

# Scale and spatial structure effects on tree size distributions: implications for growth and yield modelling

Oscar García

**Abstract:** Diameter and other size distributions are extensively used in growth modelling. These are usually obtained from sample plot data and assumed to apply both at the stand level, of interest for production planning, and at the forest patch level, the level relevant for tree growth interactions. However, spatial correlation can cause distribution parameters and their estimates to vary with the extent of ground considered. Using mapped tree data from four forest stands in central Canada, it is shown that differences in DBH variance with plot size can be substantial. In addition, size correlations between neighbouring trees were positive, rather than negative as implied by current distance-dependent growth models. Biases in mean DBH are also found. It is proved that plot totals and frequencies are unbiased estimates of stand parameters, but variances and some other statistics are not. The expected variance is expressed in terms of plot size and shape and of second-order stand spatial structure properties. Some possible approaches for reducing bias in stand-level variance estimates are discussed, and the desirability of modelling microsite or genetic spatial correlations in individual-tree simulators is pointed out.

**Résumé :** Les distributions de diamètres et d'autres dimensions sont très fréquemment employées pour modéliser la croissance. Ces distributions sont habituellement obtenues à partir de données provenant de placettes-échantillons et on assume qu'elles s'appliquent à la fois à l'échelle du peuplement, qui est utile pour la planification de la production, et à l'échelle d'îlots d'arbres, qui est appropriée pour étudier les interactions de la croissance entre les arbres. Toutefois, à cause de la corrélation spatiale, les paramètres de la distribution et la valeur estimée de ces paramètres peuvent varier en fonction de la superficie du terrain qui est considéré. À l'aide de données provenant d'arbres cartographiés dans quatre peuplements situés dans le centre du Canada, il est démontré que les différences dans la variance du DHP peuvent être substantielles selon la taille des placettes. De plus, les corrélations de taille entre les arbres voisins sont positives et non négatives comme l'assument les modèles actuels de croissance qui sont dépendants des distances. Des biais ont aussi été trouvés dans le cas du DHP moyen. Il est démontré que les totaux et les fréquences dans les placettes sont des estimations non biaisées des paramètres de peuplement mais que les variances et d'autres paramètres statistiques le sont. La variance attendue est exprimée en termes de taille et de forme de placette, ainsi que de propriétés de second ordre quant à la structure du peuplement. Certaines approches permettant de réduire les biais dans l'estimation des variances à l'échelle du peuplement sont abordées et l'opportunité de modéliser la corrélation spatiale d'ordre génétique ou à l'échelle du microsite dans les simulateurs d'arbres est soulignée.

[Traduit par la Rédaction]

## 1. Introduction

Distributions of tree sizes, especially of diameter at breast height (DBH), have been extensively studied and used in forestry at least since the time of de Liocourt (1898). A search for diameter or DBH distribution(s) in the TREECD database (now the Forest Science Database; <http://www.cabi-publishing.org/AbstractDatabases.asp?SubjectArea=&PID=114>) for February 2000 to October 2004 returns 938 literature references. Distributions are estimated on sample plots and are used in predicting tree growth, as well as to estimate the yield in forest products of various types and sizes. Only a few studies, however, have questioned the applicability of plot-level estimates

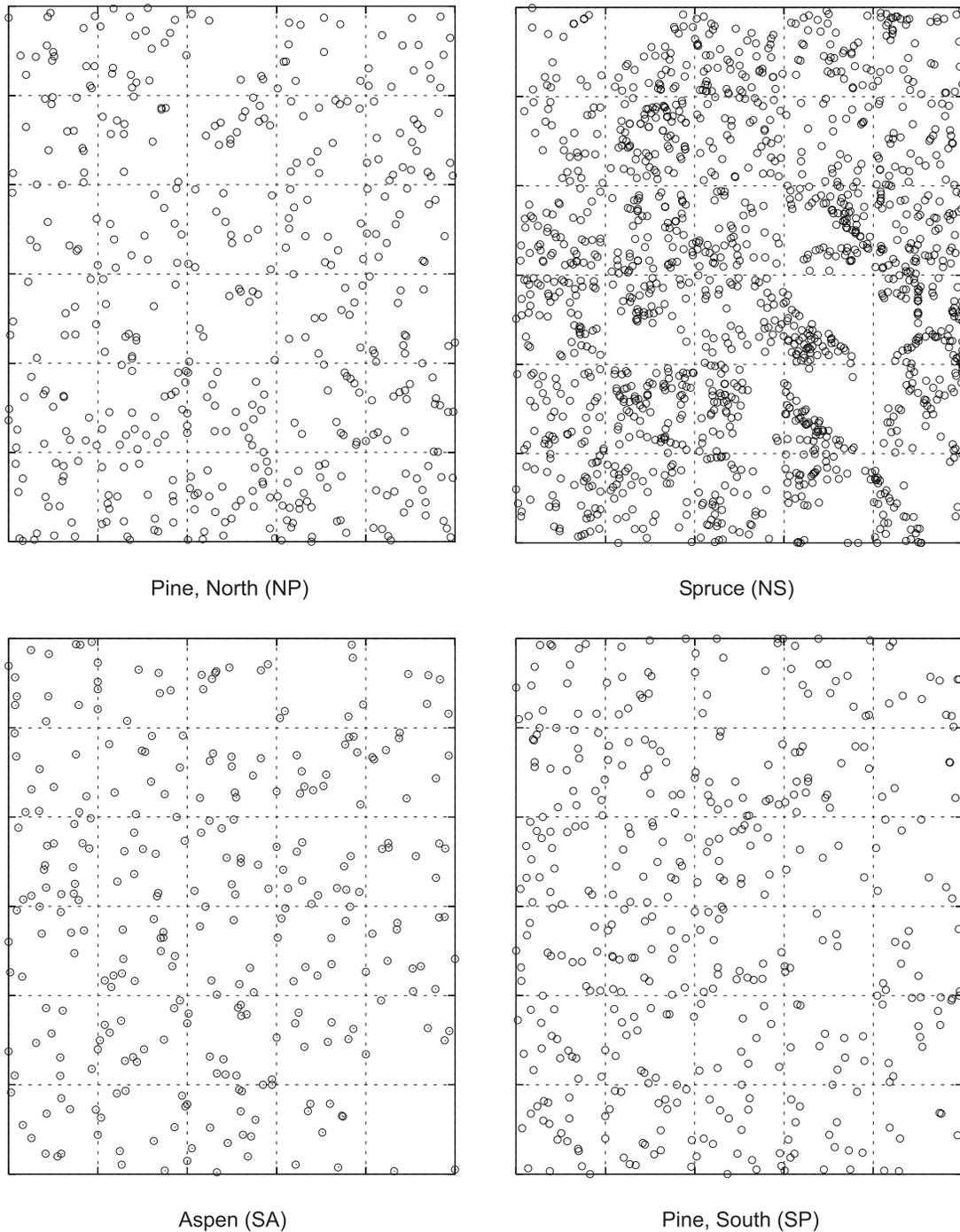
to spatial scales ranging from patches of competing trees to entire stands or compartments (Magnussen 1989; García 1992; Fox et al. 2001). Some authors have noted effects of plot size on individual-tree growth model estimates, but attributed them to error variables regression biases (Jaakkola 1967; Stage and Wykoff 1998; Hynynen and Ojansuu 2003; Lappi 2005).

Distance-independent, individual-tree growth models (Munro 1974; Vanclay 1994) ignore spatial structure and area effects. Competition indices in distance-dependent models induce negative size correlations among nearby trees, causing size variability to be higher in smaller plots than in larger plots. On the other hand, as asserted by a reviewer, it may be "intuitively clear" that variability in larger areas tends to be higher than in smaller areas, as larger areas have a greater probability of including different conditions. In fact, both situations occur, possibly with one or the other dominating over different spatial scales, or changing with the age of the stand (Matérn 1960; Bachacou and Decourt 1976; Kenkel et al. 1989; Magnussen 1989; García 1992).

Received 10 February 2006. Accepted 5 May 2006. Published on the NRC Research Press Web site at <http://cjfr.nrc.ca/> on 18 January 2007.

**O. García.** University of Northern British Columbia, 3333 University Way, Prince George, BC V2N 4Z9, Canada (e-mail: [garcia@unbc.ca](mailto:garcia@unbc.ca)).

**Fig. 1.** Mapped tree locations for the four data sets. Dashed lines show the 10 m grid. In the spruce, some regeneration seems to have concentrated on or around fallen stems.



**Table 1.** Statistics for the mapped data.

Site code	BOREAS I.D.	Species	Trees per hectare	Basal area (m <sup>2</sup> /ha)	Arith. mean DBH (cm)
NP	NSA-OJP	Jack pine	1507	15.3	10.9
NS	NSA-OBS	Black spruce	4787	35.4	9.1
SA	SSA-9OA	Aspen	983	37.5	21.6
SP	SSA-OJP	Jack pine	1403	18.6	12.4

To be specific, this paper focusses mainly on DBH variances estimated on single random plots. The variance turns out to be mathematically tractable, with explicit results, but qualitatively similar conclusions can be expected for other distribution characteristics like higher moments, quantiles, and many fitted distribution model parameters. Some of the results, although generally ignored in growth modelling, may seem obvious to forest inventory specialists. Unlike in inventory, however, single-plot estimates are normally used; later in this paper is a discussion of methods applicable in the infrequent instances when a number of random plots in the same stand are available. Moreover, inventory and survey sampling deal mostly with estimating means or totals. The estimation of population variances (as opposed to the variance of estimators) is not treated in the standard literature, therefore, the relevant theory is presented here in some detail. There are important differences: (i) a plot's basal area per hectare is an unbiased estimate of the population value, while in general the DBH variance is not. (ii) In design-based inference, spatial correlation is irrelevant for the estimation of means and totals (Gregoire 1998), but it affects the estimation of more complex quantities such as variances.

In the first part of the article, mapped tree data are used to quantify the magnitude of spatial scale effects. There were substantial area-related differences in DBH variance owing to the spatial correlation of tree diameters. It was also found that single-plot means per tree can systematically overestimate the stand average, although the relative bias is generally smaller and the underlying causes are different. The sampling theory is developed in the second part. It is shown that linear functions of individual tree values, e.g., sums per hectare (and even stand tables) are estimated without bias by the respective plot statistics. But in the presence of spatial correlation, the expectations of nonlinear functions differ, as is the case of variances. The expected plot variance is related to the variogram, and graphs of expectation over plot size for the field data are calculated. Finally, some possible bias reduction approaches are discussed.

## 2. Empirical results

### 2.1. Data

The example data are from the Boreal Ecosystem–Atmosphere Study (BOREAS), an intensive remote-sensing and field study in central Canada focusing on exchanges of energy, water, heat, CO<sub>2</sub> and trace gases between the boreal forest and the atmosphere (Rich and Fournier 1999). Four same-size, single-species plots are available. Coordinates, DBH, and other variables were measured for all trees taller than 2 m on a 50 m × 60 m area, subdivided in the field into subplots on a 10 m grid. Dead trees were ignored (Fig. 1).

Site characteristics and their BOREAS TE-23 identifiers are listed in Table 1. NSA and SSA refer to the northern study area in Manitoba and the southern study area in Saskatchewan, respectively. The forests are described as old jack pine (OJP), old black spruce (OBS), and old aspen (OA).

For the purposes of this paper, each of the four data sets are referred to as a site, a *stand*, or total area, and *plot* is any contiguous subset of trees. In particular, the next two sections deal with the thirty 10 m × 10 m plots.

**Table 2.** Total and plot DBH variances.

Site	Total (cm <sup>2</sup> )	Plot (cm <sup>2</sup> )	Ratio
Pine, north (NP)	10.16	8.56	0.84
Spruce (NS)	10.74	10.22	0.95
Aspen (SA)	18.41	13.25	0.72
Pine, south (SP)	13.51	10.28	0.76

**Note:** Plot values are the means of 30 plot variances.

**Table 3.** Results of randomization tests for no spatial structure effects on plot variance.

Site	Observed	Perm. range	<i>p</i> value
NP	8.56	8.81–11.66	<10 <sup>-5</sup>
NS	10.22	10.10–11.44	2 × 10 <sup>-4</sup>
SA	13.25	14.75–26.04	<10 <sup>-5</sup>
SP	10.28	11.09–16.29	<10 <sup>-5</sup>

**Note:** From 100 000 permutations.

### 2.2. Plot versus total variances

Calculated finite-population total and plot variances are compared in Table 2. Plot values are means over the 30 plots.

Subplot to plot variance ratios can serve as indicators of spatial structure, as suggested for point processes by Diggle (1979). Unlike in the theory of point processes, here spatial structure does not refer to the pattern of tree locations, but to the spatial relationships of tree size conditional on the tree positions. If competition causes neighbouring tree sizes to differ more than in the stand average, one would expect ratios > 1 for small subplots. If the effect of microsite or genetic similarities predominates, ratios would be < 1. Other examples of variance ratios are given by García (1992, 1998) and García and Batho (2005).

These within-plot variances are related to the between-plot variances used in forest inventory (although the contexts are different). Let  $y_{ij}$  denote the size of tree  $i$  in plot  $j$ . Sums of squares can be partitioned according to the following analysis of variance (ANOVA)

	df	SS	MSS
Within	$N - m$	$\sum_j^m \sum_i^{n_j} (y_{ij} - \bar{y}_j)^2$	$\overline{S_w^2}$
Between	$m - 1$	$\sum_j^m n_j (\bar{y}_j - \bar{y})^2$	$\bar{n} S_b^2$
Total	$N - 1$	$\sum_j^m \sum_i^{n_j} (y_{ij} - \bar{y})^2$	$S^2$

(García 1992; Cochran 1963, for fixed  $n_j$ ). Here  $S^2$  is the total site variance for a total of  $N$  trees and  $\overline{S_w^2}$  is the mean variance within the  $m = 30$  plots (slightly different from those in Table 2 in that  $\overline{S_w^2}$  is weighted by degrees of freedom). It is seen that the within-plot variance has a counterpart in  $S_b^2$ , the variance of the plot means. Although the plot size effect on  $\overline{S_w^2}$  has been largely ignored, the effect on  $S_b^2$  is well-known (e.g., Whittle 1956; Cochran 1963; Sukhatme and Sukhatme 1970). In forest inventory  $S_b^2$  is usually (but not always) found to be larger than  $S^2/n$ , the expected value in the absence of spatial structure (Matérn 1960; Loetsch and Haller 1964; Magnussen 1989; García 1992).

**Table 4.** Total and plot DBH means.

Site	Total (cm)	Plot (cm)	Perm. range (cm)	<i>p</i> value
NP	10.92	11.38	10.70–11.15	<10 <sup>-5</sup>
NS	9.13	9.40	8.99–9.28	<10 <sup>-5</sup>
SA	21.63	22.01	21.18–22.18	0.048
SP	12.45	12.98	12.09–12.81	<10 <sup>-5</sup>

**Note:** Plot values are the average of 30 plot means. Randomization tests were based on 100 000 simulations.

The statistical significance of the variance differences in Table 2 can be assessed in at least two ways. A “within” to “between” mean square ratio, with an asymptotic *F* distribution under the hypothesis of no spatial structure, can be computed from the ANOVA table above. This produced *p* values ranging from  $2 \times 10^{-21}$  for NS to  $4 \times 10^{-13}$  for SA. Alternatively, a nonparametric randomization test can be performed by calculating the mean plot variances after permuting at random the diameters among the tree locations. A *p* value is computed as the proportion of values smaller than the observed one. Results are shown in Table 3. In addition, the simulations confirmed that in the absence of spatial structure the plot and total variances are the same, with the mean of the simulations and the observed values agreeing to two or three decimal places.

### 2.3. Means

Plot means were also significantly different from the overall means (Table 4). In typical growth modelling applications, a single-plot mean is calculated; the tabulated plot values represent an average of these under repeated sampling.

A plot mean is the ratio of a sum of diameters and a tree count, both random variables under sampling. It is therefore biased as an estimate of the population mean per tree. With several plots and proper weighting, the bias is generally assumed to be unimportant (Cochran 1963; Sukhatme and Sukhatme 1970), but in this instance the ratio sample size is one. The approximately 5% difference seems to be caused by the negative correlation between DBH and local stand density (Fig. 2). In the randomization tests, where this correlation does not exist, there was practically no difference.

### 2.4. Short-range variability

A measure of size variability more relevant to tree growth might consider size differences among interacting trees, rather than those over entire plots or stands. The following identity is used for the variance

$$[1] \quad S^2 \equiv \frac{\sum (Y_i - \bar{Y})^2}{N - 1} = \frac{\sum_i \sum_j (Y_i - Y_j)^2}{2N(N - 1)} \\ = \frac{\sum_{i < j} (Y_i - Y_j)^2}{N(N - 1)}$$

The variance is seen as a mean of squared differences between pairs of data points. It seems reasonable to measure short-range variability by a similar mean, but restricted to pairs of competing or neighbouring trees.

To define neighbours, Brown's area potentially available (APA, Brown 1965) was used. A tree APA contains the ground points that are closer to that tree than to any other tree (Fig. 3).

**Table 5.** DBH variability among neighbouring trees.

Site	Total (cm <sup>2</sup> )	Plot (cm <sup>2</sup> )	Neighbours (cm <sup>2</sup> )
NP	10.16	7.63	7.27
NS	10.74	9.74	9.82
SA	18.41	13.25	14.72
SP	13.51	10.41	9.95

The APAs are also known as Dirichlet cells, and their boundaries as Voronoi or Thiessen polygons (Aurenhammer 1991). Neighbours are those trees that share a polygon side; they are joined by lines in the Delaunay triangulation (Fig. 3).

The triangulations were computed with Fortune's algorithm from Netlib (<http://www.netlib.org/voronoi/sweep2>). Pairs on the outside boundary were excluded from eq. 1 to reduce edge effects. The neighbour variances are compared with the total and plot variances in Table 5. They are close to the plot variances, as might be expected from the small plot size, but considerably lower than the total.

It is also interesting to examine the correlation of neighbouring tree diameters. A direct calculation gives 0.31, 0.10, 0.23, and 0.27, for NP, NS, SA, and SP, respectively. Neighbour sizes are positively correlated here, contrary to the assumption in distance-dependent models. Similarly, Kenkel et al. (1989), among others, have found positive DBH spatial correlations. This might help in explaining why competition indices often do not predict increments better than simple nonspatial stand density measures (e.g. Lorimer 1983; Martin and Ek 1984; Barclay and Layton 1990; Vanclay 1994, p. 61, 69, and 159).

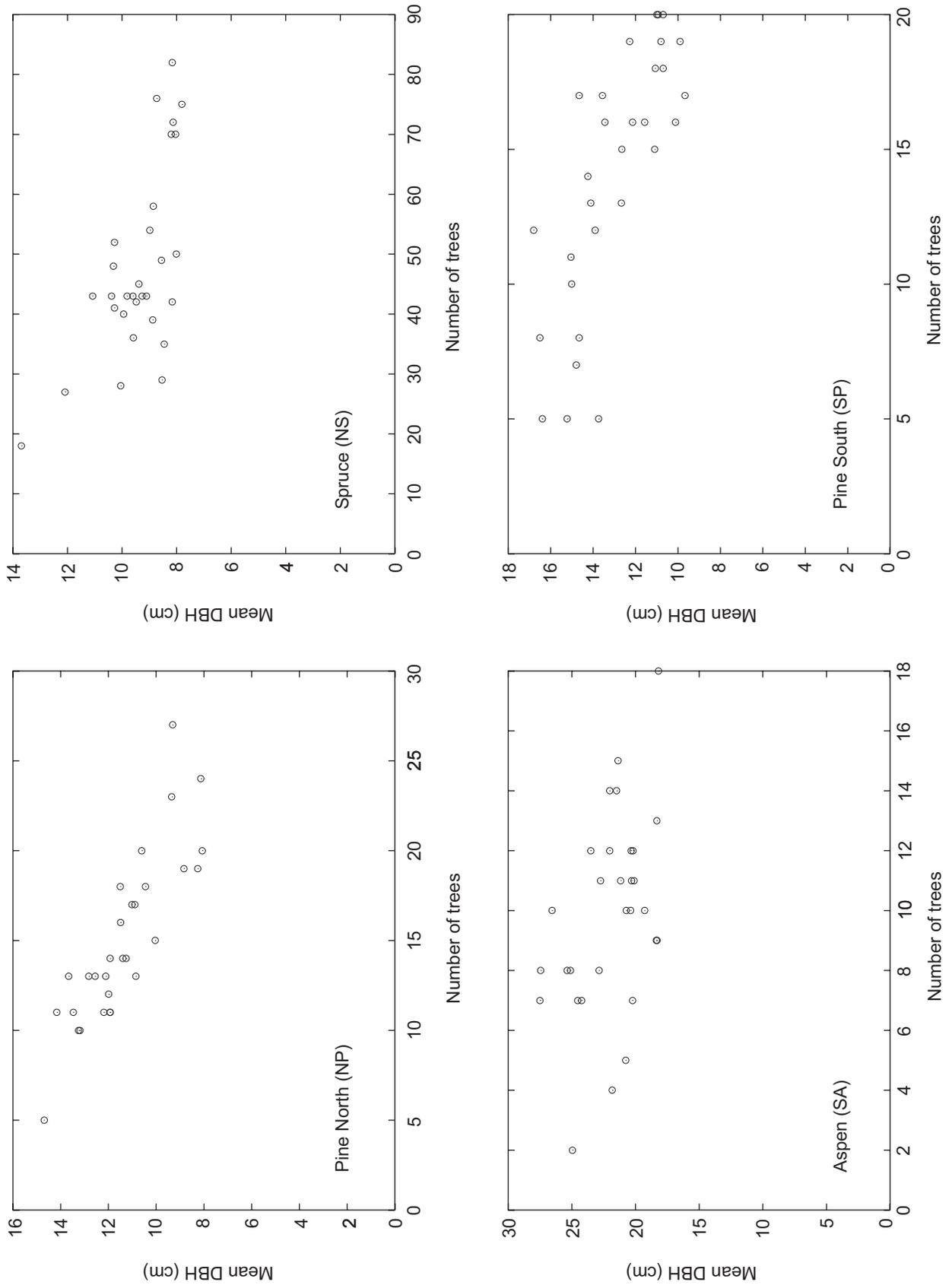
## 3. Theory

### 3.1. Inference

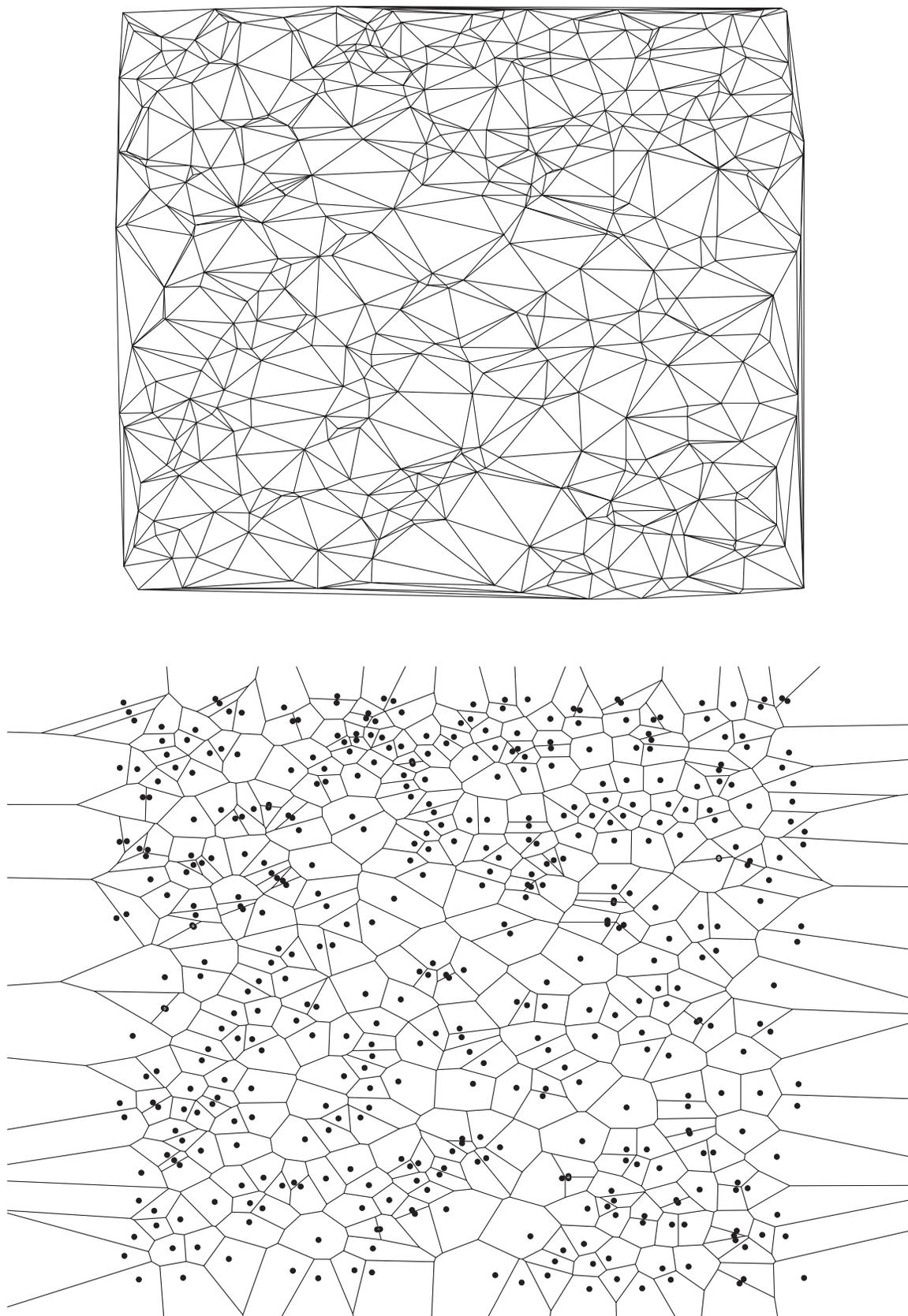
The relationship between size distributions and spatial structure can be studied through spatial statistical models (Matérn 1960; Whittle 1956; Ripley 1981; Magnussen 1989), or through classical finite population sampling methods (Cochran 1963; Sukhatme and Sukhatme 1970); that is, using model- or design-based inference (Gregoire 1998). In the spirit of Matheron's “transitive methods”, the second approach was used, “to see how far it is possible to go without appealing to [hypothetical probabilistic interpretations]” (Matheron 1971). The population is fixed, and the population distribution is the discrete one giving equal weights to the population items.

A comparison and evaluation of model- and design-based inference is outside the scope of this article. It may be useful, however, to point out that many misunderstandings seem to arise from the fact that the nature of the statistical inference involved is quite different. In model-based approaches, the inference is about parameters of an assumed model, given the observed sample; if the model is “true”, that reflects on the population. Design-based inference focuses instead on the long-run properties of a sampling strategy (sampling design plus estimator) under repeated sampling. If the strategy has “good” properties in the long run, its application to a particular sample is likely to produce good results. In other words, design-based inference looks at repeated sampling from a fixed population, while model-based inference is based on one observed sample from a hypothetical “random” population.

Fig. 2. Relationship between plot DBH and density.



**Fig. 3.** Voronoi tessellation (bottom) and Delaunay triangulation (top) for the NP data set.



In what follows, no assumptions are made about tree sizes or spatial patterns. Population properties such as distributions or variograms refer to the real (finite) population, not to a model. The only source of probability is the random location of sample plots.

**3.2. A counterexample**

Some of the lack of awareness about the effects of spatial structure on size distributions might be caused by taking certain statements out of context. For instance, in reviewing the literature, Gregoire (1998) says that under design-based inference “spatial correlation is an irrelevant issue” (p. 1433), and he comments repeatedly on the irrelevance of spatial structure and of pairwise inclusion probabilities. This is perfectly true under his stated objective of estimating totals. It is not true, however, for estimates of more complex quantities such as population variances, fitted distribution parameters, or order statistics. This can be proven with a simple counterexample where exact calculations are possible.

Consider a row of trees at 1 m spacing, with sizes alternating between 1 and 3: . . . 1 3 1 3 1 3 . . . The population may extend indefinitely in both directions, or one may think of an even number  $N$  of trees arranged in a circle. The mean is clearly 2. The deviations from the mean are either 1 or  $-1$ , so that the infinite population variance, i. e., the mean of the squared deviations, is 1; for a finite population the variance is  $N/(N-1)$ .

It can be shown that if trees are selected independently at random, the expected sample variance equals that of the population. Consider, however, a single one-dimensional sample “plot”  $n$  metres long, containing  $n$  contiguous trees. If  $n$  is even, there is only one possible outcome, an equal number of ones and threes. The sample mean is then 2, and the sample variance is  $n/(n-1)$ . If  $n$  is odd, we have to average over two possible patterns; the expected sample mean is again 2, and the expected sample variance turns out to be  $(n+1)/n$  in both cases. The calculated sample variances for the first few plot sizes are

$n$	2	3	4	5	6	7	...	$\infty$
$s^2$	2.00	1.33	1.33	1.20	1.20	1.14	...	1.00

The differences in variance are due to the spatial correlation and to the pairwise inclusion probabilities changing with the distance between trees. As seen in sections 2.3 and 3.4, even the means are affected by spatial structure when  $n$  varies.

**3.3. General sampling relationships**

Using customary notation, let a sample  $(y_1, y_2, \dots, y_n)$  be selected from a population  $(Y_1, Y_2, \dots, Y_N)$  under a sampling design with inclusion probabilities  $\pi_i$  and pairwise inclusion probabilities  $\pi_{ij}$ . That is,  $\pi_i$  is the probability of the population element  $i$  appearing in the sample and  $\pi_{ij}$  is the probability of observing both  $i$  and  $j$ . For instance, the  $Y_i$  may be tree diameters, and the sample may be the list of diameters in a sample plot. The sample size  $n$  may be fixed, as in the previous example and in the lattice plantations of Magnussen (1989), or may vary across samples, as in the BOREAS plots.

For any function  $g$ , the expected value of a sample sum is

$$[2] \quad E \left[ \sum_{i=1}^n g(y_i) \right] = \sum_{i=1}^N \pi_i g(Y_i)$$

Similarly, for any two-variable function  $g$

$$[3] \quad E \left[ \sum_{i=1}^n \sum_{j=1}^n g(y_i, y_j) \right] = \sum_{i=1}^N \sum_{j=1}^N \pi_{ij} g(Y_i, Y_j)$$

Stuart (1963) gives these relationships for the case of fixed  $n$ , but they are also valid for variable  $n$ : let  $\alpha_i$  be 1 if element  $i$  is in the sample, and 0 otherwise; then, the left-hand side of eq. 2 is

$$E \left[ \sum_{i=1}^n g(y_i) \right] = E \left[ \sum_{i=1}^N \alpha_i g(Y_i) \right] \\ = \sum_{i=1}^N E[\alpha_i] g(Y_i) = \sum_{i=1}^N \pi_i g(Y_i)$$

The proof of eq. 3 is analogous.

Making  $g = 1$  and noting that  $\pi_{ii} = \pi_i$ , it is found that

$$[4] \quad \sum_{i=1}^N \pi_i = E[n], \quad \sum_{i<j}^N \pi_{ij} = E[n(n-1)/2]$$

extending the formulae in Cochran (1963) and Stuart (1963) to variable  $n$ .

If all  $\pi_i \neq 0$ , the substitution  $g \rightarrow g/\pi_i$  in eq. 2 gives the Horvitz–Thompson estimator for variable-probability sampling

$$E \left[ \sum_{i=1}^n g(y_i) / \pi_i \right] = \sum_{i=1}^N g(Y_i)$$

That is, an average weighted by the inverse inclusion probabilities is an unbiased estimate of the population total. An analogous result obtained from eq. 3 will be useful later

$$[5] \quad E \left[ \sum_{i=1}^n \sum_{j=1}^n g(y_i, y_j) / \pi_{ij} \right] = \sum_{i=1}^N \sum_{j=1}^N g(Y_i, Y_j)$$

Note also the special cases

$$[6] \quad E \left[ \sum_{i=1}^n 1/\pi_i \right] = N, \quad E \left[ \sum_{i<j}^n 1/\pi_{ij} \right] = N(N-1)/2$$

As pointed out by Fellegi in the discussion to Stuart (1963), much of finite population sampling theory can be obtained as special cases of the above by substituting the inclusion probabilities appropriate to each sampling design.

**3.4. Estimating totals, means, and frequencies**

There are two common models for fixed-size plot sampling in forest inventory, let us call them models I and II. Model I assumes that the possible sample plots, typically square or rectangular, constitute a tessellation of the whole sampled region, as in the examples of section 2 (Loetsch and Haller 1964; Husch et al. 2003). The classical cluster sampling (more specifically area sampling) theory applies (Cochran 1963; Sukhatme and

Sukhatme 1970). In model II, plots are located uniformly at random anywhere within the region (e.g., Gregoire 1998). With typical inventory sample sizes both models lead to similar numerical results. Assume that edge effects are negligible or that they are taken care of in an appropriate way.

In both models, the probability of any tree being included in a plot of area  $a$  is the same,  $\pi_i = a/A$ , where  $A$  is the area of the sampled region. It follows from eq. 2 that any total per unit area is estimated without bias by the corresponding plot quantity.

Apart from quantities such as basal area per hectare ( $g(Y_i) \propto Y_i^2$ ), by choosing the appropriate  $g$  one obtains less obvious cases. With  $g(Y_i) = 1$  for all  $Y_i$ , it is seen that the number of trees per unit area is unbiased. The population distribution function on an area basis, i.e., the number  $F(z)$  of trees per hectare smaller than or equal to  $z$ , can be written in terms of the indicator function

$$I(y, z) = \begin{cases} 1 & \text{if } y \leq z \\ 0 & \text{otherwise} \end{cases}$$

as

$$F(z) = \frac{1}{A} \sum_{i=1}^N I(Y_i, z)$$

Making  $g(Y_i) = I(Y_i, z)$  shows that the expected value of the corresponding sample distribution function per unit area equals that of the population for any  $z$ . Any linear function of  $F(z)$  is also estimated without bias, in particular, the numbers of trees per hectare by size classes (stand tables).

Unless  $n$  is fixed, however, plot sample means per tree are biased as estimates of the population mean; they correspond to the estimation of a ratio (Cochran 1963; Sukhatme and Sukhatme 1970). The bias is related to the variability and to the correlation between the numerator and denominator (Sukhatme and Sukhatme 1970, sections 4.3–4.4) and, as seen in the examples, it can be appreciable. Where  $n$  is less variable, as in managed stands and (or) larger plots, the bias might not be important.

### 3.5. Variances

A good approximation to the expected plot variance for different plot sizes can be calculated without resorting to extensive simulations.

By taking  $g(Y_i, Y_j) = (Y_i - Y_j)^2$ , eqs. 1 and 3 give

$$\begin{aligned} E[s^2] &= E\left[\frac{\sum_{i=1}^n \sum_{j=1}^n (y_i - y_j)^2}{2n(n-1)}\right] \\ &\approx \frac{E\left[\sum_{i=1}^n \sum_{j=1}^n (y_i - y_j)^2\right]}{2E[n(n-1)]} \\ &= \frac{\sum_{i=1}^N \sum_{j=1}^N \pi_{ij} (Y_i - Y_j)^2}{2E[n(n-1)]} \end{aligned}$$

and from eq. 4,

$$[7] \quad E[s^2] \approx \frac{\sum_{i < j} \pi_{ij} (Y_i - Y_j)^2 / 2}{\sum_{i < j} \pi_{ij}}$$

Sums in eq. 7 are over all the pairs of trees in the population. The approximation of the expected ratio by the ratio of expectations can be good because, unlike in section 2.3, no appreciable correlation between plot variances and numbers of trees is observed.

Under both models I and II, the  $\pi_{ij}$  vary strongly with the distance between trees  $i$  and  $j$ . In particular, a model I random sample plot can only contain pairs belonging to a same tile of the tessellation of the stand into plots;  $\pi_{ij}$  is 0 otherwise. García (1992) relates model I sample variances to intracluster correlations. Here, the case of one circular sample plot under model II is further analyzed.

Obviously, pairs of trees further apart than the plot diameter  $D$  can not appear in the sample. Otherwise, it is seen that a pair would be included in any plot whose centre lies within the intersection of two circles of diameter  $D$  centred at the tree locations (Fig. 4). Therefore, ignoring any edge effects, the pairwise inclusion probability  $\pi_{ij}$  is proportional to the overlap of a circle of diameter  $D$  and a copy of it displaced by the intertree distance  $d_{ij}$ . For any figure this is called a *distance function* by Matérn (1960), a *geometrical covariogram* by Matheron (1971), or a *set covariance function* by Stoyan and Stoyan (1994). For the circle it depends only on the distance and circle size

$$[8] \quad K(d, D) = \frac{D^2}{2} \left[ \arccos \frac{d}{D} - \frac{d}{D} \sqrt{1 - \left(\frac{d}{D}\right)^2} \right] \text{ if } d < D, 0 \text{ otherwise}$$

The squared differences in eq. 7, as a function of distance, correspond to the population variogram (or semivariogram)

$$[9] \quad \gamma(d_{ij}) = E\left[(Y_i - Y_j)^2\right] / 2$$

(e.g., Venables and Ripley 2002). The expectation is an average over any repeated distances. Variograms for the example data sets are shown in Fig. 5, with distances rounded to the nearest metre.

Equation 7 can therefore be written in terms of the set covariance function and the variogram

$$[10] \quad E[s^2] \approx \frac{\sum_{i < j} K(d_{ij}, D) \gamma(d_{ij})}{\sum_{i < j} K(d_{ij}, D)}$$

For our purposes, computation is simplified by omitting the expectation in eq. 9, and by including only the pairs with  $d_{ij} < D$ .

Figure 6 shows the results for the BOREAS example data sets. The approach can be extended to rectangular or other plot shapes, with fixed or random plot orientation. Similar calculations for the between-plots variance might be useful in forest inventory design.

As an alternative to the variogram, spatial structure is often described by autocovariance or autocorrelation functions. These are related by

$$\begin{aligned} \text{variogram} &= \text{variance} - \text{autocovariance} \\ &= \text{variance} \times (1 - \text{autocorrelation}) \end{aligned}$$

**Fig. 4.** Plots with centres in the shaded region include the pair of trees  $i$  and  $j$ . The shaded area  $K$  for plots of diameter  $D$  is shown in the graph and in eq. 8 as a function of the intertree distance  $d$ .

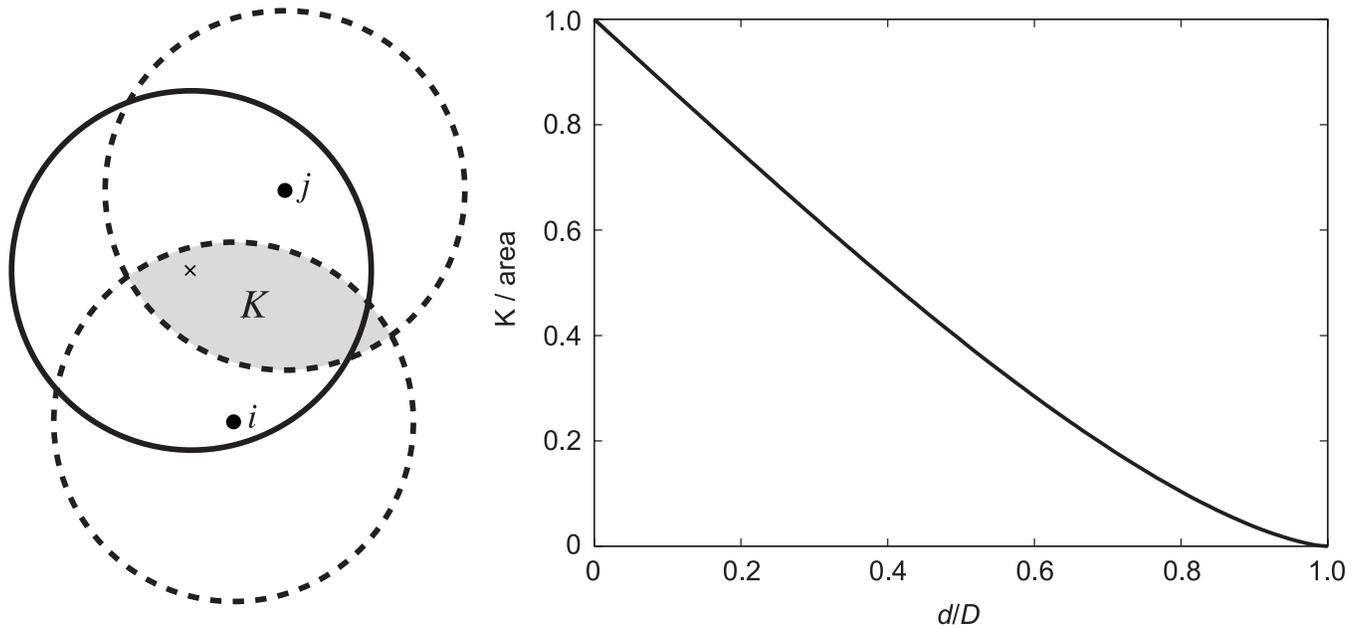


Figure 5 shows positive autocorrelations, that can be interpreted as microsite or genetic autocorrelations masking any competition effects. Competition is apparent in SP, where the variance decreases over a range of small plot sizes. Bachacou and Decourt (1976) and García (1992) give examples of negative short-range autocorrelations where competition dominates. In these, the variance for small plots can be higher than the stand variance.

**4. Bias reduction**

When there are several independent sample plots it is possible to reduce the bias in population estimates. For means per tree, the ratio of the sum over all plots to the total number of observed trees is preferred to the average of the individual plot estimates (Cochran 1963; Sukhatme and Sukhatme 1970).

García (1992) derives an ANOVA-based nearly unbiased variance estimator under inventory model I. For model II one can use the analogue of the Horwitz–Thompson estimator in eq. 5. From eq. 5

$$\begin{aligned}
 & E \left[ \sum_{i < j}^n (y_i - y_j)^2 / \pi_{ij} \right] \\
 &= \frac{1}{2} E \left[ \sum_{i=1}^n \sum_{j=1}^n (y_i - y_j)^2 / \pi_{ij} \right] \\
 &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (Y_i - Y_j)^2
 \end{aligned}$$

Therefore, using eq. 6

$$[11] \quad \hat{s}^2 = \frac{\sum_{i < j}^n (y_i - y_j)^2 / \pi_{ij}}{\sum_{i < j}^n 1 / \pi_{ij}}$$

is a nearly unbiased estimator of the population variance (eq. 1).

It is found that the inclusion probabilities are, ignoring any edge effects

$$[12] \quad \begin{aligned}
 \pi_i &= 1 - (1 - a/A)^m \\
 \pi_{ij} &= 1 - 2(1 - a/A)^m + (1 - 2a/A + K_{ij}/A)^m
 \end{aligned}$$

where  $a$  is the area of each of  $m$  plots selected over an area  $A$  and  $K_{ij}$  is the value of the plot set covariance function for the distance between trees  $i$  and  $j$  (e.g., eq. 8 for circular plots).

The method is not directly applicable to single plots because in that case the inclusion probabilities are zero for pairs of trees further apart than the plot diameter. However, a truncated version of eq. 11, neglecting autocorrelation at larger distances, might be useful. The truncation point could be chosen to try to balance bias and precision. Details would require further research.

**5. Conclusions**

The marginal sampling distribution for single-tree values from random sample plots is the same as the population distribution. Therefore, plot calculations of totals per hectare, frequencies, and other linear functions of the distribution function are unbiased estimates of the corresponding population values.

Fig. 5. DBH variograms with distances grouped in 1 m classes.

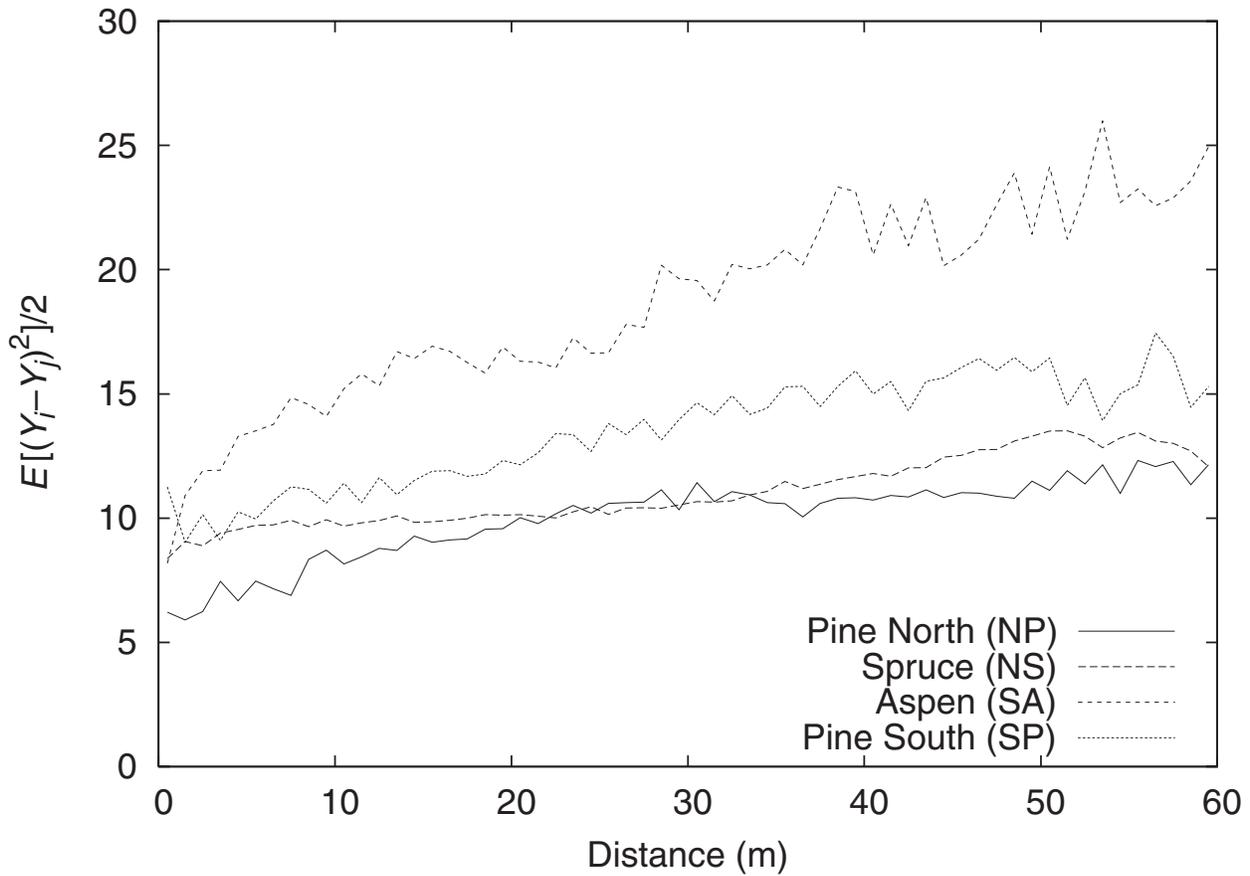
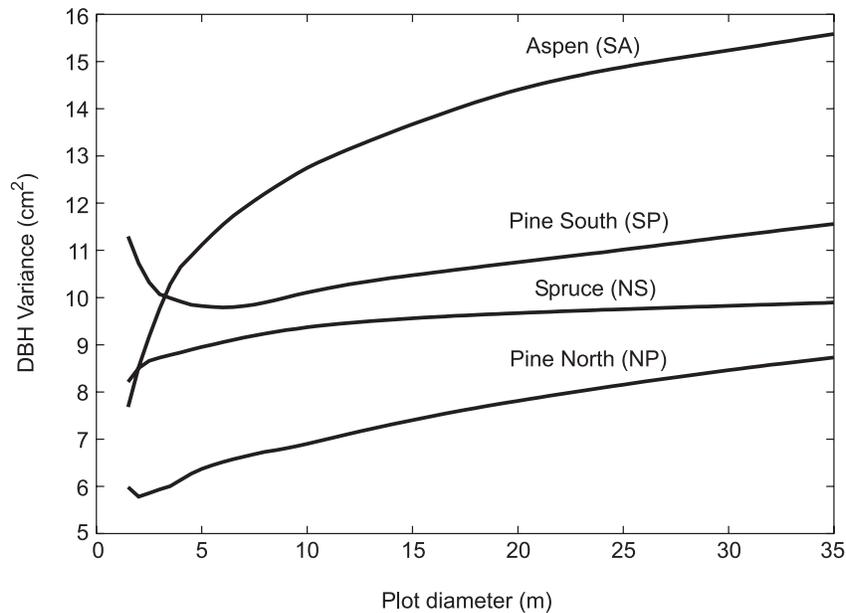


Fig. 6. Calculated circular plot variance for varying plot size.



However, values for different trees in a plot are not statistically independent, and their joint distribution varies with plot size and stand spatial structure. Hence, the expectation of quantities involving nonlinear functions of values from several trees will

vary. These include variances, higher moments, order statistics, and most fitted distribution parameters. In the simple case of the variance, which involves squared differences for pairs of trees, the expectation can be expressed in terms of plot size and shape

and of second-order spatial structure properties. In practical situations, the differences across plot sizes can be substantial.

The fact that a full distribution function can be estimated without bias, but statistics calculated from it are biased, can be counterintuitive. This may have contributed to the neglect of these limitations of DBH and other size distributions in the literature.

Expected means per tree derived from plot data also vary with plot size, although for a different reason. The differences might be important for small plots and natural stands with irregular spatial patterns.

Bias in estimating stand-level variances can be largely eliminated if several random plots are available in the stand. Related methods might be useful with single plots.

Field observations contradict current individual-tree growth model assumptions about short-range spatial structure and tree interactions. It should be possible to obtain more realistic distance-dependent simulations by incorporating microsite, and in some instances genetic, spatial correlations.

## Acknowledgements

I am grateful to Keith Rennolls for stimulating discussions, and to Tim Gregoire, Kim Iles, Steen Magnussen, and Chip Scott for helpful comments on earlier versions of the manuscript.

## References

- Aurenhammer, F. 1991. Voronoi diagrams — A survey of a fundamental geometric data structure. *ACM Comput. Surveys*, **23**: 345–405.
- Bachacou, J., and Decourt, N. 1976. Etude de la compétition dans des plantations régulières à l'aide de variogrammes. *Ann. Sci. For.* **33**: 177–198.
- Barclay, H.J., and Layton, C.R. 1990. Growth and mortality in managed Douglas-fir: relation to a competition index. *For. Ecol. Manage.* **36**: 187–204.
- Brown, G.S. 1965. Point density in stems per acre. *N.Z. For. Res. Note* 38, New Zealand Forest Service.
- Cochran, W.G. 1963. *Sampling Techniques*. 2nd ed. Wiley, New York.
- de Liocourt, F. 1898. De l'aménagement des sapinières. *Bulletin trimestriel, Société forestière de Franche-Comté et Belfort*, pp. 396–409. (English translation by Maria Nygren, The School of Natural Resources, University of Missouri-Columbia, August 2001. Available from <http://www.snr.missouri.edu/silviculture/>).
- Diggle, P.J. 1979. On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics*, **35**: 87–101.
- Fox, J.C., Ades, P.K., and Bi, H. 2001. Stochastic structure and individual-tree growth models. *For. Ecol. Manage.* **154**: 261–276.
- García, O. 1992. What is a diameter distribution? *In Proceedings of the Symposium on Integrated Forest Management Information Systems — An international symposium. Edited by M. Minowa and S. Tsuyuki. Japan Society of Forest Planning Press, Tokyo, Japan.* pp. 11–29. Available from <http://web.unbc.ca/garcia/publ/japan.pdf>.
- García, O. 1998. Estimating top height with variable plot sizes. *Can. J. For. Res.* **28**: 1509–1517.
- García, O., and Batho, A. 2005. Top height estimation in lodgepole pine sample plots. *West. J. Appl. For.* **20**: 64–68.
- Gregoire, T.G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. *Can. J. For. Res.* **28**: 1429–1447.
- Husch, B., Beers, T.W., and Kershaw, J.A., Jr. 2003. *Forest mensuration*. 4th ed. Wiley, New York.
- Hynynen, J., and Ojansuu, R. 2003. Impact of plot size on individual-tree competition measures for growth and yield simulators. *Can. J. For. Res.* **33**: 455–465.
- Jaakkola, S. 1967. On the use of variable size plots for increment research. *In Proceedings of IUFRO XIV Congress, Section 25, German Union of Forest Research Organizations (DVFFA), Munich. International Union of Forestry Research Organizations (IUFRO), Vienna Austria.* pp. 371–378.
- Kenkel, N.C., Hoskins, J.A., and Hoskins, W.D. 1989. Local competition in a naturally established jack pine stand. *Can. J. Bot.* **67**: 2630–2635.
- Lappi, J. 2005. Plot size related measurement error bias in tree growth models. *Can. J. For. Res.* **35**: 1031–1040.
- Loetsch F., and Haller, K. 1964. *Forest inventory. Vol. I.* BLV Verlagsgesellschaft, Munich.
- Lorimer, C.G. 1983. Test of age-independent competition indices for individual trees in natural hardwood stands. *For. Ecol. Manage.* **6**: 343–360.
- Magnussen, S. 1989. Inter-plant interactions and their influence on within and among plot variances. *Scand. J. For. Res.* **4**: 369–377.
- Martin, G.L., and Ek, A.R. 1984. A comparison of competition measures and growth models for predicting plantation red pine diameter and height growth. *For. Sci.* **30**: 731–743.
- Matérn, B. 1960. Spatial variation. *Meddelanden från Statens Skogsforskningsinstitut*, **49**, No. 5. (Reprinted by Springer-Verlag, 1986).
- Matheron, G. 1971. The theory of regionalized variables and its application. *Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau 5, École Nationale Supérieure des Mines de Paris.*
- Munro, D.D. 1974. Forest growth models: a prognosis. *In Growth Models for Tree and Stand Simulation. Royal College of Forestry, Stockholm, Sweden. Research Note 30.* pp. 7–21.
- Rich, P.M., and Fournier, R. 1999. BOREAS TE-23 map plot data [online]. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee. Available from <http://www.daac.ornl.gov>.
- Ripley, B.D. 1981. *Spatial statistics*. Wiley, New York.
- Stage, A.R., and Wykoff, W.R. 1998. Adapting distance-independent forest growth models to represent spatial variability: effect of sampling design on model coefficients. *For. Sci.* **44**: 224–238.
- Stoyan, D., and Stoyan, H. 1994. *Fractals, random shapes and point fields — Methods of geometrical statistics*. John Wiley & Sons, Ltd., Chichester, UK.
- Stuart, A. 1963. Some remarks on sampling with unequal probabilities. *Bull. Int. Statistical Inst.* **40**: 773–779.
- Sukhatme, P.V., and Sukhatme, B.V. 1970. *Sampling theory of surveys with applications*. Iowa State University Press, Ames, Iowa.
- Vanclay, J.K. 1994. *Modelling forest growth and yield: applications to mixed tropical forests*. CABI International, Wallingford, Oxfordshire, UK.
- Venables, W.N., and Ripley, B.D. 2002. *Modern applied statistics with S*. 4th ed. Springer, New York.
- Whittle, P. 1956. On the variation of yield variance with plot size. *Biometrika*, **43**: 337–343.