Statistical Strategy for Inventorying and Monitoring the Ecosystem Resources of the State of Jalisco at Multiple Scales and Resolution Levels

Robin M. Reich, Professor, Forest Biometry, Colorado State University, Ft Collins, CO
Hans T. Schreuder, retired Mathematical Statistician Rocky Mountain Research Station, USDA Forest Service, USA

Abstract—*The sampling strategy involving both statistical and in-place inventory information is presented for the natural resources project of the Green Belt area (Centuron Verde) in the Mexican state of Jalisco. The sampling designs used were a grid based ground sample of a 90x90 m plot and a two-stage stratified sample of 30 x 30 m plots. The data collected were used to present strategic information for the green belt area as well as mapped information on crops and forest conditions. The experience gained in this study will be used to plan a statewide inventory for the states of Jalisco and Colima, Mexico and ultimately for the entire country of Mexico.*

Introduction

In Mexico, individual agencies address the extent, status, and trends of selected natural resources in response to their specific missions. No agency, or group of agencies, examines the interactions and interdependence among multiple natural resource components from an integrated ecosystem perspective. Available information often is at scales that have limited value for state planning and policy-making. While useful for strategic national level planning, this information has limited use for supporting regional and local decision-making resource applications. Similarly, there is no integrated program to periodically assess multiple natural resources at regional and local scales, and at multiple resolution levels. Accordingly, the ecosystem resource monitoring initiative, of which this document is a part, will enable comprehensive assessments of Jalisco's natural resources for the management of their sustainability.

This ecosystem resource monitoring initiative addresses a number of critical issues that stakeholders in the states of Jalisco and elsewhere in Mexico are facing to insure the environmental sustainability for present and future generations. Land management agencies charged with this responsibility have come to the realization that current data and information available are insufficient to confront successfully the ecological and economic challenges of ecosystem resource sustainability. In light of the above, the purpose of this technical document is to describe the statistical approach for inventorying and monitoring the ecosystem resources of the Mexican state of Jalisco at multiple scales and resolution levels. Design-based inference is used to ensure a minimum number of assumptions.

Overview of the Sampling Designs

Unlike many National level inventories, such as the US Department of Agriculture, Forest Service Forest Inventory and Analysis program (FIA), the goal of the Jalisco pilot project (JCPP) is to design an integrated inventory system that provides both traditional assessments of population totals and means as well as spatially realistic maps that describe the location and distribution of various attributes of the population. In other words, an inventory is designed that meets strategic, management, and local needs. The products produced by the inventory include reliable estimates of means and totals for important ecosystem characteristics and maps describing the spatial and temporal properties of the ecosystem.

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 (forest, rangeland, agriculture, wildlife, water, soils, biodiversity, etc.) one needs some baseline data for comparison. Because one is generally dealing with complex systems, it is not wise to focus on only one or two variables for ecological monitoring purposes. Also, analyzing these variables independently of one another may lead to incorrect conclusions because

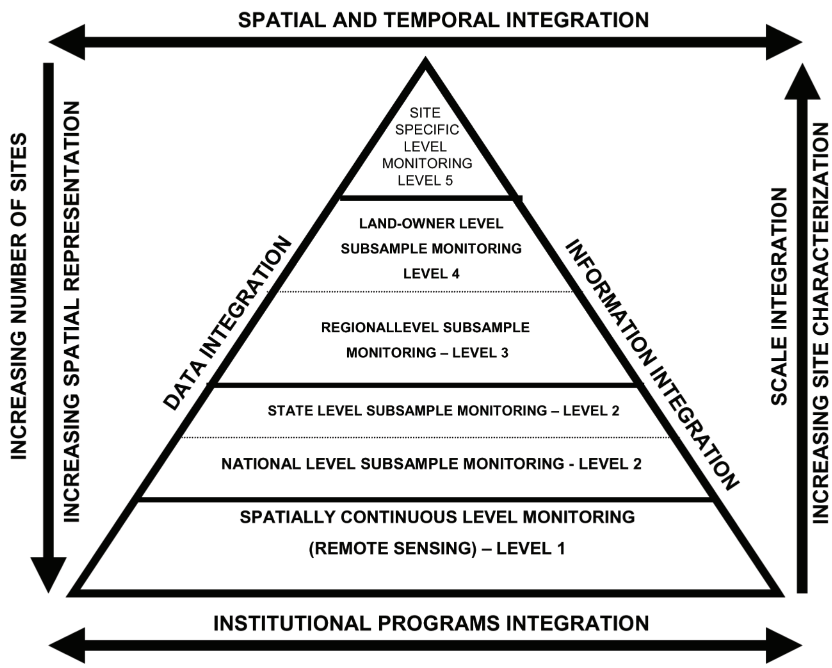

Figure 1. Conceptual model for integration of monitoring design and institutional processes.

of their inter-dependencies. One approach is to model the spatial relationship between key indicator variables. In ecosystem resource management, for example, this information can then be used to identify forest habitats that are either conducive or a deterrent to the presence of ecologically important plant and/or animal species. 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 the spatial dependency (Cliff and Ord 1981). The proposed ecosystem resource monitoring system will rely on information collected at different spatial scales of resolution and sampling intensities (fig. 1) to provide detailed information at the local level for ecosystem resource planning and management purposes.

The JCPP is an integrated inventory and monitoring program that is designed to address numerous objectives at different spatial scales. The components of the inventory are broken down in 5 components, which are referred to as 'levels', each addressing different spatial scale and information needs. This document provides a brief overview of all five levels of the inventory and a detailed description of Levels 1 and 2. Other documents will address Levels 3, 4 and 5 in more detail.

Spatially Continuous Monitoring (Level 1)

Landsat Thematic Mapper (TM) data will be used to provide a complete and uniform census of individual Environmental Accounting Units (EAU) across jurisdictional domains (i.e., private lands, federal and state lands, ejidos, communities, municipalities, regions, etc.). An EAU is a watershed used to assess ecological and human activity assessment as well as for resource decision making and planning applications This approach will provide data, or measurements collected as a series of contiguous and simultaneous measures across land tenure units. It will also provide the capability of monitoring EAU's for changes in spectral and spatial characteristics that can be applied over a range of spatial and temporal scales appropriate for addressing specific ecosystem resource issues.

Remote sensing data have been used primarily to develop static maps of vegetation cover or land use to guide management decisions in planning applications. However, a current component missing in ecosystem resource monitoring is its use for change detection and spatial modeling of ecosystem properties. When compared to traditional plot data, remotely sensed data can be quite helpful in detecting changes in the land base. Many critical indicator variables such as site productivity, presence or absence of threatened or endangered species, and biological diversity are simply not being measured with sufficient spatial coverage and frequency to allow evaluation of current and future trends. Unlike previous applications of remote sensing in monitoring ecosystem resources in which remote sensing imagery is treated as the primary variable of interest, we propose a slightly different approach. In our approach, Landsat TM data is treated as an auxiliary variable while the field data is treated as the primary variable of interest in describing specific properties of the region at any desired spatial scale (i.e., 1 to 30 m). Appropriate use of the approach can greatly increase the value of outputs from monitoring programs (Metzger 1997).

The decision to use Landsat TM data over other remote sensors (e.g., SPOT (10 m resolution), LEWIS (5 to 300 m resolution), Space Imaging (1 to 4 m resolution), Earthwatch (3 to 15 m resolution), Earthwatch Quickbird (1 to 4 m resolution), OrbView 2 (1100 m resolution), OrbView-3 (1 to 8 m resolution), SPOT 4 Vegetation Sensor (1000 m resolution), MODIS (250 to 1000 m resolution)) is based on the following reasons. The cost of acquiring data from these high resolutions sensors is significantly higher than what it currently costs to acquire Landsat TM data. Users of these sensors have shown that in terms of modeling large-scale spatial variability (Metzger 1997), Landsat TM data is superior to currently available SPOT data even though the latter has a finer resolution, but is also more expensive. Also the resolution associated with some of these new sensors are not the same across different spectral bands making it difficult to integrate with field data with the same level of precision. For this approach to work, it is important to have a wide range of spectral data at a consistent resolution. An additional benefit is that the spatial modeler involved in this work has considerable experience in working with TM. Therefore, Landsat TM data will be used as the remote sensing sensor of choice until some of these other sensors have been thoroughly evaluated.

In addition to the Landsat 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 30 m spatial resolution so that an average elevation, slope and aspect can be assigned to each pixel

A separate ground sampling effort was also undertaken to collect data that will be used to construct a classified map of the cover types of Mexico. These ground data comprise approximately 2600 purposively chosen points. At each point, the GPS coordinates and the cover type are recorded. These data were used to determine the spectral signature of the TM data and landscape characteristics associated with each cover type. This will allow us to generate a preliminary land cover map.

Design-based Inference for Inventory and Monitoring (Level 2)

The development of the sampling and plot designs is complicated by the variety of indicators to be assessed, 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 level objectives for ecosystem monitoring and large-scale assessments, a traditional grid-based sampling design is used. The plot locations are treated as a simple random sample of the landscape and post-stratification is employed as a variance reduction technique. This will be referred to as the Tier 1 inventory.

The next level of the survey is designed to meet State level inventory objectives. This is referred to as the Tier 2 inventory. The only feature that distinguishes the Tier 1 and 2 inventories is the increased sampling intensity in Tier 2.

Model-based Inference for Monitoring and Inventory (Levels 3-4)

For the remaining objectives involving estimation at local scales (levels 3 – 4), the plot data from the Tier 1 and 2 inventories will be enhanced with additional ground plots to provide information needed to develop geostatistical models. These models will be used to describe the location and distribution of various resource attributes and to estimate key attributes at all locations within the sampled population.

To ensure that the spatial variability in the study area is captured, additional sample plots will be located to enhance the spatial models for use in local management applications. The location of these plots can be based on a pre-stratification scheme that locates plots in areas with the highest errors in the spatial models or where the spectrally-derived strata may not capture all of the unique spatial features on the landscape. The primary sampling unit (psu) is a 30 m x 30 m square.

Analytical Issues

The analysis of the Tier 1 and 2 inventory data is straightforward once the data have been edited and stored in an electronic file. An analysis program for Level 1 and 2 data incorporating the estimators discussed later in this paper is now being developed in the computer language R. Analytical techniques for Levels 3 and 4 of the inventory is more complicated, but software and training is available.

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

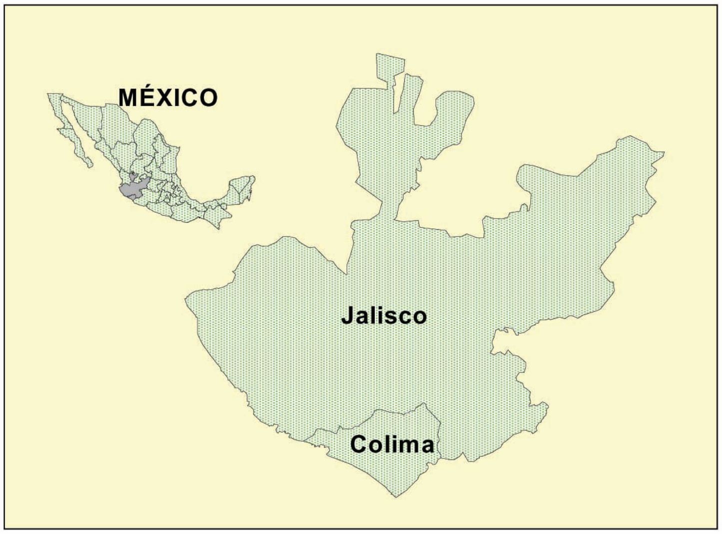

Figure 2. Geographic Location of Pilot Study Area in Mexico.

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. 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 (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.

The Design and Estimation for the Tier 1 and 2 Inventories

There is no "perfect" solution for the sample design and estimators for a large-scale inventory. Limitations associated with remote sensing, GPS, and the amount of time required by the field crews for data collection influence the design of the survey (Williams and Eriksson 2002, Overton and Stehman 1996). The sampling strategy chosen for the JCPP is an efficient and flexible strategy.

An efficient sampling design for collecting information over large geographical areas is a systematic grid with equal spacing and a random start. It has the advantage of spreading the sample units uniformly throughout the population. However, the systematic grid design tends to be the least efficient for spatial modeling because the equal spacing between plots provides less information for modeling purposes than a design with uneven spacing. In order to avoid periodicity in the resource and to permit plots to occur at random distances for spatial modeling, the plots will be located randomly within hexagonal cells formed by a triangular grid. This grid of hexagon cells for the Tier 1 inventory will uniformly cover all of Mexico with the size of each hexagon being approximately 900 km^2. To facilitate across border assessments, the grid for Mexico is an extension of the one developed for the FIA inventory in the U.S. The triangular grid can be used to intensify the Tier 1 sample for meeting State level inventory needs. This State level intensification is the Tier 2 component of the inventory. For the pilot study, the Tier 2 inventory is a 36-fold intensification of the Tier 1 grid.

A ground plot is located within each hexagon cell in accordance with a uniform distribution. The field crews will locate all of the ground plots at the UTM coordinates at the center of the TM pixel given to them – accurate location of the points is important both for spatial modeling as well as future relocation of these permanent plots. The grid density will be set first to meet state needs. It will then be intensified to meet other multiple scale needs, based on available funding. The intensity can also be altered in each of the main hydrological regions or municipalities. Plot locations will be kept secret. The opportunity to intensify also exists for local areas within land tenure units, EAUs, or administrative units, as funding is made available.

The other major advantage of using Landsat TM data is that the pixels can be used to construct an area frame comprised of equal area sample units. The advantage of this area frame approach (Husch and others 1982, Eriksson 1995) is that it is a true equal probability sample regardless of the type of stratification used, with the inclusion probabilities of the sample units (e.g., trees, CWD, and other vegetation) along the boundaries of different subpopulations not needing adjustment. This is especially advantageous when the sample data will be post-stratified, because every division (stratification) of the population creates new subpopulations with new boundaries. These new boundaries would require adjustments to the inclusion probabilities of all population elements (trees) along the new boundaries and it is unlikely the data needed for these could or would be collected (see Williams and Eriksson 2002 for a discussion of this topic).

For estimation, the plots will be post-stratified using the classified Level 1 Landsat TM data. At a minimum, the entire population and each EAU will be stratified into areas that are predominantly forest or non-forest. The plots will also be assigned to these strata, and then the appropriate results for post-stratified random sampling will be applied (Cochran 1977, Schreuder and others 1993). The estimators are given below. For large areas, the sample size in each stratum will be approximately proportional to the size of each stratum.

Description of the Ground Sample Units

The primary sampling units (psu) are a 90 m x 90 m (fig. 3) and a 30 m x 30 m (fig. 4) and are constructed with 3 x 3 blocks of Landsat TM pixels. Both psus will be sub-sampled by five 10m x 10m 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. PSU locations will be verified using a Global Positioning System (GPS) with an estimated accuracy of within 3m.

In order to maintain an equal probability sample design and because the locational and registration errors in remote sensing and GIS technologies prevent an accurate subdivision of a pixel, a psu is considered either totally within a population or totally out of a population (or stratum). Thus the population boundaries are redefined

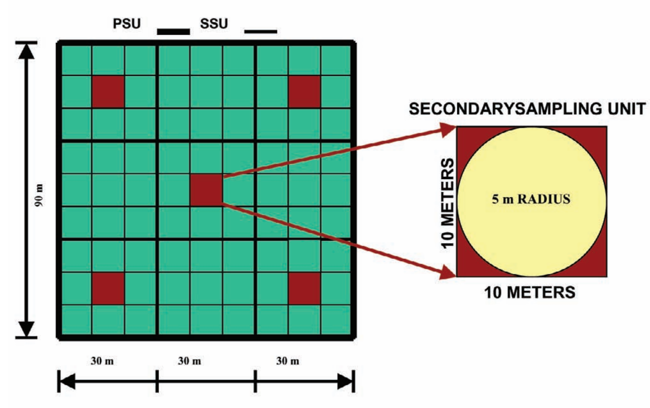

Figure 3. Plot Layout for the 90 m x 90 m Primary (PSU) and Secondary Sampling (SSU) Units.

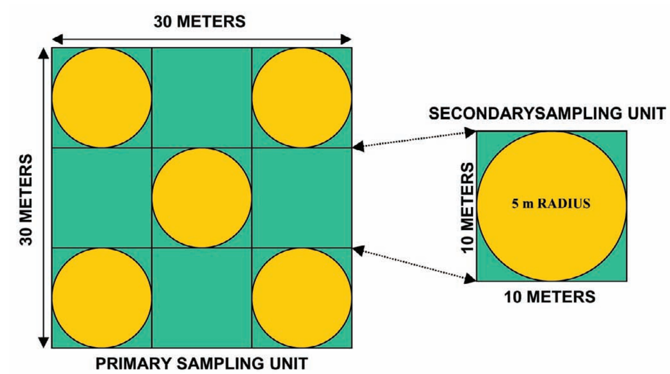

Figure 4. Plot Layout for the 30 m x 30 m Primary (PSU) and Secondary Sampling (SSU) Units.

to follow the area frame population boundaries created by the edges of the psus. A psu is ``in" when the center of the psu falls inside the population or stratum boundary. This creates a jagged stair stepped edge along the population boundaries, so a small portion of the actual landmass of an area segment will lie within the boundary of another area segment. The size of the difference should be minimal, generally less than 0.05 percent for the water sheds in the Green Belt.

Because these will be permanent plots, the psu center will be monumented on the ground. A sample of five of the 81 ssus will be selected for measurement, using a circular plot of 5 meter radius. This plot will be referred to as the 5 m plot. One 5m plot will be located at the psu center. The other four 5 m plots will be located at fixed angle directions and constant distance with respect to the psu center (fig. 3).

Several kinds of subplots will be located within each of the 5 m plots (fig. 5) 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 indicator measurement manuals. Saplings (2.5 cm < DBH

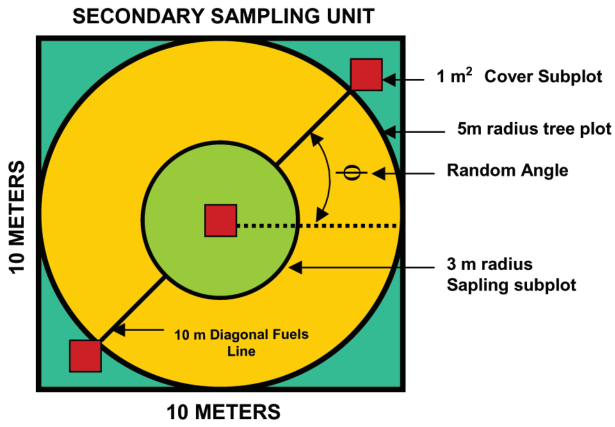

Figure 5. Layout of Secondary Sampling Unit.

< 12.5 cm) will be measured on a circular plot (3m radius) co-located at the center of each tree subplot. The term '3 m plot' will be used to denote this plot. Within each of the 5 m plots will be 3 square plots, each measuring 1 m x 1m. These will be referred to as 1 m quadrat. The first 1 m quadrat will be located at the center of the 5 m plot. The remaining two are located 6m from the center plot, on a diagonal of the 5 m plot (fig 4). Seedlings (height > 30 cm and DBH < 2.5 cm) will be sampled on three 1m quadrats. In addition to counting seedlings, the percent cover of herbaceous plants, shrubs, and tree species < 30 cm tall will be recorded.

To estimate fuel loadings, a 10 m transect with a random orientation will be established. This will be referred to as the 10 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 10 m transect will be counted and their cross-sectional areas measured by genus. Medium size woody materials (2.5 to 7.4 cm in diameter) that are intersecting the first 5 m of the transect will be counted. In addition, the first, center, and last 1m of the diagonal transect will be used to count fine woody materials (0.01-2.4 cm in diameter). 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.

Soil attributes will be observed on each 5 m radius plot. Any destructive soil samples will be collected on the west side (270 degree Azimuth) of the primary sampling unit 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.

Selection of the Ground Plot Locations and the Assignment of Measurement Periods

Two procedures were used to locate the sample plots in the field. In the first procedure, plot locations were arranged on a triangular systematic grid with a random start, with plot locations falling at the intersection of a triangular grid with approximately 27 km point spacing. This grid allows for a four-year time span between remeasurement of each ground plot on an interpenetrating rotating panel design. Thus, the entire grid and each individual panel is evenly distributed within the area, with a sampling intensity of approximately one ground plot for every 6800 hectares (167,000 acres) for the FHM program.

The JCPP will adopt the same sampling intensity as FIA with one ground plot for every 2400 hectares and maintain across border consistency by meshing the hexagon grid covering Mexico with the existing grid covering the U.S. However, with the increased focus on spatial modeling, the placement of ground plots within each hexagon will not be on the systematic grid. Instead, the plot location within each hexagon will be completely random. This is done by generating a random (x, y)-coordinate that falls within each hexagon. The justification for this practice is that the estimation of correlograms and variograms is difficult when ground plot locations are equally spaced, especially when the scale of spatial patterns is expected to be much finer than the spacing of the grid. For the purpose of estimating population means and totals, this sample is still treated as a simple random sample of the land base, as is common with most environmental surveys. The justification for this practice is summarized by Ripley (1981. pp. 19-21).

The purpose of the hexagon grid is two-fold. In the first place it is used to establish a pseudo-systematic sample (with a random start) of ground plots across the JCPP study area. Secondly it is used to allocate ground plots to each yearly panel so that a good spatial representation of all conditions is represented in each panel. The hexagon grid has no other use in the estimation process. The resulting sample can be best described as a hybrid of centric- and nonaligned-systematic sampling (Ripley 1981, pp.19-21).

The assignment of each plot to a panel is loosely based on the original design used by Forest Health Monitoring (FHM) and the current FIA design (fig. 6). The EMAP grid was developed for interpenetrating rotating panels of 3,4,7,9, and 11 years or any multiple. The FIA design is designed for sampling periods of 5, 7 or 10 years. The state of Jalisco tentatively plans a five year period of remeasurement.

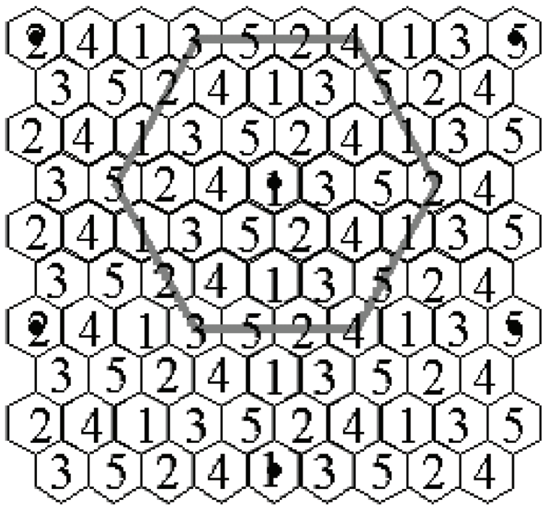

Figure 6. Numbering of each hexagon for the panel assignment.

In the second approach, A two-stage sampling design is employed. In the first stage, the pilot study area was stratified by vegetation type (e.g., temperate forest, tropical forests, grasslands, mesquite forests, agricultural lands, etc.). Strata were defined using a detailed vegetation map of the pilot study area developed using an independent set of point data. Each stratum has a known size and is used as weights to obtain area-wide estimates. The number of sample plots within each stratum were allocated proportional to the size of the stratum and the variability within stratum. In the second phase, Landsat TM data was 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 were randomly located within each spectral class. This ensures that the sample plots will cover the spectral variability associated with the Landsat TM image, which is essential for spatially interpolating the sample data.

Estimators for the TIER 1 and 2 Inventories

The estimators for the Tier 1 and 2 inventories are simple. The purpose of this section is to present the estimators in detail. This allows the readers to progress only as far into the notation as necessary. Further justification for this approach and simulation results can be found in Williams and Patterson (2003).

One topic that will not be discussed in detail is the method used to assign stratum values to each 90 m X90 m psu. Instead, it is assumed that an effective stratification algorithm will be developed for various regions that will combine the individual pixel attributes, cover map information, and other auxiliary information contained within the area frame. The algorithm may reflect regional forest and landscape characteristics and may depend on the type of analysis being preformed. However, at this point in time there is too little known about available information to define a single technique for the study area.

Slightly different approaches are used to derive the estimators of various cover types versus the estimators associated with vegetation. Thus, these topics will be treated separately.

Estimators for total area of land by cover type or condition classes

To estimate the total area of forest or other condition class, the five 5 m plots of the psu are viewed as the second stage of a two-stage cluster sample. Multistage cluster plot sampling results are used to divide each psu into M=81 secondary sampling units, from which a sample of size m=5 is chosen (fig. 3).

While the estimators derived from the smaller plots sizes (e.g., the 3 m and 1 m plots) can be combined with the one derived from the 5 m plots, the nested structure of the these plots will produce highly correlated estimates. This correlation between the estimators of forest area is such that essentially no reduction in variance exists when estimating forest areas on anything other than the largest plot used (Williams and Patterson 2002). Thus, the area of forest or condition class is estimated using only the information from the largest plot size implemented.

Regardless of the plot size or the method of assignment to a stratum for each psu, the goal is to estimate the proportion of specific condition classes in each stratum. To accomplish this, the n ground plots are treated as a post-stratified sample with a random sample size of n_h in each stratum. For each ground plot, the proportion of the psu covered by forest is either estimated by a subsample of ssus or by measuring the proportion of the psu covering the condition class of interest. Thus, given that the assumptions are reasonable, the estimator of the area of forest is a post-stratified, two-stage cluster sample (Cochran 1977), given by

$$\hat{A}_F = A\hat{p}_F = A\sum_{h=1}^{H}\frac{N_h}{N}\left[\sum_{i=1}^{n_h}\sum_{j=1}^{m}\frac{p_{hij}^F}{mn_h}\right] \qquad (1)$$

where p^F_{hij} is the proportion of forest in the j^{th} 5m subplot of the i^{th} PSU within the h^{th} stratum.

Since the variance of $\hat{A}_F$ is equal to $A^2 Var(\hat{p}_F)$ suffices to determine $Var[\hat{p}_F]$. The derivation of the variance combines results from the variance of a simple random post-stratified sample and the variance of a stratified two-stage sample (Cochran (1977), §5A.9 and §10.9). Conditioning on the n_h's, $Var\left[\hat{p}_F\right]=Var[E[\hat{p}_F|n_h]]+E[Var[\hat{p}_F|n_h]]$. Since for fixed stratum sample sizes, $\hat{p}_F$ is an unbiased estimator of p_F, the first term is zero. For fixed n_h,

$$Var\left(\hat{p}_F\right)=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left[\left(\frac{1}{n_h}-\frac{1}{N_h}\right)S_{1h}^2+\frac{1-m/M}{mn_h}S_{2h}^2\right]$$
$$=\sum_{h=1}^{H}\frac{1}{n_h}\left\{\left(\frac{N_h}{N}\right)^2\left[S_{1h}^2-\left(\frac{1}{m}-\frac{1}{M}\right)S_{2h}^2\right]\right\}-\frac{N_h}{N^2}S_{1h}^2 \tag{2}$$

(Cochran 1977). The next task is to find the expectation of $1/n_h$. Assuming the sample size is sufficiently large to preclude n_h=0 then to order n^{-2},

$$E\left[\frac{1}{n_h}\right]=\frac{1}{n(N_h/N}+\frac{1-N_h/N}{n^2(N_h/N)^2}$$

Using results derived in Cochran (1977) it can be shown that an unbiased estimator for the above approximation of $Var\left[\hat{p}_F\right]$ is given by

$$\text{var}\left[\hat{p}_F\right]=\sum_{h=1}^{H}\frac{1}{n}\frac{N_h}{N}\left[\left(1-\frac{n}{N}\right)s_{1h}^2+\frac{n}{N}\left(\frac{1}{m}-\frac{1}{M}\right)s_{2h}^2\right]+\left(1-\frac{N_h}{N}\right)\frac{1}{n^2}s_{1h}^2 \tag{3}$$

where s_{1h}^2 and s_{2h}^2 are the between cluster and within cluster sample variances for stratum h (Cochran 1977). Due to the large number of psus (N) in comparison to sample size (n) the second term within the square brackets is small relative to the first term for both the Tier 1 and 2 inventories.

Vegetation or tree estimators

In this section, the estimator that utilizes the tree information from the entire ground plot is given. Unlike the estimator for forest area, where the assumption is made that the psu is divided into M secondary sampling units, a different approach is taken. This is because the tree and other vegetation attributes are gathered from transect samples and across the 1 m, 3 m, and 5 m plots, which would require different M values and the calculation of the correlations between the estimators derived from each of the plot sizes. Thus, the proposed estimators of tree and other vegetation attributes within each psu are expressed as a triareal sampling design comprised of three different sized subplots (Husch and others 1982). The estimator for the total within each psu is expressed using a Horvitz-Thompson estimator. Two assumptions are necessary. The first is that systematic nature of the three subplots within each psu is equivalent to having the points randomly located within the psu. In the literature, this assumption has been justified by assuming that the location of trees in the forest has a random pattern within the psu (de Vries 1986). When this assumption is not realistic the variance estimator tends to over-estimate the variance. The second assumption is that none of the points falls close enough to the boundary of the psu so no adjustments to the inclusion probabilities of trees are required. This allows a constant inclusion probability to be assigned to all trees tallied on each of the three types of subplots. This assumption is necessary because it is impractical to obtain the actual probabilities of selection for border trees.

The n ground plots are treated as a post-stratified sample of psus, with a random sample size n_h in each stratum. Within each psu, a tree-centered circle, or 1 m quadrat in the case of seedlings, (Husch and others 1982) is placed about each tree, the size of the circle is determined by the diameter at breast height limits, which in turn determine the inclusion probabilities. Then a tree is included in a sample drawn by a single point if the point lies within the tree-centered circle. Let a^{1m}, a^{3m}, and a^{5m} denote the area of the 1m quadrat-, 3 m-, and 5m radius plots respectively and let a_{psu}=8100m^2 denote the area of the psu. The probability that a tree is included in the sample drawn by a single point within the psu depends on the size of the tree and is defined as:

$$\pi = \begin{matrix} a^{1m}/a_{PSU} \\ a^{3m}/a_{PSU} \\ a^{5m}/a_{PSU}\,, \end{matrix}$$

where a^{1m} is the area of the 1 m quadrat, a^{3m} is the area of the 3 m radius plot, and a^{5m} is the area of the 5 m radius plot.

The estimator given here is for the 3 m and 5 m radius plots, where m=5 plots are sampled within each psu. The estimator for the 1 m quadrats is defined analogously with the exception being the change in sample size to m=15. Estimation for the psu is based on the selection of 5 points within the psu and measurements of the subplots of 3 m and 5 m radius respectively centered at these points. Let T_{hij} be the number of trees tallied at the j^{th} point of psu i in stratum h. Then the estimator of the total of the tree attributes within the psu for the j^{th} point of psu i, in stratum h, is given by

$\hat{Y}_{hij}=\sum_{t=1}^{T_{hij}}\frac{Y_{hijt}}{\pi_{hijt}}$ with and respectively the value of interest and the probability of selecting tree t at point j, psu i, in stratum h.

with the estimator for the total of the PSU derived from the m=5 points being

$$\tilde{\hat{Y}}_{hi} = \frac{1}{m}\sum_{j=1}^{m}\hat{Y}_{hij}.$$

An unbiased estimator of $Var\left[\tilde{Y}_{hi}\right]$ is given by

$$var\left[\tilde{Y}_{hi}\right] = (m(m-1))^{-1}\sum_{j=1}^{m}(\hat{Y}_{hij} - \tilde{Y}_{hi})^2.$$

This estimator is derived by summing across the psus and strata. Then the estimator is

$$\hat{Y} = \sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{i=1}^{n_h}\tilde{Y}_{hi} \tag{4}$$

An approximation of the variance of the estimator is given by

$$Var\left[\hat{Y}\right] \approx \sum_{h=1}^{H} N^2\left[\frac{N_h}{N}\frac{1}{n} + (1-\frac{N_h}{N})\frac{1}{n^2}\right]\bar{V}\left[\tilde{Y}_{hi}\right]$$
$$+N^2\left[\frac{N_h}{N}(\frac{1}{n}-\frac{1}{N})(1\text{-}\frac{N_h}{N})\ \frac{1}{n^2}\right]S_h^2 \qquad [5]$$

where $\bar{V}\left[\hat{Y}_{hi}\right] = \sum_{i=1}^{N_h} Var\left[\tilde{Y}_{hi}\right] / N_h$ and

$$S_h^2 = \sum_{i=1}^{N_h}(Y_{hi} - \bar{Y}_h)^2 / (N_h - 1).$$

An unbiased estimate of the above approximation of $Var[\hat{Y}]$ is given by

$$\text{var}[\hat{Y}] = \sum_{h=1}^{H} N^2\left[\frac{N_h}{N}\frac{1}{n} + (1-\frac{N_h}{N})\frac{1}{n^2}\right]\text{var}[\tilde{Y}_{hi} + N^2$$
$$\left[\frac{N_h}{N}(\frac{1}{n}-\frac{1}{N}) + (1-\frac{N_h}{N})\frac{1}{n^2}\right]\tilde{s}_h^2 \qquad ,[6]$$

where $\text{var}[\hat{Y}_{hi}] = \sum_{i=1}^{n_h}\text{var}\left[\tilde{Y}_{hi}\right] / n_h$ and

$$\tilde{s}_h^2 = \sum_{i=1}^{n_h}(\tilde{Y}_{hi} - \tilde{\tilde{Y}}_h)^2 / (n_h - 1).$$

The above estimation assumes that there are no plots or subplots that cannot be measured. A nuisance that arises with almost any type of field sampling with plots is that parts of the psu or subplots will fall outside the population of interest or are inaccessible to sampling either because of difficulty of terrain or to reach part of the sample in a practical manner we need to trespass on private land. Not all landowners will allow the crews on or across their land even to measure other lands. How to treat inaccessible plots or subplots is described in the appendix. The following estimator deals with the situation of missing plots or subplots. Post-stratification, as considered above, is not shown in the below formulation since it is a straightforward extension of what is given by simply applying the estimator shown to each stratum.

The most appropriate estimator to be used in estimating population totals with missing subplots is:

$$\hat{Y}_s = A\left[\sum_{i=1}^{n}\sum_{j=1}^{n_s}\sum_{k=1}^{n_{ss}}\hat{A}_{ijk}Y_{ijk}\right] / \left[\sum_{i=1}^{n}\sum_{j=1}^{n_s}\sum_{k=1}^{n_{ss}}\hat{A}_{ijk}\right] \tag{7}$$

and

$$\hat{Y} = A\left[\sum_{i=1}^{n}\sum_{j=1}^{n_s}\hat{A}_{ij}Y_{ij}\right] / \left[\sum_{i=1}^{n}\sum_{j=1}^{n_s}\hat{A}_{ij}\right] \tag{8}$$

where $\hat{A}_{ij}$ and $\hat{Y}_{ij}$ are the estimated sampled area and value of interest respectively in subplot j of plot i, n is the number of plots in the sample, n_s is the number of subplots in the sample for plot i where $\hat{Y}_s$ is the estimator for subplot size s and subplot ss ,and m_i is the number of subplots in the sample for plot i (Max and others 1996). We recommend the use of bootstrap variance estimators here as discussed in Schreuder and others (1993).

Line intersect sampling estimators

The line intersect sample (LIS) is drawn using m=5 randomly oriented lines with the centers covering each of the 5 m plots. The estimator employed is the LIS estimator as described in Kaiser (1983) and Gregoire (1998), with the key difference being that the sample locations are a systematic sample subsample of the psu rather than five random locations. There should be little if any detectable bias in the resulting estimators, but the variance estimator should over-estimate the true sample variance. The LIS estimator for the amount of woody debris on the psu estimated from the m=5 L=10m lines with random orientation is given by

$$Y_{hi} = \frac{1}{m}\sum_{j=1}^{m}Y_{hij} = \frac{1}{m}\sum_{j=1}^{m}\frac{a_{PSU}\pi}{2L}\sum_{k=1}^{P_j}\frac{g_{hijk}}{\cos\delta_{hijk}}$$

where g_{hijk} is the cross-sectional area of the log measured perpendicular to the long axis of the log (m^3) for the k^{th} piece of debris on the j^{th} transect, P_j is the number of pieces of debris intersected by the j^{th} line, and δ_{hijk} is the slope of the log. In relatively flat terrain, δ_{hijk} , may be set to zero. However, a bias in excess of 10 percent can occur whenever the slope exceeds 25 percent. The derivation and further details for this estimator can be found in Kaiser (1983 example 2c).

The variance estimator given in eq. [6] applies.

Other estimators

While area and vegetation totals are the primary forest attributes that are estimated, other estimators are also of interest. The first class are mean per tree estimators (average number of conks or disease agents per trees). The form of this estimator is $\bar{\hat{Y}} = \hat{Y} / \hat{N}$, where $\hat{N}$ is the estimator of the number of trees (eq. [4] with Y_{hijt}=1). Another class of estimators is used to assess change. These have the form where and are the estimators at times 2 and 1, respectively.

Variance estimators for the mean per tree and difference estimators are

$$\mathrm{var}(\bar{\hat{Y}}) \approx \frac{1}{\hat{N}}\left[\mathrm{var}(\hat{Y}) + \frac{\hat{Y}}{\hat{N}}\mathrm{var}(\hat{N}) - 2\frac{\hat{Y}^2}{\hat{N}^2}\mathrm{cov}(\hat{Y},\hat{N})\right] \quad (10)$$

and

$$\mathrm{var}(\Delta\hat{Y}) = \mathrm{var}(\hat{Y}_2) + \mathrm{var}(\hat{Y}_1) - 2\,\mathrm{cov}(\hat{Y}_2,\hat{Y}_1) \quad (11)$$

Alternative estimators

At this point in time there are a number of alternative estimators that may offer advantages over the estimators given above.

In various circumstances we may have complete knowledge on a covariate associated with the variable of interest for which we know all the values in the population or we can get those with relative ease. Usually this information is combined with the information on the variable of interest measured on a sub-sample of the units in the population. Denoting by y= variable of interest and x= covariate, numerous estimators are possible. We focus only on the generalized regression and the ratio-of means estimators, the others are generally not desirable.

A very general, efficient estimator is the generalized regression estimator developed by C.E. Sarndal:

$$\hat{Y}_{gr} = \sum_{i=1}^{n} y_1 / \pi_i + a_{gr}(N - \sum_{i=1}^{n} 1/\pi_i) + b_{gr}(X - \sum_{i=1}^{n} x_i/\pi_i) = \sum_{i=1}^{N}\hat{y}_i + \sum_{i=1}^{n} e_i/\pi_i \quad (12)$$

where

$\hat{y}_i a_{gr} + b_{gr} x_i, e_i = y_i - \hat{y}_i$,

$\hat{y}_i = a_{gr} + b_{gr} x_i$, $e_i = y_i - \hat{y}_i$,

$$a_{gr} = \{\sum_{i=1}^{n} 1/(\pi_i v_i) - b_{gr}\sum_{i=1}^{n} x_i/(\pi_i v_{i)}\} / \sum_{i=1}^{n} 1/(\pi_i v_i)$$

$$b_{gr} = \{\sum_{i=1}^{n} 1/(\pi_i v_i)\sum_{i=1}^{n} x_i y_i/(\pi_i v_i) - \sum_{i=1}^{n} y_i / (v_i\pi_i)\sum_{i=1}^{n} x_i/(v_i\pi_i)\} / \{\sum_{i=1}^{n} 1/(\pi_i v_i)\sum_{i=1}^{n} x_i^2/(v_i\pi_i) - [\sum_{i=1}^{n} x_i/(v_i\pi_i)]^2\}$$

with variance:

$$V(\hat{Y}_{gr}) = (1/2)\sum_{i \neq j}^{N}(\pi_i\pi_j - \pi_{ij})(e_i/\pi_i - e_j/\pi_j)^2 \quad (13)$$

with 2 possible variance estimators given in Schreuder and others (1993) but we recommend using a bootstrap variance estimator here. This is the estimator to be used when possible.

Another estimator that is not a special case of the above is called the generalized ratio of means estimator:

$$\hat{Y}_{rm} = (\sum_{i=1}^{n} y_i/\pi_i / \sum_{i=1}^{n} x_i/\pi_i)X = (\hat{Y}_{HT}/\hat{X}_{HT})X \quad (14)$$

with approximate variance

$$V(\hat{Y}_{rm}) = V(\hat{Y}_{HT}) - 2R\,Cov(\hat{Y}_{HT},\hat{Y}_{HT}) + R^2 V(\hat{X}_{HT}) \cdot \quad (15)$$

There is a good discussion in Schreuder and others (1993) but again, we recommend using a bootstrap variance estimator. Both the generalized regression and the ratio-of-means estimator are asymptotically unbiased.

There will be opportunities to use $\hat{Y}_{gr}$ in the pilot study. With the heavy emphasis on remote sensing platforms we will examine what we can do with that in relation to the ground sampling to obtain improved estimates for some parameters.

Variance estimators can also be derived that utilize re-sampling techniques, with the most common being bootstrap and jackknife estimators. Bootstrap samples are generated by repeating the sampling design with replacement at the same sample sizes as the actual sample design and then generating estimates for each bootstrap sample. The variance between those estimates is the variance estimate for the estimator of interest. Usually 2000 bootstrap samples are generated. This is applicable for all estimators above.

Summary and Recommendations

The statistical strategies presented in this document significantly benefit from the experience that exists in Canada and the United States. In light of how rapid technology has been evolving, these strategies incorporate new design elements so that remote sensing and land measurement protocols processes can be unified for purposes of making reliable statistical inferences. Consistency of data collection protocols at national and state levels is essential for this strategy to work properly and generate estimates of known confidence. Its implementation on the Pilot Study Area produced valuable information for improving design elements and data analysis approaches. How the two strategies can be integrated even better and to determine what would be the best plot to use to constitute some of the questions to be addressed after the full analyses of the data from this project.

References

Cliff, A.D. and Ord, J.K. 1981. Spatial processes: models and applications. Pion, London. 266 p.

Cochran, W.G. 1977. Sampling Techniques. 3rd ed. John Wiley and Sons, New York. 428p.

De Vries, P. 1986. Sampling Theory for Forest Inventory. John Wiley & Sons, New York.

Efron, B. and Tibshirani, R.J. 1993. An introduction to the bootstrap. Chapman and Hall, NY. 436 p.

Eriksson, M. 1995. Design-based approaches to horizontal point sampling. For. Sci. 41:890-907.

Gregoire, T.G. and Scott, C.T. 1990. Sampling at the stand boundary: A comparison of the statistical performance among eight methods. In: Proc. XIX World Forestry Congress IUFRO. Montreal, Canada. Aug. 5-11, 1990. Publ. No. FWS-3-90, Virginia Polytechnic Inst and Univ, Blacksburg, VA, pp 78-85.

Husch, B., C.I. Miller and T.W. Beers. 1982. Forest Mensuration 3rd ed. J. Wiley and Sons, New York. 402p.

Johnson, D. S., M. S. Williams, and R. L. Czaplewski. 2002. A test of some estimators for rolling samples using Forest Inventory and Analysis data. For. Sci. 49:50-63.

Kaiser, L., 1983. Unbiased estimation in line-intersect sampling. Biometrics 39:965-976.

LePage, R. and Billard, L. eds.1992. Exploring the limits of bootstrap. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics Section. NY. 426 p.

Max, T.A., Schreuder, H.T., Hazard, J.W., Oswald, D.D., Teply, J. and J. Alegria. 1996. The Pacific Northwest Region Vegetation and Monitoring System. USDA Forest Service PNW Res Paper RP-493. 22p

Metzger, K. 1997. Modeling forest stand structure to a ten meter resolution using Landsat TM data. M.S. Thesis, Colo. State Univ. Ft. Collins, CO. 123 p.

Overton, W.S. and Stehman, S.V. 1996. Desirable design characteristics for long-term monitoring of ecological variables. Env Ecol Stat 3:349=361.

Schreuder, H. T., T. G. Gregoire, and G. B. Wood. 1993. Sampling Methods for Multiresource Forest Inventory. John Wiley and Sons, NY 446 p.

Schreuder, H. T, Ernst, R. and H. Ramirez-Maldonado. 2004. Statistical techniques for sampling and monitoring natural resources. USDA Forest Service RMRS Gen Tech. Rep RMRS-GTR-126. 111 p.

Williams, M. S., and Eriksson, M. 2002. Comparing the two paradigms for fixed-area sampling in large-scale inventories. For. Ecol. Manage.168: 135-148.

Williams, M. S., and P. L. Patterson 2003. Comparison of area estimation techniques for established surveys using remotely sensed and ground data. For. Sci. 49: 392-401.