

Spatial Statistical and Modeling Strategy for Inventorying and Monitoring Ecosystem Resources at Multiple Scales and Resolution Levels

Reich, Robin M, Professor of Forest Biometry and Spatial Statistics, College of Natural Resources, Colorado State University

Aguirre-Bravo, C., Research Coordinator for the Americas, Rocky Mountain Research Station, USDA Forest Service

Williams, M.S., Mathematical Statistician, Rocky Mountain Research Station, USDA Forest Service

Abstract—A statistical strategy for spatial estimation and modeling of natural and environmental resource variables and indicators is presented. This strategy is part of an inventory and monitoring pilot study that is being carried out in the Mexican states of Jalisco and Colima. Fine spatial resolution estimates of key variables and indicators are outputs that will allow the modeling of complex ecological conditions relevant to resource planners and managers for supporting decision making processes at multiple scale levels. Several procedures for model evaluation and multiscale spatial estimation are also key components of this strategy. Point and spatial statistical estimates will be evaluated so that issues of accuracy and precision can be properly addressed. Products from the application of this strategy will be reported at multiple scales. Final recommendations for field implementation will be made in light of the evaluations of the study results.

Introduction

Growing demands for geospatially explicit information are emerging as a result of complex sustainability challenges. This plus the technological changes that are taking place are accelerating the rate at which traditional approaches to statistical estimation and modeling are being transformed to meet the new needs of the Geospatial Information Age. Driven by these trends, experts and institutions everywhere are continuously reassessing and redirecting their programs and technical capabilities. Institutions that have relevant research and monitoring programs for the assessment and sustainable management of ecosystem resources are at the forefront of implementing the necessary technological transformations.

Over the years, data and information for land management and environmental protection applications have been generated by a variety of means to meet institutional needs for planning and decision making processes. In forestry and natural resources, for example, institutions in most countries have a variety of research and monitoring programs, several with long operational histories. Different sampling strategies (Frayer and Furnival 1999) using various remote sensing technologies (Holgren and Thuresson 1998) and field measurement protocols are common among these institutional programs. Typical

outputs include national and state tabular statistics for describing specific target populations and their related cartography. Now that geographical information systems are widely available, assessment results and a variety of information from these programs have the potential of being reported within a geospatial framework for large-scale strategic applications.

While the intent of national inventory and monitoring programs is to generate statistical summaries and cartography for strategic purposes, it is clear that this information has limited value for tactical and operational applications. These programs, being content-biased due to the nature of their systematic sampling design, can not account for the variety of spatial pattern and ecological conditions that exist at small scales of spatial resolution. For local spatial contexts, where humans interact with ecosystem resources and make a variety of management decisions, it is critical to know where the resources are located, their extent and condition, and the intensity and direction of their ecological change. To effectively address these and other related questions, the data and information provided to local planners and decision makers must be available at multiple levels of spatial resolution (Aguirre 2001).

In light of the above, a spatially balanced estimation and modeling strategy is required to generate geospatial

data and information that meet local stakeholder expectations. Pixel-level statistical modeling opens new opportunities for describing the complexity of ecosystem resource attributes at multiple resolution levels and for advancing the designs of current inventory and monitoring programs. Outputs from spatial statistical models can also be used to develop estimates of population attributes and their measures of central tendency at multiple geographic scales. Due to these and other advantages, spatially explicit products have a very high potential utility for supporting planning, management and decision making processes. Generating these products and making them available may be one of the most defining technological innovations for land management and environmental protection of institutions in the 21st century.

The objective of this paper is to present a spatial statistical modeling approach for inventorying and monitoring ecosystem resources so that the resulting outputs can be used for a variety of multi-scale applications, particularly for local operational contexts. In addition, the paper documents the spatial statistical modeling approach recommended for the Mexican states of Jalisco and Colima's pilot study project on monitoring and assessment for the sustainable management of ecosystem resources.

Integrated Monitoring Framework

In designing an integrated multi-resource inventory and monitoring system to evaluate the condition and change of variables and indicators for sustainable ecosystem resource management (for example, forest, rangeland, agriculture, wildlife, water, soils, biodiversity, etc.) one needs some baseline data for comparison. Given that we are dealing with complex systems, it is not wise to select one or two variables for ecological monitoring purposes. Also, analyzing variables independently of one another may lead to incorrect conclusions because of their spatial inter-dependencies. Statistical estimates and modeling processes are significantly influenced by the spatial patterns of relationships between and among variables. The spatial variability and arrangements of attributes to be measured are important factors to consider in choosing the proper sampling strategy. Techniques commonly used in describing spatial relationships between two or more variables include regression analysis and a variety of geo-statistical procedures that take into consideration spatial and temporal dependencies (Cliff and Ord 1981).

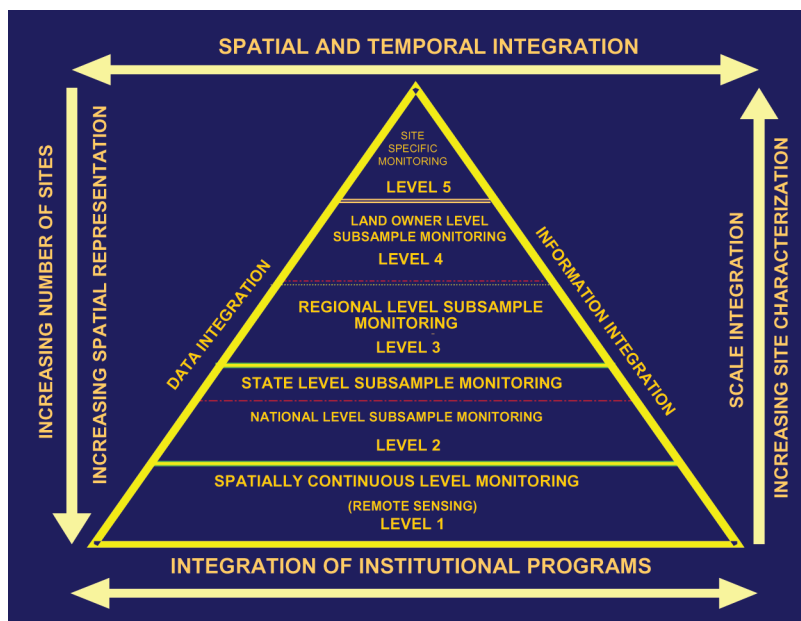


Figure 1. Conceptual model for integration of monitoring design and institutional processes.

The proposed framework for integrated ecosystem resource monitoring will rely on information collected at different spatial scales of resolution and sampling intensities and designs to provide detailed information for regional state and local levels for ecosystem resource planning and management purposes (fig. 1). At each monitoring level, particularly at the local levels (Level 4-5), field measurement protocols and plot designs must be compatible with those used at state and regional geographic scales (Levels 2-3). Remote sensing data from high resolution sensors cuts across all possible monitoring scales (Level 1). National level monitoring assessments will be generated using statistical procedures that are compatible with spatial modeling at smaller scales. Central to this integrated approach is its advantage for optimizing the use of field data from multiple sources when meeting interoperability criteria, thereby minimizing cost and maximizing utility of products from inventory and monitoring programs.

An important feature of this integrated framework is the products and outputs that will be developed at each level and their uses at higher levels of the inventory. At the lowest level, a land cover map will be generated for the entire study area. This map will be constructed from the combination satellite imagery, digital elevation models and a large data set of inexpensive ground information (Level 1). This map will provide general information on the extent and spatial location of the major and minor cover types found in the study area. This map will also be used in Level 2 for area frame construction and as a tool to post-stratify data derived from a systematic sample of permanent ground plots that are collected

for the purpose of long-term monitoring and estimating forest resources at the National and State level (Level 2). While the estimates derived from Level 2 will be design-unbiased and efficient, the small sample size and systematic spacing of ground plots is generally poor for spatial modeling purposes. To address these deficiencies, Level 3 of the inventory will use a stratified sampling scheme to ensure that ground data will be collected in all of the land categories of interest. The Level 2 and 3 ground data will be used, in conjunction with the Level 1 map to develop spatial models that describe the land resources structure of the study area. The goal of Level 4 is to identify areas where the spatial models are not performing well and to collect additional data for the purpose of refining the models in these locations. Thus, the Level 4 data will be a purposive sample of ground plots for the purpose model refinement. Level 5 is reserved for special studies. This may include intensively sampled monitoring locations, but little can be said about the types of analyses performed at this level due to the unknown nature of the issues.

Pilot Study Area

The Pilot Study Area consists of the Mexican southwestern states of Jalisco and Colima with a continental area of approximately nine million hectares (twenty million acres). Though Jalisco is larger in area (90 percent), the state of Colima (10 percent) plays a very distinctive role in the economy of the whole region and diversifies the Pilot Study Area considerably. Four major ecological regions provide the natural resources and environmental conditions that make this region one of the most prosperous in Mexico (fig. 2). The eco-regions are the transversal neo-volcanic system, the southern Sierra Madre, the Southern and Western Pacific Coastal Plain and Hills and Canyons, and the Mexican High Plateau. Linked to these ecological regions, there are several important Hydrological Regions (HR) that drain to the Pacific Ocean (HR12 Lerma-Santiago, HR13 Huicicila, HR14 Ameca, HR15 Costa de Jalisco, HR16 Armeria-Coahuayana, HR18 Balsas, and HR37 El Salado). One of the watersheds, the Lerma-Santiago Hydrological Region is connected to Chapala Lake, the most important source of water for the City of Guadalajara.

Precipitation ranges from roughly 300 mm/year in some locations to more than 1200 mm/year in the higher elevations, with the principal precipitation coming in summer monsoons. The ecological systems of this region cut across the boundaries of other Mexican states. For example, several major watersheds drain through the tropical and subtropical forests of the state of Colima.



Figure 2. Location of the states of Jalisco and Colima, Mexico.

Mostly in the state of Jalisco, water from surface and underground sources is heavily used for agriculture and industrial activities, though a significant portion goes to meet the domestic needs of approximately ten million people. While on average Colima is humid, water in the state of Jalisco is a critically limiting resource that threatens the sustainability of urban and rural ecological and economic systems. Most of the land (85 percent) in the state of Jalisco is privately owned. Small private landowners are the main driving force of economic development in agriculture, forestry, and rangeland economic activities. In contrast to Colima, for example, a small portion of Jalisco's land is owned by ejidos (10 percent), communities (3 percent), and the government (2 percent). Recently, as a result of trade liberalization brought about by NAFTA policies, new industries have been established in these two states and natural resource utilization has increased due to higher population growth rates.

The region's biophysical heterogeneity blends itself to bring about unique habitat conditions for a large diversity of plant and animal species. Within its boundaries, there are a significant number of species of mammals and birds, many of which are severely threatened by human activities. Some of the plant and animal species are endemic to specific locations within the ecological regions that comprise the Pilot Study Area. Extensive areas of pine-oak forest are home to "specialty" birds such as the thick-billed parrot, the Mexican-spotted owl, and woodpeckers. It is thought that habitat loss is the single most important element affecting bird populations in this ecosystem complex. Not much is known about how (in other words, what, when, where, why) plant and animal species are being impacted by human activities. Water and other biological resources are an integral part of these

ecological regions whose services transcend geopolitical domains and jurisdictions.

Data Sources and Description

Data are derived from various sources and using a number of different sampling protocols. One common feature is that data collection and analysis will be designed for a 10 m spatial resolution, meaning that all data will be scaled and stored on a 10 m grid system covering the study area.

GIS Data

GIS grids of elevation, slope, and aspect will be developed from digital elevation models. Grid coverages for each topographic variable will be resampled (Resample function, nearest neighbor, Grid Module (ARC/INFO®, ESRI 1995) to provide a 10 m spatial resolution.

Landsat TM Data

Landsat Thematic Mapper (TM) data contains 8 spectral bands. The data comprise 11 Landsat scenes that are radiometrically and geometrically corrected. Grids of spectral bands 1-8 of a cloud-free, 2002 and 2003 Landsat TM image will be resampled to a 10-m spatial resolution as above and averaged by moving a 3 x 3 pixel window (FOCALMEAN, Grid Module; ARC/INFO®, ESRI 1995) over the resampled grids. Each 10 m x 10 m pixel of resampled Landsat data will therefore represent an average of the surrounding 30 m x 30 m pixels, except for the central 10-m pixel of the original 30 m Landsat pixel, whose value will not change. Resampling is important because not all of the sampling units will fall within spectrally distinct areas; some plots may land in transition zones between spectral classes. Averaging of the Landsat information reduces potential registration errors and better reflects changes in forest structure and vegetative types measured on the ground. Use of other remote sensors (for example, SPOT, MODIS, IKONOS, etc.) will also be investigated as part of this study.

Landcover Point Data

To develop a detailed vegetation map of the pilot study area, point data will be collected throughout the two states to identify major vegetation types. To date, approximately 750 points have been visited. Field crews will identify land areas that clearly meet the definition of each cover type. At the location of each sample point, a Global Positioning System (GPS) is used to obtain the UTM coordinates of the sample points as well as

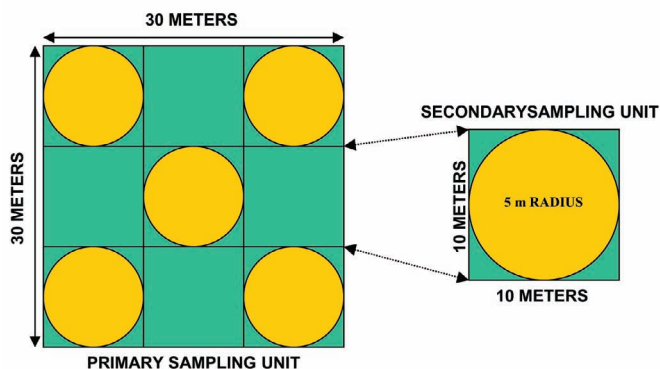


Figure 3. Plot Layout for Primary and Secondary Sampling Units.

information on the dominant vegetation type. The accuracy of the GPS coordinates is approximately 3 m.

Ground Plot Data

The primary sampling unit (PSU) is 30 m x 30 m (fig. 3) square plot corresponding to the size of an individual pixel on a Landsat TM image and consists of nine 10 m x 10 m secondary sampling units (SSUs). Each primary sampling unit will be centered on the coordinates assigned to it and will be laid out in a north-south, east-west manner. The location of each PSU will be verified using a GPS with an estimated accuracy of within 3m.

Because these will be permanent plots, the PSU center will be monumented on the ground. Five of the nine SSU's will be selected for detailed measurement, using a circular plot of 5 meters radius. SSU-1 will be located at the PSU center. The other four SSUs will be located in the four corners of the PSU (fig. 3).

The decision to use a 100 m² SSU is based on study by Reich and others (1992) to determine the optimal plot size for measuring coniferous forests (in other words, tree diameters and tree heights) in El Salto, Durango, Mexico. Results suggest that in highly aggregated stands ($c = 0.052$, table 1) in which individual trees occur in clumps, it is better to sample a small number of trees on each plot by using a small plot size and spreading the plots over a large proportion of the forest, rather than sampling fewer number of plots using a larger plot size (table 1). As the spatial distribution of trees approaches that of a random spatial pattern ($c = 0.5$) the optimal plot size increases. Similar results were observed by Reich and Arvanitis (1992). Both of these studies suggest that the spatial distribution of trees is the most important factor influencing the selection of an optimal plot size. Because of the difficulty in determining the spatial distribution of individual trees, Reich and Arvanitis (1992) developed a technique for estimating the spatial distribution of various stand characteristics

Table 1. Optimal plot size that minimizes total survey time with an allowable error of 10 percent at the 95 percent confidence level, by stand type near El Salto, Durango, Mexico (Reich and others 1992).

Spatial Distribution	Degree of Aggregation (c)	Stocking Level	Number of Stands	Trees/ha	Optimal Plot Size (m ²)
Single Storied Stands					
Aggregated	0.292	Low	1	65.1	115
Aggregated	0.054	Medium	8	762.6	15
Aggregated	0.292	Low	3	268.6	250
Two Storied Stands					
Aggregated	0.054	Low	12	327.2	10
Aggregated	0.054	Medium	23	1085.2	10
Aggregated	0.292	High	11	2478.2	205
Aggregated	0.054	Low	1	647.7	25
Aggregated	0.292	Medium	4	1244.1	31

using simple counts of “in” trees on either variable or fixed area plots.

Several kinds of subplots will be located within each of the 5 m radius plots (fig. 4) and different measurements will be made on each plot type. All large trees (>12.5 cm DBH) will be measured on each of the 5 m plots. Observed attributes will be specified in the field sampling and indicators measurement manuals. Saplings (2.5 cm < DBH < 12.5 cm) will be measured on a circular plot (3m radius) co-located at the center of each tree subplot. Within each of the 5 m radius plots will be 3 square plots, each measuring 1 m x 1 m. The first 1 m² quadrat will be located at the center of the 5 m radius plot. The remaining two 1 m² are located 6 m from the center plot, on a diagonal of the 5 m radius plot (fig. 4). Seedlings (height > 30 cm and DBH < 2.5 cm) will be sampled on the three 1 m² quadrats. In addition to counting seedlings, the percent cover of herbaceous plants, shrubs, and tree species < 30 cm tall will be recorded.

On all nine of the SSUs, a spherical densiometer will be used to estimate canopy closure while an angle gauge will be used to estimate basal area by species. This

information will be used to correlate the detailed vegetation and soils data collected on the five SSUs.

To estimate fuel loadings, a 14.14 m transect will be established diagonally across each of the 5 m radius plots, proceeding at 45 degrees (fig. 4). This will be referred to as the 14 m transect. Line intersect techniques will be used to estimate fuel loadings of large woody material (sound and rotten) > 7.5 cm in diameter. All large woody material intersecting the 14 m transect will be counted and their cross-sectional areas measured by genus. Small woody material (0-0.6 cm, 0.6-2.4 cm, 2.4-7.5 cm) will be counted on a diagonal transect on the three 1 m² plots. In each case, the mean height of fuels in each sampled diameter class, as well as the slope of the diagonal transect will be measured, and reported, respectively. Soils attributes will be observed on each 5 m radius plot. Any destructive soil samples will be collected outside the west side (270 degree Azimuth) of the primary sampling unit and at a distance of 5 meters of the plot boundary line.

Most of the indicator variables are compatible with those used by the USDA Forest Service and Canadian ecosystem resource monitoring programs. Other indicator variables can be integrated into this pilot study as resources become available and the need dictates to ensure comparability and interoperability of indicators with participating government agencies from the USA and Canada.

Sampling Design

The development of the sampling and plot designs is complicated by the diversity of variables and indicators to be assessed, and the need to assess the ecosystem resources at a range of scales, the need to monitor the indicators over time, and the need to do so efficiently. To meet national and state level objectives for ecosystem resource assessments while providing information needed to develop geostatistical models to estimate key attributes at local scales, a stratified random sampling design will be employed. Stratification generally provide

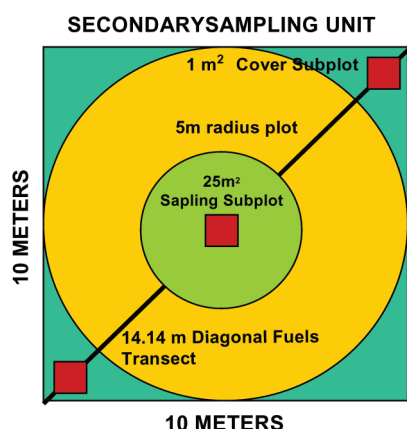


Figure 4. Layout of Tree and Cover Subplots of SSUs.

more precise estimates compared to a simple random or systematic sample of the same size, while providing estimates of population parameters for individual strata (Schreuder and others 1993).

In the first phase, the pilot study area will be stratified by vegetation type (for example, temperate forest, tropical forests, grasslands, mesquite forests, agricultural lands, etc.). Strata will be defined using a detailed vegetation map of the pilot study area developed using the independent set of point data. Each stratum will have a known size and will be used as weights to obtain area-wide estimates. The number of sample plots within stratum will be allocated proportional to the size of the stratum and the variability within stratum. In the second phase, Landsat TM data will be used to obtain an unsupervised classification of the spectral variability associated with each of the dominant vegetation types, or stratum identified in phase one. The number of spectral classes, or strata in the second stage, will vary, depending on the spectral variability observed within each stratum. An equal number of sample plots will be randomly located within each spectral class. This will ensure that the sample plots will cover the spectral variability associated with the Landsat TM image which is essential for spatially interpolating the sample data. The field crews will locate the plots at the UTM coordinates given to them – accurate location of the points is important both for spatial modeling as well as to future relocation of these permanent plots. Plot locations will be kept secret. The opportunity also exists to intensify for local areas within land tenure units, MAUs, or administrative units, as budgeting allows.

Modeling Methods

Vegetation Map

The vegetation map of the pilot study area will be constructed using the Landsat TM, climatic data, vegetation point data, and field sample data. A stepwise decision tree (Breiman and others 1984, Friedl and Brodley 1997, De'Ath and Fabricus 2000) will be used to identify independent variables (Landsat TM bands, elevation, slope, or aspect) that are important in discriminating among vegetation types. The decision tree uses a binary partitioning algorithm that maximizes the dissimilarities among groups to compare all possible splits among the independent variables and splits within each independent variable to partition the data into new subsets. Once the algorithm partitions the data into new subsets, new relationships are developed to split the new subsets. The algorithm recursively splits the data in each subset until either the subset is homogeneous or the subset contains

too few observations (< 5) to be split further. To prevent over fitting the data, a pruning algorithm (Friedl and Brodley 1997) will be used to eliminate subsets that were fit to noise in the data. Decision tree criteria will then be used as 'training' statistics to classifying the 2002 and 2003 Landsat image (fig. 5).

Spatial Modeling

Ecosystem resource attributes and indicators measured on the sample plots (in other words, canopy closure, basal area, fuel loadings, soil texture, understory vegetation, density of seedling/saplings, etc.) will be modeled to a 10 m spatial resolution using procedures developed by Joy and Reich (2002). Multiple regression analysis will be used to develop a trend surface (TS) model to explore the coarse-scale variability (in other words, non-stochastic mean structure) in continuous measures of forest structure as a function of elevation, slope, aspect, landform, and Landsat TM bands. To account for interactions between vegetation types and other independent variables, dummy variables will be introduced in the models as interactions with elevation, slope, aspect, landform, and data from the eight Landsat bands. For each component of forest structure modeled, a stepwise procedure will be used to identify the best subset of independent variables (main effects and interactions) to include in the TS models.

To describe the fine-scale spatial variability (in other words, residuals associated with the TS models) in ecosystem resource attributes and indicators will be modeled using binary regression trees (RT). The RT is a non-parametric approach to regression that compares all possible splits among the independent (continuous) variables using a binary partitioning algorithm that maximizes the dissimilarities among groups. Once the algorithm partitions the data into new subsets, new relationships are developed, assessed, and split into new subsets. The algorithm recursively splits the data in each subset until either the subset is homogeneous or the subset contains too few observations (for example, < 5) to be split further. Interpolation using RTs is relatively insensitive to sparse data (Joy and Reich 2002). Independent variables considered in the RT will include elevation, slope, aspect, landform, Landsat TM band readings, and vegetation type, the latter being treated as a categorical variable. To avoid over-fitting the RTs, a 10-fold cross-validation procedure (Efron and Tibshirani 1993) will be used to identify the tree size (in other words, number of terminal nodes) that minimizes the total deviance (in other words, error) associated with the trees.

Semi-variograms which describe how the sample variance changes as a function of distance will be used to evaluate spatial dependencies among the residuals from the various models. If the residuals exhibited

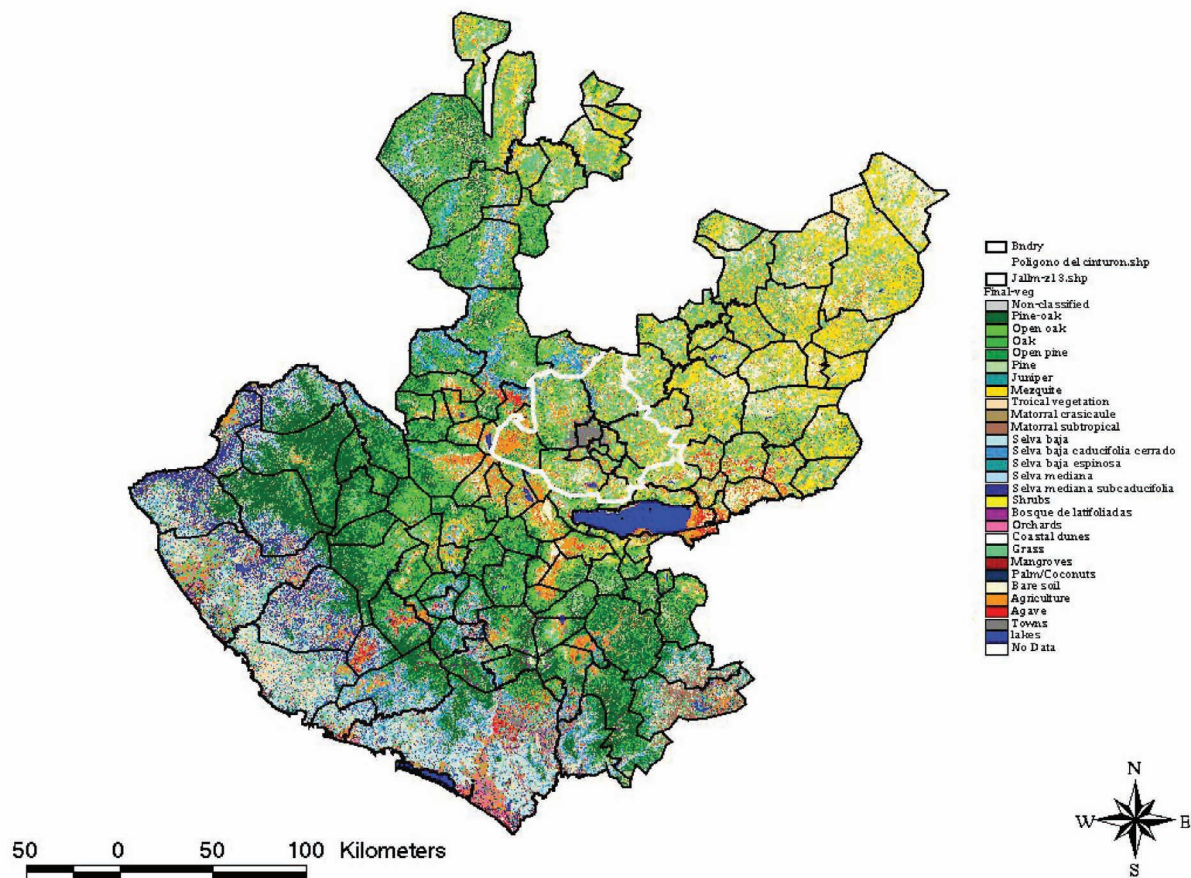


Figure 5. Preliminary vegetation map of the states of Jalisco and Colima, Mexico. The vegetation map is based on point data collected at 2000 locations and a 2002 Landsat TM imagery. The missing Landsat TM images for 2002 will be acquired and used in developing the final vegetation map of the study area.

spatial dependencies, a spatial autoregressive (SAR) model will be used to obtain generalized least squares (GLS) estimates of the regression coefficients associated with the TS model (Upton and Fingleton 1985). The model residuals will be reevaluated to ensure the removal of the spatial dependencies. In fitting the SAR models, a spatial weight matrix (in other words, a block diagonal matrix) based on inverse distance weighting will be used to represent the spatial dependencies among the PSUs and SSUs.

Grids representing the various components of forest structure will be generated for the best fitting TS model using the model's parameter estimates. Similarly, grids representing the error in each TS model will be generated by passing each grid for the appropriate independent variable through the RTs. The final predicted surfaces for each component of forest structure will be obtained from the sum of the TS and RT grids.

Model Evaluation

The effectiveness of the final models will be evaluated using a goodness-of-prediction statistic (G) (Agterburg 1984, Guisan and Zimmermann 2000, Kravchenko and

Bullock 1999, Schloeder and others 2001). The G-value, measures how effective a prediction might be relative to that which could have been derived by using the sample mean (Agterburg 1984):

$$G = \left(1 - \frac{\sum_{i=1}^n [z_i - \hat{z}_i]^2}{\sum_{i=1}^n [z_i - \bar{z}]^2} \right), \quad [1]$$

where Z_i is the observed value of the i th observation, $\hat{Z}_i$ is the predicted value of the i th observation, and $\bar{Z}$ is the sample mean. A G-value equal to 1 indicates perfect prediction, a positive value indicates a more reliable model than if one had used the sample mean. A negative value indicates a less reliable model than if one had used the sample mean, and a value of zero indicates that the sample mean should be used to estimate Z .

A 10-fold cross-validation (Efron and Tibshirani 1993) will be used to estimate the prediction error for each variable modeled. The data will be split into $K=10$ parts consisting of approximately 15 sample plots. For each k th part, the TS and RT models are fitted to the remaining $K-1=9$ parts of the data. The fitted model is used to predict the k th (in other words, removed) part of

the data. This process is repeated 10 times so that each observation is excluded from the model construction step and its response predicted.

To evaluate the effectiveness of the models, we will compute various measures of the prediction error. Prediction bias (Williams 1997) will be calculated for each validation data set as a percentage of the true value. Accuracy (Kravchenko and Bullock 1999) will be measured by the mean absolute error (MAE), which is a measure of the sum of residuals (in other words, actual minus predicted) and the root mean squared error (RMSE), which is a measure of the square root of the sum of squared residuals. Small MAE values indicate models with few errors, while small values of RMSE indicate more accurate predictions on a point-by-point basis. To assess the estimation uncertainty in the models (Isaaks and Srivastava 1989) the estimation error variance (EEV), $\hat{\sigma}_i^{2(-k(i))}$ for each observation in the kth part of the data will be calculated:

$$\hat{\sigma}_i^{2(-k(i))} = MSE^* \left[(X_i^{-k(i)})' (X^*{}' X^*)^{-1} (X_i^{-k(i)}) + 1 \right] + 2MSE(RT) + 2COV(\hat{Y}, \hat{\eta}) \quad [2]$$

where MSE^* is the regression mean squared error for the TS model fitted using K-1 parts of the data, X^* is a matrix of independent variables associated with the K-1 parts of the data, $X_i^{-k(i)}$ is a vector of independent variables associated with the ith observation in the kth part of the data, $\hat{Var}^{(RT)}$ is the mean squared error of the RT used to describe the error in the TS model, and $COV(\hat{Y}, \hat{\eta})$ is the covariance between the estimated values, $\hat{Y}$, from the TS model and the predicted residuals, $\hat{\eta}$, from the RT for the K-1 parts of the data. The consistency between the EEV, $\hat{\sigma}_i^{2(-k(i))}$, and the observed estimation errors (in other words, true errors), $e_i^{-k(i)} = (Z_i - \hat{Z}_i^{-k(i)})$, will be calculated using the standard mean squared error (SMSE) (Havesi and others 1992):

$$SMSE = \frac{1}{n} \sum_{i=1}^n \frac{(e_i^{-k(i)})^2}{\hat{\sigma}_i^{2(-k(i))}} \quad [3]$$

EEVs are assumed consistent with true errors if the SMSE falls within the interval $\left[1 \pm 2(2/n)^{-1/2}\right]$ (Havesi and others 1992). Paired t-tests ($\alpha = 0.05$) will be used to test for differences between the mean estimation errors and zero.

Data Collection and Model Building Phases

Data collection and model building will be carried out simultaneously to ensure the development of the most reliable models.

Phase I. In this phase, point data will be collected throughout the pilot study area to identify both the major and minor vegetation types. This information will be used to develop a preliminary vegetation map of the pilot study area (see section on Vegetation Map). The preliminary vegetation map will be used to identify strata for the purpose of locating sample plots in the field (fig. 5).

Phase II: In this phase, one-third of the sample plots will be located in the field and measured. In addition, point data will also be collected. The point data along with the classification of the field plots will be used to update the vegetation map of the pilot study area. Preliminary models will be developed for key indicator variables such as canopy closure to identify geographical regions or vegetation types within the pilot study area that have large errors associated with their estimation. This information will be used to allocate the next group of sample plots to various strata.

Phases III and IV: The steps outlined in Phase II will be repeated until all of the sample plots have been located in the field and measured.

Phase V: The point data collected in Phases I-IV along with the classification of the sample plots measured in Phases II-IV will be used to develop the final vegetation map of the pilot study area. Also during this phase, spatial models will be developed for all of the ecosystem resource attributes and indicators variables measured on the sample plots (see section on Spatial Modeling).

Multi-Scale Estimation (Model-Based)

In addition to being able to assess the level of uncertainty associated with the spatial models, it is also important that the models are capable of providing estimates at any spatial scale or level of support. It is also important that we are able to place bounds on the error of estimation. To accomplish this it is important that the PSU remain intact as much as possible by not splitting them in half. This may not be possible near boundaries, and in such cases, the formula presented below will have to be modified to take into consideration PSU of unequal sizes. To demonstrate this concept, assume one is interested in estimating the mean (for example, canopy closure, basal area, height understory vegetation etc.) per SSU within a specified geographical unit and place a bound on the error of estimation. Assume the area of

interest contains n PSUs consisting of $m = 9$ SSU's. The modeled surfaces are used to provide an estimate ($\hat{\bar{Z}}$) on each of the nm SSU's, along with the model prediction variance ($\hat{\sigma}^2$) using Eq. 2. An estimate of the mean value per SSU ($\hat{\bar{Z}}_{sp}$) is given by:

$$\hat{\bar{Z}}_{sp} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \hat{Z}_{ij} = \frac{1}{n} \sum_{i=1}^n \hat{\bar{Z}}_i \quad [4]$$

where $\hat{\bar{Z}}_{ij}$ is the estimated value on the j th SSU from PSU i , and $\hat{\bar{Z}}_i$ is the average for the i th PSU. If PSUs of the same size are sampled, the total sum of squares associated with estimating the mean can be partitioned into the within-PSU sum of squares (SSW) and the between-PSU sum of squares (SSB) (Scheaffer and others 1996). With appropriate divisors, these sum of squares become the usual mean squares of an analysis of variance. The within-PSU mean square (MSW) is given by

$$MSW = \frac{SSW}{n(m-1)} = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (Z_{ij} - \bar{Z}_i)^2 \approx \frac{1}{nm^2} \sum_{i=1}^n \sum_{j=1}^m \hat{\sigma}_{ij}^2 \quad [5]$$

where $\frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (Z_{ij} - \bar{Z}_i)^2$ is the MSW one would typically use in cluster sampling and $\frac{1}{nm^2} \sum_{i=1}^n \sum_{j=1}^m \hat{\sigma}_{ij}^2$ is its equivalent using the EEV formula (Joy and Reich 2002). The between-PSU mean square (MSB) is given by:

$$MSB = \frac{SSB}{n-1} = \frac{m}{n-1} \sum_{i=1}^n (\bar{Z}_i - \bar{Z}_{sp})^2 \approx \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \hat{\sigma}_{ij}^2 \quad [6]$$

where $\frac{m}{n-1} \sum_{i=1}^n (\bar{Z}_i - \bar{Z}_{sp})^2$ is the general formula for calculating the MSB and $\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \hat{\sigma}_{ij}^2$ is its equivalent using the EEV formula (Joy and Reich 2002). The MSB can be used to calculate the variance of $\hat{\bar{Z}}_{sp}$ as follows:

$$\hat{V}(\hat{\bar{Z}}_{sp}) = \frac{MSB}{nm} \quad [7]$$

Using these relationships it is possible to obtain local estimates of any of the modeled variables to any spatial scale along with their corresponding estimates of the variance.

Global Estimation (Sampling Design-Based)

The field data may also be used to obtain global estimates of the mean and variance for the states of Jalisco and Colima for individual vegetation types. Within a given vegetation type, i ($i=1,2,\dots,L$) an estimate of the mean and variance of some attribute, z , can be obtained using the formula for a stratified random sample (Cochran 1977, Schreuder and others 1993):

$$\bar{z}_i = \frac{1}{N_i} \sum_{j=1}^c N_{ij} \bar{z}_{ij} \quad [8]$$

$$\hat{V}(\bar{z}_i) = \frac{1}{N_i^2} \sum_{j=1}^c N_{ij}^2 \left(\frac{N_{ij} - n_{ij}}{N_{ij}} \right) \frac{s_{ij}^2}{n_{ij}} \quad [9]$$

where N_{ij} is the number of PSUs in the j th spectral class ($j = 1, 2, \dots, C$), $N_i = \sum_{j=1}^C N_{ij}$ is the number of PSUs in the i th vegetation type, n_{ij} is the sample size in the j th spectral class in the i th vegetation type, s_{ij}^2 is the sample variance of the j th spectral class in the i th vegetation type, and $\bar{z}_{ij}$ is the sample mean for the j th spectral class in the i th vegetation class.

The state-wide estimates of the mean and variance of the variable of interest are again obtained using the formula for a stratified random sample (Cochran 1977, Schreuder and others 1993):

$$\bar{z} = \frac{1}{N} \sum_{i=1}^L N_i \bar{z}_i = \frac{1}{N} \sum_{i=1}^L \sum_{j=1}^C N_{ij} \bar{z}_{ij} \quad [10]$$

$$\hat{V}(\bar{z}) = \frac{1}{N^2} \sum_{i=1}^L N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \hat{V}(\bar{z}_i) = \frac{1}{N^2} \sum_{i=1}^L \left(\frac{N_i - n_i}{N_i} \right) \sum_{j=1}^C N_{ij}^2 \left(\frac{N_{ij} - n_{ij}}{N_{ij}} \right) \frac{s_{ij}^2}{n_{ij}} \quad [11]$$

where N is the total number of PSUs in the states of Jalisco and Colima and $n_i = \sum_{j=1}^C n_{ij}$ is the sample size in the i th vegetation class.

These formula can be modified to provide estimates of the mean and variance for the SSUs.

Plot Remeasurement

Sample plots will be remeasured on a cycle of a one-to-five years with an average of 25 percent of the plots being remeasured in a given year. The rate of remeasurement

will be based on the temporal variability associated with the various vegetation types. For example, agricultural areas would be expected to change very rapidly from one year to the next, as compared to the mesquite forests which are very stable over time.

In the second year, a new cloud free, Landsat TM imagery will be acquired of the pilot study area. The Landsat imagery will be normalized with respect to the Landsat imagery used in the initial survey. The two Landsat images will be differenced to identify areas in which the spectral characteristics have changed. Cluster analysis will be used to stratify the pilot study area into five to ten strata with similar changes in the spectral variability. Based on their spectral properties, the sample plots will be assigned to one of the five to ten strata representing changes in the landscape. Within each stratum, sample plots will be randomly selected, without replacement, for remeasurement. The proportion of sample plots selected from each stratum will depend on the number of sample plots assigned to a given stratum. If there are no sample plots assigned to a particular stratum, there is an opportunity to establish new sample plots to expand the database used to make inferences about the resources within the pilot study area.

Spatial-Temporal Modeling

To model the changes in ecosystem resource attributes and indicators over time, first order differencing will be used (Brockwell and Davis 1991). This first order difference is defined as

$$\Delta z_t = z_t - z_{t-1} \quad [12]$$

where z_t describes the process at time t . The changes observed on the remeasured sample plots will be modeled as a function of changes in the spectral bands associated with the sample plots, elevation, slope, aspect, and vegetation type. The approach used in the modeling will be similar to the one used in developing the original models. An estimate of the process at time t will be obtained by adding the predicted surface of change to the predicted surface of the process at time $t-1$:

$$\hat{z}_t = \hat{z}_{t-1} + \Delta \hat{z}_t \quad [13]$$

In subsequent years, it may be necessary to use higher order differences to eliminate quadratic or higher order trends.

Identifying Micro-Ecological Management Units

Resource managers are constantly trying to improve the way they manage the natural resources under their care. Typically, the area of interest is sub-divided into management units, or stands, based on certain characteristics, such as canopy closure and/or species composition,

and then each area is managed on an individual basis. Unfortunately, the definitions used in the creation of these management units, or stands, may not be compatible with different management objectives.

Using the techniques discussed earlier, resource managers can generate response surfaces representing important resource attributes (in other words, canopy closure, basal area, volume growth, fuel loadings, biomass, understory vegetation, etc.) under their management. Using a collection of these surfaces to represent certain ecological or management conditions (in other words, diversity of resident and migratory birds, species richness, wildlife habitat suitability, volume production, fire hazard, etc.) one can apply a multivariate spatial clustering algorithm to identify "micro-ecological" units that have similar spatial characteristics. Thus, the management units identified for the production of volume may be different from those identified to maximize the diversity of resident and migratory birds, and so on. The algorithm applies a k-means clustering algorithm to the selected response surfaces, and clusters the individual pixels of the response surfaces into k clusters. K-means is a nonhierarchical clustering method that uses nearest centroid sorting to iteratively minimize the Euclidian distance between cluster means (Hartigan 1975).

Conclusions

The science and art of spatial statistics and modeling open new opportunities to advance the systems for inventorying and monitoring ecosystem resources and the environment. In research and other applications, these technologies provide a flexible framework for integrating multiple sources of data and information for spatial modeling at multiple scales and resolution. Integrating field data and remote sensed data through a geostatistical-based approach brings about significant gains in statistical and economic efficiency. However, for the achievement of successful results, it is essential to take into account a variety of technical considerations when using these technologies for practical applications.

Statistical estimates and modeling processes are significantly influenced by the spatial patterns that exist between and among variables of interest. The spatial variability and arrangements of these attributes are important factors to consider in choosing the proper sampling strategy. If the sampling design does not capture the spatial variability in the data it may not be possible to spatially interpolate the field data. It is also important that the field data be collected at the desired spatial resolution. For example, if the field data is collected on a systematic grid, it may not be possible to spatially interpolate the

data to a finer spatial resolution, especially, if the scales of pattern are smaller than the grid spacing used to collect the data. If Landsat imagery is being used in the interpolation process, it is also important that the sample plot corresponds as closely as possible to the size and shape of the pixels in the imagery. This tends to minimize the errors associated with what is being measured on the ground and what the satellite senses.

In addition to be able to spatially interpolate the field data, it is important to evaluate the individual models as to their predictive performance. This provides useful information to the users in terms of the accuracy and precision of estimates in areas not sampled.

The Jalisco-Colima Pilot Study constitutes a test-bed for using and learning about the application of these new technologies. While these techniques have been applied to smaller areas (< 370,000 ha) their performance when applied to more diverse and larger geographical areas is generally unknown.

References

- Aguirre-Bravo, C. 2001. Conceptual Framework for Inventorying and Monitoring the State of Jalisco's Ecosystem Resources at Multiple Scales and Resolution Levels. FIPRODEFO, Secretary of Rural Development, State Government of Jalisco, Mexico. 40p.
- Agterburg, F. P. 1984. Trend surface analysis. In: Spatial statistics and models, G. L. Gaile and C.J. Willmott (eds.). Reidel, Dordrecht, The Netherlands, pp. 147-171.
- Brockwell, P. J., and R. A. Davis. 1991. Time series: Theory and Methods. Springer, New York. 577p.
- Brown, J. K. 1974. A planar intersect method for sampling fuel volume and surface area. *Forest Science* 17: 96-102.
- Cliff, A., and J. K. Ord. 1981. Spatial Processes, Models and Applications. Pion, Ltd. London.
- Cochran, W. G. 1977. Sampling Techniques. 3rd ed. John Wiley and Sons, New York. 428p.
- De'Ath, G., and K. E. Fabricus. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81: 3178-3192.
- Efron, B., and R.J. Tibshirani. 1993. *Tibshirani. An introduction to the bootstrap*. New York, Chapman and Hall.
- ESRI. 1995. ARC/INFO® Software and on-line help manual. Environmental Research Institute, Inc., Redlands, CA.
- Frayser W.E., and G.M. Furnival. 1999. Forest Survey Sampling Designs: A History. *Journal of Applied Forestry*. 97(12): 4-8.
- Friedl, M. A., and C. E. Brodley. 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sensing and the Environment* 61: 399-409.
- Guisan, A., and N.E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 47-186.
- Hartigan, J. A. 1975. *Clustering algorithms*. John Wiley and Sons, New York, 351p.
- Hevesi, J. A., J. D. Istok and A. L. Flint. 1992. Precipitation estimation in mountainous terrain using multivariate geostatistics. Part I: structural analysis. *Journal of Applied Meteorology* 31: 661-676.
- Holmgren, P., and T. Thuresson. 1998. Satellite Remote Sensing for Forestry Planning: A Review. *Scand. J. For. Res.* 13: 90-110.
- Isaaks, E. H., and R.M. Srivastava. 1989. *An introduction to applied geostatistics*. New York, Oxford University Press.
- Joy, S. M., and R. M. Reich. 2002. Modeling forest structure on the Kaibab National Forest in Arizona. *Forest Science*, In review.
- Kravchenko, A., and D. G. Bullock. 1999. A comparative study of interpolation methods for mapping soil properties. *Agronomy Journal* 91: 393-400.
- Reiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression trees* (Belmont, California: Wadsworth Ind. Group).
- Reich, R. M., and L. G. Arvanitis. 1992. Sampling unit, spatial distribution of trees, and precision. *North. J. Appl. For.* 9:3-6.
- Reich, R. M., C. Aguirre-Bravo, and M. Iqbal. 1992. Optimal plot size for sampling coniferous forests in El Salto, Durango, Mexico. *Agrociencia* 2:93-106.
- Schreuder, H. T., T. G. Gregoire, and G. B. Wood. 1993. *Sampling methods for multiresource forest inventory*. John Wiley and Sons, New York. 446p.
- Schreuder, H. T., Williams, M. S., Aguirre-Bravo, C., Patterson, P. L., and H. Ramirez. 2003. Statistical strategy for inventorying and monitoring the ecosystem resources of the states of Jalisco and Colima at multiple scales and resolution levels.
- Schloeder, C. A., N.E. Zimmermann and M.J. Jacobs. 2001. Comparison of methods for interpolating soil properties using limited data. *American Society of Soil Science Journal*. 65:470-479.
- Upton, G. J. G., and B. Fingleton. 1985. *Spatial data analysis by example*. Vol. 1, Point pattern and quantitative data. New York, John Wiley and Sons.
- Williams, M. S. 1997. A regression technique accounting for heteroscedastic and asymmetric error. *Journal of Agriculture, Biology and Environmental Statistics* 2:108-129.