BACK

Beyond Assistance: The Role of Models in Forest Inventory

Timothy G Gregoire
Yale University
timothy.gregoire@yale.edu

Stephen V Stehman
SUNY - ESF
svstehma@syr.edu

The bole volume of a tree is too onerous to measure in routine forest inventory, therefore a model is used to predict its volume. The same is true of its aboveground biomass. This practice has implications for the statistical properties of estimators of per unit area volume and biomass. Their properties differ from those of analogous estimators of tree frequency and basal area which do not rely on a presumed model. In this presentation we make an initial attempt to articulate and understand the statistical implications for design-based inference resulting from a reliance on models in this context. The relevance of this investigation stems from the current high concern with the use of allometric equations to predict aboveground biomass of trees as part of the MRV process in implementing the REDD+ program of the United Nations.

Introduction

Following the appearance of *Foundations of Inference in Survey Sampling* in 1977 by Cassell, Särndal, & Wretman, coupled with the later publication of *Model Assisted Survey Sampling* in 1992 by Särndal, Swensson, & Wretman, many of us in forest biometry suddenly were confronted by the difference in using a presumed model as the basis for inference and using it to assist estimation in conventional sampling, where inference has long been based on the sampling design. For many of us the “yellow book” especially had a formative influence on our understanding of this distinction. Ideas that seem so rudimentary today were less than obvious nearly a quarter century ago.

Since that time, I and many others have written for audiences such as the one assembled here on the essential difference between design-based and model-based inference in forest sampling. Two recent examples include the 2009 article by Stehman in *Remote Sensing of Environment* about “Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover changes from remote sensing.” McRoberts writing also in RSE in 2010 addressed “Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data.”

Wherever professional forestry is well established around the globe, I believe that those who specialize in forest inventory have a very good grasp of precepts of sampling and the manner in which the sampling design influences the assessment of sampling variation, the consequent construction of confidence intervals, and other tools of statistical inference within the design-based framework. When model predictions are used in place of actual measurements for the purpose of estimating population parameters, our collective understanding becomes obscured.

In model-assisted estimation one or more covariates are used to model the variable of interest, y. The model is not of inherent interest nor is it used to predict the value of y for

unsampled elements of the population, but rather to improve the precision of the estimator of aggregate y, say τ. In model-assisted estimation, as advanced by the authors cited above, the reference distribution for probability statements about the likely value of τ is the distrbution of all possible estimates permissible under the sampling design.

The situation we consider in this presentation arises when the values of the variable of interest are predicted rather than measured.

Forest inventory

In *Sampling Strategies for Natural Resources and the Environment* (2008), Valentine and I wrote about equal probability sampling with fixed-area plots, as well as Bitterlich sampling and line intersect sampling which select trees with unequal probability. Once trees are selected at one location, the Horwitz-Thompson estimator may be used to estimate a design-unbiased estimator of number of trees or basal area for the entire population. Since the sampling is replicated at a number, say m, of different sampling locations, the arithmetic average of the m estimates is likewise design-unbiased. To the best of my knowledge Barabesi & Fattorini (1998) introduced the term “replicated estimator” for this average, a nomenclature that we adopted in our book. Moreover they assert that the Central Limit Theorem assures that the replicated estimator converges in distribution to a Gaussian distributed random variable.

Bole volume or tree biomass almost always has been predicted by using a fitted regression model in a non-assisted fashion. The usual methods of forest inventory rely on a previously fitted model (formerly, a row and column table) to predict the bole volume of the standing tree. It is the fact that volume is predicted not measured that impacts purely design-based inference for volume-related estimates of the population in a manner that is difficult to discern. The same is true of individual tree biomass predictions and their effect on the estimation of aggregate biomass.

As a brief side commentary, it was Lew Grosenbaugh's distrust in the accuracy of so-called volume equations that prompted him to devise 3P sampling in the early 1960s as an alternative. This was a novel method of unequal probability sampling, despite its superficial resemblance to Hajek's (1958) Poisson sampling, which was presented in Hajek's posthumously published 1981 sampling text as an alternative form of list sampling. With a few exceptions, the adoption of 3P has never been widespread.

In 1986 Gregoire, Valentine and Furnival wrote about the “Estimation of bole volume by importance sampling” as an alternative sampling procedure that would likewise enable design-unbiased estimation of volume. In our exposition of importance sampling (IS) for this purpose, a model is used to assist estimation in the conventional sense articulated in the yellow book. This was followed by a series of related papers, yet the method never was adopted in practice.

Nor was a method of three stage sampling by the same authors in 1993, wherein trees were selected proportional to basal area in stage one; in stage two, a subset was selected

with probability proportional to height; and in stage three, IS was applied to the trees selected in the second stage.

Consequences of volume/biomass prediction

Preliminary reflections

To set the stage to explore the impact of model-based predictions on design-based inference, consider a population, $\mathcal{U}$, of N trees of interest occupying a region $\mathcal{A}$ with land area $A = |\mathcal{A}|$. The aggregate basal area of the N trees is denoted by G, and aggregate bole volume (or aboveground biomass) by B.

The HT estimator based on the sample selected at location $\mathcal{P}_s \in \mathcal{A}$ is

$$\widehat{\tau}_s = \sum_{k \in \mathcal{P}_s} \frac{y_k}{\pi_k},$$

where $k \in \mathcal{P}_s$ indicates inclusion in the sample with probability π_k. The restriction that $\mathcal{P}_s \in \mathcal{A}$ is not strictly necessary, yet it conforms to widespread practice. It is important, however, that $k \in \mathcal{A}$, inasmuch as this is integrally associated with the population of interest.

When $y_k = 1 \, \forall k \in \mathcal{U}$, then $E_p[\widehat{\tau}_s] = N$.

Likewise $E_p[\widehat{\tau}_s] = G$ when $y_k = g_k \, \forall k \in \mathcal{U}$, where g_k signifies the basal area of the kth tree in $\mathcal{U}$. In making this assertion we are presuming that the measurement of tree diameter and subsequent computation of its circular cross-sectional area introduces negligible error in g_k. In this regard, Matérn's 1956 monograph “On the geometry of the cross-section of a stem” is well worth reading again.

When $y_k = b_k$ is obtained by 3P sampling with negligible measurement error of stem volume/biomass, b_k, then $E_p[\widehat{\tau}_s] = B$. Or, when $y_k = b_k$ is replaced by $y_k = \widehat{b}_k$, an estimate of bole volume (biomass) from applying IS as a second or third stage, then it is straightforward to show that $E_p[\widehat{\tau}_s] = B$. Henceforth, I will speak in terms of tree's volume, with the tacit understanding that I could refer as well to biomass.

In contrast to the above, $E_p[\widehat{\tau}_s] \neq B$ when $y_k = \tilde{b}_k \, \forall k \in \mathcal{U}$, where $\tilde{b}_k$ signifies the tree's volume predicted from a fitted regression model. The reason for the bias in the latter case is that $\tilde{b}_k \neq b_k$ coupled with the absence of any sampling error in the prediction $\tilde{b}_k$ that becomes nil when averaged over all possible samples of a single tree. Indeed, there is no sampling variation in the prediction of volume for the kth bole.

Moreover the design-variance of $\widehat{\tau}_s$, denoted as $V_p[\widehat{\tau}_s]$, when $y_k = b_k$ differs from what it is when $y_k = \tilde{b}_k$, and it is difficult to say, in general, which is the more precise.

Deducing the effects on $E_p[\widehat{\tau}_s]$ and $V_p[\widehat{\tau}_s]$

Not very long ago, I adopted the point of view that the HT estimator of aggregate volume, B, was unbiased conditionally on the correctness of the model of bole volume that was being used for prediction. Because of the ambiguity of model "correctness", I no longer find that a satisfying or convincing viewpoint.

In discussion with my co-author, we speculated whether the deviation of actual from predicted volume might usefully be viewed as a measurement error problem. To follow this line of inquiry, it might be helpful to identify the sources of the error, and more importantly to trace how they impact the design-based bias, variance, and MSE of the erstwhile HT estimator of B. A version of this has been done before, notably by Cunia (1987) in a proceedings paper entitled "Error of forest inventory estimates: its main components". The limiting aspect of that contribution is the failure to discern between statistical properties reckoned with respect to a presumed model and those that are reckoned with respect to the sampling design. To our way of thinking, an expression of the variance of an estimator in which the terms are second moments of different distributions does not make sense.

The pm approach

In the yellow book, Särndal et al devote Chapter 16 to the problem of measurement errors. While they speculate on various sources and causes of such errors, they do not explicitly treat the case where the error arises from a model-based prediciton of the true value. This may be nothing more than a curiosity, but surely the applictions with which we in this audience are involved must have analogs in other fields, do they not? It may be a moot point, inasmuch as the approach advanced by these authors may have utility to our investigation. To paraphrase from p. 603 of the yellow book, "The statistical properties of $\widehat{\tau}_s$ can only be studied by making assumptions about the errors. It become necessary to formulate stochastic measurement error models."

The sample survey "is viewed as a two-stage process with each stage contributing randomness." In stage one, the sample is selected, and stochastic structure is given by the sampling design, symbolized by p. In stage two, the measurement procedure generates an observed value y_k, and the stochastic structure is given by the measurement error model, m.

That strikes me as plausible, and I wish to see whether it leads to a useful solution to the problem we face with volume/biomass predictions at the tree level. Based on my prior investigation related to the work of our LiDAR team of researchers (Ståhl et al, 2011) since 2004, I am skeptical in the absence of so many simplifying assumptions as to render the approach nugatory. Nonetheless, let me explain the way advanced by Särndal et al further. For the expected value of $\widehat{\tau}_s$ in this two-stage view, they present

$$E_{pm}[\widehat{\tau}_s] = E_p[E_m[\widehat{\tau}_s|s]] \qquad (1)$$

where $E_m[\widehat{\tau}_s|s]$ is the expectation conditionally on the observed sample with respect to the measurement error model m, and $E_p[\bullet]$ is the usual design-based expectation, so that

$E_{pm}[\,\widehat{\tau}_s]$ signifies the expectation with respect to the sampling design and measurement model jointly. They refer to this as the pm expectation. The pm bias is

$$B_{pm}[\widehat{\tau}_s] = E_{pm}[\,\widehat{\tau}_s] - \tau, \tag{2}$$

where τ is the population parameter of interest, that is, the value being estimated by $\widehat{\tau}_s$

Similarly, the pm variance is,

$$V_{pm}[\,\widehat{\tau}_s] = E_p[V_m[\,\widehat{\tau}_s|s]] \; + V_p[E_m[\,\widehat{\tau}_s|s]]\,. \tag{3}$$

In forestry, we have seen an identical construction in the work of Daniel Mandallaz.

The pm mean square error

$$MSE_{pm}[\,\widehat{\tau}_s] = E_{pm}\big[\,(\widehat{\tau}_s - \tau)^2\big]. \tag{4}$$

The measurement error model considered in the yellow book stipulate moments that do not depend on the observed sample, namely

$$\mu_k = E_m[\,y_k|s]. \tag{5}$$

$$\sigma_k = V_m[\,y_k|s] \tag{6}$$

$$\sigma_{kk'} = C_m[\,y_k, y_{k'}|s] \; = E_m[y_k y_{k'}|s] - \mu_k \mu_{k'}. \tag{7}$$

Let's see where the pm framework leads us when using the proffered model for measurement error, arising from using a fitted regression model to predict bole volume because of the infeasability of actually measuring bole volume.

Applying *pm* moments

A general model for bole volume, b, of an individual tree is

$$h(b) = g(X';\,\beta) \; + \; \epsilon_k \tag{8}$$

where X'_k is a q-element covariate vector.

When both $h()$ and $g()$ are the identity functions, then (6) is the familiar linear model

$$b = X'\beta \; + \; \epsilon_k. \tag{9}$$

A nonlinear model arises when $h()$ is the identity function but $g()$ is nonlinear in β, for example

$$b = \beta_1 X_1^{\beta_2} \; + \; \epsilon_k. \tag{10}$$

Another special case arises when $h()$ is a nonlinear transformation (reciprocal, logarithmic, square root, Box-Cox) of b, for example

$$\ln(b) = \beta_0 + \beta_1 \ln(X_1) + \epsilon_k \tag{11}$$

Assessed with respect to the presumed model,

$$E_\xi[\, b\,] = \; h^{-1}(g(X'; \, \beta) \; + \; \epsilon) \tag{12}$$

Strictly speaking, model bias arises when the actual mean, namely $\phi_{b|X}$, does not coincide with $E_\xi[\, b\,]$.

Generalized linear models offer an attractive alternative to (12), but are not considered in this presentation.

The fitted model is used to predict the value of b_k by

$$\tilde{b}_k = \; h^{-1}\Big(g(X'_k; \, \widehat{\beta})\Big) \tag{11}$$

where β is estimated with data extraneous to the inventory sample. For sake of later reference, let the model-fitting data be labelled Ω.

Let me restate the task at hand.

From the measurements taken at sampling location $\mathcal{P}_s \in \mathcal{A}$, τ is estimated by the HT estimator

$$\widehat{\tau}_s = \sum_{k \in \mathcal{P}_s} \frac{y_k}{\pi_k},$$

in which $y_k = \tilde{b}_k$ from the previously fitted model to predict bole volume. Treating the difference between actual and predicted volume, namely $b_k - \tilde{b}_k$, as measurement error, we ask whether the yellow book's measurement error model and pm strategy for inference is a useful one to enable us to deduce the statistical properties of $\widehat{\tau}_s$.

I think there are two critical disjunctures that impede our efforts, and thereby prevent us from immediate adoption.

To understand the first impediment, let me cite again from p. 607 of Särndal et al (1992):

> It is important to have a clear notion of the frequency interpretation of the simple measurement model m given above. There is a given probability sample s and a given measurement procedure that generates an observed value for each element $k \in s$. Suppose measurements could be independently repeated many times on the same sample s, thus generating a long series of measurements on each element $k \in s$. The observed y-values for a particular k would not necessarily be the same in all repetitions, but would vary in a random fashion around a long run mean value μ_k and with a long run variance σ_k^2.

That is not true of the error with which we are dealing, $b_k - \tilde{b}_k$. Unless one considers errors in the measurement of tree diameter that causes $\tilde{b}_k$ to be different for each dbh measurment, $\tilde{b}_k$ will not change. The difference $b_k - \tilde{b}_k$ with which we are dealing does not have a stochastic component of the sort imagined in the yellow book's Chapter 16. It has, instead, a distinctly model-based component which is largely determined by the fitting data, Ω.

Secondly, in the design-based framework, the mantra is that the N elements comprising the population are fixed, and their y values likewise are fixed. Yet in the modeling framework, the b values are considered to be random variables with expected values that vary smoothly with covariate X. The prediction of b_k by $\tilde{b}_k = h^{-1}\Big(g(X_k'; \widehat{\beta})\Big)$ in forest inventory is a convenient tool. Yet it seems reasonable to ask what the model-based properties of $\tilde{b}_k$ have to do with the design-based properties of

$$\widehat{\tau}_s = \sum_{k \in \mathcal{P}_s} \frac{\tilde{b}_k}{\pi_k} \tag{12}$$

as an estimator of τ. The $\tilde{b}_k$ could have been my seat-of-the-pants guesses, and the essential question would be unchanged: how does the use of a value that is not measured affect the design-based properties of $\widehat{\tau}_s$?

Reconsider the measurement error model

While not abandoning the pm approach yet, my current direction is to look at the HT estimator of the prediction error, $\delta_k = b_k - \tilde{b}_k$, itself.

At the risk of complicating our notation for the sake of greater explicitness, let me use

$$\widehat{\tau}_s(b) = \sum_{k \in \mathcal{P}_s} \frac{b_k}{\pi_k} \tag{13}$$

$$\widehat{\tau}_s(\tilde{b}) = \sum_{k \in \mathcal{P}_s} \frac{\tilde{b}_k}{\pi_k} \tag{14}$$

$$\widehat{\tau}_s(\delta) = \sum_{k \in \mathcal{P}_s} \frac{\delta_k}{\pi_k} \tag{15}$$

Therefore

$$\widehat{\tau}_s(\tilde{b}) = \widehat{\tau}_s(b) + \widehat{\tau}_s(\delta) \tag{16}$$

Expressing (16) in words, the usual estiamtor of aggregate volume $\widehat{\tau}_s(\tilde{b})$ decomposed as the sum of the estimator, $\widehat{\tau}_s(b)$, we would like to use, plus an unbiased estimator of the population difference in the actual versus predicted volume.

That latter difference might be attributable to applying the linear model in (9), namely $b = X'\beta + \epsilon_k$, to a subportion of the physiographic growing region represented in Ω exhibits morphologically different volume accretion patterns from the population as a whole. Or it might be attriabutable to applying it to a single species in the suite of species in Ω to which the model was fitted.

Or the expectation of $\widehat{\tau}_s(\delta)$ might be due to the bias in the estimates of β after having fiited a nonlinear mean function, such as $b = \beta_1 X_1^{\beta_2} + \epsilon_k$ in (10).

Or the expectation of $\widehat{\tau}_s(\delta)$ might be due to the back-transformation bias after having fitted a model with a nonlinear transformation , such as $\ln(b) = \beta_0 + \beta_1 \ln(X_1) + \epsilon_k$, as in (11).

The possible utility of (16), $\widehat{\tau}_s(\tilde{b}) = \widehat{\tau}_s(b) + \widehat{\tau}_s(\delta)$, is that it separates what we know how to deal with, namely $\widehat{\tau}_s(b)$, and allows us to focus squarely on the causes and consequences of the prediction error, δ.

Evidently, much more needs to be done before evidence of its utility can be presented.

References

Barabesi, L and Fattorini, L. 1998. The use of replicated plot, line and point sampling for estimating species abundance and ecological diversity. *Environmental and Ecological Statistics* 5:353-370.

Cassel, C-M, Särndal, C-E, Wretman, J. H. 1977. *Foundations of Inference in Survey Sampling.* Wiley: New York. 192 p.

Cunia, T. 1987. Error of forest inventory estimates: its main components. IN: *Estimating Tree Biomass Regressions and Their Error.* Proceedings of the Workshop on Tree Biomass Regression Functions and their contribution to the error of forest inventory estimates. (E. H. Wharton & T. Cunia, eds). USDA Forest Service NE-GTR-117.

Gregoire, T. G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research* 28:1429-1447.

Gregoire, T. G., H. T. Valentine, and G. M. Furnival. 1986. Estimation of bole volume by importance sampling. *Canadian Journal of Forest Research* 16:554 557.

Gregoire, T. G., H. T. Valentine and G. M. Furnival. 1993. Two-stage and three-stage sampling strategies to estimate aggregate bole volume in the forest. Ilvessalo Symposium on National Forest Inventories proceedings of the International Union of Forest Research Organizations, S4.02 conference, 17-21 August 1992, Helsinki, Finland (A. Nyyssonen, S. Poso, J. Rautala, ed.) The Finnish Forest Research Institute Research Paper 444. pp. 201-211

Gregoire, T. G. and Valentine, H. T. 2008. *Sampling Strategies for Natural Resources and the Environment.* Chapman & Hall/CRC: Boca Raton.

Hajek, J . 1958. Some contributions to the theory of probability sampling. *Bulletin of the Institute of International Statistics*. 36:127-134.

Hajek, J . 1981. *Sampling from a finite population*. Marcel Dekker: New York. 247 p.

Matérn, B. 1956. “On the geometry of the cross-section of a stem”. *Meddelanden från statens skogsforskningsinstitut, Stockholm* 46(11) 28p.

Särndal, C-E, Swensson, B. & Wretman, J. 1992 *Model Assisted Survey Sampling* Springer-Verlag: New York. 694p.

Stehman, S. 2009. Model-assisted estimation as a unifying framework for estiamting the area of land cover and land-cover changes from remtoe sensing. *Remote Sensing of Environment* 113:2455-2462.

McRoberts, R. E. 2010. Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. *Remote Sensing of Environment* 114: 1017-1025.

Ståhl, G., Holm, S., Gregoire, T. G., Gobakken, T., Næsset, E., & Nelson, R. 2011. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway . Canadian Journal of Forest Research 41:96-107.