ORIGINAL PAPER

An adaptive composite density estimator for *k*-tree sampling

Steen Magnussen · Lutz Fehrman · William J. Platt

Abstract Density estimators for *k*-tree distance sampling are sensitive to the amount of extra Poisson variance in distances to the *k*th tree. To lessen this sensitivity, we propose an adaptive composite estimator (COM). In simulated sampling from 16 test populations, a three-component composite density estimator (COM)–with weights determined by a multinomial logistic function of four readily available ancillary variables–was identified as superior in terms of average relative absolute bias. Results from a different set of nine validation populations–with widely different stem densities and spatial patterns of tree locations—confirmed that relative root mean squared errors (RRMSE) of COM were, on average, considerably lower than those obtained with the three-component *k*-tree density estimators. The RRMSE performance of COM improved with increasing values of *k*. With $k = 6$ and sample sizes of 10, 20, and 30, the average relative bias of COM was between −5 and 5% in seven validation populations but in an open low-density savanna-like population bias reached −12% (1979 data) and 7% (1996 data). For $k = 6$ and $n = 10$, the RRMSE of COM was, in six of the nine validation populations, within 3.3 percentage points of the RRMSE for sampling with fixed-area plots. Jackknife estimates of the precision of COM estimates of density were negatively biased, leading to under-coverage (7%) of computed 95% confidence intervals.

Keywords Plotless sampling · Fixed-count sampling · Forest inventory · Spatial pattern · Over dispersion · Bias · Root mean square error · Coverage

Communicated by G. Kaendler.

S. Magnussen (✉)
Natural Resources Canada, Canadian Forest Service,
506 West Burnside Rd., Victoria, BC V8Z 1M5, Canada
e-mail: steen.magnussen@nrcan.gc.ca

L. Fehrman
Department of Forest Inventory and Remote Sensing, Faculty
of Forest Sciences and Forest Ecology, Georg-August-
Universität Göttingen, Büsgenweg 5, 37077 Göttingen, Germany

W. J. Platt
Department of Biological Sciences 202 LSB, Louisiana State
University, Baton Rouge, LA 70803, USA

Introduction

Distance sampling (also known as plotless or fixed-count distance methods) is an attractive option for ecological surveys (Augustin et al. 2009; Kint et al. 2004; Sandrine et al. 2003; Sheil et al. 2003). Expediency, cost considerations, and flexibility allow adaptations to specific circumstances and objectives of a survey (Augustin et al. 2009). Distance sampling in multi-layered or multi-species populations, for example, can be tailored for simultaneous multi-species viz. multi-objective surveys, thereby greatly reducing sampling variation for less prevalent species and objects of interest (Delince 1986; Picard et al. 2005). Forest inventory by *k*-tree (fixed-count viz. plotless) distance sampling methods offers additional advantages of a uniform workload at every sample location, as well as ease of implementation in highly dense forests and difficult terrain (Delince 1986; Essed 1956; Lessard et al. 1994, 2002). A less recognized—but equally important—advantage emerges in complex stands when an inventory is expected to provide reliable estimates for an a priori specified list of structural elements and species (Sheil et al. 2003). Forest health surveys would be an example (Augustin et al. 2009; Ferretti 1997). Conversely, *k*-tree distance sampling

($k \leq 8$) is less attractive when searching for the kth tree becomes time-consuming as it is in low-density populations. Censoring to a maximum search distance requires a likelihood approach to estimation in order to handle the censoring correctly (Zhang and Chambers 2004, p. 226) and large sample sizes for an accurate estimate of the proportion of censored samples.

The bias in estimates of k-tree estimates of density—and thus all per unit area estimates derived from density—remains the major obstacle to application (Engeman et al. 1994; Kleinn and Vilčko 2006b; Schreuder 2004), although Jonsson et al. (1992) found the bias in per unit area estimates to be of little practical concern. Lynch and Wittver (2003) report on low bias in area-weighted per tree averages obtained by the k-tree distance method. Barabesi (2001) proposed an asymptotically nearly unbiased nonparametric density estimator using point-to-plant sampling. Kleinn and Vilčko (2006a) demonstrated an unbiased estimator for k-tree sampling's inclusion probabilities using a geometric solution for obtaining unbiased inclusion probabilities. Yet their method is impractical for field work. An effective, yet impractical, bias-reduction seems possible through intensive stochastic simulations of spatial patterns that emulate the distribution of observed point-to-trees and tree-to-tree distances (Nothdurft et al. 2010).

There have been many attempts to improve distance methods (for examples, Barabesi and Marcheselli 2004; Byth 1982; Cox et al. 1997; Kendall and Moran 1963; Kleinn and Vilčko 2006b; Lewis 1975; Parker 1979; Patil et al. 1982; Persson 1964; Picard et al. 2005; Pollard 1971; Pollard et al. 2002; Prodan 1968). Recently, Kronenfeld (2009) has suggested to exploit the commonly observed trend of a decline in bias as k increases to obtain an asymptotic estimate of density from a non-linear model fitted to a set of ordered distances. Results suggest an impractically large k-value before the asymptote can be estimated with an acceptable accuracy. Overall, progress has been modest with the exception of otherwise inefficient point-to-nearest-plant distance estimators (Engeman et al. 1994; Pollard 1971).

An extensive evaluation of 19 k-tree distance estimators in simulated sampling from 22 populations (Magnussen et al. 2008a) revealed the estimator proposed by Kleinn and Vilčko (KLEINN viz. KV) as overall best in terms of average relative bias when k was 6. For a sample size (n) greater than 10, a 'GammaPoisson' (GAMPOI viz. GP) model-based estimator by Magnussen et al. (2008b) ranked second. For $k = 4$ and $k = 6$, the overall averages of root mean square errors (RMSE) of KV and GP were practically identical. It was noted that the performance of KV, GP, and that of three other top-ranked estimators was sensitive to the spatial pattern of tree locations in the sampled populations. There were also many examples where the bias of

two top-ranked estimators had opposite signs, which, everything else being equal, suggest an opportunity to produce a less biased estimator by forming a linear combination of estimators with positive and negative bias (Peña 1997).

Based on these observations, we hypothesize that an adaptive composite estimator of density (COM)—defined as a weighted sum of two or more k-tree density estimators (with sum-to-one constraints on the weights)—will be less sensitive—in terms of bias and root mean squared error—to the spatial pattern of a sampled population than a conventional k-tree density estimator. The adaptive property of COM is implemented by expressing the component weights as a multinomial logistic function (Greene 2003, pp. 857–862) of readily available ancillary variables that captures both design-features and departures from complete spatial randomness (*csr*). To test this hypothesis, we evaluated bias of 31 adaptive COM estimators in simulated simple random sampling from 16 test populations. Each COM was a weighted linear sum of two to five k-tree density estimators. Only five top-ranking estimators (MORISITA, PERSSON, KV, ORBIT, GP in Table 3 of Magnussen et al. 2008a) were considered. The COM with the lowest average absolute relative bias across a suite of 16 test population, 18 sampling designs, and 1000 replications was chosen and then validated in the same design settings but using nine different populations with contrasting stem densities (savannah to very dense) and spatial patterns of tree locations (uniform to highly clustered). This study details validation results for the final choice of COM estimator. The performance of COM is benchmarked against results from fixed-area plot sampling and those for the individual component estimators.

Materials and methods

An adaptive composite estimator of density for k-tree distance sampling

The objective is to find an adaptive composite estimator (COM) of stem density (λ, unit : stems m^{-2}) that is robust in the sense that it performs well in populations with contrasting densities and spatial distribution of trees (objects). The estimator must be applicable to simple random sampling (SRS) of the distance to the kth nearest tree $(r_{k,i}, i = 1, \ldots, n)$ from n randomly selected sample locations in the population of interest (=sample frame).

Our search for a robust COM was limited to weighted averages of two to five k-tree density estimators and to the five estimators identified by Magnussen et al. (2008a) as most promising in terms of root mean square errors. The

five estimators are detailed below. All candidate COM estimators have the form

$$\hat{\lambda}_{\text{COM}}(n_c, \mathbf{X}) = \sum_{c=1}^{n_c} \hat{w}_c(\mathbf{X})$$
$$\times \hat{\lambda}_c, 0 \le \hat{w}_c(\mathbf{X}) \le 1, \sum_{c=1}^{n_c} \hat{w}_c(\mathbf{X}) = 1 \qquad (1)$$

where n_c is the number of component estimators in COM, $\hat{w}_c(\mathbf{X})$ is the weight given to $\hat{\lambda}_c$ the cth component estimate of density $\hat{\lambda}_c$, $\mathbf{X}$ is a vector of a readily available ancillary variables ($\mathbf{X}$). The weights are predictions from a multinomial logistic function. In practical applications with the proposed COM, the multinomial logistic function will be known in advance (see "Results" section). An estimate of the sampling variance of COM is obtained via a leave-one-out jackknife procedure (Cochran 1977, p. 178).

The five component estimators considered here are (1) MORISITA (Morisita 1954), (2) PE (Persson 1964), (3) KV (Kleinn and Vilčko 2006b), (4) ORBIT (Magnussen et al. 2008a), and (5) GP (Magnussen et al. 2008b). Note the original KV estimator requires a measurement of $k + 1$ distances. To make all component estimators equal in terms of field effort, we adapted the KV estimator to k distance measures. The MORISITA estimator is

$$\hat{\lambda}_{\text{MORISITA}}(k) = \pi^{-1} \Gamma(k + 0.5)^2 \Gamma(k)^{-2} \bar{r}_k^{-2} \qquad (2)$$

where $\bar{r}_k$ is the sample mean of the n distances between a sample location and the kth nearest tree, The PE estimator (Persson 1964) was designed to be robust against departures from spatial randomness

$$\hat{\lambda}_{\text{PE}}(k) = k\pi^{-1} \operatorname{median}\left[r_k^2\right]^{-1}. \qquad (3)$$

The KV estimator by Kleinn and Vilčko (2006b) is used in a slightly improved form (Magnussen et al. 2008a)

$$\hat{\lambda}_{\text{KV}}(k) = (k - 1)\pi^{-1} \bar{r}_k^{-1} \bar{r}_{k-1}^{-1}. \qquad (4)$$

Location-specific estimates of density, if desired, can be obtained by replacing the means in (4) by their local values $\left(r_{k,i}, r_{k-1,i}, i = 1, \ldots n\right)$. The ORBIT estimator of stem density is (Magnussen et al. 2008a)

$$\hat{\lambda}_{\text{ORBIT}}(k) = \frac{n \times k}{\pi \sum_{i=1}^n r_{k,i} \hat{r}_{k+1,i} - \sqrt{\log(2)} r_{1,i}^2} \qquad (5)$$

where $\hat{r}_{k+1,i}$ is a model-based (regression) estimate of the distance to the $(k + 1)$th tree at the ith sample location. The GP is a model-based estimator (Magnussen et al. 2008b) that assumes a gamma distribution (with parameters α and β) of density across space in a population of interest. Conditional on the local density, the local distribution of tree locations is Poisson (viz. *csr*). The GP estimator is

$$\hat{\lambda}_{\text{GP}}(k) = \hat{\alpha}_k \times \hat{\beta}_k \qquad (6)$$

where $\hat{\alpha}_k$ and $\hat{\beta}_k$ are maximum likelihood estimates obtained from the n sampled k-tree distances $\left(r_{k,i}, i = 1, \ldots, n\right)$.

All possible (26) distinct composite estimators with 2,…,5 component estimators were evaluated. Each is a weighted sum of two, three, four, or five component estimators. The performance of a candidate COM estimator was measured in terms of its relative mean absolute bias $\left(\text{RMAB} = \left|\hat{\lambda}_{\text{COM}} - \lambda_{\text{TRUE}}\right| \times \lambda_{\text{TRUE}}^{-1}\right)$ in simulated SRS across 16 test populations (populations listed in rows 1–6, 8, 11–12, 14–19, and 22 in Table 2 of Magnussen et al. 2008a), sample sizes (n) of 10, 20, and 30, $k = 3,…,8$, and 1000 replications (*rep*) of each setting of population (16) and sampling design n (3) × k (6). The COM estimator with the lowest RMAB became our chosen composite estimator.

For each tested composite estimator, the adaptive weights $\hat{w}_c(\mathbf{X})$ in (1) were determined as the weights that minimized the RMAB of $\hat{\lambda}_{\text{COM}}$ across all test sites, replicates within test sites, and design settings for n and k. Specifically, we determined—a set of regression coefficients in n_c multinomial logistic functions (Greene 2003, pp. 857–862) of the ancillary variables ($\mathbf{X}$) so that

$$\hat{w}_c(\mathbf{X}) = \begin{cases} \left(1 + \sum_{i=2}^{n_c} \exp\left[\hat{\beta}_i' \mathbf{X}\right]\right)^{-1} & \text{if } c = 1 \\ \exp\left[\hat{\beta}_i' \mathbf{X}\right]\left(1 + \sum_{c=2}^{n_c} \exp\left[\hat{\beta}_i' \mathbf{X}\right]\right)^{-1} & \text{if } c > 1 \end{cases} \qquad (7)$$

where $\mathbf{X}$ is a $p \times 1$ row vector of ancillary variables and $\hat{\beta}_c'$ is the transpose of a $p \times 1$ row vector of maximum likelihood estimates of the logistic regression coefficients for components $c = 2,…, n_c$ in COM. By definition, the regression coefficients for the first component (base component, $c = 1$) in a composite estimator are set to zero, which automatically satisfy the sum-to-one constraint on the component weights. Preliminary analyses suggested the following vector of population and design specific ancillary variables $\mathbf{X} = \{1, MV, VR, k, n\}$ where MV is the ratio of the sample mean of r_k to its variance, VR is the ratio of the observed variance of the n distances to the kth tree to the variance expected under complete spatial randomness (*csr*) of tree locations, and finally k and n are as defined earlier. Due to the large amount of data, and in order to avoid confounding the regression coefficients in (7) with random replicate effects (Wolfinger and O'Connell 1993) we determined 1,000 replicate-specific estimates of the regression coefficients in (7) and used their average to compute the final weights.

To compute VR from sample data, one needs an estimate of the variance of r_k expected under *csr* which we have

derived by standard techniques from the density function of r_k (Pollard 1971, Eq. 18). Specifically

$$\text{var}(r_k|\lambda, csr) = \left(\Gamma[k]\Gamma[k+1] - \Gamma[k+0.5]^2 \right)$$
$$\times (\pi \times \lambda(k-1)!\Gamma[k])^{-1} \qquad (8)$$

where λ is the stem density and Γ denotes the gamma function. Since λ is unknown, we use in its place a maximum likelihood estimate $\hat{\lambda}_{\text{MLE}} = (n \times k - 1)\pi^{-1} \times \bar{r}_k^{-2}$ of λ (Pollard 1971).

Fixed-area plot estimator of stem density (benchmark)

For SRS k-tree sampling with n fixed-area circular plots with a radius of R_k and an area of A_k fixed so that the average plot contains k trees, a design-unbiased estimator of stem density is

$$\hat{\lambda}_{\text{FIX}} = \bar{n}_{\text{trees}}(A_k) \times A_k^{-1} \qquad (9)$$

where $\bar{n}_{\text{trees}}$ is the average number of trees within a distance of $R_k = \sqrt{A_k\pi^{-1}}$ from a sample location. Results with this estimator serve as benchmarks for the chosen COM estimator and its individual component estimators. Estimates of sampling variance of $\hat{\lambda}_{\text{FIX}}$ were obtained by standard methods (Cochran 1977, p. 23). The number of trees in a fixed-area plot intersected by one or two population boundaries was multiplied by the inverse to the fraction of the plot area located inside the population boundaries (Schreuder et al. 1993, p. 300 (tree concentric method)).

Validation of the adaptive composite estimator of density

The performance of the chosen COM estimator is validated in simulated SRS distance sampling with $k = 3,…,8$ and $n = 10, 30$, and 50 sample locations in nine populations of forest trees with known stem locations within a known rectangular area. A sample size of 10 is deemed relevant for stand-level inventories (Druckenbrod et al. 2005; Lessard et al. 1994; Lynch and Rusydi 1999; Payandeh and Ek 1986), while sample sizes of 30 and 50 are relevant for inventories of larger populations. The k-value has a large effect on bias of k-tree density estimators; it generally decreases as k increases (Kleinn and Vilčko 2006b; Kronenfeld 2009; Lessard et al. 1994; Magnussen et al. 2008a; Steinke and Hennenberg 2006). Our restriction to $k \leq 8$ is based on considerations regarding practical field implementation (Steinke and Hennenberg 2006).

Performance of the chosen COM and its component estimators are assessed in terms of bias (estimated density minus the actual density) and root mean squared error (RMSE, the square root of the sum of the Monte Carlo estimate of sampling variance and the square of the estimate of bias). Bias and RMSE are both expressed in percent of the actual density. We also assess estimates of precision (standard error) and coverage of sample-based 95% confidence intervals for the true density. For nearly unbiased estimates, it can be expected that the interval covers (includes) the true density at a rate of 0.95 in repeated sampling. All estimates of precision are based on the leave-one-out jackknife procedure (Cochran 1977, p. 178).

Each sampling design was repeated 1,000 times in each of the nine validation populations. With this number of replications, the Monte Carlo error (Koehler et al. 2009) for $n = 10$ on an estimate of a stem density in the range from 0.04 to 0.10 trees m^{-2} was less than 0.4%.

Validation populations

Nine forest tree populations with widely different densities and patterns of stem locations were used exclusively for validation purposes. Their names and main characteristics with respect to distance sampling are as follows:

RANDOM   Is a 300 m × 300 m (9.0 ha) area with 7,200 randomly located trees ($\lambda = 0.0800$ trees m^{-2}).

MATERN   Is a 300 m × 300 m (9.0 ha) area with clusters of trees. Each of 1,800 randomly located parent trees produces trees located at random within a circle with a random radius between 2 m and 8 m and centered on the location of the parent. The number (m) of offspring per parent follows a beta-binomial distribution (Garren et al. 2000) with parameters $(\alpha, \beta, m) = (5, 51, 90)$. The resulting expected value of 8.0 (± 4.3) offspring per parent yields an expected density of $\lambda = 0.1607$ *trees m^{-2}*. The actual density in the simulated point pattern was 0.1592. When sampling with fixed-area plots, the number of trees per plot will follow a negative binomial distribution (Eberhardt 1967; Picard et al. 2005).

UNIFORM   Is a 300 m × 300 m (9.0 ha) area with 2699 trees   ($\lambda = 0.0300$ trees m^{-2})   distributed at random according to a hard-core process (Bondesson and Fahlten 2003) that ensures a separation between any pair of trees of at least 4 m. The actual mean distance between nearest neighbors is 4.61 m.

BCI   Is a 1000 m × 500 m (50 ha) Smithsonian research plot on Barro Colorado Island. The area contains a rich old-growth wet tropical forest (Condit and Hubbell 1998; Hubbell et al. 1999). Data from the 1981–1983 census with (x, y) coordinates of 368 035 mapped trees of woody

species are used ($\lambda = 0.7361$ trees m^{-2}). Only trees with a minimum height of 1.3 m are included.

FUSHAN   Is a 500 m × 500 m (25 ha) Forest Dynamics Plot in a sub-montane evergreen broadleaf forest in northeastern Taiwan (established in 2004, Su et al. 2007). The plot holds 111 853 trees with a height >1.3 m and known locations ($\lambda = 0.4474$ trees m^{-2}).

MIOMBO   Is a 340 m × 420 m (14.28 ha) plot in sub-tropical savanna in the Miombo Woodlands of northern Zambia (Kleinn and Vilčko 2006b). Mapped locations are known for 5713 trees with a diameter of 7 cm or greater at breast height (1.3 m) ($\lambda = 0.0400$ trees m^{-2}).

TALL PINES (3)   Is a central 450 m × 600 m (27 ha) part of the 39.4 ha surveyed portion of the old-growth 80-ha Wade Tract pine savanna in Thomas County, Georgia, USA (Platt et al. 1988). The 1979 population of trees with a diameter at breast height (1.37 m) of 2 cm or greater was divided into larger ($\geq$66 cm diameter at 1.5 m) and smaller trees. The threshold of 66 cm is the maximum likelihood estimate of the diameter that separates old-growth trees from cohorts of younger trees. The population of larger trees, TP79A, had 1255 pines alive in 1979 ($\lambda = 0.0046$ trees m^{-2}). The population of smaller trees, TP79B, had 2231 mature and immature pines alive in 1979 ($\lambda = 0.0083$ trees m^{-2}) in an irregular pattern with higher
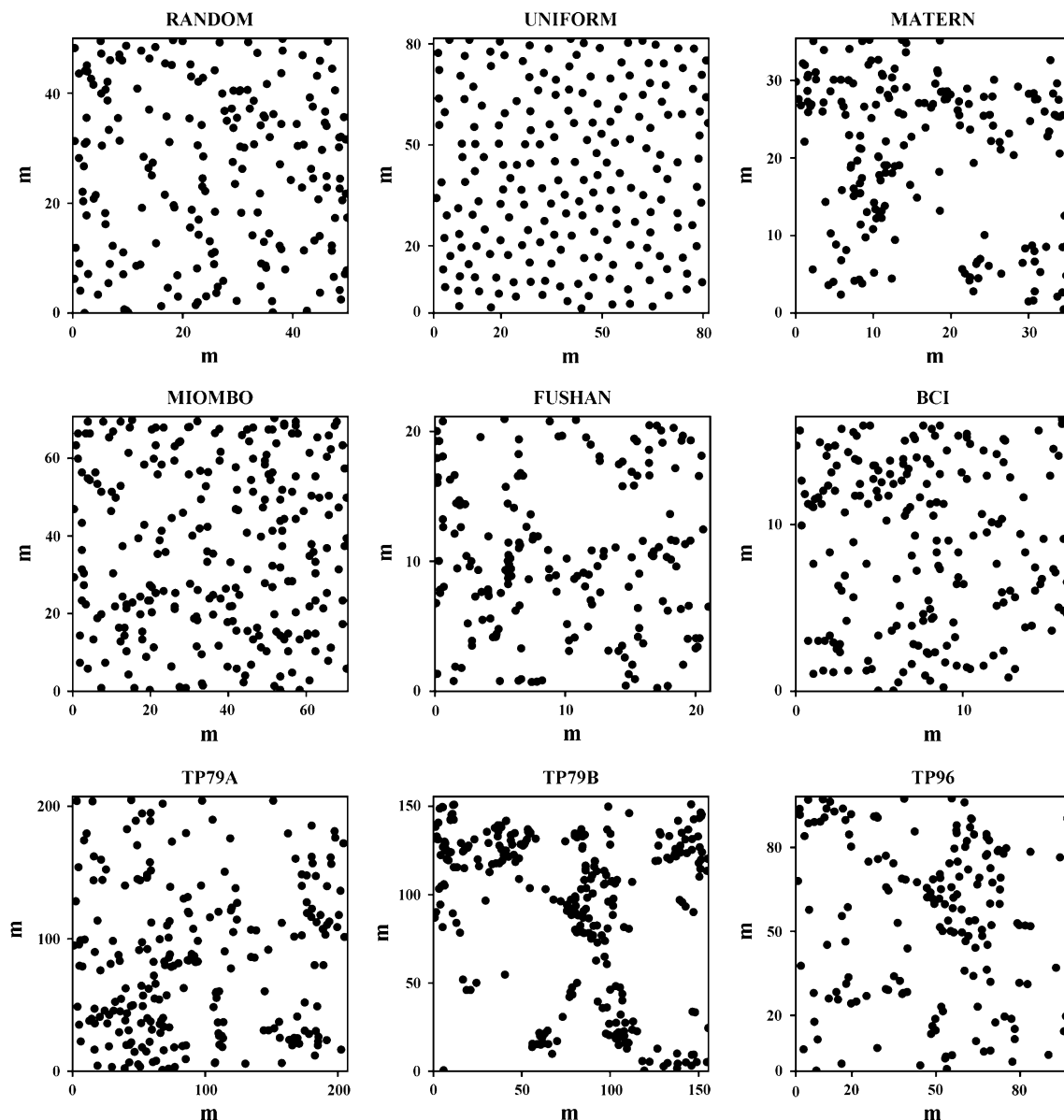

Fig. 1 Maps of tree locations in the nine validation populations. Each map represents the centermost part of each study area with approximately 200 trees

Fig. 2 Relative frequencies of trees per plot in fixed-area circular plots holding an average of 6 trees. Results are based on 6000 replications. Expectations under the assumption of complete spatial randomness are indicated by black dots
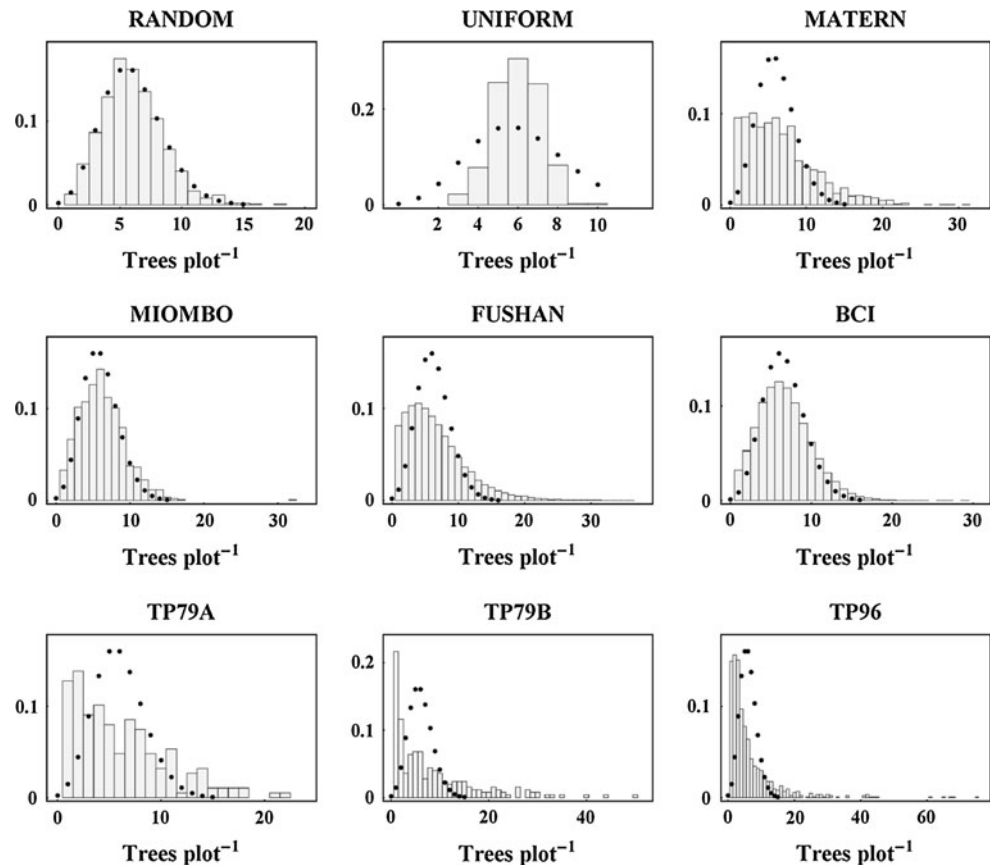

Fig. 3 Non-parametric kernel-based estimates of the probability density distribution of the distance (r_k) to the $k = 6$ nearest tree from a random location (results are based on 6000 distances). The distribution under the assumption of complete spatial randomness is indicated by a dashed line
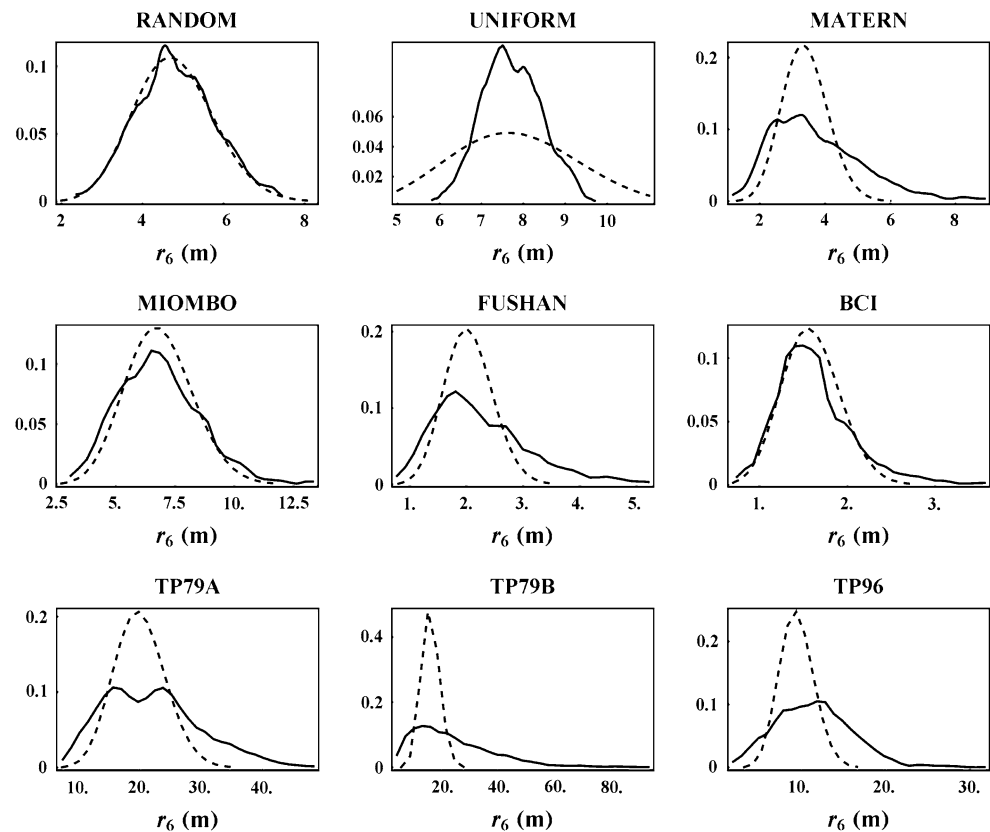
Fig. 4 Ripley's $L_4(h)$ function plotted against h (inter-point distance) for six validation sites (only a central square with approx. 200 trees is used for analysis)
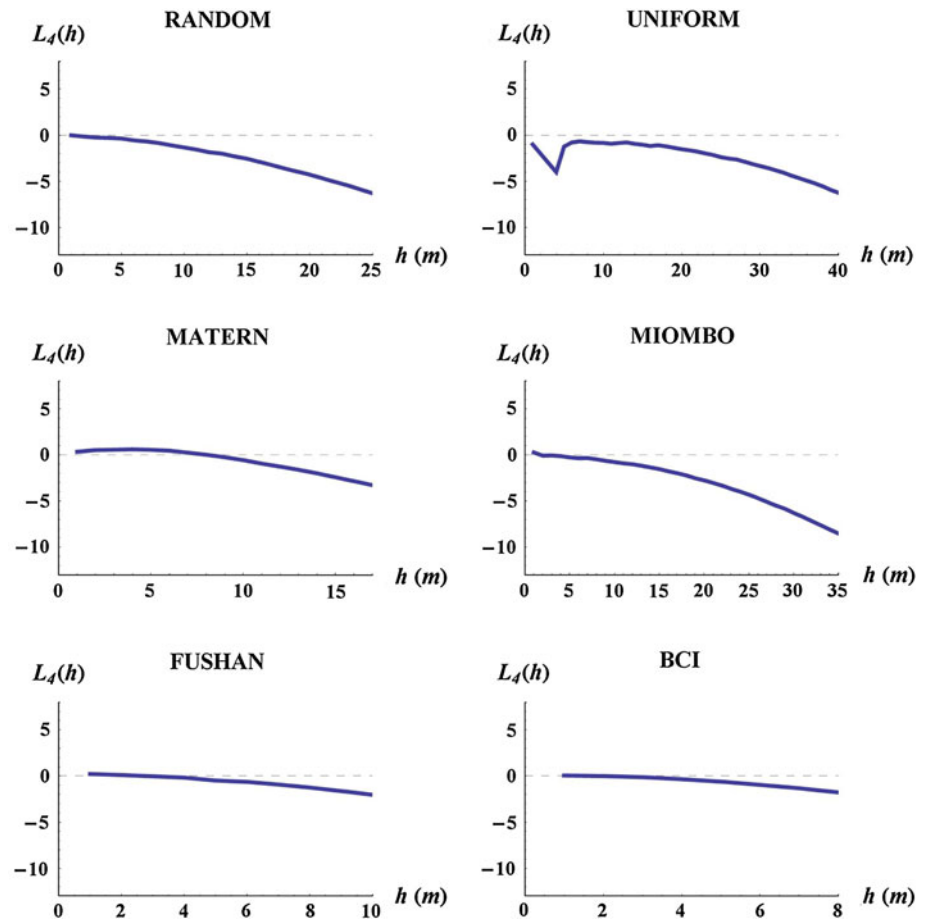
concentrations along old roads. The third population (TP96) is comprised of 5,706 pines alive in 1996, following a hurricane (1985) and a recruitment episode (mast year of 1987) ($\lambda = 0.0208$ trees m^{-2}).

The nine validation populations differed markedly in their spatial distribution of trees. Mapped locations of approximately 200 trees within a central portion of each study area captured essential features (Fig. 1). A summary of the variation in tree density within a population (site) is provided in Fig. 2. It shows the relative frequency distributions of tree counts in square fixed-area plots holding an average of $k = 6$ trees per plot. The distribution expected under the assumption of complete spatial randomness is also provided. For this plot size, a significant departure from *csr* (Pearson's Chi-squared test, Lloyd 1999) was noted in UNIFORM ($\hat{X}_9 = 22.7, P = 0.01$), MATERN ($\hat{X}_{12} = 135.0, P = 0.00$), FUSHAN ($\hat{X}_{16} = 1058.8, P = 0.00$), and TP96 ($\hat{X}_{11} = 130.5, P = 0.00$). The empirical distributions of the distance to the kth trees and the distributions expected under an assumption of *csr* (Cressie 1993, p. 637) are shown in Fig. 3. All but RANDOM had cumulative distribution functions that departed significantly ($D_{max} > 0.08, P < 0.01$) from the expectation under *csr* [test: Kolmogorov–Smirnov, Conover (1980, p. 436)].

Further insight about the spatial point patterns are provided by Ripley's L_4 function (Cressie 1993, p. 616). Figures 4 and 5 capture trends (derived for the central part of each population with approximately 200 trees) in the statistic $\hat{L}_4(h)^{0.5}\pi^{-0.5} - h$ where h is the inter-point distance. Under clustering (at distances less than h), this statistics is positive and vice versa under regularity.

Results

Across the 16 test populations and 18 design settings, the composite k-tree density estimator $w_{GP}\lambda_{GP} + w_{KV}\lambda_{KV} + w_{PE}\lambda_{PE}$—with adaptive weights determined by a multinomial logistic function of four ancillary variables—had the smallest average absolute (relative) bias (14.7%) among the 26 tested composite estimators. Runner-ups were two tied composites of (GP, PE and ORBIT) and (GP, PE, and MORISITA) each with a RMAB of 14.9%. A two-component estimator (GP and ORBIT) ranked fourth with an RMAB of 15.2%. Dropping GP out of a composite estimator resulted in an increase in RMAB of approximately 4%. Final estimates of the regression coefficients in
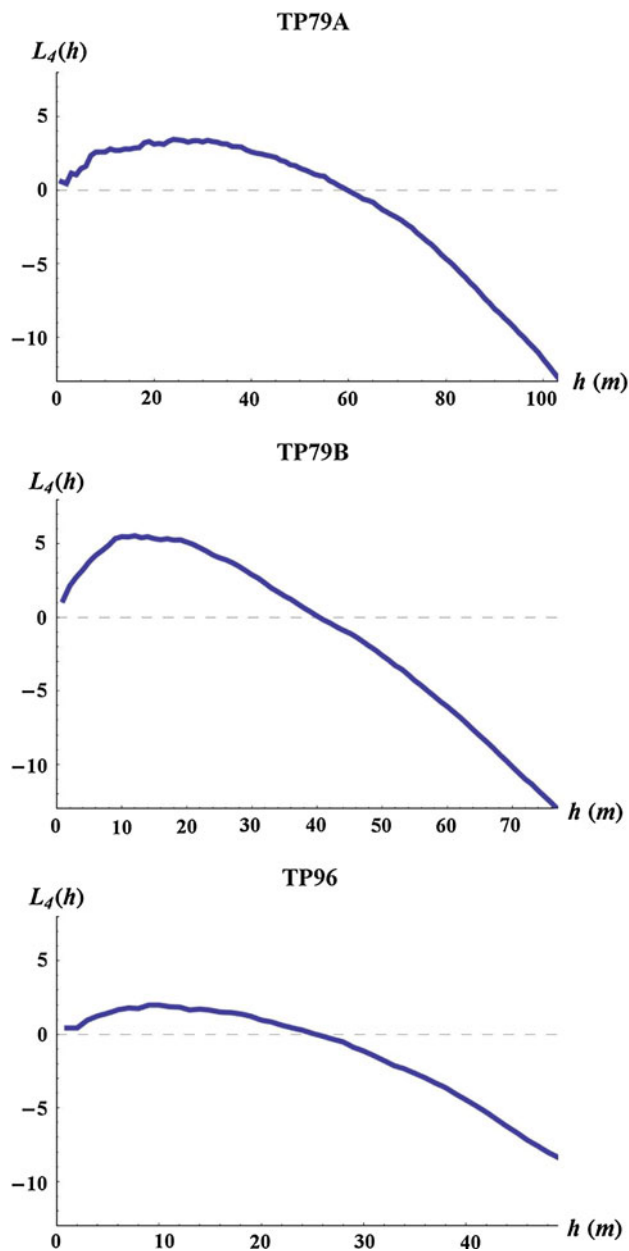
Fig. 5 Ripley's $L_4(h)$ function plotted against h (inter-point distance) for three Tall Pines sites (only a central square with approx. 200 trees is used for analysis)

the multinomial logistic function in (7) are in Table 1. All regressors for KV were statistically significant ($P < 0.05$). For PE the regressors for n, k, and MV were significant ($P < 0.05$). The weight $\hat{w}_{GP}$ increased slightly (0.01) for a unit increase in n and k, increased by 0.11 for a unit increase in MV, and declined by 0.07 for a unit increase in VR. These changes were, of course, offset by an opposite change in the sum of $\hat{w}_{KV}$ and $\hat{w}_{PE}$. Weights given to GP varied from 0.00 to 0.96 with a mean of 0.48. Corresponding figures for KV were 0.00–0.71 and 0.38, and for PE 0.02–1.00 and 0.21.

Table 2 lists the relative bias of the five candidate k-tree estimators in the nine validation populations. Relative bias of GP, KV, PE, and hence COM typically declined in absolute value with increasing k (Fig. 6). The exceptions are for COM in TP79A and GP in TP79B. For the unbiased FIX, our estimates of bias were always less than 0.05%.

In RANDOM, the bias of the proposed COM estimator is between -2% ($k = 8$) and 2% ($k = 4$). COM is less biased than any of its constituent estimators. The average weights (in RANDOM) given to KV, PE, and GP were 0.31, 0.48, and 0.23. In this population, the estimator by MORISITA would be unbiased.

In MATERN, the variance of observed distances was between 4 and 5 times larger than expected under *csr*. Over dispersion generally increased with k. The relative bias of COM was between 3% ($k = 4$) and 6% ($k = 6$) (Fig. 6). Bias of COM is similar to that of GP, but for $k = 6$, 8 about 1% greater than the bias of PE. The GP estimator received a weight that increased with k (from 0.75 to 0.86) while PE-weights declined accordingly (0.23–0.13).

In UNIFORM, the KV estimator is nearly unbiased (Table 2, Fig. 6). Its relative bias is consistently below 0.5%. Accordingly, the KV component in COM received the largest weights (mean 0.80), while PE received an average weight of 0.19 with the remaining 0.01 to the strongly biased GP estimator. The bias of COM is therefore intermediate to that of KV and PE but substantially closer to KV than to PE. The observed variance of r_k was only about one-fifth of the variance expected under *csr*.

Table 1 Maximum likelihood estimates of the regression coefficients in the multinomial logistic model (7) fitted to data obtained from simulated k-tree sampling in 16 test populations

Component	Dependent variable	Regressors (ancillary variables, **X**)				
		Constant	n	k	MV	VR
$\hat{\lambda}_{GP}$	$\hat{w}_{GP}$	0	0	0	0	0
$\hat{\lambda}_{KV}$	$\hat{w}_{KV}$	-0.79 (0.35)	-0.24 (0.04)	-0.55 (0.12)	2.31 (0.24)	-1.03 (0.26)
$\hat{\lambda}_{PE}$	$\hat{w}_{PE}$	-0.50 (0.35)	-0.15 (0.04)	-0.47 (0.13)	1.81 (0.27)	0.22 (0.18)

The multinomial logistic model predicts the weights given to k-tree density estimators GP, KV, and PE as a function of ancillary variables. Estimates of the standard error of a regression coefficient are in parentheses

Table 2 Monte Carlo estimates of bias (in percent of actual density) of estimators FIX, the five candidate *k*-tree estimators MORISITA, PE, GP, KV, PE, and the final choice of COM in nine validation populations

Estimator	k	RANDOM	MATERN	UNIFORM	BCI	FUSHAN	MIOMBO	TP79A	TP79B	TP96
FIX	4	0.1	−0.2	−0.2	0.1	−0.1	0.0	0.2	−0.2	−0.3
	6	0.1	−0.1	0.0	−0.2	0.1	0.1	0.0	0.1	0.0
	8	0.0	0.3	0.0	−0.1	0.1	0.0	0.0	0.0	0.1
MORISITA	4	−1.6	−21.6	2.5	−6.7	−22.6	−10.8	−27.5	−62.0	−38.3
	6	−3.4	−17.5	0.0	−5.8	−21.3	−8.5	−24.7	−57.4	−34.7
	8	−3.8	−14.8	−1.5	−5.8	−21.9	−8.2	−23.4	−53.3	−31.7
PE	4	9.4	−4.3	13.9	6.6	−3.3	8.1	−2.8	−33.0	−26.8
	6	6.2	−4.2	9.0	3.7	−5.2	5.4	−4.3	−30.4	−12.3
	8	4.1	−4.0	6.4	2.5	−6.2	4.3	−4.9	−25.0	−5.5
KV	4	−6.2	−30.8	0.2	−14.8	−28.4	−9.8	−25.4	−57.5	−40.7
	6	−3.9	−23.6	−0.4	−12.0	−24.4	−6.6	−20.3	−49.0	−35.4
	8	−2.9	−19.1	−0.3	−10.2	−22.2	−5.2	−17.9	−43.1	−31.6
ORBIT	4	−1.5	−18.7	−5.1	−7.0	−20.6	−6.0	−27.4	−60.5	−35.8
	6	−4.3	−17.5	−5.4	−8.0	−21.8	−6.7	−27.0	−59.4	−36.2
	8	−5.1	−15.6	−5.8	−7.4	−22.0	−6.9	−26.5	−57.0	−34.5
GP	4	12.9	4.3	22.9	4.1	1.7	14.9	5.8	2.1	−8.2
	6	10.9	6.5	8.0	3.0	2.6	13.8	6.8	10.5	−1.8
	8	9.3	8.0	8.1	2.7	2.8	9.8	6.5	17.0	2.9
COM	4	2.1	3.4	1.4	−2.0	−6.1	−1.4	−13.6	−13.6	−1.1
	6	−0.6	4.7	−0.6	−2.4	−5.0	−2.2	−11.8	−1.2	7.1
	8	−2.0	6.0	−1.8	−2.0	−6.1	−3.3	−11.4	6.0	12.7

Results are averages over sample sizes $n = 10$, 20, and 30

In MIOMBO, the bias of COM fluctuated between −3% ($k = 8$) and −1.4% ($k = 4$). COM was uniformly less biased than KV, PE, and GP (Table 2, Fig. 6). The KV component of COM received an average weight of approximately 0.5, while the remainder was split roughly equally between GP and PE.

In FUSHAN, the bias of COM fluctuates between −5% and −6%. Both GP and PE are less biased than COM for $k \leq 5$, but for $k \geq 6$, the bias of COM is close to that of PE. Thus, COM is a considerable improvement over KV, MORISITA, and ORBIT for all k. GP receives an increased weight as k increases (0.31 with $k = 3$, and 0.57 with $k = 8$) due to a parallel increase in VR (3.4–5.2). Bias results from BCI were, by and large, intermediate to those from MATERN and FUSHAN (Table 2, Fig. 6). COM bias remained relatively low (−2 to −3%).

In the Tall Pines, the performance of COM varied considerably among TP79A, TP79B, and TP96. In TP79A, the bias of COM is approximately halfway between the bias of KV (−25% for $k = 4$ and −18% for $k = 8$) and GP (6% for $k = 4$ and 7% for $k = 8$). Estimators PE or GP are both superior to COM for all tested k-values and would have been a better choice for small k-values (3 and 4). In TP96B, COM has the least bias of any k-tree estimator for $k \geq 6$ and is ranked 2nd to GP for smaller k-values. Estimators KV, MORISITA, ORBIT, and PE have large bias.

TP96 is another example of how COM adapts toward the best of the component estimators (GP with an average weight of 0.93) with a resulting performance distinctly better than possible with KV, PE, MORISITA, or ORBIT. The COM estimator is actually second to FIX for $k = 4$ and $k = 5$.

Relative root mean squared errors (RRMSE) are in Table 3 for $n = 10$ and in Fig. 7 for $n = 20$. As expected—given the optimization criterion applied to the composite estimator—we see that RRMSE of COM is less variable across site than the RRMSEs of its constituent components (GP, KV, PE). For $n = 10$ (Table 3), the RRMSE for COM is within 2.5 percentage points from those for FIX in sites RANDOM, UNIFORM, BCI, FUSHAN, and MIOMBO. The five candidate estimators achieve RRMSE values approximately at par with those for FIX in at most one or two sites. For $n = 20$ (Fig. 7), RRMSE values of COM are, with the exception of results from TP96, again less variable across sites than seen in RRMSE values for GP, KV, and PE. FIX was generally superior for all k-values but for $k \geq 5$ COM is a close second in RANDOM, UNIFORM, and BCI.

Monte Carlo estimates of precision (standard error) for COM were, on average, about 13% above corresponding FIX estimates (from +4% in FUSHAN to +21% in MATERN). As a rule, the difference decreased with increasing n and k.
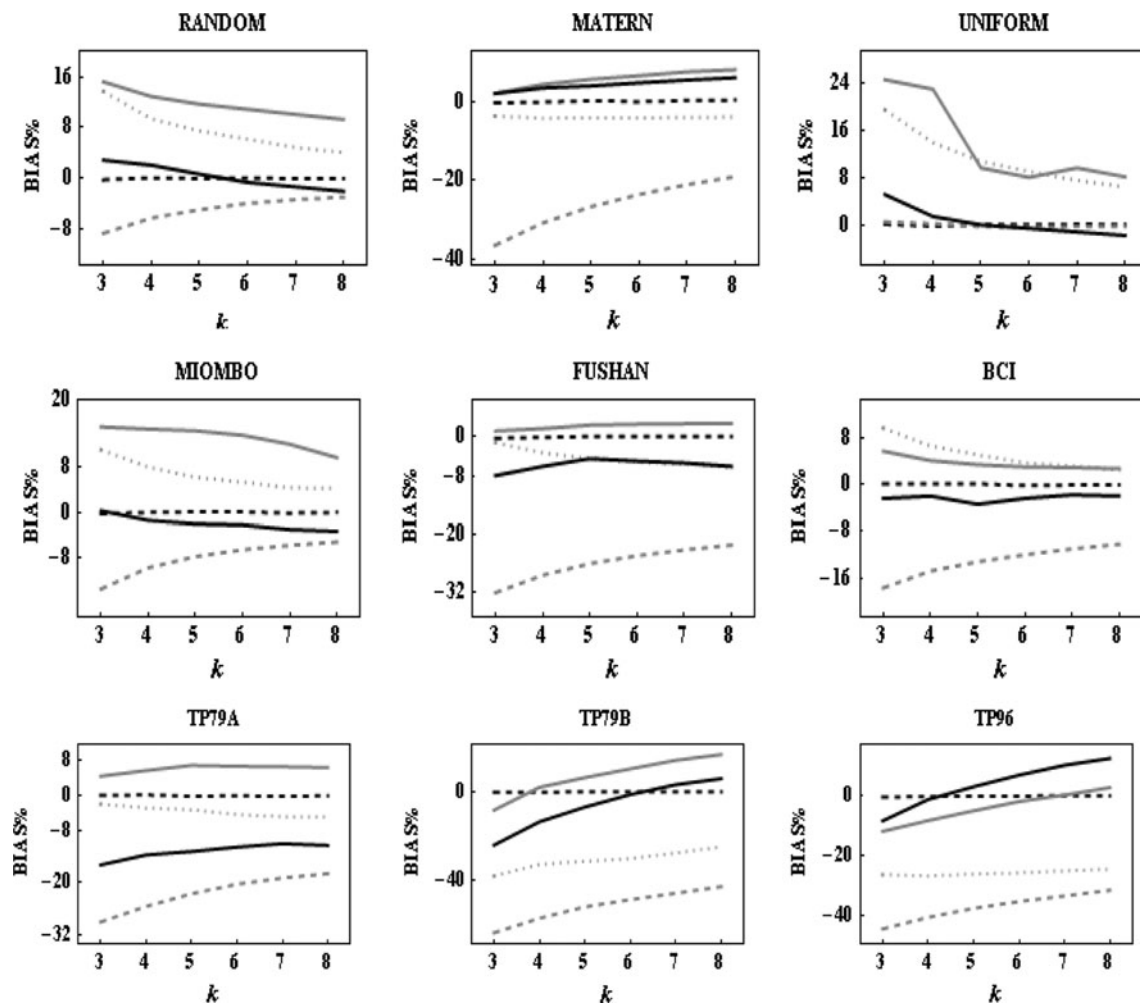
Fig. 6 Trends in estimates of relative bias (BIAS%) for $k = 3,\ldots, 8$ of estimators FIX (*black-dashed*), COM (*black-full*), GP (*gray-full*), KV (*gray-dashed*), and PE (*gray-dotted*). Results are averages over sample sizes of 10, 20, and 30 and 1000 replications

Jackknife estimates of precision for COM estimates of density underestimated, as a rule, the variability in the Monte Carlo results. The underestimation varied between 16 and -29%. Computed confidence intervals were therefore too short with an average under-coverage of 7% (Table 4). Under-coverage of KV-based and GP-based intervals was in most cases greater. As expected, the jackknife procedure generally failed to give realistic estimates of precision and confidence intervals for PE estimates of density (Efron and Tibshirani 1993, p. 148).

Discussion and conclusions

A long recognized problem for k-tree sampling is biased estimators of density and resulting bias in all per unit area estimates of forest attributes (Steinke and Hennenberg 2006). The problem has been resistant to a practical solution, however, at least for practically relevant k-values (>4) and sample sizes of 5–20 typical in stand-level forest

inventories (Kleinn and Vilčko 2006b; Lynch and Rusydi 1999). For larger sample sizes (>30), the RMSEs of the GP estimator proposed by Magnussen et al. (2008a, b) were just 2% higher than for sampling with fixed-area plots. As far as we can judge, our COM estimator is superior to the asymptotic estimator proposed by Kronenfeld (2009), at least for $k \leq 8$. A recent study suggests that a k-tree density estimator (RDE) based on a stochastic simulation of tree locations that adhere to the distribution of the $n \times k \times (k - 1)$ observed distances affords a nearly ($<2\%$) unbiased estimation of stem density (Nothdurft et al. 2010). A direct comparison between the bias of COM and RDE is not possible, but the results indicate a better performance of RDE. A comparison of root mean squared errors is less conclusive. A cautious interpretation suggests a tie. The need for intensive computer simulations is currently a disadvantage for RDE, at least in terms of practical implementation.

The proposed k-tree density estimator (COM) appears to perform well in terms of bias and RMSE when k is at least

Table 3 Monte Carlo estimates of RRMSE of estimators FIX, the five candidate *k*-tree estimators MORISITA, PE, KV, ORBIT, and GP, and the final choice of COM

Estimator	k	RANDOM	MATERN	UNIFORM	BCI	FUSHAN	MIOMBO	TP79A	TP79B	TP96
FIX	4	15.6	27.6	7.8	18.4	26.3	18.4	25.2	44.7	49.4
	6	13.1	25.0	6.9	17.4	24.5	15.4	22.7	40.4	45.0
	8	11.5	23.1	5.0	15.1	23.2	13.6	21.1	37.0	42.1
MORISITA	4	11.4	28.7	7.1	14.6	27.4	19.9	31.5	63.6	41.4
	6	11.8	25.7	6.6	14,2	26.3	18.7	29.1	59.1	38.1
	8	11.9	23.9	6.8	14.3	26.5	18.6	27.9	55.2	35.4
PE	4	23.8	33.9	17.8	24.5	31.9	26.1	34.6	49.1	36.4
	6	20.3	29.4	12.2	21.9	28.0	19.8	29.9	46.7	35.1
	8	16.3	26.1	9.2	19.4	25.2	17.0	27.5	43.8	32.5
KV	4	17.0	36.3	7.9	22.8	34.7	20.5	31.5	58.7	42.8
	6	13.4	29.9	5.9	19.6	30.7	16.4	27.3	51.3	38.0
	8	11.5	26.0	4.9	17.6	28.7	14.1	25.0	46.2	47.9
ORBIT	4	13.5	26.0	9.5	16.6	26.4	17.7	32.0	62.4	39.8
	6	14.1	25.1	9.6	17.0	27.3	17.9	31.6	61.3	40.1
	8	14.3	23.8	9.8	16.7	29.0	18.1	31.3	59.0	38.6
GP	4	24.6	35.0	24.9	22.5	31.7	30.2	33.9	55.5	39.3
	6	19.9	30.2	11.4	18.6	28.1	24.1	30.3	56.5	44.0
	8	17.4	28.2	9.6	16.6	26.0	20.2	26.9	56.3	47.9
COM	4	17.5	36.5	8.2	18.5	27.4	20.5	31.8	54.2	52.3
	6	13.8	27.8	5.7	18.3	23.1	18.7	29.2	53.7	53.3
	8	12.1	28.3	5.4	15.8	21.8	15.6	26.5	54.1	61.0

Results are averages across 1,000 replications for nine validation populations and a sample size $n = 10$

5, and density is above 100 trees ha^{-1}. In clustered, low-density populations, the performance appears too variable to be useful in practice. However, in clustered savanna-like populations, the incentive to use distance sampling is a priori not strong. The advantage of COM is a more stable performance across populations than is possible with a conventional *k*-tree density estimator. The traditional approach to create a robust estimator is to find an optimal (fixed) linear combination of existing estimators (Byth 1982; Clayton and Cox 1986; Cox 1976; Delince 1986). Our explorations in 16 test populations clearly suggested that the optimal weight given to a specific component estimator changed from one population to the next. To make the composition of COM adapt to the sample population, we use component weights obtained from a multinomial logistic function of two design variables (*n*, *k*) and two indicators of over dispersion in the sampled distance data. We recognize that indicators of over dispersion may be poorly estimated from small samples. A key to the (relative) success of COM is that the GP and KV often have a bias in opposite directions. The combined contributions of GP and KV to COM dominate the performance of COM.

A composite estimator is by design a compromise. As such, it occasionally performs worse than one of its component estimators and possibly one of the many non-included distance sampling density estimators. A composite estimator is therefore intended for application across a fairly wide range of forest conditions rather than a single well defined spatial pattern. In the latter case, an estimator that is optimal for the particular situation would be more appropriate. For example, if it is known that the spatial pattern is nearly random, we would not hesitate to recommend MORISITA's estimator (Morisita 1954). In populations with a very regular inter-tree spacing (as in plantations and thinned stands), the KV estimator would be an excellent choice. The ancillary variable *VR*, or rather its logarithm, provides a good indicator of the expected performance (absolute bias) of GP. As log(*VR*) increases, the bias of GP declines at a rate of 0.07 ($P < 0.01$), whereas the bias of KV is increasing at a rate of 0.07 ($P < 0.01$). A decision to use, for example, one of the estimators GP, KV, and PE based solely on observed values of *VR* and *MV* could be marred by a large sampling error in these variables (Cressie 1993, p. 613). In applications with $k \geq 5$, a closed-canopy stand, and an unknown spatial pattern of tree locations, the COM estimator can produce results with an acceptable level of (expected) bias ($\leq 6\%$) and an RMSE not too much higher than achievable with fixed-area plot sampling.
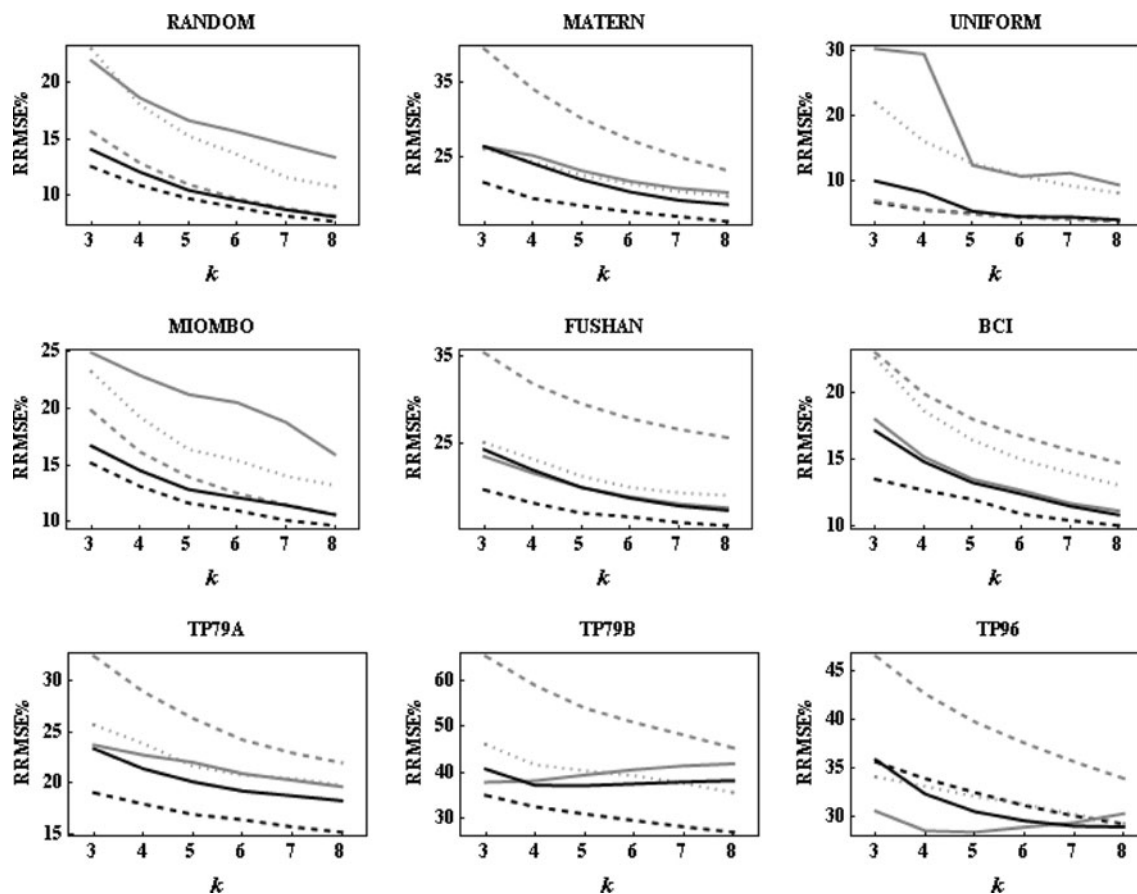
Fig. 7 Trends in relative root mean squared error (RRMSE%) for $k = 3,\ldots, 8$, and $n = 20$ of estimators FIX (*black-dashed*), COM (*black-full*), GP (*gray-full*), KV (*gray-dashed*), and PE (*gray-dotted*). Results are averages over 1000 replications

Table 4 Coverage rates of 95% confidence intervals achieved with estimators COM, KV, GP, and FIX ($n = 10$, $k = 6$)

Population	COM	KV	GP	FIX
RANDOM	0.85	0.87	0.80	0.90
MATERN	0.86	0.70	0.85	0.90
UNIFORM	0.86	0.87	0.74	0.88
BCI	0.87	0.81	0.87	0.89
FUSHAN	0.85	0.71	0.88	0.88
MIOMBO	0.88	0.86	0.84	0.86
TP79A	0.86	0.77	0.86	0.87
TP79B	0.89	0.25	0.87	0.88
TP96	0.86	0.42	0.89	0.89

Our chosen jackknife estimator of sampling variance will need further improvements to address the negative bias and persistent under-coverage of 95% confidence intervals. Computationally, the jackknife approach is simple compared to, for example, the delta technique (Särndal et al. 1992, p. 172). We surmise that the jackknife approach underestimates the correlations between the n jackknife pseudo values of $\hat{\lambda}_{GP}$, $\hat{\lambda}_{KV}$, and $\hat{\lambda}_{PE}$. However, we do not know how to address this problem correctly. A bootstrap approach would be expensive due to the excessive time it would require to complete a large number of estimations of $\hat{\lambda}_{GP}$.

The inherent promise of k-tree sampling remains attractive and warrants a continued effort toward elimination of the bias problem. The new COM estimator is but one small step toward this goal. Inspired by the stochastic resampling approach proposed by Nothdurft et al. (2010), we have begun to investigate the performance of an estimator in which n distance differences between the $(k - 1)$th and kth nearest tree are permutated at random in order to generate B replications of n 'synthetic' distances to the kth tree. Preliminary results are encouraging and suggest that future progress may well lie in the direction of resampling-based estimators.

References

Augustin NH, Musio M, von Wilpert K, Kublin E, Wood SN, Schumacher M (2009) Modeling spatiotemporal forest health monitoring data. J Am Stat Assoc 104:899–911. doi:10.1198/jasa.2009.ap07058

Barabesi L (2001) A design-based approach to the estimation of plant density using point-to-plant sampling. J Agric Biol Environ Stat 6:89–98

Barabesi L, Marcheselli M (2004) Design-based ranked set sampling using auxiliary variables. Environ Ecol Stat 11:415–430

Bondesson L, Fahlten J (2003) Mean and variance of vacancy for hard-core disc processes and applications. Scand J Stat 30:797–816

Byth K (1982) On robust distance-based intensity estimators. Biometrics 38:127–135

Clayton G, Cox TF (1986) Some robust density estimators for spatial point processes. Biometrics 42:753–767

Cochran WG (1977) Sampling techniques. Wiley, New York

Condit R, Hubbell SP (1998) Tropical forest census plots: methods and results from Barro Colorado island, panama and a comparison with other plots. Springer, Berlin

Conover WJ (1980) Practical nonparametric statistics. Wiley, New York

Cox TF (1976) The robust estimation of the density of a forest stand using a new conditioned distance method. Biometrika 63:493–499

Cox DD, Cox LH, Ensor KB (1997) Spatial sampling and the environment: some issues and directions. Environ Ecol Stat 4:219–233

Cressie NAC (1993) Statistics for spatial data, Revised edition 2 edn. Wiley, New York

Delince J (1986) Robust density estimation through distance measurements. Ecology 67:1576–1581

Druckenbrod DL, Shugart HH, Davies I (2005) Spatial pattern and process in forest stands within the Virginia piedmont. J Veg Sci 16:37–48

Eberhardt LL (1967) Some developments in 'distance sampling'. Biometrics 27:207–216

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, Boca Raton

Engeman RM, Sugihara RT, Dusenberry WE (1994) A comparison of plotless density estimators using Monte Carlo simulation. Ecology 75:1769–1779

Essed FE (1956) A quick, simple and at the same time accurate method for estimating the total volume and the increment per cent of even-aged stands. In Forester 82:261–263

Ferretti M (1997) Forest health assessment and monitoring–issues for consideration. Environ Monit Assess 48:45–72

Garren ST, Smith RL, Piegorsch WW (2000) On a likelihood-based goodness-of-fit test of the beta-binomial model. Biometrics 56:947–949

Greene WH (2003) Econometric analysis, 5th edn. Prentice Hall, Upper Saddle River

Hubbell SP, Foster RB, O'Brien ST, Harms KE, Condit R, Wechsler B, Wright SJ, SLd Lao (1999) Light gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. Science 283:554–557

Jonsson B, Holm S, Kallur H (1992) A forest inventory method based on density-adapted circular plot size. Scand J For Res 7:405–421

Kendall MG, Moran PP (1963) Geometrical probability. Charles Griffin, London

Kint V, De Wulf R, Noël L (2004) Evaluation of sampling methods for the estimation of structural indices in forest stands. Ecol Model 180:461–476

Kleinn C, Vilčko F (2006a) Design-unbiased estimation for point-to-tree distance sampling. Can J For Res 36:1407–1414

Kleinn C, Vilčko F (2006b) A new empirical approach for estimation in k-tree sampling. For Ecol Manage 237:522–533

Koehler E, Brown E, Haneuse J-PA (2009) On the assessment of Monte Carlo error in simulation-based statistical analyses. Am Stat 63:155–162

Kronenfeld BJ (2009) A plotless density estimator based on the asymptotic limit of ordered distance estimation values. For Sci 55:283–292

Lessard VC, Reed D, Monkevich N (1994) Comparing n-tree distance sampling with point and plot sampling in northern Michigan forest types. N J Appl For 11:12–16

Lessard VC, Drummer TD, Reed DDA (2002) Precision of density estimates from fixed-radius plots compared to n-tree distance sampling. For Sci 48:1–6

Lewis SM (1975) Robust estimation of density for a two-dimensional point-process. Biometrika 62:519–521

Lloyd CJ (1999) Analysis of categorical variables. Wiley, New York

Lynch TB, Rusydi R (1999) Distance sampling for forest inventory in Indonesian teak plantations. For Ecol Manage 113:215–221

Lynch TB, Wittver RF (2003) n-Tree distance sampling for per-tree estimates with application to unequal-sized cluster sampling of increment core data. Can J For Res 33:1189–1195

Magnussen S, Kleinn C, Picard N (2008a) Two new density estimators for distance sampling. Eur J For Res 127:213–224

Magnussen S, Picard N, Kleinn C (2008b) A Gamma-Poisson distribution of point to k nearest event distance. For Sci 54:429–441

Morisita M (1954) Estimation of population density by spacing method. Contributions from the department of biology, faculty of science. Mem Fac Sci Kyushu Univ Ser E 1:187–197

Nothdurft A, Saborowski J, Nuske RS, Stoyan D (2010) Density estimation based on k-tree sampling and point pattern reconstruction. Can J For Res 40:953–967

Parker KR (1979) Density estimation by variable area transect. J Wildl Manage 43:484–492

Patil SA, Kovner JL, Burnham KP (1982) Optimum nonparametric estimation of population density based on ordered distances. Biometrics 35:597–604

Payandeh B, Ek AR (1986) Distance methods and density estimators. Can J For Res 16:918–924

Peña D (1997) Combining information in statistical modeling. Am Stat 51:326–332

Persson O (1964) Distance methods: the use of distance measurements in the estimation of seedling density and open space frequency. Studia Forestalia Suecica 15:1–68

Picard N, Kouyaté AM, Dessard H (2005) Tree density estimations using a distance method in Mali Savanna. For Sci 31:7–18

Platt WJ, Evans GW, Rathbun SL (1988) The population dynamics of a long-lived conifer (*Pinus palustris*). Am Nat 131:491–525

Pollard JH (1971) On distance estimators of density in randomly distributed forests. Biometrics 27:991–1002

Pollard JH, Palka D, Buckland ST (2002) Adaptive line transect sampling. Biometrics 58:862–870

Prodan M (1968) Punktstichprobe für die Forsteinrichtung (in German). Forst- und Holzwirt 23:225–226

Sandrine R, Philippe S, Michel A (2003) Applying distance-sampling methods to spotlight counts of red foxes. J Appl Ecol 40:32–43

Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York

Schreuder HT (2004) Sampling using a fixed number of trees per plot. USDA Forest Service, CO RMRSFC, RMRS-RN-17

Schreuder HT, Gregoire TG, Wood GB (1993) Sampling methods for multiresource forest inventory. Wiley, New York

Sheil D, Ducey MJ, Sidiyasa K, Samsoedin I (2003) A new type of sample unit for the efficient assessment of diverse tree communities in complex forest landscapes. J Trop For Sci 15:117–135

Steinke I, Hennenberg KJ (2006) On the power of plotless density estimators for statistical comparisons of plant populations. Can J Bot 84:421–433

Su SH, Chang-Yang CH, Lu CL, Tsui CC, Lin TT, Lin CL, Chiou WL, Kuan LH, Chen ZS, Hsieh CF (2007) Fushan subtropical forest dynamics plot: tree species characteristics and distribution patterns. Taipei

Wolfinger R, O'Connell MA (1993) Generalized linear mixed models: a pseudo-likelihood approach. J Stat Comput Simul 48:233–243

Zhang L-C, Chambers RL (2004) Small area estimates for cross classifications. J Roy Stat Soc Ser B 66:479–496