Distributions and spatial structure

Oscar García

University of Northern British Columbia
Prince George, BC, Canada V2N 4Z9. E-mail: garcia@unbc.ca

Abstract

Diameter 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 causes distribution parameters and their estimates to vary with the extent of ground considered. I examine the magnitude of these effects on some data sets, investigate the theory of the relationships between distributions and spatial structure, and explore bias reduction strategies.

_____

Introduction

Distributions of tree sizes, especially of diameter at breast height (dbh), have been extensively studied and used in forestry. They are estimated from sample-plot data, and used in predicting tree growth, or to estimate the output of forest products of various kinds or sizes. Surprisingly, only a handful of studies seem to have addressed the applicability of such a distribution to spatial scales ranging from patches of competing trees to entire stands or compartments (Magnussen, 1989; García, 1992; Fox et al., 2001). If tree sizes are spatially correlated the distributions should vary with the extent of land considered. I show that plot totals and frequencies are unbiased estimates of stand parameters, but variances and other statistics are not. Attention is focused largely on diameters and variances, but the conclusions are more general.

Empirical Results

Data

The effects of spatial structure are illustrated with mapped tree data from the Boreal Ecosystem-Atmosphere Study (BOREAS), an intensive remote-sensing and field study in central Canada (Rich and Fournier, 1999). Diameters at breast height (dbh) in four 50×60 m single-species "stands" were used. The species are jack pine, black spruce, and trembling aspen. Each stand is subdivided into 30 "plots" on a 10 m grid (Figure 1).

Variances

Calculated finite-population total (stand) and plot variances are compared in Table 1. Plot values are averages over the 30 plots.
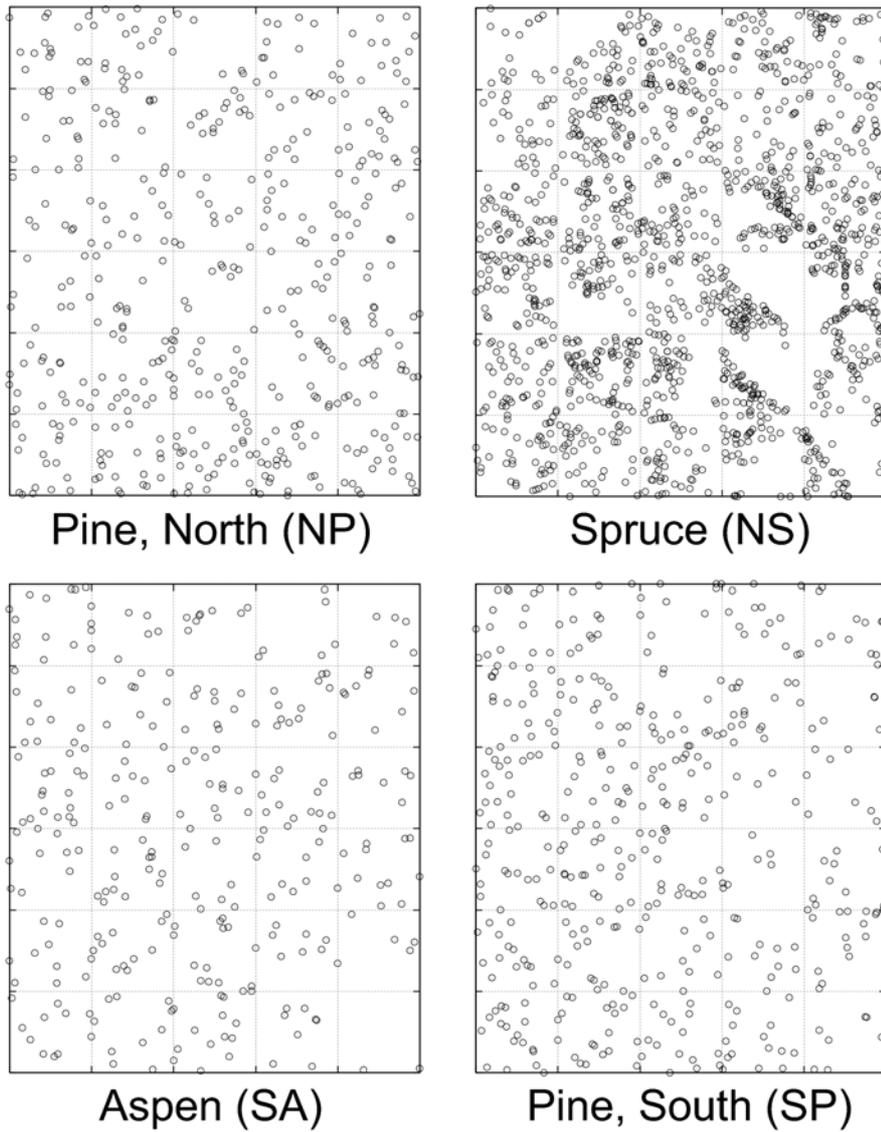
Figure 1: Tree locations for the four data sets. Dashed lines show the 10 m grid. In the spruce some regeneration seems to have concentrated on or around fallen stems.

Table 1: Stand and plot dbh variances (cm^2), and results of permutation tests for no spatial structure effect on plot variances. 100,000 permutations.

Site	Stand	Plot	Ratio	Perm. range	P-value
Pine, North (NP)	10.16	8.56	0.84	8.81 - 11.66	$<10^{-5}$
Spruce (NS)	10.74	10.22	0.95	10.10 - 11.44	2×10^{-4}
Aspen (SA)	18.41	13.25	0.72	14.75 - 26.04	$<10^{-5}$
Pine, South (SP)	13.51	10.28	0.76	11.09 - 16.29	$<10^{-5}$

The ratio of plot to total variance can serve as an indicator of spatial structure, as suggested for point processes by Diggle (1979)[1] . If competition causes neighbouring tree sizes to differ more than in the stand average, one would expect ratios larger than 1 for small plots. If the effect of micro-site or genetic similarities predominates, ratios would be smaller than 1. Other examples of variance ratios are given by García (1992, 1998) and García and Batho (2005).

To assess the statistical significance of the variance differences, a nonparametric permutation test was performed by calculating the mean plot variances after permuting at random the diameters among the tree locations. The *P*-value, based on 100,000 permutations, is the proportion of values smaller than the observed one (Table 1). The simulated samples also confirmed that in the absence of spatial structure the plot and total variances are essentially the same.

By partitioning sums of squares in an analysis of variance (ANOVA), García (1992) showed that these within-plot variances are a counterpart to the between-plot variance of plot means, used in survey sampling and forest inventory. Although the area effect on within-plot variances has been largely ignored, the effect on between-plot variances is well known (e.g. Whittle, 1956; Cochran, 1963; Sukhatme and Sukhatme, 1970). Forest inventory findings imply that plot to stand variance ratios are usually, but not always, smaller than 1 (Matérn, 1960; Loetsch and Haller, 1964; Magnussen, 1989; García, 1992).

Means

Perhaps surprisingly, plot means were also significantly different from the stand means, larger by about 5%. 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. The bias, however, is generally assumed to be unimportant (Cochran, 1963; Sukhatme and Sukhatme, 1970). In this instance the difference is caused by a negative correlation between dbh and local stand density. In permutation tests, where this correlation does not exist, there was practically no difference.

Short-range variability

For growth prediction, size differences among interacting trees should be relevant, rather than those over entire sample plots or stands. I make use of the following identity for the variance:

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

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

To define neighbours I used Brown's area potentially available (APA; Brown, 1965). A tree APA, also known as a Dirichlet cell or Voronoi polygon, contains the ground points that are closer to that tree than to any other tree (Figure 2). Neighbours are those trees that share a polygon side; they are joined by lines in the Delaunay triangulation.

---

[1] 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 sizes conditional on the tree positions.

Pairs on the outside boundary were excluded. The neighbour variances are compared to the stand and plot variances in Table 2. They are close to the plot variances, as might be expected from the small plot size, but considerably lower than those for the stand.
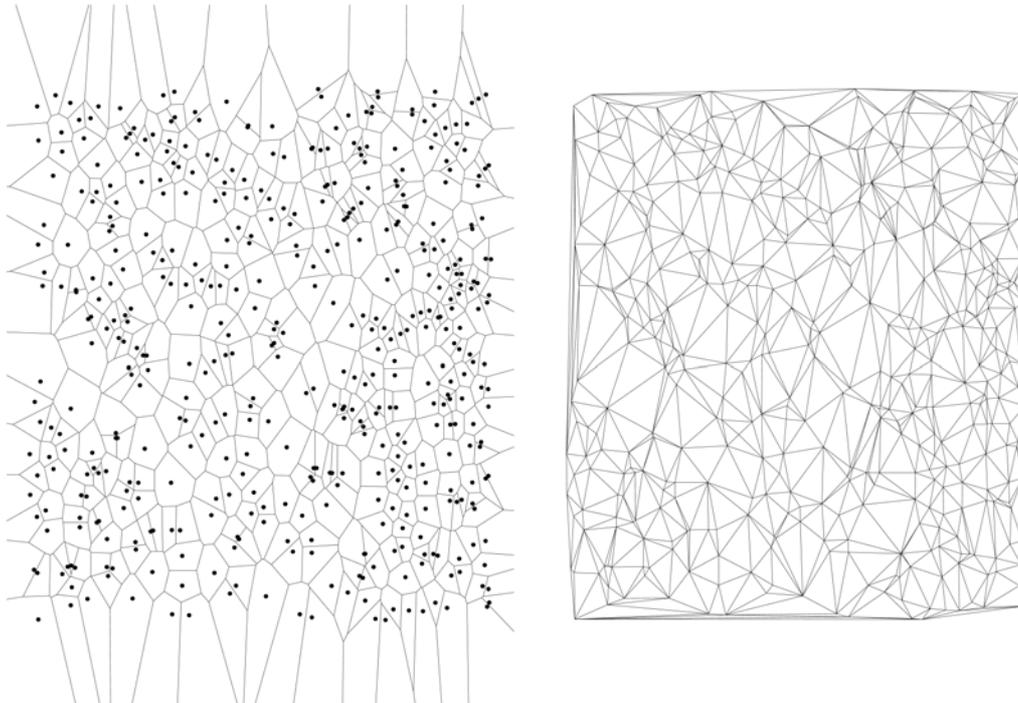

Figure 2: Voronoi tessellation (left), and Delaunay triangulation (right) for the NP data set

Table 2: Dbh variability among neighbouring trees

Site	Stand	Plot	Neighbours
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

Note that distance-independent individual-tree growth models ignore any distribution differences with spatial scale. The competition indices in distance-dependent models cause short-range variances to be higher.

Theory

Here I use design-based inference, as in classical finite population sampling (Cochran, 1963; Sukhatme and Sukhatme, 1970), rather than the model-based inference of spatial statistics (Matérn, 1960; Whittle, 1956; Magnussen, 1989). Although more restricted in applicability, design-based inferences do not depend on the "validity" of hypothetical population models (Gregoire, 1998).

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 designed-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 true for his main stated objective of estimating totals. It is not true for estimates of more complex quantities such as variances, fitted distribution parameters, or order statistics. This can be proven with a simple counter-example where exact calculations are possible.

Consider a row of trees at 1 m spacing, with sizes alternating between 1 and 3:

$$\ldots 1 \; 3 \; 1 \; 3 \; 1 \; 3 \; 1 \; 3 \ldots$$

The population may extend indefinitely in both directions, or we may think of an even number N of trees arranged on a circle. The mean is clearly 2. The deviations from the mean are either 1 or -1, so that the infinite population variance, the mean of the squared deviations, is 1; for a finite population it is $N/(N\text{-}1)$.

If trees are selected independently at random, the expected sample variance equals that of the population. Consider, however, a one-dimensional sample "plot" n metres long, containing n contiguous trees. It is found that the sample mean is 2, and the sample variance is $n/(n\text{-}1)$ if n is even, or $(n+1)/n$ if n is odd. For the first few plot sizes the variances are:

n	2	3	4	5	6	7	…	∞
s^2	2.00	1.33	1.33	1.20	1.20	1.14	…	1.00

The differences in variance are due to spatial correlation, and to the pairwise inclusion probabilities changing with the distance between trees. As seen before, even means are affected by spatial structure when n is random.

In what follows only the main results are given, details and proofs are available from the author.

I distinguish two models for fixed-size plot sampling, call them I and II. Model I assumes that the possible sample plots are a partition of the whole sampled region (a frame), as in the examples above. In Model II plots are located uniformly at random anywhere within the region (e.g., Gregoire, 1998). Assume that edge effects are negligible, or that they are taken care of in an appropriate way.

Totals, frequencies, means

In both model I and II, the probability π_i of any tree being included in a plot of a given size is the same. It follows from general variable-probability sampling relationships that any total per unit area is estimated without bias by the corresponding plot quantity.

Apart from quantities such as basal area per hectare, the result applies to less obvious cases, such as the number of trees per unit area, and the (finite) distribution function on an area basis, i. e., the number $F(z)$ of trees per hectare smaller than or equal to z. That is, for any fixed z, the expected value of the sample empirical distribution function per unit area equals the population distribution function per unit area. 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 the number n of trees in a plot is fixed, however, 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 and, as seen in the examples, it can be appreciable. With less variable n, as in managed stands and/or larger plots, the bias might not be important.

Variances

It is shown that the expected plot variance for different plot sizes can be calculated without resorting to extensive simulations. Under both models I and II, the joint inclusion probabilities π_{ij} vary strongly with the distance between trees i and j. In particular, a Model I random sample plot can only contain pairs of trees belonging to the same element of the sampling frame; π_{ij} is 0 otherwise. García (1992) relates Model I sample variances to intracluster correlations. Here I analyze further the case of one circular sample plot under Model II.

For a given plot shape, the joint inclusion probability for a pair of trees is proportional to a function of the (vector) inter-tree distance, variously called a *distance function* (Matérn, 1960), *geometrical covariogram* (Matheron, 1971), or *set covariance function* (Stoyan and Stoyan, 1994). For a circular plot it depends only on the absolute scalar distance and the plot size, and it is obviously zero for distances larger than the plot diameter (García, 1992).

Using equation (1) and known variable-probability sampling relationships, the expected sample variance can be written as a convolution of the set covariance function and the population semi-variogram $\gamma(d_{ij}) = E[(Y_i - Y_j)^2]/2$. Figure 3 shows the results for the BOREAS example data sets.
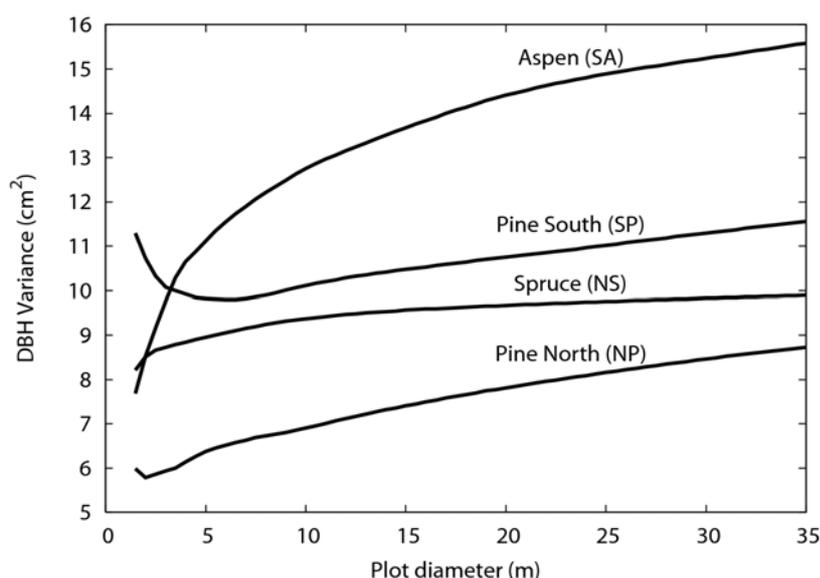

Figure 3: Calculated circular plot variance for varying plot size.

<u>Bias reduction</u>

When there are several independent sample plots it is possible to reduce the bias in population variance estimates. García (1992) derives an ANOVA-based nearly unbiased variance estimator under inventory Model I. For Model II, an analogue of the Horwitz-Thompson estimator based on joint inclusion probabilities can be used.

The method is not applicable to single plots because in that case the inclusion probabilities for some (many) pairs of trees are zero.

Conclusions

Tree size distributions can only be meaningful at a particular spatial scale. Even then, observed tree sizes are not independently distributed. In general, statistics calculated from them vary with sample plot size, which should be specified. Distribution properties are not the same at the stand, plot, and forest patch levels.

Expected values of totals per hectare, frequencies, and other "single-point" quantities are independent of sample area. This is true for any linear function of the distribution function, including the empirical distribution itself. However, expected values of variances, higher moments, fitted distribution parameters, and of other "multi-point" quantities vary with the extent of terrain considered.

The fact that a distribution function can be estimated without bias, but statistics calculated from it are biased, can be counter-intuitive. This may have contributed to a generalized misuse of dbh and other size distributions in the literature.

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

Reference

Brown, G. S. (1965) Point density in stems per acre. New Zealand Forestry Research Notes 38, NZ Forest Service.

Cochran, W. G. (1963) *Sampling Techniques*. Second Edn. Wiley.

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. *Forest Ecology and Management* 154, 261–276.

García, O. (1992) What is a diameter distribution? In: Minowa, M. and Tsuyuki, S. (eds.) *Proceedings of the Symposium on Integrated Forest Management Information Systems — An International Symposium*. Japan Society of Forest Planning Press, pp. 11–29. http://web.unbc.ca/~garcia/publ/japan.pdf.

García, O. (1998) Estimating top height with variable plot sizes. *Canadian Journal of Forest Research* 28, 1509–1517.

García, O. and Batho, A. (2005) Top height estimation in lodgepole pine sample plots. *Western Journal of Applied Forestry* 20, 64–68.

Gregoire, T. G. (1998) Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research* 28, 1429–1447.

Loetsch, F. and Haller, K. (1964) *Forest Inventory*, Vol. I. BLV Verlagsgesellschaft.

Magnussen, S. (1989) Inter-plant interactions and their influence on within and among plot variances. *Scandinavian Journal of Forest Research* 4, 369–377.

Matérn, B. (1960) Spatial variation. *Meddelanden frøan Statens Skogsforskningsinstitut* 49. (Reprinted by Springer-Verlag, 1986).

Matheron, G. (1971) The theory of regionalized variables and its applications. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau 5, École Nationale Supérieure des Mines de Paris.

Rich, P. M. and Fournier, R. (1999) BOREAS TE-23 map plot data. Data set available on-line (http://www.daac.ornl.gov) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.

Stoyan, D. and Stoyan, H. (1994) Fractals, Random Shapes and Point Fields — Methods of Geometrical Statistics. Wiley.

Sukhatme, P. V. and Sukhatme, B. V. (1970) *Sampling Theory of Surveys with Applications.* Iowa State University Press.

Whittle, P. (1956) On the variation of yield variance with plot size. *Biometrika* 43, 337–343.