Calculation of Upper Confidence Bounds on Not-sampled Vegetation Types using a Systematic Grid Sample: An Application to Map Unit definition for Existing Vegetation Maps

Paul L. Patterson[1], Mark Finco[2]

Abstract: *This paper explores the information FIA data can produce regarding forest types that were not sampled and develops the equations necessary to define the upper confidence bounds on not-sampled forest types. The problem is reduced to a Bernoulli variable. This simplification allows the upper confidence bounds to be calculated based on Cochran (1977). Examples are provided that demonstrate how the resultant equations are relevant to creating mid-level vegetation maps by assisting in the development of statistically defensible map units.*

Keywords: Map unit, remote sensing, FIA, dominance type, grid sampling, confidence bounds, mid-level map

Background

Mid-level vegetation maps are important information sources for forest planning, habitat management, and many other resource management activities (Brohman and Bryant 2005). One of the first and most important steps in creating any map is defining the classes, or map units, which will be mapped.

In the Forest Service, several of the Regions have adopted similar map unit design processes that integrate regionally defined dominance type definitions with Forest Inventory & Analysis (FIA) data. Very generally, the first step of the process estimates dominance type composition over an area of interest using the Regional Dominance Type Classification (RDTC).

The second step of the process aggregates the RDTC classes into map units. This is generally done by setting a minimum abundance criterion (e.g., must comprise at least 3% of the area being mapped) for any map unit and hierarchically aggregating (i.e., create ecologically rational groups) the RDTC classes into map units, so that the map units meet the specified minimum abundance criterion. Some Forest Service Regions take the process a step further

[1] Mathematical Statistician, Interior West FIA, Ft. Collins, CO
[2] Senior Analyst, Remote Sensing Applications Center, Salt Lake City, UT

In: McWilliams, Will; Moisen, Gretchen; Czaplewski, Ray, comps. 2009. 2008 Forest Inventory and Analysis (FIA) Symposium; October 21-23, 2008: Park City, UT. Proc. RMRS-P-56CD. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 1 CD.

by considering “mappability” or “separability” of the map units in the context of the imagery and other geospatial data being used in the mapping process.

Problem Statement

FIA data is a sample of forest characteristics, having approximately one sample per 6000 acres of forest land. Traditionally, FIA data are used to describe forest composition based on the dominance types that were sampled. Because FIA is a sample, however, minor components of the landscape (in terms of abundance, not necessarily management or ecological importance) may not be represented in the FIA sample. A good example of this is riparian areas, which are ecologically important, but generally cover a very small percentage of the landscape.

This paper asks a non-traditional question of the FIA sample (or any other spatially balanced sample). Specifically, this paper explores what information FIA data can produce regarding dominance types or map units that were not sampled. The paper shows how sample data can be used to define the upper confidence bounds on not-sampled dominance types. This allows the analyst to state their confidence ($(1-\alpha)100\%$) that the proportion of a not-sampled dominance type is less than some proportion (η) of the region of interest (R). By doing this, an upper confidence bound is determined for proportion of R in not-sampled dominance types.

Statistical Derivation

We will assume the dominance type classification is based on data gathered on a plot, that the plots are a uniform shape. We are not assuming the plot design is the national FIA plot design, only that the plot design is same for all plots. Our proposed method for constructing the upper confidence bound is in the context of the finite sampling paradigm. Our population unit will be a co-located square that contains the plot. The square is only used in the construction of the statistical model; the size of the square is based on two constraints; first it is big enough to contain the square and second the square is small enough so that it can classified based on the data gathered on the plot. For example, the FIA plot characterizes and is contained in a co-located 300x300 foot square, but the FIA plot in most landscapes will not characterize a co-located square that is a quarter mile on a side. Let R_{sq} denote the square. If the region R is large enough, it is reasonable to assume we can tessellate the region R with the square R_{sq}. We further assume that the sample can be viewed as a simple random sample of the tessellation. As pointed out by Milne (1959) and Ripley (1981) the periodicity of a systematic grid, such as used by FIA, coinciding with a systematic land feature is of little concern. So the FIA sample grid can be viewed as simple random sample. The population characteristic of interest is whether each square in the tessellation is classified as being of the rare dominance types. This is a Bernoulli variable; 0 when R_{sq} is not one of the rare dominance types and 1 when R_{sq} is one of the rare dominance types.

With the above background, we now outline how an upper $(1-\alpha)100\,\%$ confidence bound can be calculated based on Cochran (1977) sections 3.4, 3.5, and 3.6.

An estimate of the proportion (P) of region (R) that is classified as being in one of the rare dominance types is given by $p = \frac{a}{n}$, where n is the number of plots in the sample and a is the number of plots in the sample that are classified as one of the rare dominance types. Since the population characteristic is a Bernoulli variable we can find the frequency distribution of the estimate p. If we assume the population size N is much larger than the sample size n, then we can use the binomial approximation for the frequency distribution of p (see Cochran section 3.4). The population size N is equal to A/A_{sq}, where A and A_{sq} are the area of R and R_{sq} respectively.

We can now calculate an upper $(1-\alpha)100\,\%$ confidence bound for the proportion of R that is classified as being in one of the rare dominance types. It is the largest p such that

$$\sum_{i=0}^{a} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \geq \alpha \qquad [1]$$

If a is greater than zero, then one solves equation [1] for p using numerical techniques; the case $a > 0$ is not of interest in this paper. If no sample plots are classified as being one of the rare dominance types, then a is equal to zero and equation [1] has a simple form which has the following solution:

$$p^0 (1-p)^n \geq \alpha \Leftrightarrow p \leq 1-\alpha^{1/n} \qquad [2]$$

If no FIA plots are classified as being in the rare types, then a $(1-\alpha)100\,\%$ upper confidence bound for proportion of rare categories in R is $1-\alpha^{1/n}$. That is we are $(1-\alpha)100\,\%$ confident that the proportion of land in R that is classified as being in one of the rare types is less than $1-\alpha^{1/n}$.

A word of caution about the sample size should be noted when using this method. To illustrate using an example, suppose the region of interest (R) is the state of Utah, and the rare types are types that occur within forested lands, it may not be appropriate to use the number of plots in the state of Utah for n. It may be more appropriate to restrict to the subpopulation of forested lands in Utah. It is okay to do so (Cochran, section 3.10), but we need the size of subpopulation N' (in this example forested lands) to be much larger than n', the number of sample elements in the subpopulation, and we need to use n' instead of n in equation [2]. Our $(1-\alpha)100\%$ upper confidence bound for proportion of rare categories in the sub-population is then $1-\alpha^{1/n'}$.

Application Example

In this example, assume we are mapping an 800,000 acre National Forest. The National Forest has approximately 200 FIA plots, of which 100 plots are in forested land. When the regional dominance type definitions are applied to the FIA data, we estimate the proportion of each sampled dominance type on the forest lands (note that the non-forest dominance types would not be relevant using FIA data). Based on the number of plots on forested lands we estimate there are 400,000 forested acres, so N' is much larger than $n' = 100$. Using the formula [2], with $n' = 100$ the following table is generated relating to the forest types that were not sampled.

Table 1: Upper estimate of rare types as a function of desired statistical confidence.

Sample Size	Confidence	Upper Confidence Bound for Percentage of Rare Types
(n')	$(1-\alpha)100$%	$(1-\alpha^{1/n'})100$%
100	98%	3.8%
100	96%	3.2%
100	95%	3.0%
100	94%	2.8%
100	92%	2.5%
100	90%	2.3%

Using this table, we can state that we are 95% confident that the FIA sample did not miss a dominance type that comprised more than 3.0 % of the forested area. Actually, we can make a slightly stronger statement that we are 95% confident that all non-sampled dominance types combined comprise less than 3.0% of the forested area.

Conversely, we could use the same formulas to determine the number of plots (at the $(1-\alpha)100\%$ confidence level) required to not miss a dominance type that is at least $p100\%$ of the forest area. Solving the equation [2] for n we get

$$n = \frac{\ln(\alpha)}{\ln(1-p)} \qquad [3]$$

For example, if the requirement is to be 95% confident that a dominance type that comprises 1% (or greater) of the forest area was not missed, then n is 298. For the fictitious National Forest discussed above, this means that the standard FIA sample would have to be intensified to include an additional 198 plots in the forested area. Figure 1 illustrates sample size as a function of the minimum percentage of the dominance type of that is not be missed at a confidence level 95%. Note the rapid increase in sample size as the percentage of the region covered by the rare type decreases from 3.0%.

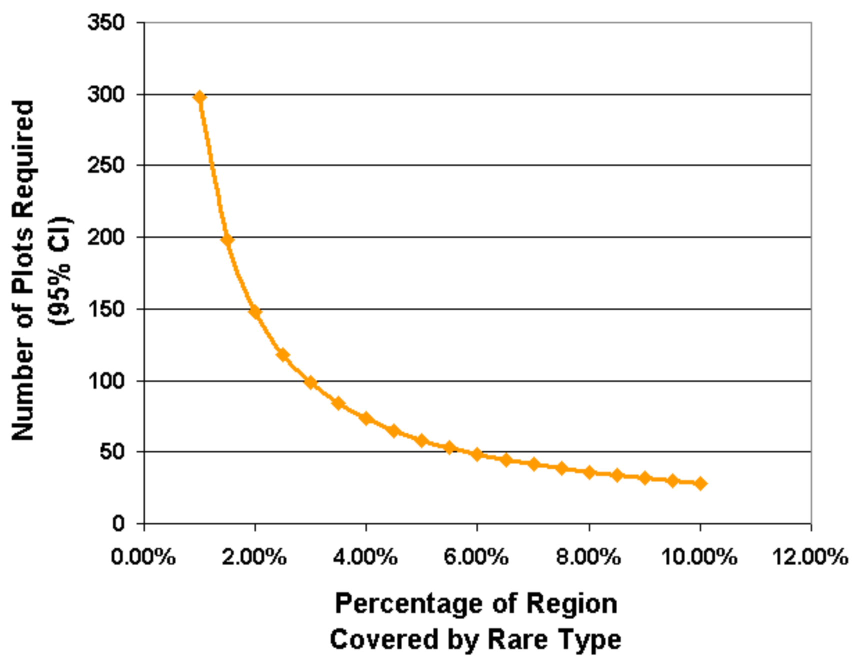

Figure 1. Sample size requirements as a function of percentage of region covered by the rare type.

Conclusions

The primary resource that land management agencies manage is existing vegetation. Vegetation impacts most resource management concerns and impacts issues ranging from timber harvest to threatened and endangered species. Therefore, it is critical for the Forest Service and other agencies to understand what vegetation exists, how it is spatially distributed on the landscape, and how much of it there is to effectively accomplish their land management responsibilities.

The idea that vegetation classification systems, existing vegetation maps, and inventory data need to be well integrated has long been recognized within some Forest Service regions and the rest are quickly adopting integrated classification, mapping and inventory programs. In all of these efforts, FIA and locally intensified inventory data have played a critical role. This paper provides insight into some previously unanswered questions that relate to the integration of existing vegetation maps and vegetation inventories.

References

Brohman, R.; Bryant, L. eds. 2005. Existing Vegetation Classification and Mapping Technical Guide. Gen. Tech. Rep. WO–67. Washington, DC: U.S. Department of Agriculture Forest Service, Ecosystem Management Coordination Staff. 305 p.

Cochran, W.G. 1977. Sampling Techniques. Ed. 3. Wiley, New York. 428 p.

Milne, A. 1959. The centric area-sample treated as a random sample. Biometrics. 15: 270-297.

Ripley, B.D. 1981. Spatial statistics. Wiley, New York. 252 p.