Constructing a virtual forest: Using hierarchical nearest neighbor imputation to generate simulated tree lists

Kevin R. Gehringer and Eric C. Turnblom

Abstract:

A nearest neighbors method for generating simulated tree lists has been developed. The method employs an implicit two-scale hierarchy to incorporate information from a coarse scale representing the distribution of stand attributes across a region and a fine scale representing the distribution of tree attributes within a stand. The tree list generation method was implemented and tested using data from untreated, naturally regenerated and planted forests in western Oregon, western Washington, and southern British Columbia west of the Cascade Mountains. Simulated tree lists were generated from stand scale attributes for each of the actual tree lists in the data. Distributions of stand scale and tree scale attributes were estimated and used to compare the simulated and actual tree lists. At the stand scale distributions of quadratic mean diameter and average height for the simulated and actual stands were in very good agreement, having approximately 98% of their probability mass in common for each attribute. At the tree scale, comparisons of the distributions of diameter at breast height, height, and species composition between the simulated and actual stands were more variable, with approximately 84% of the simulated stands identified as statistically similar to their respective actual stands.

1. Introduction

Modern forest management systems typically use tree lists, minimally a set of compatible diameter at breast height (DBH) and height measurements with an indication of tree species, to describe a stand whose development is to be simulated or used to estimate forest resources for planning (Mitchell, 1975, Wykoff et al., 1982, Wykoff, 1986, Donnelly, 1997, Hann et al., 1997). Tree lists obtained from measurements of existing plots or stands, and for which complete individual tree measurements are available to define the initial condition, are highly advantageous, and their use allows forest growth simulations and other analyses to begin with the ground truth. In many situations complete individual tree measurement data are not available and the ability to generate a realistic, simulated tree list that is

Kevin R. Gehringer.[1] Biometrics Northwest LLC, Redmond, WA, 98053
Eric C. Turnblom. School of Environmental and Forest Sciences, University of Washington, Seattle, WA 98195-2100

[1]Corresponding author (e-mail: krg@biometricsnw.com).

unknown **99**: 1–26 (2014) DOI: 10.1139/*Zxx-xxx* Published by NRC Research Press

representative of an actual stand, using a small number of stand attributes, e.g., site index, age, stand density, and average tree size, is highly desirable.

Procedures for generating simulated tree lists using an implicit two-scale relationship and a nearest neighbors algorithm are described. The procedures represent a novel imputation framework that has been specialized for a two-scale, hierarchical, nearest neighbors tree list generation (HNNTLG) problem. The implicit two-scale relationship is represented as two nested spatial scales: a coarse scale for the distribution of stand attributes across a geographic region and a coupled fine scale for the distribution of individual tree attributes within a stand.

The design of the HNNTLG procedures and development of software implementing them (Gehringer, 2001, Gehringer and Turnblom, 2001) were guided by the following criteria. (1) Simulated trees should be generated as multidimensional objects directly. (2) Simulated trees should be physically realizable, and the tree list generation methodology should only generate *realistic* tree attributes. (3) The stand representation should be flexible enough to allow arbitrary tree size distributions and species composition. (4) The addition of new data should be easy to perform and should not significantly change the global classification used to identify similar stands and generate tree lists, except possibly by the creation of a new stand class, to ensure consistency as new data are added. (5) A pseudorandom number generation framework should be used to generate a simulated stand or tree list, permitting the generation of independent, but similar, tree lists for stands whose attributes map to the same class.

2. Methods

Fundamental to pseudorandom number generation is selecting a suitable representation for a probability density function (PDF) $f(x)$ or a cumulative distribution function (CDF) $F(x) = \int_{-\infty}^{x} f(y)dy$ from which random values may be generated. A suitable representation for $f(x)$ or $F(x)$ is one that can represent the types of distributions encountered in practice (Temesgen, 2003, Temesgen et al., 2003, LeMay and Temesgen, 2005) while facilitating the development of an algorithm to generate random values from the underlying distribution.

2.1. Representing the HNNTLG distribution

A forested region may be represented as a mosaic of more or less distinct forest stands that are distinguished by their physical attributes and their dominant vegetation types, e.g., stand age, site index, soil characteristics, site quality, and the numbers, sizes and species of trees and other vegetation. Each stand may be represented by a multivariate PDF $f(x)$ describing the joint distribution of its stand and tree scale attributes as a vector x. A mixture distribution (Titterington et al., 1985, Borders and Patterson, 1990, Biging et al., 1994) provides a natural representation for such a joint distribution across a forested region as in Equation 1,

$$f(x) = \sum_{i=1}^{N} \alpha_i f_i(x) \tag{1}$$

where $x = (x_1, x_2, \ldots, x_d)$ is a d-dimensional vector of coupled stand and tree attributes, N is the number of forest stands, $f_i(x)$ are the joint PDFs for the stand and tree scale attributes of the individual forest stands, and $\alpha_i > 0$ are weights giving the proportion of the landscape represented by each stand, with $\sum_{i=1}^{N} \alpha_i = 1$, making f itself a PDF.

A distribution of tree (fine) scale attributes describing the structural conditions within a forest stand is ultimately required to generate a tree list that is representative of the stand (coarse) scale attributes. Let $x = (x^s, x^t)$ split the d-dimensional attribute vector x into its stand scale $x^s = (x_1^s, x_2^s, \ldots, x_{d_s}^s)$ and tree scale $x^t = (x_1^t, x_2^t, \ldots, x_{d_t}^t)$ attributes, with $d_s > 0$, $d_t > 0$, and $d_s + d_t = d$. Using this notation, the mixture density $f(x)$ and its component PDFs $f_i(x)$ may be written as $f(x^s, x^t)$ and

$f_i(x^s, x^t)$, respectively, making the stand scale and tree scale attributes explicit. The distribution of tree attributes for any component density $f_i(x^s, x^t)$ in the mixture density is given by its marginal density of tree attributes, as in Equation 2.

$$f_i^{\text{tree}}(x^t) = \int f_i(x^s, x^t)\, dx^s \tag{2}$$

A distribution of tree attributes can, theoretically at least, be determined for any component density $f_i(x^s, x^t)$ and subsequently used to generate simulated tree attribute vectors $\hat{x}^t$. To bridge theory to practice two difficulties must be overcome. First, the overall mixture density $f(x^s, x^t)$, its weights α_i, and its component densities $f_i(x^s, x^t)$ are unknown. This, it will turn out, is not a significant obstacle; a direct representation of the mixture density is not necessary. Second, a procedure to index the component density functions $f_i(x^s, x^t)$ and their associated marginal densities of tree attributes $f_i^{\text{tree}}(x^t)$ is needed, and should be derived directly from the stand scale attribute vectors x^s to select component density functions $f_i(x^s, x^t)$ from which simulated tree scale attributes may be generated.

An index of component density functions $f_i(x^s, x^t)$ may be derived by first considering the support sets $S_i = \{(x^s, x^t) \mid f_i(x^s, x^t) > 0\}$ generated by the component density functions, where $S = \bigcup_{i=1}^{N} S_i$ is the support set of the mixture density. Unique index points representing subsets of the stand scale attribute space may be obtained by directly partitioning the d_s-dimensional subset of stand scale attribute vectors x^s in the support set S, $S|_{x^s}$, and then reformulating the mixture density and the indexing problem using the specified partition and a conditioning argument.

Let $B = \{B_1, B_2, \ldots, B_M\}$ be a partition containing the d_s-dimensional subset of stand scale attributes $S|_{x^s}$ within the support S of the mixture density $f(x^s, x^t)$. By definition, the subsets B_m, $m = 1, 2, \ldots, M$ are disjoint , $B_i \cap B_j = \emptyset$ for $i \neq j$, and their union is assumed to contain the support for the stand scale attributes $S|_{x^s} \subset \bigcup_{m=1}^{M} B_m$. Component density functions for a mixture density having the partition sets B_m as their corresponding support sets are obtained by conditioning the mixture density function $f(x^s, x^t)$ using the sets B_m in the partition B to compute conditional density functions

$$f_{B_m}(x^s, x^t) = \begin{cases} \frac{1}{P_{B_m}} f(x^s, x^t) & \text{for } x^s \in B_m \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $P_{B_m} = \text{Prob}(\{x \mid x^s \in B_m\})$ is the probability that a point (x^s, x^t) has its stand scale attributes in the set B_m. The partition conditioned mixture density

$$f_B(x^s, x^t) = \sum_{m=1}^{M} \beta_m f_{B_m}(x^s, x^t) \tag{4}$$

is obtained by summing the partition conditioned component density functions using mixing weights $\beta_m = P_{B_m}$. Note that $f_B(x^s, x^t) = f(x^s, x^t)$ for all values of $x = (x^s, x^t)$. A distribution of tree attributes for any set B_m in the partition is determined by computing the marginal distribution of tree attributes for that set as in Equation 5,

$$f_{B_m}^{\text{tree}}(x^t) = \frac{1}{P_{B_m}} \int_{\{x \mid x^s \in B_m\}} f(x^s, x^t)\, dx^s \tag{5}$$

and an index for the partition conditioned component densities may be obtained by choosing a representative index point $b_m = (b_{m1}, b_{m2}, \ldots, b_{md_s})$ from each of the sets B_m in the partition B.

Changing the focus from the component density functions and their support sets to a partitioning of the stand scale support of the mixture density, and the resulting partition conditioned mixture density, has made the indexing problem much simpler. Further, relationships between the stand scale attributes

and the tree scale attributes implicitly defined by each component density function $f_i(x^s, x^t)$ have been retained by the partition conditioned mixture density: it blends the tree scale attribute distributions from all component density functions whose stand scale marginal support sets $S_i|_{x^s}$ have nonempty intersection with a partition set B_m in the partition B. This is a complicated way of stating the assumption that forest stands having similar stand (course) scale attributes are also similar at the tree (fine) scale, provided that enough relevant stand scale attributes are used to describe them, or alternatively, if forest stands may be classified using stand scale attributes, then the distributions of tree scale attributes for stands assigned to the same class are statistically similar. The partition B simply provides pigeon holes B_m for the classification of stands like the bins of a histogram.

2.1.1. Including discrete-valued attributes

The inclusion of discrete-valued attributes within the vector of stand scale and tree scale attributes has not been addressed. The notation used implies that attribute values are continuous. This was done for notational convenience given the understanding that discrete attribute values act as conditioning events that partition the continuous attributes to generate continuous probability density functions. Discrete attributes simply increase the number of terms in a mixture density representation.

2.2. HNNTLG implementation

A nonparametric representation for the regional mixture distribution of stand scale and tree scale attributes and their respective probability density functions was used in the HNNTLG implementation to allow the data to *speak for itself* in determining the shape of the mixture distribution $f(x^s, x^t)$ and the partition conditioned mixture distribution $f_B(x^s, x^t)$. The implementation has two stages. The first stage uses the stand and tree scale attributes from sampled stands to build a mapping associating the tree attributes from the sampled stands with the partition sets B_m for a stand scale partition B. The second stage uses the mapping created by the first stage with a specified stand scale attribute vector y^s to identify similar partition sets whose associated tree attributes are then used to define a canonical stand that is used to generate simulated trees. Differences in the treatment of discrete and continuous attribute values are described as they occur.

2.3. Partitioning the stand scale attributes

Let $h = (h_1, h_2, \ldots, h_{d_s})$ be a vector of bin widths for a d_s dimensional vector of stand scale attributes x^s, some of which may be discrete-valued, where $h_j = 0$ for discrete-valued attributes and $h_j > 0$ for continuous attributes. Define b_m as the center of a bin having edge lengths h by setting $b_{mj} = x_j^s$ if x_j^s is discrete or setting $b_{mj} = h_j \left(\lfloor \frac{x_j^s}{h_j} \rfloor + \frac{1}{2} \right)$ if x_j^s is continuous, where $\lfloor x \rfloor$ is the floor function returning the largest integer less than or equal to x. A partition of the stand scale attribute space can then be generated as a sequence of multidimensional bins B_m where $x^s \in B_m$ if $x_j^s = b_{mj}$ for discrete attribute values and $x_j^s \in [b_{mj} - \frac{h_j}{2}, b_{mj} + \frac{h_j}{2})$ for continuous attribute values. The bin centers b_m uniquely identify each bin and are the index points.

2.4. Computing stand scale similarity scores

The similarity between two stand scale attribute vectors x^s and y^s for the nearest neighbor selection was computed using a bin width weighted Euclidean distance between the vectors as in Equation 6, with weights $w_j = 10000$ if x_j^s is discrete or $w_j = \frac{1}{h_j^2}$ if x_j^s is continuous, $j = 1, 2, \ldots, d_s$.

$$S(x^s, y^s) = \left(\sum_{j=1}^{d_s} w_j (x_j^s - y_j^s)^2 \right)^{\frac{1}{2}} \tag{6}$$

The weight of 10000 was used for all discrete-valued stand attributes to give a clear indication of differences between discrete attribute values. The weighting of the continuous stand scale attributes was chosen to provide a measure of the distance between two attribute vectors in terms of the bin widths h_j, standardizing the distance measures for the continuous attributes.

2.5. Building the stand scale to tree scale mapping

The algorithm for the stand scale to tree scale mapping is best described using a collection of data sets from sampled forest stands rather than by the stand scale and tree scale attribute vectors x^s and x^t and their distributions. Some additional notation to describe a sample data set and a collection of sample data sets will prove useful. Let $X = (X^s, X^t)$ be a sample obtained from a forest stand, where X^s represents an observed d_s-dimensional vector of stand attributes and X^t represents a matrix of d_t-dimensional vectors of observed tree attributes X_l^t, $l = 1, 2, \ldots, n$ obtained from n sampled trees. The vector of stand scale attributes X^s was assumed to have been drawn from the distribution $f_{B_k}^{\text{stand}}(x^s)$ for some partition set B_k, $X^s \sim f_{B_k}^{\text{stand}}(x^s)$, and the vectors of observed tree attributes X_l^t were, similarly, assumed to have been drawn from $f_{B_k}^{\text{tree}}(x^t)$, $X_l^t \sim f_{B_k}^{\text{tree}}(x^t)$.

Let $X_i = (X_i^s, X_i^t)$, $i = 1, 2, \ldots, N$ be the observed stand and tree attributes from a collection of N sampled forest stands, with X_i^t containing attributes from n_i sampled trees X_{il}^t, $l = 1, 2, \ldots, n_i$. The type of stand, e.g., planted, natural, or thinned, etc., the dimensions of the stand and tree scale attribute vectors, d_s and d_t, respectively, and the particular attributes measured are assumed to be the same for all N sampled stands. Using the partition defined above, a mapping between the stand scale bin centers b_m and the measured tree attributes X_i^t for each sampled stand is now straightforward to describe. For each bin B_m in the partition B, keep a list $T_m = (X_{i_1}^t, X_{i_2}^t, \ldots, X_{i_{n_m}}^t)$ of the $N_{T_m} = \sum_{i=i_1}^{i_{n_m}} n_i$ tree attribute vectors from the n_m samples whose stand scale attributes $X_{i_1}^s, X_{i_2}^s, \ldots, X_{i_{n_m}}^s$ map to the partition bin center b_m. With this mapping it is then easy to move from a vector of stand attributes x^s to a partition bin center b_m, and then to the tree attributes T_m associated with that bin. The tree attributes associated with a partition bin define a canonical stand whose joint distribution of tree attributes is representative of the forest conditions within that particular bin.

2.6. Generating simulated tree attributes

A simulated vector of tree attributes $\hat{y}^t$ that is representative of trees within forest stands similar to a specified vector of stand attributes y^s is generated in two steps. First, the $k \leq K_{\max}$ partition bins $B_{m_1}, B_{m_2}, \ldots, B_{m_k}$ having bin centers $b_{m_1}, b_{m_2}, \ldots, b_{m_k}$ with similarity scores $S(y^s, b_m) < S_{\max}$ are selected, and the tree attribute vectors $T_{m_1}, T_{m_2}, \ldots, T_{m_k}$ associated with these partition bins are concatenated to form a canonical tree list T containing $N_T = \sum_{m=m_1}^{m_k} N_{T_m}$ tree attribute vectors. For convenience, the tree attribute vectors in the canonical tree list T are relabeled as $T_l = (T_{l1}, T_{l2}, \ldots, T_{ld_t})$, $l = 1, 2, \ldots, N_T$. Second, the tree attribute vectors T_l, $l = 1, 2, \ldots, N_T$ in the canonical tree list T are partitioned into their continuous and discrete components, T_l^C and T_l^D, respectively, to obtain canonical lists of their continuous and discrete tree attributes T^C and T^D. A combination of a bootstrap method (Efron, 1982, Efron and Tibshirani, 1998) for the discrete attributes and the SIMDAT algorithm (Taylor and Thompson, 1986, Thompson, 2000) for the continuous attributes was then used to compute a simulated tree scale attribute vector $\hat{T} = (\hat{T}^C, \hat{T}^D)$ using the following algorithm for each desired tree. (1) Generate a random integer r, $1 \leq r \leq N_T$ to identify a reference tree $T_r = (T_r^C, T_r^D)$. (2) Assign the discrete attribute values from the reference tree to the simulated tree $\hat{T}^D = T_r^D$. (3) Generate the continuous attribute vector $\hat{T}^C$ for the simulated tree using the SIMDAT algorithm and the continuous attributes from the reference tree T_r^C and its K_T nearest neighbors from the continuous tree attributes in the canonicalcanonical tree list T^C.

Values for the maximum number of partition bins $K_{\max}$ and the maximum similarity score $S_{\max}$ may vary, depending on the application, but values of $K_{\max} = 5$ and $S_{\max} = 5.0$ were used here to select

the partition bins and create the canonical tree lists. Fewer than $K_{\max}$ partition bins may be selected depending on the value chosen for $S_{\max}$ and the sparsity of the partition bin centers near the desired stand attribute vector y^s. The maximum similarity score $S_{\max}$ filters out partition bins having discrete stand scale attributes that do not agree with those in the stand attribute vector y^s and partition bins that are far from y^s. A value of $K_T = 10$ has been shown to give good results with the SIMDAT algorithm (Taylor and Thompson, 1986, Thompson, 2000), and this value was used here. If fewer than K_T trees were available, all of the available trees were used to generate the vector of simulated tree attributes $\hat{T}^C$. As is the case for the sample size and the bin width or number of bins for histograms and a variety of other nonparametric probability density estimators, an asymptotic dependence exists between K_T, the smoothing or averaging parameter, and the sample size N_T, requiring K_T to increase as N_T increases (Silverman, 1986, Gehringer, 1990, Thompson and Tapia, 1990, Gehringer and Redner, 1992, Redner and Gehringer, 1994, Redner, 1999), but in practice, a fixed value of K_T may be used to good effect (Taylor and Thompson, 1986, Thompson, 2000).

A preprocessing step may be applied to the tree attribute data in the canonical tree list T prior to generating simulated trees. The preprocessing step may be used to fill in missing values in the tree attribute data, to remove outliers, or to filter the tree attribute data to better approximate a particular set of stand scale conditions, such as species composition. A postprocessing step may also be applied to the simulated tree attributes $\hat{y}^t$ that are generated, individually or as a group, for example to add additional variability or to compute derived tree scale quantities that were not simulated, e.g., crown width and crown ratio, or to compute additional stand scale attributes.

2.7. Validation and goodness of fit testing

To test the two-scale HNNTLG procedures just described a suite of programs, the tree list generation database (TLGDB) (Gehringer and Turnblom, 2001, Gehringer, 2001), was developed for the Stand Management Cooperative (SMC) at the University of Washington (Maguire et al., 1991). The TLGDB was designed to generate simulated tree lists for pure or mixed species stands having Douglas-fir (*Pseudotsuga menziesii*) and western hemlock (*Tsuga Heterophylla*) as dominant or codominant species, with other associated tree species found throughout Oregon, Washington, and southern British Columbia west of the Cascade Mountains. Two treatment regimes for managed stands were implemented in the TLGDB: untreated stands and thinned stands. Only untreated stands are considered here.

The stand and tree scale attributes in the TLGDB are presented in Table 1 along with their types, discrete or continuous, and the partition bin widths that were used for the continuous stand scale attributes. At the stand scale, stand type was the only discrete attribute used, and stand density measured as trees per hectare (TPH), QMD, and mean tree height (H) were the continuous attributes used to generate the partition. Five stand types were used: pure Douglas-fir or western hemlock, having at least 75% of the basal area in the dominant species; Douglas-fir or western hemlock dominant, having at least 50% of the basal area in the dominant species; and mixture, having less than 50% of the basal area as Douglas-fir or western hemlock, or having a different dominant species. At the tree scale, diameter at breast height (DBH) and height were continuous attributes measured or estimated for each tree, and species was the discrete attribute.

A simulated sample stand contains compatible DBH, height, and species values generated for each simulated tree in an actual sample. Values for the stand scale attributes total age, site index, stand origin, and plot size were copied from the description of a desired sample stand. Stand type was implicitly defined by the generated tree list. A preprocessing step that restricted pure stands to the dominant species in the canonical tree list was implemented to allow the generation of 100% pure Douglas-fir or western hemlock stands. A postprocessing step that added a small amount of random variability u to the simulated tree heights, was also implemented. The random variable u was generated from the interval $(-2\,\hat{H}_{\text{MAI}}, 2\,\hat{H}_{\text{MAI}})$ where $\hat{H}_{\text{MAI}}$ was the mean annual height increment for a simulated tree.

2.7.1. Data

Regional stand measurement data were provided by a variety of organizations that are listed in the acknowledgements. Given the number of data sources, a myriad of data collection and sampling strategies were undoubtedly employed, with differing assumptions and methodologies. A variety of unknown height-diameter relationships were also used to estimate missing tree heights. Even if complete descriptions of the data collection histories, sampling protocols, and other statistical procedures had been available, it would have been logistically unrealistic to reconcile discrepancies. Therefore no attempt was made to address the statistical compatibility of the sampling strategies or other procedures that were employed to obtain the data sets used. The data sets were assumed to be compatible and their tree lists were used *as is* for this analysis. Several straightforward data screening procedures were employed: only trees from sample plots that were at least 0.0405 hectares were used, and plots having undefined or missing values for key attributes such as stand origin, site index, etc., were omitted from the analysis.

The screened data consisted of DBH and height measurements with an indication of tree species for 573,036 individual trees from a total of 5209 untreated planted or naturally regenerated sample plots distributed throughout the region of interest. The breakdown by stand type is given in Table 2. Pure Douglas-fir stands represented 65.3% of the sample plots, with pure western hemlock stands representing 15.4%, Douglas-fir dominant stands accounting for 13.3%, and with western hemlock dominant and mixed stands forming the remaining 6.0%. Only standing live trees were included and they were used with their expansion factors to compute QMD, mean height, and TPH values from each included sample stand for the stand scale partition. The 5209 sampled stands produced 2865 unique partition bins, and Table 3 provides a summary, by stand type, of the continuous stand scale attributes derived from the partition bin centers b_m.

2.7.2. Goodness of fit methods and criteria

To assess the performance of the HNNTLG procedures a simulated sample stand was generated for the stand scale attributes of the actual sampled stands used to populate a TLGDB. In the simulated stands only the sample plot size and number of sample trees to be generated were matched. No attempt was made to match the numerical attributes of the actual stands in this analysis other than using the preprocessing step for 100% pure stands. The integrated absolute error (IAE), defined in Equation 7, was the nonparametric goodness of fit (GOF) statistic used to compare the DBH, height, and species composition distributions for each pair of actual and simulated sample stands, and to compare the regional distributions of QMD and average height across all actual and simulated sample stands.

$$\text{IAE} = \int_{-\infty}^{\infty} |f(x) - g(x)|\, dx. \tag{7}$$

The IAE computes differences in probability mass between two PDFs f and g, producing values in the interval $[0, 2]$, with zero indicating that f and g are identical and two indicating that the functions have no overlap. The IAE value automatically incorporates distribution shape, and it has a direct analog for discrete-valued PDFs defined over distinct values, x_i, $i = 1, 2, \ldots, n$: IAE $= \sum_{i=1}^{n} |f(x_i) - g(x_i)|$. This allows the same GOF statistic to be used for continuous and discrete-valued density functions, an important consideration here since both continuous and discrete-valued distributions are present at the tree scale. The IAE has been used to compute an index comparing diameter distributions (Borders and Patterson, 1990, Reynolds et al., 1988), it has been used effectively to compare nonparametric estimates of probability density functions (Gehringer, 1990, Gehringer and Redner, 1992), and it may be the natural GOF statistic for comparing probability density functions (Devroye and Gyorfi, 1985).

An IAE cutoff value is needed to differentiate between PDFs that are similar or different, playing a role similar to the critical value determined from the α-level in a classical goodness of fit or hypothesis

test. A simulation study to approximate the distribution of IAE values under the null hypothesis of equal PDFs was performed to determine a cutoff value using the standard normal distribution and the range of sample sizes from the sampled stands. Based on the simulation, an IAE cutoff value of 0.50 was selected; this allows a maximum difference of 25% of the total probability mass for the PDFs of similar simulated and actual samples.

For each pair of actual and simulated stands IAE values were computed for the DBH and height distributions and the species composition. The IAE values for DBH and height were computed by estimating the underlying PDFs from each actual or simulated sample stand (Gehringer, 1990, Gehringer and Redner, 1992, Redner and Gehringer, 1994, Redner, 1999) and then numerically integrating Equation 7 for each attribute. The IAE values for species composition were computed directly using empirical estimates of the proportions for each tree species in the samples. An IAE value greater than 0.50 for an attribute was assumed to indicate an error for that attribute. An overall error rate was also computed by performing a logical OR operation on the outcomes from the three individual tree attributes: if an error was tallied for at least one of DBH, height, or species composition for a particular pair of actual and simulated stands, then that stand contributed to the overall error rate.

Testing scenarios corresponding to the four possible tree list generation modes were considered: default, varied heights (Var-H), pure stands, and pure stands with varied heights (pure/Var-H). Figures are provided only for the default scenario since results for the other three testing scenarios were similar. Tables summarizing the GOF results for all four testing scenarios are presented. Results are summarized in three different ways. First, DBH, height, species composition, and total error rates are presented for each testing scenario, with histograms of the computed IAE values for the default scenario. These comparisons examine the within stand, or tree scale, agreement between the actual and simulated samples, with the histograms providing a graphical indication of the error rates. Second, scatter plots of simulated and actual QMD and mean height are presented with R^2 values and results from a simple linear regression. Third, the distributions of QMD and mean height across all actual and simulated stands are compared by computing their respective IAE values, nonparametric estimates of their PDFs, bias and RMSE values. The latter comparisons provide an indication of how well the QMD and average height attributes are reproduced by the simulated stands, examining the between stand, or stand scale, consistency of the simulated stands, providing an indication of their agreement across the region of interest.

3. Results

Empirical error rates for the untreated stands and the default testing scenario are presented in Table 4, and histograms of the computed IAE values for DBH (top), height (middle), and species composition (bottom) are presented in Figure 1. The figures clearly show that the bulk of the actual *vs.* simulated stand comparisons produced IAE values that were less than the IAE cutoff value of 0.50 indicating that the majority of simulated untreated stands were similar to their actual counterparts. The empirical error rates for the default testing scenario were 1.54% for the DBH distributions, 5.26% for the height distributions, 12.31% for species composition, and 15.84% overall. These results, and those for the other testing scenarios, are consistent with estimates of natural variability for ”forests that appear identical” of 10% to 20% (Botkin, 1993).

The larger error rate obtained for species composition, relative to those for the DBH and height distributions, is due to the discrete nature of the species distributions for each stand and the small numbers of sample trees generated for each simulated sample plot. The greater magnitude of the height error rate, relative to that of the DBH distributions, may be explained, in part, by the use of height-diameter relationships to estimate missing tree heights: this produces a reduction in observed variability for tree heights in the sampled plots, giving narrower height distributions. A summary of the empirical error rate results for the other three testing scenarios is provided in Table 4 for comparison.

As can be seen in Table 4, approximately 84% of the untreated stands were similar for each of the

four testing scenarios demonstrating very good agreement between the actual and simulated stands at the tree scale. The results for the 100% pure scenario were slightly better than those for the default scenario, having a total error rate of 15.32% *vs.* 15.84%, with the varied height and pure/varied height scenarios performing slightly worse than the default scenario, having total error rates of 16.28% and 16.07%, respectively. These results were consistent with expectations, since the 100% pure scenarios reduced the variability in the simulated stands by eliminating all but one species for single species samples, while the two varied height scenarios increased the variability of the height values in the simulated stands. These effects may be seen in Table 4, where the three individual error rates for the pure scenario are smaller than those of the default scenario, and the height error rates for the varied height and pure/varied height scenarios are greater than those obtained for the default and pure scenarios, respectively.

Plots of the simulated *vs.* actual QMD (left) and mean height (right) data for the default scenario are presented in Figure 2. An examination of the plots clearly indicates a strong linear relationship between the actual and simulated values for both QMD and mean height. A simple linear regression analysis produced regression lines and R^2 values of $y = 0.4406 + 0.9731x$ and $R^2 = 0.9673$ for QMD and $y = 0.1659 + 0.9921x$ and $R^2 = 0.8705$ for mean height using all 5209 points. The intercepts and slopes were all statistically significantly different from their expected values of zero (0) and one (1), but the intercept values were on the order of typical measurement errors on individual trees. This is a statistical significance *vs.* biological relevance issue, and the small differences between the computed intercepts and slopes and their expected values are highly unlikely to be biologically relevant. Three data points, potential outliers, are visible below the main mass of data in the QMD and mean height plots creating leverage and moving the intercepts upward and away from zero. These results indicate strong agreement between the actual and simulated untreated stands for QMD and mean height values, but with greater variability for mean height indicated by the smaller R^2 value.

The three potential outliers in Figure 2 are not visible as outliers when mean height is plotted against QMD as in Figure 3. These three points represent three consecutive measurements of a single sample plot that appear at the extreme upper right of the actual joint QMD-mean height data, and each of these plots occurs in its own partition bin. These plots have five measured trees, and one tree has DBH and height values that are 20 cm and 5 msmaller than those of the four other trees. Further, because these points are on the extreme edge of the data, all of their nearest neighbors will have smaller aggregate and individual tree values, causing the downward shift seen in Figure 2. Adding additional sample plots with larger trees, using fewer than five of the nearest partition bins, or using only the isolated partition bin for each sample, a feature of the TLGDB that was not discussed, would help to resolve this issue.

Nonparametric probability density estimates of the actual and simulated QMD (top) and mean height (bottom) distributions for the default scenario are presented in Figure 4. The IAE values for these distributions and the default scenario were 0.0297 and 0.0289, respectively, indicating that the actual and simulated distributions were almost identical, having approximately 98% of their total probability mass in common. The two-dimensional joint distributions for QMD and mean height, see Figure 3, were also estimated for the actual and simulated samples for each scenario to compute IAE values, obtaining a value of 0.0657 for the default scenario, indicating that the two distributions had almost 97% of their probability mass in common. The full range of values for the actual QMD and mean height distributions were also spanned by the marginal and joint distributions for the simulated values. Results for the other three testing scenarios are presented in Table 5 for comparison, along with bias, RMSE, and R^2 values for each scenario.

4. Discussion

The primary benefits of the HNNTLG procedures for tree list generation are the use of an implicit two-scale mixture distribution derived relationship between coupled coarse and fine scales, a direct

partitioning of the coarse scale attributes, a mapping associating fine scale attributes with uniquely determined partition sets to provide a classification of the stands and a corresponding conditioning of the fine scale attributes, and the use of the SIMDAT algorithm to generate simulated tree measurements spanning the full range of the observed measurements for the sampled stands. Use of the mixture distribution formulation and the partition provides a variety of benefits, including a reduction in the number of nearest neighbor points that must be considered to find good matches, the automatic stratification of multivariate data with discrete categories and continuous attributes (Eskelson et al., 2008, McRoberts, 2009), and it provides regularization and consistency of prediction for new data via localization and conditioning of the attributes, reducing the potential for negative impacts due to a bias-variance trade off or over fitting in regression models (Eskelson et al., 2009b). While the partition and conditioning used here were hierarchical, they need not be: the same methods and formulation work for attributes at the same scale or level, e.g., using TPH, QMD, and average height to impute missing basal area per hectare values. The partition also need not be a simple rectilinear division of the data space. Finally, use of the SIMDAT algorithm allows the full range of sampled data to be reproduced in the simulated samples, avoiding the reduction in variability of imputed values obtained by a simple averaging of the k nearest neighbors (Eskelson et al., 2009b).

The HNNTLG procedures produce a nonparametric estimate of a likelihood or probability distribution, making the imputed results asymptotically unbiased and statistically consistent, relative to the unknown probability distribution under standard assumptions (Silverman, 1986, Thompson and Tapia, 1990, Redner, 1999, Redner and Gehringer, 1994, Gehringer and Redner, 1992, Gehringer, 1990). The HNNTLG procedures may be used to consistently generate simulated tree lists, while allowing the addition of new data to an existing partition or creating a refinement of the coarse scale attribute partition if necessary, e.g., when the total number of samples or the number of samples associated with the partition bins exceeds some threshold. The need for a partition refinement as the sample size increases is a theoretical requirement for convergence to the unknown probability distribution as the sample sizes increases (Silverman, 1986, Thompson and Tapia, 1990, Redner, 1999, Redner and Gehringer, 1994, Gehringer and Redner, 1992, Gehringer, 1990), but in practice the need to refine the partition is not critical.

The fundamental problem of imputation involves estimating or approximating, in some way, a probability distribution relating two sets of associated attributes. A subset of data points is assumed to have observed values for all attributes from both sets, with the remaining points having observed values for only the first set of attributes, and the problem is to fill in the missing values. The explicit formulation of the tree list imputation problem in the context of a joint mixture distribution of related attributes and the partition induced conditioning establish theoretical and computational frameworks for imputation methods, including distance weighted nearest neighbor methods and imputation methods based on classification and regression trees (Temesgen et al., 2008, Eskelson et al., 2009a). The flexibility of the HNNTLG approach is demonstrated by showing that a class of distance weighted nearest neighbor imputation methods may be formulated as a limiting case within the context of the HNNTLG joint mixture distribution and partitioning imputation framework.

4.1. Comparison with distance weighted methods

The HNNTLG framework for nearest neighbor tree list imputation methods, includes distance weighted methods using Euclidean or Mahalanobis distance or the most similar neighbor (MSN) method (Moeur and Stage, 1995) whose weights are determined by a canonical correlation analysis. The HNNTLG procedures become a straightforward distance weighted method if the partition bins are small enough to isolate each sample stand in its own bin and if $K_{\max} = 1$ and $K_T = 1$. Under these conditions the HNNTLG procedures identify the nearest partition bin or sample stand and copy the associated tree list. If $K_{\max} = 1$, $K_T > 1$, then the most similar stand is selected and a simulated tree list is generated using the sampled trees, a capability not supported by most distance weighted meth-

ods that typically copy attribute values from the closest sample point or average the attribute values from the k nearest points. Clearly, the distance weighted methods could generate simulated tree lists by using the SIMDAT algorithm (Taylor and Thompson, 1986, Thompson, 2000).

The HNNTLG procedures and three distance weighted nearest neighbor imputation methods are compared by estimating basal area per hectare (BAPH) values ($m^2 ha^{-1}$) using TPH, QMD, and mean tree height stratified by stand type, the same attributes used to generate the HNNTLG partition. The distance metrics used were Euclidean, Mahalanobis, and MSN weighted distances, with the Mahalanobis and MSN weights computed within each stand type stratum. Two nearest neighbor selection strategies were used: select only the nearest neighbor (NN) and select the k-nearest neighbors (k-NN), averaging the k BAPH values associated with the nearest stands in each stratum. A value of $k = 5$ was chosen for compatibility with the value of $K_{\max} = 5$, the number of partition bins used in the HNNTLG procedures to generate the canonical stand. For each distance measure, all available stands within a stratum were used to find the nearest neighbor or k-nearest neighbors, a cross-validation approach (Efron, 1982, Efron and Tibshirani, 1998), maximizing the opportunity for the distance weighted methods to find a good match. The HNNTLG procedures used the simulated trees generated for each stand to compute the imputed BAPH values while the BAPH values for the distance weighted methods were computed using the actual BAPH values for the nearest stand(s).

Results of the comparison are presented in Table 6, and indicate that the HNNTLG procedures are quite competitive. The HNNTLG procedures produced the largest magnitude mean bias of 0.3547, followed, in decreasing magnitude, by Euclidean k-NN, Mahalanobis k-NN, Euclidean NN, MSN NN, Mahalanobis NN, and MSN k-NN methods, having mean bias values of -0.2266, -0.0751, 0.0676, 0.0560, -0.0450, and -0.0257, respectively. The Mahalanobis NN and k-NN distance methods gave the smallest magnitude RMSE values, 4.3681 and 4.3668, respectively, followed, in increasing magnitude, by the HNNTLG procedures, Euclidean NN, the Euclidean k-NN, the MSN k-NN, and the MSN NN methods, with RMSE values of 6.2321, 8.8908, 9.1786, 10.2631, 13.1139, respectively. Finally, the Mahalanobis NN and k-NN distance methods gave the largest magnitude R^2 values of 0.9449 and 0.9445, respectively, followed, in decreasing order, by the HNNTLG procedures, Euclidean NN, Euclidean k-NN, MSN k-NN, and MSN NN methods, with R^2 values of 0.8894, 0.7770, 0.7579, 0.6976, and 0.5642, respectively.

The Mahalanobis distance weighted methods outperformed all of the other imputation methods, having small bias, the smallest RMSE values, and the largest R^2 values, with the HNNTLG procedures in a very competitive third place, and the other methods trailing behind, having larger RMSE values and smaller R^2values. Three factors directly influenced the ranking results for the top three methods. First, the Mahalanobis distance weighting is effectively optimal for this problem. Second, the data set contained a number of stands with multiple sample plots, so there are some very good matches of TPH, QMD, mean height, and BAPH in the overall data set. Third, no attempt was made to generate good or optimal tree lists using the HNNTLG method, so there are some relatively poor matches that degrade its performance.

The MSN NN and MSN k-NN methods performed rather poorly for these data in terms of variability, having the largest RMSE values and the smallest R^2 values. The small mean bias values obtained for the MSN NN and MSN k-NN methods clearly do not tell the whole story: variability also matters. The emphasis of the statistically weighted MSN method on maximizing correlations via a canonical correlation analysis, and concomitantly minimizing bias, globally across an entire data set or strata within a data set, likely increases the variability observed due to a bias-variance trade off: the minimum MSE typically does not occur for the minimum bias.

4.2. GOF testing procedures

The IAE cutoff value of 0.50 used for the GOF comparisons may be too small. Individual tree DBH measurement data from 30 juvenile, pure Douglas-fir stands were examined to estimate an upper

bound for the IAE cutoff value. Within each Douglas-fir stand two to four permanent study plots were available, each having one to three commensurate measurement dates across the sample plots, providing a total of 352 within stand comparisons to compute IAE values. A maximum IAE cutoff value of approximately 0.60 was obtained after screening outliers, indicating that two samples drawn from the same stand could have differences of up to 30% in probability mass for their DBH distributions. If an IAE cutoff value of 0.60 were used here, total error rates less than 10% would have been obtained, so the IAE cutoff value of 0.50 was reasonable, while somewhat conservative.

4.3. Future work and extensions

The TLGDB currently supports only pure Douglas-fir, pure western hemlock, and mixed species stands containing one of these two species as a dominant component. Extending the TLGDB to allow arbitrary dominant tree species, and possibly even understory vegetation, is highly desirable. These extensions would permit a broader use of the HNNTLG procedures for generating simulated forest attributes as well as providing a mechanism to add more realism to simulations of forest development and ecology.

The number of treatment types and treatment combinations should also be expanded in the TLGDB. Currently only untreated and thinned stands may be represented in a TLGDB, with support for multiple thinning events. Stands having fertilization, pruning, or combinations of thinning, pruning, and fertilization could also be supported. A greater variety of distance measures should also be made available to increase the flexibility of the TLGDB software. Different distance measures can produce different partition boundaries which could be beneficial for some problems, particularly if multiple, nested, two-scale relationships are used, for example, geographic location $\rightarrow$ site and stand attributes, site and stand attributes $\rightarrow$ tree attributes, and tree attributes $\rightarrow$ branch attributes.

Finally, when generating simulated tree scale attributes it may also be desirable to condition the continuous attributes using one or more discrete attributes, e.g., tree species, during the tree generation procedure by restricting the tree DBH and height measurements to the species of the tree currently being generated, provided that there is enough data to permit it. This would, then, allow differences in height–diameter relationships among tree species to be used. A straightforward version of this filtering was done to obtain the 100% pure stands of Douglas-fir and western hemlock.

5. Conclusions

Procedures for using an implicit two-scale relationship with a nearest neighbors algorithm to generate simulated trees representative of forest conditions specified using a stand scale description have been described and shown to perform well. The primary benefits of these procedures are their use of a direct partitioning of the stand scale attributes to guarantee local consistency, their use of actual tree measurement data to generate simulated trees to guarantee that simulated tree attributes are realistic and biologically achievable, and their ability to simulate within stand variability. The procedures may be used to generate forest stands for use with growth and yield or forest simulation models or to provide estimates of forest characteristics and their potential variability across sparsely sampled regions.

6. Acknowledgements

Initial development of the Tree List Generation Database was funded by the Stand Management Cooperative (SMC) in the College of Forest Resources at the University of Washington, Seattle, WA. We would like to thank the SMC for their support. We would also like to thank Jim Flewelling and Temesgen Hailemariam for their reviews of the tree list generation database documentation.

Data for this work were provided by: The British Columbia Ministry of Forests, The Canadian Forest Service, Oregon State University, Port Blakely Tree Farms, The Regional Forest Nutrition Research Project (RFNRP), The Stand Management Cooperative (SMC) at The University of Washington, The

U.S. Forest Service Pacific Northwest Research Station, The Washington State Department of Natural Resources, and The Weyerhaeuser Company, and we thank them for their support.

An earlier version of this work was presented at the 2006 Nearest Neighbors Workshop ”Meeting in the Middle” held August 28-30 at the University of Minnesota, Minneapolis, MN. We would like to thank the organizers and participants of this workshop for their interest and many helpful comments.

References

Biging, G., Robards, T., Turnblom, E., and VanDeusen, P. (1994). The predictive models and procedures used in the forest stand generator (STAG). *Hilgardia*, 61(1):36 pp.

Borders, B. E. and Patterson, W. D. (1990). Projecting stand tables: A comparison of the weibull diameter distribution method, a percentile-based projection method, and a basal area growth projection method. *For. Sci.*, 36(2):413–424.

Botkin, D. B. (1993). *Forest Dynamics: An Ecological Model*. Oxford University Press.

Devroye, L. and Gyorfi, L. (1985). *Nonparametric Density Estimation: the L_1 View*. Wiley, New York.

Donnelly, D. (1997). *Pacific Northwest coast variant of the forest vegetation simulator*. Addison-Wesley. Available from http://www.fs.fed.us/fmsc/ftp/fvs/docs/overviews/pnvar.pdf [Last Accessed 16 May 2008].

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. CBMS-NSF Regional Conference Series in Applied Mathematics 38. SIAM.

Efron, B. and Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman & Hall/CRC.

Eskelson, B., Temesgen, H., and Barrett, T. (2009a). Estimating cavity tree and snag abundance using negative binomial regression models and nearest neighbor imputation methods. *Canadian Journal of Forest Research*, 39:1749–1765.

Eskelson, B. N. I., Temesgen, H., and Barrett, T. M. (2008). Comparison of stratified and non-stratified most similar neighbour imputation for estimating stand tables. *Forestry*, 81(2):125–134.

Eskelson, B. N. I., Temesgen, H., LeMay, V., Barrett, T. M., Crookston, N. L., and Hudak, A. T. (2009b). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research*, 24:235–246.

Gehringer, K. R. (1990). Nonparametric probability density estimation using normalized B-Splines. Master's thesis, The University of Tulsa.

Gehringer, K. R. (2001). *New shoots: A tree list generation database tutorial*. Stand Management Cooperative, College of Forest Resources, University of Washington, Seattle, Box 352100, Seattle, WA 98195-2100.

Gehringer, K. R. and Redner, R. A. (1992). Nonparametric probability density estimation using normalized B-splines. *Commun. Stat. B-Simul.*, 21(3):849–878.

Gehringer, K. R. and Turnblom, E. C. (2001). *Tree list generation database user's guide and reference manual*. Stand Management Cooperative, College of Forest Resources, University of Washington, Seattle, Box 352100, Seattle, WA 98195-2100.

Hann, D., Hester, A., and Olsen, C. (1997). *ORGANON User's manual Edition 6.0*. Dept. Forest Resources, Oregon State University, Corvallis, OR 97331-5703.

LeMay, V. and Temesgen, H. (2005). Comparison of nearest neighbor methods for estimating basal area and stems per ha using aerial auxiliary variables. *For. Sci.*, 51(2):109–119.

Maguire, D. A., Bennett, W. S., Kershaw Jr., J. A., Gonyea, R., and Chappell, H. N. (1991). Establishment report stand management cooperative silviculture project field installations. Technical report, Stand Management Cooperative, College of Forest Resources, University of Washington, Room 164, Bloedel Hall, Box 352100, Seattle, WA, 98195-2100.

McRoberts, R. E. (2009). A two step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. *Remote Sensing of Environment*, 113:532–545.

Mitchell, K. J. (1975). *Dynamics and Simulated Yield of Douglas-fir*. Monograph 17. For. Sci.

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

Redner, R. A. (1999). Convergence rates for uniform B-spline density estimators. I. One dimension. *SIAM J. Sci. Comput.*, 20(6):1929–1953 (electronic).

Redner, R. A. and Gehringer, K. (1994). Function estimation using partitions of unity. *Commun. Stat. A-Theor.*, 23(7):2059–2078.

Reynolds, Jr., M. R., Burk, T. E., and Huang, W.-C. (1988). Goodness-of-fit tests and model selection procedures for diameter distribution models. *For. Sci.*, 34(2):373–399.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman Hall.

Taylor, M. S. and Thompson, J. R. (1986). A data based algorithm for the generation of random vectors. *Comput. Stat. Data An.*, 4:93–101.

Temesgen, H. (2003). Estimating tree-lists from aerial information: a comparison of a parametric and most similar neighbor approaches. *Scand. J. For. Res.*, 18:279–288.

Temesgen, H., Barrett, T., and Latta, G. (2008). Estimating cavity tree abundance using nearest neighbor imputation methods for western oregon and washington forests. *Silva Fennica*, 42(3):337–354.

Temesgen, H., LeMay, V., Marshall, P., and Froese, K. (2003). Imputing tree-lists from aerial attributes for complex stands of british columbia. *Forest Ecol. Manag.*, 177:277–285.

Thompson, J. R. (2000). *Simulation: A modeler's approach*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Thompson, J. R. and Tapia, R. A. (1990). *Nonparametric Function Estimation, Modeling, and Simulation*. SIAM: Society for Industrial and Applied Mathematics.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons.

Wykoff, W. R. (1986). *Supplement to the User's Guide for the stand PROGNOSIS model - Version 5.0*. USDA FS Intermountain For. and Range Exp. Stn., Ogden, UT. Gen. Tech. Rep. INT-208.

Wykoff, W. R., Crookston, N., and Stage, A. (1982). *User's guide to the stand prognosis model*. USDA FS Intermountain For. and Range Exp. Stn., Ogden, UT. Gen. Tech. Rep. INT-133.

Table 1. Stand and tree scale attributes and partition bin widths.

Attribute	Scale	Type	Bin width h_j
Stand type	Stand	Discrete	0
Stand density (TPH)	Stand	Continuous	100
QMD (cm)	Stand	Continuous	2.5
Mean height (m)	Stand	Continuous	1.0
DBH (cm)	Tree	Continuous	N/A
Height (m)	Tree	Continuous	N/A
Species	Tree	Discrete	N/A

Table 2. Stand type summary for untreated stands.

Stand type	Count	Percent
Pure Douglas-fir	3404	65.3
Pure western hemlock	800	15.4
Douglas-fir dominant	691	13.3
Western hemlock dominant	178	3.4
Mixture	136	2.6
Total	5209	100.0

Table 3. Stand scale data coverage summary, by stand type, for untreated stands.

Stand type (count)	Attribute	Mean	S.D.	Min.	Med.	Max.
Pure DF (1539)	QMD (cm)	23.3	12.3	3.8	21.3	73.8
	Mean height (m)	19.9	9.7	3.5	18.5	51.5
	TPH	1506.2	1175.4	150.0	1150.0	9750.0
Pure WH (543)	QMD(cm)	18.6	8.4	6.3	16.3	41.3
	Mean height (m)	19.1	8.5	4.5	18.5	40.5
	TPH	3307.6	2647.1	650.0	2350.0	14050.0
DF dominant (507)	QMD (cm)	19.5	8.4	3.8	16.3	48.8
	Mean height (m)	17.0	7.3	3.5	15.5	42.5
	TPH	2086.9	1345.6	150.0	1750.0	9550.0
WH dominant (151)	QMD (cm)	21.8	6.4	8.8	21.3	41.3
	Mean height (m)	20.8	6.8	8.5	21.5	34.5
	TPH	1983.1	1071.7	650.0	1650.0	5450.0
Mixture (125)	QMD (cm)	23.4	6.1	8.8	23.8	36.3
	Mean height (m)	21.1	5.4	9.5	21.5	31.5
	TPH	1348.4	1081.1	250.0	850.0	4750.0

Table 4. Error rates for DBH, height, and species GOF tests. Values are the proportion of stands misclassified for each variable. The total column gives the proportion of stands which were misclassified for at least one of DBH, height, or species.

Scenario	DBH	Height	Species	Total
Default	0.0154	0.0526	0.1231	0.1584
Pure	0.0138	0.0507	0.1175	0.1532
Var-H	0.0148	0.0543	0.1257	0.1628
Pure/Var-H	0.0150	0.0582	0.1192	0.1607

Table 5. Overall QMD and mean height distribution IAE values, bias, RMSE, and R^2 values for comparisons of actual and simulated untreated stands for all four scenarios.

Scenario	QMD (cm)				Mean height (m)				Joint
	IAE	Bias	RMSE	R^2	IAE	Bias	RMSE	R^2	IAE
Default	0.0297	0.1320	1.6419	0.9673	0.0289	-0.0143	0.8775	0.8705	0.0657
Pure	0.0220	0.0107	1.6162	0.9696	0.0221	-0.1438	0.9530	0.8753	0.0635
Var-H	0.0240	0.1297	1.6051	0.9679	0.0238	-0.0206	0.8950	0.8704	0.0647
Pure/Var-H	0.0191	0.0127	1.6232	0.9686	0.0236	-0.1401	0.9793	0.8727	0.0658

Table 6. Comparison of the HNNTLG default scenario, Euclidean distance, MSN canonical correlation weighted distance, and Mahalanobis distance to impute BAPH values using TPH, QMD, and average height stratified by stand type.

Method	Bias	RMSE	R^2
HNNTLG	0.3547	6.2321	0.8894
Euclidean NN	0.0676	8.8908	0.7770
Euclidean k-NN	-0.2266	9.1786	0.7579
MSN NN	0.0560	13.1139	0.5642
MSN k-NN	-0.0257	10.2631	0.6976
Mahalanobis NN	-0.0450	4.3681	0.9449
Mahalanobis k-NN	-0.0751	4.3998	0.9445

Fig. 1. Histograms of IAE values for DBH (top), height (middle), and species composition in untreated stands for the default scenario. The vertical line at 0.50 represents the boundary between distributions that are similar (IAE < 0.50) and different (IAE ≥ 0.50). The error rates were 1.54%, 5.26%, and 12.31% for DBH, height, and species, respectively.

Fig. 2. Plots of simulated *vs.* actual QMD (left) and mean height (right) values from untreated stands for the default scenario, with regression lines and R^2 values of $y = 0.4406 + 0.9731x$ and $R^2 = 0.9673$ for QMD and $y = 0.1659 + 0.9921x$ and $R^2 = 0.8705$ for mean height.

Fig. 3. Actual and simulated mean height *vs.* QMD from untreated stands for the default scenario and all sample plots, with a 2-dimensional IAE value of 0.0657.

Fig. 4. Nonparametric probability density function estimates for the actual and simulated QMD (top) and mean height (bottom) distributions for all untreated stands and the default scenario. The IAE values are 0.0289 and 0.0297 for QMD and mean height, respectively.

Figure 1

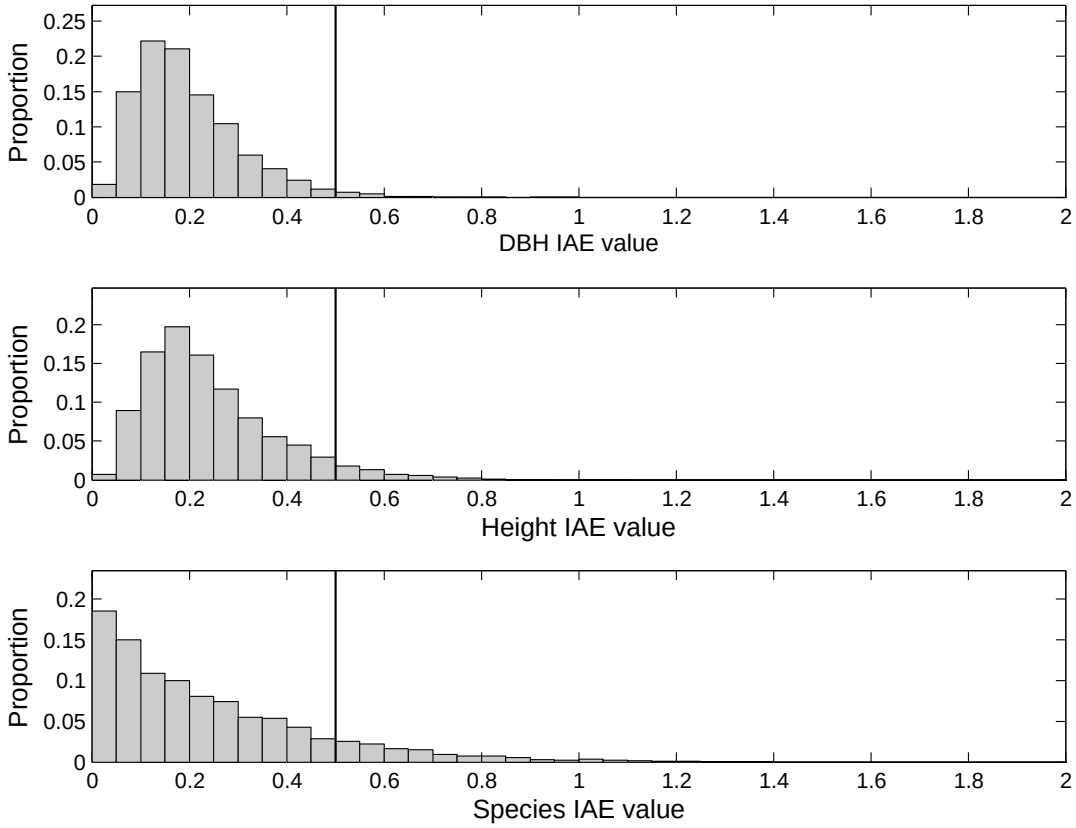

Proportion
DBH IAE value
Proportion
Height IAE value
Proportion
Species IAE value

Figure 2

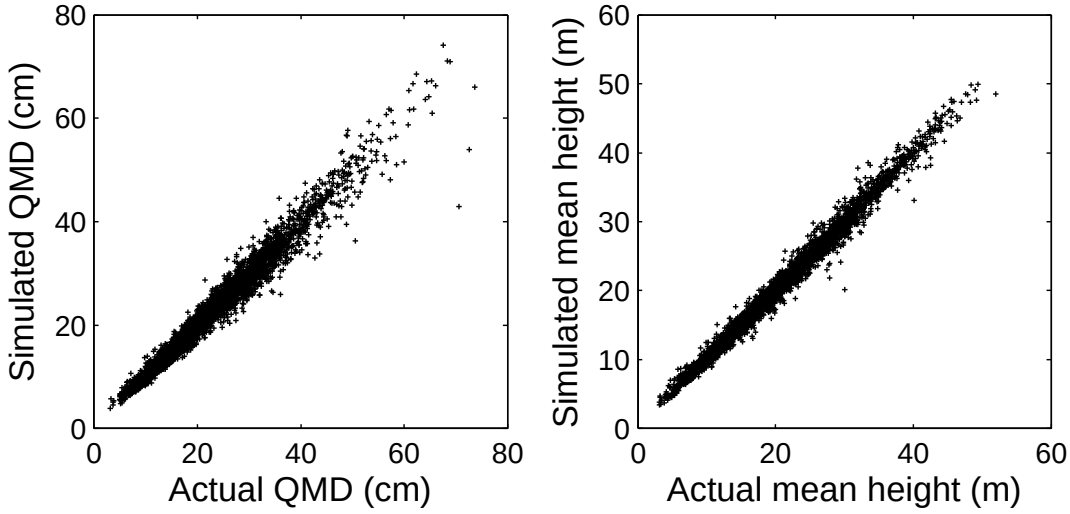

Simulated QMD (cm)
Actual QMD (cm)
0
20
40
60
80
Simulated mean height (m)
Actual mean height (m)
0
10
20
30
40
50
60

Figure 3

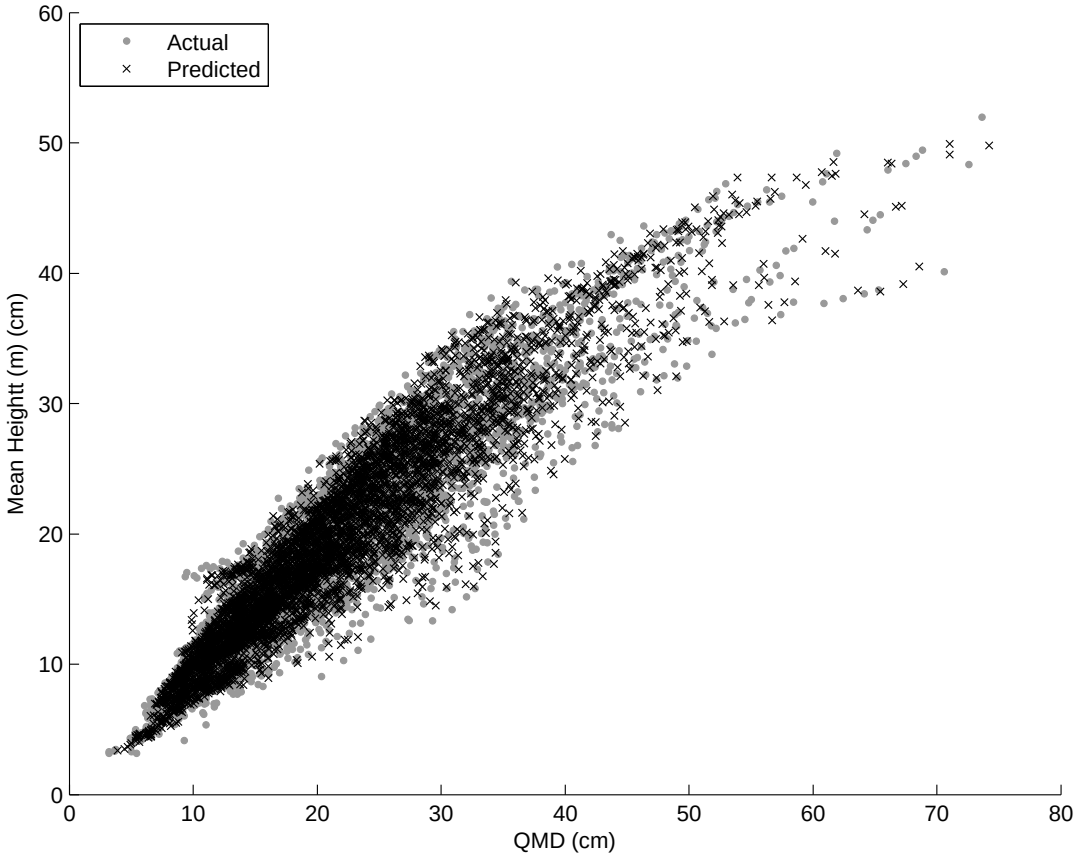
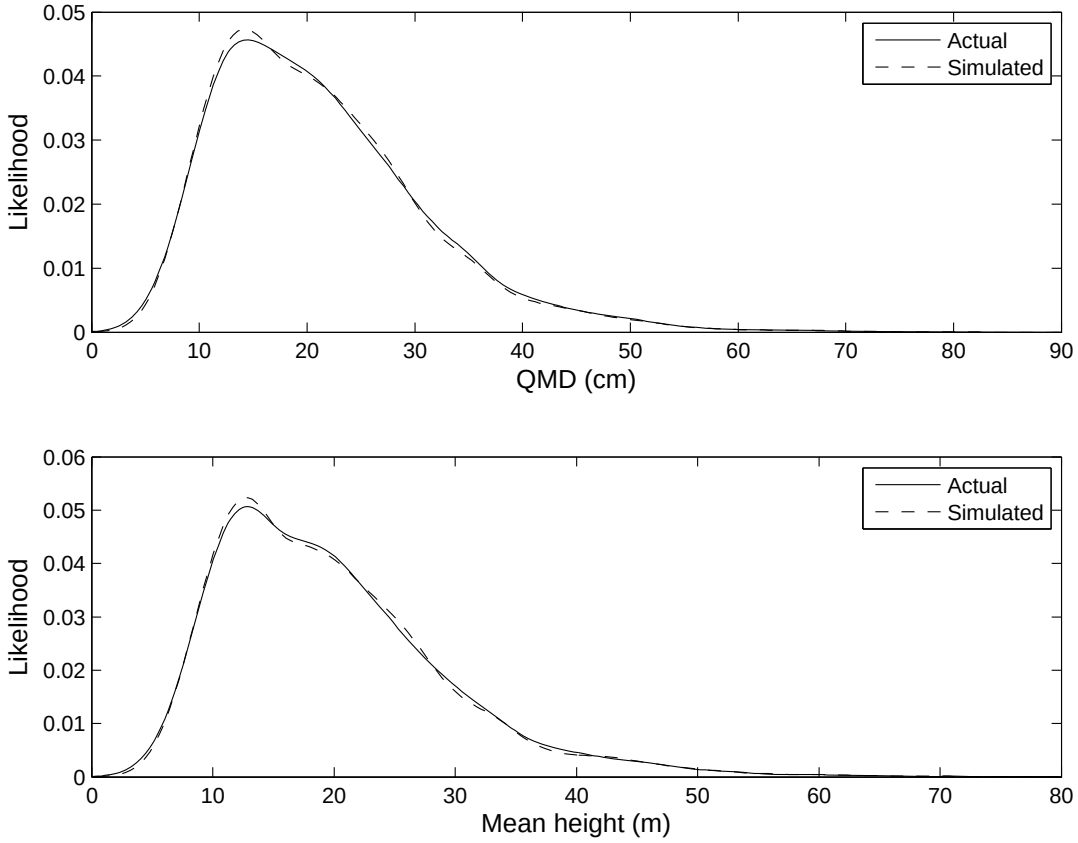

Actual
Predicted
Mean Heightt (m) (cm)
QMD (cm)
0
10
20
30
40
50
60
70
80

Figure 4


Actual
Simulated
Likelihood
QMD (cm)
0.05
0.04
0.03
0.02
0.01
0
0 10 20 30 40 50 60 70 80 90
Actual
Simulated
Likelihood
Mean height (m)
0.06
0.05
0.04
0.03
0.02
0.01
0
0 10 20 30 40 50 60 70 80