ISSN 1946-7664. MCFNS 2011
Vol. 3, Issue 2, pp. 98–101
Mathematical and Computational
Forestry & Natural-Resource Sciences

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: Mar. 12, 2010
Accepted: Apr. 29, 2011
Published: Aug. 28, 2011
Last Correction: Aug. 5, 2011

CALCULATION OF UPPER CONFIDENCE BOUNDS ON PROPORTION OF AREA CONTAINING NOT-SAMPLED VEGETATION TYPES: AN APPLICATION TO MAP UNIT DEFINITION FOR EXISTING VEGETATION MAPS

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

[1] *U.S. Forest Service, Rocky Mountain Research Station, Fort Collins, CO 80526 USA*

[2] *U.S. Forest Service, Remote Sensing Applications Center, Salt Lake City, UT 84119 USA*

Abstract. This paper explores the information forest inventory 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

1 Introduction

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. Typically, mid-level vegetation maps have map unit descriptions for vegetation type class and other vegetation structural variable classes, such as tree size and canopy closure. Therefore, the pixels or polygons in a mid-level map have one and only one map unit label. These labels are synoptic for the each mapped spatial unit (pixel or polygon) and are summarized at the plot level for the purposes of creating a vegetation map.

The Forest Inventory and Analysis program (FIA) of the U.S. Forest Service conducts the national forest inventory of the United States of America (USA). In the U. S. Forest Service, several of the Regions have adopted similar map unit design processes that integrate regionally defined dominance type definitions with FIA data. FIA data is a sample of forest characteristics, having approximately one sample per 2428 ha of forest land, which have traditionally been used to describe forest composition in tabular reports (Thompson et al., 2005).

Very generally, the first step of the mapping process passes the plot data through a dichotomous decision tree and summarizes it at the plot level with respect to vegetation type. These plot level summaries are used to estimate dominance type composition over an area of interest using a Regional Dominance Type Classification (RDTC). Readers familiar with the FIA mapped plot configuration (Bechtold and Patterson, 2005) should note that multiple condition plots would receive a single label using this approach.

The second step of the mapping process aggregates the RDTC classes into map units. This is generally done by setting a minimum abundance criterion for any map unit (e.g., a map unit must comprise at least 3% of the mapped area) 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 by considering the statistical separability of the map units in the context of imagery and other geospatial data being used as predictors.

Many examples of the mid-level vegetation mapping programs exist within the Forest Service. For example, Gillham et al. (2007) describe how field data, remote sensing imagery, and ancillary geospatial layers were integrated using rule-based predictive models to create a mid-level map for the Bridger-Teton National

Forest. Since this map's production, it has been useful for addressing resource and land management issues, such as a fuels reduction project in the Buffalo Valley (Buffalo Ranger District, 2009). Recently, FIA state reports, such as Thompson et al. (2005), have included maps that are built in a similar manner.

Regardless of the mapping project, because FIA data is observed for elements of a sample, 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. The standard approach to understanding these resources is to pre-stratify the area and intensify the sample within a stratum. This is an expensive means to understand more about a landscape component that you know nothing about its abundance.

This paper asks a non-traditional question of the FIA sample (or any other spatially balanced sample), which has not been addressed in the literature within the context of vegetation mapping. 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 and that the plots are uniformly shaped. We are not assuming the plot configuration is the national FIA plot configuration (Bechtold and Patterson, 2005), only that the plot configuration is uniform 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 plot 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 90-m x 90-m square, but the FIA plot in most landscapes will not characterize a co-located square that is a half a kilometer 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. The FIA sample can be treated as a simple random sample (Bechtold and Patterson, 2005, page 25). 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.

An upper confidence bound for the presence of a rare event has been addressed in other fields. For completeness, we outline one construction of an upper confidence bound using the model characteristics developed above (for details see Cochran, 1977, sections 3.4, 3.5, and 3.6).

An estimate of the proportion of region (R) that is classified as being in one of the rare dominance types is given by $p = \frac{b}{n}$, where n is the number of plots in the sample and b is the number of plots in the sample that are classified as one of the rare dominance types. 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. The population size N is equal to A/A_{sq}, where A and A_{sq} are the area of R and R_{sq} respectively. An upper $(1-\alpha)\,100\%$ confidence bound for the proportion of R that is classified as being in one of the rare dominance types is the largest p such that

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

If b is greater than zero, then one solves (1) for p using numerical techniques; the case $b > 0$ is not of interest in this paper. If no sample plots are classified as being one of the rare dominance types, then b is equal to zero and (1) has a simple form with 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 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 in the interior west of the USA, and the rare types are types that occur within forested lands. Using the number of plots in the state of Utah for n gives an upper confidence bound for the proportion of the rare type that occur in the state of Utah, where what may be of interest is an upper confidence bound for the proportion of rare types that occur in forested lands in Utah; so it may be more appropriate to restrict to the subpopulation of forested lands in Utah. It is alright 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 plots in the subpopulation. The subpopulation is defined by the map units, so the plots in the subpopulation are the subset of plots that lie in the forested map units. We need to use n' instead of n in (2); the $(1-\alpha)\,100\%$ upper confidence bound for proportion of rare categories in the sub-population is then $1-\alpha^{1/n'}$.

2 Application Example

In this example, assume we are mapping a fictional 400,000 ha National Forest and assume the National Forest has 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 200,000 forested ha, so N' is much larger than $n' = 100$. Using (2) with $n' = 100$, Table 1 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 for sample size $n' = 100$.

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

Using Table 1, 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 (2) for n' we get

$$n' = \frac{\ln(\alpha)}{\ln(1-p)} \tag{3}$$

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%.

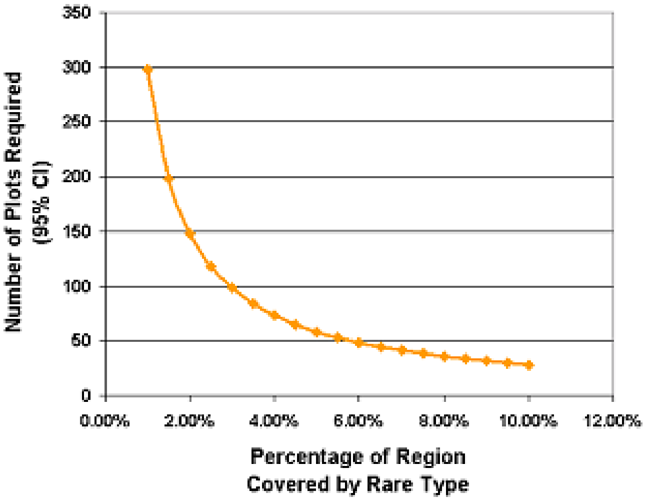

Figure 1: Sample size requirements as a function of percentage of region covered by the rare type.

3 Conclusions

The idea that vegetation classification systems, existing vegetation maps, and inventory data need to be well integrated has long been recognized. Within the mapping community, FIA and locally intensified inventory data have played a critical role in map unit design. Until now, however, mappers could say little about the vegetation components that were not in the sample. This paper provides insight into how statistically sound and quantitative statements can be made regarding the minor and unsampled landscape components. Analysts may now state their confidence that the proportion of a not-sampled dominance type is less than some proportion of the region of interest, which places an upper confidence bound on the proportion of not-sampled map units.

Acknowlededements

The authors would like to thank the reviewers and the editor for their constructive comments and valuable suggestions.

References

Bechtold, W.A.; Patterson P.L. Editors. (2005). The enhanced Forest Inventory and Analysis program-national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station. http://www.srs.fs.usda.gov/pubs/gtr/gtr_srs080/gtr_srs080.pdf. Lasted accessed August 2, 2011

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. http://www.fs.fed.us/emc/rig/includes/VEG_guide .pdf. Lasted accessed August 2, 2011.

Buffalo Ranger District (2009). Buffalo Valley Fuels Reduction Project – Draft Environmental Assessment, U.S. Department of Agriculture Forest Service, Bridger-Teton National Forest, Buffalo Ranger District, October 2009. 100p.

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

Gillham, J.; Goetz, W.; Fisk, H.; Lachowski, H.; Davy, L. 2007. Existing vegetation summary: Bridger-Teton National Forest. RSAC-0091-TECH1. Salt Lake City, UT: U.S. Department of Agriculture, Forest Service, Remote Sensing Applications Center. 124 p.

Thompson, Michael T.; DeBlander, Larry T.; Blackard, Jock A. 2005. Wyoming's Forests, 2002. Resour. Bull. RMRS-RB-6. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 148 p. http://www.fs.fed.us/rm/ogden/pdfs/wyoming_state .pdf. Lasted accessed August 2, 2011.