On Estimation in *k*-tree Sampling

Christoph Kleinn and František Vilčko[1]

Abstract.—The plot design known as *k*-tree sampling involves taking the *k* nearest trees from a selected sample point as sample trees. While this plot design is very practical and easily applied in the field for moderate values of *k*, unbiased estimation remains a problem. In this article, we give a brief introduction to the history of distance-based techniques in forest inventory sampling, present a new and simple approximation technique for estimation, and describe how to eventually develop a design-unbiased estimator. This article draws on two manuscripts that were recently published (Kleinn and Vilčko 2006a, Kleinn and Vilčko 2006b), in which more details are elaborated.

Introduction

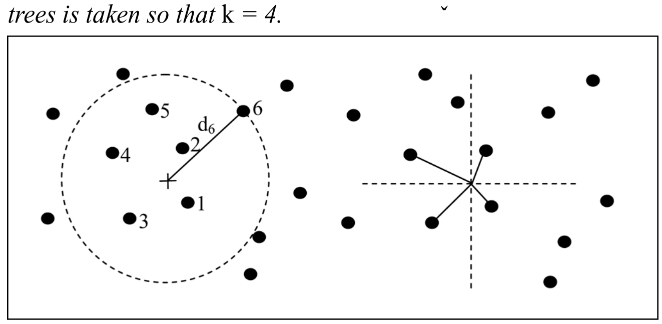
The plot design known as *k*-tree sampling, in which from a sample point the *k* nearest trees are taken as sample trees, is a practical response design if *k* is not too big. We call this approach here "classical *k*-tree sampling" to distinguish it from variations such as the point-centered quarter method (fig. 1) or T-square sampling. (e.g., Krebs 1999).

Estimation for *k*-tree sampling is frequently done in a design-based manner with expansion factors that "expand" the per-plot observation to per-hectare values. One of the frequently used estimation approaches for classical *k*-tree sampling is to use the distance to the *k* nearest tree as radius of a virtual circle plot and calculate a per-plot expansion factor. Another approach is to take the mean distance to the *k* tree from all *n* sample points to calculate an overall expansion factor to be applied to all *n* sample points.

It had long been known, however, that *k*-tree sampling is not an unbiased estimator, but leads on average to a systematic overestimation of the population parameters. From simulation studies on different populations, Payandeh and Ek (1986) suggest that the relatively rare application of *k*-tree sampling in forest inventory has to do with the lack of an unbiased estimator. While some authors see minor problems regarding application of *k*-tree sampling because it is practical and because the bias in the commonly used estimators was found to be modest in many cases (Krebs 1999), others tend to advise against it when unbiased estimation is an issue because it violates basic principles of statistical sampling (e.g., Mandallaz 1995, Schreuder 2004).

Empirical approaches for estimation have been investigated and various techniques are available. One may distinguish two major groups of estimators: (1) design-based estimators that attempt to find from the *k*-tree sample a suitable plot size that allows good extrapolation, and (2) approaches under model assumptions in which estimation depends on the spatial pattern that needs to be captured and described from the sample. Picard *et al.* (2005) give a comprehensive overview of many of these approaches.

Figure 1.—*Two strategies of* k*-tree sampling. Left: the "classical"* k*-tree plot, in which the* k *trees nearest to a sample point (+) are taken as sample trees. Right: the point-centered quarter method, in which the space around the sample point is subdivided into four quadrants; in each of those the nearest trees is taken so that* k *= 4.*



Source: Kleinn and Vilčko (2006a).

[1] Institute of Forest Management, Georg-August-Universität Göttingen, Büsgenweg 5, D-37077 Göttingen. E-mail: ckleinn@gwdg.de.

In this article, we give a brief overview of the history of *k*-tree sampling and present a new and simple empirical approximation technique for the classical *k*-tree plot that is elaborated in Kleinn and Vilčko (2006a). The way toward designing unbiased estimation is shown in the last section. More details about that approach are in Kleinn and Vilcko (2006b).

On the History of *k*-tree Sampling

Loetsch *et al.* (1973) state that the first use of distance techniques in forestry applications was mentioned in the book "Forstmathematik" (forest mathematics) by König (1835). König (1835), in fact, had recognized and elaborated with empirical results that stand attributes such as number of stems and basal area depended on inter-tree distances. He developed a model (presented as a table) with which he determined basal area per hectare as a function of mean stand diameter and average distance between trees. This model obviously was not a point-to-tree distance technique, however, but a tree-to-tree distance technique. These tree-to-tree distance techniques were further developed for forest inventory in the 1940s and 1950s (among others, Bauersachs 1942, Köhler 1952, Weck 1953). Essed (1957) analyzed these tree-to-tree distance techniques and explicitly pointed to the problem of systematic overestimation.

According to a literature review, Stoffels (1955) was among the first to elaborate point-to-tree distance sampling in forest inventory. His target attribute was number of stems per hectare (density). He investigated three-tree sampling and recommended to count tree number three only half (meaning that in the three-tree sample there were actually only two and a half trees counted), which was a simple empirical way to attempt compensating for the then unexplainable systematic overestimation. In Germany, with the studies of Prodan (1968) and Schöpfer (1969a, 1969b), *k*-tree sampling was broadly introduced into practical application of forest management inventories. Those authors recommended $k = 6$ because they found it to be a practical number for application and good in terms of statistical performance. Prodan (1968) knew about the systematic overestimation with the simple expansion factor approach that became clear in simulation studies in test stands. To correct for that bias he recommended taking the attributes of the sixth tree only half because that tree was only half contained in the sample plot. While this ad-hoc approach is difficult to justify in theoretical terms, various simulation studies have shown that it works reasonably well under many conditions (Lessard *et al.* 1994, Payandeh and Ek 1986).

In general, *k*-tree sampling has not been as readily used for forest inventory as fixed area plots and relascope sampling. A number of recent applications have been found, however, many of them under difficult conditions in tropical forested landscapes: Hall (1991, Afromontane catchment forests); Lynch and Rusdyi (1999, Indonesian teak plantations); Sheil *et al.* (2003, East Kalimantan natural forests); and Picard *et al.* (2005, Mali savannah). The test data used for simulations in this present study come from the Miombo woodlands in Northern Zambia.

A New and Simple Technique for Estimation in Classical *k*-tree Sampling

The systematic overestimation of the expansion factor-based estimator for the classical *k*-tree plot has been described and illustrated early by Essed (1957). By taking the distance to the *k* tree as a radius of a virtual sample plot, one defines systematically the smallest possible circular sample plot for the contained *k* trees and, therefore, the largest possible expansion factor—which leads immediately to the observed systematic overestimation. Using the distance to the (k+1) tree as plot radius for a *k*-tree plot, the expansion factor (and therefore the estimations) would be smaller and thus lead to systematic underestimation.

If using the distance to the *k* tree as plot radius produces a systematic overestimation and the distance to the (k+1) tree causes a systematic underestimation, we may conclude that the "true" (i.e., adequate for estimation) circular plot radius must be in between.

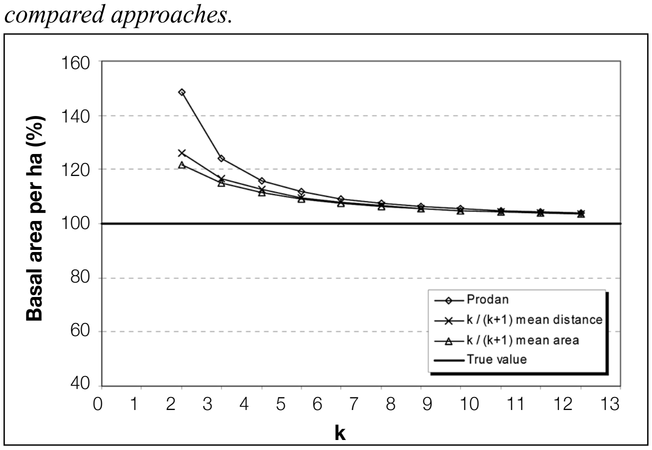
Our idea is simply to use an average plot size from the two distances to the *k* and *k*+1 tree, in which we tested two approaches of calculating the plot radius:

(1) The radius is calculated as arithmetic mean of the distances d_k to the *k* and d_{k+1} to the *k*+1 tree.

(2) The radius is calculated as geometric mean of the circular plot areas from $r = d_k$ and $r = d_{k+1}$. This may be geometrically interpreted as the arithmetic mean of the circle plots with radii d_k and d_{k+1}.

The bias of the these two approaches in comparison to Prodan's (1968) approach is given in figure 2, in which results of a simulation study using a tree map are shown. With all three estimators, a clear positive bias exists, which is, however, smaller for our two approaches, particularly for small values of *k*. For about *k* = 5 onward, the bias for the three approaches is about the same. We should mention here that Prodan (1968) presented his approach only for *k* = 6. We applied it here in a manner analogous to *k* =2..12. Approach (1), in which the plot radius is calculated from the arithmetic mean of d_k and d_{k+1}, produces consistently a smaller bias than approach (2), although the differences are small.

Figure 2.—*Bias of estimating basal area from* k*-tree plots with* k *= 2..12 with different estimators. The two approaches introduced here are contrasted to Prodan's (1968) approach in which the* k *tree is counted half so that the* k*-tree sample actually becomes a (*k*-0.5)-tree sample. While Prodan (1968) proposed that approach for* k *= 6 only, we applied it here to* k *= 2..12. The results are from simulations on a tree map from the Miombo woodlands in Northern Zambia. Our new approaches exhibit smaller bias, in particular for small values of* k*. For about* k *> 6, the bias is about the same for all three compared approaches.*



Source: Kleinn and Vilčko (2006a).

Of course, a simulation study on but one tree map is not an evidence of general superiority, but it may be an indication of promising performance. Kleinn and Vilčko (2006a) present additional simulations with other maps with different spatial patterns with similar results.

Seeing it from a practical point of view and in comparison to Prodan's (1968) approach, for the new approaches one must make one more measurement: the distance to the *k*+1 tree. This measurement adds some additional effort, because the *k*+1 tree must be determined. For relatively small values of *k*, however, this additional effort is expected to be small.

Toward a Design Unbiased Estimator

In Kleinn and Vilčko (2006b), the authors develop a design unbiased estimator for the classical *k*-tree plot. The approach draws on the inclusion zone concept, in which a polygon is drawn around each sample tree with the area of the polygon a measure for the inclusion probability of this particular tree. Once the inclusion probability of all sample trees is known, the Horwitz-Thompson estimator can be used to obtain an unbiased estimator.

The inclusion zone approach is closely linked to the infinite population approach (Eriksson 1995, Mandallaz 1991), also referred to as continuous population approach (Williams 2001). In these approaches, a forest area is considered an infinite population of sample points of which a subset is selected as a sample. That means that the dimensionless points are the sampling elements, and not trees or plot areas. The value that is being assigned to a dimensionless point comes from the surrounding trees. It is the plot design that defines how these trees around the sample point are to be selected. For fixed-area circle plots, for example, all trees up to a defined distance from the sample point are included. In relascope sampling, this distance is not constant but depends on tree diameter and basal area factor. In *k*-tree sampling it is the first, second, etc. *k* nearest tree to the sample point that are included and that determine the value assigned to this particular sample point.

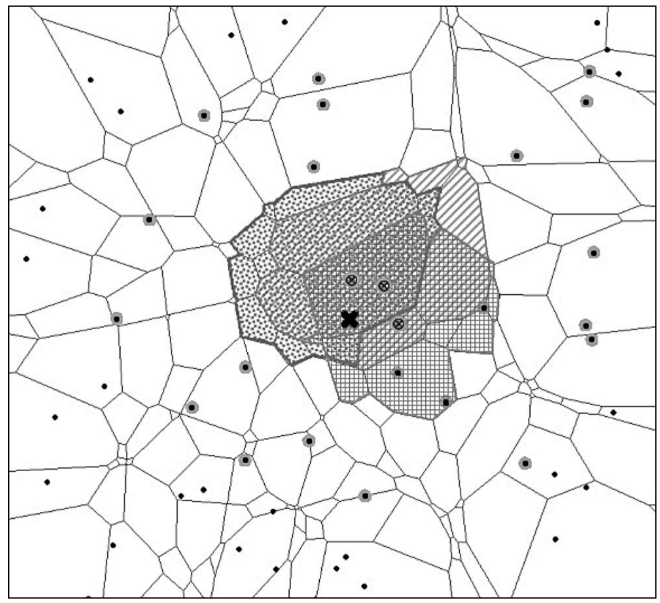
It is more instructive here, however, not to follow that described sample-point-centered approach, but to use a tree-centered approach (Husch *et al.* 1993), which leads immediately to the definition of inclusion zones. Around each tree we build an inclusion zone such that this particular tree is selected by a sample point if it falls into that inclusion zone. It is then obvious, by application of basic principles of geometric probabilities, that the area of this inclusion zone (divided by the total area of the inventory region) defines the probability of selection of that particular tree from which the inclusion probability also can be derived for probabilistic sampling approaches.

Size and shape of the inclusion zone is exclusively defined by the plot design that is being used. For fixed-area circular sample plots, the inclusion zones are circles centered around the trees and with the same size as the sample plot. In relascope sampling, inclusion zones are also circular but size is proportional to the tree's basal area.

To build an unbiased estimator for any plot design, it is sufficient to search for the individual inclusion zones of all sampled trees. Eventually, for *k*-tree sampling, that means that we must find, around an individual sampled "target" tree, the area in which a sample point that falls there has the target tree as nearest, second-nearest, etc. *k* nearest neighbor. For $k = 1$ the solution is simple; the searched inclusion zone polygon are the commonly known Voronoi diagrams or Dirichlet polygons, which have been used in different contexts in forestry (e.g., Lowell 1997, Moore *et al.* 1973, Overton and Stehman 1996).

In Kleinn and Vilčko (2006b), the authors elaborate on inclusion zones for $k > 1$. These inclusion zones contain the set of all points around the target tree for which this particular tree is either the first, second, etc. *k* neighbour. Those polygons are called higher order Voronoi diagrams. Okabe *et al.* (1999) describe approaches for their construction. To do so, the tree positions of neighboring trees must be known; i.e., mapped up to a certain distance. Figure 3 illustrates the approach for $k = 3$, depicting the inclusion zone for all three sample trees. In this case, the coordinates of 15 trees need to be mapped to determine this inclusion zone.

Figure 3.—*Inclusion zones for three trees in a* k*-tree plot for* k *= 3. The sample point is marked by x. The three circled small x's are the three nearest trees. The three differently hatched polygons are the inclusion zones for these trees. Tree positions are marked as dots. To determine the inclusion zones for the* k *= 3 trees, the positions of all trees marked with bold gray dots need to be known.*



Source: Kleinn and Vilčko (2006b).

Shape and size of the inclusion zone depends exclusively on tree positions and not on any attribute value of the target tree. That means that a considerable quantity of additional measurements needs to be done to be able to build the inclusion zones. The proper distance around the sample trees that these position measurements need to be done still has not been determined.

When the inclusion zones of all *k* sample trees are known, then also the inclusion *probabilities* are known, and the Horwitz-Thompson estimator is immediately an unbiased estimator. While application of the Horwitz-Thompson estimator is cumbersome for calculation, Valentine *et al.* (2001) suggest an easier way: one imagines that the tree-specific value of the target attribute is distributed evenly over the inclusion zone, thus forming a density that is constant over the entire inclusion zone. At a selected sample point, one observes the density values of all those inclusion zones that contain the sample point. The sum of these density values is the observation that is used at that point. Another simple approach for calculation would be the one that is commonly used in relascope sampling building on the tree-specific expansion factors.

Conclusions

The inclusion zone approach allows building a design-unbiased estimator (elaborated in detail in Kleinn and Vilčko 2006b). The inclusion zones in *k*-tree sampling are built as higher order Voronoi polygons. Their size and shape vary and depend exclusively on the position of the surrounding trees, a significant difference from inclusion zones of other plot designs. With respect to the expected precision, it is important to note that the size of the inclusion zone—and therefore the inclusion probability as well—is not proportional to any tree attribute. Therefore, it is expected that overall performance of *k*-tree sampling is inferior to other plot designs. This hypothesis, however, is currently being researched by simulation studies. With an unbiased estimator available, it is now possible for the first time to compare *k*-tree sampling to other plot designs, in which also for *k*-tree sampling an unbiased estimator can be used (e.g., Lessard *et al.* 2002, Payandeh and Ek 1986).

Whether our approach will be of relevance for practical field application depends on whether it will be possible to do the required tree mapping around the sample trees with a reasonable amount of effort. Such research is currently ongoing in the research group of the authors; from a selected sample point, polar coordinates of neighboring trees are determined by electronic compass and laser distance meter. The measurement devices are linked to a computer that calculates immediately the Voronoi polygons and indicates whether these polygons change when more and farther trees are included into the mapping; if the polygons do not change any more then tree mapping can be stopped.

It is likely, however, that approximations to estimation will continue to be of great practical relevance. Therefore, a simple approximation approach has been presented in the first half of this paper (elaborated in more detail in Kleinn and Vilčko [2006a]). In addition, it is a subject of research whether simple methods could approximately determine the size of the inclusion zones; for example, by simple regression modeling with the distance to the *k* trees being the independent input variables.

Acknowledgments

Parts of this study were supported by the German Research Council DFG (project KL894/7-1). The authors gratefully acknowledge this support. The authors also thank Dr. Zucchini, Institute of Statistics and Econometrics, Georg-August-Universität Göttingen, for helpful discussions and suggestions.

The authors thank Dr. George Gertner for a critical review, which improved and streamlined the manuscript.

Literature Cited

Bauersachs, E. 1942. Bestandesmassenaufnahme nach dem Mittelstammverfahren des zweitkleinsten Stammabstandes. Forstwiss. Centralbl. 64: 182-186. In German.

Eriksson, M. 1995. Design-based approaches to horizontal-point-sampling. Forest Science. 41(4): 890-907.

Essed, D. 1957. Estimation of standing timber. Wageningen. 60 p. Ph.D. dissertation.

Hall, J.B. 1991. Multiple-nearest-tree sampling in an ecological survey of Afromontane catchment forest. Forest Ecology and Management. 42: 245-266.

Husch, B.; Miller, C.I.; Beers, T.W. 1993. Forest mensuration. 3rd ed. Melbourne, FL: Krieger. 402 p.

Kleinn, C.; Vilčko, F. 2006a. A new empirical approximation for estimation in *k*-tree sampling. Forest Ecology and Management. 237(2): 522-533.

Kleinn, C.; Vilčko, F. 2006b. Design unbiased estimation for point-to-tree distance sampling. Canadian Journal of Forest Research. 36(6): 1407-1414.

Köhler, A. 1952. Vorratsermittlung in Buchenbeständen nach Stammdurchmesser und Stammabstand. Allg. Forst- u. Jagdztg. 123: 69-74. In German.

Krebs, C.J. 1999. Ecological methodology. 2nd ed. San Francisco: Addison Wesley Longman. 620 p.

Lessard, V.C.; Drummer, T.D.; Reed, D.D. 2002. Precision of density estimates from fixed-radius plots compared to *n*-tree distance sampling. Forest Science. 48(1): 1-6.

Lessard, V.C.; Reed, D.D.; Monkevich, N. 1994. Comparing *n*-tree distance sampling with point and plot sampling in northern Michigan forest types. Northern Journal of Applied Forestry. 11(1): 12-16.

Loetsch F.; Zöhrer, F.; Haller, K.E. 1973. Forest inventory. Vol. II. BLV Verlagsgesellschaft München, Bern, Wien. 469 p. In German.

Lowell, K.E. 1997. An empirical evaluation of spatially based forest inventory samples. Canadian Journal of Forest Research. 27: 352-360.

Lynch, T.B.; Rusydi, R. 1999. Distance sampling for forest inventory in Indonesian teak plantations. Forest Ecology and Management. 113: 215-221.

Mandallaz, D. 1991. A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models. Diss. Techn. Wiss. ETH Zürich, Nr. 9378. 257p. http://e-collection.ethbib.ethz.ch/show?type=diss&nr=9378. (13 September 2006).

Mandallaz, D. 1995. "Les hazard fait bien les choses": statistische Methoden für die Waldinventur. Schweiz. Z. Forstwes. 146(12): 1015-1032. In German.

Okabe, A.; Boots, B.; Sugihara, K.; Chiu, S.N. 1999. Spatial tessellations—concepts and applications of Voronoi diagrams. 2nd ed. New York: John Wiley. 671 p.

Overton, W.S.; Stehman, S.V. 1996. The Horwitz-Thompson Theorem as a unifying perspective for probability sampling: with examples from natural resource sampling. The American Statistician. 49(3): 261-269.

Payandeh, B.; Ek, A.R. 1986. Distance methods and density estimators. Canadian Journal of Forest Research. 16: 918-924.

Picard, N.; Kouyaté, A.M.; Dessard, H. 2005. Tree density estimations using a distance method in Mali savanna. Forest Science. 51(1): 7-18.

Prodan, M. 1968. Punktstichprobe für die Forsteinrichtung. Forst- und Holzwirt. 23(11): 225-226. In German.

Schöpfer, W. 1969a. Die 6-Baum-Stichprobe in der Forsteinrichtung. Teil 1. AFZ. 25: 533-536. In German.

Schöpfer, W. 1969b. Die 6-Baum-Stichprobe in der Forsteinrichtung. Teil 2. AFZ. 26: 588-591. In German.

Stoffels, A. 1955. Die Genauigkeit der Bestimmung der Stammzahl pro Hektar durch Messung von Stammabständen. Forstwiss. Centralbl. 74: 211-218. In German.

Valentine, H.T.; Gove, J.H.; Gregoire, T.G. 2001. Monte Carlo approaches to sampling forested tracts with lines and points. Candian Journal of Forest Research. 31: 1410-1424.

Weck, J. 1953. Untersuchung über die Brauchbarkeit und Genauigkeit eines Verfahrens der Bestandesmessung unter Verwendung von Stammabständen. Forstarchiv. 24: 257-260. In German.

Williams, S.W. 2001. Comparison of estimation techniques for a forest inventory in which double sampling for stratification is used. Forest Science. 47(4): 563-576.