United States
Department of
Agriculture

Forest Service

**Northeastern Forest
Experiment Station**

Research Note NE-345

Generalized Variance Function Applications In Forestry

James Alegria
Charles T. Scott

Abstract

Adequately predicting the sampling errors of tabular data can reduce printing costs by eliminating the need to publish separate sampling error tables. Two generalized variance functions (GVFs) found in the literature and three GVFs derived for this study were evaluated for their ability to predict the sampling error of tabular forestry estimates. The recommended GVFs for most tables are either a GVF which incorporated the sampling errors of the row and column totals or a nonlinear GVF when the sampling errors are not published. Tables composed with one sampling intensity and containing data from a multinomial distribution can be represented by a simple linear estimator.

Large surveys such as those conducted by USDA Forest Inventory and Analysis (FIA) Projects generate large amounts of information displayed in tabular form. Most often, the tables are published with few indications of the associated precision of the estimate. The lack of some measure of precision is not an oversight but is due to costs of publishing twice as many tables and the desire to maximize readability of the report. If a measure of precision is presented, it is often the sampling error in percent:

$$SE(T_{ij}) = 100(var(T_{ij}))^{.5} / T_{ij}$$

where T_{ij} is the cell estimate and $var(T_{ij})$ is the sample variance of the estimate. The sampling error is in the form of a simple approximation formula (Hines and Vissage, 1988) or a summary table of selected sampling errors. The sampling errors selected for publication are usually the errors of the table totals or subtotals. These can inadvertently lead to the assumption that the sampling errors of the individual cells within the table are of a similar magnitude. While this may be true, it is more common that the sampling error of a row or table total is smaller than the sampling error of an individual cell estimate. The cell sampling error can be many times greater than the sampling error of its row or column total. The objective of this study is to present GVFs as a cost-effective alternative to printing sampling error tables, and to increase the amount of information available in publications that currently do not present sampling errors for each cell in a table.

Methods

Models

The models investigated were selected from the literature or derived. For inclusion in the study, the model had to contain variables that were published regularly or that could easily be added to existing tables. They are the estimated totals of the individual cells, (T_{ij}), the row totals, ($T_{i.}$), the column totals, ($T_{.j}$), the grand total ($T_{..}$), and the sampling errors of the row and column totals, $SE(T_{.j})$ and $SE(T_{i.})$. The "." signifies the summation over that row or column subscript. The sampling errors for the row and column totals

could be printed as an additional column and row in the table. Hence, a model that utilized the row and column sampling errors was included in this study.

The models investigated were:

$$SE(T_{ij}) = SE(T_{..})(T_{..}/T_{ij})^{.5} \quad (1)$$

$$SE(T_{ij}) = (b_0 + b_1(1/T_{ij}))^{.5} \quad (2)$$

$$SE(T_{ij}) = b_0 + b_1(1/T_{ij})^{.5} \quad (3)$$

$$SE(T_{ij}) = b_1 T_{ij}^{b_2} T_{i.}^{b_3} T_{.j}^{b_4} \quad (4)$$

$$SE(T_{ij}) = b_0 + b_1 SE(T_{.j})(T_{.j}/T_{ij})^{.5} + b_2 SE(T_{i.})(T_{i.}/T_{ij})^{.5} \quad (5)$$

Model 1 is used by some FIA projects (Hines and Vissage, 1988) and has no unknown parameters that can be estimated from the data in the table. It assumes that the sampling error is inversely proportional to the square root of the cell value. It is also a special case of (3) with $b_0 = 0.0$ and $b_1 = SE(T_{..})/(T_{..})^{.5}$.

Model 2 has been used in large surveys by the Current Population Survey and the National Health Interview Survey in a related form (Valliant, 1987). Model 3 is the general form of (1). It allows for an intercept and does not force b_1 to equal a specific value but allows the data to determine the slope coefficient.

Model 4 was derived to account for nonlinear relationships between the variables T_{ij}, $T_{.j}$, and $T_{i.}$ and the dependant $SE(T_{ij})$. Replacing b_2 with $-.5$ and eliminating $T_{.j}$ and $T_{i.}$, the model becomes:

$$SE(T_{ij}) = b_1 T_{ij}^{-.5} \quad (6)$$

and is identical to Model 3 if b_0 is equal to zero.

Model 5 was derived to incorporate as much information as possible about the variability of the cell sampling errors without printing all of the sampling errors. It uses information from the sum of the rows and columns as well as the marginal sampling errors. The sampling errors of the row and column totals must be published for this model to be used.

Tests

The models were evaluated on a portion of the tables produced by the Northeastern Forest Experiment Station's FIA unit for the 1989 Kentucky inventory. Fifteen tables were divided into two types: those that contained continuous variables such as board-foot and cubic-foot volumes and those that were from multinomial distributions. An example of a multinomial variable is forest ownership in acres. The plot is either located on one of several ownership categories or it is not. The weighted proportion of plot occurrences in a category is multiplied by the total acreage in that county (assumed not to have error). Although acreage is usually considered as a continuous variable, it is treated as a constant. Other area tables in this study were derived in a similar manner.

The coefficient of determination (R^2) was the measure of performance. For Model 1 it was calculated as if it were a linear regression model; that is:

$$y_1 = SE(T_{..})(T_{..}/T_{ij})^{.5}$$

$$R_1^2 = 1 - \text{var}(SE(T_{ij}) - y_1)/\text{var}(SE(T_{ij})).$$

Similarly, the R^2s for the nonlinear models were:

$$R_i^2 = 1 - \text{SS corrected/MSE} \qquad i = 2,4$$

where SS is the sum of squares and MSE is the mean square error.

Sampling errors greater than or equal to 100 percent were eliminated from the regressions since they represent cells with a single non-zero observation. However, we believe that the models can be applied to these cells, resulting in an improved estimate of the variance.

Results

The proportion of variation explained, R^2, for each of the five models is shown in Table 1. Model 1 consistently explained less variation than the other models, and actually introduced more error than it explained for the continuous tables. Thus, Table 1 shows R^2 values of zero.

Model 2 performed well with the area tables, producing R^2 values between 0.81 and 0.99. This is the only model for which a theoretical basis has been shown to exist for two stage sampling used in populations surveys (Valliant 1987). Valliant also shows how this model should do well for continuous variables exhibiting gamma distribution of X(b,1) where b and 1 are population parameters, so long as the variables meet certain conditions. One condition is that the second parameter must be constant across all strata in the population, an improbability with tables composed of two or more sampling intensities.

Consequently, Model 2 did not perform well with the other variables. The R^2 values dropped to 0.49 to 0.84 for the board-foot, cubic-foot and number-of-stems tables.

With one exception, Model 3 performed as well as or better than Model 2 and equaled Model 4 for several of the tables. Its lowest R^2 was 0.87 for the area tables, but Model 4 was considerably better for tables involving number of stems per acre.

The range R^2 values for area tables in model 4 was 0.90 to 0.99. It dropped to 0.67 to 0.92 for the continuous tables. The overall performance of Model 4 was second only to

Table 1. Proportion of variation explained, R^2, for each of the four models applied to 15 different tables from 1989 Kentucky forest survey

Attribute	Rows	Columns	R^2 for Model — 1	2	3	4	5
Forest area	County	Owner class	0.40	0.81	0.89	0.90	0.96
Forest area	County	Forest type	0.43	0.87	0.93	0.95	0.98
Forest area	County	Stand size	0.41	0.81	0.87	0.90	0.93
Forest area	County	Site class	0.42	0.86	0.92	0.92	0.97
Forest area	County	Stocking class	0.41	0.85	0.92	0.92	0.98
Forest area	Forest type	Owner class	0.50	0.98	0.98	0.98	0.97
Forest area	Owner class	Stocking class	0.49	0.99	0.99	0.99	0.99
Forest area	Forest type	Stand size	0.48	0.99	0.97	0.98	0.98
Net cubic-foot volume	Species	Diameter class	0.00	0.76	0.78	0.82	0.84
Cubic-foot vol. in Sawlog	Species	Diameter class	0.00	0.77	0.79	0.83	0.83
Board-foot volume	Species	Diameter class	0.00	0.72	0.74	0.78	0.84
Net cubic-foot volume	Tree class	Species group	0.00	0.84	0.87	0.92	0.88
Board-foot volume	Species	Log grade	0.00	0.66	0.72	0.79	0.81
No. live trees	Species	Diameter class	0.00	0.49	0.53	0.69	0.88
No. growing-stock trees	Species	Diameter class	0.00	0.49	0.53	0.67	0.87

Model 5. For the area tables, the estimate for the coefficient b_2 ranged from −.5 to −.6, and b_3 and b_4 were either not significantly different from zero or barely significant at the 0.05 level.

Model 5 had the best overall performance with R^2 values of 0.93 to 0.99 for the area tables and 0.81 to 0.88 for the remaining tables. However, it performed substantially better than Models 3 and 4 only for the two tables involving number of trees per acre. The marginally better performance was attributed to the additional information in the model from the sampling errors of the row and column totals.

Many tables in forestry are constructed using more than one sampling intensity. This alters the relationship of cell estimate to cell variance, which is the basis for generalized

variance functions. Most of the models were able to predict well the estimated cell variances for the area tables. The area tables are derived from information on the entire plot with each attribute recorded only once per plot. This is in contrast to the other tables where data are recorded on individual trees in various combinations of plot designs. The problem is especially acute when the table is separated into diameter classes. When more than one fixed plot size is used, the decision whether to record information from a tree is based on the tree's distance from the plot center and the diameter. So the diameter classes in a table also correspond to plot size. The relationship between the cell totals and the sampling errors is not a "smooth curve" but actually two or more curves because sampling error increases as plot size decreases.

The problem of estimating a table composed of more than one fixed plot size becomes apparent when the table is divided into diameter classes. The table for number of live trees per acre contains three fixed plot sizes and the columns are divided into diameter classes. Deciding which plot a tree belongs to depends on the diameter of the stem as well as its distance from the plot center, so the column headings divided the tables in plot sizes as well as diameter classes. Three separate equations, one for each plot size, was fitted to the table using Model 3; the R^2 improved from 0.53 to 0.82, indicating that the GVF differs for each of the three diameter-class groups.

A model's ability to predict the sampling errors of "mixed" sampling intensities depends on the amount of information about the table within the model and its degree of flexibility. Model 5 included information concerning the sampling errors of $T_{.j}$ and $T_{i.}$ and consequently performed better than the other models. Model 4 was second because of its ability to adjust the power of the exponents to better estimate a compromise coefficient through the collection of points produced by the different sampling intensities.

Conclusion

Models 2 through 5 performed well for the area tables. If their performance can be classified as about equal, then two factors remain in deciding which model to use. The first is ease of use for the reader — the fewer the number of coefficients to multiply and the fewer variables to locate in the table, the better. The second factor is ease in implementing the regression procedure into existing software and any maintenance of the procedure that may be required by the producers of the tables. Nonlinear models are more difficult to implement because of the initial estimates of the parameters for each table or groups of tables and the possibility of changing the initial values from one data set to another. Changing the estimates can become burdensome, especially when many tables are produced as a part of a larger program requiring considerable computer resources. On the basis of these factors, Model 3 is recommended.

For tables constructed from more than one sampling intensity, Model 5 is recommended if $SE(T_{.j})$ and $SE(T_{i.})$ are published. Otherwise, Model 4 is recommended so long as the problems of implementing a nonlinear subroutine are not considered significant. Model 3 is a consideration should Models 4 and 5 be deemed unsuitable.

Including generalized variance functions with tables has been shown to be feasible even with tables composed of varying sampling intensities. GVF's have the advantage of saving costs as compared to publishing separate tables of sampling errors, adding utility to those publications that currently do not publish sampling errors for all tables, and maintaining a high degree of readability in the final product.

Literature Cited

Hines, F. Dee; Vissage, John S. 1988. **Forest statistics for Arkansas counties — 1988.** Resour. Bull. SO-141. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station. 68 p.

Valliant, R. 1987. **Generalized variance functions in stratified two-stage sampling.** Journal of the American Statistical Association. 82: 499-508.

JAMES ALEGRIA is a forester, Northeastern Forest Experiment Station, Radnor, Pennsylvania 19087.

CHARLES T. SCOTT is a research forester and Project Leader, Northeastern Forest Experiment Station, Delaware, Ohio 43015.

MANUSCRIPT RECEIVED FOR PUBLICATION 18 MARCH 1991

Northeastern Forest Experiment Station
5 Radnor Corporate Center
100 Matsonford Road, Suite 200
P.O. Box 6775
Radnor, Pennsylvania 19087

April 1991

U. S. GOVERNMENT PRINTING OFFICE: 1991/548-088/20021