Journal of Forestry Research (2011) 22(2): 159−166
DOI 10.1007/s11676-011-0144-1

ORIGINAL PAPER

Empirical evaluation of confidence and prediction intervals for spatial models of forest structure in Jalisco, Mexico

Robin M. Reich • C. Aguirre-Bravo • Vanessa A. Bravo • Martin Mendoza Briseño

Received: 2010-07-29; Accepted: 2010-11-11

Abstract: In recent years there has been an increasing interest in developing spatial statistical models for data sets that are seemingly spatially independent. This lack of spatial structure makes it difficult, if not impossible to use optimal predictors such as ordinary kriging for modeling the spatial variability in the data. In many instances, the data still contain a wealth of information that could be used to gain flexibility and precision in estimation. In this paper we propose using a combination of regression analysis to describe the large-scale spatial variability in a set of survey data and a tree-based stratification design to enhance the estimation process of the small-scale spatial variability. With this approach, sample units (i.e., pixel of a satellite image) are classified with respect to predictions of error attributes into homogeneous classes, and the classes are then used as strata in the stratified analysis. Independent variables used as a basis of stratification included terrain data and satellite imagery. A decision rule was used to identify a tree size that minimized the error in estimating the variance of the mean response and prediction uncertainties at new spatial locations. This approach was applied to a set of n=937 forested plots from a state-wide inventory conducted in 2006 in the Mexican State of Jalisco. The final models accounted for 62% to 82% of the variability observed in canopy closure (%), basal area ($m^2 \cdot ha^{-1}$), cubic volumes ($m^3 \cdot ha^{-1}$) and biomass ($t \cdot ha^{-1}$) on the sample plots. The spatial models provided unbiased estimates and when averaged over all sample units in the population, estimates of forest structure were very close to those obtained using classical estimates based on the sampling strategy used in the state-wide inventory. The spatial models also provided unbiased estimates of model variances leading to confidence and prediction coverage rates close to the 0.95 nominal rate.

Key Words: tree-based stratified design; generalized least squares; standardized mean squared error; Landsat-7 ETM+

Introduction

Forest inventory data collected over large geographical regions represent a combination of several spatial phenomena of different origin and appear as complex spatial patterns at different scales. For example, the large-scale variability in forest productivity may be influenced by strong environmental gradients (i.e., elevation, precipitation, temperature), while the variability in forest productivity at the stand level may be influenced by differences in slope, aspect, soil texture, nutrient availability, pH, depth to bedrock, etc. Disturbances (i.e., fire, logging, insects, diseases, grazing, agriculture, etc.) also play an important role in the distribution and structure of the forest resources.

To model such complex spatial patterns, it is generally assumed that the data can be decomposed 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 is 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 are modeled using the covariance structure (Cressie 1991). In surveys covering large geographical regions, the residuals from the regression model may lack spatial structure because of large distances separating the location of the sample data. This lack of spatial structure makes it difficult, if not impossible to use optimal predictors such as ordinary kriging for modeling the small-scale spatial variability in a set of data. 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).

The online version is available at http://www.springerlink.com

Robin M. Reich (✉) • Vanessa A. Bravo
Department of Forest and Rangeland Stewardship, Colorado State University, Fort Collins, CO 80523-1472 USA. Tel: 970-491-6980;
Fax: 970-491-6754; Email: robin@warnercnr.colostate.edu

C. Aguirre-Bravo
USDA Forest Service, Arlington, VA 22209-2131 USA

Martin Mendoza Briseño
Clegio de Postgraduados,Campus Montecillos, Texcoco, Mexico.

Responsible editor: Chai Ruihai

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 (i.e., image pixel) are classified with respect to the predictions of error attributes into homogeneous classes, and the classes are then used as strata in the analysis. Bloch and Segal (1989) first proposed the use of binary classification trees 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, Benedtti 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.

In the tree-based stratified design, strata are formed by dividing the sample data into finer and finer partitions using a binary partitioning algorithm that maximizes the dissimilarities among strata (Breiman et al. 1984). The procedure is sequential and determines a path from a null stratification, a single stratum containing all the sample data, to the extreme in which each sample observation represents a stratum (Benedtti et al. 2005). Once the algorithm partitions the data into new strata, new relationships are developed, assessed, and split into new strata. Given that the partitioning algorithm is designed to produce strata with small deviance, the strata are sub-optimal in the sense that they do not minimize the variance of the estimated stratified mean (Cocchi et al. 2002).

Since the size of the tree, or number of strata, is not limited in the growing process, the final tree may be more complex than necessary to describe the data (Breiman et al. 1984). Since this technique has a tendency to over fit, Breiman et al. (1984) recommended fitting excessively large trees and then pruning the trees back to an optimal size. Pruning is done by cross-validation (Stone 1974) with a cost-complexity function that penalizes predictions made with excessively large trees (Esposito et al. 1997; Ribic and Miller 1998; O'Connor and Wagner 2004). In general, pruning methods aim to simplify decision trees that over-fit the data with reasonable loss in the goodness-of-fit of the model.

Because of its popularity, numerous methods have been proposed for selecting an optimal tree size (Esposito et al. 1997). A comparison of some of the well known pruning algorithms is provided by Esposito et al. (1997). Most of these approaches however, try to produce a less complex tree that is easily interpreted. However, when regression trees are used for describing the small-scale variability in a set of data 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 one might also be interested in constructing confidence intervals for the mean response as well as constructing predictions intervals for estimates at new locations in which no data is available. It is rather straightforward to calculate both confidence and prediction intervals for a regression model (Neter et al. 1975), but how does one take into considerations the variability associated with the regression tree component of the model? The variance associated with each terminal node of a regression tree is influenced by the size of the tree. This variance generally decreases with increasing tree size. The number of observations associated with each terminal node also influences the estimate of the variance. In some instances having a large number of observations at each terminal node may provide better estimates of the variance than having a small number of observations.

This study focuses on data from a state-wide inventory of the natural resources (e.g., forest, grasslands, agriculture) in the state of Jalisco, Mexico (Reich et al. 2008b). We consider if it is possible to enhance the accuracy and precision of models describing the spatial variability in forest structure (canopy closure (%), basal area ($m^2 \cdot ha^{-1}$), cubic volume ($m^3 \cdot ha^{-1}$) and biomass ($t \cdot ha^{-1}$)) by accounting for the error structure using post-stratified estimators based on efficiently defined strata.

Methods

Study site

The state of Jalisco is located in western Mexico between 22°45' and 18°55' N latitude and 101°28' and 105°42' W longitude and contains an area of approximately eight million hectares. 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: (1) The first is the *sub-humid tropical zone* located along the Pacific coast and is characterized by high temperatures, monsoon rains during summer months (730−1 200 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 2 000 m and up to 4 000 m near the Colima volcano in the southern part of the state. In the northern part of this zone, the forests are mesic, while in the south the forests are slightly dyer. Soils are shallow and are derived from metamorphic and volcanic rocks. (2) At higher elevations the *sub-humid temperate zone* covers the greatest portion of the state. Pine, oak and mixed deciduous hardwood forests dominate this zone (1 000−2 600 m). Average annual rainfall ranges from 900−1500 mm. Soils are derived from volcanic rock and have a high content of organic matter. This zone gradually changes to third ecological region. (3) An *arid and semi-arid zone* that has a low annual precipitation of 400mm or less and 8 to 12 dry months. Dominant vegetation includes mesquite-acacia and xerophitic scrub. Soils in this region are shallow and derived from igneous rocks and have a low content of organic matter.

Field data

The data used in this study are from an inventory and monitoring program designed to provide regional and local estimates of the natural resources (e.g., forests, grasslands, agriculture) in the state of Jalisco, Mexico (Reich et al. 2008b). A total of 1 424 permanent plots were located throughout the state using a stratified design that took into consideration the climatic variability within the state and spectral variability of the land cover. At each location a 30m × 30m primary sampling unit corresponding to the spatial resolution of a Landsat 7 ETM+ image was centered on the coordinates assigned to it and laid out in a north-south, east-west manner. Plot locations were verified using a Global Positioning System (GPS) with an estimated accuracy of ± 3 m. Each plot was sub-divided into nine 10m × 10m secondary sampling units (ssu). Five of the nine ssu's were systematically selected for detailed measurement (FIPRODEFO 2004).

Large trees (≥12.5 cm DBH) were measured for diameter at breast height (DBH) and total tree height (m) and recorded by species on each of the five ssu's using a circular plot with a radius of 5m. Tree diameters were used to estimate the basal area (m^2 ha^{-1}) on each sample plot. On all nine of the ssu's a spherical densiometer was used to estimate the average percent canopy closure on each sample plot. Six sets of regression equations were used to estimate cubic volumes ($m^3 \cdot ha^{-1}$) as a function of DBH and total tree height: (1) pines (19 tree species), (2) other conifers (10 tree species), (3) oaks (43 tree species), (4) tropical, industrial (56 tree species), (5) tropical, non-industrial (107 tree species), and (6) other hardwoods (303 tree species). Cubic volumes for other species groups were assumed to be zero. Standing tree biomass ($t \cdot ha^{-1}$) was calculated for all tree species. Standing tree biomass of tropical trees were estimated using regression equations for three major climatic regions: dry (< 1 500 mm of rain per year), moist (1 500 – 4 000 mm of rain per year) and wet (> 4000 mm of rain per year) (Brown et al. 1989; Martinez-Yrizar et al. 1992; Brown and Iverson 1992). These regions coincide in general with those adopted to describe the vegetation and ecology in the state (Reich et al. 2008a; Rzedowski 1978). For tropical palms a regression equation developed by Rich (1986) was used. Biomass estimates for coniferous trees were based on regression equations developed from a database of trees from the USA, India and Puerto Rica (Brown et al. 1989).

Spatial data

Information on the spectral variability of forested vegetation and topography were taken from satellite imagery and a Digital Elevation Model (DEM) of the state. Ten cloud-free Landsat7 ETM+ images obtained during the months of January through March, 2004, were joined together to create a seamless image of the state. The satellite imagery consisted of nine spectral bands (spectral bands 1-5, 6L (low gain), 6H (high gain, see USGS-EROS Data Center web site for more information), 7 and 8). Spectral bands 6L and 6H were thermal bands (57-m resolution), while band 8 was a panchromatic image (15-*m* resolution). These latter three bands were resampled to a 30-m spatial resolution using nearest neighbor techniques. The DEM was obtained from the National Elevation Dataset (NED) as a seamless ArcInfo (ESRI 1995) grid at a 90-m resolution (U.S. Geological Survey (USGS), Gesch et al. 2002). The DEM was resampled to a 30m spatial resolution using bilinear techniques. Raster surfaces of percent slope and aspect were obtained from the Digital Elevation Model. A raster surface of the land cover types was provided by the state of Jalisco.

Model development

For each component of forest structure, a stepwise AIC procedure (Venables and Ripley 2002) was used to identify a subset of independent variables to include in the regression model that minimized the AIC (Akaike 1973). Independent variables considered for inclusion in the regression models included elevation, slope, aspect, Landsat 7 ETM+ bands and forest type. Regression coefficients and variances were estimated using generalized linear model theory (McCullagh and Nelder 1989). For the regression analysis, cubic volumes and biomass were square root transformed to obtain symmetric distributions.

Canopy closure was allowed to enter the basal area model as a predictor, while in the cubic volume and biomass models both canopy closure and basal area could potentially enter the models as predictor variable(s). Since canopy closure and basal area are random variables and correlated with the errors in estimating cubic volumes and/or biomass, the observed values of canopy closure and basal area were substituted with their linear expectations. This removed the randomness of using canopy closure and/or basal area as a predictor variable in the cubic volume and/or biomass models. The errors associated with estimating the dependent variable under these conditions, have zero mean, constant variance and are uncorrelated (Theil 1971).

The small-scale variability (i.e., estimated errors from the regression models) in forest structure was modeled using a tree-based stratified design. Independent variables considered in the stratification included elevation, slope, aspect, Landsat 7 ETM+ bands and forest type. To evaluate the effectiveness of modeling the small-scale variability in forest structure using a tree-based stratified design, different binary regression trees were fit to the residuals from the GLM models. This was accomplished by varying two parameters that controlled the recursive portioning algorithm used to construct the tree. The first parameter *minsize* was used to define the stratum size at which the last split was performed. If *minsize* = 5 (the default value) the algorithm continues to partition the data into stratum if there are at least 5 observations in a given stratum. The parameter *minsize*, was initially set to take on the values of 5, 15 and 25 and then increased in increments of 5 if no optimal tree structure was identified. The second parameter *best* is an integer that was used to control the number of strata, or number of terminal nodes in the tree. If there was no partitioning of the requested size, the next largest partitioning was returned. The number of strata was varied from a minimum of 10 strata, to the maximum number of strata possible, in increments of 5 strata. The maximum number of strata is related to the parameter *minsize* and has an upper bound ~*nobs*/*minsize*, where *nobs* is the number of observations in the data set.

Cross-validation

A 10-fold cross validation (Stone 1974) was used to evaluate the predictive performance of the fitted models (GLM model + regression tree). The data were split into *K*=10 parts consisting of approximately 94 sample plots. For each part, the models were fitted to the remaining *K*-1=9 parts of the data. The fitted model was used to predict the part of the data removed from the modeling process. This process was repeated 10 times so that each observation was excluded from the model fitting step and its response predicted. The prediction errors are then inferred from the predicted minus actual values. While it may be desirable to assess the uncertainty in the models using an independent data set, this may not always be feasible because of time and cost constraints. The cross-validation procedures used in this paper have become a popular method of assessing accuracy and prediction since the articles by Stone (1974) and Geisser (1975).

During the cross-validation, 95% prediction intervals were calculated for the prediction data sets, assuming normality. Confidence intervals for the mean response were also computed for the fitted models. Coverage rates were calculated as the proportion of intervals that covered the observed data. The coverage rate only provides information on whether an observation was included in the interval or not. Additional information can be gained by checking estimates of the variances for unbiasedness. The standardized mean squared error (SMSE) was used to test the null hypothesis of equal variance (Hevesi et al. 1992):

$$\text{SMSE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varepsilon}(s_i)}{\operatorname{var}(\hat{z}(s_i))}. \qquad (1)$$

where $\hat{\varepsilon}(s_i) = (z(s_i) - \hat{z}(s_i))$, is the true error and $\operatorname{var}(\hat{z}(s_i))$ is the estimated variance. The estimated variances were assumed consistent with the true errors if the SMSE fell within the interval $\left[1 \pm 2(2/n)^{1/2}\right]$ (Hevesi et al. 1992).

The effectiveness of the fitted models (GLM model + regression tree) were evaluated using a goodness-of-prediction statistic (G) (Agterberg 1984; Kravchenko and Bullock 1999; Guisan and Zimmermann 2000; Schloeder et al. 2001). The G-value is a measure of the effectiveness a prediction might be relative to that which could have been derived using the sample mean (Agterberg 1984). A G-value equal to 1 indicates perfect prediction, a positive value indicates the model estimates are more reliable than if one had used the sample mean, a negative value indicates the model estimate is less reliable than if one had used the sample mean, and a value of zero indicates that the model produces estimates equivalent to the sample mean. Various measures of the prediction errors were also computed.

A decision rule was adapted to identify a tree size that minimized the error in estimating the variance of the mean response and the prediction variance

$$R = \sqrt{(SMSE_M - 1)^2 + (SMSE_P - 1)^2 + MSEP/(df - n)} \qquad (2)$$

where SMSE$_M$ is the standardized mean square error of the variance of the mean response and SMSE$_P$ is the standardized mean square error of 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 GLM model and n is the number of terminal nodes in the tree. The last term in this equation is a penalty for using a regression tree with an excessive number of terminal nodes, or strata.

Comparison with state-wide inventory

Maps representing the components of forest structure were generated for the models selected to minimize the error in estimating the uncertainty in the spatial estimates. Maps displaying the uncertainty in the spatial estimates were also generated. State-wide estimates of the mean response and prediction variances were obtained by averaging estimates over all 30 m × 30 m pixels in the state (N = 95,693,043). These estimates were compared to classical estimates obtained using the same set of data (Reich et al. 2008b).

Results

Models of forest stand structure

The decision rule identified 55 strata for the canopy closure model and 54 for the basal area model and with a *minsize* of 25 observations (Table 1). The basal area model accounted for 63% of the variability observed on the sample plots while the canopy model accounted for 62% of the variability observed in canopy closure on the sample plots. In contrast, the models for cubic volume and biomass had strata sizes of 85 and 95, respectively and with a *minsize* of 5 observations. The volume model had a G-statistic of 78% and 82% for the biomass model. The slight difference in the G-statistic is because not all trees on a sample plot had a volume, but all trees had a biomass. If a tree species was not economically important no volume estimates were recoded. This weakened the correlation between the volumes observed on the sample plots and the set of predictor variables. Volumes and biomass vary across climatic regions (i.e., tropical, temperate, semi-arid) and position on the landscape (i.e., elevation, slope and aspect). Forest types in the tropical, temperate and semi-arid regions have unique spectral properties which are reflected in the satellite imagery. The models are able to capture this variability by partitioning the data in a large number of small homogeneous strata. Sample plots with the same canopy closure or basal area can have different spectral properties depending on, for example the forest type, stocking level and average tree size and thus, making it more difficult to model using the spectral reflectance and terrain data. This results in a coarser partitioning of the data into a small number of large strata. This coarser partitioning of the data results in a lower G-statistic for these two models.

The confidence and prediction coverage rates for all models were close to the 0.95 nominal coverage rate. The standardized mean squared errors were close to their expected value of unity indicating the variance estimates were consistent with the true errors. The standardized mean squared errors ranged from 0.96 to 1.04. The histograms and plots of the predicted vs. observed values (not shown) did not show any trends suggesting any systematic bias in the models.

Table 1. Fit statistics for the spatial statistical model of selected variables describing forest structure.

Statistic	Canopy Closure (%)	Basal Area ($m^2 \cdot ha^{-1}$)	Cubic Volume ($m^3 \cdot ha^{-1}$)	Biomass ($t \cdot ha^{-1}$)
Sample Size	937	937	937	937
Minsize	25	25	5	5
Tree size	55	54	85	95
G-statistic	0.630	0.622	0.791	0.816
Standardized Mean Squared Error – Model	1.047	0.964	0.997	0.969
Standardized Mean Squared Error – Prediction	1.063	0.968	1.014	0.980
Mean Squared Error Prediction	568.22	67.44	16.09	15.79
0.95 Coverage Rate – Model	0.95	0.95	0.96	0.96
0.95 Coverage Rate – Prediction	0.94	0.95	0.95	0.95
Cost Complexity Function	0.724	0.125	0.033	0.049

Comparison with the state-wide inventory

The final models were used to predict basal area, canopy closure, cubic volumes and biomass along with their prediction variances for all pixels in the state. Fig. 1 displays the spatial variability in estimates of basal area for the state of Jalisco. State-wide estimates were obtained by averaging across all pixels. These estimates were compared to classical estimates obtained using the same data. The classical estimates take into consideration the two-way stratification with differential weighting due to selection probabilities and adjustments for post stratification (Table 2). Estimates obtained from the spatial models were very close to the estimates obtained using the classical approach. Given that both approaches rely on the spectral variability of the land cover in the estimation process it is not surprising the two approaches provide almost identical results.

The percent sampling error for estimating the parameters of forest structure using the classical approach ranged from 9.3% for canopy closure to 17.3% for biomass (Table 2). The percent sampling errors were within the goals set by the state. The minimum, maximum and average prediction standard deviations associated with estimating forest structure using the spatial models are also summarized in Table 2 for comparison with the statewide inventory.

Additional comparisons of the two approaches were made for the three climatic regions (e.g., tropical, temperate, and semi-arid) and 12 economic regions in the state (Reich et al. 2009). The only significant difference observed in estimates for the three climatic zones was for canopy closure in the tropical region. The spatial model predicted denser canopies compared to estimates from the classical approach. At the regional level, there was no significant difference between the classical approach and the spatial estimates for all variables in the temperate climatic zone. Estimates from the spatial models suggest the classical approach provided larger estimates in some regions in the semi-arid climatic zone and smaller estimates in some regions in the tropical climatic zone. The agreement between the classical approach and spatial estimates at the regional, climatic and state level suggest the spatial models may be capable of providing reliable estimates of population parameters for any geographical region, large or small, in the state (Reich et al., 2009). This is a major limitation of the classical approach.

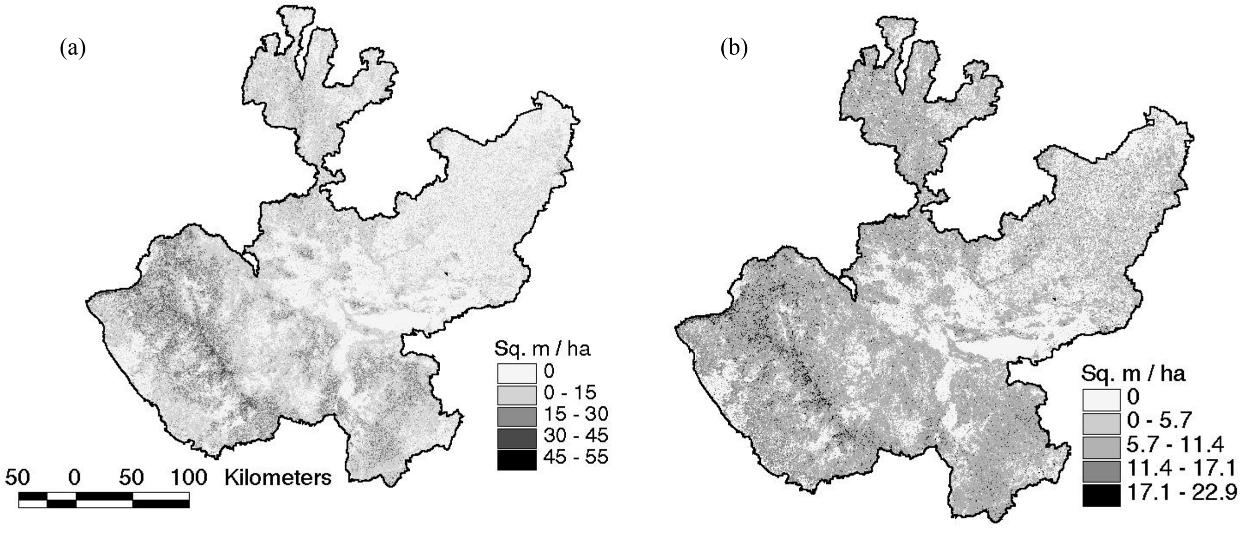

Fig. 1 Spatial distribution of (a) predicted basal area (m^2/ha) and (b) associated prediction standard deviations in the state of Jalisco Mexico.

Table 2. Comparison of estimated means from the state-wide inventory for selected variables describing forest structure and the corresponding estimates from the spatial statistical models.

Variable	Classical Model			Spatial Model					
	Mean	Std. Error[1]	%SE	Mean	Min	Max	Mean[2] $\mathrm{s.d.}(\hat{z}_i)$	Min $\mathrm{s.d.}(\hat{z}_i)$	Max $\mathrm{s.d.}(\hat{z}_i)$
Canopy Closure (%)	48.9	2.28	9.3	52.2	0	100	23.8	16.7	45.6
Basal Area ($m^2 \cdot ha^{-1}$)	7.9	0.53	13.2	8.0	0	51.8	16.9	11.0	27.7
Cubic Volume ($m^3 \cdot ha^{-1}$)	46.2	4.08	17.3	47.2	0	526	17.5	11.1	145.1
Biomass ($t \cdot ha^{-1}$)	58.2	5.09	17.1	56.1	0	120.7	13.4	0.0	63.8

[1] Standard error of the mean (n= 937); [2] Prediction standard deviation for an individual sample unit.

Discussion

The results of this study provided evidence that the use of regression tree-based stratification is adequate for describing the small-scale variability in components of forest structure in the state of Jalisco, Mexico. When combined with a regression model to describe the large-scale variability in forest structure, the final models accounted for 62% to 82% of the variability observed in forest structure on the sample plots. While the regression models alone provided unbiased estimates of the mean response, the variance estimates were not accurate. This leads to inadequate confidence and prediction intervals. To account for the small-scale variability in forest structure, the errors of the regression model were stratified into homogenous classes using terrain and satellite imagery as the basis of stratification. With this approach, sample units (i.e., image pixel) are classified with respect to the predictions of error attributes into homogeneous classes, and the classes are then used as strata in the analysis.

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 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 tree-based algorithm is able to select the most powerful variables for the construction of strata (Cocchi et al. 2002). A disadvantage of this approach is that it requires large sample sizes to ensure the tree-based algorithm is capable of capturing the complex relationship that exists between the variable of interest and its predictors.

The uncertainty of modeling the spatial variability in forest stand structure has been quantified on the basis of two contributing sources of error. The first component is due to the uncertainty in modeling the large-scale variability using multiple regression models, and the second component is due to the uncertainty associated with the stratification scheme used to account for the small-scale variability. The error in spatial predictions can be presented as maps showing the computed estimation errors as well as place prediction intervals around our estimates. The uncertainty in spatial predictions varied considerably, both spatially and for the different models of forest structure. These maps provide information on the location of the main source of uncertainties and indicate where improvements in the model may be realized.

The results of this study highlight the value of quantifying the uncertainty in spatial predictions. The large prediction variances are for estimating the variability of an individual sample unit, and the large prediction variances should not be attributed to the inadequacy of the approach to handle spatially correlated data or due to the quality of the spatial data. Although the tree-based stratified design provides reliable estimates of model variances in this study, theoretical considerations imply this is not always the case. The use of regression trees to model small-scale variability is based on the assumption that a relationship exists between residuals from the regression models and the set of predictors available for stratification. In the present study, this assumption was supported by the results. If there is no relationship, this approach may fail.

The results reported here are comparable to those obtained by Reich et al. (2004) to model the spatial distribution of forest fuel loadings on 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 observed on the field plots. Estimates of the $SMSE_P$ 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. Half of the models had coverage rates less than the nominal 0.95 rate, suggesting the prediction intervals may not be large enough to insure 95% prediction intervals around estimates. These results stress the importance of having a decision rule capable of identifying the appropriate tree structure to ensure unbiased estimates of the variances.

To identify the optimum tree structure (i.e., number and size of the strata) simulations were carried out 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.

Conclusion

To support assessment and monitoring efforts, ecosystem resource managers require spatially explicit information concerning the status of key indicator variables. The models developed in this study can be used to predict unknown population values from the relationships developed from regional survey data. The uncertainty associated with estimates of population parameters can be reduced considerably through the use of information contained in explanatory variables used in the spatial models. The pattern in the fitted surfaces is thus generated by the know patterns in the explanatory variables and can be used to strengthen the spatial inference from the sample data.

Results of this study indicate that the proposed modeling approach described here can simulate the spatial variability of forest stand structure using field data from a state-wide inventory. The spatially generated estimates at the regional and state level were found to generally match the observed data. The models also provided a description of the distribution of the response variables throughout the state. The accuracy and precision of the models are obviously limited by the requirement for large amounts of data of good quality. The validation of the spatial models using an independent set of data is extremely difficult and costly to perform, and may not be realistic. If such data are not available the spatial models may be used as an investigative tool to target additional data collection and identification of response variables requiring improvement. The modeling approach described here can enhance the efficiency and effectiveness of investigating the behavior of spatially distributed phenomena.

References

Agterberg FP. 1984. Trend surface analysis. In: G.L. Gaile and C.J. Willmott, (eds.), *Spatial statistics and model.* Reidel: Dordrecht, The Netherlands. pp. 147–171.

Akaike H. 1973. Information theory and and extension of the maximum likelihood principle. In: N. Petrov and F. Csaki (eds), *Second International Symposioum on Information Theory*. Hungarian Academy Sciences, Budapest, Hungary, pp. 268-281. Repreinted 1992 in *Breakthroughs in Statistics*, S. Kotz and N. Johnson (eds), 1:610-624, Springer Verlag, New York, New York, USA .

Bloch DA, Segal MR. 1989. Empirical comparison of approaches to forming strata – using classification trees to adjust for covariates. *J Amer Statist Assoc,* **84**: 896–905.

Benedetti R, Espa G, Lafratta G. 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. p.17.

Brown S, Gillespie AJR, Lugo AE. 1989. Biomass estimation methods for tropical forests with applications to forest inventory data. *For Sci,* **35**: 881–902.

Brown S, Inverson LR. 1992. Biomass estimates for tropical forests of South and Southeast Asia. *World Resource Review,* **4**: 366–384.

Brieman L, Freidman J, Olshen R, Stone C. 1984. *Classification and Regression trees*. Pacific Grove, CA: Wadsworth and Brooks, p.358.

Carroll SS, Pearson D. 2000. Detecting and modeling spatial and temporal dependence in conservation biology. *Conservation Biology,* **14**: 1893–1897.

Cocchi D, Fabrizi E, Raggi M, Trivisano C. 2002. Regression trees based stratification: an application to the analysis of the Italian post enumeration survey. In: *Proceedings of the International Conference on Improving Surveys*, August 25-28,200, Copenhagen, Denmark. http://www.icis.dk/ICIS-papers/B2_5_2.pdf.

Cressie N. 1991. *Statistics for spatial data*. New York: John Wiley and Sons, 928 pp.

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

Esposito F, Malerba D, Semerao G. 1997. A comparative analysis of methods pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **19**: 476–491.

FIPRODEFO 2004. Manual para la toma de datos de campo: Proyecto de inventario y monitoreo de los recusos natural es de Jalisco. Version 2.0 Deciembre 2004. FIPRODEFO, Guadalajara, Mexico

Geisser S. 1975. The predictive sample reuse method with applications. *J. of American Statistical Association,* **70**: 320–328.

Gesch D, Oimoen M, Greenlee S, Nelson C, Steuck M, Tyler D. 2002. The national elevation dataset. *Photogrammetric Engineering & Remote Sensing,* **68**: 5–32.

Guisan A, Zimmermann NE. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling,* **135**: 47–186.

Hevesi JA, Istok JD, Flint AL. 1992. Precipitation estimation in mountainous terrain using multivariate geostatistics. Part I: structural analysis. *Journal of Applied Meteorology,* **31**: 661–676.

Kravchenko A, Bullock DG. 1999. A comparative study of interpolation methods for mapping soil properties. *Agronomy Journal*, **91**:393–400.

Martinez-Yrizar, Sarukha AJ, Perez-Jimenez A, Rincon E, Maass JM, Solis-Magallanes A, Cervantes L. 1992. Above-ground phytomass of a tropical deciduous forest on the coast of Jalisco, Mexico. *J Tropical Ecology,* **8**: 87–96.

McCullagh P, Nelder JA. 1989. Generalized linear models. 2nd ed. London: Chapman and Hall, 511 pp.

Michaelsen J, Schimel DS, Friedl MA, Davis FW, Dubyah RC. 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *J Veg Science,* **5**: 673–686.

Neter J, Wasserman W, Kutner MH. 1985. *Applied linear statistical models*. Homewwod, IL: Irwin, 1396 pp.

O'Connor, R.J., Wagner, T.L. 2004. A test of regression-tree model of species distribution. *The Auk,* **121**: 604–609.

Reich RM, Lundquist JE, Bravo VA. 2004. Spatial models for estimating fuel loads in the Black Hills, South Dakota, USA. *Int J Wildland Fire,* **13**: 119-129.

Reich RM, Aguirrie-Bravo C, Bravo VA. 2008a. New approach for modeling climatic data with applications in modeling tree species distributions in the states of Jalisco and Colima, Mexico. *Journal of Arid Environments*, **72**: 1343–1357.

Reich RM, Aguirrie-Bravo C, Mendoza-Briseno MA. 2008b. An innovative approach to inventory and monitoring of natural resources in the Mexican State of Jalisco. *Environ. Monit. Assess*, **146**: 383–396.

Reich RM, Aguirre-Bravo C. 2009. Small-area estimation of forest stand structure in Jalisco, Mexico. *J Forestry. Research*, **20**(4): 285–292.

Reich RM, Bonham DC, Aguirrie-Brav C, Chazaro-Basañeza M. 2010. Patterns of tree species richness in Jalisco, Mexico: relation to topography, climate and forest structure. *Plant Ecology*, **210**: 67–84. DOI 10.1007/s11258-010-9738-5.

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

Ribic CA, Miller TW. 1998. Evaluation of alternative model selection criteria in the analysis of unimodal response curves using CART. *J. Applied Statistics,* **25**:685–698.

Rich PM. 1986. Mechanical architecture of arborescent rain forest palms. *Principles,* **30**:117–131.

Schloeder CA, Zimmermann NE, Jacobs MJ. 2001. Comparison of methods for interpolating soil properties using limited data. *American Society of Soil Science Journal,* **65**: 470–479.

Stone M. 1974. Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Society B,* **36**:111−147.

Theil H. 1971. *Principles of econometrics*. John Wiley and Sons, New York. 736 pp.

Venables WN, Ripley BD. 2002, *Modern Applied Statistics with S.* New York: Springer (4th ed), p.495.