Modeling the Spatial Distribution of Soil Texture in the State of Jalisco, Mexico

Nantachai Pongpattananurak
Dep. of Conservation
Faculty of Forestry
Kasetsart Univ.
Bangkok, Thailand 10900

Robin M. Reich
Dep. of Forest and Rangeland Stewardship
Colorado State Univ.
Fort Collins, CO 80523-1472

R. Khosla*
Dep. of Soil and Crop Sciences
Colorado State Univ.
Fort Collins, CO 80523–1170
and
Precision Agriculture Research
King Saud Univ.
Riyadh, Saudi Arabia

C. Aguirre-Bravo
Small Businesses Sustainability Coordination Acquisition Management - Washington Office
1621 N. Kent Street
Suite 707
Arlington, VA 22209-2132

Information on the spatial variability of soil attributes, such as soil texture, is crucial for a wide range of decisions in ecosystem and agricultural research and management. Surveys designed to collect such information across large geographic regions may not capture the fine to moderate scales of variation in soil properties. As a result, soil properties may exhibit spatial dependencies at scales smaller than the scale at which sampling was performed. This lack of spatial structure in the sample data makes it difficult, if not impossible, to use optimal predictors such as ordinary kriging for modeling spatial variability in a data set. We developed a new approach for modeling soil texture fractions across large geographic regions at a fine spatial resolution. The method of three-stage least squares is used to model the large-scale variability in soil texture, while the small-scale variability is modeled using multivariate regression trees. This approach was used to model the spatial distribution of soil textural classes in the Mexican state of Jalisco. Independent variables used in the modeling process included terrain data, climatic data, and satellite imagery. Our results indicate that the sand model accounted for 62% of the variability observed in the sample plots, while the clay model accounted for 56% of the observed variability. Maps of soil attributes obtained from this study can serve as a useful surrogate (i) explaining the spatial variability in soil attributes across large geographic regions and (ii) supporting applications of precision forestry and agriculture for site-specific management across both small and large geographic regions.

Abbreviations: 3SLS, three-stage least squares; GIS, geographic information system; MRT, multivariate regression tree; MSEP, mean squared error of prediction; SMSE, standardized mean squared error.

Forests in the state of Jalisco, Mexico, contain diverse and unique communities of endemic species of plants, animals, reptiles, and amphibians. In particular, the tropical dry forests, located along the Pacific coast, have among the highest species richness of tropical dry forests in the world. Furthermore, there are more endemic tree species in these dry forests than elsewhere in the neotropics (Challenger, 1998). Tropical dry forests are considered endangered worldwide (Cué-Bär et al., 2006) and the ones in the state of Jalisco were given high priority to establish a network of 1442 permanent sample plots to collect baseline data on the biotic and abiotic resources in the state. The availability of baseline information is central to the assessment and decision processes associated with ecosystem sustainability in the state.

The soils in the state exhibit complex spatial patterns because of discontinuities in soil processes or other localized effects (Ettema et al., 1998). These patterns are largely a function of historical factors, such as the exploitation of the forest

Soil Sci. Soc. Am. J. 76:199–209
Posted online 11 Nov. 2011
doi:10.2136/sssaj2011.0180
Received 17 May 2011.
*Corresponding author (raj.khosla@colostate.edu).
© Soil Science Society of America, 5585 Guilford Rd., Madison WI 53711 USA
All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

resources through logging, agricultural activities, fire, and climatic patterns, factors identified as important in influencing the present ecological system of the region (Bailey, 1996; Katsalirou et al., 2010). Resource availability has long been recognized as important in determining patterns in the diversity of organisms (Grime, 1979; Tilman, 1982). The main factors determining spatial heterogeneity at the local level are resource availability and response to environmental variables that have a direct impact on plant growth or on available resources. In times of uncertainty, information on the current health of the natural resources is important both ecologically and economically. The need for basic information about soil resource heterogeneity is central to support a variety of management decisions. As part of this effort, soil texture was identified as being critical to the management of ecosystem sustainability in the state. Soil physical properties such as soil texture have a direct effect on the water holding capacity, cation exchange capacity, nutrient availability, N loss, as well as other soil processes and conditions (Mzuku et al., 2005). Therefore, an adequate understanding of the spatial variability in soil texture is essential.

One approach to modeling such complex relationships is to decompose the set of sample data representing the spatial processes under study into two components: a mean structure representing the large-scale variation and a stochastic-dependent structure representing the small-scale variation (Cressie, 1991). If the data are not on a lattice, trend surface or regression analysis can be used to describe the large-scale spatial variability, while the small-scale spatial variability represented by the residuals from the regression model can be modeled using the covariance structure or a semivariogram (Cressie, 1993).

The accurate assessment of soil properties requires a sampling strategy that captures the characteristics of many different soil variables at the appropriate spatial scale (Fortin et al., 1989). Often, however, resource surveys are designed based on the resources available, and the sampling strategies used may or may not capture fine to moderate scales of variation in the soil properties but rather the trend in soil properties at larger scales that are related to long-term patterns of climate, parent material, landform, and vegetation. Soil properties may exhibit spatial dependencies at scales smaller than the scale at which sampling was performed (Cambardella et al., 1994). In such situations, the use of optimal predictors such as kriging for describing the small-scale spatial variability in a set of data will be somewhat disappointing. The estimation procedure will be more like a simple averaging of the neighboring data values (Isaaks and Srivastava, 1989). While this is a simple solution to the kriging system, it is not a desirable solution to the problem, and it would be inappropriate to use geostatistics in these situations.

In many instances, the residuals from the trend surface or regression model still contain a wealth of information that could be used to gain flexibility and precision in estimation. Failure to model this structure may lead to an inferior model, inaccurate predictions, and inappropriate conclusions (Carroll and Pearson, 2000). While geostatistical approaches may not provide the desired solution, there is always an interest in exploring new approaches.

One approach to enhance the estimation process is to use stratified estimation with terrain and satellite imagery as the basis of stratification. With this approach, sample units are classified into homogeneous classes with respect to the predictions of error attributes, and the classes are then used as strata in the analysis. Bloch and Segal (1989) first proposed the use of binary classification trees (Breiman et al., 1984) in forming strata to adjust for covariates. Michaelsen et al. (1994) used regression tree analysis to produce a site stratification of a tall grass prairie to facilitate in the design and allocation of ground sampling efforts. The stratification was derived using a digital elevation model and land use–land cover data. More recently, Benedetti et al. (2005) and Cocchi et al. (2002) explored a tree-based strategy to partition municipalities into strata based on geographical region, population size, and census track data. Reich et al. (2008a) used a tree-based stratified design to model the spatial variation in temperature and precipitation in the state of Jalisco, Mexico. With this approach, sample units (i.e., pixels of a satellite image) are classified with respect to predictions of residuals from a regression model into homogeneous classes, and the classes were then used as strata in describing the small-scale variability in the climate data.

Another challenge associated with modeling soil texture is the implied constraint that estimates of the sand, silt, and clay percentages sum to 100%. Most models of soil texture found in the literature concentrated on only one or two components (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) used compositional ordinary kriging to simultaneously model soil texture while constraining the three fractions of soil texture to sum to 100%. Unfortunately, if the sample design does not capture the spatial structure in soil properties in the region of interest (because of limitations associated with the sampling design and not necessarily the soil property itself), such an approach would not work. As an alternative, the method of three-stage least squares (3SLS) (Zellner and Theil, 1962) could be used to fit a system of linear equations to describe the large-scale variability in soil texture while ensuring that the three fractions of soil texture sum to 100%. To ensure that estimates from the system of equations are consistent with the observed values, it would be necessary to use a multivariate tree-based (De'ath, 2002) stratification scheme to describe the small-scale variability in soil texture.

To implement the modeling approach described above requires a set of spatial data for use as independent variables in the various models. Spatial data layers are available to characterize the vegetation, climate, and topography in the state of Jalisco. Because of potential classification errors in the vegetation maps, the vegetation in the state has been characterized using Landsat-7 ETM+ satellite imagery. These spatial data layers have the potential to be useful for developing relationships with soil properties, such as soil texture, and may give insight into the influence of

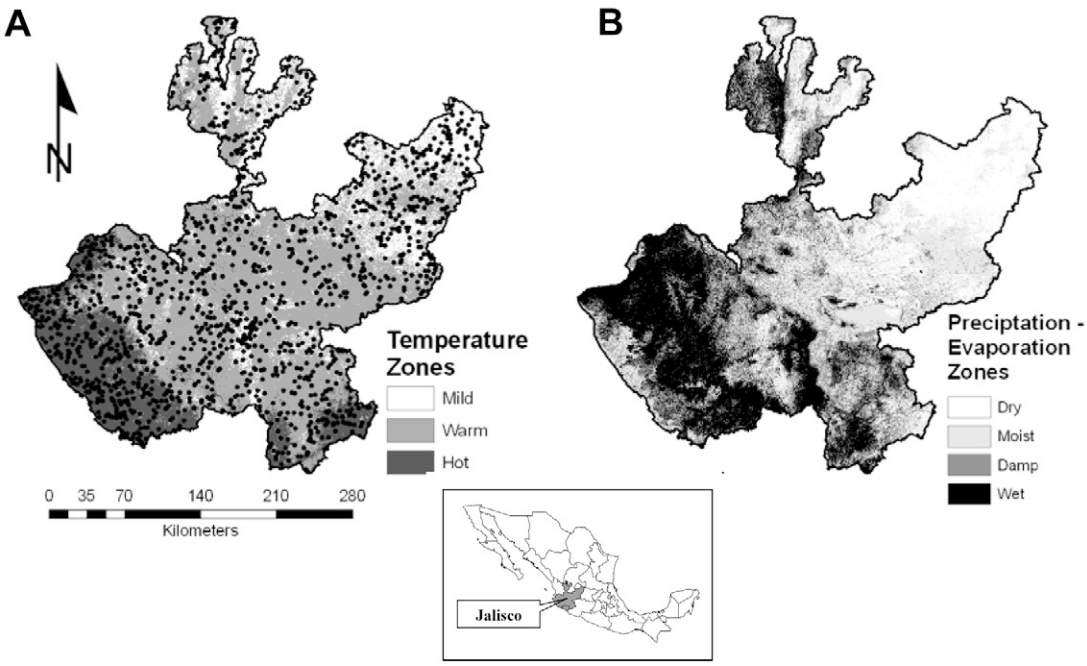

Fig. 1. Statewide distribution of (A) temperature and (B) precipitation zones for Jalisco, Mexico. The circles overlaid on the temperature zones represent the locations of 1427 soil samples.

landscape attributes on soil properties in the state (Odeh et al., 1995; Scull, 2010).

The objectives of this study were: (i) to use soils data from the statewide inventory of Jalisco to develop models to predict soil texture based on landscape features; (ii) to evaluate the predictive performance of the models using cross-validation procedures; and (iii) to produce fine-scale maps of soil texture using spatial data of the independent variables.

MATERIALS AND METHODS

Site Description

The state of Jalisco is located in western Mexico between 22°45′ and 18°55′ N and 101°28′ and 105°42′ W and contains an area of approximately 8 million ha (Fig. 1). Climatic variation in the region is influenced by an interaction between westerly winds of maritime air masses and the effects of mountain ranges.

The state can be divided into three broad ecological regions. The first is the subhumid tropical zone located along the Pacific coast and is characterized by high temperatures, monsoon rains during summer months (730–1200 mm), and an annual dry period that ranges from 5 to 9 mo. Tropical dry forests dominate the region and occur on terrain with elevations from sea level to 2000 m and up to 4000 m near the Colima volcano. In the northern part of this zone, the forests are mesic, while in the south the forests are slightly drier. Soils are shallow and derive from metamorphic and volcanic rocks.

At higher elevations and to the east is the subhumid temperate zone that covers the greatest portion of the state. Pine, oak, and mixed deciduous hardwood forests dominate this zone (1000–2600 m). The average annual rainfall ranges from 900 to 1500 mm (Perry, 1991). Soils derive from volcanic rock and have a high content of organic matter. This zone gradually changes to third ecological region in the eastern part of the state, an arid and semiarid zone that has a low annual precipitation of 400 mm or less and 8 to 12 dry mo. The dominant vegetation includes mesquite–acacia and xerophytic scrub. Soils in this region are shallow, derived from igneous rocks, and have a low content of organic matter.

Field Data

Soil samples were collected from the surface (0–10-cm depth) from 1427 out of 1442 30- by 30-m sample plots established in 2006 as part of an inventory and monitoring program to characterize the natural resources in the state (Fig. 1) (Reich et al., 2008b). Sample locations were georeferenced using a global positioning system (GPSmap 76CSx, Garmin, Olathe, KS). Sample plots were located using a stratified design that took into consideration the climatic variability (i.e., temperature, precipitation, and evaporation) and the spectral reflectance (Landsat-7 ETM+ imagery) of the landscape to ensure that samples were acquired from a wide range of ecological conditions. Soil samples were analyzed for basic soil physical and chemical properties, including soil texture and pH. Soil textural components (sand, silt, and clay) were determined using the Bouyoucos hydrometer method (Gee and Bauder, 1986; Jones, 2002) and soil pH was determined by the electrode method (Thomas, 1996).

Satellite Imagery and Digital Elevation Model

Spectral information for the state was obtained from 10 cloud-free Landsat-7 ETM+ images acquired during the months of January through March 2004. The winter imagery allowed better differentiation between major vegetation types (i.e., tropical dry forests, pine–oak forests, grasslands, etc.) than imagery acquired from other seasons that year. Individual scenes were

normalized (Hall et al., 1991) to account for differences in acquisition dates. The thermal (57-m resolution) and panchromatic (15-m resolution) bands were resampled to a 30-m spatial resolution using the nearest neighbor resampling technique (Muukkonen and Heiskanen, 2005). Nearest neighbor resampling was used because it better maintained the original reflectance values while providing sufficient accuracy and reducing the potential introduction of unwanted geometric distortions in areas with no ground control points to provide precise control (Muukkonen and Heiskanen, 2005). Finally, the 10 satellite images were joined together to create a seamless mosaic image of the state. Ground reflectance values of the Landsat-7 ETM+ bands were extracted for the 1427 sample locations.

A digital elevation model (DEM) of the state was obtained from the USGS National Elevation Dataset as a seamless raster surface at 90-m resolution (Gesch et al., 2002). The DEM was resampled to a 30-m resolution using a bilinear interpolation technique (Edenius et al., 2003) to produce a more continuous surface reflecting gradual changes in elevation at a 30-m spatial resolution and to correspond with the spatial resolution of the field data and other spatial layers. Estimates of elevation, slope, and aspect from the DEM were extracted for each sample location.

Climate Zones

Geographic information system (GIS) raster layers representing the long-term average monthly temperature (°C), precipitation (mm), and evaporation (mm) were available at a 30-m resolution from previous work (Reich et al., 2008a). The raster layers were developed using Mexico's climatological network (SICLIM [Velázquez-Alvarez and Balancan-S, 2000]), which included 256 weather stations from across the state. The models used to create the raster layers used a combination of satellite imagery and topographic data to describe the spatial variability in the climate data and accounted for 45 to 85% of the variability in the monthly mean temperature, precipitation, and evaporation. The raster layers were used to define three temperature zones and four precipitation–evaporation zones, which when combined defined 12 unique climatic zones (Fig. 1) (Reich et al., 2008a). The temperature and precipitation–evaporation zones were based on a histogram equalization approach that produced a uniform distribution of temperatures and precipitation–evaporation across the state (Acharya and Ray, 2005). These zones coincide in general with those used to describe vegetation patterns in the state (Reich et al., 2010). The raster layer of climate zones was used to place each sample location into one of 12 unique climate zones.

Modeling Approach

Statistical Model

A statistical framework was used to model soil texture in this study. Let $z(s_i)$ represent a sample value of the target variable z at spatial location s_i. Also, assume that the sample data contain a set of explanatory variables 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 of p known explanatory variables $x_j(s_i)$:

$$z(s_i)=\beta_0+\sum_{j=1}^{p}x_j(s_i)\beta_j+\eta(s_i) \qquad [1]$$

where β_j ($j = 0, ..., p$) 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 $C(x,y) = \mathrm{cov}[\eta(x_i), \eta(y_j)]$. The error term in Eq. [1] is unknown because the true model is unknown. Once the model parameters have been estimated, the residuals of the regression model 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

$$\eta(s_i)=\hat{\eta}(s_i)+\mu(s_i) \qquad [2]$$

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

$$\hat{\eta}(s_i)=f\left[x(s_i)\right]+\delta(s_i) \qquad [3]$$

with $E[\delta(s_i)|x(s_i)] = 0$; $f[x(s_i)]$ is a deterministic function used to define the stratification and $\delta(s_i)$ is a zero-mean random error term (Cocchi et al., 2002). Combining Eq. [2] and [3] results in

$$\eta(s_i)=f\left[x(s_i)\right]+\varepsilon(s_i) \qquad [4]$$

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; Benedetti et al., 2005). The mean function $f[x(s_i)$ is estimated by $\hat{f}$ using the recursive partitioning method introduced by Breiman et al. (1989). Combining Eq. [1] and [4], the full model describing the spatial variability in the sample data is given by

$$z(s_i)=\beta_0+\sum_{j=1}^{p}x(s_i)\beta_j+f\left[x(s_i)\right]+\varepsilon(s_i) \qquad [5]$$

where $f[x(s_i)$ is a process explaining the small-scale variability in the data set.

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

$$\mathrm{var}\left[z(s_i)\right]=\mathrm{var}\left[\hat{\eta}(s_i)\right]+\mathrm{var}\left[\hat{\delta}(s_i)\right] \qquad [6]$$

where $\mathrm{var}[\hat{\eta}(s_i)]$ reflects the uncertainty in estimating the parameters of the regression model and $\mathrm{var}[\hat{\delta}(s_i)]$ reflects the uncertainty in estimating the error ($\hat{\eta}$) in the regression model, which was calculated using the sample variance within each strata (Cochran, 1977). The use of the sample variance as a measure of the uncertainty in estimating the error in the regression model is justified by the fact that the best estimator, given that the sam-

ple data belong to the stratum, is the sample variance. The sample variance is thus the mean square error of this best predictor.

The variance associated with an estimate at a new location, s_0, can be written as

$$\mathrm{var}\left[z(s_0)\right]=\mathrm{var}\left[\hat{\eta}(s_0)\right] + \mathrm{var}\left[z(s_0)\right]+\mathrm{var}\left[\hat{\delta}(s_0)\right] \quad [7]$$

where the additional term, var[$z(s_0)$], reflects the random variation at a new location, s_0.

Modeling the Large-Scale Spatial Variability

Multiple regression analysis was used to develop a model to describe the large-scale variability in the sand, silt, and clay fractions as a function of topography (elevation, slope, and aspect), the climate zone, and the spectral variability of the ground cover in the state. The climate zone was treated as a categorical variable, while the topographic data and the bands of the satellite imagery were treated as continuous variables in the regression models. For each component of soil texture, a stepwise model selection by Akaike's Information Criteria (Venables and Ripley, 2002) was used to identify the best subset of independent variables to include in the regression models. This process identified the following functional form of a system of equations to describe the large-scale spatial variability in the sand, silt, and clay fractions of soil texture:

$$\begin{aligned} \mathrm{Sand} &= f(\mathrm{topography+climate+satellite}) \\ \mathrm{Clay} &= f(\mathrm{sand+topography+climate+satellite}) \\ \mathrm{Silt} &= 100-\mathrm{sand}-\mathrm{clay} \end{aligned} \quad [8]$$

Sand exhibited the strongest correlation with the set of available independent variables (i.e., topography, climatic zone, and satellite imagery). Sand was also highly correlated with clay and was included in the clay model as a predictor. Silt was modeled as a function of the estimated values for sand and clay to ensure that the estimates of sand, silt, and clay summed to 100%. Regression coefficients for the system of equations (Zellner and Theil, 1962) were estimated using 3SLS, which combines two-stage least squares with seemingly unrelated regression. Two-stage least squares is a method of using dependent variables as independent variables on the right-hand side of a regression model, while seemingly unrelated regression is a technique for fitting a system of multiple regression equations with cross-equation parameter restrictions and correlated error terms.

Modeling the Small-Scale Spatial Variability

The small-scale spatial variability in sand, silt, and clay were modeled using a regression-tree-based stratified design (Cocchi et al., 2002; Benedetti et al., 2005; Reich et al., 2008a). With this approach, sample units (i.e., pixels of the GIS layers) are classified into homogeneous classes with respect to predictions of the residuals from the 3SLS model, and the classes are then used as strata to describe the small-scale variability. Independent variables considered in the binary regression trees included elevation, slope, aspect, and the satellite imagery.

To ensure that the estimates of sand, silt, and clay summed to 100%, it was important to use the same stratification scheme in estimating the mean residual errors for the sand, silt, and clay models. To accomplish this, a binary multivariate regression tree (MRT) (De'ath, 2002, 2011) was used to simultaneously partition the residuals from the 3SLS model into strata. The recursive algorithm splits the three sets of residuals into strata by comparing all possible splits among the independent (continuous) variables using a binary partitioning algorithm that maximizes the dissimilarities among the strata. Once the algorithm partitions the residuals into new strata, new relationships are developed, assessed, and split into new strata. The algorithm recursively splits the set of residuals in each stratum until either the stratum is homogeneous or it contains too few observations to be split further.

The optimal (regression) tree structure (i.e., the size and number of strata) was determined by evaluating the predictive performance of the fitted model while varying two parameters that controlled the recursive portioning algorithm of the MRT (De'ath, 2011). The size of the strata was identified by varying the parameter minsize. If minsize = 5 (the default value), the algorithm continues to partition the data into strata if there are at least five observations in a given stratum. The parameter minsize was set to 5, 10, and 25, and then increased in increments of 10 if an optimal tree structure was not identified. The number of strata was selected using the parameter *best*, which is an integer specifying the size (i.e., the number of strata) of a specific partitioning of the data. If there is no partitioning of the requested size, the next largest partitioning is returned. The number of strata varied from a minimum of 10 to the maximum number of strata possible in increments of five strata.

Cross-Validation

A 10-fold cross-validation (Efron and Tibshrani, 1993) was used to evaluate the predictive performance of the fitted soil textural models (3SLS + MRT). During the cross-validation, 95% prediction intervals were calculated for the prediction data sets, assuming normality:

$$\hat{z}(s_0)\pm 1.96\sqrt{\mathrm{var}\left[\hat{z}(s_0)\right]} \quad [9]$$

where $\hat{z}$ is an estimate of sand, silt, or clay at the new location s_0, and var[$\hat{z}(s_0)$] is the prediction variance (Eq. [7]). Confidence intervals for individual observations were also computed (Eq. [6]) for the fitted models. Coverage rates were calculated as the proportion of intervals that covered the observed data.

The standardized mean squared error (SMSE) was used to test the null hypothesis that the variance estimates were unbiased (Hevesi et al., 1992) and was calculated using

$$\mathrm{SMSE}=\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varepsilon}(s_i)^2}{\mathrm{var}\left[\hat{z}(s_i)\right]} \quad [10]$$

where $\hat{\varepsilon}(s_i) = [z(s_i) - \hat{z}(s_i)]$ is the true error and var[$\hat{z}(s_i)$] is the estimated variance. The estimated variance was assumed consis-

tent with the true errors if the SMSE fell within the interval $[1 \pm 2(2/n)^{1/2}]$ (Hevesi et al., 1992).

The effectiveness of the fitted soil textural model (3SLS + MRT) was evaluated using a goodness-of-prediction statistic (G) (Agterberg, 1984; Kravchenko and Bullock, 1999; Guisan and Zimmermann, 2000; Schloeder et al., 2001). The value of G is a measure of the effectiveness of a prediction relative to that which could have been derived using the sample mean (Agterberg, 1984). A G value of 1 indicates perfect prediction, a positive value indicates that the model estimates are more reliable than if the sample mean had been used, a negative value indicates that the model estimate is less reliable than if the sample mean had been used, and a value of zero indicates that the model produces estimates equivalent to the sample mean. The prediction bias (Williams, 1997) was calculated for each validation data set as a percentage of the true value, while the prediction uncertainty was measured by the mean squared error of prediction (MSEP).

Numerous methods are available for selecting an optimal size for regression trees (Esposito et al., 1997). Most of these approaches try to produce a less complex tree that is easy to interpret. When modeling the small-scale variability in a set of data, however, it is not clear what the optimum tree size should be, especially if there is interest in making inferences about the final model. In addition to providing point estimates, there might also be interest in constructing confidence intervals for the mean response as well as constructing predictions intervals for estimates at new locations for which no data are available. It is rather straightforward to calculate both confidence and prediction intervals for a regression model (Neter et al., 1985), but how is the variability associated with the regression tree component of the model taken into consideration? The variance associated with the terminal nodes of a regression tree is influenced by the size of the tree, or number of strata, and generally decreases with increasing tree size. The number of observations assigned to a stratum also influences the estimate of the variance. In some situations, having a large number of observations within strata may provide better estimates of the variance than having a small number of observations.

A decision rule was adopted to identify a regression tree structure that minimized the error in estimating the variance of the mean response and the prediction variance, defined as

$$C=\sqrt{(\mathrm{SMSE_M}-1)^2+(\mathrm{SMSE_P}-1)^2}+\frac{\mathrm{MSEP}}{\mathrm{df}-L} \qquad [11]$$

where C is the cost complexity, $\mathrm{SMSE_M}$ is the standardized mean squared error of the variance of the mean response and $\mathrm{SMSE_P}$ is the standardized mean squared error of the prediction variance, MSEP is obtained from the 10-fold cross-validation, df is the degrees of freedom of the 3SLS model, and L is the number of terminal nodes in the regression tree. The last term in the equation is a penalty for using a regression tree with an excessive number of terminal nodes or strata.

Raster Layers of Soil Texture

Raster layers representing the large- and small-scale variability in soil texture were created by building expressions using the estimated regression coefficients and logical tree structure to process the set of raster layers representing the explanatory variables in the various models. The raster layers representing the large- and small-scale variability where combined to create the final surfaces of sand, silt, and clay. Raster layers of the prediction errors were also developed. The estimated surfaces of sand, silt, and clay were used to classify each pixel in the state into one of 12 soil textural classes (Soil Survey Division Staff, 1993).

RESULTS AND DISCUSSION

Influence of Regression Tree Structure on Soil Textural Model Performance

Figure 2 displays the influence of the stratum size and the number of strata on the performance of the soil textural model for clay content. The poorest performance was for the null stratification scenario of having only one stratum. The 3SLS regression model significantly underestimated the variances of the mean response and significantly overestimated the prediction variance, as measured by the SMSE (Fig. 2A). The biases in estimating the variances resulted in coverage rates other than the nominal rate of 0.95 (Fig. 2D). Because the models do not take into consideration the small-scale variability in the data, the overall fit of the models, as measured by the G statistic, were at a minimum (Fig. 2B).

Increasing the number of strata has several effects. By partitioning the errors into finer and finer partitions, or strata, the algorithm attempts to minimize the variability within each stratum while maximizing the overall fit of the models. Depending on the partitioning used to create the strata, the variability within a stratum may vary considerably. This variability reflects the uncertainty in estimating the response variable within a given stratum. The net result is that both the model and prediction SMSE values approach unity as the number of strata increase (Fig. 2A). Accounting for the small-scale variability ensures unbiased estimates in the prediction variance and the variance associated with the mean response. As the SMSE approaches unity, the coverage rates approach the nominal rate of 0.95 (Fig. 2D). The MSEP also increased as the number of strata increased. When the prediction SMSE equals unity, the MSEP is a measure of the uncertainty in the model (Fig. 2C).

Increasing the minimum stratum size decreases the bias in estimating the variance of the mean response. The decrease in the model SMSE is due to an increase in the variance component associated with the small-scale variability and not to an increase in the overall fit of the models. Increasing the minimum stratum size at which the last split is performed results in a coarser partitioning of the data. This coarser partitioning results in strata that have more variability associated with them, thus increasing the uncertainty associated with estimates.

An advantage of the tree-based stratified design for describing the small-scale variability in a set of data is that it does not require distributional assumptions about the variable of interest or

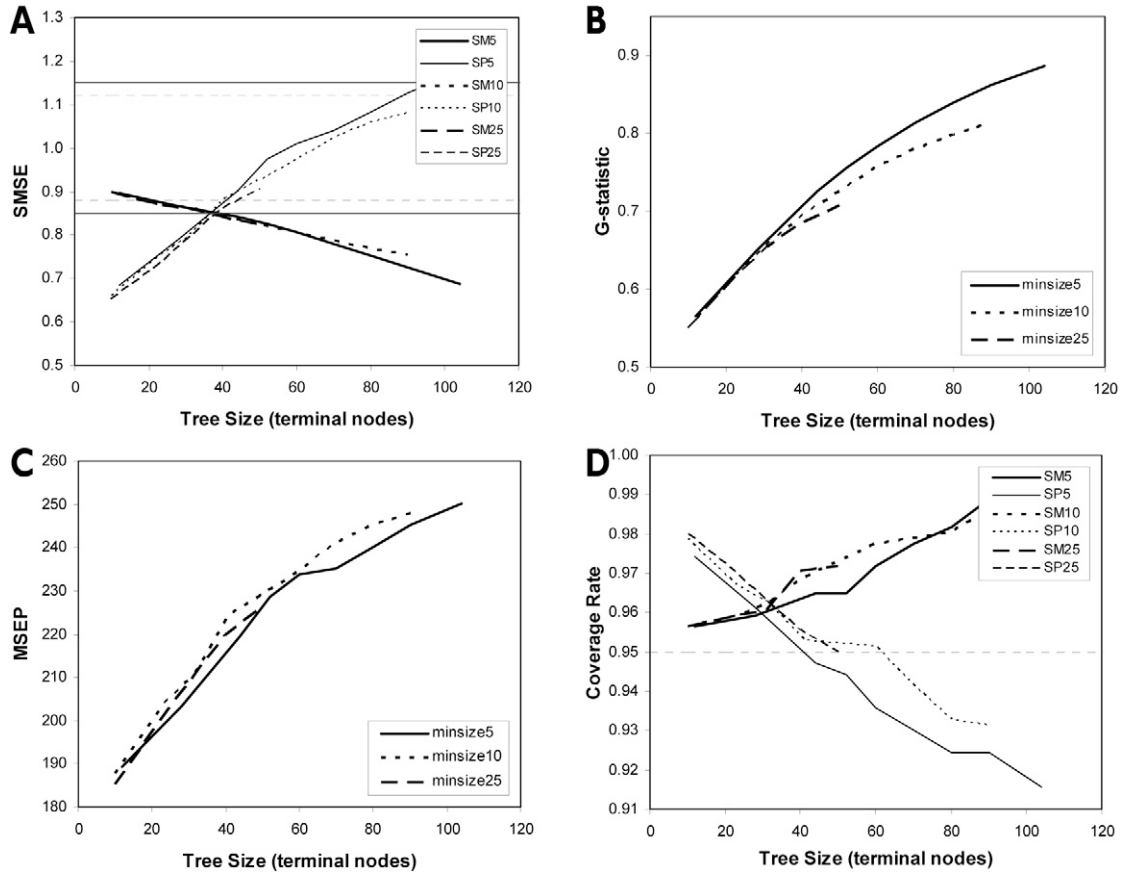

Fig. 2. Influence of the minimum number of observations at a given terminal node (minsize = 5, 10, and 25) and regression tree size on (A) the standardized mean squared error (SMSE) for the mean response (SM) and predictions at a new location (SP), (B) the *G* statistic, representing the total variability of the mean response accounted for by the model, (C) the mean squared error of prediction (MSEP), and (D) 0.95 coverage rates for the mean response and predictions at new locations for the clay model (three-stage least squares + multivariate regression tree). In (A), the two sets of straight lines represent the region in which the variance estimates are unbiased. The inner bound represents the 95% confidence interval while the outer bound represents the 99% confidence interval.

any hypothesis regarding the functional form of the relationship between the variable of interest and its predictors (O'Connor and Wagner, 2004). Moreover, when many auxiliary variables are available, the regression tree-based algorithm is able to select the most powerful variables for the construction of strata (Cocchi et al., 2002; Benedetti et al., 2005). A disadvantage of this approach is that it requires large sample sizes to ensure that the tree-based algorithm is capable of capturing the complex relationship that exists between the variable of interest and its predictors. Although the regression-tree-based stratified design provided reliable estimates of model variances in this study, theoretical considerations imply that this is not always the case. The use of regression trees to model small-scale variability assumes that a relationship exists between the residuals from the regression model and the set of predictors available for stratification. If there is no relationship, this approach may fail.

To identify the optimum tree structure (i.e., number and size of the strata), simulations were performed in conjunction with a 10-fold cross-validation to evaluate the predictive performance of the models. The study tried to cover many aspects and potential problems in identifying the optimal tree structure by the use of simulations; however, it is difficult to say anything definitive about the behavior of this approach from a few simulation studies. It does appear, however, that the use of regression trees to describe the small-scale variability in a data set provides reliable results. This is supported by the fact that variance estimates for the final models were unbiased and coverage rates were quite close to the 95% nominal rate.

Soil Texture Model

The decision rule (Eq. [11]) identified 41 strata for the sand model and 12 for the clay model, with a minimum stratum size of 25 and five observations, respectively (Table 1). Because the MRT algorithm simultaneously partitioned the residuals from the 3SLS model into homogeneous strata, it was not possible to obtain the best regression tree structure for each soil texture component. Unbiased estimates of the variances for the mean response and new predictions of sand and clay contents were not achieved using the same regression tree structure. Based on the fit statistics from the cross-validation, the regression tree with 41 terminal nodes and a minimum stratum size of 25 observations provided the best overall fit.

The sand model accounted for 62% of the variability observed in the sample plots, while the clay model accounted for only 56% of the variability observed in the sample plots. The confidence and prediction coverage rates for all models were close to

Table 1. Fit statistics for the three-stage least squares + multivariate regression tree model for sand and clay under two stratification schemes for the state of Jalisco, Mexico, including the minimum number of observations in the terminal nodes of the regression tree (minsize), the number of strata in the tree (tree size), R^2 of the three-stage least squares model, the total variability of the mean response accounted for by the model (*G* statistic), the standardized mean squared error of the mean response ($SMSE_M$) and prediction ($SMSE_P$), the mean squared error of the prediction (MSEP), the coverage rate of the mean response (CR-M) and predictions (CR-P) at the α = 0.05 level of significance, and the cost complexity (*C*).

Soil attribute	minsize	Tree size	R^2	*G* statistic	$SMSE_M$	$SMSE_P$	MSEP	CR-M	CR-P	*C*
Sand	5	12	0.43	0.52	0.89	0.67*	191.48	0.96	0.98*	0.65
Clay	5	12	0.39	0.46	0.97	1.04	114.78	0.96	0.95	0.23
Sand	25	41	0.43	0.62	0.85	0.84	218.80	0.97*	0.96	0.58
Clay	25	41	0.39	0.56	0.96	1.40*	130.85	0.96	0.91*	0.62

* Significantly different from 1 for SMSEM and SMSEP and significantly different from 0.95 for CR-M and CR-P at the a = 0.05 level.

the 95% nominal coverage rate. Except for the clay model, the standardized mean squared errors were close to their expected value of unity, indicating that the variance estimates were consistent with the true errors. The standardized mean squared errors ranged from 0.84 to 1.40. Histograms and plots of the predicted vs. observed values (not shown) did not show any trends suggesting any systematic bias in the models. Figure 3A and 3B compare the observed and predicted values of sand, silt, and clay plotted on a soil textural triangle. The graphs indicate a good agreement between the observed and estimated soil textures.

The results reported here are comparable with those obtained by Reich et al. (2004) to model the spatial distribution of forest fuel loadings in the Black Hills National Forest in southwestern South Dakota. Their models described 55 to 71% of the spatial variability in the components of forest fuels (e.g., depth of duff, depth of litter, 1-h fuels, 10-h fuels, and 100-h fuels) ob-

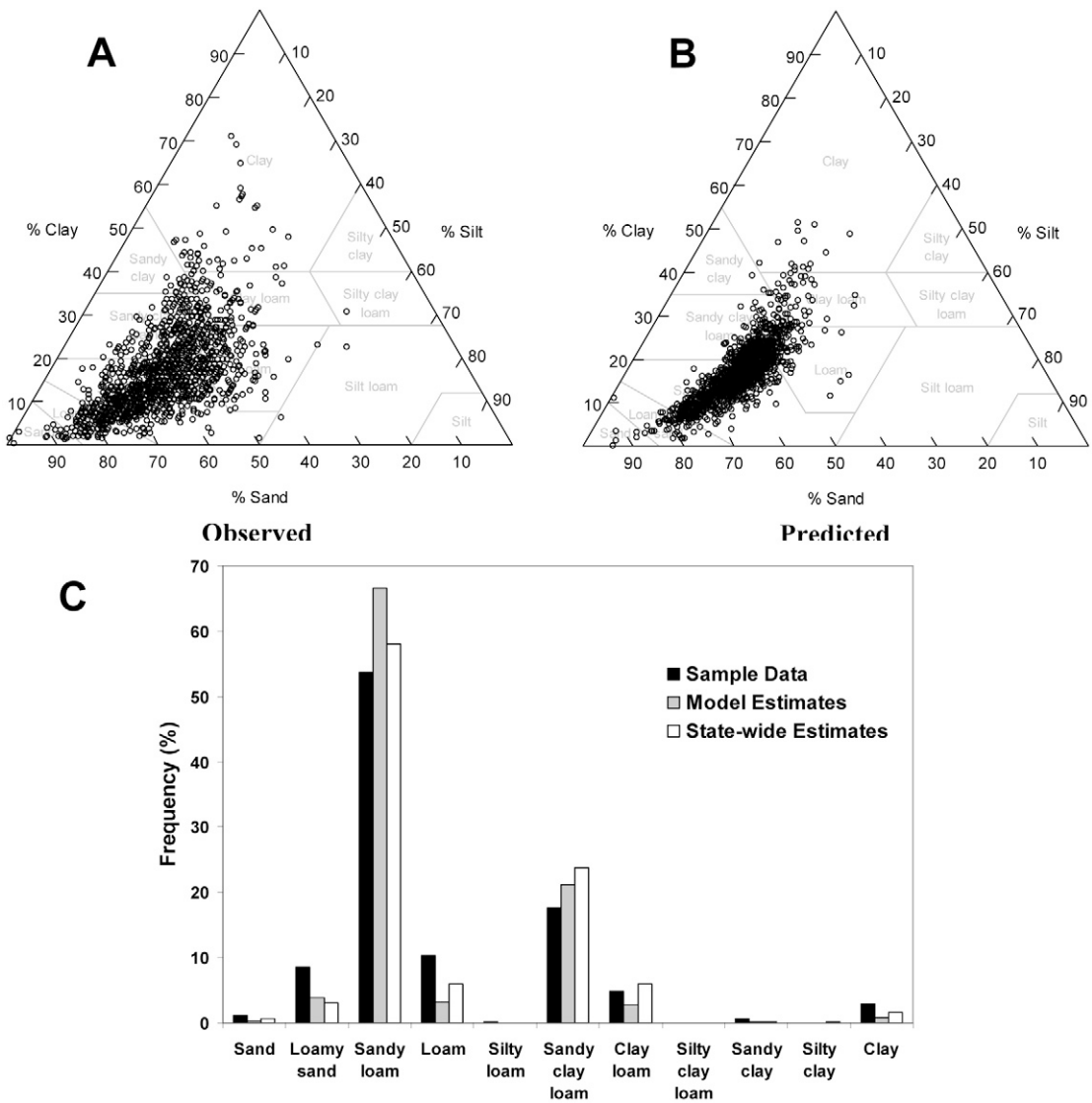

Fig. 3. Comparison of (A) observed and (B) predicted soil textural classes for the sample data superimposed on the soil textural triangle, and (C) frequency distribution of soil textural classes for the observed sample data and predicted values at the sample and state levels.

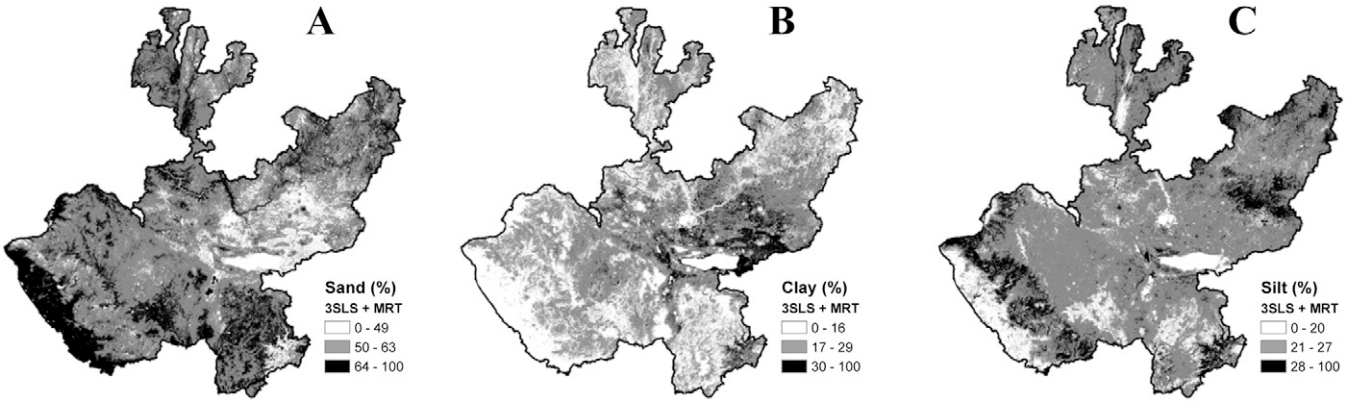

Fig. 4. Spatial distribution of predicted (A) sand, (B) clay and (C) silt in the state of Jalisco, Mexico, using the three-stage least squares + multivariate regression tree (3SLS + MRT) model.

served on the field plots. Estimates of the SMSE showed that the computed prediction variances were statistically consistent with the true errors for all models except one. The 0.95 prediction coverage rates ranged from a low of 0.90 to a high of 0.99.

Spatial Distribution of Soil Texture

Figure 4 displays the predictive surfaces of sand, silt and clay contents, while Fig. 5 displays the standard deviation surfaces for the sand and clay models. Also developed was a map representing the 12 soil textural classes (not shown). The two soil textural classes of sandy loam and sandy clay loam dominate the soils in the state. The sandy clay loam soils occur in the coastal region dominated by tropical dry forests and in the semiarid region in the eastern part of the state. A mosaic of sandy clay loam and clay loam soils characterizes the central portion of the state dominated by grasslands and agricultural lands. Figure 3C compares the relative frequency distribution of the observed soil textural classes of the sample data with the predicted soil textural classes of the sample data and the state as a whole. No significant differences were observed between the distribution of observed and predicted soil textural classes on the sample plots based on a chi-square goodness-of-fit test (P value = 0.14). The distribution of predicted soil textural classes for the state (n = 94,792,466 pixels or sample units) was not significantly different from the distribution of soil textural classes observed on the sample plots (P value = 0.53).

CONCLUSIONS

Finding a suitable statistical approach for describing the spatial distribution in a set of data is not an easy task. This study demonstrated the use of 3SLS + MRT in developing reliable models for the prediction of soil texture across a large geographical region with a fine-scale spatial resolution while constraining the estimates of sand, silt, and clay to sum to 100%. The technique of combining linear regression models with a binary regression-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. The

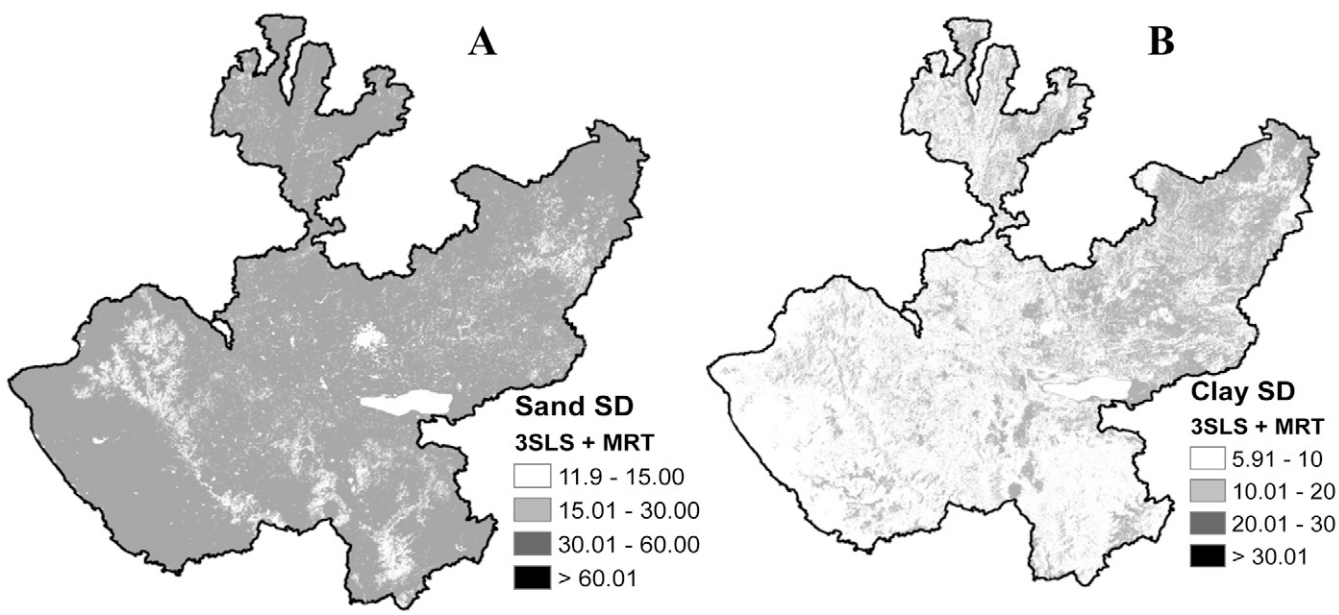

Fig. 5. Spatial distribution of the prediction standard deviation for (A) sand and (B) clay in the state of Jalisco, Mexico, using the three-stage least squares + multivariate regression tree (3SLS + MRT) model.

error surfaces developed in this study can be used to construct confidence intervals and prediction intervals for individual observations. The error 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.

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 regions. Maps of soil attributes provided comprehensive information on the spatial variability of soil properties for the entire state at a fine spatial resolution (i.e., 30 by30 m). The GIS layers of soil attributes developed in this study can support the applications of precision forestry and agriculture, such as managing soil fertility and crop production inputs for site-specific management, across both small and large geographic regions.

ACKNOWLEDGMENTS

We acknowledge the financial and in-kind support by the Mexican state of Jalisco's Secretary of Rural Development and its Sustainability Forestry Department, Jalisco's Agriculture Advisory Board, the U.S. Forest Service, Colorado State University, and the Universidad de Guadalajara, Mexico. Each of these institutions, through their field and laboratory technicians and scientists, provided significant support for this study.

REFERENCES

Acharya, T., and A.K. Ray. 2005. Image processing: Principles and applications. Wiley-Interscience, New York.

Agterberg, F.P. 1984. Trend surface analysis. p. 147–171. *In* G.L. Gaile and C.J. Willmott (ed.) Spatial statistics and models. D. Reidel, Dordrecht, the Netherlands.

Bailey, R.G. 1996. Ecosystem geography. Springer-Verlag, New York.

Benedetti, R., G. Espa, and G. Lafratta. 2005. A tree-based approach to forming strata in multipurpose business surveys. Disc. Pap. 5. Dip. di Economia, Univ. Degli Studi di Trento, Trento, Italy.

Bloch, D.A., and M.R. Segal. 1989. Empirical comparison of approaches to forming strata: Using classification trees to adjust for covariates. J. Am. Stat. Assoc. 84:897–905. doi:10.2307/2290064

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. Classification and regression trees. Wadsworth & Brooks, Monterey, CA.

Cambardella, C.A., T.B. Moorman, J.M. Novak, T.B. Parkin, K.L. Karlen, R.F. Turco, and A.E. Konopka. 1994. Field-scale variability of soil properties in central Iowa soils. Soil Sci. Soc. Am. J. 58:1501–1511. doi:10.2136/sssaj1994.03615995005800050033x

Carroll, S.S., and D.L. Pearson. 2000. Detecting and modeling spatial and temporal dependence in conservation biology. Conserv. Biol. 14:1893–1897.

Challenger, A. 1998. Utlizacion y conservacion de los ecosistemas terrestres de Mexico: Pasado, presente y futuro. Conabio, IBUNAM y Agrupacion Sierra Madre, Mexico City.

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. *In* Proc. Int. Conf. on Improving Surveys, Copenhagen, Denmark. 25–28 Aug. 2002. Available at citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.9270&rep=rep1&type=pdf (verified 21 Oct. 2011). Pennsylvania State Univ., University Park.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. John Wiley & Sons, New York.

Cressie, N. 1991. Geostatistical analysis of spatial data. p. 87–108. *In* Spatial statistics and digital image analysis. Natl. Acad. Press, Washington, DC.

Cressie, N. 1993. Statistics for spatial data. John Wiley & Sons, New York.

Cué-Bär, E.M., J.L. Villaseñor, J.J. Monrroe, and G.I. Manríquez. 2006. Identifying priority areas for conservation in Mexican tropical deciduous forest based on tree species. Interciencia 31:712–719.

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

De'ath, G. 2011. Package 'mvpart.' Available at cran.r-project.org/web/packages/mvpart/mvpart.pdf (verified 21 Oct. 2011). Inst. for Statistics and Mathematics, Vienna Univ. of Economics and Business.

Edenius, L., C.P. Vencatasawmy, P. Sandstrom, and U. Dahlberg. 2003. Combining satellite imagery and ancillary data to map snowbed vegetation important to reindeer *Rangifer tarandus*. Arct. Antarct. Alp. Res. 35:150–157. doi:10.1657/1523-0430(2003)035[0150:CSIAAD]2.0.CO;2

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

Esposito, F., D. Malerba, and G. Semerao. 1997. A comparative analysis of methods for pruning decision trees. IEEE Trans. Pattern Anal. Mach. Intell. 19:476–491. doi:10.1109/34.589207

Ettema, C.H., D.C. Coleman, G. Velidis, R. Lowrance, and S.L. Rathbun. 1998. Spatiotemporal distributions of bacterivorous nematodes and soil resources in a restored riparian wetland. Ecology 79:2721–2734. doi:10.1890/0012-9658(1998)079[2721:SDOBNA]2.0.CO;2

Fortin, M.-J., P. Drapeau, and P. Legendre. 1989. Spatial autocorrelation and sampling design in plant ecology. Vegetatio 83:209–222. doi:10.1007/BF00031693

Gee, G.W., and J.W. Bauder. 1986. Particle size analysis. p. 404–408. *In* A. Klute (ed.) Methods of soil analysis. Part 1. Physical and mineralogical methods. 2nd ed. Agron. Monogr. 9. ASA and SSSA, Madison, WI.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler. 2002. The National Elevation Dataset. Photogramm. Eng. Remote Sens. 68:5–11.

Grime, J.P. 1979. Plant strategies and vegetation process. John Wiley & Sons, Chichester, UK.

Guisan, A., and N.E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecol. Modell. 135:147–186. doi:10.1016/S0304-3800(00)00354-9

Hall, F.G., D.E. Strebel, J.E. Nickeson, and S.J. Gotz. 1991. Radiometic reflection: Toward a common radiometric response among multidata multisensor images. Remote Sens. Environ. 35:11–27. doi:10.1016/0034-4257(91)90062-B

Hevesi, J.A., J.D. Istok, and A.L. Flint. 1992. Precipitation estimation in mountainous terrain using multivariate geostatistics: I. Structural analysis. J. Appl. Meteorol. 31:661–676. doi:10.1175/1520-0450(1992)031<0661:PEIMTU>2.0.CO;2

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

Jones, J.B. 2002. Agronomic handbook: Management of crops, soils and their fertility. CRC Press, Boca Raton, FL.

Katsalirou, E., S. Deng, D.L. Nofziger, A. Gerakis, and S.D. Fuhlendorf. 2010. Spatial structure of microbial biomass and activity in prairie soil ecosystems. Eur. J. Soil Biol. 46:181–189. doi:10.1016/j.ejsobi.2010.04.005

Kravchenko, A., and D.G. Bullock. 1999. A comparative study of interpolation methods for mapping soil properties. Agron. J. 91:393–400. doi:10.2134/agronj1999.00021962009100030007x

Michaelsen, J., D.S. Schimel, M.A. Friedl, F.W. Davis, and R.C. Dubyah. 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. J. Veg. Sci. 5:673–686. doi:10.2307/3235882

Muukkonen, P., and J. Heiskanen. 2005. Estimating biomass for boreal forests using ASTER satellite data combined with standwise inventory data. Remote Sens. Environ. 99:434–447. doi:10.1016/j.rse.2005.09.011

Mzuku, M., R. Khosla, R. Reich, D. Inman, F. Smith, and L. MacDonald. 2005. Spatial variability of measured soil properties across site-specific management zones. Soil Sci. Soc. Am. J. 69:1572–1579. doi:10.2136/sssaj2005.0062

Neter, J., W. Wasserman, and M.H. Kutner. 1985. Applied linear statistical models. Richard D. Irwin, Homewood, IL.

O'Connor, R.J., and T.L. Wagner. 2004. A test of a regression-tree model of species distribution. Auk 121:604–609. doi:10.1642/0004-8038(2004)121[0604:ATOARM]2.0.CO;2

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 67:215–226.

Perry, J.P., Jr. 1991. The pines of Mexico and Central America. Timber Press, Portland, OR.

Reich, R.M., C. Aguirre-Bravo, and V.A. Bravo. 2008a. New approach for

modeling climatic data with applications in modeling tree species distributions in the states of Jalisco and Colima, Mexico. J. Arid Environ. 72:1343–1357. doi:10.1016/j.jaridenv.2008.02.004

Reich, R.M., C. Aguirre-Bravo, and M. Chazaro-Basañeza. 2010. Patterns of tree species richness in Jalisco, Mexico: Relation to topography, climate and forest structure. Plant Ecol. 210:67–84. doi:10.1007/s11258-010-9738-5

Reich, R.M., C. Aguirre-Bravo, and M.A. Mendoza-Briseño. 2008b. An innovative approach to inventory and monitoring of natural resources in the Mexican state of Jalisco. Environ. Monit. Assess. 146:383–396. doi:10.1007/s10661-007-0086-4

Reich, R.M., J.E. Lundquist, and V.A. Bravo. 2004. Spatial models for estimating fuel loads in the Black Hills, South Dakota, USA. Int. J. Wildland Fire 13:119–129. doi:10.1071/WF02049

Schloeder, C.A., N.E. Zimmermann, and M.J. Jacobs. 2001. Comparison of methods for interpolating soil properties using limited data. Soil Sci. Soc. Am. J. 65:470–479. doi:10.2136/sssaj2001.652470x

Scull, P. 2010. A top-down approach to the state factor paradigm for use in macroscale soil analysis. Ann. Assoc. Am. Geogr. 100:1–12.

Soil Survey Division Staff. 1993. Soil survey manual. Agric. Handbk. 18. U.S. Gov. Print. Office, Washington, DC.

Thomas, G.W. 1996. Soil pH and soil acidity p. 475–490. *In* D.L. Sparks (ed.) Methods of soil analysis. Part 3. Chemical methods. SSSA Book Ser. 5. SSSA, Madison, WI.

Tilman, D.L. 1982. Resource competition and community structure. Princeton Univ. Press, Princeton, NJ.

van Meirvenne, M., and I. van Cleemput. 2006. Pedometrical techniques for soil texture mapping at different scales. p. 323–341. *In* S. Grunwald (ed.) Environmental soil-landscape modeling: Geographic information technologies and pedometrics. CRC Press, Boca Raton, FL.

Velázquez-Alvarez, J., and A. Balancan-S. 2000. SICLIM 1.0: Sistema de Informacion Climatológica. Base de Datos, Periodo 1921–1990. Inst. Mexicano de Tecnología del Agua, Morelos, México.

Venables, W.N., and B.D. Ripley. 2002. Modern applied statistics with S. 4th ed. Springer, New York.

Williams, M.S. 1997. A regression technique accounting for heteroscedastic and asymmetric error. J. Agric. Biol. Environ. Stat. 2:108–129. doi:10.2307/1400643

Zellner, A., and H. Theil. 1962. Three-stage least squares: Simultaneous estimation of simultaneous equations. Econometrica 30:54–78. doi:10.2307/1911287