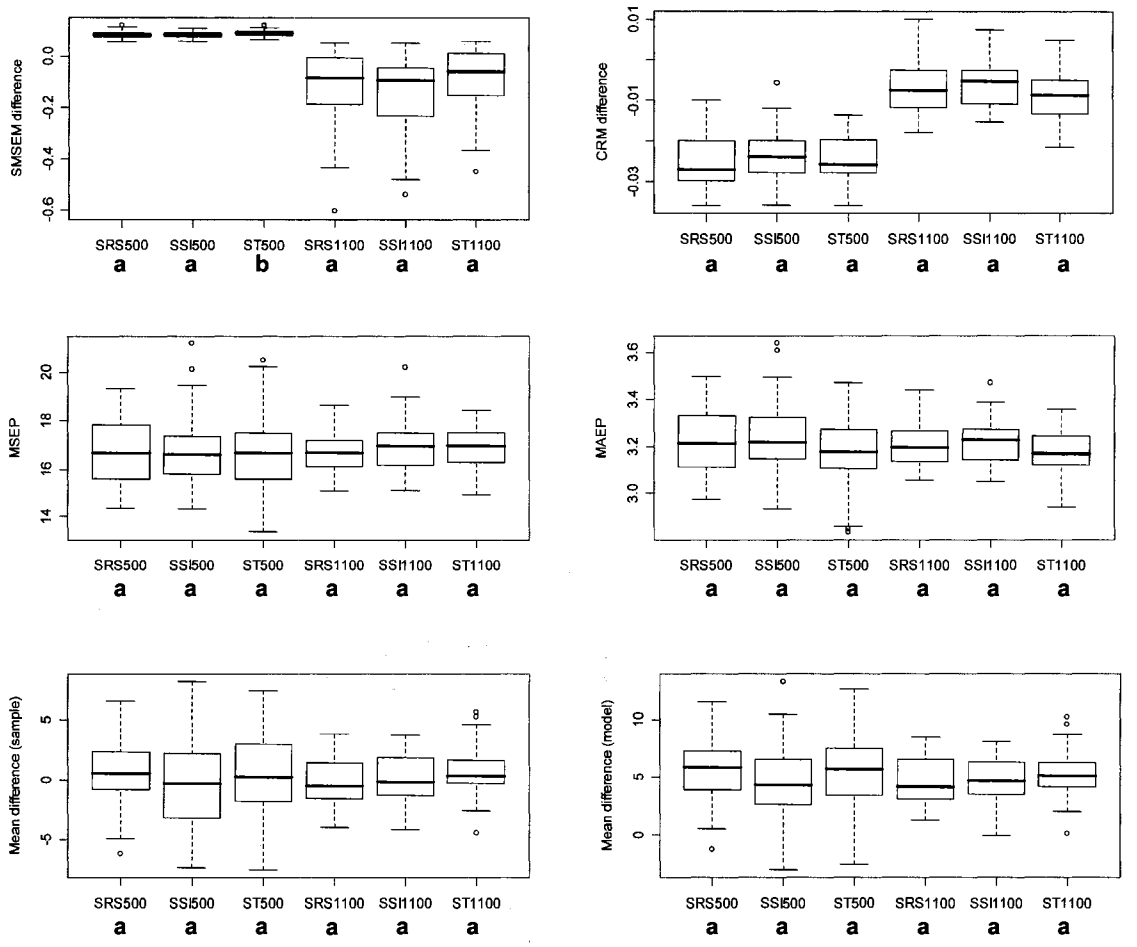


Figure 18. Box plots comparing model statistics for simple random sampling (SRS), systematic random sampling (SSI), and stratified random sampling (ST) using a sample size of 500 and 1100. The letter below the plots indicates a pair-wise comparison among the sampling designs using MRPP (SMSEM = standardize mean square error of the model, CRM = confidence coverage rate of the model, MSEP = prediction mean square error, MAEP = mean absolute error of prediction).

Reliable estimates of the population mean were obtained from the three sampling designs and two sample sizes. Based on a two-tailed *t*-test, the sample means were not significantly different from the simulated population mean ($\alpha = 0.05$) for both sample sizes (Table 5 and 6). MRPP indicated no differences among the means for the three sampling designs and two sample sizes (Figure 18).

Estimated population means obtained from the spatial models were significantly lower than the true population mean (Table 5 and 6). All three sampling designs systematically underestimated the population means by approximately 5 tones/ha for both sample sizes. This bias is due to the process used in creating hypothetical surface of biomass. Table 7 provides evidence that the hypothetical surface used in this study had larger amounts of biomass than the original biomass surface with respect to the three climatic zones, or the state as a whole. The process of adding the random noise and truncating negative estimates of biomass to zero resulted in a higher proportion of pixels with zero biomass, while pixels with a positive biomass increased. The net effect was to increase the average biomass for the state by 3.8 tones/ha. The range and the mean of hypothetical surface also suggested that many extreme values of biomass were introduced into the surface. As a result, the sample data over-sampled areas with little to no biomass and under-sampled area with high biomass resulting in a systematic bias in the estimates. In spite of this systematic bias, it does not detract from the results presented in this study.

Table 7. Comparison of the original biomass surface to the hypothetical biomass surface.

Climatic Zone	Pixels with Biomass > 0 (%)	Pixels with Biomass > 0 (%)	Range (tones/ha)	Range (tones/ha)
Original Biomass Surface				
Semi-arid	1.89	16.63	621.19	45.20
Temperate	2.12	56.42	699.96	59.01
Tropical	0.37	22.58	708.41	86.16
Entire State	4.38	95.62	708.41	62.66
Hypothetical Biomass Surface				
Semi-arid	4.64	13.88	681.00	50.64
Temperate	10.84	47.69	779.41	62.99
Tropical	2.34	20.60	827.74	88.28
Entire State	17.83	82.17	827.74	66.46

Comparing Predicted Biomass Surfaces

The final surfaces of predicted forest tree biomass were created based on models developed from data obtained using three different sampling designs and two sample sizes. To facilitate comparison, models were selected with similar R^2 values and G-statistic. The means and associated mean square errors for the three sampling designs are summarized in Table 8 by climatic zones. In general, the systematic design tended to produce reliable estimate of the mean biomass for the semi-arid and temperate zone, while the stratified design provided more reliable estimates of the mean biomass in the temperate and tropical zone. However, the spatial model based on the stratified design was best in terms of the MSE.

Figure 19 and 20 shows the distribution of errors (truth – predicted) throughout the state for the two sample sizes. Negative values indicate an overestimation of biomass while positive values indicate an underestimation. The three sampling designs displayed similar trends of underestimation in the tropical zones which are characterized as having high biomass. However, the underestimation of biomass was generally confined to small areas when using the stratified design. In the semi-arid and temperate zone, the stratified design overestimated less when compared to the other two designs.

Table 8. Summary statistics for the final predictive surfaces of biomass (tones/ha) based on simple random sample (SRS), systematic sampling (SSI), and stratified random sampling (ST) using a sample size of 500 and 1100.

Sampling Designs	Statistics	Sample Size	$R^{2\varsigma}$	G^{ς}	Climatic Zone		
					Semi-arid	Temperate	Tropical
SRS	**Mean**	**500**	0.500	0.759	41.88	55.58	74.67
	MSE$^{\varsigma}$	**500**			3182.62	3859.95	6668.99
SSI	**Mean**	**500**	0.502	0.760	50.22	57.95	64.77
	MSE	**500**			3116.57	3944.19	6835.96
ST	**Mean**	**500**	0.526	0.759	42.48	53.89	83.47
	MSE	**500**			2714.11	3266.19	6085.39
SRS	**Mean**	**1100**	0.480	0.762	42.64	60.33	74.48
	MSE	**1100**			7096.35	8510.59	7062.07
SSI	**Mean**	**1100**	0.475	0.761	52.17	60.64	79.02
	MSE	**1100**			4129.71	4791.41	4502.07
ST	**Mean**	**1100**	0.487	0.761	46.35	59.95	75.89
	MSE	**1100**			2842.53	3542.11	2866.42
Population Mean (tones/ha)					50.64	62.99	88.28
Number of Pixels					9578434	30281203	11868912

$^{\varsigma}$MSE = mean square error of the difference between the hypothetical surface and the predictive surface of biomass, R^2 = proportion of the observed variability accounted for by the ordinary least square model, G = proportion of the observed variability accounted for by the ordinary least square model plus the binary regression tree.

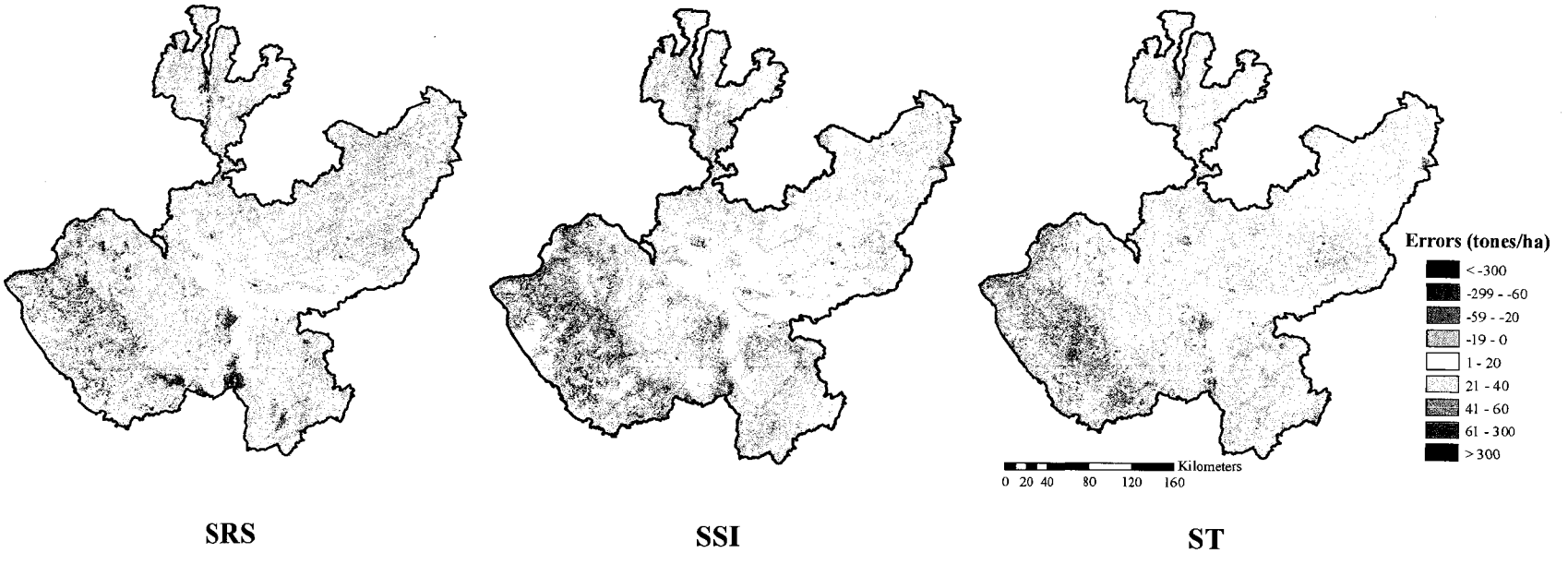


Figure 19. Spatial distribution of raw errors (truth – predicted) obtained from simple random sampling (SRS), systematic sampling (SSI) and stratified random sampling (ST) using the sample size of 500.

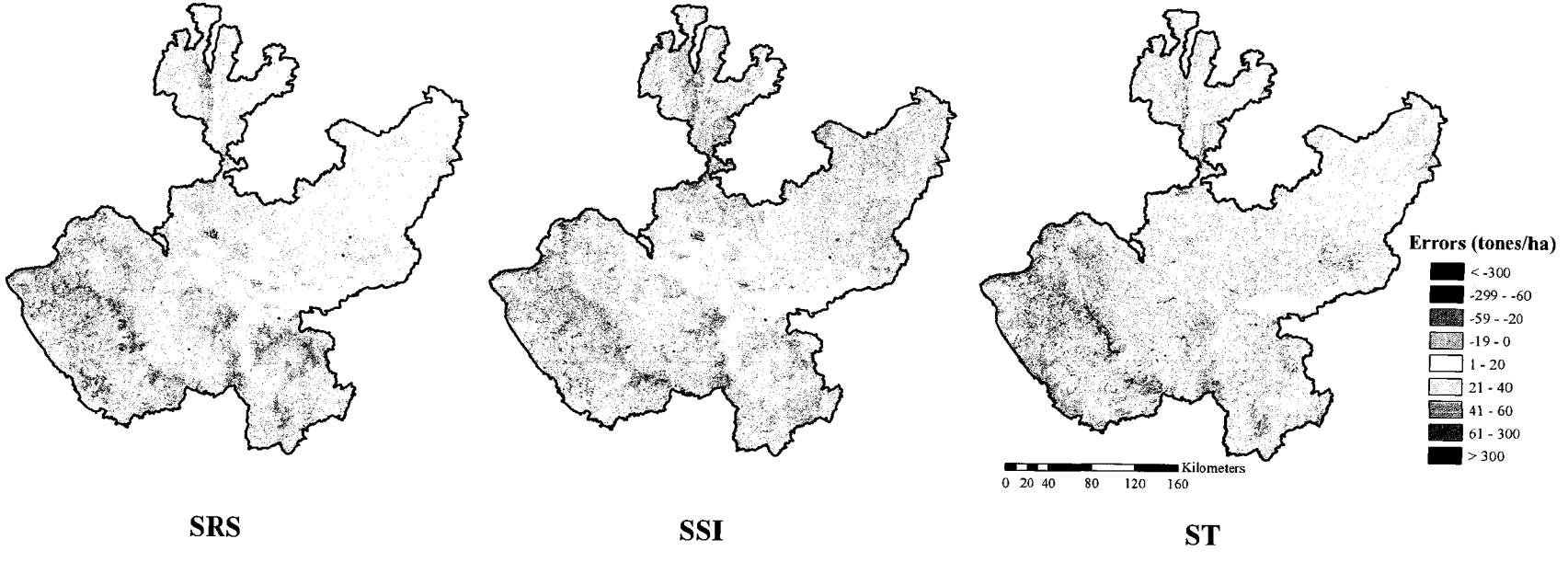


Figure 20. Spatial distribution of raw errors (truth – predicted) obtained from simple random sampling (SRS), systematic sampling (SSI) and stratified random sampling (ST) using the sample size of 1100.

DISCUSSION

In this study, three sampling designs were evaluated in terms of developing a model to describe the spatial distribution of forest tree biomass in the State of Jalisco, Mexico. Model evaluation indicated that stratifying the population based on the spectral properties of the vegetation produced a better model than if this information was ignored when designing the survey. All three sampling designs provided unbiased estimates of the variance for the two sample sizes evaluated.

Both the simple random and systematic designs allocated samples proportional to the size of the spectral classes resulting in some spectral classes being under-sampled and other over-sampled. Thus, these two designs may not be capturing the extent of the spectral variability in the population. This is especially true for the forest types that occur along the ridge tops in the western portion of the state. By allocating the sample uniformly across all spectral classes, the stratified design was able to capture more of the variability in the landscape for the same sampling effort when compared to the simple random and systematic design. With a sample size of 500 and 1100, the simulation study suggested that on the average the stratified design represented 12 to 14 out of 15 vegetation classes that occurred in the state, compared to 10 and 11 vegetation classes for the simple random and systematic designs, respectively.

While the spatial models produced biased estimates of the mean biomass, this was attributed primarily to the method used in adding the random noise and does not detract from the results obtained in this study. It should be pointed out that this systematic bias was not present in the original forest tree biomass model developed by Reich et al. (2008).

The agreement between the sample mean from the state-wide inventory (55.18 tones/ha) and the spatial model (56.10 tones/ha) were consistent with one another.

Stevens and Olsen (2004) and Stevens (2006) suggested that a systematic sample tends to spatially balance sample locations throughout a large spatial domain and therefore should potentially provide better statistical estimates. However, the results of this study indicated that both systematic and simple random sampling behaved the same in terms of their predictive performance. The results also suggested that the stratified design provided the best estimates of forest tree biomass.

Very few studies have compared methods of modeling spatial data over large geographical regions in which the data lacks strong spatial dependency. In some situations, a geostatistical model such as kriging suggests ways to allocate the samples in order to capture the spatial dependency in the data. Most studies dealing with natural resources (e.g., Pettitt and Mcbarney 1993, Papritz and Webster 1995, Brus and De Gruijter 1997, Jardim and Ribeiro 2007) pay little to no attention on how to collect the sample data for developing a predictive model. Thus, direct comparisons of the results of this study to other studies are not applicable. However the results from this study agreed somewhat with a study by Paprttz and Webster (1995). The authors pointed out that with a limited sample size, stratified random samples provided accurate and precise estimates of soil attributes when applying kriging.

CONCLUSION

Digital mapping of natural resource is important for decision making and management of ecosystems and natural resources. The spatial distribution of natural resource attributes can be developed using reliable spatial modeling techniques. Numerous sampling strategies have been suggested by research scientists to collect spatial data to develop such models. The results of this study provide evidence that the allocation of sample units based on spectral variability of the landscape could improve the predictive performance of certain types of spatial models. The use of satellite imagery provides detailed information on the spatial variability on the variable of interest throughout the landscape. The use of a stratified design based on prior knowledge of the spectral variability of the population of interest should increase the accuracy and precision of the statistical estimates of the population. The approach advocated in this study could benefit research scientists as well as managers interested in studying a variety of natural resources phenomenon that occur over large geographical regions.

REFERENCES

Agterberg, F. P. 1984. Trend surface analysis. In: Spatial Statistics and models, G. L. Gaile and C. J. Willmott (eds.). Reidel, Dordrecht, The Netherlands, pp. 147-171.

Akaike, H. 1969. Fitting autoregressive models for prediction. Annals of the Institute for Statistics and Mathematics 21, 243–247.

ESRI. 2006. ArcGIS Version 9.2. Windows 2000 /XP Server Operating System. Build 800. Redlands, CA.

ERDAS, Inc. 1999. Earth Resource Data Analysis System Field Guide, Fifth Edition. ERDAS, Inc., Atlanta, Georgia, 672 pp.

Benedetti, R., G. Espa, and G. Lafratta. 2005. A tree-based approach to forming strata in multipurpose business surveys. Discussion Paper No. 5 2005, Dipartimento di Economia, Universita Degli Studi di Trento, Trento, Italy. 17 pp.

Breiman, L., J. H. Friedman, R.A. Olshen, and C. J. Stone. 1984. Classification and regression. Tress. Wadsworth International Group, Belmont, CA. 368 pp.

Brus, D. J. and J. J. de Gruijter. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1-44.

Chambers, J. M. and T. J. Hastie. 1992. Statistical models in S. Wadsworth and Brooks/Cole. CA. 608 pp.

Cocchi, D., E. Fabrizi, M. Raggi and C. Trivisano. 2002. Regression trees based stratification: An application to the analysis of the Italian post enumeration survey. Proceedings of the International Conference on Improving Surveys, August 25–28, 2002, Copenhagen, Denmark. [online] URL: http://www.icis.dk/ICIS-papers/B2_5_2.pdf.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. Wiley and Sons, New York. pp. 428.

Conquest, L.L. 2003. Model-assisted sampling approaches in the sampling of natural resources. American Statistical Association, Section of Statistics and the Environment, News letter: V.5, No 1.

Efron, B. and R. J. Tibshrani. 1993. An introduction to the bootstrap. Chapman and Hall. New York, 456 pp.

ESRI. 1995. ARC/INFO Software and on-line help manual. Environmental Research Institute, Inc., Redlands, CA.

ESRI. 2006. Environmental Systems Research Institute, Inc., 380 New York St., Readlands, CA 97393. USA.

Fishman, G.S. 1995. Monte Carlo: Concepts, Algorithms, and Applications, Springer Verlag, New York, USA. 728 pp.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, D. Tyler. 2002. The national elevation dataset. Photogrammetric Engineering & Remote Sensing 68, 5-11.

Guisan, A. and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 47-186.

Haining, R. 2003. Spatial data analysis: theory and practice. Cambridge University Press, Cambridge. 432 pp.

Hevesi, J. A., J. D. Istok, and A. L. Flint. 1992. Precipitation estimation in mountainous terrain using multivariate geostatistics. Part I: structural analysis. Journal of Applied Meteorology 31, 661-676.

Insightful Corp. 2000. S-Plus 2000 for Windows. Professional Edition. Lucent Technologies, Inc.

Jagger, 2005. [online] URL: http://www.biostat.wustl.edu/arcives/html/s-news/2005-03/msg0053.html

Jardim E. and P. J. Ribeiro Jr.. 2007. Geostatistical assessment of sampling designs for Portuguese bottom trawl surveys. Fisheries Research 85, 239-247.

Johnson, E. W. 2000. Forest sampling desk reference. CRC Press, Florida. 985 pp.

Kravchenko, A and D. G. Bullock. 1999. A comparative study of interpolation methods for mapping soil properties. Agronomy Journal 91, 393-400.

Lark, R. M. and B. R. Cullis. 2004. Model-based analysis using REML for inference from systematically sampled data on soil. European Journal of Soil Science 55, 799-813.

Manly, B. F. J. 1998. Randomization, Bootsrap and Monte Carlo Methods in Biology. Chapman & Hall. London. 399 pp.

McCullagh, P. and J.A. Nelder, 1989. Generalized linear models. 2nd ed. Chapman and Hall, London p. 511 pp.

Mielke Jr., P. W. and K. J. Berry. 2001. Permutation methods: A distance function approach. Springer-Verlag. New York. 352 pp.

Paprttz A. and R. Webster. 1995. Estimating temporal change in soil monitoring: I. Statistical theory. European Journal of Soil Science 46, 1-12.

Pettitt A. N. and A. B. McBratney. 1993. Sampling Designs for Estimating Spatial Variance Components. Applied Statistics 42, 185-209.

Rabus, B., M. Eineder, A. Roth, and R. Bamler. 2003. The shuttle radar topography mission—a new class of digital elevation models acquired by spaceborne radar. ISPRS Journal of Photogrammetry & Remote Sensing 57, 241- 262.

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. [online] URL: http://www.R-project.org.

Reich, M. R., J. E. Lundquist, and V. A. Bravo. 2004. Spatial models for estimation fuel loads in the Black Hills, South Dakota, USA. International Journal of Wildland Fire 13, 119-129.

Reich, R. M. and R. A. Davis. 2007a. Quantitative spatial analysis: course note for NR/ST523. Colorado State University, Fort Collins, CO, 481 pp.

Reich, R. M., and R. A. Davis. 2007b. 'On-line spatial library for R package.' (Colorado State University: Fort Collins, CO). [online] URL: http://www.warnercnr.colostate.edu/~robin/

Reich, R.M. and C. Aguirrie-Bravo. 2008. Empirical Evaluation of Confidence and Prediction Intervals for Spatial Models of Forest Structure in Jalisco, Mexico. Journal of Environmental Statistitics. (In review)

Reich, R.M., C. Aguirrie-Bravo and M. A. Mendoza Briseño. 2008. An innovative approach to inventory and monitoring of natural resources in the Mexican State of Jalisco. Journal Environmental Monitoring and Assessment. (In press)

Rzedowski, J. 1978. Vegetación de México. Editorial Limusa. México, D.F, Mexico.

Scheaffer, R. L., W. Mendenhall III, and L. Ott. 2006. Elementary Survey Sampling (6^{th} Ed.). Duxbury Press. California. 180 pp.

Schloeder, C. A., N. E. Zimmermann and M. J. Jacobs. 2001. Comparison of methods for interpolating soil properties using limited data. American Society of Soil Science Journal 65, 470-479.

Stevens D. L. Jr. 2006. Spatial properties of design-based versus model-based approaches to environmental sampling. Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, July 5-7, Lisbon, Portugal. [online] URL: http://www.spatial-curacy.org/2006/PDF/Stevens2006accuracy.pdf

Stevens D. L. Jr. and A. R. Olsen. 2003. Variance estimation for spatially balanced samples of environmental resources. Environmetrics 14, 593–610.

Stevens D. L. Jr. and A. R. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99, 262–278.

Theobald, D. M., D. L. Jr. Stevens, D. White, N. S. Urquhart, A. R. Olsen and J. B. Norman. 2007. Using GIS to Generate Spatially Balanced Random Survey Designs for Natural Resource Applications. Journal Environmental Management 40, 134-146.

Venables, W. M. and B. D. Ripley. 1999. Modern Applied Statistics with S-PLUS. Springer-Verlag, New York. 501 pp.

Warren, W.G. 1998. Spatial analysis of marine populations: factors to be considered. Canadian Special Publication, Fish and Aquatic Science. 125:21-28.

Zimmerman, G. M., H. Goetz and P. W. Mielke Jr. 1985. Use of an improved statistical method for group comparisons to study effects of prairie fire. Ecology 66(2), 606-611.

CHAPTER 3: OPTIMAL PLOT SIZE FOR ESTIMATING TREE BASAL AREA, TREE DENSITY AND SPECIES ABUNDANCE FOR A SEASONAL DRY EVERGREEN FOREST IN THAILAND

ABSTRACT

No specific plot size is optimal for estimating all variables of interest in a forest inventory. Finding an optimal plot size is critical in designing a cost efficient forest inventory. This study focused on how plot size and sampling intensity influenced estimates of tree basal area, tree density and number of tree species in a seasonal dry evergreen forest in Thailand. The data used in the study comes from a mapped 50 ha plot in which the location, size and species of all trees with a DBH > 1 cm were known. The results of a simulation study indicated that plots ranging in size from 5 m x 5 m to 50 m x 50 m provided unbiased estimates of basal area and tree density irrespective of the sampling intensity. Nonparametric estimator of the total number of tree species provided reliable estimates when using a large number of small plots. Equations are presented to express the total time, or cost associated with estimation basal area/ha, tree/ha, and number of tree species as a function of plot size. These equations are used to estimate optimal plot size for different tract sizes, coefficient of variations and percent sampling errors. Increasing the variability within a population decreased the optimal plot size, while increasing the allowable error increased the optimal plot size. Larger tract sizes required a fewer number of larger plot sizes to minimize the increased cost associated with travel time.

INTRODUCTION

Uneven-aged stands of tropical forests are diverse in both species composition and their structure. Because of this diversity, it is not straightforward on how best to sample these forests to obtain reliable estimates of things such as timber volume, basal area, tree density or even the number of tree species. One critical aspect of this is the selection of a plot size, and is one plot size optimal for all conditions. The goal of any survey is to make an inference about a population based on a representative sample selected from the population. Should one use many small plots or a few large plots? The answer to this question is not straightforward, and needs to be addressed to ensure the most efficient use of the resources available.

Seasonal dry evergreen forests in Thailand are one of the most valuable forest type. This forest type contributes not only to the socio-economic well-being of the local Thai people but is also a crucial component of the tropical ecosystem. Due to rapidly decreasing landbase associated with this forest type, numerous studies (e.g., Baker 1997 and 2001, Bunyavejchewin 1986 and 1999, Bunyavejchewin et al. 2001) have been conducted to understand its ecological function and process. Unfortunately, previous studies have rarely focused on the most efficient method of collecting this type of information. In many situations, the size of the plots chosen are based on the preference of the researchers and/or because of a traditional protocol without any scientific support. Obviously, finding the most appropriate plot size with respect to plot measurement time and travel time has never been attempted in a seasonal dry evergreen forest in Thailand. Thus, the main objectives of this study are:

1) evaluate the influence of plot size on the statistical properties of estimates of tree basal area, tree density and number of tree species in a seasonal dry evergreen forest, and
2) develop a set of equations to estimate the total cost of a survey as a function of plot size, sampling variability and the desired sampling error.

To achieve these goals, a permanent 50 ha plot representing a seasonal dry evergreen forest in Hua Kha Khaeng (HKK) Wildlife Sanctuary, Thailand, was used in this study.

METHODS

Study Area

Huai Kha Khaeng Wildlife Sanctuary (HKK) is one of the 17 protected areas forming the Western Forest Complex (WFC) of Thailand. The HKK covers an area of 2780 km^2 in west central Thailand. The region is characterized by a 5-6 month dry season extending from November to April. Mean annual rainfall is approximately 1400 mm. Located within this region is one of several large-scale permanent Forest Dynamic Plots (FDP) which is part of a larger network of permanent plots established under the guidance of the Center for Tropical Forest Science, Smithsonian Tropical Research Institute. The HKK permanent plot is 50 ha (500 m x 1000 m) in size and is located at 15°40′ N latitude and 99°10′ E longitude, about 4 km west of Kapook Kapieng Ranger Station in the northern part of HKK (Figure 21). The location of the plot was chosen to represent the climatic conditions of seasonal dry evergreen forests in Southeast Asia. Elevations on the plot range from 549 m to 638 m. The plot is oriented with the long axis aligned in the north-south direction (Figure 22) (Bunyavejchewin et al. 1998)

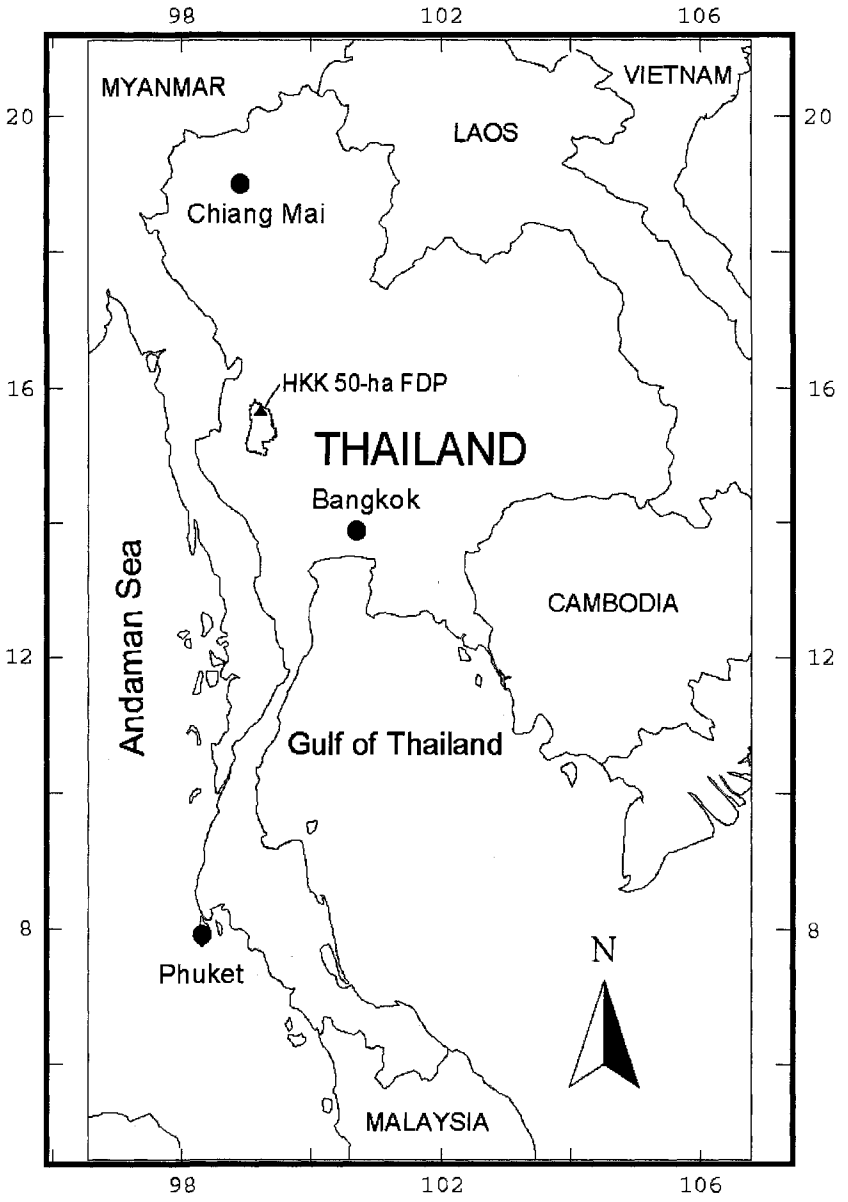


Figure 21. The permanent 50 ha plot is located in Huai Kha Khaeng (HKK) Wildlife Sanctuary, western Thailand.

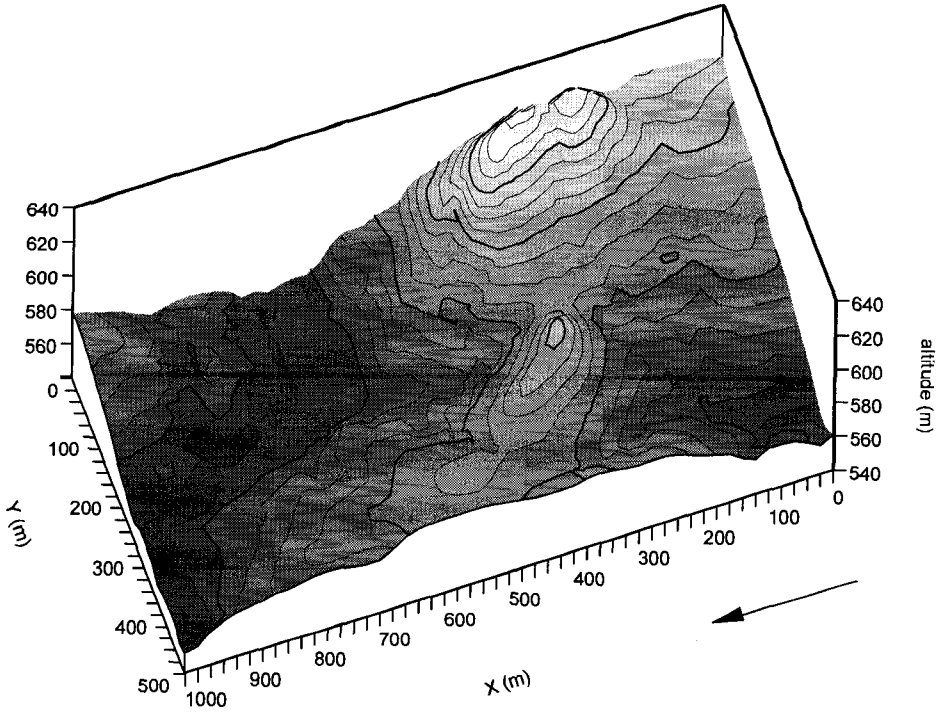


Figure 22. The 50 ha forest dynamics plot in 3D. Contour lines represent a 5 m interval of elevation.

Tree Census

The FDP census was conducted in 1994. All tree standing woody plants with a DBH (diameter at breast height) greater than 1 cm measured 1.30 m above ground, were tagged and mapped following a standard protocol (Manokaran et al. 1990, Condit 1998, and Bunyavejchewin et al. 2001). Estimates of basal area/ha and number of trees/ha were based on trees with a DBH grater than 10 cm while the total number of tree species was based on all trees with a DBH greater than 1 cm.

The floristic structure of the HKK FDP was documented by Bunyavejchewin (2001). The plot contained 248 species of 164 genera and 61 families. The five most common families in terms of tree basal area were *Dipterocarpaceae* (21.95%), *Annonaceae* (19.36%), *Lauraceae* (7.81%), *Euphorbiaceae* (4.13%), and *Sapindaceae* (5.53%) respectively. *Annonaceae* (20.87%), *Eurphorbiaceae* (18.95%), *Sapindaceae* (12.30%), *Rubiaceae* (6.09) and *Lauraceae* (5.69%) were the five most abundant families based on tree density. The five most diverse families included *Euphobiaceae* (12.08%), *Moraceae* (7.00%), *Leguminoseae* (6.23%), *Rubiaceae* (5.06%), and *Sapindaceae* (4.28%) respectively. Seven species of *Dipterocarpaceae* forming the upper canopy and emergent layers included *Anisoptera costata*, *Dipterocarpus alatus*, *D. obtusifolius*, *Hopea odorata*, *Shorea siamensis*, *S, roxburghii* and *Vatica cinerea.*

Plot Configuration

Five plot sizes were evaluated in this study: 5 m x 5 m, 10 m x 10 m, 20 m x 20 m, 25 m x 25 m, and 50 m x 50 m. The 50 ha plot was sub-divided into N = 20000, 5000, 1250, 800 and 200 non-overlapping disjoint region corresponding to the five plot sizes,

respectively. To facilitate comparison among the various plot sizes, the proportion of the 50 ha plot sampled was fixed at 0.5%, 1.0%, 2.0%, 5.0%, 10.0%, and 15.0%. Sample sizes associated with the various plot sizes and sample intensities are summarized in Table 9.

Simulation Study

For a given plot size and sample intensity, a random sample of n plots were selected without replacement to obtain estimates of basal area per ha, trees/ha and number of tree species. Estimates of the mean ($\bar{y}$) basal area/ha and trees/ha were computed as follows:

$$\bar{y} = \frac{1}{an}\sum_{i=1}^{n} y_i \tag{45}$$

with estimated variance:

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{a}\right)^2 \frac{s^2}{n} \tag{46}$$

where, y_i is the i^{th} observation, $s^2 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ is the sample variance and a is the plot size in hectares (Cochran 1977).

Estimating the number of tree species in the population is not as simple as just counting the number of species on the sample plots or calculating an arithmetic mean as in the case of estimating basal area/ha or number of trees/ha. Bunge and Fitzpatrick (1993, 1995), Chao and Lee (1992), and Schreuder et al. (1999) presented some useful nonparametric estimators for estimating the total number of tree species in temperate forest in the U.S. using field data. In this study, the applicability of these nonparametric estimators were calculated for use in the tropical forests of Thailand.

Table 9. Sample sizes associated with the different sampling intensity and plot sizes.

Plot Size (m x m)	Sampling Intensity (%)					
	0.5	**1.0**	**2.0**	**5.0**	**10.0**	**15.0**
5 x 5	100	200	400	1000	2000	3000
10 x 10	25	50	100	250	500	750
20 x 20	7	13	25	63	125	188
25 x 25	4	8	16	40	80	120
50 x 50	-	2	4	10	20	30

Nonparametric estimators evaluated in this study were

$$CCOVf = \frac{c_s}{1 - fc_1 / n}, \tag{47}$$

$$CM2f = c_s + fc_1^2 / (2c_2), \tag{48}$$

$$CM3f = f\frac{c_s}{1 - c_1 / n} + \frac{fc_1(cv)^2}{1 - c_1 / n}, \tag{49}$$

$$CP1f = \frac{c_s}{1 - f\exp(-2c_2 / c_1)}, \tag{50}$$

$$CMBf = c_s + \frac{(n-1)c_1 f}{n}, \tag{51}$$

where c_s is the total number of species counted on the sample plots, c_1 is the number of species occurring only once, c_2 is the number of species occurring twice, $\pi_{sp} = \frac{n_{sp}}{n}(100)$ is the percentage of plots species *sp* occurred on, $f = \left(\frac{N-n}{N}\right)$ is a finite population correction factor (*fpc*), and

$$\mathrm{cv}^2 = \frac{\sum(\pi_{sp} - \hat{\pi})^2 / (c_s - 1)}{\hat{\pi}^2}, \text{ with } \hat{\pi} = \sum \pi_{sp} / c_s \tag{52}$$

If one is sampling an infinitely large population, one can ignore the finite population correction factor. The nonparametric estimators were evaluated with and without the *fpc*. In selecting an estimator, it is desirable to have one that does not overestimate the true number of tree species on the 50 ha plot and should be as close to the true value as possible.

Since no valid formula is available for calculating the variance associated with these nonparametric estimators, a bootstrap procedure was used to estimate the sample

variance (Smith and Belle 1984, Efron and Tibshirana 1994, Shao and Tu 1995, Manly 1998). Two-hundred interactions were used to estimate the sample variances associated with the nonparametric estimators.

To evaluate the statistical properties of the nonparametric estimators for estimating the number of tree species, Monte Carlo simulation (Fishman 1995) were used to sample the 50 ha plot $M = 20000$ times for the 5 m x 5 m plot size and $M = 50000$ times for all other plot sizes. An overall sample mean ($\hat{\mu}$) for the M simulations was calculated as

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} \bar{y}_i \ . \tag{53}$$

where $\bar{y}_i$ is the sample mean for the i^{th} simulation. The bias of the estimator was computed as follows:

$$\text{Bias (\%)} = \left(\frac{\frac{1}{M}\sum_{i=1}^{M} \bar{y}_i - \mu}{\mu} \right) 100 \tag{54}$$

where μ is the true number of tree species on the 50 ha plot. To asses any bias associated with estimating the sample variance, the mean variance

$$\hat{\bar{V}}(\bar{y}) = \frac{1}{M}\sum_{i=1}^{M} \hat{V}(\bar{y}_i). \tag{55}$$

and the variance of the means

$$s_{\bar{y}}^2 = \frac{\sum_{i=1}^{M} (\bar{y}_i - \hat{\mu})^2}{M-1}. \tag{56}$$

were computed. This latter variance was assumed to be an estimate of the true variance. An *F*-test was used to test if the ratio of these two variances differed significantly from one.

Optimal Plot Size

In practice, most surveys are constrained by a fixed budget limiting the number of sample plots that can be established and measured. In some instances, one has the opportunity to design a survey using a plot size that will minimize the total cost of the survey and still maintain the desired level of precision.

The total time of a survey for a given tract size can be described as a function of the average travel time between plots, the average plot measurement time and number of sample plots as follows:

$$T_i = n_i(v_i + m_i) \tag{57}$$

where T_i is the total time of the survey using a plot of size Q_i, n_i is the sample size for a plot of size Q_i, v_i is the average travel time between plots, and m_i is the average measurement time for a plot of size Q_i. Gambill et al. (1985) demonstrated the procedure of determining an optimal plot size using fixed-area plot while Reich and Arvantis (1992) used the same procedure for variable plot sampling. For an infinitely large population one can calculate the sample size using the follow formula:

$$n_i = \frac{CV_i^2 t^2}{E^2} \tag{58}$$

where, CV_i is the coefficient of variation associated with a plot of size Q_i, t is the Student's t value and E is the allowable sampling error (%).

The coefficient of variation generally decreases with increasing plot size, thus requiring fewer samples to achieve the same percent sampling error. The relationship between the coefficient of variation and plot size can be expressed as follows:

$$CV_i = kQ_i^c \tag{59}$$

where c and k are constants. Reich and Arvantis (1992) noted that the c parameter indicates the degree of randomness for a given forest stand characteristic.
A c - coefficient of 0.5 suggested that the variable of interest is randomly distributed. If c is smaller than 0.5, then sampling variability will increase faster than a random population. This implies that the variable of interest is aggregated. A c - coefficient greater than 0.5 implies that the variable of interest has a regular distribution.

To express the relationship in how the coefficient of variation changes with a change in plot size, consider the follow ratio: where the subscripts i and j link the plot size Q to its coefficient of variation CV:

$$\frac{CV_i}{CV_j} = \frac{Q_i^c}{Q_j^c}. \tag{60}$$

Rearranging this equation, one can develop a relationship describing how a change in plot size changes the coefficient of variation:

$$CV_i = CV_j\left(\frac{Q_i^c}{Q_j^c}\right). \tag{61}$$

Substituting this relationship into Eq. 58, and simplifying, the formula for sample size becomes

$$n_i = \omega Q_i^{2c} \tag{62}$$

where $\omega = \frac{t^2 CV_j^2}{E^2 Q_j^{2c}}$.

Assuming a systematic sample with a square spacing, the distance between sample plots is given by

$$d_i = \sqrt{\frac{10000W}{n_i}} \tag{63}$$

where, W is a target tract size, in ha, and n_i is the sample size. The average travel time between plots (sec) is given by

$$v_i = \frac{d_i}{S} \tag{64}$$

where S is the rate of travel (m/sec). Substituting Eq. 62 and Eq. 63 into this equation, it is possible to express the rate of travel as a fuction of plot size:

$$v_i = \frac{100}{Q_i^c}\sqrt{\frac{WS}{\omega}}. \tag{65}$$

Gambill et al. (1985) showed that plot measurement time is nonlinearly related to a plot of size Q_i as follows

$$m_i = B_0 z_i^{B_1} (\ln z_i)^{B_2} \tag{66}$$

where B_0, B_1 and B_2 are regression coefficients, $z_i = \frac{Q_i e}{Q_j}$, and e is the base for natural logs (or equivalent to 2.71828). Finally substituting Eq. 62, Eq. 65 and Eq. 66 into Eq. 57, the total time, or cost of a survey for a plot of size Q_i can be expressed as follows:

$$T_i = 100\omega\sqrt{\frac{WS}{\omega}}Q_i^c + \omega Q_i^{2c} B_0 z_i^{B_1} (\ln z_i)^{B_2}. \tag{67}$$

To model the relationship between plot measurement time and plot size (Eq. 66), a series of concentric plots measuring 5 m x 5 m, 10 m x 10 m, 20 m x 20 m, 25 m x 25 m and 50 m x 50 m ($n = 4$ for each plot size) were randomly allocated in the HKK dry evergreen forest. The time required to estimate basal area/ha, number of trees/ha, and total number of tree species were recorded for each plot size. The average rate of travel time was determined by establishing four different routes of 100 m long within the HKK dry evergreen forest and recording the time required to walk the 100 meters. Routes were subjectively selected to represent different conditions of accessibility (e.g., amount of ground cover and slope) in the dry evergreen forest types. To model the relationship between plot size and the coefficient of variation, 20 concentric plots of size 5 m x 5 m, 10 m x 10 m, 20 m x 20 m, 25 m x 25 m and 50 m x 50 m were randomly located within the 50 ha plot. For each plot size the coefficient of variation associated with estimating basal area/ha, trees/ha and number of tree species was estimated.

RESULTS

Basal Area per Hectare

Based on the census of the HKK FDP in 1994, the population mean basal area for woody trees $\geq$ 10 cm in DBH was 29.31 m^2 per ha. In general, all plot sizes irrespective of the sampling intensity provided unbiased estimates of basal area (Table 10, Figure 23a). However, the smallest plot size of 5 m x 5 m with a 0.5% sampling intensity ($n = 100$) showed the highest bias. The mean variance decreased with increasing sampling intensity for all plot sizes (Figure 23b). Variances ratios were not significantly different from one, except for the 50 m x 50 m plot with a 1% sampling intensity ($n = 2$) (Figure 23c).

Table 10. Influence of plot size and sampling intensity in estimate the mean and variance of basal area and tree density using Monte Carlo simulations.

Plot Size		Sampling Intensity (%)					
(m x m)	Estimator[1]	0.5	1	2	5	10	15
				Basal Area (m^2/ha)			
5 x 5	Mean	29.52	29.30	29.29	29.32	29.32	-
	Variance	78.13	36.79	18.35	7.05	3.36	-
10 x 10	Mean	29.26	29.35	29.28	29.31	29.31	29.30
	Variance	69.52	36.03	17.57	6.88	3.26	2.05
20 x 20	Mean	29.33	29.29	29.30	29.34	29.32	29.31
	Variance	62.43	33.93	17.49	6.76	3.21	2.01
25 x 25	Mean	29.26	29.31	29.29	29.31	29.30	29.31
	Variance	70.15	35.80	17.62	6.84	3.24	2.04
50 x 50	Mean	-	29.34	29.33	29.31	29.30	29.30
	Variance	-	44.70	22.02	8.57	4.03	2.55
				Density (trees/ha)			
5 x 5	Mean	439.10	438.54	438.23	438.40	438.26	-
	Variance	1714.62	851.47	421.16	163.32	77.36	-
10 x 10	Mean	438.35	438.47	438.29	438.30	438.26	438.21
	Variance	1845.20	916.76	453.91	175.81	83.29	52.46
20 x 20	Mean	438.26	438.22	438.28	438.30	438.26	438.31
	Variance	2284.61	1228.72	630.14	242.88	115.75	72.78
25 x 25	Mean	438.15	438.65	438.17	438.26	438.16	438.29
	Variance	3098.22	1535.09	763.20	295.74	140.17	88.14
50 x 50	Mean	-	438.69	438.32	438.32	438.36	438.26
	Variance	-	3225.05	1603.76	628.90	296.24	186.61

[1]Estimated means and variance for the 5 m x 5 m plot was based on 20000 simulations, while all other estimates were based on 50000 simulations.

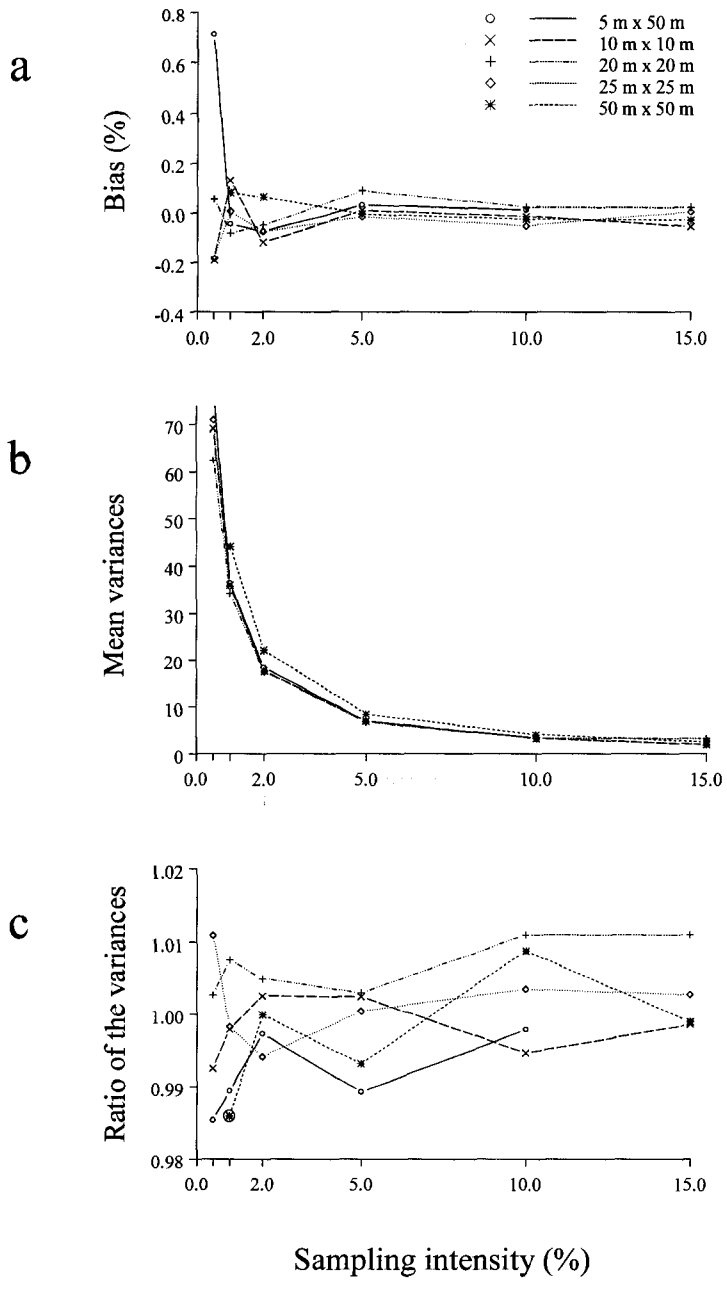


Figure 23. Influence of plot size and sampling intensity on a) percent bias, b) estimated sample variance, and c) ratio of the mean variance to the variance of means for estimating basal area per hectare. Significant differences are indicated by a circle.

Number of Trees per Hectare

The tree density determined from the census data was 438.28 trees/ha. Similar to basal area/ha, all plot sizes and sampling intensities provided unbiased estimates of tree density (Table 10 and Figure 24a). Similarly, the estimated variances decreased with increasing plot size for all sampling intensities (Figure 24b). The 20 m x 20 m plot size with a 0.5% (n = 7) and 2% (n = 25) sampling intensity yielded biased estimate of the variance (Figure 24c).

Number of Tree Species

A total of 244 species were identified on the 50 ha plot. Candidate estimators were chosen in terms of their accuracy and the fact that they did not overestimate the true number of tree species. Only two estimators, *CM3f* and *CP1f* satisfied these conditions. Consequently, the results will focus on only these two estimators. Summary statistics for *CM3f* and *CP1f* are given in Table 11.

CM3f: The percent bias associated with *CM3f* for different plot sizes and sampling intensities are shown in Table 11 and Figure 25a. Plot sizes and sampling intensity that yielded biased estimates of the total number of tree species are marked by a circle. This statistical significance was based on a z-test. In general, small sample sizes underestimated the number of tree species irrespective of the plot size. Estimates associated with the 5 m x 5 m and 10 m x 10 m were unbiased at a sampling intensity of 2% and 5% but were biased with sampling intensities of 10% and 15%. Estimates using plot sizes of 20 m x 20 m and 25 m x 25 m were unbiased at sampling intensities exceeding 10%. The 50 m x 50 m plot size underestimated the number of tree species for all sampling intensities.

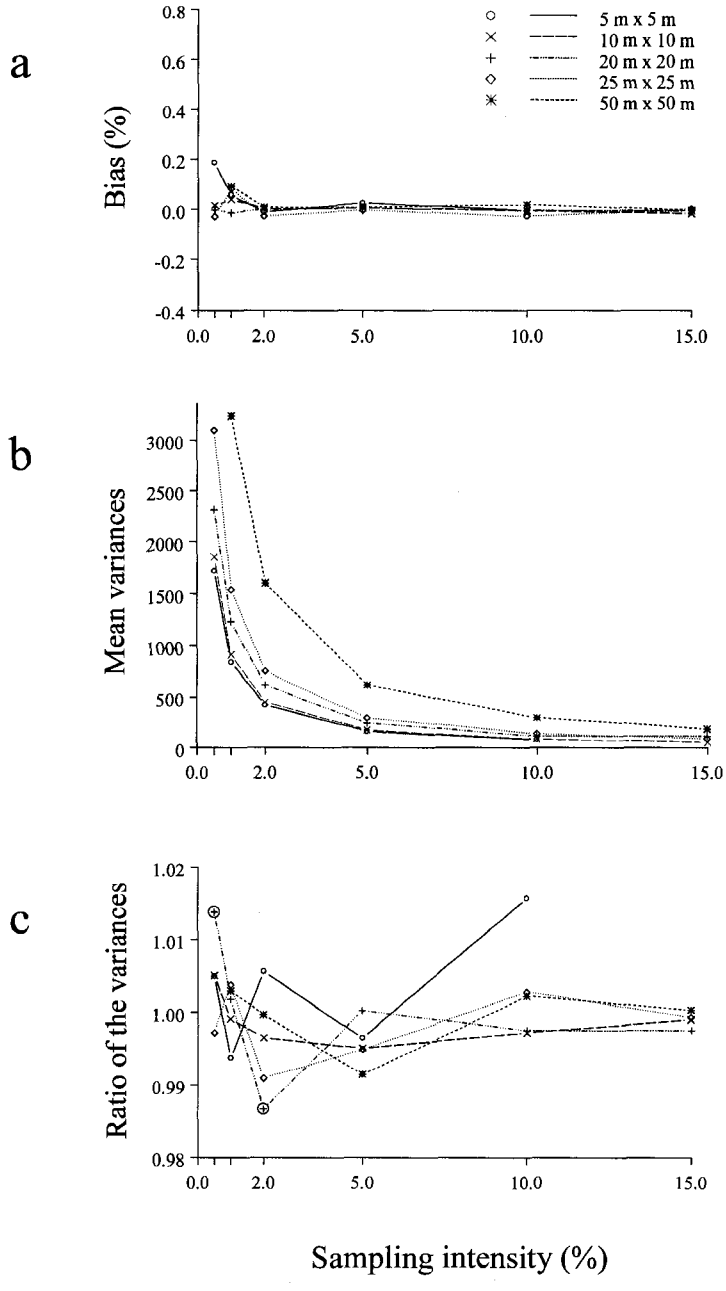


Figure 24. Influence of plot size and sampling intensity on a) percent bias, b) estimated sample variance, and c) ratio of the mean variance to the variance of means for estimating a number of trees per hectare. Significant differences are indicated by a circle.

Table 11. Influence of plot size and sampling intensity in estimating the mean and variance of the number of trees species using Monte Carlo simulations.

Plot Size (m x m)	Estimator[1]	Sampling Intensity (%)					
		0.5	1	2	5	10	15
				CM3f			
5 x 5	**Mean**	113.81	158.74	208.54	270.47	310.13	-
	Variance	206.60	336.85	490.34	667.87	749.31	-
10 x 10	**Mean**	93.25	139.07	183.89	241.76	279.76	297.73
	Variance	50.83	212.44	322.70	441.35	496.96	498.73
20 x 20	**Mean**	72.59	98.34	138.38	204.07	239.19	256.57
	Variance	38.11	40.06	46.14	220.31	252.88	253.74
25 x 25	**Mean**	65.92	90.45	123.56	190.18	226.56	243.87
	Variance	41.24	40.31	40.09	114.97	189.93	188.04
50 x 50	**Mean**	-	81.37	107.36	146.00	181.25	206.10
	Variance	-	62.64	53.83	42.35	37.14	36.34
				CP1f			
5 x 5	**Mean**	138.33	167.52	193.28	233.23	261.95	-
	Variance	1684.02	1695.50	1563.72	1753.45	1776.26	-
10 x 10	**Mean**	116.48	164.27	190.62	232.15	276.85	272.12
	Variance	618.12	1548.83	1516.13	1702.10	2803.80	1542.07
20 x 20	**Mean**	80.51	104.31	149.06	230.28	256.01	268.04
	Variance	50.64	65.16	275.71	1661.49	1530.53	1404.73
25 x 25	**Mean**	75.01	99.98	129.32	223.29	255.34	267.22
	Variance	55.55	51.93	62.23	1163.83	1526.43	1357.49
50 x 50	**Mean**	-	93.96	122.65	161.93	194.38	233.48
	Variance	-	83.55	71.73	53.73	91.22	356.01

[1]Estimated means and variance for the 5 m x 5 m plot was based on 20000 simulations, while all other estimates were based on 50000 simulations.

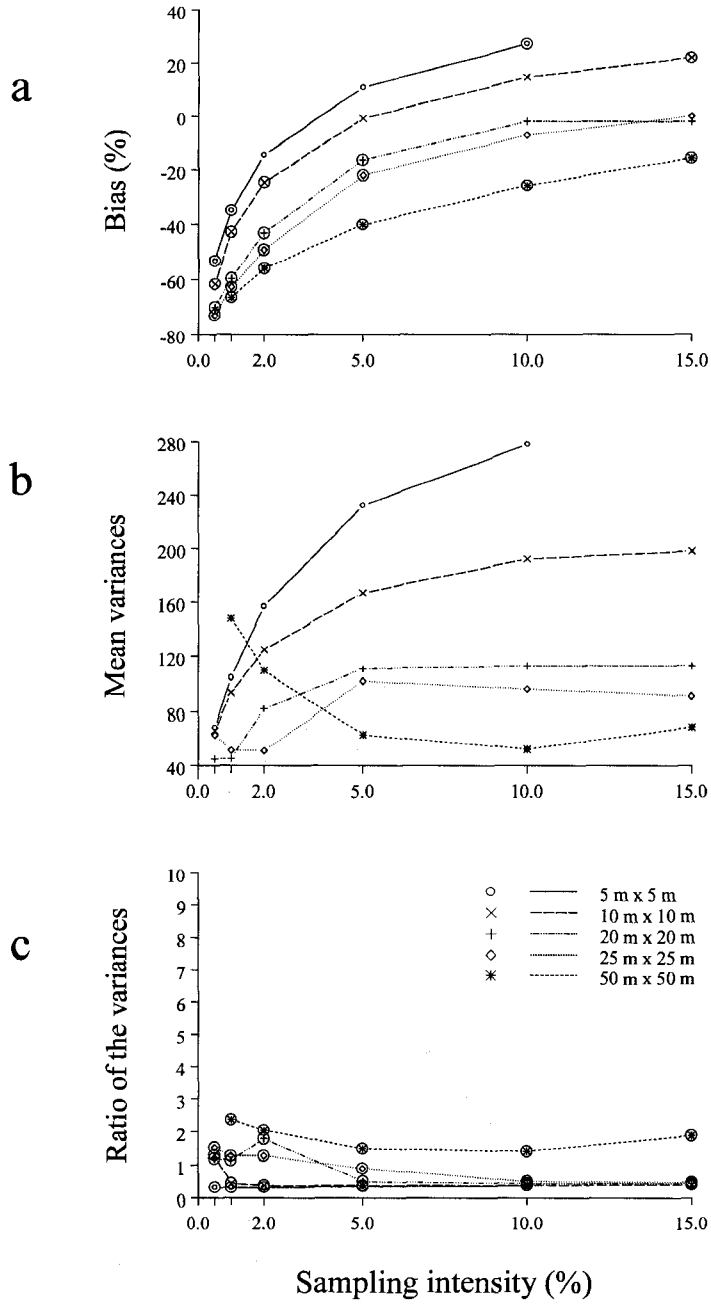


Figure 25. Influence of plot size and sampling intensity on a) percent bias, b) estimated sample variance, and c) ratio of the mean variance to the variance of means for estimating total number of trees species using the nonparametric estimator, *CM3f*. Significant differences are indicated by a circle.

Variance estimates generally increased with increasing plot sizes for all sampling intensities (Table 11, Figure 25b). However, the largest plot size (50 m x 50 m) revealed the opposite trend in that the variance decreased with increasing sampling intensity. Variance estimates were fairly consistent when the sampling intensity exceeded 5% for all plot sizes. An *F*- test for the ratio of variances (Figure 25c), indicated that all estimates were biased for all sample sizes and sampling intensities. However, the variance ratios were fairly stable for all plot sizes and sampling intensities.

***CP1f*:** The relationships between the bias of *CP1f* and sampling intensity (Figure 26a) showed patterns similar to the bias observed for *CM3f*. In general, small sample sizes underestimated the number of tree species irrespective of this plot size. Estimates of the total number of true species were consistent and unbiased when using plot sizes smaller than 50 m x 50 m and with sampling intensity greater than 5%.

According to Figure 26b, there was more variability associated with estimating the number of tree species when using *CP1f*. Compared to *CM3f* (Table 11 and Figure 26b), for small plot sizes of 5 m x 5 m, 10 m x 10 m, and 20 m x 20 m variance estimates were consistent with sampling intensity over 2%. The larger plot sizes displayed more variability than the smaller plot size for all sampling intensity.

The *F*-test for the ratio of variances indicated that all estimated variances associated with *CP1f* were biased (Figure 26c). The small plot sizes of 5 m x 5 m and 10 m x 10 m clearly showed an overestimation of the true variance at the lower sampling intensities, while the larger plot sizes tended to overestimate the variance for all sampling intensities. Comparing the variance ratios obtained from *CP1f* and *CM3f*, the study

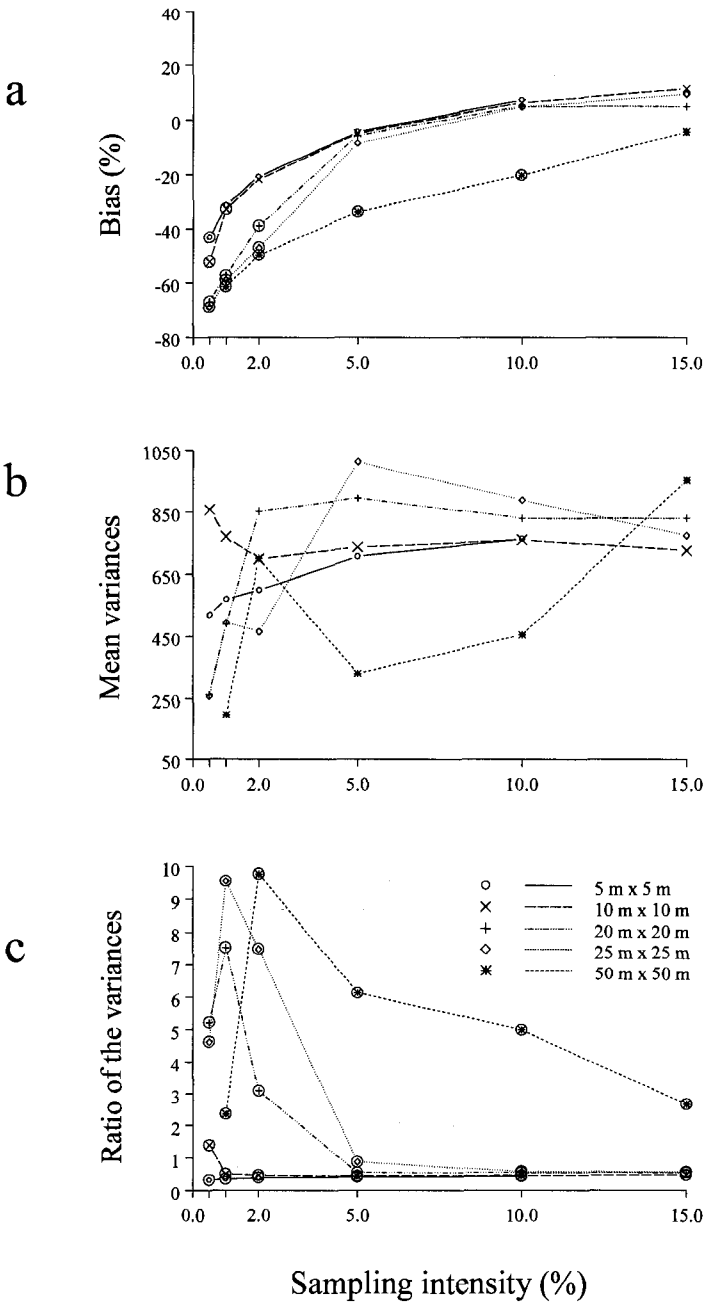


Figure 26. Influence of plot size and sampling intensity on a) percent bias, b) estimated sample variance, and c) ratio of the mean variance to the variance of means for estimating total number of trees species using the nonparametric estimator, *CP1f*. Significant differences are indicated by a circle.

found that the estimated variances for *CM3f* were more consistent than those of *CP1f* even though both were biased.

Optimal Plot Size

Coefficients of Variation and Plot Size

The mean and coefficient of variation based on 20 concentric plots are summarized in Table 12. The relationship between the coefficient of variation and plot size are depicted in Figure 27. The fitted regression models explained at least 98% of the variability in the coefficient of variation as a function of plot size (Table 13). In addition, the *c*-coefficients were all less than 0.5 (Table 13) implying that the spatial pattern of these variables was aggregated (Reich and Arvantis 1992).

Plot Measurement Time and Time Traveling

Plot measurement time of tree DBH, number of trees and counting the number of tree species on each plot size are reported in Table 14. Obviously, identifying tree species consumed more time than measuring tree DBH or counting trees, particularly for the larger plot sizes. The parameter estimates for the logarithmic models to predict plot measurement times for the three variables are shown in Table 15. All three models had R^2 values greater than 0.98 indicating a good fit between the measurement time and plot size.

The average travel time was estimated at 0.733 m/sec based on the four routes randomly established in the HKK seasonal dry evergreen forest. The four sample routes covered slopes ranging between 6 and 14 degrees and represented high, moderate and low amounts of ground cover in this forest type.

Table 12. Estimated sample means and coefficients of variation (CV) obtained from a series of concentric plot sizes (n = 20) randomly located in the 50 ha permanent plot.

Plot Size (m x m)	Plot Area (m x m)	Basal Area/ha		No. of Trees/ha		No. of Tree Species	
		Mean (m^2/ha)	CV (%)	Mean (trees/ha)	CV (%)	Mean (species)	CV (%)
5 x 5	**0.0025**	28.11	168.69	480	88.03	1.05	84.48
10 x 10	**0.0100**	23.46	88.49	440	46.87	3.75	48.07
20 x 20	**0.0400**	26.27	52.78	450	32.19	11.35	28.05
25 x 25	**0.0625**	25.49	37.58	449	27.63	15.10	21.26
50 x 50	**0.2500**	30.13	25.79	460	17.88	35.80	15.79

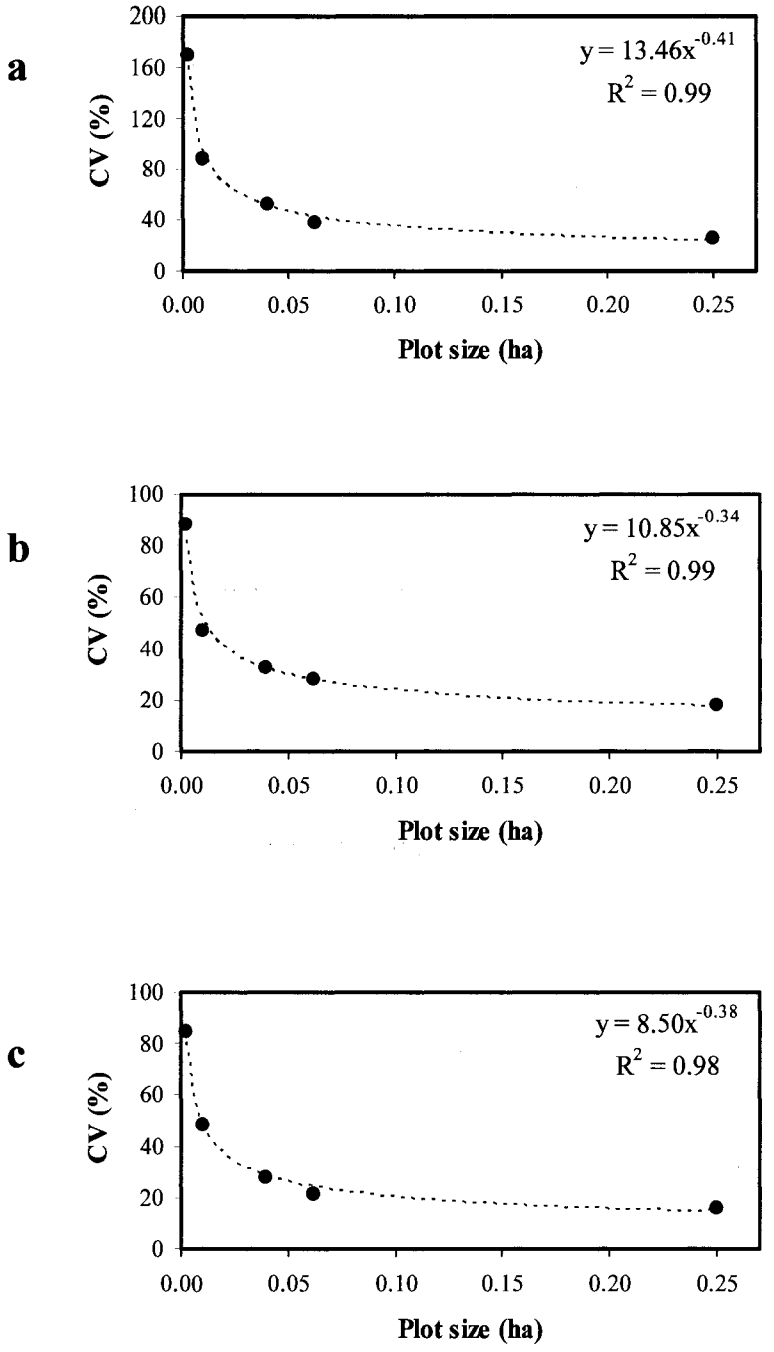


Figure 27. Relationship between the coefficient of variation and plot size associated with estimates of a) basal area/ha, b) trees/ha and c) number of tree species. The dotted lines are the fitted logarithmic regression models.

Table 13. Estimates of the *c*-coefficients and associated R^2 values for the logarithmic models describing the relationship between coefficient of variation and plot size.

Variables	Sample Size	*c*-coefficient	R^2
Basal Area per ha	20	-0.415	0.989
No. of Trees per ha	20	-0.338	0.990
No. of Tree Species	20	-0.376	0.985

Table 14. Plot measurement time of three stand characteristics for different plot sizes established in the HKK seasonal dry evergreen forest.

Plot Size (m x m)	Plot Area (ha)	Sample Size	Counting Trees (sec)	Counting Species (sec)	DBH Measurement (sec)
5 x 5	**0.0025**	4	101	123	111
10 x 10	**0.01**	4	179	264	217
20 x 20	**0.04**	4	480	819	635
25 x 25	**0.0625**	4	594	1124	836
50 x 50	**0.25**	4	4331	6454	5300

Table 15. Estimated regression coefficients and R^2 values for the logarithmic models for estimating plot measuring time as a function of plot size.

Variables	Sample Size	R^2	B_0	B_1	B_2
Basal Area per ha	5	0.996	3.26	1.48	-1.80
No. of Trees per ha	5	0.984	3.09	1.58	-2.14
No. of Tree Species	5	0.995	3.43	1.41	-1.52

Computing Optimal Plot Sizes

Using the relationship developed in the previous section, Eq. 23 was solved to identify the optimal plot size that minimized the total time required to estimate basal area/ha, trees/ha or counting the number of tree species on the sample plots. The following sections describe how the optimal plot size changes for each of theses variables assuming different tract sizes, the desired percent sampling error and coefficient of variation.

Optimal plot size for basal area per hectare

Assume a preliminary sample using a 5 m x 5 m plot size had a coefficient of variation of 100% in a 12500 ha stand. If the rate of travel is 0.733 m/sec and the desired percent sampling is 15% at the 0.95 level of confidence, the equation expressing the total time of a survey to estimate basal area/ha is given by:

$$T_i = 10419.25933 Q_i^{-0.41448} + 3.867199 Q_i^{-0.82960} z_i^{1.48138} (\ln z_i)^{-1.79630}. \quad (24)$$

Solving this equation iteratively for different plot sizes (Q_i), it is found that n = 2 plots measuring 76 m x 76 m would minimize the total cost of the survey for estimating basal area with a 15% sampling error at the 0.95 level of confidence

Table 16 and Figure 28 summarize how the optimal plot and associated sample size changes for different tract size (W), percent sampling errors (E) and initial coefficient of variations (CV_j) for a 5 m x 5 m sample plot.

For a given coefficient of variation for a 5 m x 5 m plot as the desired allowable error decreased the required sample size increased, thus requiring smaller plot sizes in order to minimize the total cost of the survey. Likewise as the variability associated with

Table 16. Optimal plot sizes and associated sample size for estimating basal area that minimize the total cost of the survey.

		Optimal Plot Size (m x m)						
Tract Size		CV[2] (%)						
(ha)	E[1] (%)	25	50	75	100	125	150	175
100	5	24.8	18.7	16.5	15.4	14.7	14.2	13.9
100	10	34.9	24.8	20.9	18.7	17.4	16.5	15.8
100	15	43.3	30.2	24.8	21.9	20.0	18.7	17.8
100	20	50.6	34.9	28.5	24.8	22.5	20.9	19.6
500	5	37.0	26.2	21.8	19.5	18.0	17.0	16.3
500	10	53.8	37.0	30.1	26.2	23.6	21.8	20.5
500	15	67.2	46.0	37.0	31.9	28.6	26.2	24.4
500	20	78.8	53.8	43.2	37.0	33.0	30.1	27.9
2500	5	57.2	39.3	31.8	27.6	24.8	22.9	21.4
2500	10	83.8	57.2	45.8	39.3	34.9	31.8	29.5
2500	15	104.9	71.5	57.2	48.9	43.3	39.3	36.2
2500	20	123.0	83.8	67.0	57.2	50.6	45.8	42.2
12500	5	89.2	60.8	48.7	41.7	37.0	33.7	31.1
12500	10	130.9	89.2	71.2	60.8	53.8	48.7	44.8
12500	15	163.7	111.6	89.2	76.0	67.2	60.8	55.8
12500	20	191.7	130.9	104.6	89.2	78.8	71.2	65.4
		Optimal Sample Size (plots)						
Tract Size		CV (%)						
(ha)	E (%)	25	50	75	100	125	150	175
100	5	7	43	120	239	402	611	864
100	10	2	7	21	43	76	120	174
100	15	2	3	7	15	27	43	64
100	20	2	2	4	7	13	21	31
500	5	4	25	75	161	286	453	662
500	10	2	4	12	25	46	75	114
500	15	2	2	4	8	15	25	38
500	20	2	2	2	4	7	12	17
2500	5	2	13	41	91	168	277	421
2500	10	2	2	6	13	24	41	63
2500	15	2	2	2	4	8	13	20
2500	20	2	2	2	2	4	6	9
12500	5	2	7	20	46	87	146	227
12500	10	2	2	3	7	12	20	31
12500	15	2	2	2	2	4	7	10
12500	20	2	2	2	2	2	3	5

[1]E is an allowable sampling error. [2]CV is a coefficient of variation.

Note: The student's t value of 1.96 ($\alpha = 0.05$) was used to determined an optimal plot size.

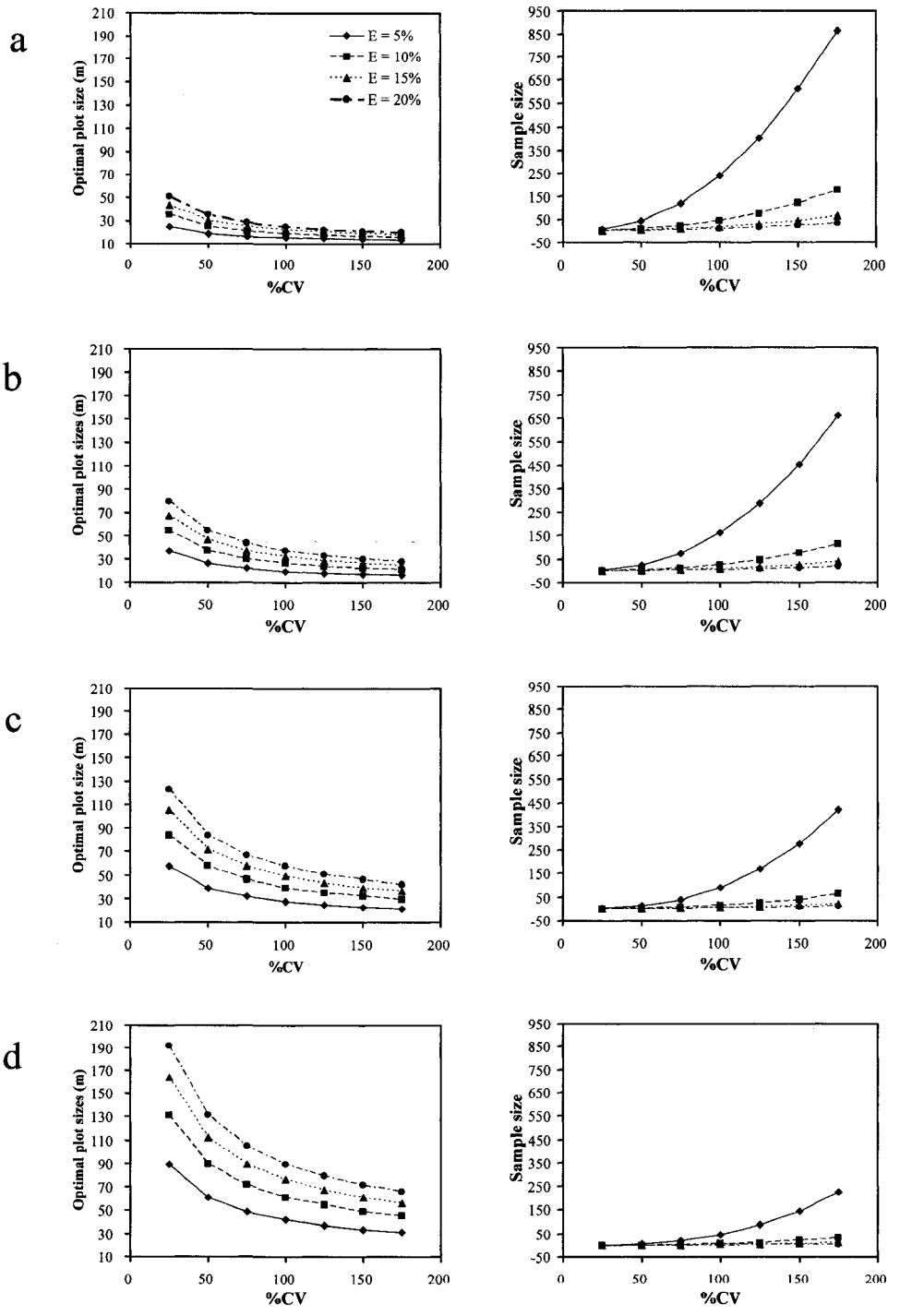


Figure 28. The relationship between optimal plot size and coefficient of variation for different tract sizes (a = 100 ha, b = 500 ha, c = 2500 ha and d = 12500 ha) and percent sampling errors to estimate basal area/ha.

the initial sample increased, the required number of sample required to achieve the desired allowable error increased thus requiring smaller sample plots to offset the increased cost associated with measuring more plots. As the size of the area being surveyed increased, the distance between samples increased thereby requiring larger sample plots to offset the increased travel time.

Optimal plot sizes for tree density

Assume a preliminary sample using a 5 m x 5 m plot size had a coefficient of variation of 100% in a 12500 ha stand. If the rate of travel is 0.733 m/sec and the desired percent sampling is 15% at the 0.95 level of confidence, the equation expressing the total time of a survey to estimate tree density is given by:

$$T_i = 16477.63219 Q_i^{-0.33830} + 9.16018 Q_i^{-0.67660} z_i^{1.57548} (\ln z_i)^{-2.14059}$$

Solving this equation iteratively for different plot sizes (Q_i), it is found that n = 8 plots measuring 50.4 m x 50.4 m would minimize the total cost of the survey for estimating tree density with a 15% sampling error at the 0.95 level of confidence

Table 17 and Figure 29 summarize how the optimal plot and associated sample size changes for different tract size (W), percent sampling errors (E) and initial coefficient of variations (CV_j) for a 5 m x 5 m sample plot.

For a given coefficient of variation for a 5 m x 5 m plot as the desired allowable error decreased the required sample size increased, thus requiring smaller plot sizes in order to minimize the total cost of the survey. Likewise as the variability associated with the initial sample increased, the required number of sample required to achieve the desired allowable error increased thus requiring smaller sample plots to offset the

Table 17. Optimal plot sizes and associated sample size for estimating tree density that minimize the total cost of the survey.

		Optimal Plot Size (m x m)						
Tract Size		**CV^2 (%)**						
(ha)	**E^1 (%)**	**25**	**50**	**75**	**100**	**125**	**150**	**175**
100	**5**	18.0	14.0	12.6	11.9	11.5	11.2	11.0
100	**10**	24.6	18.0	15.4	14.0	13.2	12.6	12.2
100	**15**	30.0	21.5	18.0	16.1	14.8	14.0	13.4
100	**20**	34.7	24.6	20.4	18.0	16.5	15.4	14.6
500	**5**	26.0	18.8	16.0	14.5	13.6	13.0	12.5
500	**10**	36.7	26.0	21.4	18.8	17.2	16.0	15.2
500	**15**	45.0	31.7	26.0	22.6	20.4	18.8	17.7
500	**20**	52.1	36.7	29.9	26.0	23.3	21.4	20.0
2500	**5**	38.8	27.4	22.6	19.8	18.0	16.7	15.8
2500	**10**	55.1	38.8	31.6	27.4	24.6	22.6	21.0
2500	**15**	67.6	47.6	38.8	33.6	30.0	27.4	25.4
2500	**20**	78.1	55.1	44.9	38.8	34.7	31.6	29.3
12500	**5**	58.3	41.0	33.5	29.0	26.0	23.8	22.1
12500	**10**	82.6	58.3	47.5	41.0	36.7	33.5	31.0
12500	**15**	101.2	71.5	58.3	50.4	45.0	41.0	38.0
12500	**20**	116.7	82.6	67.4	58.3	52.1	47.5	43.9
		Optimal Sample Size (plots)						
Tract Size		**CV (%)**						
(ha)	**E (%)**	**25**	**50**	**75**	**100**	**125**	**150**	**175**
100	**5**	18	96	248	475	775	1158	1611
100	**10**	3	18	48	96	162	248	352
100	**15**	2	6	18	36	62	96	138
100	**20**	2	3	9	18	30	48	70
500	**5**	11	64	179	365	623	953	1364
500	**10**	2	11	31	64	114	179	263
500	**15**	2	4	11	23	40	64	95
500	**20**	2	2	5	11	19	31	46
2500	**5**	7	39	113	239	426	676	994
2500	**10**	2	7	18	39	70	113	169
2500	**15**	2	3	7	13	24	39	58
2500	**20**	2	2	3	7	11	18	27
12500	**5**	4	23	67	143	259	419	630
12500	**10**	2	4	11	23	41	67	100
12500	**15**	2	2	4	8	14	23	34
12500	**20**	2	2	2	4	7	11	16

[1]E is an allowable sampling error. [2]CV is a coefficient of variation.

Note: The student's t value of 1.96 ($\alpha = 0.05$) was used to determined an optimal plot size.

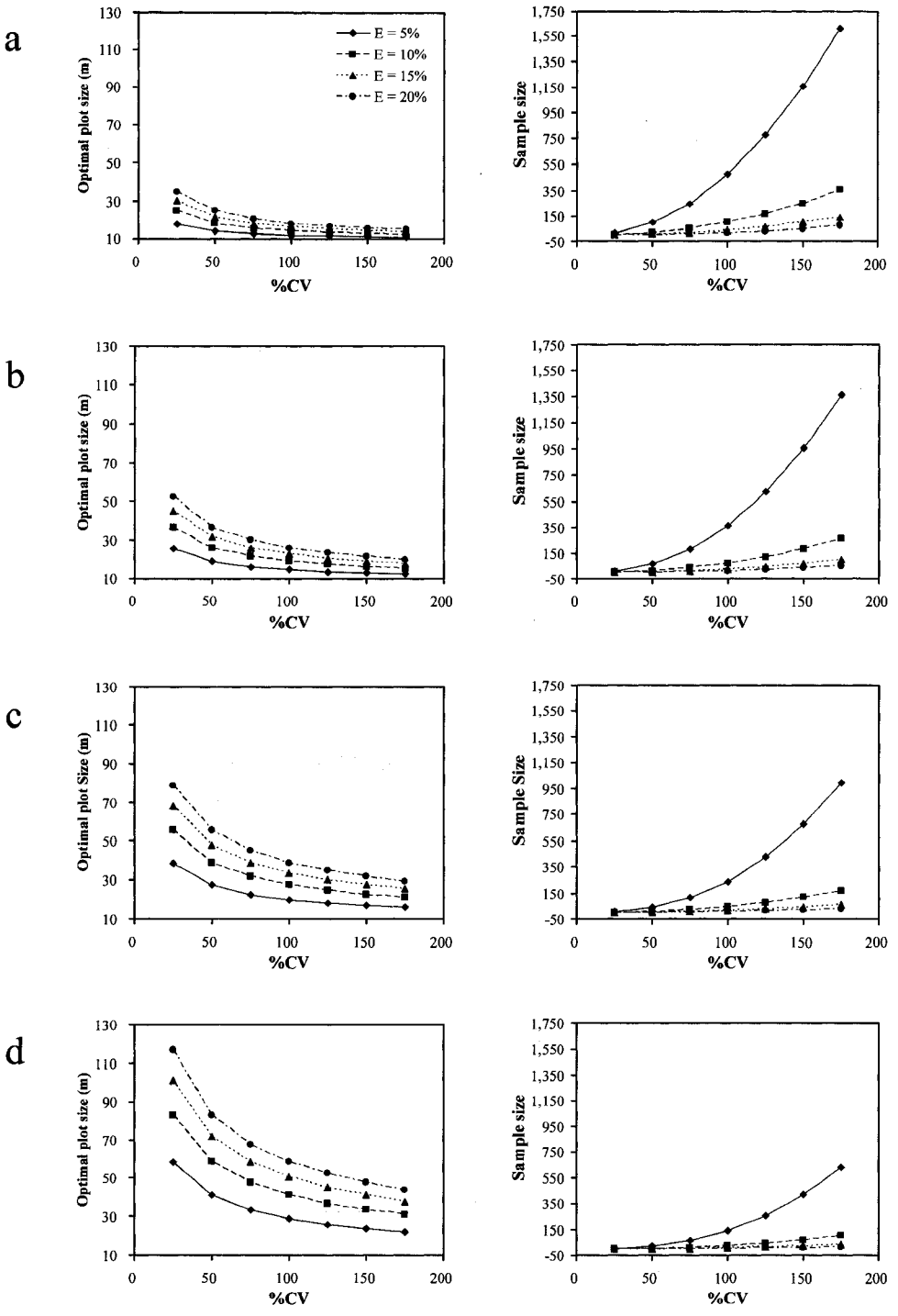


Figure 29. The relationship between optimal plot size and coefficient of variation for different tract sizes (a = 100 ha, b = 500 ha, c = 2500 ha and d = 12500 ha) and percent sampling errors to estimate tree density.

increased cost associated with measuring more plots. As the size of the area being surveyed increased, the distance between samples increased thereby required larger sample plots to offset the increased travel time.

Optimal plot sizes of total tree species

Assume a preliminary sample using a 5 m x 5 m plot size had a coefficient of variation of 100% in a 12500 ha stand. If the rate of travel is 0.733 m/sec and the desired percent sampling is 15% at the 0.95 level of confidence, the equation expressing the total time of a survey to estimate the total number of tree species is given by:

$$T_i = 13153.98794 Q_i^{-0.37590} + 6.477655 Q_i^{-0.7518} z_i^{1.40524} (\ln z_i)^{-1.51607}$$

Solving this equation iteratively for different plot sizes (Q_i), it is found that n = 5 plots measuring 53 m x 53 m would minimize the total cost of the survey for estimating the total number of tree species with a 15% sampling error at the 0.95 level of confidence

Table 18 and Figure 30 summarize how the optimal plot and associated sample size changes for different tract size (W), percent sampling errors (E) and initial coefficient of variations (CV_j) for a 5 m x 5 m sample plot.

For a given coefficient of variation for a 5 m x 5 m plot as the desired allowable error decreased the required sample size increased, thus requiring smaller plot sizes in order to minimize the total cost of the survey. Likewise as the variability associated with the initial sample increased, the required number of sample required to achieve the desired allowable error increased thus requiring smaller sample plots to offset the increased cost associated with measuring more plots. As the size of the area being surveyed increased, the distance between samples increased thereby required larger

Table 18. Optimal plot sizes and associated sample size for estimating the total number of tree species that minimize the total cost of the survey.

		Optimal Plot Size (m x m)						
Tract Size		CV[2] (%)						
(ha)	E[1] (%)	25	50	75	100	125	150	175
100	5	17.3	13.5	12.2	11.5	11.1	10.9	10.7
100	10	24.1	17.3	14.8	13.5	12.7	12.2	11.8
100	15	29.9	20.9	17.3	15.4	14.2	13.5	12.9
100	20	35.0	24.1	19.7	17.3	15.8	14.8	14.0
500	5	25.6	18.2	15.4	13.9	13.0	12.4	12.0
500	10	37.3	25.6	20.8	18.2	16.6	15.4	14.6
500	15	46.8	31.8	25.6	22.1	19.8	18.2	17.0
500	20	55.0	37.3	29.8	25.6	22.8	20.8	19.3
2500	5	39.6	27.1	22.0	19.2	17.3	16.1	15.2
2500	10	58.6	39.6	31.7	27.1	24.1	22.0	20.4
2500	15	73.8	49.8	39.6	33.8	29.9	27.1	25.0
2500	20	86.8	58.6	46.6	39.6	35.0	31.7	29.1
12500	5	62.4	42.2	33.7	28.8	25.6	23.3	21.5
12500	10	92.5	62.4	49.6	42.2	37.3	33.7	31.0
12500	15	116.4	78.6	62.4	53.0	46.8	42.2	38.7
12500	20	136.9	92.5	73.5	62.4	55.0	49.6	45.5
		Optimal Sample Size (plots)						
Tract Size		CV (%)						
(ha)	E (%)	25	50	75	100	125	150	175
100	5	15	87	228	440	721	1,077	1,504
100	10	3	15	43	87	148	228	324
100	15	2	5	15	32	56	87	126
100	20	2	3	7	15	27	43	63
500	5	9	56	160	330	569	878	1,256
500	10	2	9	26	56	100	160	236
500	15	2	3	9	19	34	56	83
500	20	2	2	4	9	16	26	39
2500	5	5	31	94	204	371	598	888
2500	10	2	5	14	31	57	94	143
2500	15	2	2	5	10	19	31	47
2500	20	2	2	2	5	9	14	21
12500	5	3	16	50	111	207	343	524
12500	10	2	3	7	16	30	50	76
12500	15	2	2	3	5	10	16	25
12500	20	2	2	2	3	5	7	11

[1]E is an allowable sampling error. [2]CV is a coefficient of variation.

Note: The student's t value of 1.96 ($\alpha = 0.05$) was used to determined an optimal plot size.

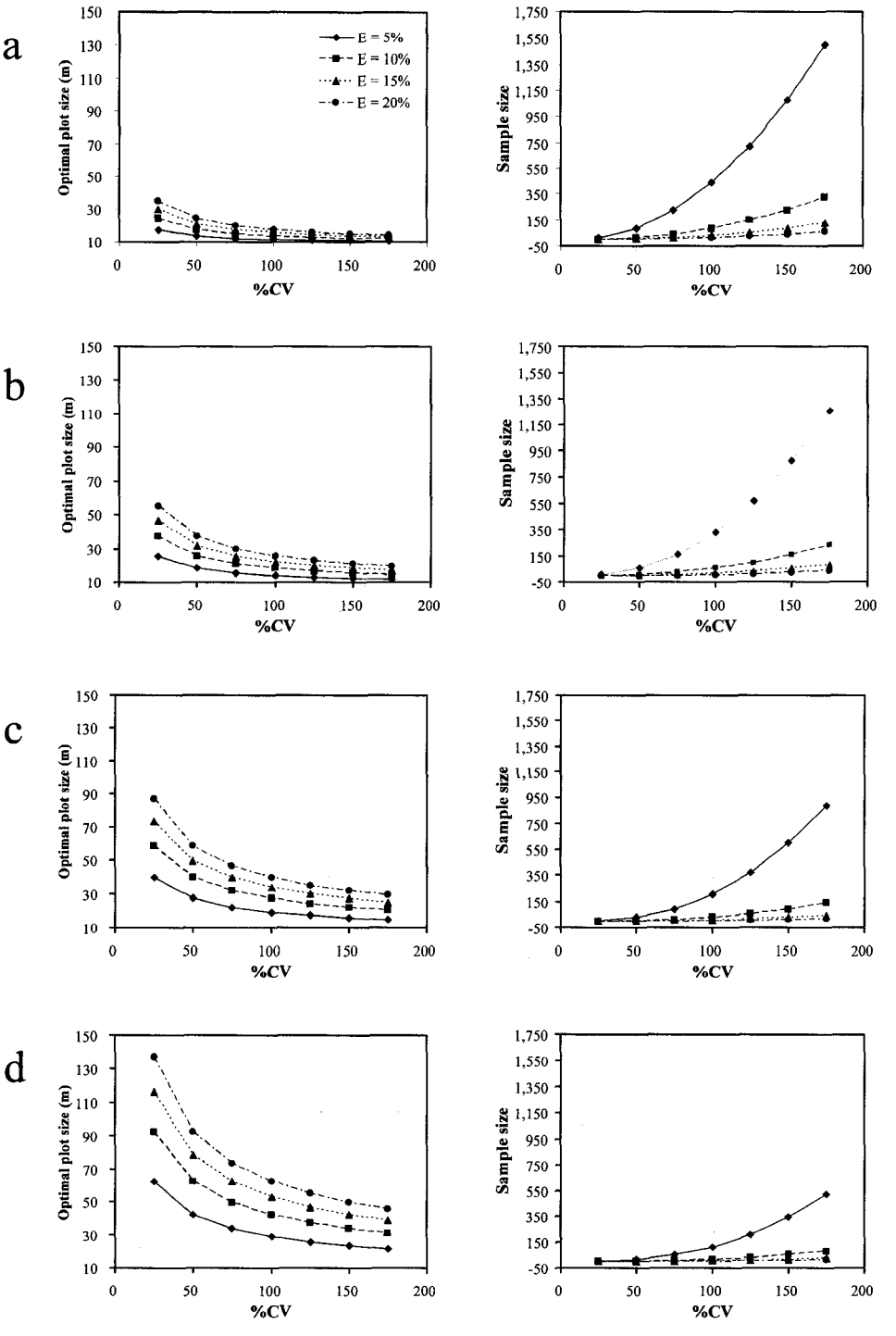


Figure 30. The relationship between optimal plot size and coefficient of variation for different tract sizes (a = 100 ha, b = 500 ha, c = 2500 ha and d = 12500 ha) and percent sampling errors to estimate the total number of tree species.

sample plots to offset the increased travel time.

DISCUSSION

A simulation study was carried out using a 50 ha permanent plot located in the Huai Kha Kheang Wildlife Sanctuary in western Thailand. While it may have been desirable to replicate this study using additional permanent plots, this plot is the only one available in Thailand representing the forest structure and species compositions of seasonal dry evergreen forests in Thailand. This limitation should not detract from the results presented in this study.

The results of this study confirmed that estimated means and sample variances for basal area/ha and trees/ha were unbiased for all plot sizes and sampling intensities evaluated in this study. For estimating the number of tree species, the nonparametric estimators *CM3f* and *CP1f* provided unbiased estimates when using small plot sizes with large sample sizes. For all plot sizes and sampling intensities, both estimators provided biased estimates of the sample variance. However, *CM3f* yielded more consistent variance estimates across all plot sizes and sampling intensities. In general, variance estimates for all three variables were consistent when the sampling intensity exceeds 2%. The variance estimates associated with estimating the number of tree species agreed with the study by Kenkel and Podani (1991) in Central Canada. The authors found that the efficiency in estimating the variance can be improved by using larger plot sizes. For predicting the number of tree species, the results of this study suggest that using large plot sizes tends to underestimate the total number of tree species and provide less efficient estimates of the variance for both *CM3f* and *CP1f*.

In practice, the design of a forest survey is limited by the cost which influences the number of sample plots that can be established of a given size. The spatial pattern of the variable being measured directly affects the variability in the population. Thus, in order to design the most cost efficient survey it is important to select an optimal plot size that will yield the most cost efficient estimate at minimal cost. However, in selecting an optimal plot size, one must take consideration not only the time required to measure the plot but also the travel time. The results of this study indicated an inverse relationship between plot size and sample size. For example, if it were required to use a larger sample size in order to achieve specified allowable error, one would be required to use a larger number of smaller plots to compensate for the increased cost associated with measuring more plots. Similar trends were observed between plot size and sample size as the variability in the population changed as well as the size of the population being surveyed.

Based on the results of this study, plot size and sampling intensity did not influence the reliability of the statistical estimates for basal area and tree density. As a result, the suggested optimal plot sizes developed in this study can be used with confidence for estimating tree basal area and tree densities. To estimate the total number of tree species in a population, a small plot size with a large sample size improved the efficiency of the estimates. This is opposite of the results reported by Archaux et al. (2007). The authors suggested that for estimating the number of tree species larger plot sizes are more reliable than smaller one. This is particular true only when counting the number of species on given sample plot. However, the nonparametric estimators are efficient when using a small plot size which accounts for more variation in the abundance of tree species found in the population.

CONCLUSION

Information from the permanent 50 ha plot in seasonal dry evergreen forest in Thailand provides a unique opportunity for an in-depth study of how plot size and sampling intensity influence ones ability to efficiently estimate forest stand parameters. The results of a Monte Carlo simulation study confirm that estimates of basal area and tree density were unbiased for any plot size and sampling intensity. In addition, selected nonparametric estimators of the number of tree species provided good estimate for small plot sizes and large sample sizes. Variance estimates were generally biased for all plot sizes and sampling intensities.

To conduct forest inventories in the future, an optimal plot size for estimating tree basal area, tree density, and species abundance can be determined from the equations developed in this study. The equations take into consideration the variability associated with the characteristics of interest, while minimizing the total cost of survey which normally depends on the sample size, plot measurement time and travel time. To utilize these equations, preliminary data associated with using a 5 m x 5 m plot are required to prime the equation. The suggested optimal plot sizes are provided in this study as a look-up table.

There are many subtypes of this forest type in Thailand which vary by regions, climate, parent material and landform. While, the results of this study area only applicable to seasonal dry evergreen forests in the west central part of Thailand, however, the results of this study could potentially be used as a guideline in designing forest surveys in other parts of the country.

REFERENCES

Archaux, F., L. Berges, and R. Chevalier 2007. Are plant censuses carried out on small quadrats more reliable than on larger ones? Plant Ecol. 188, 179-190

Baker, P. J. 1997. Seedling establishment and growth across forest types in an evergree/deciduous forest mosaic in western Thailand. Natural History Bulletin of the Siam Society 45(1), 17-41.

Baker, P. J. 2001. Age structure and stand dynamics of a seasonal tropical forest in western Thailand. Dissertation. University of Washington. 388 pp.

Bunge, J. and M. Fitzpatrick. 1993. Estimating the number of species: a review. J. Amer. Stat. Assoc. 88, 364-373.

Bunge, J. and M. Fitzpatrick. 1995. Comparison of three estimators of the number of species. J. Appl. Stat. 22, 45-59.

Bunyavejchewin, S. 1986. Ecological studies of tropical semi-evergreen rain forest at Sakaerat, Nakhon Ratchasima, Northeast Thailand. 1. Vegetation patterns. Natural History Bulletin of the Siam Society 34(1), 35-57.

Bunyavejchewin, S. 1999. Structure and dynamics in seasonal dry evergreen forest in northeastern Thailand. Journal of Vegetation Science 10, 787-792.

Bunyavejchewin, S., J. V. LaFranke, P. Pattapong, M. Kanzaki, A. Itoh, T. Yamakura, and P.S. Ashton. 1998. Topograpic analysis of a large-scale research plot in seasonal dry evergreen forest at Huai Kha Khaeng Wildlife Sanctury, Thailand. Tropics 8, 1-16.

Bunyavejchewin, S., P. J. Baker, J. V. LaFrankie, and P. S. Ashton. 2001. Stand Structure of a seasonal dry evergreen forest at the Huai Khaeng Wildlife Sanctuatry, western Thailand. Natural History Bulletin of the Siam Society 49, 89-106.

Chao, A. and S-M. Lee. 1992. Estimation the number of classes via sample coverage. J. Amer. Stat. Assoc. 87, 210-217.

Cochran, W.G. 1977. Sampling Techniques. John Wiley and Sons, New York, 428 pp.

Condit, R. 1998. Tropical Forest Census Plots: Methods and Results from Barro Colorado Island, Panama and a Comparison with Other Plots. Springer-Verlag and R. Landes Company. 211pp.

Efron, B. and R. J. Tibshirana. 1994. An Introduction to the Bootstrap. Chapman & Hall, San Francisco, USA. 436 pp.

Fishman, G.S. 1995. Monte Carlo: Concepts, Algorithms, and Applications, Springer Verlag, New York, USA. 728 pp.

Gambill, C. W., H. V. Wiant, Jr. and D. O. Yandle. 1985. Optimal plot size and AF. Forest Sci. 31, 587-594.

Kenkel, N. C. and J. Podani. 1991. Plot size and estimation efficiency in plant community studies. J. Veg. Sci. 2, 539-544.

Manly, B. F. J. 1998. Randomization, Bootsrap and Monte Carlo Methods in Biology. Chapman & Hall. London. 399 pp.

Manokaran, N., J. V. LaFrankie, K. M. Kochummen, E. S. Quah, J. E. Klahn, P. S. Ashton, and S. P. Hubbell. 1990. Methodology for the fifty hectare research plot at Pasoh Forest Reserve. Research Pamphlet No. 104. Forest Research Institute Malaysia. 69 pp.

Reich, R. M. and Arvantis, L. G. 1992. Sampling unit, spatial distribution of trees, and precision. Northern Journal of Applied Forestry 9, 3-6.

Schereuder, H.T., M.S. Williams, and R. M. Reich. 1999. Estimation the number of tree species in a forest community using survey data. Environmental Monitoring and Assessment 56, 293-303.

Shao, J. and D. Tu. 1995. The Jackknife and Bootrap. Springer-Verlag, New York, USA. 540 pp.

Smith, E. P. and G. V. Belle. 1984. Nonparametric Estimation of Species Richness. Biometric 40, 119-129.