The effects of uncertainty in individual tree volume model predictions on large area estimates of forest volume

Ronald E. McRoberts

Northern Research Station, U.S. Forest Service, Saint Paul Minnesota 55038 USA.

Abstract

Greenhouse gas inventories for the forestry sector rely on large area estimates of tree biomass. These estimates are calculated by predicting volumes of individual trees, multiplying by a biomass conversion factor, adding individual tree biomass estimates at the plot level, and then averaging over plots to obtain the large area estimates. However, the uncertainty in the model predictions is generally ignored with the result that the precision of the large area estimates is over-estimated. The study objective was to estimate the effects of model-related uncertainties on large area volumes estimates for study areas in the state of Minnesota, USA, and the state of Santa Catarina, Brazil. Monte Carlo simulation approaches were used because of the complexities associated with multiple sources of uncertainty and the nonlinear nature of the models. The effects of uncertainty in model predictions on the large area volumes estimates were small, although the results depend heavily on sample size and quality of fit of models.

Keywords: Residual uncertainty, parameter uncertainty, nonlinear model.

1. Introduction

Statistical models are routinely used by forest monitoring programs to predict volume, biomass, and carbon for individual trees using measurements of species, diameter, and height as independent variables. When the individual tree model predictions are aggregated at plot-level, the effects of uncertainty in the model predictions are routinely ignored. Failure to account for these effects leads to erroneously optimistic precision for large area estimates. From a rigorous statistical perspective, this practice cannot be justified, although from a practical perspective it can perhaps be justified if the effects of model prediction uncertainty are negligible relative to the effects of other sources of uncertainty such as plot-to-plot variability.

2. Data

For each of two study areas, one in Minnesota, USA, and one in Santa Catarina, Brazil, two datasets were acquired. The *calibration datasets* consisted of very precise estimates of volume and measurements of diameter and height for representative species-specific samples of individual trees. The *estimation datasets* consisted of observations of species and measurements of height and diameter for individual trees on arrays of forest inventory plots. The Minnesota estimation dataset was acquired using a tessellated random sampling design, and the Santa Catarina estimation dataset was acquired using a systematic sampling design.

3. Methods

An allometric model of the relationship between individual tree volume (V) as the dependent variable and diameter at-breast-height (dbh) and height (ht) as the independent variables was formulated as,

$$V = \beta_0 \cdot dbh^{\beta_1} \cdot ht^{\beta_2} + \varepsilon, \quad (1)$$

where the β s are parameters to be estimated, and ε is a random residual with mean 0. Before model fitting, natural logarithmic (ln) transformations of both dependent and independent variables were calculated, and the model was reformulated as,

$$\ln(V) = \alpha_0 + \alpha_1 \cdot \ln(dbh) + \alpha_2 \cdot \ln(ht) + \varepsilon \quad (2)$$

where the α s are the parameters to be estimated. The advantages of the transformations were that the model could be expressed in linear form which facilitates estimation of the parameters, and heteroskedasticity was removed, thereby eliminating the necessity of weighted regressions. On the original scale, predictions for the i^{th} tree were calculated as,

$$\hat{V}_i = \exp\left[\hat{\alpha}_0 + \hat{\alpha}_1 \cdot \ln(dbh_i) + \hat{\alpha}_2 \cdot \ln(h_i) + \frac{\hat{\sigma}_{res}}{2}\right] \quad (3)$$

where $\hat{\sigma}_{res}$ is the residual standard deviation calculated on the ln-ln scale, and $\frac{\hat{\sigma}_{res}}{2}$ compensates for bias that accrues when transforming from the ln-ln scale back to the original scale (Baskerville, 1972). The quality of model fit was assessed using R^2 on both the ln-ln and original scales. Models were constructed for individual species or species groups and for the non-specific aggregation of all species.

The study focused on the effects on estimates of large area mean volume per unit area of model prediction uncertainty from two sources, model residual uncertainty and parameter uncertainty. The heteroskedastic residual uncertainty on the original scale was assessed using 3-step procedure: (i) the pairs $(V_i, \hat{V}_i)$ were ordered with respect to the model prediction, $\hat{V}_i$; (ii) the pairs were aggregated into groups of size at least 25; and (iii) within each group, g, the mean of the observations $\overline{V}_g$, the mean of the predictions $\overline{\hat{V}}_g$, and the standard deviation $\hat{\sigma}_g$ of the residuals $\varepsilon_i = V_i - \hat{V}_i$, were calculated. The relationship between the group standard deviations and the group prediction means was represented using the model,

$$\hat{\sigma}_g = \gamma_1 \cdot \overline{\hat{V}}_g^{\ \gamma_2} + \varepsilon \quad (4)$$

where the γ s are parameters to be estimated.

Uncertainty in the model parameter estimates was assessed on the ln-ln scale using a 4-step Monte Carlo approach: (i) the transformed dataset was aggregated into dbh size classes; (ii) each dbh size class was resampled with replacement until the original class sample size was achieved; (iii) the model was fit to the resampled data and the parameters were estimated, and (iv) the procedure was replicated 5,000 times. The resulting distribution of parameter estimates represented the uncertainty in estimates of the parameters on the ln-ln scale.

A 5-step Monte Carlo simulation procedure was used to estimate the effects of model residual uncertainty and parameter uncertainty on the uncertainty of large area estimates of mean volume per unit area.

Step 1. For the k^{th} replication a set of model parameter estimates, $\hat{\alpha}^k$, was randomly selected from the distribution previously constructed.

Step 2. For the i^{th} tree on the j^{th} plot in the estimation dataset, a volume observation was simulated using the parameter estimates from Step (1) as,

$$V_{ij}^k = \exp\left[\hat{\alpha}_0^k + \hat{\alpha}_1^k \cdot \ln\left(dbh_{ij}\right) + \hat{\alpha}_2^k \cdot \ln\left(h_{ij}\right) + \frac{\hat{\sigma}_{res}}{2}\right] + \varepsilon_{ij} \qquad (5)$$

where $\hat{\sigma}_{res}$ is as described following Equation (3) and ε_{ij} is a residual randomly selected from a Gaussian $(0, \hat{\sigma}_i^2)$ distribution for which $\hat{\sigma}_i$ is calculated using Equation (4) with $\exp\left[\hat{\alpha}_0^k + \hat{\alpha}_1^k \cdot \ln\left(dbh_{ij}\right) + \hat{\alpha}_2^k \cdot \ln\left(h_{ij}\right) + \frac{\hat{\sigma}_{res}}{2}\right]$ as the predictor variable.

Step 3. The total volume for the k^{th} replication for the j^{th} plot in the estimation dataset was calculated as $V_j^k = \sum_{i=1}^{n_j} V_{ij}^k$ where n_j is the number of trees on the plot.

Step 4. The mean and variance of the mean over all plots for the k^{th} replication were calculated as,

$$\overline{V}^k = \frac{1}{n_{plot}} \sum_{j=1}^{n_{plot}} V_j^k \qquad (6)$$

and

$$\hat{Var}\left(\overline{V}^k\right) = \frac{1}{n_{plot}\left(n_{plot} - 1\right)} \sum_{j=1}^{n_{plot}} \left(V_j^k - \overline{V}^k\right)^2 \qquad (7)$$

where n_{plot} is the number of plots.

Step 5. Steps (1)-(4) were replicated and the mean and variance over replications were calculated as per Rubin (1987),

$$\hat{\mu} = \frac{1}{n_{rep}} \sum_{k=1}^{n_{rep}} \overline{V}^k \qquad (8)$$

and

$$\hat{Var}\left(\hat{\mu}\right) = \left(1 + \frac{1}{n_{rep}}\right) \cdot W_1 + W_2 \qquad (9)$$

where $W_1 = \frac{1}{n_{rep} - 1} \sum_{k=1}^{n_{rep}} \left(\overline{V}^k - \hat{\mu}\right)^2$ and $W_2 = \frac{1}{n_{rep}} \sum_{k=1}^{n_{rep}} \hat{Var}\left(\overline{V}^k\right)$ are the among- and within-replications variances, respectively, and n_{rep} is the number of replications. The replications continued until $\hat{\mu}$ and $\hat{Var}\left(\hat{\mu}\right)$ stabilized.

3. Results and discussion

For the Minnesota calibration dataset, the mean number of observations per species/species group was 66, and the total over all species was 2102. Fits of models produced $0.95 \leq R^2 \leq 0.99$ for the species/species groups and $R^2=0.97$ for the non-specific aggregation of the data. For the Santa Catarina calibration dataset, the mean number of observations per species/species group was 111, and the total over all species was 2119. The larger number of observations per species/species group for Santa Catarina than for Minnesota is because many more species were aggregated into species groups for Santa Catarina than for Minnesota. Fits of models produced $0.89 \leq R^2 \leq 0.98$ for the species/species groups and $R^2=0.92$ for the non-specific aggregation of the data.

For both study areas, the combined effects of model residual and parameter uncertainties on large area estimates were negligible (Table 1). However, for fewer observations per species/species group and/or poorer qualities of fit, the effects might not be negligible. For both study areas, the advantages of using species/species group models rather than the non-specific models were minimal with respect to estimates of both means and standard errors (SE). For the Minnesota study area, the smaller SEs for the non-specific model are attributed to the much larger calibration dataset sizes. The larger SEs for Santa Catarina than for Minnesota are attributed to the smaller Santa Catarina R^2 values and greater forest diversity.

Table 1. Estimates of mean volume per unit area (m^3/ha) and standard errors (SE).

Source of uncertainty		Minnesota				Santa Catarina			
Parameter	Residual	Species		Non-specific		Species		Non-specific	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
No	No	92.60	1.75	90.40	1.70	101.15	2.91	102.02	2.94
No	Yes	92.60	1.75	90.40	1.70	101.15	2.91	102.02	2.95
Yes	No	92.61	1.77	90.41	1.71	101.36	2.95	102.00	2.97
Yes	Yes	92.61	1.78	90.41	1.71	101.36	2.96	102.01	2.97

4. Conclusions

Two primary conclusions may be drawn. First, the negligible effects of residual and parameter uncertainties on the large area volume estimates provide practical justification for the practice of ignoring the effects. Second, if species-specific estimates are not required, the non-specific models produce similar estimates of both means and precision as species/species group models. Additional details for the Minnesota study are reported in McRoberts and Westfall (2014) and for the Santa Catarina study in McRoberts *et al*. (in review).

References

Baskerville, G.L. 1972. "Use of logarithmic regression in the estimation of plant biomass". *Canadian Journal of Forest Research* 2(1): 49-53.

McRoberts, R.E., and Westfall, J.A. 2014. "The effects of uncertainty in model predictions of individual tree volume on large area volume estimates". *Forest Science* 60: 34-43.

McRoberts, R.E., Moser, P., Oliveira, L.Z., & Vibrans, A.C. (in review). "The effects of uncertainty in individual tree volume model predictions". *Canadian Journal of Forest Research*.

Rubin, D.B. 1987. *Multiple imputation in non-response surveys.* Wiley, New York. 287 p.