United States
Department
of Agriculture

Forest Service

**Rocky Mountain
Research Station**

General Technical
Report RMRS-GTR-66

October 2000

Annual Design-Based Estimation for the Annualized Inventories of Forest Inventory and Analysis: Sample Size Determination

Hans T. Schreuder
Jin-Mann S. Lin
John Teply

Abstract

Schreuder, Hans T.; Lin, Jin-Mann S.; Teply, John. 2000. **Annual design-based estimation for the annualized inventories of forest inventory and analysis: sample size determination.** Gen. Tech. Rep. RMRS-GTR-66. Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 3 p.

The Forest Inventory and Analysis units in the USDA Forest Service have been mandated by Congress to go to an annualized inventory where a certain percentage of plots, say 20 percent, will be measured in each State each year. Although this will result in an annual sample size that will be too small for reliable inference for many areas, it is a sufficiently large sample to draw annual inferences for large areas in terms of what is there and how much has changed on an annual basis. This is extremely useful in estimating the state of the forest in the country over time in terms of health, timber supply, and so forth.

The formula for estimating sample size required is: $n = t^2 s^2 / \left(p^2 \hat{\theta}^2\right)$ where s^2 is the estimated variance for parameter estimates $\hat{\theta}$ or $\hat{\theta} = \hat{\theta}_2 - \hat{\theta}_1$, and p is the desired level of precision expressed as percent of $\hat{\theta}$. We conclude that for p = 5 percent, based on sampling 1 ha plots on all National Forests in Oregon and Washington, that a sample size of 9,400 plots with a corresponding area of 22,560,000 ha, about the size of the State of Idaho, is sufficient to draw annual inferences. Alternatively, an area of 1,000,000 ha requires a sample size of 417 plots. Clearly p = 5 percent is a tough specification. If we go to p = 10 percent, which may be adequate for many purposes, we require one-fourth the sample size above, or 2,350 plots representing an area of about 5,640,000 ha, about the size of West Virginia (24,070 square miles). With this specification, a sample of 2,350 plots would give adequate precision.

Keywords: reliable probabilistic estimation, sample size determination, area required, key parameters

Authors

Hans T. Schreuder is a Mathematical Statistician, Rocky Mountain Research Station, USDA Forest Service, located at the Natural Resources Research Center, 2150 Centre Avenue, Building A, Fort Collins, CO 80526.

Jin-Mann S. Lin is a Graduate Research Assistant, Department of Statistics, Colorado State University, Fort Collins, CO.

John Teply, formerly Forester (retired), Region 6, Portland, Oregon.

Rocky Mountain Research Station
324 25th Street
Ogden, UT 84401

Annual Design-Based Estimation for the Annualized Inventories of Forest Inventory and Analysis: Sample Size Determination

Hans T. Schreuder
Jin-Mann S. Lin
John Teply

Introduction

The USDA Forest Service has been mandated by Congress to go to an annualized inventory (Agricultural Education Reform Act of 1998), measuring 20 percent of the plots in each state each year. The plot used is a 1 ha circular plot subsampled by four 1⁄24 acre circular subplots. Forest Inventory and Analysis (FIA) will be using a simple average of measured plots, called the moving average, to publish statistics for each state every 5 years. Although mandated to measure 20 percent per year it appears that this percentage will only be measured each year in the South; other areas may only have 14 or 10 percent measured each year. Multiple imputation and related model-based methods have been proposed to predict values for plots not remeasured in a given year to generate a clean-looking database that is easy to use by user groups. Schreuder and Reich (1998) note that public databases should be defensible even in court, and it would be better to draw inferences only on actual data using design-based (also called probabilistic) methods until procedures such as multiple imputations have been thoroughly tested.

This note recommends sample sizes and corresponding areas that are large enough so that probabilistic inferences can be safely drawn without resorting to models.

Methods

By design-based inference we typically mean that inference is assumed to be **essentially** model-free—essentially, because one can argue that any kind of inference assumes some underlying model. In general there are no design-based alternatives to the moving average method right now given the limitations of sample size, and we should encourage proper use of model-based methods by various user groups that may want to make their own assumptions to generate the complete data sets they want to use. Because results are likely to be contested bitterly, we need the best defensible estimates possible.

The problem then is: We have x percent of the plots measured each year, presumably x = 20 percent. For the last year of data, what sample size is needed to generate estimates of acceptable precision for all variables of interest? We assume we have t consecutive years of data, where t ≤5. This is where the idea of a simple average is proposed using the data from the t years of the annualized inventory. Basically then, if we have t years of data with t ≤5, using the simple average, we generate estimates for the period covered by the t years. This approach will be used at this time for publishing 5 year state reports by FIA. But users will use annual data. To do so from a scientifically satisfactory point of view we ask: For what sample sizes or areas can we use a straight probabilistic approach?

We want to determine that sample size n that will give us the following desirable characteristics in estimating a parameter θ:

1. An estimate of θ, $\hat{\theta}$, such that the standard error (se) of $\hat{\theta}$ is less than or equal to p1 percent of $\hat{\theta}$.
2. Estimates of $\Delta\theta, \Delta\hat{\theta} = \hat{\theta}_2 - \hat{\theta}_1$ such that the se of $\Delta\hat{\theta}$ is less than or equal to p2% of $\Delta\hat{\theta}$.
3. For area and area change estimation we may be able to use regression estimators $\hat{\theta}_{regr}$ where the sample size will depend on the strength of the correlation between ground (y) and remote sensing information (x) on the 1-ha plots.

The standard formula for sample size determination is (Schreuder and others 1993): $n = t^2 s^2 / \left(p^2 \hat{\theta}^2\right)$ where t is the tabulated t-value for the desired p level, s^2 is the estimated variance for parameter estimates $\hat{\theta}$ or $\hat{\theta} = \hat{\theta}_2 - \hat{\theta}_1$, and p is either p1 or p2. Similarly for the regression estimator in 3 above:

$$n = t^2 s^2\left(\hat{\theta}\right)\left(1 - \rho^2\right) / \left(\hat{\theta}^2_{regr} p^2\right)$$ where *r* is the correlation between y and x.

For determination of 1 above, we will concentrate on current basal area (G), number of trees excluding seedlings (N), number of trees including seedlings (N_s), mortality (M), and mortality including seedlings (M_s). These variables represent key variables of interest as well as some (M and M_s) that are likely to be as variable as any in the list of parameters of interest. All plots will be considered as measured, regardless of the fact that some plots may be partially or entirely nonforest.

For 2 above we need measured plots at two times close together to be able to estimate se ($\Delta\hat{\theta}$) for a period mimicking the annualized inventory time period of 1 year as much as possible, concentrating on ΔG, ΔN, ΔM, ΔN_s, and ΔM_s. These variables represent key variables of interest as well as some (ΔM and ΔM_s) that are likely to be as variable as any in the list of parameters of interest. We assume all plots to be measured 2 years apart even though some were remeasured over shorter and longer (from 1 to 3 years) periods. Paired differences are used for change calculations.

There are no data available yet for 3 above.

To determine sample size for 1 above, we used the entire data set of 1 ha plots available for Region 6 of the Forest Service, a total of 10,720 plots, and also used subsets of these data for each National Forest. We also used the one available panel of remeasured data sets available for most forests for recent change in the parameters of interest. This data set was used because it is the most intensive data available for an area, recent 2 year change data are available for a subset of the plots, and the same basic 1 ha primary sampling unit (psu) is used currently by FIA (FIA had used other plot and sample designs in the past; Birdsey and Schreuder 1992). Plots in Region 6 were established on a 2.74 km (1.7 mile) grid, with a 1 ha psu subsampled at five locations with the following subplots: the whole 1 ha plot for trees larger than 121.9 cm (47.9 inches) in the western part of the State and larger than 81.0 cm (31.9 inches) in the eastern part of the State, 0.076 ha (1⁄5 acre) for trees between 32.9 and 121.7 cm (13.0-47.9 inches) in the west and between 32.9 and 81.0 cm (13.0-31.9 inches) in the east, .020 ha (1⁄24 acre) for trees between 7.6 and 32.8 cm (3.0-12.9 inches), and 0.004 ha (1⁄100 acre) for seedlings (trees less than 7.4 cm or 3 inches in d.b.h.) (Max and others 1996).

The results are likely to be somewhat conservative because a grid sample as used is often more efficient than the simple random sampling assumed for variance calculations. We also ignored possible stratification into National Forests and the fact that in wilderness areas the grid intensity was on a 3.4 mile grid rather than the 1.7 mile grid used elsewhere. Both the stratification and the not weighting of the 3.4 mile grid plot locations can be safely ignored for our purposes.

Results are shown in table 1. For G, N, M, N_s, and M_s, these results are intuitively reasonable. The within the 1 ha plot sampling in Region 6 is geared to larger trees to some degree (three subplot sizes with larger subplots for the larger trees) so that one would expect generally to have better efficiency in estimating G relative to N. This is true for all forests. Number of trees including seedlings, N_s, Mortality, M, and

Table 1—Summary listing of smallest and largest coefficients of variation with required acreages for p = 0.05 for key parameters.

Variable	CV smallest	Hectares (acres)	Forest	CV largest	Hectares (acres)	Forest
G	55	1,171,200 (2,928,000)	Mt. Hood	94	3,369,600 (8,424,000)	Olympic
N	68	1,785,600 (4,464,000)	Umpqua	118	5,258,400 (13,146,000)	Wallowa-Whitman
M	110	4,617,600 (11,544,000)	Siuslaw	176	11,923,200 (29,808,000)	Malheur
N_s	55	1,152,000 (2,880,000)	Rogue River	112	4,800,000 (12,000,000)	Wallowa-Whitman
M_s	107	4,406,400 (11,016,000)	Umpqua	159	9,753,600 (24,384,000)	Wallowa-Whitman
ΔG	58	1,286,400 (3,216,000)	Malheur	122	5,683,200 (14,208,000)	Gifford Pinchot
ΔN	62	1,468,800 (3,672,000)	Willamette	95	3,475,200 (8,688,000)	Wallowa-Whitman
ΔM	103	4,099,200 (10,248 ,000)	Mt. Hood	242	22,425,600 (56,064,000)	Malheur
ΔN_s	48	902,400 (2,256,000)	Siskiyou	106	4,329,600 (10,824,000)	Wallowa-Whitman
ΔM_s	104	4,128,000 (10,320,00)	Gifford Pinchot	200	15,379,200 (38,448,000)	Malheur

mortality including seedlings, M_s, should be estimated less efficiently because they are less frequent or more highly peaked in occurrence. For a 5 percent precision criterion, the required sample size of G estimates varies from 480 for Rogue River to 1,404 for the Olympic Forest. For N they vary from 744 for the Umpqua to 2,236 for the Wallowa-Whitman. For N_s they vary from 480 for Rogue River to 2,000 for Wallowa-Whitman, for M from 1,924 for the Siuslaw to 4,968 for the Malheur, and for M_s from 1,836 for the Umpqua to 4,064 for the Wallowa-Whitman. Based on these results it is then clear that a sample of size 5,000 (rounded up from 4,968 for convenience of recollection) is quite adequate for sampling for these parameters of interest given the 5 percent criterion for precision. This translates to an area of 6,000 x 5,000 = 30,000,000 acres or about 12,000,000 ha.

The results for ΔG, ΔN, ΔM, ΔN_s, and ΔM_s are less consistent in that ΔG is usually estimated less efficiently than ΔN. This is plausible but not as readily obvious as for G verses N because of the possible dynamics involved over time. Mortality can fluctuate considerably over time, and this shows up clearly in both M and M_s. For ΔG sample sizes required vary from 536 for the Malheur to 2,368 for the Gifford Pinchot, from 612 for the Willamette to 1,448 for the Wallowa-Whitman for ΔN, from 1,708 for the Mt. Hood to 9,344 for the Malheur with ΔM, from 376 for the Siskiyou to 1,804 for the Wallowa-Whitman with ΔN_s, and from 1,720 for the Gifford-Pinchot to 6,408 for the Malheur forest with ΔM_s. Based on these results then it is clear that a sample of size 9,344 (say 9,400 for ease of recollection) is quite adequate for estimating these parameters, corresponding to an area of size 6,000 x 9,400 = 56,400,000 acres or about 22,560,000 ha, about the size of the State of Idaho. Clearly less stringent criteria than p = 5 percent may be used frequently resulting in a much smaller sample size requirement. With p = 10 percent for example, a sample of size 2,350 would be adequate, about the size of West Virginia.

References

Birdsey, R.A.; Schreuder, H.T. 1992. An overview of forest inventory and analysis estimation procedures in the eastern U.S.—with an emphasis on components of change. USDA Forest Service, Rocky Mountain Forest and Range Experiment Station. Gen.Tech. Rep. RM-214. 11 p.

Max, T.A; Schreuder, H.T.; Hazard, J.W.; Teply, J; Alegria, J. 1996. The Region 6 Vegetation Inventory and Monitoring System. USDA Forest Service, Pacific Northwest Station. Res. Pap. PNW-RP-493. 22 p.

Schreuder, H.T.; Reich, R. 1998. Data estimation and prediction for natural resources public data. USDA Forest Service, Rocky Mountain Research Station. Res. Note RMRS-RN-2. 5 p.

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please specify the publication title and General Technical Report number.

Fort Collins Service Center

Telephone	(970) 498-1392
FAX	(970) 498-1396
E-mail	rschneider@fs.fed.us
Web site	http://www.fs.fed.us/rm
Mailing Address	Publications Distribution Rocky Mountain Research Station 240 West Prospect Road Fort Collins, CO 80526

RMRS
ROCKY MOUNTAIN RESEARCH STATION

The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals.

Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Research Locations

Flagstaff, Arizona	Reno, Nevada
Fort Collins, Colorado*	Albuquerque, New Mexico
Boise, Idaho	Rapid City, South Dakota
Moscow, Idaho	Logan, Utah
Bozeman, Montana	Ogden, Utah
Missoula, Montana	Provo, Utah
Lincoln, Nebraska	Laramie, Wyoming

*Station Headquarters, Natural Resources Research Center, 2150 Centre Avenue, Building A, Fort Collins, CO 80526