STOCHASTICALLY GENERATING TREE DIAMETER LISTS TO POPULATE FOREST STANDS BASED ON THE LINKAGE VARIABLES FOREST TYPE AND STAND AGE

Bernard R. Parresol and F. Thomas Lloyd
USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, North Carolina 28804-3454

KEY WORDS: Forest Structure, Imputation, Simulation, Stochastic Modeling, Variable-radius Sampling

Abstract

Forest inventory data were used to develop a stand-age-driven, stochastic predictor of unit-area, frequency-weighted lists of breast high tree diameters (DBH). The average of mean statistics from 40 simulation prediction sets of an independent 78-plot validation dataset differed from the observed validation means by 0.5 cm for DBH, and by 12 trees/h for density. The 40-simulation average of standard deviation, quartile range, maximum value and minimum value differed from the validation dataset, respectively, by 0.3, 1.3, 0.6 and 1.5 cm for DBH, and 10, 42, 29, and 54 trees/h for density. In addition, test statistics were also computed individually for each of the 40 single simulations of the 78-plot validation dataset. In all cases, the test statistics supported the null hypothesis of no difference between simulated and observed DBH lists. When power of these hypothesis test statistics was set to 80%, the calculated minimum detectable differences were still reasonably small at 2.7 cm for mean DBH and 90 trees/h for stocking. Also, the shape and dispersion of simulated mean-DBH/density scatter graphs were similar to the same scatter graph from the observed, validation dataset.

Introduction

Sample inventories are a relatively abundant data source that routinely contains multi-variable forest attribute information useful to GIS analyses. However, the spatial density of plots in sample inventories is, by design, generally too sparse to directly populate databases associated with most GIS layers of forest stand boundaries. Moeur and Stage (1995) used canonical correlation theory on a set of variables common to both the inventory and the existing GIS databases (hereafter referred to as linkage variables) to assign intact, unmodified plot-level data from the inventory database to a GIS database that is linked to a stand polygon layer. This research has a similar goal, except the focus here is on using the linkage variables and the inventory database to parameterize a stochastic predictor of within-polygon variation of unit-area, frequency-weighted breast high tree diameters (DBH). The application potential of this research is illustrated by the methodological role it played in a larger project.

The objective of the larger project was to simulate, over time, the effects of alternative location-specific partial-cut and clearcut-regeneration harvesting on the amount of suitable Red-cockaded Woodpecker (RCW) (*Picoides borealis* Vieillot) habitat available from some forested area. This required predicting frequency-weighted DBH lists for assignment to GIS pixels representing land areas of a given size, and then approximating stand polygons by grouping contiguous pixels that fall within the polygons. These DBH lists were subsequently used to calculate wood product yields from simulated harvesting, to identify stands suitable for RCW nesting sites, to spatially delimit RCW foraging areas around nesting sites, and to serve as input to a forest growth and mortality simulator (Teck et al. 1996). The forest growth and mortality simulator, in conjunction with the GIS database, provided a dynamic and spatially specific description of vegetation charge over a 20- to 30-year period. Alternative, temporal- and location-specific silvicultural treatment scenarios were then applied to the simulated forest structure, and their cumulative affects over time on wood product outputs and the forest-wide suitability RCW habitat were evaluated.

A Generalized Predictor

Model fitting starts by forming relative-size-based subsets made up of individual records of the *p*th ranked DBH (and associated plot-level linkage variables) from each plot. For notation purposes, the subset of trees made up of the largest DBH from each plot was assigned the rank $p = 1$, the subset made up of the second largest trees was assigned rank $p = 2$, etc. The linear model for the relationship between DBH and the vector of linkage variables ($\boldsymbol{\lambda}$) is

$$d_{pi} = \boldsymbol{\beta}_p' \boldsymbol{\lambda}_i + \varepsilon_{pi};\ p = 1, 2, \ldots, q;\ i = 1, 2, \ldots, s \qquad (1)$$

where the linkage variables are made up of plot/stand-level attributes such as stand age, stocking, and site productivity, or other surrogate measures of these variables, d_{pi} is the DBH of the *p*th size-ranked tree from the *i*th plot, $\boldsymbol{\beta}_p$ is a *k*-element vector of regression coefficients for the *p*th subset, ε_{pi} is the error term for the *i*th tree in the *p*th subset (where $E[\varepsilon_{pi}] = 0$ and $\mathrm{Var}[\varepsilon_{pi}] = \sigma_p^2$), *s* is the number of plots in the inventory database that contain *p* sampled trees (*s* decreases as *p* increases), and *q* is the largest

number of trees sampled at a single plot location. The stochastic predictor of the *p*th-ranked DBH is

$$\hat{d}_p = \hat{\boldsymbol{\beta}}_p' \boldsymbol{\lambda}_0 + \left\{ F_{U_1}^{-1} \left[\boldsymbol{\lambda}_0' \left(\boldsymbol{\Lambda}' \boldsymbol{\Lambda} \right)^{-1} \boldsymbol{\lambda}_0 \right]^{1/2} + F_{U_2}^{-1} \right\} \hat{\sigma}_p \quad (2)$$

where $\boldsymbol{\Lambda}$ is the *s* by *k* matrix of linkage variables from the *p*th subset, $\boldsymbol{\lambda}_0$ is the value of the linkage variable(s) for which stochastic predictions are being made, F_U^{-1} is the inverse of the standard normal distribution function, the U_r's are independent uniform random variates on the interval [0,1], and $\hat{\sigma}_p$ is the root mean square error from the fits of Equation (1) to the rank-based subsets. This model was developed for merchantable-sized trees (≥ 12.7 cm) that are sampled with probability proportional to their DBH, therefore the per hectare frequency associated with each $\hat{d}_p$ is

$$n_p = \frac{\text{BAF} \times 40{,}000}{\pi \hat{d}_p^2} \quad (3)$$

where BAF is the basal area factor of the prism used in selecting the sample trees (Grosenbaugh 1952).

The Predictor Used in This Investigation

The specific form of Equation (1) and (2) depends on the linkage variables that are available. The following expression of Equation (2),

$$\hat{d}_p = \hat{\bar{d}}_p + F_{U_1}^{-1} \sqrt{\hat{\mathrm{V}}(\hat{\bar{d}}_p)} + F_{U_2}^{-1} \hat{\sigma}_p \quad (4)$$

assumes that only stand age is available as the linkage variable, with $\hat{\bar{d}}_p = \hat{\beta}_p (a_0 - 3)$, a_0 the stand age of the trees in the DBH list being predicted, the a_i's are the ages from the model-fitting dataset, $\sqrt{\hat{\mathrm{V}}(\hat{\bar{d}}_p)} = \sqrt{((a_0 - 3)^2 / \sum_{i=1}^{s} a_i^2)\, \hat{\sigma}_p^2}$ (Draper and Smith 1966), $\hat{\sigma}_p = \sqrt{\hat{\mathrm{V}}(\varepsilon_{pi})}$, and F_U^{-1}, U_1, and U_2 are as already defined. It was assumed for this dataset that it takes, on the average, three years for a tree to attain 1.37 m in height from a seed, two years from a planted 1-year-old seedling, and three years for spouts.

Estimating Trees per Hectare

The next problem is to decide when to stop adding trees to the list. That is to say, we need to decide the number of trees per hectare. For each of the inventory plots, the total per-hectare number trees were determined. We used Poisson regression to develop predictors for the total trees per hectare as a function of stand age.

$$\ln t_i = \beta_0 + \beta_1 a_i + \varepsilon_i \quad (5)$$

where t_i is the trees per hectare for the *i*th plot, a_i is the stand age on the *i*th plot, the β's are model parameters and ε is model error. Poisson regression differs from linear least squares regression in assuming that at each value of the independent variable (age), the dependent variable (trees per hectare) is Poisson distributed as opposed to being normally distributed. In addition, the dependent variable is assumed to be a count (discrete variable). The Poisson distribution has only one parameter, commonly called mu (μ). In this distribution the mean equals the variance, hence μ = mean = variance. The value of *t* is computed as $\hat{t} = \exp(\hat{\beta}_0 + \hat{\beta}_1 a)$. Of course, all stands of the same age are not identical. Stochastic components are required to mimic natural variability. The stochastic predictor used for trees per hectare ($\hat{t}_{\text{sto}}$) is

$$\begin{aligned} \hat{\mu} &= \exp(\hat{\beta}_0 + \hat{\beta}_1 a + F_U^{-1} S_{\hat{t}}) \\ \hat{t}_{\text{sto}} &= P_{\hat{\mu}} \end{aligned} \quad (6)$$

where $S_{\hat{t}}$ is the standard error of the prediction computed as $S_{\hat{t}} = \sqrt{\mathbf{a}'\mathbf{V}\mathbf{a}}$, the vector $\mathbf{a}'$ is [1 a], the matrix $\mathbf{V}$ is the covariance matrix of the parameter estimates from Equation (5), F_U^{-1} is as previously defined, and $P_{\hat{\mu}}$ is a Poisson random variable with mean and variance equal to $\hat{\mu}$.

To build the tree list you start by predicting trees per hectare for the stand using Equation (6). You then generate a DBH with rank equal to 1 from Equation (4), its per-hectare frequency from Equation (3) and subtract the predicted per-hectare frequency for trees of rank 1 from the total trees per hectare. Continue this process down through the rankings until the predicted total number of trees per hectare is used up, that is, $\sum n_p = \hat{t}_{\text{sto}}$. This is the tree list describing the forest structure associated with that age and forest type.

Stochastic Species Assignment

Besides stand age, plots were classified into one of five forest types as being either loblolly pine, slash pine, longleaf pine, pine-hardwood mix, or hardwood-pine mix. Equations (4) and (5) were fit for each of the five forest type groups. Within a forest type trees were assigned to one of two species groups, either a yellow pine group or a hardwood group, but the approach works for finer species groupings, even for individual

species when the model-fitting dataset is large. Empirical probabilities were computed within each size-based subset. Predicting the species group assignment for each predicted $\hat{d}_p$ is stochastically determined by partitioning the domain [0,1] of a uniform random variable into two intervals, the length of which equals the respective empirical probability for being either a pine or a hardwood. An independent uniform random variate U_p is generated for every stochastically predicted DBH $(\hat{d}_p,\ p = 1,...,q)$, and the species group into which U_p fell was assigned.

Example

The data used in this analysis comes from an inventory of 719 permanent plots located on the Savannah River Site (SRS), a 73,451-h land base (comprised of approximately 6,700 stands), operated by the Department of Energy (DOE). DOE has an inter-agency agreement with the USDA Forest Service, Region 8 to manage the natural resources on the site. As part of the Forest Service's management activities, a GIS layer of stand boundaries is maintained, along with an associated database of stand-level attributes. This database is made up of numerous variables, but only two linkage variables (stand age and forest type) were available when this research was done.

DOE contracted the Forest Inventory and Analysis (FIA) unit of the Southern Research Station, another Forest Service unit, to install the 719-plot inventory used in this research. Plot center points were laid out on an approximate 1000-m grid. The plot design was the same five-subplot layout in use by FIA unit at that time for their regional inventory. At each plot location trees 12.7-cm or larger in DBH were sampled at each of the five equidistantly spaced subplots with probability proportional to their DBH, using an 8.61-factor prism. Circular, fixed-radius, 0.00135-h plots were also established at each subplot center point for sampling trees less than 12.7 cm in DBH, but only the trees with DBH equal to or greater than 12.7 cm were used in the investigation. Also, the data from the five subplots were pooled, resulting in an effective prism factor of 1.72. The inventory plots were established and initially measured in the period 1986-88. Approximately half of them (386 plots) were remeasured in 1992. The data used here came from the 1992 inventory.

Results and Discussion

This analysis used plots from the 1992 inventory that were in the loblolly pine forest type. This 157-plot dataset was randomly split into a 78-plot validation subset, and a 79-plot model-fitting subset. The 50/50-split method of validation adequately serves the illustrative purpose of this investigation.

The largest number of trees sampled on a single plot in the model-fitting subset was 21 (the value of q in Equation (1)). Equations (4) and (5) were fit using this 79-plot subset, and validated against the 78-plot subset. The least squares estimates of $\hat{\beta}_p$ from Equation (4) are listed in Table 1 for each of the 21 rank-based data groupings. As expected, the magnitudes of $\hat{\beta}_p$ and $\hat{\sigma}_p$ decrease with increasing p. The performance of the stochastic prediction process was tested using these estimates.

Table 1. Parameter estimates in Equation (4) used to predict the 21 DBH-based order statistics.

Rank of order statistic	Number of data points	Estimate of slope $(\hat{\beta}_p)$	Root mean square $(\hat{\sigma}_p)$ *cm*
1	79	1.146628	12.324
2	79	1.037074	11.677
3	79	0.966745	11.337
4	79	0.903008	10.379
5	77	0.868421	9.382
6	76	0.821200	9.062
7	73	0.778998	9.326
8	69	0.738786	9.419
9	64	0.719658	8.810
10	59	0.691616	8.578
11	54	0.634936	8.153
12	44	0.629256	7.533
13	37	0.580774	7.799
14	26	0.570619	6.520
15	24	0.522966	6.587
16	18	0.471870	6.155
17	15	0.445366	4.414
18	11	0.425769	5.045
19	8	0.423444	4.324
20	6	0.380717	3.927
21	4	0.298526	1.807

Table 2 compares means of 40 simulation runs (each run reusing the stand ages from validation subsets as input) with the observed means from the validation dataset. These mean comparison statistics include two measures of central tendency (mean and median) and four measures of population dispersion (standard deviation, quartile range, and average minimum and maximum values). The average stochastically predicted means from the 40 simulations differed from the observed validation subsets means by 0.5 cm in DBH, 12 trees/h in density, and 0.0 m^2/h in basal area. The same 40-simulation averages for the

population-level dispersion statistics of standard deviation, quartile range, maximum value and minimum value differed from the observed validation subset means by 0.3, 1.3, 0.6 and 1.5 cm for DBH, and 10, 42, 29, and 54 trees/h for density, respectively.

Table 2. The average quadratic mean DBH, trees/h, and basal area over all 40 simulation sets compared with the observed averages from the validation dataset.

	Variable	Mean	Median	Standard deviation
		cm		
Observed	$\hat{d}_{\text{quad}}$	28.2	27.2	6.1
Predicted		27.7	26.9	5.8
		number/ha		
Observed	$\hat{t}_{\text{sto}}$	376	334	195
Predicted		388	343	205
		m^2/ha		
Observed	ba	20.7	19.8	6.4
Predicted		20.7	20.2	6.8
	Variable	Quartile range	Min.	Max.
		cm		
Observed	$\hat{d}_{\text{quad}}$	8.9	15.7	44.2
Predicted		7.6	16.3	45.7
		number/ha		
Observed	$\hat{t}_{\text{sto}}$	274	82	974
Predicted		232	111	1028
		m^2/ha		
Observed	ba	10.3	8.6	37.9
Predicted		9.0	7.5	36.9

The power computed using these averages was low (8% for $\hat{d}_{\text{quad}}$ and 7% for $\hat{t}_{\text{sto}}$), which is not surprising given closeness of the means. When the power of the comparison is set to 80%, the minimum detectable differences were 2.7 cm for $\hat{d}_{quad}$ and 90 trees/h for $\hat{t}_{\text{sto}}$. Expressed as a percentage, this says that the probability of correctly accepting the null hypothesis is high for true differences no more than 10% (2.7/27.7) of d_{quad} and 24% (90/376) of t_{sto}. Although this suggests considerable unexplained variation, the results are not bad considering that only stand age was used as a linkage variable, and furthermore, that thinning was intensively practiced on this forest, further compromising stand age as a surrogated for mean size.

In addition to investigating the average performance of repeated simulations of the validation subset, the performance of individual simulation prediction sets of the validation subsets was also tested. Part of this analysis consisted on visually comparing $(\hat{t}_{\text{sto}},\hat{d}_{\text{quad}})$ scatter graphs. Figure 1 displays the scatter graph of $(\hat{t}_{\text{sto}},\hat{d}_{\text{quad}})$ points from the validation dataset, along with the best, median, and worst fit individual prediction sets, as determined by a Euclidean distance measure. This measure was calculated as the sum of the distances between each of the 78 predicted and observed $(\hat{t}_{\text{sto}},\hat{d}_{\text{quad}})$-points. The scatter graphs of the simulations with the smallest (best fit), the median, and the largest value (worst fit) cumulative distance are plotted. The result supports the assertion that the model is performing acceptably for this key relationship.

Reineke (1933) showed how the logarithmic transformations of $\hat{d}_{\text{quad}}$ and $\hat{t}_{\text{sto}}$ linearizes the moving average relationship between them. Fisher's normalizing transformation, $0.5\times\ln((1+r)/(1-r))$, was used to test for differences between predicted and observed correlations between $\ln(\hat{d}_{\text{quad}})$ and $\ln(\hat{t}_{\text{sto}})$ (Steel and Torrie, 1980). The observed correlation coefficient is –0.78. Only two of the 40 test statistics were significant when using a Type I error rate of 0.05. This is the expected rejection rate for this error rate under the null hypothesis. Examining these statistics in terms of Type II error rate of 0.2 (80% power) shows that when the simulated average correlation is larger than the observed correlation (for sample size equal to 78), that the absolute value of minimum detectable difference is 0.09. The minimum detectable difference is 0.14 for the same power of test when the simulated average correlation is smaller than the observed. The simulated correlation estimates for the 38 non-significant comparisons range from –0.67 to –0.84.

The moving mean relationship for individual prediction sets was also examined by investigating the similarity of the regression relationship between $\ln(\hat{d}_{\text{quad}})$ and $\ln(\hat{t}_{\text{sto}})$. Forty analyses of covariance were performed, testing for differences in the slope and intercept. The analysis used $\ln(\hat{d}_{\text{quad}})$ as the dependent variable, and the $\ln(\hat{t}_{\text{sto}})$, an indicator variable I distinguishing between the predicted and validation values, and their interaction $\text{I}\times(\ln(\hat{t}_{\text{sto}}))$, as the independent variables. The null hypothesis of equal slopes was rejected 2 times in 40 at the 0.05 alpha level, and 3 times in 40 for the intercept. Again this is the expected rejection rate under the null hypothesis.

Similar results were obtained from comparing means for individual prediction sets with observed means from the validation dataset. Again, 38 of the 40

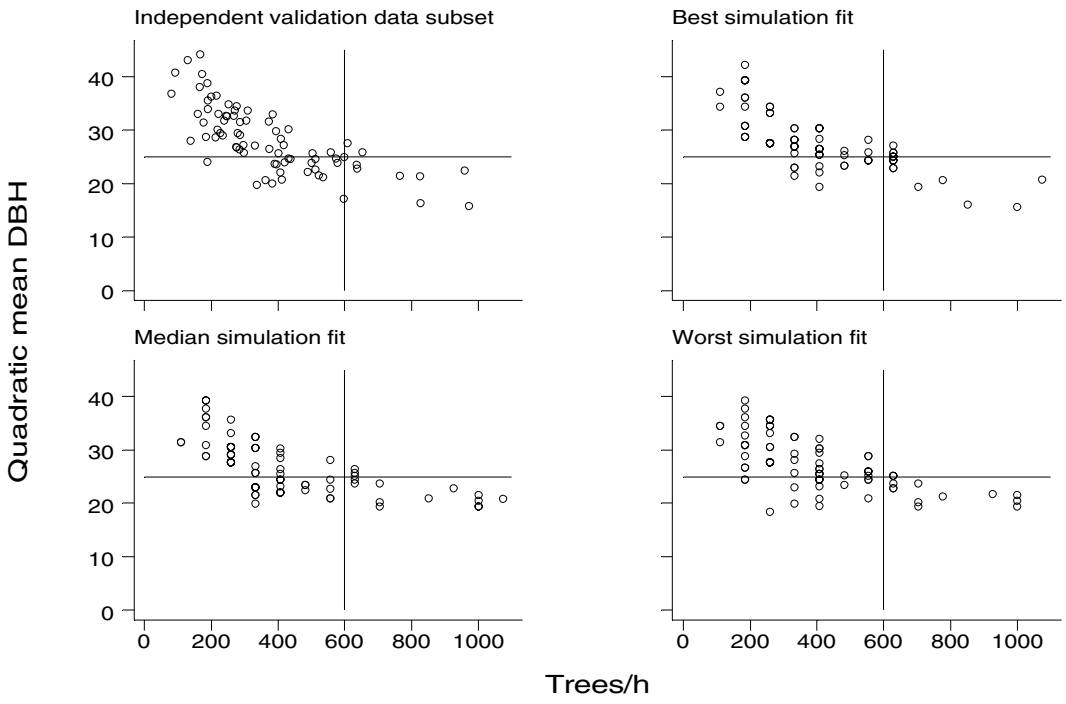

Figure 1. Scatter plots of the d_{quad}-over-t_{sto} for the best fit, median fit, and worst fit of individual simulation runs based on the cumulative Euclidean distances between the 78 predicted and observed $(\hat{t}_{sto}, \hat{d}_{quad})$ -pairs.

values of Hotelling's T^2 test statistic were non-significant (Hotelling, 1931; StataCorp, 1999). Also, test statistics for the null hypotheses of no difference between the standard deviations from individual simulation prediction sets and the observed standard deviation (Ostle, 1963; StataCorp, 1999) supported similar assertions.

Conclusions

This research presents a different application of traditional sample inventory data, designed to satisfy the data needs of new analytical tools. The objective was to use an inventory database to produce stochastic, regression-based predictions of unit-area variation in unit-area, frequency-weighted DBH lists for the purpose of modeling variation in forest structure within stand polygons in a GIS database. Because of data constraints in the databases for which the methodology was developed, this illustration used only stand age as a linkage variable. Improved performance would be expected for applications that have additional linkage variables available. Given that stand age is only a moderately strong surrogate of mean size, the model still did an accurate job of predicting mean $\hat{d}_{quad}$, $\hat{t}_{sto}$, and ba. However, the precision of the predictions was weak. This was due to the fact that stand age is at best only a moderately strong surrogate for mean size, and that its strength is further compromised when stands are thinned, as they were in this application.

The methodology assumes that variation between stands of the same age is a plausible estimate of variation within a stand. This assumption needs more rigorous examination. One possibility would be to use subplot information (when it is available as it was in this sample design) as a better data source for estimating within-stand variation. Unfortunately, the prism factor (8.61 m^2/h) in this sample design was deemed too large to produce stable results, and so the subplot data were pooled.

Further investigation is needed into testing the effect of alternative ways of reducing the number of predicted tree diameters $(\hat{d}_p, p = 1, 2, ..., i, ..., q)$ from the maximum tree list down to the i needed to predict the target trees per hectare. In this investigation, the list was simply truncated from the bottom up, that is, dropping the prediction $\hat{d}_p$ with the largest value of p, then the next largest p, etc. Some type of a probability-based elimination approach might be tested. Finally, investigations are needed into developing stochastic predictors that use additional linkages variables.

Acknowledgments
The authors extend their appreciation to John Blake, Don Van Blaricom, and Kay Franzreb for their support and helpful comments on earlier drafts of this manuscript. This research was funded by the Department of Energy, Savannah River Site, Aiken, SC.

Literature Cited

Draper, N.R. and H. Smith. 1966. Applied Regression Analysis. John Wiley & Sons, Inc., New York, New York. 407 p.

Goldberg, S. 1960. Probability-An Introduction. Prentice-Hall, Mathematics Series. Englewood Cliffs, New Jersey. 322 p.

Grosenbaugh, L.R. 1952. Plotless timber estimates – new, fast, easy. J. For. 50: 33-37.

Hotelling, H. 1931. The generalization of Student's ratio. Ann. Math. Stat. 2: 360-378.

Moeur, M. and A.R. Stage. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For. Sci. 41(2): 337-359.

Ostle, B. 1963. Statistics in Research. Iowa State University Press, Ames, Iowa. 585 p.

StataCorp. 1999. Stata Statistical Software: Release 6.0. College Station, TX: Stata Corporation.

Steel, R.G.D. and J.H. Torrie. 1980. Principles and procedures of statistics, a biometric approach. McGraw-Hill Book Co., New York, New York. 633 p.

Teck, R., M. Mouer, and B. Eav. 1996. Forecasting ecosystems with the Forest Vegetation Simulator. J. For. 94(12): 7-10.