Constructing a virtual forest: An hierarchical nearest neighbors method for generating simulated tree lists

Kevin R. Gehringer *

November 2, 2006

Abstract

A nearest neighbors method for generating simulated tree lists has been developed. The method employs an hierarchical structure to incorporate information at two implicitly defined spatial scales, a coarse scale representing the distribution of stand attributes across a region and a fine scale representing the distribution of individual 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 (coarse) scale attributes for each of the actual tree lists in the data, and comparisons of the distributions of the simulated and actual attributes at the stand (coarse) scale and the tree (fine) scale were made. At the stand (coarse) 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 in common for each attribute. At the tree (fine) 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 94% of the simulated stands being statistically indistinguishable from their respective actual stands.

1 Introduction

Many modern forest growth models and forest simulation systems use some form of tree list, minimally a set of compatible diameter at breast height (DBH) and height measurements with an indication of tree species, to describe a stand that is to be modeled, simulated, or projected into the future (Mitchell, 1975, Wykoff et al., 1982, Wykoff, 1986, Donnelly, 1997, Hann et al., 1997). An appropriate initial tree list is a necessary input for such individual tree based stand simulation systems to accurately capture and project the stand dynamics or to define an initial condition for growth and yield projections. Tree lists obtained from measurements of existing plots or stands, and for which complete individual tree measurements are available to define the initial condition of a stand, are highly advantageous, and their use allows forest growth simulations to begin with the *ground truth*. In many situations, however, individual tree measurement data are not available, e.g., if aggregated stand attributes or statistical summaries are all that is available, or if actual sample data are not available for a particular geographic location or management regime.

*Biometrics Northwest LLC, Redmond, WA

The use of tree lists as a primary input to forest stand simulators and growth and yield models is likely to persist. Improvements in growth and yield models and stand simulation systems continue to increase the detail and the spatial and temporal resolution of the models, providing greater accuracy and precision, more realistic growth predictions and simulations of stand dynamics, and incorporating physiological and ecological aspects of tree growth and stand development. The increased detail represented in forest growth and simulation models, and the improved accuracy and precision of their outputs, are important for economic analyses of forest productivity on a shrinking land base, with shorter rotations than in the past, and for ecological analyses, such as the restoration or creation of habitat for wildlife via silvicultural manipulations of managed forest stands.

The forest growth and yield and simulation models currently used in the Pacific Northwest are typically limited by the detail of the initial stand descriptions. The ability to generate a *reasonable*, realistic, simulated tree list that is 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, and would support the use of tree list based growth and yield or stand simulation systems for a broad range of modeling and silvicultural situations by meeting their detailed, tree list input requirements.

General, hierarchical, nearest neighbors tree list generation (HNNTLG) procedures for creating simulated tree lists using an implicit two-scale relationship are described. The implicit two-scale relationship is represented as two nested spatial scales: a coarse scale representing the distribution of stand attributes within a region and a fine scale representing the distribution of individual tree attributes within a stand. The objective was to develop a random stand or tree list generator, within the framework of pseudorandom number generation, using the following criteria to guide the development of the HNNTLG procedures and the design of a suite of programs, the tree list generation database (TLGDB), implementing them (Gehringer, 2001, Gehringer and Turnblom, 2001).

- Individual trees in a simulated stand or tree list should be generated as multidimensional objects directly.
- Simulated trees should be physically realizable, and the tree list generation methodology should only generate *realistic* trees.
- The stand representation should be flexible enough to represent multimodal and relatively flat tree size distributions, as well as unimodal size distributions, and arbitrary species compositions.
- The addition of data from new sampled stands should be easy to perform, and it should not modify the existing relationships among stands used to generate tree lists, except possibly by the creation of a new stand class, guaranteeing that the tree list generation procedure remains consistent as new data are added.
- The framework used to generate a simulated stand or tree list should be as similar as possible to the standard framework of pseudorandom number generation, permitting the generation of different, but similar, simulated tree lists for stands whose stand scale attributes map to the same sampled stand.

2 Methods

The approach used to develop the HNNTLG methods was modeled after the well understood problem of pseudorandom number generation. Fundamental to pseudorandom number generation is the identification of

a suitable representation for a probability density function (PDF), $f(x)$, or cumulative distribution function (CDF), $F(x) = \int_{-\infty}^{x} f(t)dt$ from which random values may be generated. A suitable representation for $f(x)$ or $F(x)$ in this context implies that the representation chosen for the PDF or CDF facilitates the development of an algorithm to generate random values from the underlying distribution.

A mixture density representation (Redner and Walker, 1984, Titterington et al., 1985, Borders and Patterson, 1990, Biging et al., 1994) for the joint distribution of stand (coarse) scale attributes from forest stands within a region and tree (fine) scale attributes within stands is described next. This representation for the joint distribution was used to derive linked approximations to the marginal stand scale and tree scale attribute distributions that were then used to generate simulated tree attributes for a specified stand description using a nearest neighbors algorithm. The specific algorithms used by the HNNTLG procedures, as implemented in the TLGDB are then described, followed by a description of a data set and the methods that were used to assess the performance the implemented HNNTLG procedures.

2.1 Representing the HNNTLG distribution

A forested region may be represented as a collection or mosaic of more or less distinct forest stands or patches that are distinguished by their physical attributes and their dominant vegetation types, e.g., stand age, site index, soil characteristics and site quality, the numbers and sizes of trees and their species compositions. Each stand or patch, then, may be represented by a multivariate PDF$f(x)$ describing the distributions of its stand and tree scale attributes in a vector x representing stand scale and tree scale attributes that may be obtained from a sampled forest stand or patch.

The joint distribution of stand and tree attributes across a region may, therefore, be represented by a mixture distribution (Redner and Walker, 1984, Titterington et al., 1985, Borders and Patterson, 1990, Biging et al., 1994) 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 patches, $f_i(x), i = 1, 2, \ldots, N$ represent the joint PDFs for the stand and tree attributes of the individual forest stands or patches in the region, and $\alpha_i > 0, i = 1, 2, \ldots, N$ are weights such that $\sum_{i=1}^{N} \alpha_i = 1$, making f itself a PDF.

The number of component density functions in the mixture density is controlled by the interpretation of the mixing weights α_i and the component density functions $f_i(x)$. Interpreting the mixing weights α_i as the probabilities of occurrence within a region of the stands or patches described by a representative set of component densities $f_i(x)$ yields a concise representation for the mixture density $f(x)$, whereas a sample based interpretation having $\alpha_i = \frac{1}{N}$ for each of N sampled stands would lead to a representation that would contain a significant amount of duplication, having multiple component densities representing similar forest stands or patches. For simplicity, the component density functions $f_i(x)$ representing each stand or patch are assumed to be unimodal, as any component density function containing multiple modes could be decomposed into a sum of unimodal density functions.

Ultimately the distribution of tree attributes representing the structural conditions within a specific forest stand or patch is required to generate a simulated tree list that is representative of the attributes of that stand or patch. Define a partition $x = (x^s, x^t)$ of a d-dimensional attribute vector x into its stand (coarse) scale attributes $x^s = (x_1^s, x_2^s, \ldots, x_{d_s}^s)$ and its tree (fine) scale attributes $x^t = (x_1^t, x_2^t, \ldots, x_{d_t}^t)$, with $d_s > 0$, $d_t > 0$, and $d_s + d_t = d$. Using this partition, 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 partition into stand scale and tree scale attributes explicit. With this notation, the distribution of tree attributes for any component density $f_i(x^s, x^t)$ in the mixture density $f(x^s, x^t)$ is simply the marginal density of tree attributes for that component, 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 may, therefore, be obtained for any component density $f_i(x^s, x^t)$, theoretically at least, and subsequently used to generate simulated tree attribute vectors x^t. To bridge theory to practice two difficulties must be overcome. First, the overall mixture density $f(x)$, its weights α_i, and its component densities $f_i(x)$ are unknown. This, it will turn out, is not a significant obstacle, as a direct representation of the mixture density is not necessary. Second, a procedure for identifying a specific component density function $f_i(x^s, x^t)$ and its associated marginal density of tree attributes $f_i^{\text{tree}}(x^t)$ is necessary to generate simulated tree attributes. An index derived directly from the attribute vectors is needed so that stand (coarse) scale descriptions can be used to select appropriate component density functions from which simulated tree (fine) scale attributes may be generated.

An index of component density functions $f_i(x^s, x^t)$ comprising a mixture density may be derived by first considering the support sets S_i generated by the individual component density functions and their corresponding stand scale marginal distributions $f_i^{\text{stand}}(x^s)$. Difficulties encountered using the support sets then led to the use of a direct partitioning and indexing of the stand scale attributes x^s in the domain of the mixture density. The direct partitioning was independent of the component density functions and was used with a conditioning argument to obtain distributions of tree scale attributes that were representative of the stand attributes within a subset of the partition. The partitioning, indexing, and conditioning procedures described next provide the framework for the implementation of the HNNTLG procedures as described in Section 2.2.

Let $f(x) = f(x^s, x^t)$ be a mixture density having N component density functions $f_i(x) = f_i(x_s, x_t)$ and mixing weights α_i representing the joint distribution of stand scale and tree scale attributes for forest stands or patches within a region as defined in Equation 1, and let $S_i = \{x \mid f_i(x) = f_i(x^s, x^t) > 0\}$ be the support for each component density function with

$$S = \bigcup_{i=1}^{N} S_i \tag{3}$$

being the support of the mixture density itself. A naive procedure for creating an index of the component densities would be to use the stand scale attributes x^s and the distributions of stand attributes from the marginal component densities $f_i^{\text{stand}}(x^s)$ to select a vector of representative stand attributes, an index point, from each of the support sets $S_i|_{x^s}$, e.g., the mode, where $S|_{x^s}$ indicates that the set S is restricted to the attributes in the vector x^s. A dictionary ordering could then be imposed on the chosen vectors from the support sets for rapid lookup using a nearest neighbor algorithm.

If all of the stand scale marginal distributions were different from one another, then this index, when used with a nearest neighbor algorithm, would uniquely identify each component density function $f_k(x^s, x^t)$ whose index point was closest to a specified stand scale attribute vector y^s representing a particular forest stand or patch. Once found, the marginal distribution of tree attributes for that component density $f_k^{\text{tree}}(x^t)$ could then be used to generate simulated tree attributes for the forest stand or patch specified by the stand attribute vector y^s. If, however, two, or more, stand scale marginal distributions were similar (statistically indistinguishable) then their index points and support sets would also be similar, and identifying the unique component density for a specified stand attribute vector y^s may not be possible. It is also possible to have a situation where index points from two or more different marginal component density functions are of equal,

or nearly equal, distance from a specified stand attribute vector y^s. This can occur since the support sets $S_i|_{x^s}$ for neighboring component density functions may overlap. If either of these situations occurs, some sort of tie-breaking decision must be made, e.g., by randomly choosing one component density or by averaging the component densities.

Ideally, it is preferable to avoid these two situations by requiring that each forest stand or patch described by a specified vector of stand attribute values y^s map to one, and only one, index point. A mapping from a specified stand description y^s to an unique index point may be obtained by partitioning the d_s-dimensional space of stand scale attributes $S|_{x^s}$ into disjoint subsets, and then reformulating the mixture density and the indexing problem based on the partition as follows.

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 $f(x) = f(x^s, x^t)$. By definition, the subsets B_m, $m = 1, 2, \ldots, M$ of a partition are disjoint or nonoverlapping, $B_i \cap B_j = \emptyset$ for $i \neq j$, and their union contains the support for the stand scale attributes $S|_{x^s}$, as in Equation 4,

$$S|_{x^s} \subset \bigcup_{m=1}^{M} B_m \tag{4}$$

guaranteeing that all of the stand scale vectors in the support set are included by the partition. Component density functions for a mixture density derived using the partition B as their corresponding support sets may then be obtained by conditioning the mixture density function $f(x) = f(x^s, x^t)$ by each of the sets B_m in the partition B to compute the conditional density functions $\hat{f}_{B_m}(x) = \hat{f}_{B_m}(x^s, x^t)$

$$\hat{f}_{B_m}(x) = \hat{f}_{B_m}(x^s, x^t) = \begin{cases} \frac{1}{P_{B_m}} f(x^s, x^t) & \text{for } x^s \in B_m \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where

$$P_{B_m} = \text{Prob}(\{x \mid x^s \in B_m\}) = \int_{\{x|x^s \in B_m\}} f(x)\, dx \tag{6}$$

is the probability that a point $x = (x^s, x^t)$ has its stand scale attributes in the set B_m. The mixture density derived from the partition B, $\hat{f}_B(x) = \hat{f}_B(x^s, x^t)$, is then obtained by summing the partition conditioned component density functions $\hat{f}_{B_m}(x) = \hat{f}_{B_m}(x^s, x^t)$ to obtain

$$\hat{f}_B(x) = \hat{f}_B(x^s, x^t) = \sum_{m=1}^{M} \hat{\beta}_m \hat{f}_{B_m}(x^s, x^t) \tag{7}$$

with mixing weights $\hat{\beta}_m = P_{B_m}$. Note that $\hat{f}_B(x) = f(x)$ for all values of x. 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 8,

$$\hat{f}^{\text{tree}}_{B_m}(x^t) = \int \hat{f}_{B_m}(x^s, x^t)\, dx^s = \frac{1}{P_{B_m}} \int_{\{x|x^s \in B_m\}} \hat{f}(x^s, x^t)\, dx^s \tag{8}$$

and an index for the conditioned component densities may be obtained by simply 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.

By shifting the emphasis 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, the indexing problem has been made much simpler, but there was a trade-off. The relationships between the stand scale attributes and the tree scale attributes that were implicitly defined by each component density

function $f_i(x) = f_i(x^s, x^t)$ have been lost. The partition conditioned mixture density, instead, provides a blending of the tree scale attribute distributions from all component density functions whose stand scale marginal support sets $S_i|_{x^s}$ that have nonempty intersection with a partition set B_m in the partition B. Decoupling the stand scale and tree scale attributes from their individual component density functions by using the partition B requires that the following assumption be made.

Assumption 1 (**A1**) Forest stands or patches having similar stand (course) scale attributes, e.g., site index, QMD, average height, species composition, etc., are also similar at the tree (fine) scale, provided that a sufficient number of stand (coarse) scale attributes are used to describe them.

Stated another way, if forest stands or patches may be classified using stand scale attributes, then the distributions of tree scale attributes for stands assigned to the same class are statistically indistinguishable. The partition B, then, simply provides a collection of pigeon holes B_m for the classification of stands much like the bins of a histogram.

2.1.1 Including discrete valued attributes

The inclusion of discrete valued attributes was mentioned when describing the vector of stand scale and tree scale attributes $x = (x_s, x_t)$, but the notation used implied that attribute values were continuous. This was done for notational convenience given th following understanding of the role played by the discrete parameters.

Consider a probability density function $f(x) = f(x^C, x^D)$ representing the distribution of a d-dimensional attribute vector x containing d_C continuous attribute values and d_D discrete attribute values, $d = d_C + d_D$, with $x = (x^C, x^D)$ partitioning the attribute vector into its continuous x^C and discrete x^D components. Let N_D represent the number of unique values X_i^D for the discrete attributes x^D, and let p_i, $i = 1, 2, \ldots, N_D$ be the probabilities of occurrence for each of the values X_i^D. Then a density function $f_C(x^C)$ depending only on the continuous attributes may be obtained by creating the mixture distribution

$$f_C(x^C) = \sum_{i=1}^{N_D} p_i f_i(x^C) \tag{9}$$

where

$$f_i(x^C) = \frac{f(x^C, x^D \mid x^D = X_i^D)}{\int f(x^C, x^D \mid x^D = X_i^D) dx^C} \tag{10}$$

is the conditional density function describing the distribution of the continuous attributes associated with the discrete value X_i^D. The discrete attribute values act as conditioning events that partition the continuous attributes, which may then be represented as continuous conditional probability density functions. The inclusion of discrete attributes, then, simply increases the number of terms in a mixture density representation of a distribution of forest attributes, providing a more detailed partitioning of the continuous attributes.

2.2 HNNTLG implementation

A nonparametric, data based representation for the regional mixture distribution of stand scale and tree scale attributes, and their respective probability density functions, was used in the implementation to allow the data to "speak for itself" in determining the shape of the regional mixture distribution $f(x)$ and the partition conditioned mixture distribution $f_B(x)$. The implementation combines a straightforward partition

of the domain of the stand scale attributes with a mapping associating individual tree data from a collection of sampled stands with the partition set B_m having the most similar index point b_m in the partition B to the stand scale attributes from each sampled plot. This process yields an approximation to the probability density function representing the conditional distribution of tree attributes within each partition set B_m via the actual tree measurement data associated with the partition sets. The tree data associated with each partition set are then used to generate simulated tree attributes having similar distributional properties to the trees associated with the partition set.

The implementation has two components. The first 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 from a stand scale partition B. The second uses a mapping created by the first component with a stand scale attribute vector to identify similar partition sets whose associated tree attributes are then 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 x^s be a d_s dimensional vector of stand scale attributes, some of which may be discrete valued, and let $h = (h_1, h_2, \ldots, h_{d_s})$ be a vector of bin widths for each attribute, where $h_j = 0$ for discrete valued attributes and $h_j > 0$ for attributes having continuous values, then an index point b_m that is the center of a bin having edge lengths h may be obtained as in Equation 11,

$$b_{mj} = \begin{cases} x_j^s & \text{if } x_j^s \text{ is discrete} \\ h_j \left(\lfloor \frac{x_j^s}{h_j} \rfloor + \frac{1}{2} \right) & \text{if } x_j^s \text{ is continuous} \end{cases}, j = 1, 2, \ldots, d_s \tag{11}$$

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 was generated as a sequence of bins B_m as defined in Equation 12 using the bin centers b_m as index points uniquely identifying each partition bin.

$$B_m = \left\{ x^s \middle| \begin{cases} x_j^s = b_{mj} & \text{if } x_j^s \text{ is discrete} \\ x_j^s \in [b_{mj} - \frac{h_j}{2}, b_{mj} + \frac{h_j}{2}) & \text{if } x_j^s \text{ is continuous} \end{cases}, j = 1, 2, \ldots, d_s \right\} \tag{12}$$

2.4 Computing stand scale similarity scores

The similarity between two stand scale attribute vectors x^s and y^s was computed using the bin width weighted Euclidean distance between the vectors as in Equation 13, with weights given in Equation 14.

$$S(x^s, y^s) = \sum_{j=1}^{d_s} w_j (x_j^s - y_j^s)^2 \tag{13}$$

$$w_j = \begin{cases} 10000 & \text{if } x_j^s \text{ is discrete} \\ \frac{1}{h_i} & \text{if } x_j^s \text{ is continuous} \end{cases}, j = 1, 2, \ldots, d_s \tag{14}$$

The value of 10000 was used as the weight for all discrete valued stand attributes, but different weights could have been assigned to different discrete attributes to reflect differences in the impact of deviations among the discrete attribute values when assessing the similarity of two stand scale attribute vectors. For the TLGDB implementation, however, a clear indication of differences occurring for discrete attribute values was desired for the similarity score. 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 distances across the continuous attributes used in the partition.

2.5 Building the stand scale to tree scale mapping

The algorithm for the stand scale to tree scale mapping is best described from the perspective of a collection of data sets from sampled forest stands rather than by using 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 is, therefore, necessary. Define $X = (X^s, X^t)$ to 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 d_t-dimensional vectors of observed tree attributes X_l^t, $l = 1, 2, \ldots, n$ obtained from the n sampled trees. The vector of stand scale attributes X^s was assumed to have been drawn from the distribution $\hat{f}_{B_k}^{\text{stand}}(x^s)$ for some partition set B_k, $X^s \sim \hat{f}_{B_k}^{\text{stand}}(x^s)$, and the vectors of observed tree attributes X_j^t were, similarly, assumed to have been drawn from $\hat{f}_{B_k}^{\text{tree}}(x^t)$, $X_j^t \sim \hat{f}_{B_k}^{\text{tree}}(x^t)$.

Let $X_1, X_2, \ldots, X_N$ represent observed stand and tree attributes $X_i = (X_i^s, X_i^t)$ from a collection of N sampled forest stands, each sample containing attributes from n_i sampled trees X_{il}^t, $l = 1, 2, \ldots, n_i$. The dimensions of the stand and tree scale attribute vectors, d_s and d_t, respectively, and the attributes measured are assumed to be the same for all N sampled stands. This implies that all of the sampled stands are of the same treatment class, e.g., untreated stands, naturally regenerated stands, planted stands, thinned stands, etc., or that the stand scale and tree scale attribute vectors contain components supporting all possible treatment types with discrete attributes acting as indicators for the different treatment types identifying the appropriate subsets of attribute values. Since treatment flags simply partition the attribute vectors into disjoint subsets, as described in Section 2.1.1, all of the sampled stands were simply assumed to be from the same treatment class.

With this notation and the partition defined in Section 2.3, a mapping between the bin centers b_m and the measured tree attributes X_i^t for each 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 from the partition bin center to the tree attributes T_m associated with that bin. The tree attributes associated with each partition bin, then, define the distribution of tree scale attributes for a stand that is representative of the forest conditions within that partition bin.

The mapping is constructed dynamically by building the partition one bin at a time, while maintaining a dictionary order on the partition bin centers b_m indexing the partition bins B_m and their associated tree attributes T_m. A sample $X_i = (X_i^s, X_i^t)$ is added to a mapping having $M \geq 0$ bins in its partition $B$by first, computing the partition bin center b from the vector of stand scale attributes for that sample, X_i^s, using Equation 11. Next, if the partition bin center b is already in the index of bin centers, say as b_m, add the tree attribute vectors X_{il}^t, $l = 1, 2, \ldots, n_i$ to the list of tree attributes T_m associated with partition bin B_m. If b is not in the index, it is a new partition bin, so add the bin center b to the index of partition bins as bin center b_{M+1}, and then add the tree attribute vectors X_{il}^t, $l = 1, 2, \ldots, n_i$ to the list of tree attributes T_{M+1} associated with the new partition bin B_{M+1}. Finally, if a new bin was added to the partition, sort the index of partition bin centers. By building the partition dynamically, the minimum number of partition bins necessary to represent the stand scale to tree scale mapping are generated for a collection of sampled stands.

An example of adding ten sampled stands to an initially empty partition is now described. The stands $X_1, X_2, \ldots, X_{10}$ were added to the partition in the order specified by their subscripts. For the example, assume that stands X_1, X_7 and X_{10} are similar, that stands X_2 and X_4 are similar, and that stands X_5 and X_6 are similar. All other stands are assumed to be different. Similar stands are mapped to the same partition

Table 1: Example of a stand scale to tree scale mapping created using ten sample stands. The stands were added to an initially empty partition, in order, from X_1 to X_{10}, and the bin centers were created in sequence, but are ordered as shown in the table.

Index position	Partition bin center	Associated tree attributes
1	b_4	$T_4 = (X_5^t, X_6^t)$
2	b_1	$T_1 = (X_1^t, X_7^t, X_{10}^t)$
3	b_5	$T_5 = (X_8^t)$
4	b_2	$T_2 = (X_2^t, X_4^t)$
5	b_3	$T_3 = (X_3^t)$
6	b_6	$T_6 = (X_9^t)$

bin center, and the partition bin centers were assumed to be in the order $b_4 < b_1 < b_5 < b_2 < b_3 < b_6$. The steps of the algorithm are shown in detail for a few stands to demonstrate how stands, and their respective tree attributes, were associated with the partition bins in the example.

To add stand X_1, first compute the partition bin center b. The partition is empty, so add b to the index as b_1, increment the number of partition bins M by one, and associate the tree attributes X_1^t with B_1. Now add stand X_2. Compute the partition bin center b, which represents a new partition bin since X_1 was different than X_2, and add b_2 to the index, increment M, and associate the tree attributes X_2^t with B_2. Stand X_3 is added next. It is different from stands X_1 and X_2, so its partition bin center b is added to the index as b_3, M is incremented, and the tree attributes X_3^t are associated with B_3. Stand X_4 is added next, but it is similar to stand X_2, so its partition bin center b is identical to b_2 which is already in the index of partition bin centers. In this case, the tree attributes X_4^t are associated with B_2. The remaining six stands are added similarly, creating three new partition bins whose centers are added to the index. The outcome, consisting of the ordered index of partition bin centers and their associated tree attribute data, is given in Table 1.

2.6 Generating simulated tree attributes

A simulated vector of tree attributes $\hat{Y}^t$ representing a tree similar to those associated with a vector of stand attributes Y^s is generated from a partition B, using its index of partition bin centers b_m, and its mapping associating sampled tree attributes with the partition bins B_m, in two steps. First, the $k \le K_{\max}$ 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 $T_{m_1}, T_{m_1}, \ldots, T_{m_k}$ 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 $T_l = (T_{l1}, T_{l2}, \ldots, T_{ld_t})$, $l = 1, 2, \ldots, N_T$. The tree attribute vectors T_l are assumed to be a random sample of size N_T from the unknown multidimensional probability density function $f_{B_m}^{\text{tree}}(x^t)$.

Second, the tree attribute vectors T_l, $l = 1, 2, \ldots, N_T$ in the canonical tree list T are split 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. Simulated tree attributes were then generated using a combination of a bootstrap method (Efron, 1982, Efron and Tibshirani, 1998, Davison and Hinkley, 2003) for the discrete attributes and the SIMDAT algorithm for the continuous attributes (Taylor and Thompson, 1986, Thompson, 2000).

A simulated tree scale attribute vector $\hat{T} = (\hat{T}^C, \hat{T}^D)$ was generated by first generating a random integer

r, $1 \leq r \leq N_T$. This identifies a reference tree $T_r = (T_r^C, T_r^D)$. Values for the discrete attributes of the simulated tree attribute vector $\hat{T}^D$ were obtained directly from the reference tree, $\hat{T}^D = T_r^D$. Values for the continuous attributes of the simulated tree attribute vector $\hat{T}^C$ require a bit more effort. Let T^C be represented as an $N_T \times d_t$ data matrix, as in Equation 15,

$$T^C = \begin{pmatrix} T_{11}^C & T_{12}^C & \cdots & T_{1d_t}^C \\ T_{21}^C & T_{22}^C & \cdots & T_{2d_t}^C \\ \vdots & \vdots & \ddots & \vdots \\ T_{N_T1}^C & T_{N_T2}^C & \cdots & T_{N_Td_t}^C \end{pmatrix} \tag{15}$$

and apply the variance normalizing transformation $Z = T^C S^{-1}$ to the tree attribute vectors, stored as the rows of T^C, where $S_{jj} = s_j$ for $j = 1, 2, \ldots, d_t$, s_j is the standard deviation of column j, and all other elements of S are zero. This normalization removes differences in variance among the d_t coordinate dimensions.

Determine the K_T nearest neighbors to the normalized attribute vector Z_r, from the reference tree, by computing the Euclidean distance from Z_r to Z_l, $D_{rl} = \sqrt{(Z_r - Z_l)^t (Z_r - Z_l)}$, for $l = 1, 2, \ldots, N_T$, and then selecting the normalized attribute vectors associated with the K_T smallest distances. The normalized reference vector Z_r is always selected as one of the nearest neighbors, as it is the closest vector to itself with Euclidean distanceequal to zero. Relabel the K_T nearest neighbors Z_l, $l = 1, 2, \ldots, K_T$. Transform the K_T nearest neighbors by subtracting their sample mean, $\bar{Z} = \frac{1}{K_T} \sum_{l=1}^{K_T} Z_l$, yielding the data vectors $Z_l' = Z_l - \bar{Z}$, $l = 1, 2, \ldots, K_T$, and compute the weighted sum in Equation 16,

$$Z' = \sum_{l=1}^{K_T} u_l Z_l' \tag{16}$$

where $u_1, u_2, \ldots, u_{K_T}$ are random variables generated from the uniform distribution in Equation 17.

$$U\left(\frac{1}{K_T} - \left[\frac{3(K_T - 1)}{K_T^2} \right]^{\frac{1}{2}}, \frac{1}{K_T} + \left[\frac{3(K_T - 1)}{K_T^2} \right]^{\frac{1}{2}} \right), \tag{17}$$

Finally, compute the continuous simulated tree attribute vector by transforming the normalized vector Z' into its original coordinates using Equation 18.

$$\hat{T}^C = \left(Z' + \bar{Z} \right) S \tag{18}$$

If the tree attribute vector $\hat{T}^C$ has appropriate values for all of its components, then $\hat{T} = (\hat{T}^C, \hat{T}^D)$, otherwise repeat the SIMDAT procedure until an acceptable vector $\hat{T}^C$ is obtained. Ultimately an acceptable simulated tree attribute vector $\hat{T} = (\hat{T}^C, \hat{T}^D)$ is obtained, and this is assigned to $\hat{Y}^t$, $\hat{Y}^t = \hat{T}$. The procedure just described is repeated, selecting a new reference tree T_r, with replacement, for each desired simulated tree attribute vector.

The uniform distribution in Equation 17 was chosen to ensure that the vectors Z generated using the SIMDAT procedure were uncorrelated and that the first and second moments, the mean and covariance, of the simulated data vectors approximated those of the original data (Taylor and Thompson, 1986, Thompson, 2000). An asymptotic dependence exists between K_T, the smoothing or averaging parameter, and the sample size N_T, as for the sample size and the bin width or number of bins for histograms and a variety of other nonparametric probability density estimators (Silverman, 1986, Gehringer, 1990, Thompson and Tapia, 1990, Gehringer and Redner, 1992, Redner and Gehringer, 1994, Redner, 1999). For practical purposes a fixed value of K_T may be chosen without incurring a significant penalty, and empirical results support this conclusion.

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 to generate the results in this paper. If fewer than K_T trees were available, all of the available trees were used to generate the vector $\hat{T}^C$.

Values for the maximum number of partition bins $K_{\max}$ and the maximum similarity score $S_{\max}$ may vary, depending on the application, but a values of $K_{\max} = 5$ and $S_{\max} = 5.0$ were used to select the partition bins and create the canonical tree lists for this paper. Fewer than $K_{\max}$ partition bins may be selected depending on the value chosen for $S_{\max}$ and sparsity of the partition bin centers near the desired stand attribute vector Y^s. The maximum similarity score $S_{\max}$ filtered out partition bins having discrete stand scale attributes that did not agree with those in the stand attribute vector Y^s and partition bins that were far from Y^s.

A preprocessing step may be applied to the tree attribute data in the canonical tree list T prior to generating simulated tree attributes. 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.

The procedures just described work directly with actual data to generate simulated tree attributes having the same, or nearly the same, covariance structure as the original data. The procedures also automatically guarantee that the tree attributes within a simulated vector will be consistent and compatible, as the simulated attributes were generated simultaneously and were derived directly from *actual* tree measurement data. The only requirement to use these procedures data representative of tree attributes for a range of stand attributes. In addition, the assumption **A2** seems warranted.

Assumption 2 (**A2**) Attributes of forest inventory data at the tree (fine) scale collected from sample plots within a larger stand or forest patch may be extended to the whole stand or patch (coarse) scale.

2.7 Validation and goodness of fit testing

To validate 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, Stand Management Cooperative, 1992). The TLGDB was designed to generate simulated tree lists for pure and mixed species stands of Douglas-fir (*Pseudotsuga menziesii*) and western hemlock (*Tsuga Heterophylla*) as the 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 are supported by the TLGDB: untreated stands and thinned stands. Only untreated stands are considered here, results for the thinned stands were comparable and are described elsewhere (Gehringer and Turnblom, 2001).

The primary tasks performed by the TLGDB programs are the addition of stand description and tree measurement data from sampled forest stands to an existing TLGDBand the generation of simulated tree lists from a TLGDB. Stand description and tree measurement data from sampled forest stands are added to a TLGDB using the `TGADD` program which builds and maintains the stand scale to tree scale partition bin mapping described in Section 2.5. Simulated tree lists are generated from a TLGDB using the `TGRAND`

Table 2: Stand and tree scale attributes in the SMC TLGDB with partition bin widths for the continuous stand attributes.

Attribute	Scale	Type	Bin width h_j
QMD (cm)	Stand	Continuous	4.0
Site Index 50 (m)	Stand	Continuous	3.0
Stand density (TPH)	Stand	Continuous	200
Stand origin	Stand	Discrete	0
Stand type	Stand	Discrete	0
Total age (years)	Stand	Continuous	4.0
DBH (cm)	Tree	Continuous	N/A
Height (m)	Tree	Continuous	N/A
Species	Tree	Discrete	N/A

program, which finds the nearest partition bins to a specified stand description and generates simulated trees as described in Section 2.6. The `TGRAND` program supports a preprocessing mode to generate 100% pure stands of Douglas-fir or western hemlock if a pure stand is desired, and a post processing mode to randomly modify or jitter the heights of simulated trees to mitigate the impact of reduced height variability caused by the use of height–diameter relationships to estimate heights in the data used to populate the TLGDB.

Data used to populate a TLGDB are provided to `TGADD` via a measurement file (MF) that contains values for a variety of stand scale attributes and a list of measured tree attributes in a "keyword equals value" file format. A random measurement file (RMF), whose format is identical to that of an MF, is generated by `TGRAND` using a description file (DF) containing values for the stand scale attributes of the desired RMF. In addition to the stand scale and tree scale attributes, an MF and RMF also contain other information, in particular keywords for the units of measurement for values within the files, and the size of the sample plot or the size of the area that a tree list in an RMF is generated to represent. A DF also contains a stand density value that is used with the area represented by an RMF to determine the number of trees to generate. An RMF is thought of as a simulated sample whose extent and tree count are controlled by the area being represented and the desired stand density.

The stand and tree scale attributes used in the SMC TLGDB are presented in Table 2 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 and stand origin were discrete attributes, and total stand age, site index at 50 years, quadratic mean diameter (QMD), and stand density measured as trees per hectare (TPH) were continuous attributes used to generate the partition. Five stand types are supported: 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. Two stand origins are also supported: planted or natural regeneration. 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. Values for the discrete attributes, whether at the stand or tree scale, were represented using unique integer codes within the TLGDB and its suite of programs.

A simulated sample stand in an RMF consists of compatible DBH, height, and species values, for each of the simulated trees, with values for the stand scale attributes total age, site index, stand origin, and plot

Table 3: Data sources for the SMC TLGDB.

British Columbia Ministry of Forests
Canadian Forest Service
Oregon State University, Corvallis, Oregon
Port Blakely Tree Farms
Regional Forest Nutrition Research Project (RFNRP), SMC
Stand Management Cooperative (SMC), University of Washington
USFS Pacific Northwest Research Station
Washington State Department of Natural Resources
Weyerhaeuser Company

size being copied from the DF to the RMF. Stand type is implicitly defined in an RMF by the generated tree list. Jittered heights were computed in `TGRAND` by adding a uniform random number u generated from the interval $(-2\hat{H}_{\mathrm{MAI}}, 2\hat{H}_{\mathrm{MAI}})$ to a simulated tree height, where $\hat{H}_{\mathrm{MAI}}$ was the mean annual height increment for the simulated tree.

2.7.1 Data

The SMC TLGDB was intended to provide a means for generating simulated tree lists representative of Douglas-fir and western hemlock forests in the Pacific Northwest, west of the Cascade Mountains, extending from southern Oregon to southern British Columbia. Regional stand measurement data were provided by the organizations listed in Table 3. The majority of data provided for the SMC TLGDB were proprietary, so plot locations are not presented.

Given the number of data sources, a myriad of data collection and sampling strategies were likely employed, with differing assumptions. A variety of unknown height-diameter relationships would also have been used to estimate heights for trees whose heights were not measured. Even if the data collection histories, data sampling protocols, and other statistical procedures were known, it would be nearly impossible to reconcile discrepancies that would undoubtedly be present. Therefore, no attempt was made to address the statistical compatibility of the sampling strategies or other procedures that were employed to obtain these data sets. The data sets were assumed to be compatible and their tree lists were used *as is*. To increase the reliability of the TLGDB only trees from sample plots that were at least 0.0405 hectares in size were used.

An indication of stand origin, natural regeneration, seeded, planted, or unknown, was usually provided for each sampled stand. The four stand origins observed in the data were mapped to the planted and natural regeneration stand types in the SMC TLGDB as follows. Stands whose origins were unknown were assumed to have been naturally regenerated. Seeded stands were also assumed to be naturally regenerated. Stand origins that were planted or natural regeneration were assigned the appropriate category. Any stand having an undefined stand origin was removed from the data set and not considered further.

The majority of the sample data contained breast height ages, but stand total ages were desired for the SMC TLGDB. Stand total ages for Douglas-fir were computed from stand breast height ages using Bruce's equations (Bruce, 1981). Stand total ages for western hemlock stands were computed in the same manner, as a breast height age to total age conversion formula for western hemlock was not readily available. The breast height age to stand total age conversion is given in Equation 19, where A_{TOT} is stand total age, A_{BH}

Table 4: Stand type and origin summary for untreated stands used to populate the TLGDB.

Stand type	Natural	Planted	Total	Percent
Pure Douglas-fir	2603	801	3404	65.3
Pure western hemlock	550	250	800	15.4
Douglas-fir dominant	479	212	691	13.3
Western hemlock dominant	112	66	178	3.4
Mixture	118	18	136	2.6
Total	3862	1347	5209	100.0

is breast height age, and SI is site index at age 50.

$$A_{\mathrm{TOT}} = A_{\mathrm{BH}} + 13.25 - \frac{\mathrm{SI}}{20} \tag{19}$$

A species specific 50 year site index value was usually provided for pure Douglas-fir and pure western hemlock stands. For mixed Douglas-fir and western hemlock stands, if two site index values were provided, the site index value used was the value for the species with the greater percentage of stand basal area. If a single species specific site index value was provided for a stand, and that species had the lower percentage of stand basal area a site index conversion equation was used to convert the given site index value to a site index value for the more dominant species (Nigh, 1995). The Douglas-fir to western hemlock site index conversion was performed using Equation 20, and Equation 21 was used to convert a western hemlock site index value to a Douglas-fir site index value.

$$\mathrm{SI}_{\mathrm{WH}} = 0.432 + 0.899 \cdot \mathrm{SI}_{\mathrm{DF}} \tag{20}$$

$$\mathrm{SI}_{\mathrm{DF}} = 0.480 + 1.11 \cdot \mathrm{SI}_{\mathrm{WH}} \tag{21}$$

In both equations the site index values $\mathrm{SI}_{\mathrm{DF}}$ and $\mathrm{SI}_{\mathrm{WH}}$ were assumed to be for a reference age of 50 years. If no site index value was provided for a sample stand, that stand was removed from the data set and not considered further.

Data consisting of compatible DBH and height measurements with an indication of tree species for 573,036 individual trees were obtained from a total of 5209 sample plots distributed throughout the region of interest. The breakdown by stand type and assigned stand origin is given in Table 4. The majority of the sample plots, 65.3%, were from pure Douglas-fir stands, with pure western hemlock stands representing 15.4%, and Douglas-fir dominant stands accounting for 13.3% of the sample plots, and the remaining 6.0% were from western hemlock dominant and mixed stands. Only standing live trees were included in the SMC TLGDB and these trees, with their expansion factors, were used to compute QMD and TPH values from each included stand for the stand scale partition.

The 5209 sampled stands produced 3139 unique partition bins. A statistical summary of the continuous stand scale attributes using the partition bin centers b_m, by stand type, is given in Table 5. The dimension of the stand attributes used for the partition introduces a combinatorial explosion in terms of presenting any sort of accessible, detailed data summary. Summarizing the data coverage in a TLGDB is further complicated by the fact that it may contain gaps, and there is no convenient manner in which to briefly summarize this information. With these limitations in mind, the summary statistics in Table 5 can only provide an incomplete overview of the data represented within the SMC TLGDB, indicating nominal coverage ranges of the continuous stand scale attributes for each stand type.

Table 5: Stand scale data coverage summary, by stand type, for untreated stands used to populate the SMC TLGDB.

Stand type (count)	Attribute	Mean	Std. Dev.	Minimum	Median	Maximum
Pure Douglas-fir (1921)	QMD (cm)	21.6	11.0	2.0	18.0	74.0
	Site Index (m)	32.6	7.1	16.5	31.5	49.5
	TPH	1651.0	1151.3	100.0	1300.0	9700.0
	Total age	49.7	18.0	6.0	46.0	126.0
Pure western hemlock (493)	QMD(cm)	17.6	7.9	6.0	14.0	42.0
	Site Index (m)	36.9	9.9	19.5	34.5	58.5
	TPH	3692.3	2852.8	700.0	2700.0	14100.0
	Total age	46.9	16.3	18.0	46.0	94.0
Douglas-fir dominant (491)	QMD (cm)	18.8	8.0	6.0	18.0	50.0
	Site Index (m)	33.6	6.7	16.5	34.5	43.5
	TPH	2210.4	1397.1	100.0	1900.0	9500.0
	Total age	49.0	16.6	14.0	46.0	106.0
Western hemlock dominant (138)	QMD (cm)	21.3	6.6	10.0	22.0	42.0
	Site Index (m)	39.6	8.0	25.5	37.5	55.5
	TPH	2056.5	1107.6	700.0	1700.0	5500.0
	Total age	48.4	15.1	18.0	46.0	86.0
Mixture (96)	QMD (cm)	22.3	6.6	10.0	22.0	38.0
	Site Index (m)	38.3	6.0	28.5	40.5	46.5
	TPH	1535.4	1182.7	300.0	1000.0	4700.0
	Total age	53.8	11.9	22.0	54.0	74.0

2.7.2 Goodness of fit methods and criteria

To assess the performance of the HNNTLG procedures and the SMC TLGDB, a simulated sample stand having the same stand scale attributes and sample plot size was generated for each of the actual sample stands used to populate the TLGDB using the five stand scale attributes listed in Table 2. A nonparametric goodness of fit (GOF) statistic was then 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.

A nonparametric GOF procedure was chosen for several reasons. First, the distributions of DBH and height are frequently multimodal and they may also be strongly skewed. Second, the classical goodness of fit testing procedures may be inappropriate in situations where sample sizes are large, large sometimes being as small as 50 or 100 sample points, spuriously indicating that two data sets are different when they were in fact obtained from identical, or indistinguishable, underlying distributions (Bickel and Doksum, 1978, Cochran, 1952). Third, The tree size distributions and the species composition within a forest stand are its fundamental measured components, and the GOF testing procedures used to compare the actual sample stands and the simulated sample stands should emphasize the shapes of the distributions of the tree sale attributes DBH, height, and species composition, as well as the regional distributions of QMD and average height. Finally, any GOF comparison that aggregates the data, e.g., by binning continuous data for a chi-squared test, reducing a data set to a few parameters such as a mean and variance for parametric testing as in a t-test, or that limits the comparison to a single point, as for the Kolmogorov-Smirnov test (Bickel and Doksum, 1978), must make less effective use of the available data than a comparison based on the entire range of values and their corresponding likelihoods.

The nonparametric GOF test statistic used here was based on the integrated absolute error between two functions f and g, which is defined in Equation 22.

$$\text{iae}(f, g) = \int_{-\infty}^{\infty} |f(x) - g(x)| \, dx. \tag{22}$$

The integrated absolute error is obviously related to the Kolmogorov-Smirnov test statistic, and it has been used to compute an index for comparing diameter distributions (Borders and Patterson, 1990, Reynolds et al., 1988). The integrated absolute error may also be the natural GOF statistic for comparing probability density functions (Devroye and Gyorfi, 1985), and it has been used effectively to compare nonparametric estimates of probability density functions (Gehringer, 1990, Gehringer and Redner, 1992)

If the functions f and g in Equation 22 are PDFs the integrated absolute error computes the difference in the location of probability mass for the two functions, producing values in the interval $[0, 2]$, with a value of zero indicating that the functions f and g are indistinguishable, and a value of two indicating that the functions f and g had no overlap. A transformation of the integrated absolute error was performed, as in Equation 23, to obtain a statistic, the p_{iae}-value, having a range that was consistent with that of p-value in the classical hypothesis testing framework.

$$p_{\text{iae}} = 1 - \frac{1}{2} \int_{-\infty}^{\infty} |f(x) - g(x)| \, dx, \tag{23}$$

The p_{iae}-value has a range from zero, indicating that two distributions have no overlap, to one, indicating that the two distributions are indistinguishable.

The p_{iae}-value may be interpreted as the proportion of probability mass that two PDFs f and g have in common, or, when multiplied by 100, the percent similarity in the total area under the curves of the two PDFs. This interpretation of the p_{iae}-value is consistent with the total error (Type I+ Type II) of

classical hypothesis tests, but unlike those test statistics the p_{iae}-value *automatically* works for multimodal and skewed distributions, both of which frequently occur in forestry data. The p_{iae}-value is also readily modified to work with discrete valued PDFs f and g defined over a finite or infinite set of distinct values, $X = \{x_1, x_2, x_3, \ldots, x_n\}$ or $X = \{x_1, x_2, x_3, \ldots\}$, as in Equation 24, allowing the same GOF statistic to be used for continuous and discrete values.

$$p_{\text{iae}} = 1 - \frac{1}{2} \sum_{\text{all } x_i} |f(x_i) - g(x_i)| \tag{24}$$

A p_{iae}-value for two samples from discrete or categorical distributions f and g was computed directly from the formula in Equation 24 using empirical estimates of the proportion for each of the values obtained from either distribution. A p_{iae}-value for two samples from continuous distributions f and g was computed by estimating the underlying PDFs for each sample (Silverman, 1986, Gehringer, 1990, Thompson and Tapia, 1990, Redner and Gehringer, 1994, Gehringer and Redner, 1992, Redner, 1999) and then using a numerical integration scheme to approximate the integral in Equation 23. Estimates $\hat{f}$ and $\hat{g}$ of the PDFs for the two samples were computed using a nonparametric technique based on cubic B-splines (Gehringer, 1990, Gehringer and Redner, 1992, Redner and Gehringer, 1994, Redner, 1999). Numerical integration of the formula in Equation 23 was performed using a straightforward midpoint based method. The interval width for the midpoint method was chosen to ensure that the numerical integrations of $\hat{f}$ and $\hat{g}$ were within $\epsilon = 0.01$ of one, the total probability for a PDF, prior to computing the p_{iae}-value, that is, $|1 - \int \hat{f}(x)dx| < \epsilon$ and $|1 - \int \hat{g}(x)dx| < \epsilon$. A histogram could also have been used to compute a p_{iae}-value, as it is a nonparametric probability density function estimator. In this case the continuous PDFs would be converted into discrete PDFs having probabilities associated with the histogram bins, and the summation formula in Equation 24 would replace the integration formula in Equation 23.

To effectively use the p_{iae}-value as a GOF test statistic an appropriate p_{iae} cutoff value, or critical p_{iae}-value, must be determined. The critical p_{iae}-value plays the same role as the critical value, or α-level, in a classical goodness of fit or hypothesis test. To guide the selection of a critical p_{iae}-value a simulation of a real sampling scenario was performed using the standard normal distribution $N(0, 1)$. The simulation consisted of drawing a random integer n uniformly distributed between 20 and 50 for a sample size, generating two independent random samples of size n from the standard normal distribution $N(0, 1)$, and computing the p_{iae}-value for the two samples using Equation 23 and the numerical integration of the nonparametric PDF estimates. This process was repeated 10000 times to approximate the distribution of the p_{iae}-valueunder the null hypothesis of indistinguishable distributions. The histogram of p_{iae}-values obtained from the simulation is presented in Figure 1. The p_{iae}-value distribution is clearly not symmetric and is strongly left skewed, with the bulk of the p_{iae}-values being greater than 0.65.

A p_{iae}-value of 0.65 should, therefore, provide a reasonable critical value to use in practice, based on the simulation just performed, with p_{iae}-values > 0.65 indicating similar distributions, and p_{iae}-values ≤ 0.65 indicating different distributions. The critical p_{iae}-value 0.65 is likely to be conservative. If the underlying distribution for two samples was multimodal, was not strongly unimodal like the normal distribution, or if the distribution had greater variability than the normal distribution, then a smaller critical p_{iae}-value would be expected when comparing random samples drawn from such a distribution. The p_{iae}-value also has a dependence on sample size, much like that of the t-test and other test statistics. As the sample size increases, the p_{iae}-value converges to a value of one under the null hypothesis of indistinguishable distributions, and for small sample sizes, the critical p_{iae}-value would be likely have a smaller value than 0.65.

For each pair of simulated and actual stands, p_{iae}-values were computed for the DBH and height distributions and the species composition and were then used to determine empirical misclassification rates. If a

Table 6: The GOF testing scenarios for the four tree list generation modes of `TGRAND`.

Scenario	100% pure stands	Jitter heights
RMF (default)	No	No
RMFJ	No	Yes
RMFP	Yes	No
RMFJP	Yes	Yes

p_{iae}-value was less than the critical value of 0.65 for a particular attribute and pair of stands, a misclassification error for that attribute occurred. A total misclassification rate was also computed by performing a logical *or*ing of the three individual classification outcomes, that is, if at least one of DBH, height, or species composition was misclassified for a particular pair of actual and simulated stands, then that stand contributed to the total misclassification rate. The empirical misclassification rates computed for each attribute, and the total misclassification rate, provide the primary form of validation for the HNNTLG methodology and the SMC TLGDB.

Four testing scenarios were considered, corresponding to the four possible tree list generation modes, listed in Table 6, that may be used with the `TGRAND` program to generate a simulated stand. The stand generation modes were determined by combinations of the pure and jitter preprocessing and postprocessing steps implemented in `TGRAND`. Figures are provided only for scenario RMF, the default `TGRAND` usage scenario, results for the other three testing scenarios were similar. Tables summarizing the GOF results for all four testing scenarios are, however, presented. The RMF testing scenario is the worst case scenario for stands generated using `TGRAND`, since the preprocessing step filters the canonical tree list to exclude all but one species and the post processing step adds variability by jittering heights.

Results were summarized in three different ways. First, DBH, height, species composition, and total misclassification rates are presented for each testing scenario. These comparisons test the within stand, or tree scale, agreement between each actual and simulated stand. Histograms of the p_{iae}-values for the DBH, height, and species composition are also presented, with the p_{iae} critical value 0.65 indicated. These figures provide a graphical sense of the misclassification rates for different p_{iae} critical values. Second, scatter plots of simulated and actual QMD and mean height are presented with linear regression coefficients and r^2 values. Third, the distributions of QMD and mean height across all actual and simulated stands are compared by computing their p_{iae}-values, nonparametric estimates of their PDFs, and their 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, testing the between stand, or stand scale, consistency of the actual and simulated stands, providing an indication of the agreement between the simulated and actual distributions of these attributes across the region.

3 Results

The empirical misclassification rate results for untreated stands and testing scenario RMF are presented in Table 7, and the histograms of p_{iae}-values fir DBH, height, and species in Figure 2, Figure 3, and Figure 4, respectively. The figures clearly show that the bulk of the actual stand *vs.* simulated stand comparisons produced p_{iae}-values that were greater than the critical p_{iae}-value of 0.65 indicating that the majority of simulated untreated stands were similar to their actual counterparts. The misclassification rates for testing scenario RMF were 0.60% for the DBH distributions, 3.69% for the height distributions, 2.61% for the species

Table 7: Misclassification 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
RMF	0.0060	0.0369	0.0261	0.0607
RMFJ	0.0075	0.0361	0.0276	0.0618
RMFP	0.0058	0.0353	0.0257	0.0587
RMFJP	0.0081	0.0353	0.0269	0.0609

Table 8: Linear regression coefficients and r^2 values for the model $y = a + bx$ applied to the simulated (y) and actual (x) QMD and average height values. A perfect fit would give values $a = 0$ and $b = 1$.

	QMD(cm)			Mean height(m)		
Scenario	Intercept (a)	Slope (b)	r^2	Intercept (a)	Slope (b)	r^2
RMF	0.7766	0.9578	0.9558	1.1419	0.9403	0.8720
RMFJ	0.8073	0.9547	0.9534	1.1248	0.9400	0.8686
RMFP	0.7223	0.9648	0.9564	1.0018	0.9535	0.8721
RMFJP	0.7743	0.9603	0.9535	0.9919	0.9529	0.8693

composition, and 6.07% overall. The larger misclassification rates for height distributions is most likely due to a reduction in the natural variation in tree heights caused by the use of height-diameter relationships to estimate missing tree heights. A summary of the misclassification rate results for the other three testing scenarios is also provided in Table 7 for comparison.

As can be seen in Table 7, all of the testing scenarios had total correct classification rates of approximately 94%, demonstrating very good agreement between the actual and simulated untreated stands at the tree scale. The results for scenario RMFP, the 100% pure scenario, were slightly better than those for scenario RMF, having a total misclassificatin rate of 5.87%, and the two jittered scenarios, RMFJ and RMFJP, performed slightly worse than scenario RMF, having total misclassification rates of 6.18% and 6.09%, respectively. These results were consistent with expectations, since the 100% pure scenario reduced the variability in the simulated stands, while the two jittered scenarios increased the variability in the simulated stands.

The QMD data used in the simple linear regression analysis for testing scenario RMF are presented in Figure 5 and Table 8. An examination of the figure clearly indicates a strong linear relationship between the actual and simulated values for this stand attribute. The QMD intercept of 0.7766, the slope of 0.9578 and the r^2 value of 0.9558 support the observation of strong linearity. The linear regression coefficients and r^2 values for the other three testing scenarios, as well as the results for scenario RMF are presented in Table 8 for comparison. These results indicate that there is very good agreement for QMD measurements between the actual and simulated untreated stands for all four testing scenarios.

The mean tree height data used in the simple linear regression analysis for testing scenario RMF are presented in Figure 6 and Table 8. Again, the figure clearly indicates a strong linear relationship between the actual and simulated values for this stand attribute. The mean height intercept of 1.1419, the slope of 0.9403 and the r^2 value of 0.8720 support the observation of strong linearity. The linear regression coefficients and r^2 values for the other three testing scenarios, as well as the results for scenario RMF are presented in Table 8 for comparison. These results indicate that there is very good agreement for mean height measurements between the actual and simulated untreated stands for all four testing scenarios, but

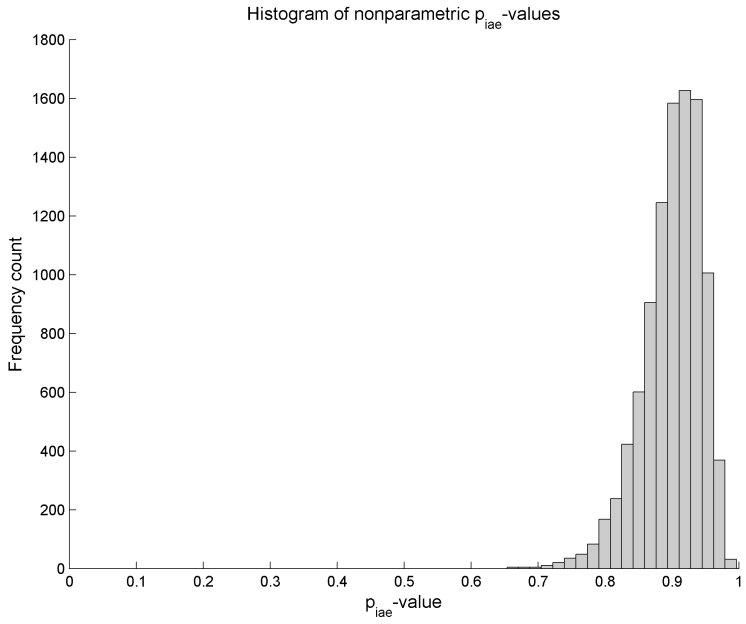

Figure 1: Histogram of 10000 p_{iae}-values generated by choosing two random samples of size n between 20 to 50 from a standard normal distribution $N(0, 1)$.

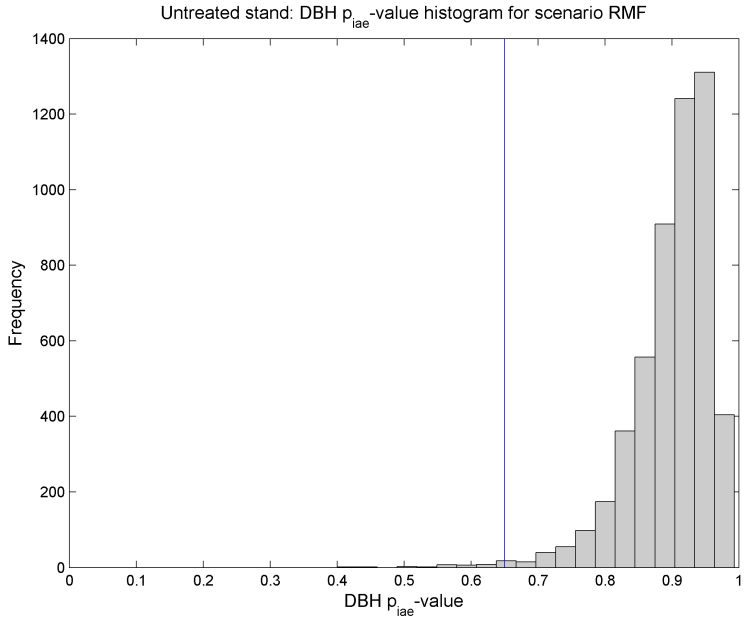

Figure 2: Histogram of p_{iae}-values for tree DBH in untreated stands for scenario RMF. The vertical line at 0.65 represents the boundary between the similar (p_{iae}-value > 0.65) and different (p_{iae}-value ≤ 0.65) distributions. The misclassification rate is 0.60%.

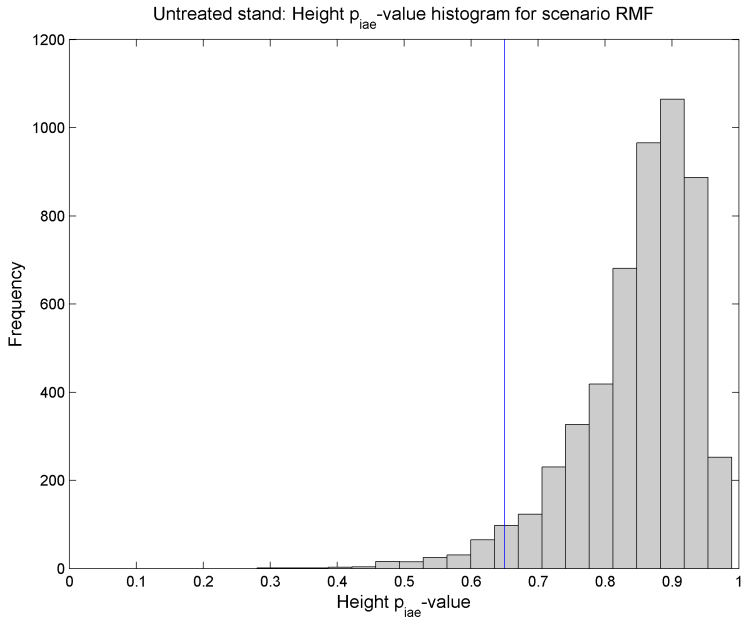

Figure 3: Histogram of p_{iae}-values for tree height in untreated stands for scenario RMF. The vertical line at 0.65 represents the boundary between the similar (p_{iae}-value > 0.65) and different (p_{iae}-value ≤ 0.65) distributions. The misclassification rate is 3.69%.

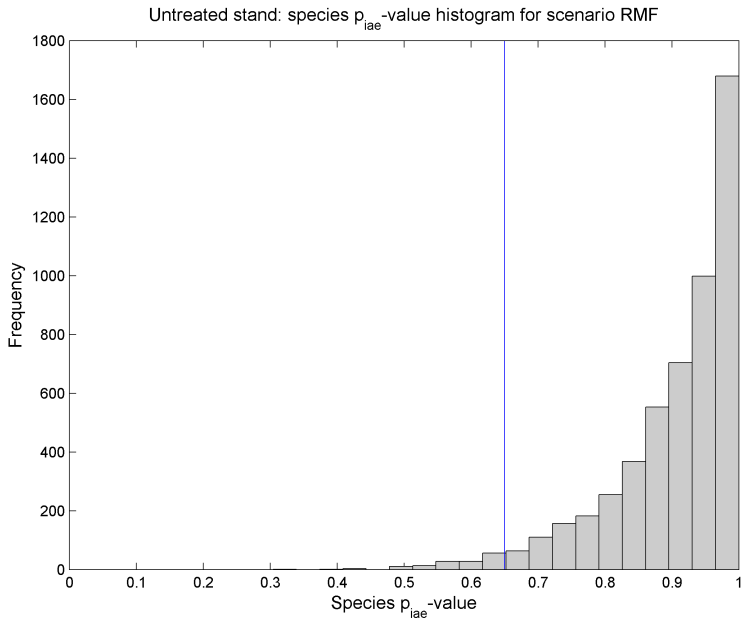

Figure 4: Histogram of p_{iae}-values for species composition in untreated stands for scenario RMF. The vertical line at 0.65 represents the boundary between the similar (p_{iae}-value ≥ 0.65) and different (p_{iae}-value < 0.65) distributions. The misclassification rate is 2.61%.

Table 9: Overall QMD and mean height distribution p_{iae}-values, bias, and RMSE for actual and simulated untreated stands for all four scenarios. See Figure 7 and Figure 8.

Scenario	QMD (cm)			Mean height (m)		
	p_{iae}-value	Bias	RMSE	p_{iae}-value	Bias	RMSE
RMF	0.9817	0.1237	1.8071	0.9825	0.0028	2.0732
RMFJ	0.9813	0.1588	1.8640	0.9811	0.0257	2.0745
RMFP	0.9854	0.0279	1.8162	0.9861	-0.1110	2.0179
RMFJP	0.9853	0.0730	1.8697	0.9859	-0.0892	2.0167

with greater variability than that seen for QMD.

Finally, nonparametric probability density estimates for the actual and simulated QMD and mean height distributions are presented in Figure 7 and Figure 8, respectively, for testing scenario RMF. The p_{iae}-values for the QMD and mean height distributions were 0.9817 and 0.9825, respectively, for scenario RMF, indicating that the actual and simulated QMD and mean height distributions were almost identical, having approximately 98% of their total probability mass in common. Results for the other three testing scenarios, as well as the results for scenario RMF, are presented in Table 9 for comparison, along with the bias and RMSE for QMD and mean height for each scenario. These results reinforce thoseobtained from the simple linear regressions, and provide another indication of the performance of the HNNTLG procedures and the SMC TLGDB. Notice, in particular, that the distribution of mean height for the simulated stands reproduces the kink in the actual mean height distribution near the value of 15 m.

4 Discussion

The primary benefits of the HNNTLG approach for tree list generation are provided by the use of an implicit two-scale relationship between coupled coarse and fine scales, with a direct partitioning of the coarse scale attributes, and a mapping associating fine scale attributes with the uniquely determined partition sets. The direct partitioning of the coarse scale attributes provides a classification of those attributes and a corresponding conditioning of the fine scale attributes that is independent of the statistical properties any particular data set. The HNNTLG method and the TLGDB were also designed to be dynamic, that is, they were designed to allow the addition of new data to an existing TLGDB partition and to allow a refinement of the coarse scale attribute partition when it became necessary, e.g., when the number of samples associated with each partition bin exceeds some threshold. The need for a partition refinement as the sample size increases is a theoretical requirement, since the coarse scale partitioning is essentially a histogram, and as the sample size increases, so must the number of bins (Silverman, 1986, Thompson and Tapia, 1990, Redner, 1999, Redner and Gehringer, 1994, Gehringer and Redner, 1992, Gehringer, 1990). In practice, however, the need to refine the partition of a TLGDB is not that critical, as it is also possible to isolate the tree attribute values for each sampled stand if desired.

The HNNTLG procedures provide a framework for nearest neighbor tree list generation methods, including the most similar neighbor (MSN) methods (Moeur and Stage, 1995). If the partition bin widths h_j are small enough, each sampled stand is isolated in its own partition bin, and if $K_{\max} = 1$ and $K_T = 1$, then the HNNTLG procedures and the TLGDB become very similar to the MSN method (Moeur and Stage, 1995), identifying the nearest partition bin and copying the associated tree list. If $K_{\max} = 1$, $K_T > 1$, and the partition bin widths are small enough to isolate each stand, then the most similar stand is selected and a

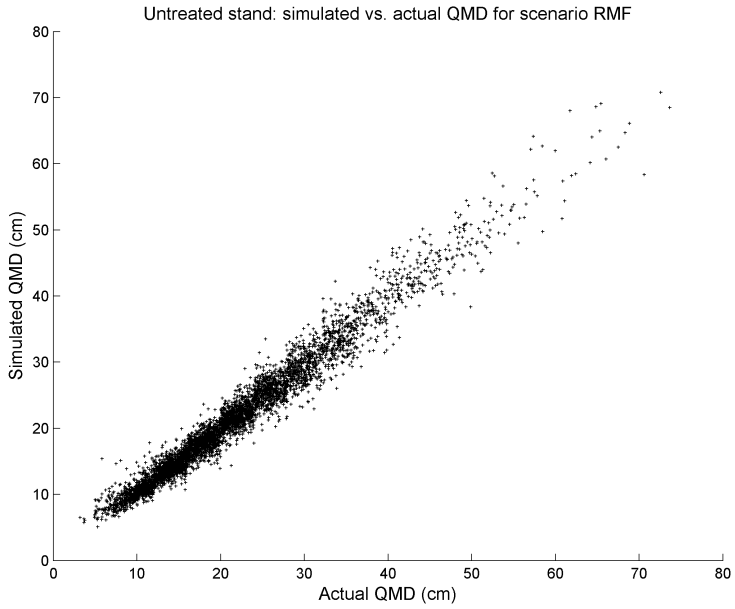

Figure 5: Plot of simulated *vs.* actual QMD values (cm) from untreated stands for scenario RMF. The estimated linear model was $y = 0.7766 + 0.9578x$ with $r^2 = 0.9558$.

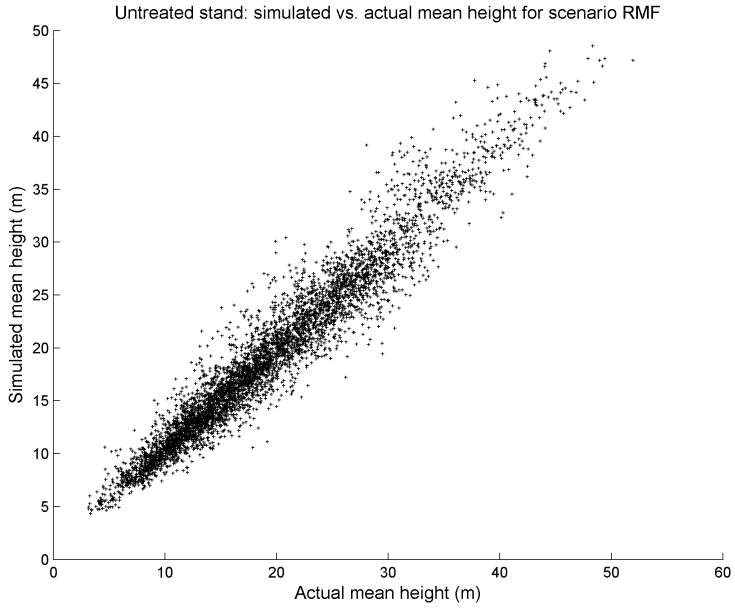

Figure 6: Plot of simulated *vs.* actual mean height values (m) from untreated stands for scenario RMF. The estimated linear model was $y = 1.1419 + 0.9403x$ with $r^2 = 0.8720$.

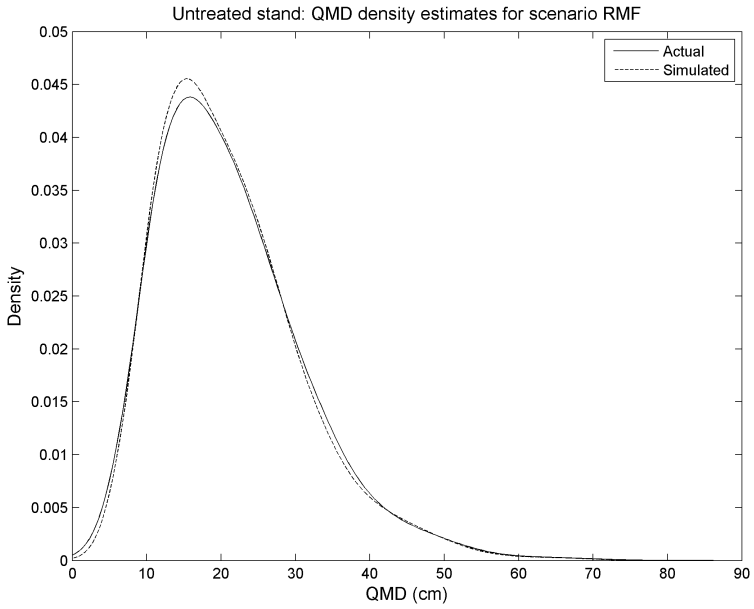

Figure 7: Nonparametric probability density function estimates for the actual and simulated QMD distributions for all untreated stands and scenario RMF. The p_{iae}-value for these distributions is 0.9817.

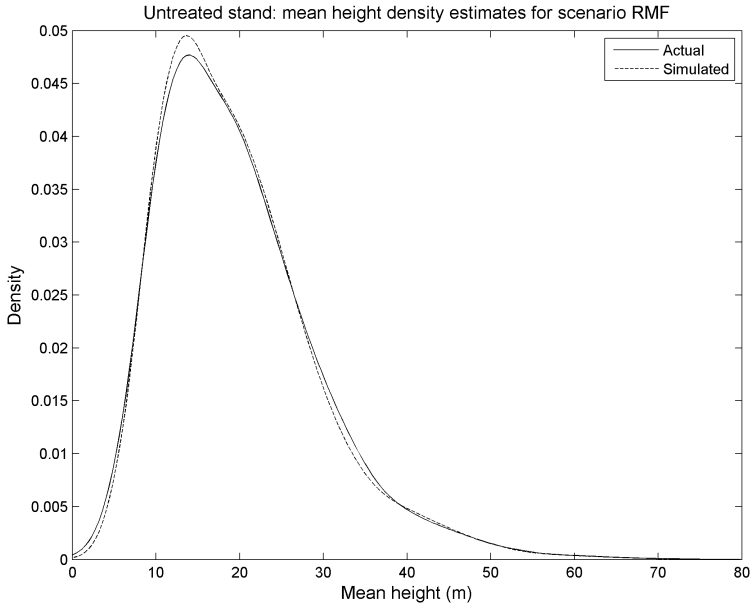

Figure 8: Nonparametric probability density function estimates for the actual and simulated mean height distributions for all untreated stands and scenario RMF. The p_{iae}-value for these distributions is 0.9825.

Table 10: Hypothetical nesting of levels for multiple two-scale relationships

Coarse scale	Mapping	Fine scale
Geographic location	→	Site and stand attributes
Site and stand attributes	→	Tree attributes
Tree attributes	→	Branch attributes

simulated tree list is generated using the associated tree list, a capability not supported by the MSN method, since it does not provide a randomization procedure, but simply copies the tree list from the most similar stand.

In developing the algorithms used to create the coarse scale to fine scale mapping and generate simulated fine scale attributes, it was not necessary to directly approximate the histogram of coarse scale attributes for the partition, that is, it was not necessary to compute estimates of the values P_{B_m}. The objective was to develop data based approximations to the distributions of tree attributes within each partition bin, using the partition to index the space of stand scale attribute values, not to represent the stand scale attribute value distribution. Empirical estimates $\hat{P}_{B_m}$ could have been computed and used as weights when computing similarity scores, allowing the relative probabilities of the partition bins to influence the scores. This would, however, entail updating the weights each time data were added, changing the relative scores and reducing the local nature of the updates, at least until enough data were added to stabilize the estimates of the partition bin probabilities $\hat{P}_{B_m}$. Given the quality of the results that were obtained, estimating the coarse scale partition bin probabilities did not seem warranted, but it is a feature that could be added to the SMC TLGDB software.

The direct partitioning of the coarse scale attributes, combined with the coarse scale to fine scale attribute mapping, also guarantees that impacts from the addition of new data or a partition refinement are local, affecting small regions within the range of the coarse scale attribute space. After the addition of new data or a refinement of a partition, a particular stand description Y^s will map to a partition bin in the stand scale attribute space that was identical to the bin it would have mapped to before the addition of new data, or it will map to a partition bin that was close to the location of the original partition bin if the partition was refined. This implies that simulated fine scale attributes will be generated consistently after new data are added to a TLGDB or after a TLGDB partition is refined. This may not be the case for statistically weighted methods such as the MSN methods (Moeur and Stage, 1995) where the addition of new data requires the determination of a new weighting, which could, subsequently and substantially, change the relative distances among the data points.

While only a two-scale relationship coupling coarse and fine scales was used, the application of the method is not limited to two scales. More scales may be included via a sequence of levels representing nested two scale relationships derived through conditioning arguments similar to those in Section 2.1. The fine scale at a higher (coarser) level being partitioned as the coarse scale at the next lower (finer) level. By nesting a sequence of two-scale relationships in this way, simulated values at one level may be generated to fill in gaps and then used to generate simulated values at a more detailed level. A possible example of using multiple two-scale nestings to represent relationships from geographic location to branch characteristics is given in Table 10.

The HNNTLG procedures and the SMC TLGDB were designed to simulate the process of taking multiple samples from one stand or from multiple stands, and to allow the introduction of variability among the samples from a single stand to approximate the within stand variability that exists. The TLGDB may be

used in conjunction with a forest growth and yield model or forest stand simulator to obtain estimates of the possible variation in the development of an actual stand through repeated simulations. Each simulation would use a different simulated stand, generated to match a particular set of stand attributes, as its input. Variation in the output from the repeated simulations should, then, provide an approximation to the variation expected for an actual stand after a similar time duration.

4.1 Within stand variability

Tree lists generated using the HNNTLG method and the TLGDB have stand scale attributes that are similar to the stand scale attributes that were specified in a statistical sense, within the region of coverage, as indicated by the nearest partition bins to a desired stand description. The stand scale attributes specified may be thought of as parameters defining a conditional distribution of tree scale attributes, but not specifying particular stand scale values for attribute values derived from simulated trees. This interpretation is consistent with the typical use of a random number generator, for example, a random number r drawn from a standard normal distribution $N(0, 1)$. The expectation in this case is that *on average*, the values of r will have a mean value of zero, and a variance equal to one, but no particular set of normally distributed random values is expected to have a mean value that exactly equals zero and a variance exactly equal to one. This interpretation is also consistent with the use of sample measurements from forest stands as indicators of average properties of the stand, knowing that the stand scale attributes vary by location within the stand being sampled, and with the idea that the TLGDB generates simulated samples of tree attributes.

If a specific set of stand or forest patch attributes are to be matched, to within some nominal tolerance, repeated draws from the TLGDB may be necessary to achieve the objective. Nonparametric probability density function estimates for DBH and height distributions of an actual stand and ten simulated stands that were generated to approximate it are presented in Figure 9, and p_{iae}-values for the DBH distribution, height distribution, and species composition of the simulated stands and the actual stand are presented in Table 11. A comparison of the simulated stands to the actual stand indicates that a small set of stand attributes, e.g., QMD, average height, and species composition, could simultaneously be matched, for example to within $\pm 10\%$, with repeated draws. If arbitrarily small tolerances are used, however, it may not be possible to generate a stand with matching attributes, so some care must be taken to choose a reasonable tolerance. The ten simulated stands for this comparison were generated with the default options of `TGRAND`, and they clearly provide reasonable approximations to the DBH and height distributions and the species composition of the actual stand. Nonzero values of the nonparametric density estimates in the figure for negative DBH or height values are an artifact of the probability density estimation technique (Gehringer, 1990, Gehringer and Redner, 1992, Redner and Gehringer, 1994, Redner, 1999). No negative tree diameters or heights were generated.

4.2 GOF testing procedures

The total misclassification rates observed for the GOF comparisons of the simulated and actual untreated stands may be conservative. First, the total misclassification rates were computed by performing a logical *or*ing using the individual misclassification rates for each of the three tree scale attributes. This method of computing the total error rates must inflate their values when compared to computing an error rate in the three dimensions DBH, height, and species simultaneously, as the individual tests are more restrictive, that is, more likely identify stands as different given the reduction in variability caused by considering each dimension independently. The impact of this reduction in variability combined with the logical *or*ing is

Table 11: DBH height, and species p_{iae}-values for 10 simulations of a particular stand. The DBH and height density function estimates are given in Figure 9.

Simulation	p_{iae}-value			Simulation	p_{iae}-value		
number	DBH	Height	Species	number	DBH	Height	Species
1	0.9302	0.8745	0.9592	6	0.9406	0.9233	0.9592
2	0.8985	0.9038	0.8980	7	0.9705	0.9441	0.8571
3	0.9539	0.9293	0.9592	8	0.9142	0.8972	0.9184
4	0.9531	0.8704	0.9592	9	0.8762	0.9100	0.9592
5	0.8733	0.8619	0.9592	10	0.9236	0.9132	0.9592

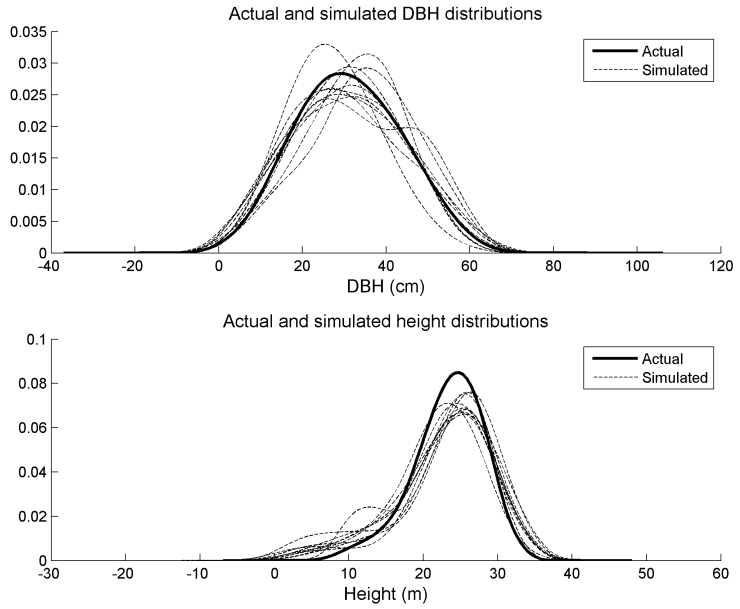

Figure 9: DBH and height density function estimates for 10 simulations of a particular stand and the DBH and height density function estimates for the actual stand. The p_{iae}-values for DBH, height, and species composition for each simulation are given in Table 11.

difficult to assess without performing a multivariate GOF test of some sort. Complicating the use of a multivariate GOF test is the fact that some attributes were continuously distributed, tree DBH and height, while the other, species, was discrete. Some sort of multivariate rank test may be useful here, ordering by the discrete attributes first, and this should be investigated further.

Second, for all of the testing scenarios, the misclassification rates for tree height distributions were greater than those for DBH distributions or species composition. An explanation of this phenomenon seems readily available: many of the tree heights were estimated from height-diameter relationships. The estimation of tree heights from a height-diameter relationships dramatically reduces the variation of tree heights relative the actual height variation within a stand at the time of its measurement. This reduction in height variability would, then, cause a higher misclassification rate for height than would otherwise be expected.

Finally, the critical p_{iae}-value of 0.65 may be too large. The standard normal distribution used in the simulation to determine a critical p_{iae}-value is well behaved, being strongly unimodal and symmetric. The DBH and height distributions for forest stands are frequently not so well behaved, being skewed or even multimodal, either of which would reduce the value of the critical p_{iae}-value. An experiment was performed to assess the magnitude of the conservative effect on the misclassification rate caused by too large a critical p_{iae}-value. Douglas-fir diameter data from the SMC were used to determine how small a critical p_{iae}-value may be for pure Douglas-fir stands, using individual tree DBH measurements from two to four permanent study plots on each of 30 pure Douglas-fir installations. The plots chosen from each installation had one to three commensurate measurement dates and provided 352 within stand comparisons which were used to compute p_{iae}-values. The SMC installations were all on commercial forest land, were naturally regenerated or planted juvenile stands, and had at least 90% Douglas-fir measured in stems per unit area. The 50 year site index for the installations ranged from 25.9 m to 44.2 m, with stand densities from approximately 250 TPH to 1850 TPH, and breast height ages ranging from three years to 21 years. A minimum p_{iae}-value of approximately 0.55 was obtained from this experiment, and if used it would have produced total misclassification rates that were less than 2%. Given this investigation, the critical p_{iae}-value of 0.65 seems reasonable, while being somewhat conservative.

4.3 Future work and extensions

The TLGDB currently supports only pure Douglas-fir stands, pure western hemlock stands, and stands containing one of these two species as a dominant component within mixed species stands. Extending the TLGDB to allow arbitrary dominant tree species, and possibly even understory vegetation, would be 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 addition of these features could prove to be useful when developing forest management strategies by enabling consideration of the impacts of within stand variability and a more complete description of the vegetation contained within a forest.

In addition to extending the species representation capabilities of the TLGDB, the number of treatment types and treatment combinations that may be represented should be increased. Currently only untreated and thinned stands, with support for multiple thinning events, may be represented in a TLGDB. Stands having fertilization, pruning, or combinations of thinning, pruning, and fertilization should also be supported.

When generating simulated tree scale attributes it may be desirable to restrict the continuous attributes by the values of one or more discrete attributes during the tree generation procedure described in Section 2.6. This would provide another level of conditioning when generating the tree scale attributes, for example, 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, take into account differences in height–diameter relationships among tree species. A straightforward version of this filtering was done to obtain the 100% pure Douglas-fir or western hemlock stands in the preprocessing step.

Finally, the HNNTLG procedures need to be extended to support multiple, nested two-scale relationships and tested within this context. A mapping from geographic location to stand scale attributes, and then from stand scale attributes to tree attributes, should be sufficient to test the general, multi-scale applicability of the procedures.

5 Conclusions

Procedures for using an implicit two-scale relationship with a nearest neighbors algorithm to generate simulated trees representative of forest conditions associated with a specified 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 at the stand scale, their use of actual tree measurement data to generate simulated trees to guarantee realistic, biologically achievable simulated tree attributes, and their ability to simulate within stand variability. The procedures are appropriate for simulating forest stands for use with growth and yield or forest simulation models or to provide estimates of forest stand characteristics and their potential variability across large, sparsely sampled regions, which commonly occur in forestry.

6 Acknowledgements

This work was funded by the Stand Management Cooperative (SMC) in the College of Forest Resources at the University of Washington, Seattle, WA. I would like to thank the SMC for their support of this work, and in particular, I want to thank Eric Turnblom. In addition, I would like to thank Jim Flewelling and Temesgen Hailemariam for their help reviewing the tree list generation database documentation.

References

P.J. Bickel and K.A. Doksum. *Mathematical Statistics*. Holden-Day, 1978.

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

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

David Bruce. Consistent height-growth and growth-rate estimates for remeasured plots. *Forest Science*, 27 (4):711–725, 1981.

William G. Cochran. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23(3):315–345, 1952.

A.C. Davison and D.V. Hinkley. *Bootstrap methods and their application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, reprint with corrections edition, 2003. Originally published 1997.

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

D.M. Donnelly. *Pacific Northwest coast variant of the forest vegetation simulator*. Addison-Wesley, 1997. Available on the Web.

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

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

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

Kevin R. Gehringer. *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, August 2001.

Kevin R. Gehringer and Richard A. Redner. Nonparametric probability density estimation using normalized B-splines. *Comm. Statist. Simulation Comput.*, 21(3):849–878, 1992.

Kevin R. Gehringer and Eric C. Turnblom. *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, August 2001.

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

D. A. Maguire, W. S. Bennett, J. A. Kershaw Jr., R. Gonyea, and H. N. Chappell. Establishment report stand management cooperative silviculture project field installations. Technical report, Stand Management Cooperative, College of Forest Resources, University of Washington, 1991.

Kenneth J. Mitchell. *Dynamics and Simulated Yield of Douglas-fir*. Monograph 17. Forest Science, 1975.

Melinda Moeur and Albert R. Stage. Most similar neighbor: An improved sampling inference procedure for natural resource planning. *Forest Science*, 41(2):337–359, 1995.

Gordon D. Nigh. The geometric mean regression line: A method for developing site index conversion equations for species in mixed stands. *Forest Science*, 41(1):84–98, 1995.

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

Richard A. Redner and Kevin Gehringer. Function estimation using partitions of unity. *Comm. Statist. Theory Methods*, 23(7):2059–2078, 1994.

Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

Marion R. Reynolds, Jr., Thomas E. Burk, and Won-Chin Huang. Goodness-of-fit tests and model selection procedures for diameter distribution models. *Forest Science*, 34(2):373–399, 1988.

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

The Stand Management Cooperative. Stand management cooperative: An integrated program of research in silviculture, forest nutrition, wood quality, and modeling. Annual report, University of Washington, College of Forest Resources, Room 164, Bloedel Hall, Box 352100, Seattle, WA, 98195-2100, 1992.

Malcolm S. Taylor and James R. Thompson. A data based algorithm for the generation of random vectors. *Computational Statistics & Data Analysis*, 4:93–101, 1986.

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

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

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

W. R. Wykoff. *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, 1986.

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