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

Kevin R. Gehringer
Biometrics Northwest LLC, Redmond WA
www.biometricsnw.com

Meeting in the Middle
2006 Nearest Neighbors Workshop August 28-30
University of Minnesota Minneapolis, MN

General problem

- Generate simulated attributes at a fine scale conditioning on attributes at a coarse scale
  - The two scales are assumed to be nested and are implicitly defined
- Objective
  - Generate simulated fine scale attributes that are representative of attributes that could be obtained from a sample collected for a specified set of coarse scale attributes

Specific problem

- Generate simulated tree attributes that are representative of the attributes associated with a specified set of stand attributes
  - The coarse scale is the forest stand
    - Site quality, age, stand density, average tree size
  - The fine scale is the tree
    - DBH, height, species, etc.
- Objective
  - Generate realistic simulated samples/tree lists

Context

- Growth and yield modeling inputs
  - Tree list based models provide greater detail
- Fill in/impute missing tree scale data
  - Forestry data are sparsely sampled
- Historic data were usually summarized
  - Expand into individual trees to improve detail
- Simulation
  - Want to obtain estimates of potential variability to support decision making and management

Requirements

- Simulated tree attributes should be generated as multidimensional objects
- Simulated trees should be physically realizable
- Representations for the stand and tree distributions should be flexible
- The addition of new data should be easy to perform

Requirements (cont.)

- The addition of new data should not affect the consistency of the simulated trees and stands/tree lists
- The framework used should be similar to that of pseudorandom number generation
  - Think of the solution as providing a *random sample generator* at the tree (fine) scale, conditioned by a set of specified stand (coarse) scale attributes

Approach

- Assume that the joint distribution of stand scale and tree scale attributes may be represented by a finite mixture distribution

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

Approach (cont.)

N is the number of component density functions

x^s is a d^s - dimensional vector of discrete and continuous stand scale attributes

x^t is a d^t - dimensional vector of discrete and continuous tree scale attributes

$f_i(x^s, x^t)$ is the component density for stand i

α_i is the mixing coefficient for stand i, and $\sum_{i=1}^{N} \alpha_i = 1$

Approach (cont.)

- The mixture density formulation allows us to compute a number of useful things for the mixture and each component density, e.g.,

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

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

$$S_i = \left\{(x^s, x^t) \mid f_i(x^s, x^t) > 0\right\} \text{and } S = \bigcup_{i=1}^{N} S_i$$

Approach (cont.)

- But

$$f_i^{\text{tree}}(x^t)$$

  is exactly the distribution needed to generate simulated tree attributes for stand i
- Find an index at the stand scale that allows an appropriate stand marginal to be identified, then use the associated tree scale marginal distribution to generate simulated trees

Approach (cont.)

- Index the component density functions by choosing a stand scale index point from their support sets, restricted to the stand scale attributes
- For example,

choose $x_i^{\text{index}} \in S_i \mid_{x^s}$

as the mode of $f_i^{\text{stand}}(x^s)$

- Use NN to find the nearest component density to a specified stand scale attribute vector
- Simulate tree attributes from $f_k^{\text{tree}}(x^t)$

Challenges

- The mixture density and its component density functions are unknown
- Indexing stand scale attributes in the support sets almost works, except for
  - Overlapping component density functions and stand scale marginal distributions
  - The likelihood of multiple similar index points or stand scale marginal distributions
  - The likelihood of index point ties using NN

Resolution

- Use a direct partitioning of the stand scale attribute space to condition the fine scale attributes and generate a stand scale index

$$B = \{B_1, B_2, \ldots, B_M\}$$

$$B_i \bigcap B_j = \varnothing, \quad i \neq j$$

$$S|_{x^s} = \bigcup_{i=1}^{N} S_i|_{x^s} \subset \bigcup_{m=1}^{M} B_m$$

Revised approach

- Using the partition sets, compute

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

$$P_{B_m} = \Pr\left(\left\{(x^s, x^t) \mid x^s \in B_m\right\}\right)$$

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

Revised approach (cont.)

- The conditional marginal distributions of tree attributes for each partition are then

$$\hat{f}_{B_m}^{\text{tree}}(x^t) = \int \hat{f}_{B_m}(x^s, x^t)dx^s$$

- They may be indexed by choosing an index point from of the stand scale partition sets

$$b_m \in B_m$$

- NN search may then be used to find a partition and the associated tree attribute distribution

Assumption 1

- Shifting to a partition simplified the problem, but at a cost
  - The connection between the stand and tree scale attributes was lost
- Assumption 1: Forest stands having similar stand (coarse) scale attributes have similar tree (fine) scale attribute distributions

Discrete valued attributes

- Discrete valued attributes do not pose any significant difficulties
  - They act as conditioning events for the continuous attributes and distributions at the stand or tree scale
  - They, therefore, provide a refinement of the partition or mixture density, adding more partition sets or terms to the mixture density sum

Implementation

- A partition for the stand scale attributes
- A stand scale similarity score for finding NN partition bins
- A mapping from partition bins to tree lists that have been associated with the bins
- A NN tree generation procedure

Implementation (cont.)

- The partition used multidimensional bins with widths

$$h_j = 0 \quad \text{if } x_j^s \text{ was discrete}$$

$$h_j > 0 \quad \text{if } x_j^s \text{ was continuous}$$

- The index points were defined by the bin centers

$$b_{mj} = \begin{cases} x_j^s & \text{if } x_j^s \text{ was discrete} \\ h_j \left(\left\lfloor \dfrac{x_j^s}{h_j} \right\rfloor + \dfrac{1}{2} \right) & \text{if } x_j^s \text{ was continuous} \end{cases}$$

Implementation (cont.)

- The partition bins were then

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

- Stand scale similarity scores were computed as

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

Implementation (cont.)

- Weights were

$$w_j = \begin{cases} 10000 & \text{if } x_j^s \text{ was discrete} \\ \dfrac{1}{h_j} & \text{if } x_j^s \text{ was continuous} \end{cases}$$

- This gives a bin width weighted Euclidean distance for the continuous attributes and imposes a large penalty on discrete attribute discrepancies

Implementation (cont.)

- A partition bin to tree data mapping was produced
  - Tree data from sampled stands were mapped to the partition containing their stand scale attributes
- Simulated trees were then generated in three steps

Implementation (cont.)

- Find the nearest partition bins to a specified stand scale attribute vector
- Create a canonical tree list by merging tree attributes from the selected bins
- Generate simulated tree attributes by randomly selecting a tree from the canonical tree list
  - Assign discrete attributes to the simulated tree
  - Use the SIMDAT algorithm (Taylor and Thompson, 1986) to generate a vector of continuous attributes

Assumption 2

- Attributes at the tree (fine) scale obtained from sampled forest stands may be extended to the stand (coarse) scale
  - Tree data from a sample may be used to populate the entire stand area

Validation

- A suite of programs implementing the described procedures was developed and tested for the Stand Management Cooperative (SMC) at the University of Washington
  - The Tree list generation database (TLGDB)
    http://depts.washington.edu/silvproj/tlghome/index.htm

- The TLGDB was designed to generate tree lists for Douglas-fir (DF) and western hemlock (WH) forests in the PNW for untreated and thinned stands, West of the Cascade Mountains
  - Only these species and mixtures of them are supported at this time

Validation (cont.)

- Test data were from a variety of sources
  - BC Ministry of Forests
  - Canadian Forest Service
  - Oregon State Univ.
  - Port Blakely Tree Farm
  - RFNRP (SMC)
  - SMC
  - USFS PNW Research Station
  - WA DNR
  - Weyerhaeuser Company
- A total of 5209 samples were used to populate the TLGDB
  - 65% were Pure DF
  - 15.4% were Pure WH
  - 13.3% were DF dominant
  - 3.4% were WH dominant
  - 2.6% were mixture
  - 74.1% were planted
  - 25.9% were natural
  - 573,036 tree records

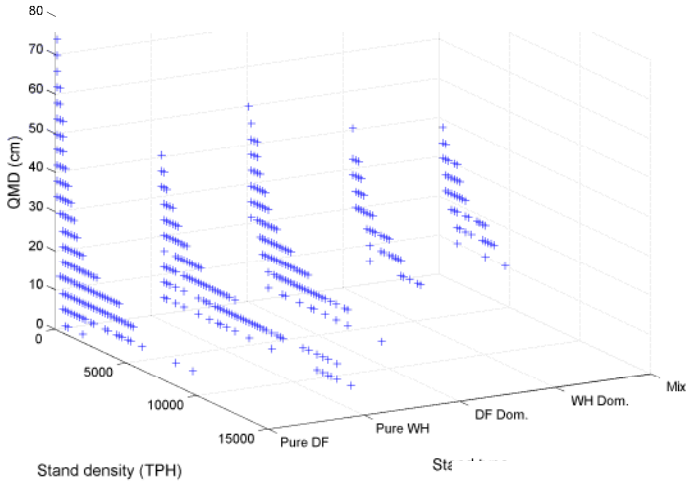
Validation (cont.)

- Stand scale attributes
  - QMD (cm): bin width = 4.0
  - 50 year site index (m): bin width = 3.0
  - Stand density (TPH): bin width =200
  - Stand total age (years): bin width = 4.0
  - Stand origin: Natural or planted
  - Stand type: Pure DF, Pure WH, DF Dominant, WH Dominant, or Mixture
- Tree scale attributes
  - DBH (cm)
  - Height (m)
  - Tree species
    - DF, WH, and other typical PNW species

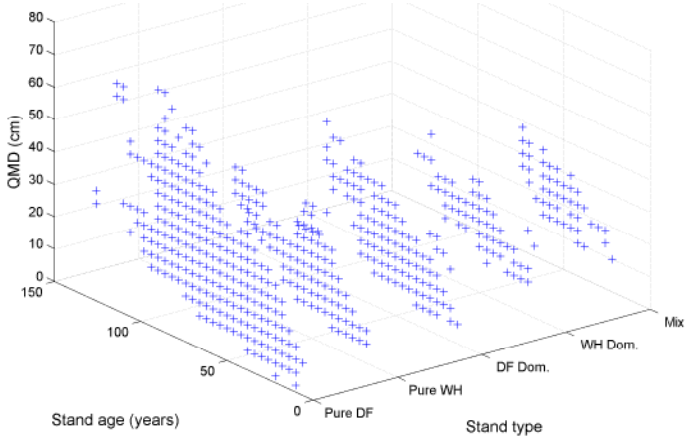
- Pure stands had > 70% BA
- Dominant stands had > 50% BA
- Mixture was anything else

Data Coverage

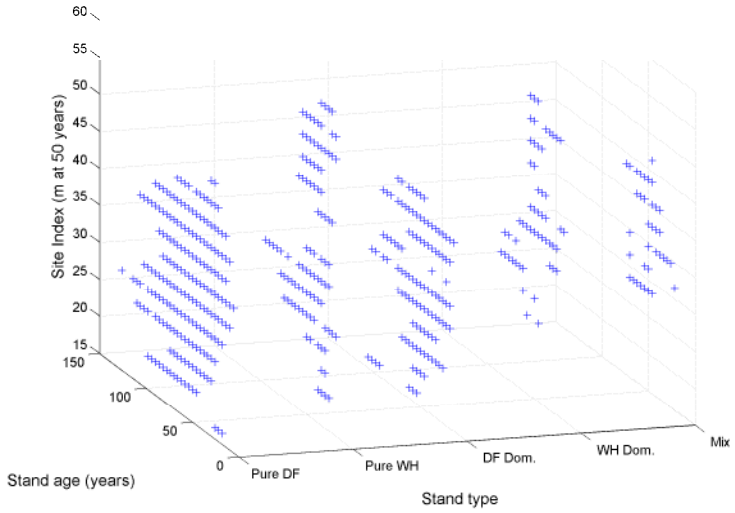
QMD vs. Stand density vs. Stand type



QMD vs. Stand age vs. Stand type



Site Index vs. Stand age vs. Stand type

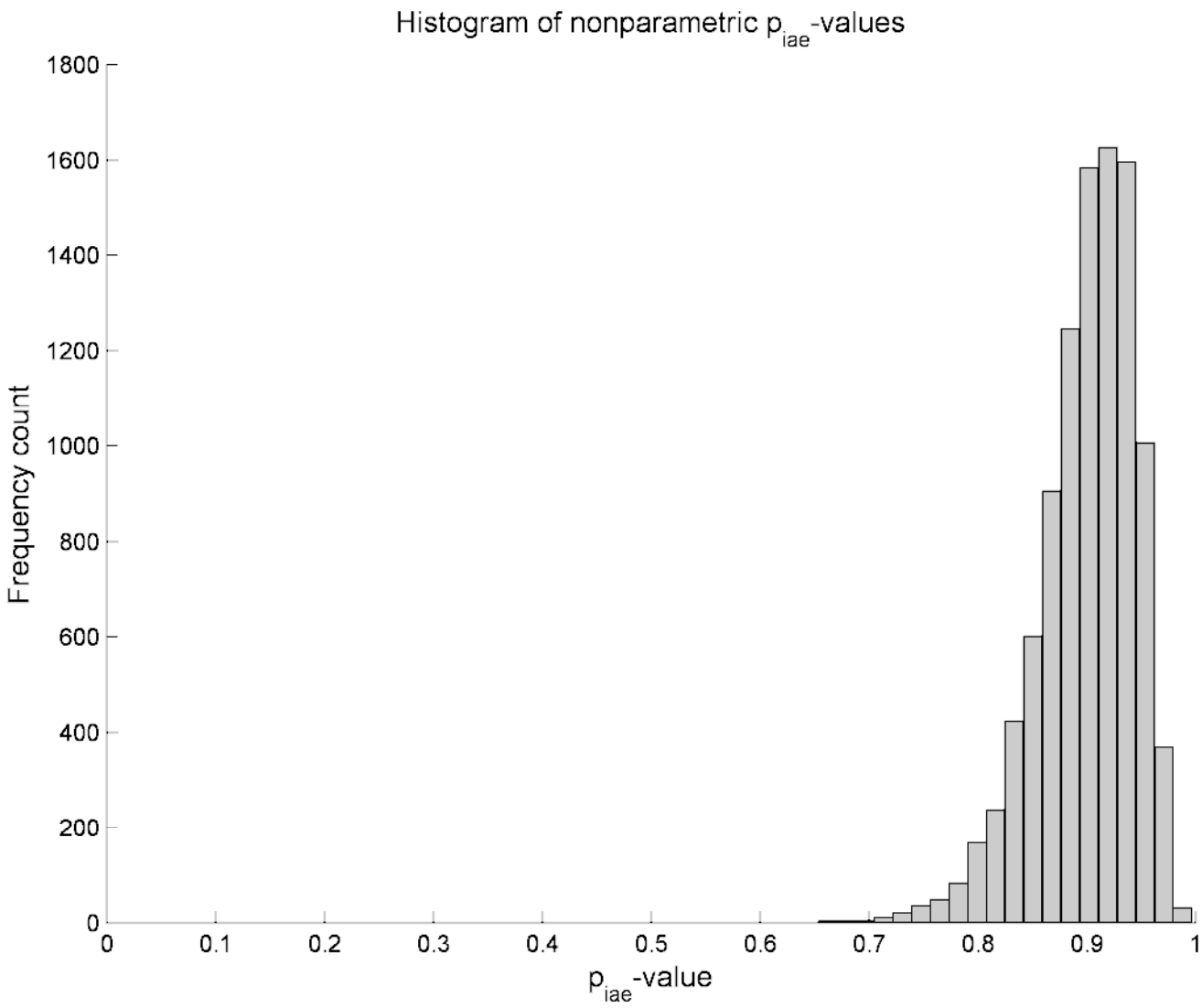


Validation (cont.)

- Goodness of fit (GOF) assessment
  - Compare actual samples to simulated samples
  - Statistic: the p_{iae}-value

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

  - This statistic measures similarity
    - A value of 1 implies indistinguishable distributions
- A cutoff value of 0.65 was chosen for rejection
  - Selected via simulation using $N(0,1)$ distribution


Histogram of nonparametric p_{iae}-values
Frequency count
p_{iae}-value
0
200
400
600
800
1000
1200
1400
1600
1800
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Validation (cont.)

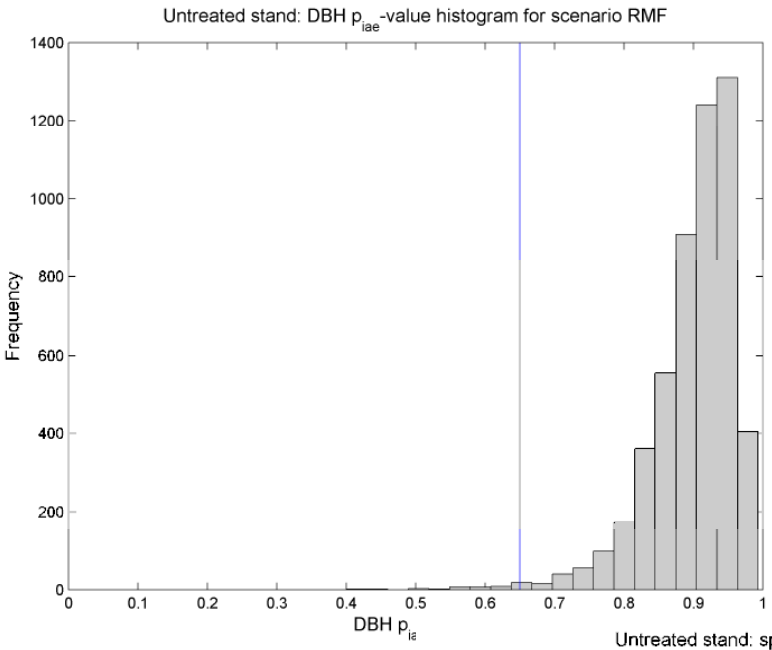
- Four assessment scenarios
  - RMF: No options (Default)
  - RMFJ: Jitter heights option
  - RMFP: Force simulated stands to be 100% pure if the original stand was 100% pure
  - RMFJP: Jitter Heights and 100% pure option
- Only results for scenario RMF for untreated stands are presented
- Other results were similar

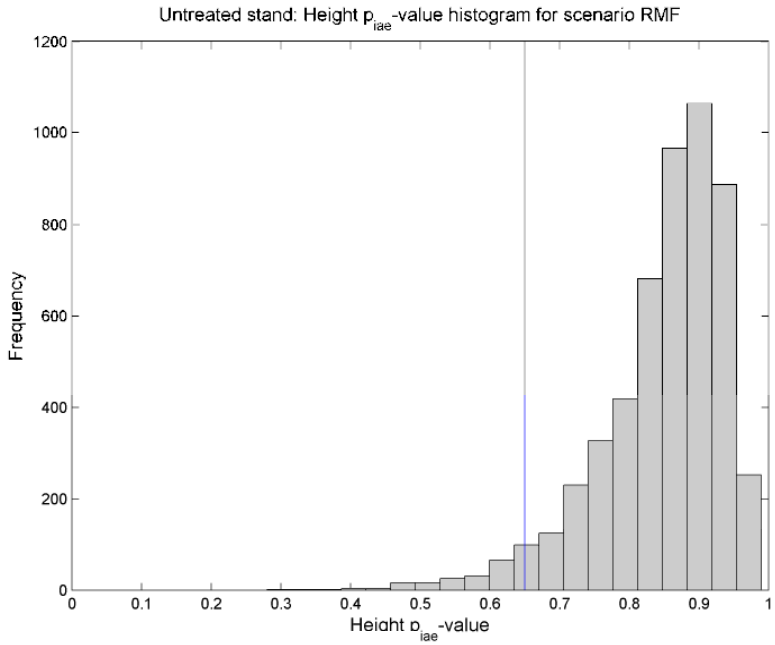
- Assessment criteria
  - Tree scale
    - Empirical rejection rates obtained from p_{iae} values for DBH, height, and species
  - Stand scale
    - Linear fits of actual and simulated QMD and average height values
    - r^2 values for actual and simulated QMD and average height
    - p_{iae}-values and estimated bias and RMSE for QMD and average height

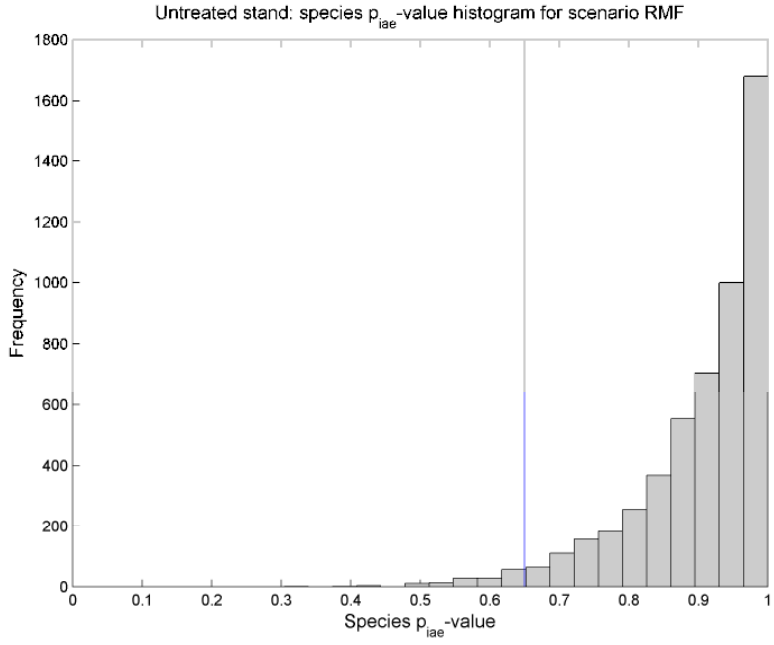
Tree scale results

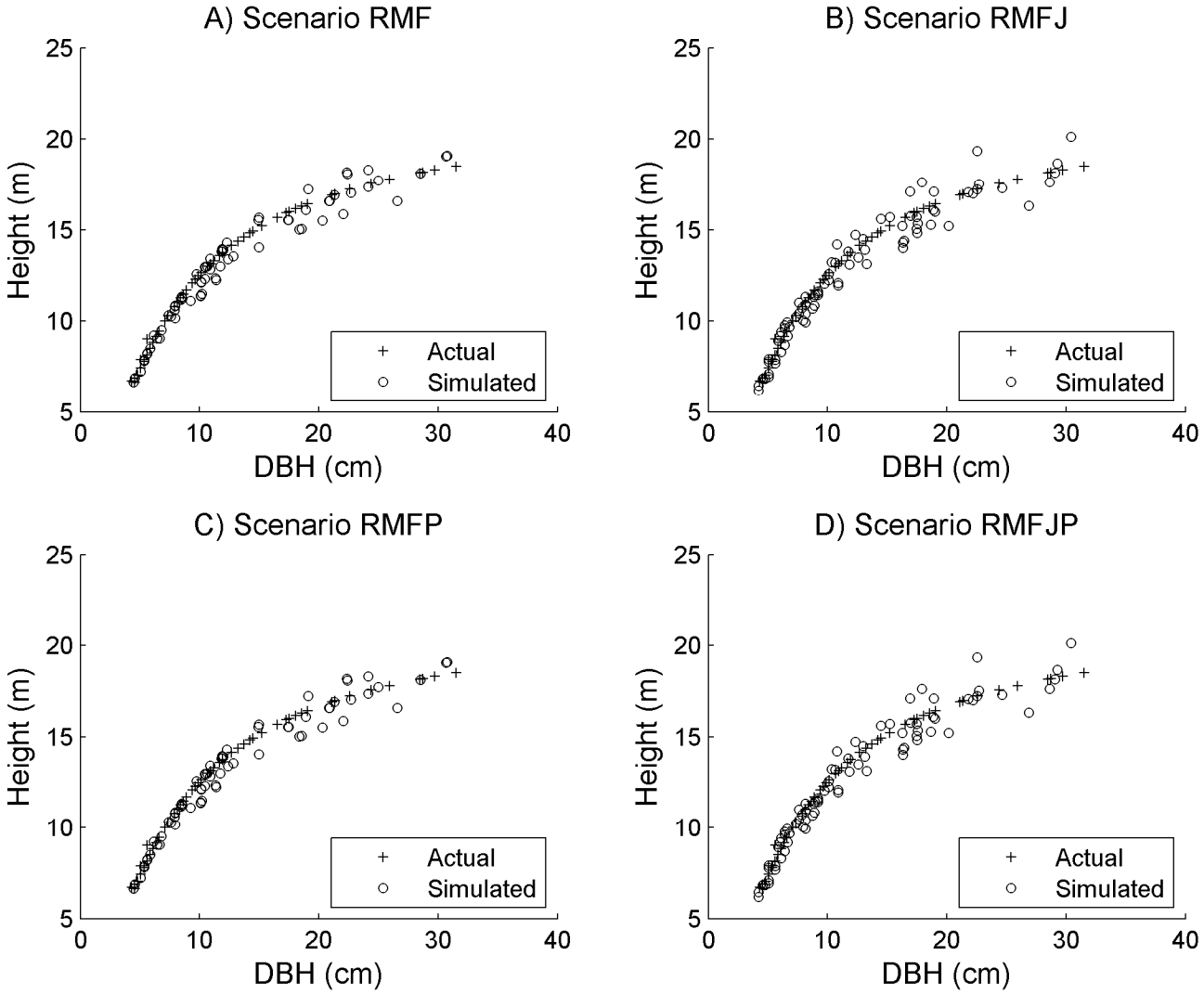
Empirical error rates (RMF)	
DBH	0.6%
Height	3.7%
Species	2.6%
Total	6.1%

- Error rate is percent of simulated stands having different distributions than their actual stands ($p_{\text{iae}} \leq 0.65$)
- Total computed as logical *or*ing of other three


Untreated stand: DBH p_{iae}-value histogram for scenario RMF
Frequency
DBH p_{ia}


Untreated stand: Height p_{iae}-value histogram for scenario RMF
Frequency
Height p_{iae}-value

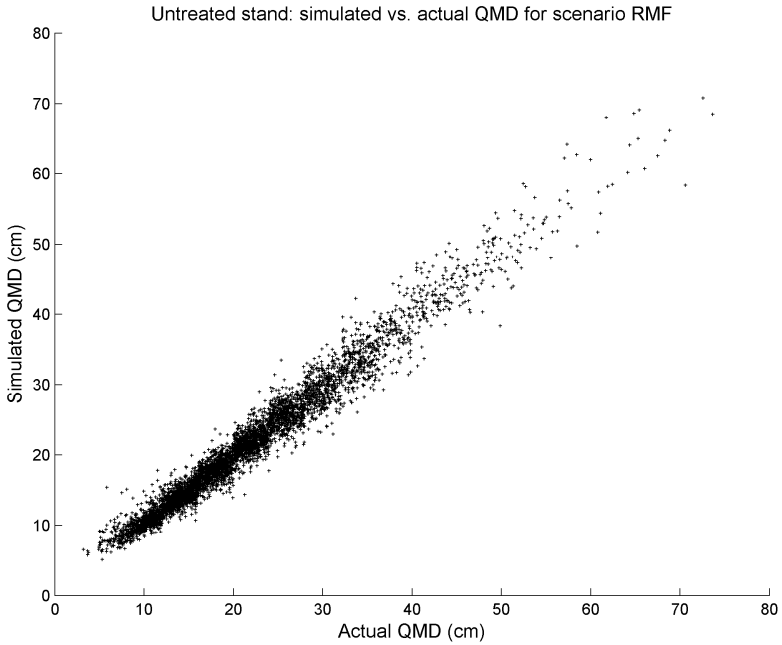

Untreated stand: species p_{iae}-value histogram for scenario RMF
Frequency
Species p_{iae}-value

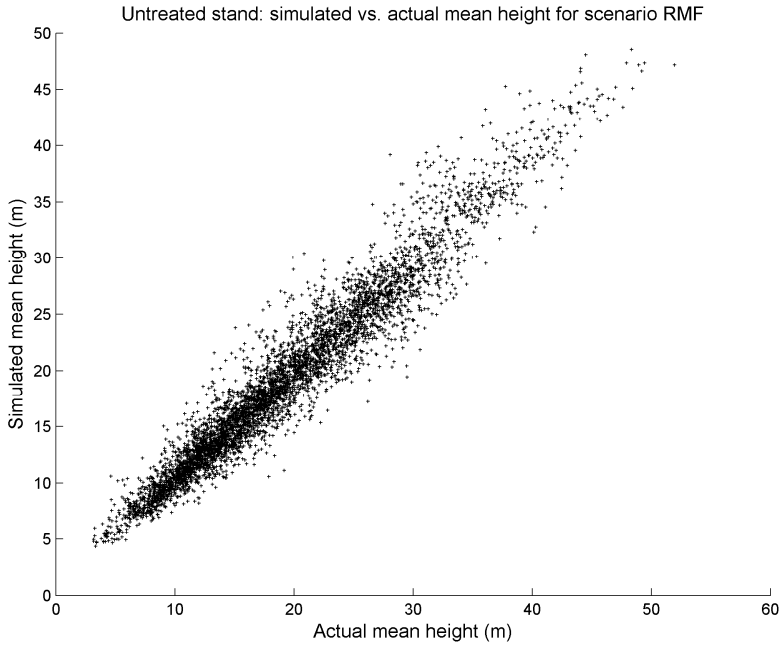

A) Scenario RMF
B) Scenario RMFJ
C) Scenario RMFP
D) Scenario RMFJP
Height (m)
DBH (cm)
Actual
Simulated
5
10
15
20
25
0
10
20
30
40

Stand scale results

Linear model fits and r^2 values (RMF)		
	QMD (cm)	Average height (m)
Intercept	0.777	1.142
Slope	0.955	0.940
r^2	0.956	0.872

Stand scale results (cont.)

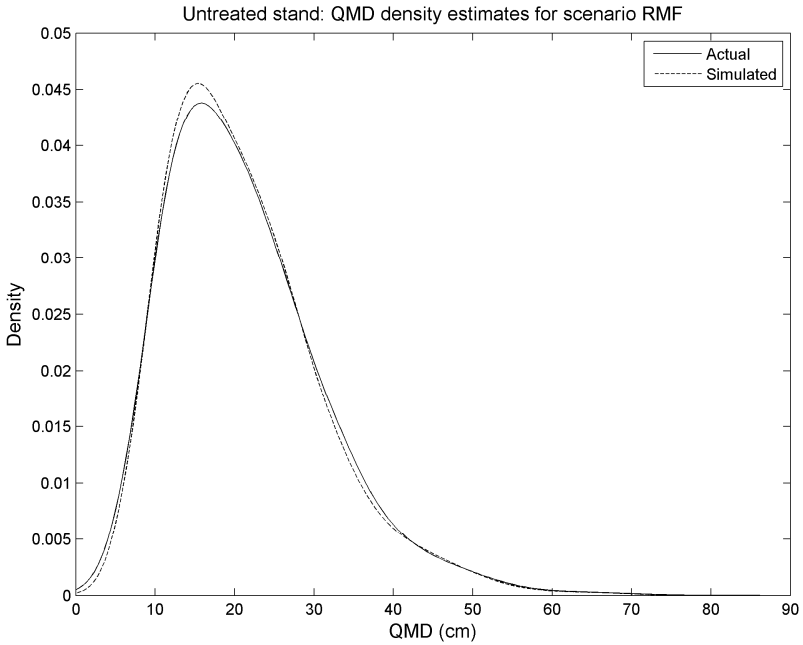


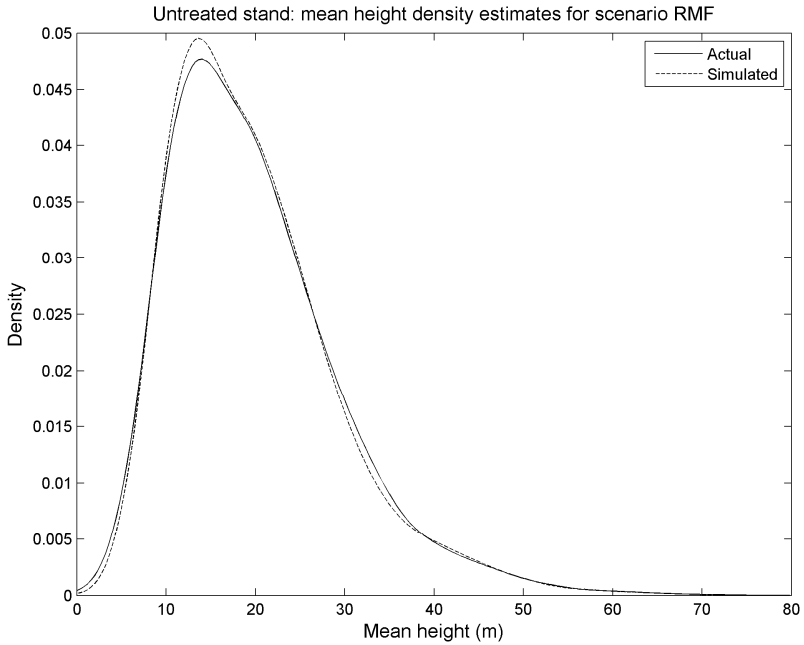


Stand scale results (cont.)

p_{iae}-values, bias, and RMSE (RMF)		
	QMD (cm)	Average height (m)
p_{iae}-value	0.982	0.9825
Bias	0.124	0.003
RMSE	1.807	2.073

Stand scale results (cont.)





Benefits

- The hierarchical NN method works well, is straightforward to implement, and was derived within a consistent mathematical framework
- The direct partitioning at the stand (coarse) scale removes the need for statistically derived relationships, e.g., CCA, between the stand (coarse) and tree (fine) scales
  - The partitioning at the stand scale essentially produces a multidimensional histogram conditioned by discrete attributes

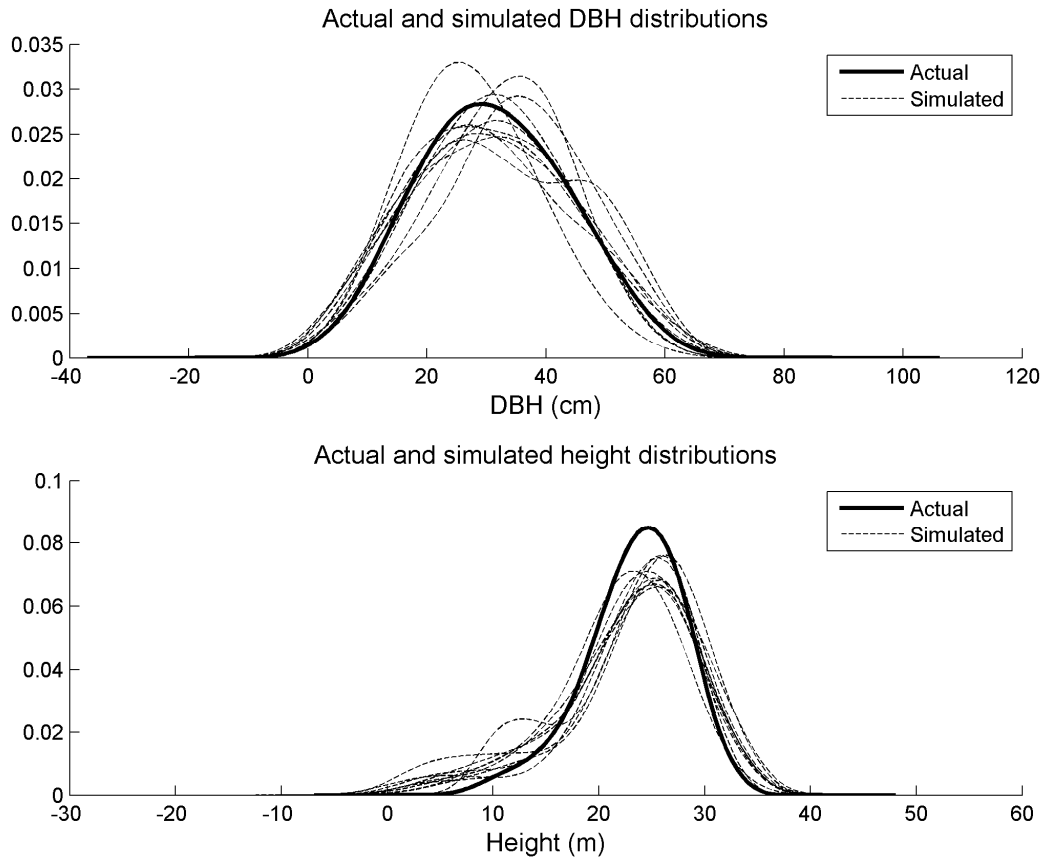
Benefits (cont.)

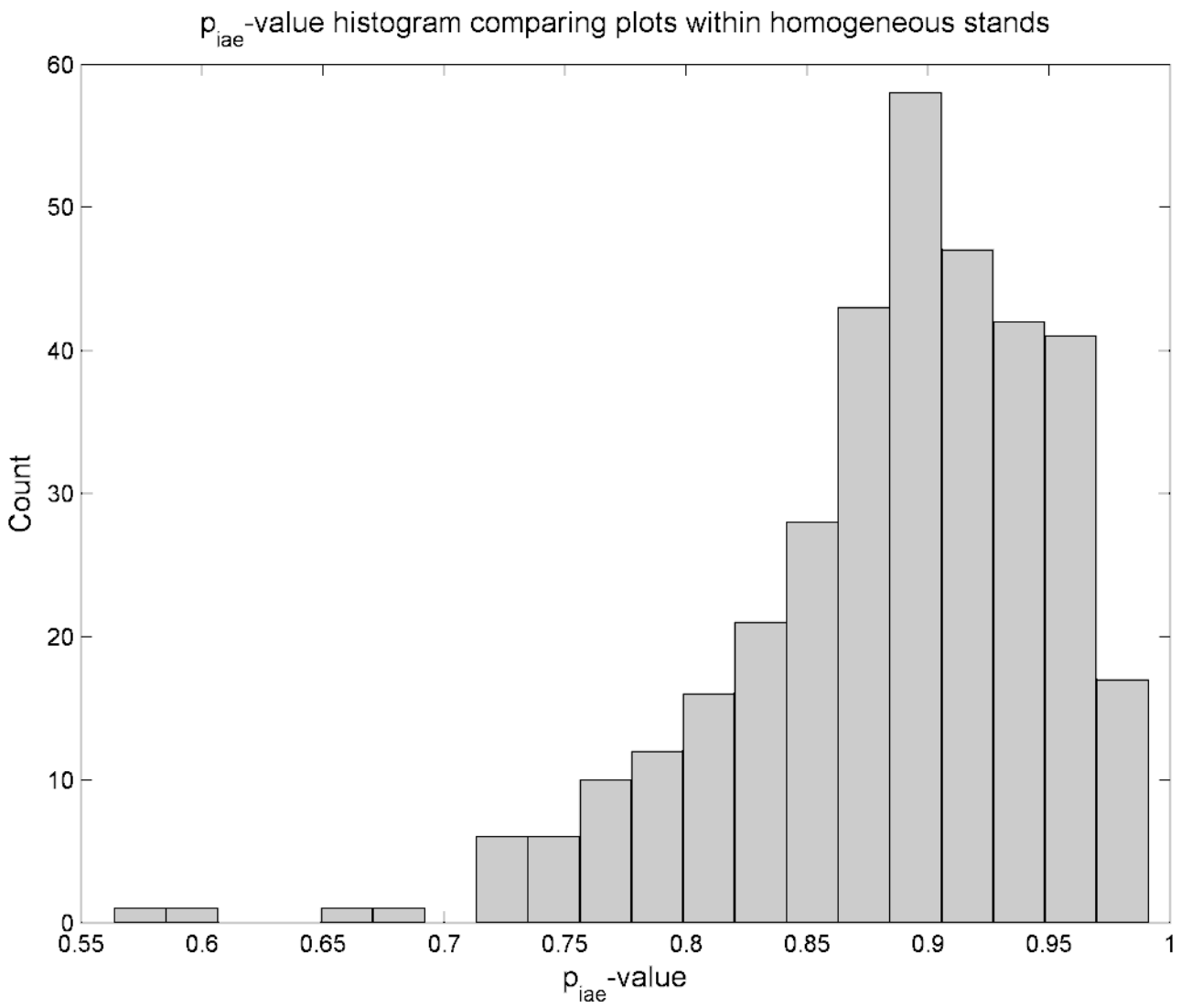
- The approach allows for the generation of within stand variability
  - Unlike methods that copy sampled tree lists
- More than two scales may be used via a nested sequence of two or more two-scale relationships, e.g.,
  - Geographic location to stand scale attributes, and stand scale attributes to tree scale attributes (as was done here)

Benefits (cont.)

- The partition is built dynamically
  - Minimizes storage space
  - New data can be added easily
- The direct partitioning guarantees local impacts when adding data or generating tree lists
  - A new bin is created or an existing bin is updated

Within stand variability example




p_{iae}-value histogram comparing plots within homogeneous stands
Count
p_{iae}-value
0
10
20
30
40
50
60
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1