Scandinavian Journal of Forest Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/sfor20

The k-nearest neighbor technique with local linear regression

Steen Magnussen[a] & Erkki Tomppo[b]

[a] Canadian Forest Service, Natural Resources Canada, 506 West Burnside Road, Victoria, BC V8Z 1M5, Canada

[b] The Finnish Forest Research Institute, PO Box 18 (Jokiniemenkuja 1), Vantaa FI-01301, Finland

Published online: 28 Jan 2014.

To cite this article: Steen Magnussen & Erkki Tomppo (2014) The k-nearest neighbor technique with local linear regression, Scandinavian Journal of Forest Research, 29:2, 120-131, DOI: 10.1080/02827581.2013.878744

To link to this article: http://dx.doi.org/10.1080/02827581.2013.878744

Scandinavian Journal of Forest Research, 2014
Vol. 29, No. 2, 120–131, http://dx.doi.org/10.1080/02827581.2013.878744

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

The *k*-nearest neighbor technique with local linear regression

Steen Magnussen[a*] and Erkki Tomppo[b]

[a]*Canadian Forest Service, Natural Resources Canada, 506 West Burnside Road, Victoria, BC V8Z 1M5, Canada;* [b]*The Finnish Forest Research Institute, PO Box 18 (Jokiniemenkuja 1), Vantaa FI-01301, Finland*

(*Received 27 June 2013; accepted 19 December 2013*)

In a standard *k*-nearest neighbor (*k*NN) technique, imputations of unit-level values in the variables of interest (**Y**) are based on the *k*-nearest neighbors in a set of reference units. Nearest is defined with respect to a distance metric in the space of auxiliary variables (**X**). This study evaluates *k*NN imputations of **Y** with a selection, by the same distance metric, of *k*-nearest locally weighted regression models. Imputations are obtained as predictions using the **X** values of the *k*-nearest neighbors in the population. In simulated random sampling from three artificial multivariate populations and two actual univariate populations and sampling units composed of a single population element or a cluster of four elements, the new *k*NN technique: (1) improved the correlation between an imputation and its actual value; (2) lowered the root mean square error (RMSE) of imputations; (3) increased the slope in regressions of actual y values regressed against their imputed values; (4) performed relatively best with k values of 4 and sample sizes of 200 or greater; (5) compared favorably with a recently proposed *k*NN calibration procedure; and (6) had a higher (15–28%) RMSE than with a simple local linear regression. Distribution matching had a consistent negative effect (+10%) on RMSE.

Keywords: bandwidth; bias; distribution matching; forest inventory; imputation; weighted least squares regression

Introduction

Maps of forest resources are important information tools for management purposes and decision-making in a context of sustainable forestry (Corona et al. 2002; Hyde et al. 2006). The *k*-nearest neighbor technique (*k*NN) has become a popular and easy-to-implement method for multivariate mapping (Chirici et al. 2012). In *k*NN, the values of one or more target variables (**Y**) are imputed for elements (e.g. pixels with a regular geometric shape and known area) in a finite population without a direct observation of **Y** (Paass 1985; Aha 1997). Imputation of **Y** in population units with missing **Y** values (i.e. $\tilde{\mathbf{Y}}_{knn}$) is based on a set of selected auxiliary variables (**X**) known for all (N) elements in the population and correlated with **Y**. In a typical *k*NN application, a sample of n elements provides paired observations of **X** and **Y**.

The sample of n elements is referred to as the reference set, while the N-n elements with no observation of **Y** are referred to as the target set (Tomppo 1991). The imputed **Y** value for an element in the target set is a fixed known function (f) of the k **Y** values in the reference set whose associated **X** values are closest – in terms of a selected distance metric – to the **X** values of the element to receive an imputation. The analyst chooses **X**, f, k, and the distance metric, usually through a combination of cross-validation procedures and ranking of goodness-of-fit statistics (McRoberts 2009).

In forestry, the *k*NN technique has appeal (Maltamo & Kangas 1998; Holmström & Fransson 2003; Maselli et al. 2005; LeMay et al. 2008; Breidenbach et al. 2010) due to (1) readily available, low-cost, remotely sensed auxiliary variables correlated with **Y**; (2) challenges encountered with alternative parametric and semi-parametric multivariate modeling approaches (Koistinen et al. 2008) due to locally varying relationships between **X** and **Y** in response to variation in species, age, forest structure, soil, and climate (Zhang & Shi 2004; Opsomer et al. 2008; McRoberts et al. 2010) that are often easier to master with the *k*NN technique; (3) flexibility in generating forest attributes ranging from tree level to landscape levels, for example, tree lists (Temesgen et al. 2003), snag lists (Eskelson, Temesgen, Lemay et al. 2009), cavity tree abundance (Temesgen et al. 2008); (4) ease of integration with inventory data with a panel structure (Eskelson, Temesgen, Barrett 2009); and (5) provision of an imputation of **Y** for each population element suited for mapping and small-area estimation problems (Tomppo 2006). The latter is likely a main reason for the appeal. For the estimation of a population total from a large sample, alternative design-based estimators (Mandallaz 2013) may be equally efficient (Haara & Kangas 2012).

In *k*NN applications, the analyst will typically choose **X**, f, k, and the distance metric with an objective of minimizing the uncertainty of an imputation. However, apart from the choice of k, the task has proven

*Corresponding author. Email: steen.magnussen@nrcan.gc.ca

difficult (Katila & Tomppo 2001; McRoberts 2009). The *curse of dimensionality* (Scott 1992, p. 27) is the main obstacle toward real progress because distances in **X** space from a target element to the n reference elements become increasingly similar as the dimension of **X** grows. Conversely, the number of reference elements with a similar distance to a target element grows exponentially with the dimension of **X**. Thus in practice f is often the operator that generates a simple average of the k-selected reference values. The distance metric is typically Euclidian (applied to standardized **X** data) or a Mahalanobis distance (Rencher 1995, p. 87).

The choice of k determines, to a large degree, the variance in $\tilde{\mathbf{Y}}_{k\mathrm{nn}}$. As k increases, the variance decreases. Hence a forest map of $\tilde{\mathbf{Y}}_{k\mathrm{nn}}$ with, say, $k = 12$, may only display a fraction of the variance seen in the reference values of **Y** (Franco-Lopez et al. 2001; McRoberts et al. 2007; Räty & Kangas 2012). This phenomenon is clearly a detractor, since the variation in a mapped attribute provides important information to managers. Several analogous methods have been proposed to rescale $\tilde{\mathbf{Y}}_{k\mathrm{nn}}$ in a way that recovers the variance in the observed sample $\mathbf{Y}_{\mathrm{s}}$ (Lister & Lister 2006). Baffetta et al. (2012) demonstrated the method known as distribution matching (DM). Individual imputations are modified to match an (unbiased) estimate of the cumulative distribution function fitted to the observed sample $\mathbf{Y}_{\mathrm{s}}$. The authors demonstrated that DM preserved the statistical properties of a kNN estimate of a population total (or average). The DM procedure is order preserving (Stern 1990) and conserves the correlation between **Y** and $\tilde{\mathbf{Y}}_{k\mathrm{nn}}$.

A calibration of kNN imputations $\left(\tilde{\mathbf{Y}}^{\mathrm{cal}}_{k\mathrm{nn}}\right)$ with a global calibration function has been shown to improve the correlation between **Y** and a kNN imputation (Magnussen, Tomppo et al. 2010) and to restore the variance toward that of $\mathbf{Y}_{\mathrm{s}}$. The restoration of variance is, however, not as efficient as with a DM. Thus a DM applied to $\tilde{\mathbf{Y}}^{\mathrm{cal}}_{k\mathrm{nn}}$ should restore the variance to the level in $\mathbf{Y}_{\mathrm{s}}$ without compromising the benefits of a calibration.

With the objective of improving the calibration method by Magnussen, Tomppo et al. (2010) – without sacrificing simplicity – this study proposes a new kNN variant $\left(\tilde{\mathbf{Y}}^{\mathrm{lr}}_{k\mathrm{nn}}\right)$ with imputations computed as the average of k predictions generated from k-weighted local linear regression models selected from a set of n reference models. Selection of the k-weighted regression models is based on the same distance metric as in a standard kNN. It is hypothesized that locally weighted linear regressions (Cleveland et al. 1988) can correct for extrapolation bias more efficiently than a global calibration function. A DM applied to $\tilde{\mathbf{Y}}^{\mathrm{lr}}_{k\mathrm{nn}}$ is also expected to achieve a restoration of variance to the level in $\mathbf{Y}_{\mathrm{s}}$ without compromising any advantage that the proposed method may have.

The proposed new kNN estimator is evaluated in simulated sampling from three artificial complex multivariate populations of size $N = 8000$ and two univariate populations with actual inventory data.

Material and methods

Notation

Symbols and their definitions are listed in Appendix 1.

Population and sampling methods

A finite population U composed of N equal-area spatial elements (e.g. pixels) $(U = \{U_1, ..., U_N\})$ is considered with the objective of estimating the population average (μ) of one or more target variables $(\mathbf{Y})$ from a without-replacement equal probability sample (s) of size n. A set of p auxiliary variables $(\mathbf{X})$ is known for every element in the population. The auxiliary variables have been selected on grounds of their ability to predict **Y**. The sample of population elements (U_{s}) is obtained by simple random sampling of either n single elements (SRS) or n compact clusters of m elements each (CLU). Thus each population contains M clusters of size m so that $N = m \times M$, and each sample is composed of $n \times m$ elements, with $m \equiv 1$ in SRS and $m \equiv 4$ in CLU.

Throughout notation is for a univariate Y; extension to a multivariate case requires no new theory.

The standard k*NN estimator*

The standard kNN estimator of Y in the ith population element (Haara et al. 1997) can be written succinctly as

$$\tilde{y}_i^{k,st} = \sum\nolimits_{j\in\Gamma_k^{U_s}(i)} w_{ij}y_j,\ i = 1,\ldots,N \tag{1}$$

where summation in (1) is over the ordered set $\Gamma_k^{U_{\mathrm{s}}}(i)$ of k elements in U_{s} with auxiliary variable values closest to $\mathbf{X}_i$, and w_{ij} is the weight given to the y value y_j of the jth selected reference element $\left(\sum_{j\in\Gamma_k^{U_s}(i)} w_{ij} = 1\right)$. Euclidean distances in **X**-space were used as the criterion for the selection of the k-nearest neighbors. The set $\Gamma_i^{U_{\mathrm{s}}}$ of nearest neighbors is ordered by increasing distance to $\mathbf{X}_i$.

A standardized Euclidean distance metric is used throughout, that is, the **X** variables have been standardized to a mean of zero and a variance of one for the computation of distance. This is also the metric used by the "Euclidean" option in the R-package "yaImpute" (Crookston & Finley 2008). In practice, the weights may be a function of distance that optimizes precision (McRoberts 2009). Here $w_{ij} = k^{-1}$.

The standard kNN estimator of the population mean of Y is denoted $\tilde{\mu}_y^{k,\mathrm{st}}$, and it is computed as the population total of $\hat{\tilde{y}}_i^{k,\mathrm{st}}$ divided by N. The expected value of $\tilde{\mu}_y^{k,\mathrm{st}}$ over all possible samples of size n is invariant to the sampling design (Baffetta et al. 2009).

The calibrated* k*NN estimator

The calibrated estimator proposed by Magnussen, Tomppo et al. (2010) is

$$\tilde{y}_i^{k,\,\mathrm{cal}} = \sum\nolimits_{j\in\Gamma_k^{Us}(i)} w_{ij}y_j \;+\hat{\Delta}_i',\; i=1,\ldots,N \qquad (2)$$

where $\hat{\Delta}_i'$ is an adjustment intended to capture the expected effect of selecting the k-nearest neighbors from the reference set $\Gamma_k^{Us}(i)$ instead of the k-nearest neighbors in U, which will be denoted $\Gamma_k^{U}(i)$. Let $\mathbf{X}_{j\in\Gamma_k^{Us}(i)}$ denote the $\mathbf{X}$ values of the k-nearest reference elements to $\boldsymbol{X}_i$ and let $\mathbf{X}_{l\in\Gamma_k^{U}(i)}$ denote the corresponding values of the k-nearest elements in the population. We have

$$\hat{\Delta}_i' = \hat{\Delta}_i - N^{-1}\sum\nolimits_{i=1}^{N}\hat{\Delta}_i, \text{ and } \hat{\Delta}_i = k^{-1}\mathbf{1}_k'\left(\mathbf{X}_{l\in\Gamma_k^{U}(i)} - \mathbf{X}_{j\in\Gamma_k^{Us}(i)}\right)\hat{\boldsymbol{\beta}}_p \qquad (3)$$

where $\mathbf{1}_k$ is a vector of k ones, and $\hat{\boldsymbol{\beta}}_p$ is the vector of p regression coefficients in a sample-based ordinary least squares regression of $\mathbf{Y}_\mathrm{s}$ on $\mathbf{X}_\mathrm{s}$. Throughout, a superscripted apostrophe denotes the transposition of a vector or a matrix. Magnussen, Tomppo et al. (2010) used decorrelated and scaled [0,1] $\mathbf{X}$ variables, and a set of constant, linear, quadratic, and cubic orthogonal Bernstein polynomials (Lorentz 1953, p. 13) to compute $\hat{\Delta}_i$. However, in the populations used in this study, there is no quadratic or cubic relationship between $\mathbf{Y}$ and $\mathbf{X}$, so the simpler method in (3) works equally well.

The calibrated kNN estimator of μ_y is $\tilde{\mu}_y^{k,\mathrm{cal}} = N^{-1}\sum_{i=1}^{N}\tilde{y}_i^{k,\mathrm{cal}}$.

The* k*NN estimator with local linear regression

In the proposed kNN estimator, imputations are generated from a set of $n\times m$ locally weighted linear regression models fitted to the sample data ($\mathbf{Y}_\mathrm{s}$, $\mathbf{X}_\mathrm{s}$). The working assumption states that imputations with this approach will benefit not only from the robustness of local linear smoothing (Chambers & Clark 2012, p. 97) but also harness benefits of a calibration.

For each of the $n\times m$ reference elements, a vector of p +1-weighted least squares regression coefficients were computed as

$$\hat{\boldsymbol{\beta}}_{p+1}^{j} = \left(\mathbf{Z}_\mathrm{s}'\mathbf{W}_j^{-1}\mathbf{Z}_\mathrm{s}\right)^{-1}\mathbf{Z}_\mathrm{s}'\mathbf{W}_j^{-1}\mathbf{Y}_\mathrm{s},\, j=1,\,\ldots,\,n \qquad (4)$$

where $\mathbf{Z}_\mathrm{s}$ is an $(n\,m)\times(p+1)$ matrix resulting from a left-concatenation of $\mathbf{X}_\mathrm{s}$ with a vector of ones (intercept), and $\mathbf{W}_j$ is an $(n\;m)\times(n\;m)$ diagonal matrix with elements $w_{j1},\,\ldots,\,w_{j\,n\times m}$ computed as

$$w_{ji} = \frac{\prod_{v=1}^{p+1}\frac{1}{\vartheta_v}K\left(\frac{x_{vj}-x_{vi}}{\vartheta_v}\right)}{\sum_{i=1}^{n}\prod_{v=1}^{p+1}\frac{1}{\vartheta_v}K\left(\frac{x_{vj}-x_{vi}}{\vartheta_v}\right)},\; i=1,\,\ldots,\;n, \qquad (5)$$

$$K(u) = \max\left(0, \tfrac{3}{4}\left(1-\tfrac{1}{5}u^2\right)\,{}^{-0.5}\!\sqrt{5}\right)$$

where x_{vj} is the value of the auxiliary variable v in the jth sample element, $K(u)$ is the Epanechnikov kernel (Silverman 1986, p. 42), and ϑ_v is the kernel bandwidth, which was chosen as (Silverman 1986, p. 45)

$$\vartheta_v = 0.9\min(\sigma_{x_v},(q_{0.75}[x_v]-q_{0.25}[x_v])1.349)\times n^{-0.2} \qquad (6)$$

The locally weighted regression models are then used to generate k predictions of y_i (i = 1, …, N) using the k vectors of $\mathbf{X}\in\Gamma_k^U(i)$ as predictors. After these preliminaries, the proposed kNN estimator is

$$\tilde{y}_i^{k,\,lr} = k^{-1}\sum\nolimits_{r=1}^{k}\mathbf{Z}'_{\Gamma_k^U(i)[r]}\hat{\boldsymbol{\beta}}_{p+1}^{\Gamma_k^{Us}(i)[r]},\; i=1,\,\ldots,\,N \qquad (7)$$

where $[r]$ denotes the rth element in an ordered set Γ. In words, $\tilde{y}_i^{k,\,\mathrm{lr}}$ is the arithmetic mean of k predictions of y_i generated from: (1) the $\mathbf{X}$ values in the k-nearest neighbors to $\mathbf{X}_i$ in the population; and (2) the associated regression coefficients in k locally weighted regression models selected on the basis of the distance between $\mathbf{X}_i$ and the n reference elements. The proposed kNN estimator of μ_y becomes $\tilde{\mu}_y^{k,\,\mathrm{lr}} = N^{-1}\sum_{i=1}^{N}\tilde{y}_i^{k,\mathrm{lr}}$.

Local linear regression

A kNN estimator with local linear regression may be no better than a simple local linear regression estimator (locreg). We therefore included a locreg estimator $\hat{y}_i^{\mathrm{locreg}}$ in order to answer this question. Computations of $\hat{y}_i^{\mathrm{locreg}}$ follow the steps outlined in Equations (4) and (5) with the exception that the subscript j runs from 1 to N. For reasons of parsimony only results of one target variable (VOL) and SRS are shown.

Distribution matching (DM)

The DM procedure is detailed by Baffetta et al. (2012). A brief excerpt follows. Let $\tilde{y}_{(1)}^{k,\,\mathrm{est}} \le \tilde{y}_{(2)}^{k,\,\mathrm{est}} \le \cdots \le \tilde{y}_{(N)}^{k,\,\mathrm{est}}$, est $=\{\mathrm{st},\mathrm{cal},\mathrm{lr}\}$ be the sequence of kNN imputations listed in ascending order (ditto for $\hat{y}_{(j)}^{\mathrm{locreg}}, j=1,\,\ldots,\,N$). Let $\tilde{F}^{k,\,\mathrm{est}}(y)$ denote an unbiased estimator of the empirical distribution function (EDF) of $\tilde{y}_i^{k,\,\mathrm{est}}$, and let $\tilde{F}_j^{k,\,\mathrm{est}}$ denote the value of $\tilde{F}^{k,\,\mathrm{est}}$ at $\tilde{F}^{k,\,\mathrm{est}}\left(\tilde{y}_{(j)}^{k,\,\mathrm{est}}\right)$. A DM kNN imputation is hereafter $\breve{y}_{(j)}^{k,\,\mathrm{est}} = \hat{F}_\mathrm{s}^{-1}\left(\tilde{F}_j^{k,\,\mathrm{est}}\right)$, where $\hat{F}_\mathrm{s}$ is an unbiased estimator of the population EDF of y. For locreg the corresponding DM estimator becomes $\hat{y}_{(j)}^{\text{locreg}} = \hat{F}_{\text{s}}^{-1}\left(\tilde{F}_j^{\text{locreg}}\right)$.

A truncated [0,1] linear interpolation function was adopted for $\hat{F}_{\text{s}}$ while $\tilde{F}_j^{k,\text{est}} = \left(j - \frac{1}{3}\right) \times \left(N + \frac{1}{3}\right)^{-1}$, which is the median unbiased quantile estimator (Hyndman & Fan 1996). The quantile function $\hat{F}_{\text{s}}^{-1}$ was truncated at a lower and upper limit determined from the minimum (maximum) of $\mathbf{Y}_{\text{s}}$ multiplied by $q_{1/N}^{t,N} \times \left(q_{1/(n\times m)}^{t,n\times m}\right)^{-1}$, where $q_p^{t,\nu}$ is the (100p)th sample percentile of a t-distribution with ν degrees of freedom.

Estimator performance

The key performance indicators for $\hat{y}_i^{k,\text{st}}$, $\hat{y}_i^{k,\text{cal}}$, $\hat{y}_i^{k,\text{lr}}$, and $\hat{y}_i^{\text{locreg}}$ are the correlation with y_i and the root mean square error of unit-level imputations $\widehat{\text{RMSE}}^{k,\text{est}} = N^{-0.5}\left(\sum_{i=1}^{N}\left(\hat{y}_i^{k,\text{est}} - y_i\right)^2\right)^{0.5}$, est $= \{\text{st}, \text{cal}, \text{lr}\}$ with a logical parallel for $\widehat{\text{RMSE}}^{\text{locreg}}$. Differences between $\hat{y}_i^{k,\text{est}}\left(\hat{y}_i^{\text{locreg}}\right)$ and y_i were also quantified and tested with a Hotelling's T^2-statistic (Rencher 1995, p. 133) under the null hypothesis of a linear relationship with a slope of one and an intercept of zero.

The estimators $\hat{\mu}_y^{k,\text{est}}\left(\hat{\mu}_y^{\text{locreg}}\right)$ are also assessed for bias and variance. Bias is estimated as the difference between the mean of 400 independent replicated estimates of $\hat{\mu}_y^{k,\text{est}}\left(\hat{\mu}_y^{\text{locreg}}\right)$ and μ_y. The efficiency of the estimators is assessed by comparing the empirical (Monte Carlo) variances of $\hat{\mu}_y^{k,\text{est}}\left(\hat{\mu}_y^{\text{locreg}}\right)$ in 400 replications of a specific sampling design (see next).

Sampling designs

Sample sizes in SRS were $n = \{50, 100, 200, 300\}$ elements. In CLU sampling, sample sizes were $n_{\text{c}} = \{12, 25, 50, 75\}$, that is, $n_{\text{c}} \times m$ = 48, 100, 200, and 300 elements. With a population size (N) of 8000, the sample fractions for SRS were 0.0063, 0.0125, 0.0250, and 0.0375. Under CLU they were, with one minor difference, identical.

The kNN estimators were evaluated with k values of 1, 2, 4, 6, 8, 10, and 12. Each of the two (*SRS, CLU*) × 4 (n) × 7 (k) = 56 settings were replicated n_{rep} = 400 times followed by a computation of the above estimators.

Case studies

The performance of the three kNN estimators (st, cal, lr) and the local linear regression estimator (locreg) was evaluated in three artificial populations and in two synthetic populations with actual data from the ninth Finnish National Forest Inventory (Tomppo et al. 2011).

The artificial multivariate populations (POP1, POP2, and POP3) of size N = 8000 elements were generated from known marginal distributions of **X** and **Y** and defined correlation coefficients between the variables in **X** and **Y**.

There are three Y variables (*Y1, Y2*, and *Y3*) in each of the artificial populations, three X variables in POP1 (*X1*, *X2*, and *X3*), and four in POP2 and POP3 (*X1*, *X2*, *X3*, and *X4*). The marginal distributions of variables in the three populations were complex in order to reflect scenarios with skewed, multi-modal, and non-Gaussian distributions in forest inventory applications. These types of distributions are not uncommon in actual forest inventories (LeMay et al. 2008; Magnussen et al. 2009). To simplify reporting, all variables were standardized to a mean of zero and a variance of one. Details of the populations are in the Appendix 2 and Table 1. Figure 1 shows the marginal distributions of standardized **X** and **Y** values.

The two Finnish population sets represent forested areas on MINERAL (MIN) and PEATLAND (PEAT) soils in the eastern part of Central Finland (North Karelia and South Savo), approximately between latitudes 61° 10′N and 63°95′N and longitudes 27°20′E and 31°10′E. A population unit is a quarter of a Landsat 7 ETM+ image pixel (approximately 12.5 m × 12.5 m in size) from path 186 and rows 16 and 17 (acquisition date: June 10, 2000). Only cloud-free pixels are used. The image data were co-registered to the national base maps (with soil strata) (Tomppo & Halme 2004 give more

Table 1. Pair-wise variable target correlations in populations POP1, POP2, and POP3. Realized correlations between two different variables in the randomly generated populations of 8000 elements may deviate by up to ≤0.02 from the target.

	X1	*X2*	*X3*	*X4*	*Y1*	*Y2*	*Y3*
POP1							
X1	1.00	0.80	0.40	–	0.10	0.20	0.05
X2		1.00	0.50	–	0.40	0.30	0.00
X3			1.00	–	0.20	0.20	0.30
Y1					1.00	0.70	0.60
Y2						1.00	0.20
POP2							
X1	1.00	0.50	0.50	0.20	0.50	0.30	0.20
X2		1.00	0.50	0.50	0.20	0.50	0.30
X3			1.00	0.50	0.50	0.20	0.50
X4				1.00	0.50	0.50	0.20
Y1					1.00	0.50	0.50
Y2						1.00	0.50
POP3							
X1	1.00	0.70	0.50	−0.20	0.40	0.30	0.00
X2		1.00	0.60	−0.20	0.50	0.50	−0.10
X3			1.00	0.30	0.30	0.20	0.10
X4				1.00	−0.20	−0.20	0.10
Y1					1.00	0.70	−0.70
Y2						1.00	−0.50

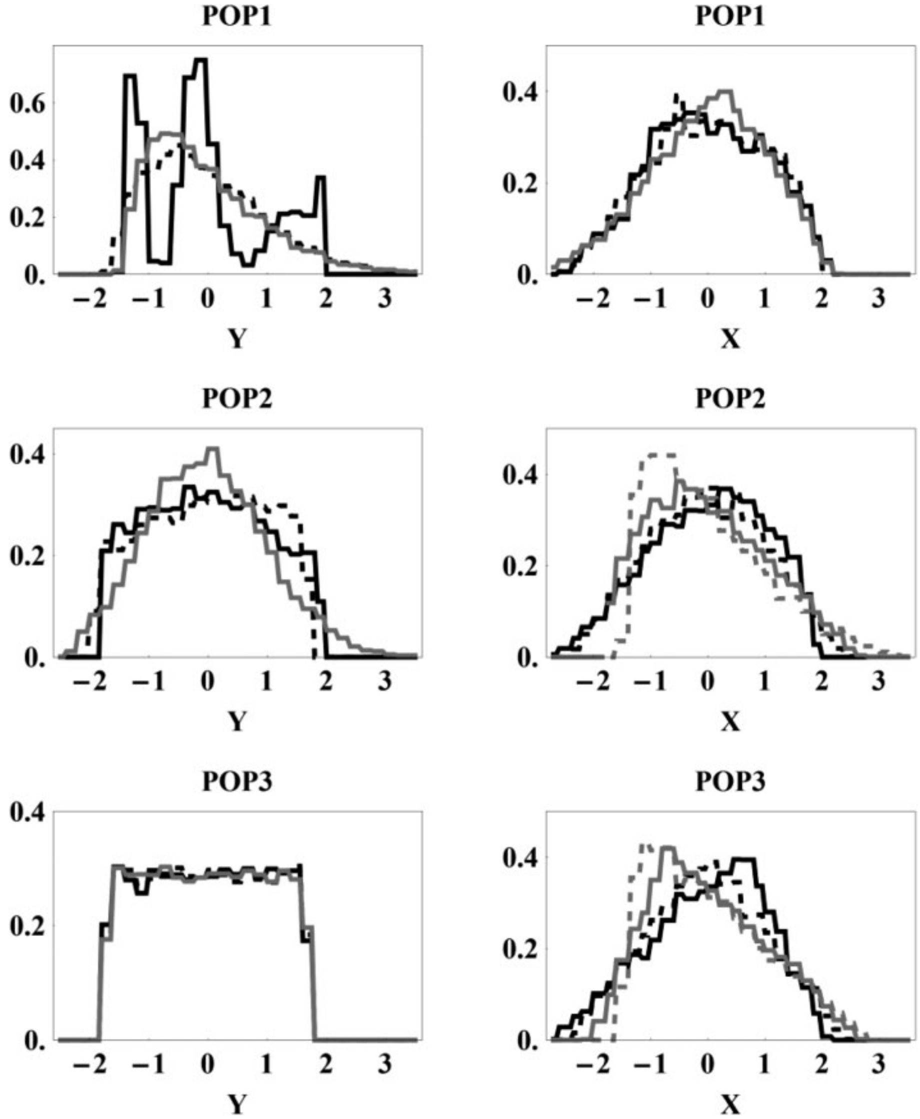

Figure 1. Marginal distributions of standardized values of **X** and **Y** variables in three simulated populations.

details). Only pixels co-located with the ninth Finnish National Forest Inventory plots are used here. The total number of pixels (N) in the MIN dataset is 5330 and 1633 in the PEAT set. Stem volume (VOL) is the target variable (Y). A forward stepwise regression analysis identified five auxiliary variables (**X**) as significant at the 2% level. They are location (easting and northing) and TM bands 3, 5, and 9. All correlation coefficients among auxiliary variables were < 0.65. Only the SRS designs were simulated for MIN and PEAT.

Results

POP1-POP3

Simple random sampling

Estimates of bias in $\widehat{\mu}_y^{k,\,\text{st}}$ and $\widehat{\mu}_y^{k,\,\text{lr}}$ were similar across populations, Y variables, sample sizes, and k values. No estimate of bias (min 0.007, max 0.002) was deemed important. Pair-wise t tests of differences in bias between $\widehat{\mu}_y^{k,\,\text{st}}$ and $\widehat{\mu}_y^{k,\,\text{lr}}$ identified a total of 13 statistically significant differences out of a possible 252 (3 populations, 3 variables, 4 sample sizes, 7 k values), which is close to the expected rate of 0.05 under a null hypothesis of no difference. The conclusion is that bias in $\widehat{\mu}_y^{k,\,\text{st}}$ is of the same magnitude as the bias of $\widehat{\mu}_y^{k,\,\text{lr}}$. No separate bias assessment was done for $\widehat{\mu}_y^{k,\,\text{cal}}$ since it is calibrated to $\widehat{\mu}_y^{k,\,\text{st}}$. DM matching had, as expected, no discernible effect (<0.2%) on bias.

The overall mean of $(\widehat{\mu}_y^{k,\,\text{st}})$ was 0.0086 compared to 0.0089 for $(\widehat{\mu}_y^{k,\,\text{lr}})$. A Levene's test (Shoemaker 2003) of equal empirical variances of $\widehat{\mu}_y^{k,\,\text{st}}$ and $\widehat{\mu}_y^{k,\,\text{lr}}$ at the 95% level of significance indicated 36 significant differences in a total of 252 comparisons. Of 36 tests with a P value < 0.05, 25 indicated that $\text{var}(\widehat{\mu}_y^{k,\,\text{st}}) \leq \text{var}(\widehat{\mu}_y^{k,\,\text{lr}})$, yet differences were less than 5%. Most (32) of the significant differences were in POP2 and POP3, where the average of var $(\hat{\mu}_y^{k,\mathrm{lr}})$ was 7% larger than the variance of var $\hat{\mu}_y^{k,\mathrm{st}}$. In POP1 var $(\hat{\mu}_y^{k,\mathrm{lr}})$ was 5% lower than var $(\hat{\mu}_y^{k,\mathrm{st}})$. Significant differences appeared to be uniformly distributed across sample sizes and k values.

The empirical variance of both $\hat{\mu}_y^{k,\mathrm{st}}$ and $\hat{\mu}_y^{k,\mathrm{lr}}$ declined with increasing sample size at the expected rate of approximately n^{-1}. A k value of 6 produced the lowest variance, although results for $k = 4$ and $k = 8$ were almost identical. The variance of $\hat{\mu}_y^{k,\mathrm{cal}}$ was, as expected, equal to the variance of $\hat{\mu}_y^{k,\mathrm{st}}$.

DM did not change the empirical variances of $\hat{\mu}_y^{k,\mathrm{est}}$ to any degree of practical importance. A DM could, for a given combination of population, variable, n, and k, increase or decrease the variance by up to 2%. When averaged over all settings the difference was less than 0.1%.

The correlation between y_i and $\hat{y}_i^{k,\mathrm{st}}$ was, overall, 0.31 but higher (0.40) for $\hat{y}_i^{k,\mathrm{lr}}$ and $\hat{y}_i^{k,\mathrm{cal}}$(0.34). The correlations varied significantly among populations and Y variables, but differences $\hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{lr}}\right) - \hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{st}}\right)$ and $\hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{lr}}\right) - \hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{cal}}\right)$ remained almost constant across combinations of population and Y variable. In all but two cases (POP3, $n = 50$, $Y1$, and $k = \{10, 12\}$) did the difference $\hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{lr}}\right) - \hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{st}}\right)$ reach statistical significance ($P < 0.05$) in paired t-tests with Fisher's z-transform of a correlation coefficient (Fisher 1915). With increasing k, the difference $\hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{lr}}\right) - \hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{st}}\right)$ decreased. For $k \leq 2$ it was approximately 0.14, and for $k \geq 10$ it was 0.04. Increasing the sample size from 50 to 300 improved $\hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{lr}}\right)$ by 0.08 but less so for $\hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{st}}\right)$ (0.05) and $\hat{\rho}\left(y_i, \hat{y}_i^{k,\mathrm{cal}}\right)$ (0.04).

The DM procedure barely altered the correlation coefficient between y and its imputed value. The largest (average) change in a given design setting was approximately ±0.008. The average change in a design setting was approximately ±0.003.

Root mean square errors of $\hat{y}_i^{k,\mathrm{lr}}$ were, overall, 9% lower than $\mathrm{RMSE}\left(\hat{y}_i^{k,\mathrm{st}}\right)$ and 4% lower than $\mathrm{RMSE}\left(\hat{y}_i^{k,\mathrm{cal}}\right)$. These relative effect-sizes were approximately constant across populations and Y variables. Local regressions achieved the largest reduction in RMSE in settings with $n \geq 200$ and $k \leq 6$, as expected from the above detailed trends in correlation coefficients. In favorable combinations of k and n, $\widehat{\mathrm{RMSE}}\left(\hat{y}_i^{k,\mathrm{lr}}\right)$ was 12% lower than $\widehat{\mathrm{RMSE}}\left(\hat{y}_i^{k,\mathrm{st}}\right)$ and 7% lower than $\widehat{\mathrm{RMSE}}\left(\hat{y}_i^{k,\mathrm{cal}}\right)$. In combinations of large k and small n, the reductions were modest (2–4%). An F-test of the hypothesis $\mathrm{RMSE}\left(\hat{y}_i^{k,\mathrm{lr}}\right) \times \widehat{\mathrm{RMSE}}\left(\hat{y}_i^{k,\mathrm{st}}\right)^{-1} = 1$ was rejected ($P < 0.05$) in all but eight cases with $n = 50$ and $k \geq 4$. Tests of $\widehat{\mathrm{RMSE}}\left(\hat{y}_i^{k,\mathrm{lr}}\right) \times \widehat{\mathrm{RMSE}}\left(\hat{y}_i^{k,\mathrm{cal}}\right)^{-1} = 1$ was only rejected when $k < 4$ and $n \leq 100$.

In this study, the DM incurred an increase in all estimates of RMSE. The increase varied from 3 to 12%, regardless of population, variable, sample size, or k value. The increase was similar for $\hat{y}_i^{k,\mathrm{st}}$ $\hat{y}_i^{k,\mathrm{lr}}$ and $\hat{y}_i^{k,\mathrm{cal}}$.

Imputations with locally weighted regression models improved the slope in regressions of y_i on its imputed value. In regressions of y_i on $\hat{y}_i^{k,\mathrm{st}}$ the average slope was 0.29; the average increased to 0.34 in regressions on $\hat{y}_i^{k,\mathrm{cal}}$ and to 0.39 in regressions on $\hat{y}_i^{k,\mathrm{lr}}$. Intercepts were correspondingly reduced toward zero. As reported for correlations and RMSE, the positive effect of imputations with local linear regressions was approximately constant across populations and Y variables. In 215 cases out of 252 possible, the slope in the regression of y_i on $\hat{y}_i^{k,\mathrm{lr}}$ was significant larger than the slope obtained with $\hat{y}_i^{k,\mathrm{cal}}$. All non-significant results were concentrated in settings with $n = 50$ and $k \geq 6$. The slope in regressions of y_i on $\hat{y}_i^{k,\mathrm{st}}$ was, in 242 of 252 cases, significantly lower than the slope obtained with $\hat{y}_i^{k,\mathrm{lr}}$. The null hypothesis of a slope of 1 and an intercept of 0 was, however, rejected in nearly all cases.

Cluster sampling

Bias estimates from CLU designs were very similar to their counterparts in SRS designs. Empirical variances of $\hat{\mu}_y^{k,\mathrm{st}}$ and $\hat{\mu}_y^{k,\mathrm{lr}}$ were due to a positive intra-cluster correlation larger (20–30%) than corresponding variances under a SRS design. In POP1 var $(\hat{\mu}_y^{k,\mathrm{lr}})$ was, in 15 cases (out of 84), significantly smaller (10–15%) than var $(\hat{\mu}_y^{k,\mathrm{st}})$. There was no instance where var $(\hat{\mu}_y^{k,\mathrm{st}})$ was significantly smaller than var $(\hat{\mu}_y^{k,\mathrm{lr}})$. However, in POP2 and POP3, a total of 91 differences (out of 168) were significantly different from zero. In more than half the cases (58), the test suggested var $(\hat{\mu}_y^{k,\mathrm{st}}) \leq$ var $(\hat{\mu}_y^{k,\mathrm{lr}})$ with a typical difference of approximately 8%. There was no apparent pattern to the occurrence of significant differences with respect to either n or k. As under SRS var $(\hat{\mu}_y^{k,\mathrm{cal}}) \cong$ var $(\hat{\mu}_y^{k,\mathrm{lr}})$.

Trends and differences in correlation coefficients were virtually identical to those reported for the SRS designs. A non-significant difference in a CLU design was also non-significant in the corresponding SRS design.

Results and trends in RMSE under a CLU design also followed those reported for SRS designs. The only but consistent difference was a slight increase of 0.003 in individual RMSE estimates, thus reflecting a lower-variance effective sample size of CLU designs (Faes et al. 2009).

MIN and PEAT (SRS, VOL)

Bias of $\tilde{\mu}_{\text{vol}}^{k,\text{est}}$ $\text{est} = \{\text{st}, \text{cal}, \text{lr}\}$ was also minor (<0.3%) in the MIN and PEAT populations from Finland. In MIN, the bias was slightly lower in $\tilde{\mu}_{\text{vol}}^{k,\text{lr}}$ than in $\tilde{\mu}_{\text{vol}}^{k,\text{st}}$, but in PEAT the opposite was true. No difference across the 56 settings was statistically significant. DM, as found in POP1-POP3, had no discernible effect on bias.

Empirical variances of $\widehat{\mu}_y^{k,\text{lr}}$ were, in both MIN and PEAT, 7–8% higher than the variances of $\widehat{\mu}_y^{k,\text{st}}$ but no difference was statistically significant at the 5% level (Levene's test). As in POP1-POP3, the DM procedure had no visible effect on the empirical variance.

The unit-level correlation between $\tilde{y}_i^{k,\text{lr}}$ and y_i was, on both sites, slightly stronger (0.55) than between $\tilde{y}_i^{k,\text{st}}$ and y_i (0.51). Trends across values of k and n were weaker, but otherwise similar in directions to those reported for POP1-POP3.

*RMSE*s of $\tilde{y}_i^{k,\text{lr}}$ were, on both sites, on average, 3% lower than RMSEs of $\tilde{y}_i^{k,\text{st}}$. Again, the largest reduction achieved by the local linear regressions was seen in results with k values of 1 and 2 (6%) where the RMSE values were slightly greater (2%) than that for larger ks. Differences in RMSEs were only statistically significant at a rate of 1:32 (all with $k = 1$ or $k = 2$), that is, not far from the expected rate under the null hypothesis of no difference. The negative effect of a DM seen in POP1-POP3 was confirmed in both MIN and PEAT.

Linear regressions of $\tilde{y}_i^{k,\text{lr}}$ against y_i had, on both sites, on average, a slope (0.64) that was slightly (0.04) larger than the slope in the regressions of $\tilde{y}_i^{k,\text{st}}$ against y_i but the differences were never significant at the 5% level. Intercepts were virtually identical. Intercepts and slopes following a DM were further away from the desired one-to-one line than before the DM.

Local linear regression

In POP1-POP3, a local linear regression under the SRS design with VOL as the dependent variable (Y) achieved in most (7 of 12) design-settings a slightly higher (0.02) correlation between the predicted and actual value of Y than possible with $y_i^{k,\text{lr}}$. In the remaining five cases the absolute difference was less than 0.01. DM had next to no effect on these correlations. However, RMSEs of $\hat{y}_i^{\text{locreg}}$ was consistently (across populations and sample-sizes) between 18 and 24% lower than the RMSEs for the k value(s) that produced the lowest estimate of RMSE of $\widehat{y}_i^{k,\text{lr}}$ in a corresponding combination of population and sample size. In every single replication, $\widehat{\text{RMSE}}\left(\hat{y}_i^{\text{locreg}}\right)$ was at least 5% lower than $\widehat{\text{RMSE}}\left(\widehat{y}_i^{k,\text{lr}}\right)$. The negative effect of a DM on RMSE was also seen in the results with locreg and it cut the RMSE differences between kNNlr and locreg to approximately one-half of the differences prior to a DM.

Results for locreg in MIN and PEAT confirmed the correlation analyses for POP1–POP3 and equally the lower RMSE (23–28% in MIN and 15–18% in PEAT) of $\hat{y}_i^{\text{locreg}}$ as well as the (approximately) same relative magnitude of the negative effect of a DM.

Discussion

The core attraction of the kNN technique to forest inventory is simplicity. Unit-level multivariate imputations can be generated in an instant. As for any other model, the performance of kNN depends critically on the auxiliary variables and their associations with the target variable(s). If associations are nonlinear, all kNN method will fail because the distance metric and the weights become nonlinear by virtue of the function to transform **X** to **Y** (Stage & Crookston 2007). With a risk of failure, an analyst should investigate the potential presence of a nonlinear relationship (Bunzel et al. 2001).

Variable selection was beyond the scope of this study, but it remains a complex challenge (McRoberts 2009; Chirici et al. 2012; Packalén et al. 2012) due to the *curse of dimensionality*, which is of particular relevance to the kNN technique (Hastie et al. 2005, ch. 2.5).

The choice of a distance metric and weighting of reference units have been viewed as important tuning parameters for optimizing the performance of the kNN technique (Tomppo & Halme 2004). Yet we still lack convincing examples showing that the performance of kNN can be improved significantly by a manipulation of these two parameters (Katila & Tomppo 2001; McRoberts 2009).

The proposed kNN method with a fixed linear model form and locally weighted linear regressions is easy to implement. It only requires the addition of a set of $n_c \times m$-weighted least squares regressions. The kNN attraction of simplicity is therefore not lost. An automatic choice of bandwidth may forgo some reduction in RMSE (Gao & Gijbels 2009), but optimizing the bandwidth for each combination of Y variable and value of k could become an onerous task.

Although the improvements achieved in this study with the proposed kNN with locally weighted linear regressions were not impressive, they were nevertheless consistent across variables and populations and confirmed that localized models may be better at exploiting local patterns in associations between **X** and **Y** than a global model (Yim et al. 2010; Räty & Kangas 2012; Ver Hoef & Temesgen 2013). The only negative side-effect of the proposed kNN technique appears to be the possibility of a slight increase in the empirical variances of the estimated population mean.

In this study, a local linear regression model for a univariate Y outperformed the best kNN variant across all populations and sample sizes. Unfortunately, one cannot extrapolate the univariate performance to the case with multiple dependent variables (Friedman 1991; Ruppert & Wand 1994). But the promising performance of locreg warrants further studies with a multivariate *Y*.

The improvement in the performance of $\widehat{y}^{k,\mathrm{lr}}$ with increasing n is important. In *k*NN applications, sample sizes are typically larger than a few hundred. One should therefore expect that the local linear regressions collectively capture the relationships between $\mathbf{X}$ and $\mathbf{Y}$ across the space of $\mathbf{X}$ or at least better than with the standard *k*NN technique.

The fact that the benefits of imputations with locally weighted linear regression were largely limited to k values between 1 and 4 is not seen as a major detractor, although practical applications with $k > 15$ are not uncommon (Franco-Lopez et al. 2001). First, when the dimension of $\mathbf{Y}$ is beyond 3, simple theoretical considerations (Hastie et al. 2005, ch. 2.5) point to the use of a small k value, at least when the efficiency of the *k*NN technique is a concern. Large k values also imply a non-trivial covariance among imputed values due to the repeated use of a reference element in multiple imputations (McRoberts et al. 2007; Magnussen et al. 2009; Magnussen, McRoberts et al. 2010). This study suggests that for imputations with locally weighted linear regressions, one should try to keep k at four or below. We base this recommendation on results that showed that: (1) empirical variance of $\tilde{\mu}_y^{k,\mathrm{lr}}$ was lowest for $k = 6$ variances but very similar for $k = 4$; (2) the gain in the strength of the correlation between $\widehat{y}_i^{k,\mathrm{lr}}$ and y_i increased with decreasing k; and (3) the decrease in RMSE of $\widehat{y}_i^{k,\mathrm{lr}}$ relative to that of $\widehat{y}_i^{k,\mathrm{st}}$ was only significant for $k \leq 4$. Also, with this technique good results can be achieved with k as low as one, in line with what has been demonstrated with the Most Similar Neighbor technique (Moeur et al. 1995; LeMay & Temesgen 2005; Hudak et al. 2008).

The benefit of a *k*NN calibration (Magnussen, Tomppo et al. 2010) was confirmed, although on a relative scale, they were smaller than expected and consistently smaller than the benefits of imputations with locally weighted linear regressions. Because the computational efforts behind $\widehat{y}_i^{k,\mathrm{cal}}$ and $\widehat{y}_i^{k,\mathrm{lr}}$ are not materially different, the proposed local linear imputation method is practical.

The performance of *k*NN imputations with locally weighted linear regressions was similar in designs with simple random sampling of single population elements and with simple random sampling of compact clusters of four population elements. The only quantitative difference was a logical consequence of the positive intra-cluster correlation in the compact clusters of four elements, a correlation that leads to a decreased (variance) efficiency and a decreased variance effective sample size (Cochran 1977, p. 240; Thiébaux & Zwiers 1984). Implementation of *k*NN with locally weighted linear regressions – in a context of cluster sampling – is no different from an implementation under a SRS sampling design.

A variance estimator for a population applicable to the standard *k*NN technique (Magnussen 2013) will not need any modification in order to accommodate an estimate obtained with a *k*NN technique that uses locally weighted linear regressions. The same holds for the calibrated *k*NN. Baffetta et al. (2012) demonstrated that this conclusion can be extended to DM *k*NN estimates. For the populations in this study, both the empirical difference variance estimator by Baffetta et al. (2009) and the jackknife variance estimator (Wolter 2007, p. 162) performed well (Magnussen 2013).

DM (Lister & Lister 2006; Baffetta et al. 2012) is arguably an effective and easy-to-implement method to achieve a variance in $\widehat{y}_i^{k,\mathrm{est}}$ that matches the variance in $\mathbf{Y}_s$. However, DM does not improve the correlation between the true and the imputed $\mathbf{Y}$ value. In a toy-like example ($N = 225$, $n = 25$) given in Baffetta et al. (2012), the RMSE of DM *k*NN imputations was no larger than the RMSE of standard *k*NN imputations. However, the inflation in RMSE seen in this study is a reason for concern as it may question the use of DM as a routine post-hoc processing method. In the study of Baffetta et al. (2012), the domains of support for the EDFs were capped to the range of $\mathbf{Y}_s$, hence no allowance was made to counter the effect of sample size on the observed range of a variable (Casella & Berger 2002, p. 231). In this study, we saw numerous examples where the range in $\mathbf{Y}_s$ was between 75 and 85% of the full range in $\mathbf{Y}$. With the proposed *k*NN technique, the benefit of a DM is reduced because the variance $\tilde{y}_i^{k,\mathrm{lr}}$ was always considerably larger than the variance of $\tilde{y}_i^{k,\mathrm{st}}$. More research is warranted on the impact of DM on RMSE and to clarify when the support domain of the EDF should or should not be restricted to the observed sample range.

References

Aha WD. 1997. Lazy learning. Artif Intell Rev. 11:7–10.

Azzalini A. 1985. A class of distributions which includes the normal ones. Scand J Stat. 12:171–178.

Baffetta F, Corona P, Fattorini L. 2012. A matching procedure to improve *k*-NN estimation of forest attribute maps. For Ecol Manage. 272:35–50.

Baffetta F, Fattorini L, Franceschii S, Corona P. 2009. Design-based approach to *k*-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sens. Environ 113:463–475.

Breidenbach J, Nothdurft A, Kändler G. 2010. Comparison of nearest neighbour approaches for small area estimation of tree species-specific forest inventory attributes in central Europe using airborne laser scanner data. Eur J For Res. 129:833–846.

Bunzel H, Kiefer NM, Vogelsang TJ. 2001. Simple robust testing of hypotheses in nonlinear models. J Am Stat Assoc. 96:1088–1096.

Casella G, Berger RL. 2002. Statistical inference. 2nd ed. Pacific Grove: Duxbury Press.

Chambers RL, Clark RG. 2012. An introduction to model-based survey sampling with applications. New York: Oxford University Press.

Chirici G, Corona P, Marchetti M, Mastronardi A, Maselli F, Bottai L, Travaglini D. 2012. *K*-NN forest: a software for the non-parametric prediction and mapping of environmental variables by the *k*-nearest neighbors algorithm. Eur J Rem. Sens. 45:433–442.

Cleveland WS, Devlin SJ, Grosse E. 1988. Regression by local fitting: methods, properties, and computational algorithms. J Econom. 37:87–114.

Cochran WG. 1977. Sampling techniques. New York: Wiley.

Corona P, Chirici G, Marchetti M. 2002. Forest ecosystem inventory and monitoring as a framework for terrestrial natural renewable resource survey programmes. Pl Biosys. 136:69–82.

Crookston NL, Finley AO. 2008. yaImpute: an R package for *k*NN imputation. J Stat Softw. 23:16.

Eskelson BNI, Temesgen H, Barrett TM. 2009. Estimating current forest attributes from paneled inventory data using plot-level imputation: a study from the Pacific Northwest. For Sci. 55:64–71.

Eskelson BNI, Temesgen H, Lemay V, Barrett TM, Crookston NL, Hudak AT. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand J For Res. 24:235–246.

Faes C, Molenberghs G, Aerts M, Verbeke G, Kenward MG. 2009. The effective sample size and an alternative small-sample degrees-of-freedom method. Am Stat. 63:389–399.

Fazar W. 1959. Program evaluation and review technique. Am Stat. 13:10–16.

Fischer M. 2010. Multivariate copulae. In: Kurowicka D, Joe H, editors. Dependence modeling. Singapore: World Scientific; p. 19–36.

Fisher RA. 1915. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika. 10:507–521.

Franco-Lopez H, Ek AR, Bauer ME. 2001. Estimation and mapping of forest stand density, volume, and cover type using the *k*-nearest neighbors method. Remote Sens Environ. 77:251–274.

Friedman JH. 1991. Multivariate adaptive regression splines. Ann Stat. 19:1–67.

Gao J, Gijbels I. 2009. Bandwidth selection in nonparametric kernel testing. J Am Stat Assoc. 103:1584–1594.

Haara A, Kangas A. 2012. Comparing K nearest neighbours methods and linear regression – is there reason to select one over the other? Math Comp Forest Nat Res Sci. 4:50–65.

Haara A, Maltamo M, Tokola T. 1997. The *k*-nearest-neighbour method for estimating basal area diameter distribution. Scand J For Res. 12:200–208.

Hastie T, Tibshirani R, Friedman J, Franklin J. 2005. The elements of statistical learning: data mining, inference and prediction. Math Intel. 27:83–85.

Holmström H, Fransson JES. 2003. Combining remotely sensed optical and radar data in *k*NN-estimation of forest variables. For Sci. 49:409–418.

Hudak AT, Crookston NL, Evans JS, Hall DE, Falkowski MJ. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sens Environ. 113:289–290.

Hyde P, Dubayah R, Walker W, Blair JB, Hofton M, Hunsaker C. 2006. Mapping forest structure for wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM+, Quickbird) synergy. Remote Sens Environ. 102:63–73.

Hyndman RJ, Fan Y. 1996. Sample quantiles in statistical packages. Am Stat. 50:361–365.

Katila M, Tomppo E. 2001. Selecting estimation parameters for the Finnish multisource National Forest Inventory. Remote Sens Environ. 76:16–32.

Koistinen P, Holmström L, Tomppo E. 2008. Smoothing methodology for predicting regional averages in multi-source forest inventory. Remote Sens Environ. 112:862–871.

LeMay V, Maedel J, Coops NC. 2008. Estimating stand structural details using nearest neighbor analyses to link ground data, forest cover maps, and Landsat imagery. Remote Sens Environ. 112:2578–2591.

LeMay V, Temesgen H. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. For Sci. 51:109–119.

Lister AJ, Lister TW. 2006. Post-modeling histogram matching of maps produced using regression trees. In: McRoberts RE, Reams GA, Van Deusen PC, McWilliams WH, editors. USFS Sixth Annual Forest Inventory and Analysis Symposium; Sep 21–24; Denver (CO); p. 111–118.

Lorentz GG. 1953. Bernstein polynomials. 2nd ed. Toronto: University of Toronto Press.

Magnussen S. 2013. An assessment of three variance estimators for the *k*-nearest neighbour technique. Silva Fenn. 47:19.

Magnussen S, Köhl M. 2006. A better alternative to Wald's test-statistic for simple goodness-of-fit tests under one-stage cluster sampling. For Ecol Manage. 221:123–132.

Magnussen S, McRoberts RE, Tomppo E. 2009. Model-based mean square error estimators for *k*-nearest neighbour predictions and applications using remotely sensed data for forest inventories. Remote Sens Environ. 113:476–488.

Magnussen S, McRoberts RE, Tomppo E. 2010. A resampling variance estimator for the *k*-nearest neighbours technique. Can J For Res. 40:648–658.

Magnussen S, Tomppo E, McRoberts RE. 2010. A model-assisted *k*-nearest neighbour approach to remove extrapolation bias. Scand J For Res. 25:174–184.

Maltamo M, Kangas A. 1998. Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Can J For Res. 28:1107–1115.

Mandallaz D. 2013. Design-based properties of some small-area estimators in forest inventory with two-phase sampling. Can J For Res. 43:441–449.

Maselli F, Chirici G, Bottai L, Corona P, Marchetti M. 2005. Estimation of Mediterranean forest attributes by the application of *k*-NN procedures to multitemporal Landsat ETM plus images. Int J Remote Sens. 26:3781–3796.

McRoberts RE. 2009. Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sens Environ. 113:489–499.

McRoberts RE, Cohen WB, Næsset E, Stehman SV, Tomppo EO. 2010. Using remotely sensed data to construct and assess forest attribute maps and related spatial products. Scand J For Res. 25:340–367.

McRoberts RE, Tomppo EO, Finley AO, Heikkinen J. 2007. Estimating areal means and variances of forest attributes using the *k*-nearest neighbors technique and satellite imagery. Remote Sens Environ. 111:466–480.

Moeur M, Crookston NL, Stage AR. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For Sci. 41:337–359.

Opsomer JD, Claeskens G, Ranalli MG, Kauermann G, Breidt FJ. 2008. Non-parametric small area estimation using penalized spline regression. J R Stat Soc Series B Stat Methodol. 70:265–286.

Paass G. 1985. Statistical record linkage methodology: state of the art and future prospects, Vol. L1 Book 2. Voorburg: International Statistical Institute (ISI).

Packalén P, Temesgen H, Maltamo M. 2012. Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. Can J Rem Sens. 38:557–569.

Räty M, Kangas A. 2012. Comparison of *k*-MSN and kriging in local prediction. For Ecol Manage. 263:47–56.

Rencher AC. 1995. Methods of multivariate analysis. New York: Wiley.

Ruppert D, Wand MP. 1994. Multivariate locally weighted least squares regression. Ann Stat. 22:1346–1370.

Scott DW. 1992. Multivariate density estimation: theory, practice and visualization. New York: Wiley.

Shoemaker LH. 2003. Fixing the *F* test for equal variances. Am Stat. 57:105–114.

Silverman BW. 1986. Density estimation for statistics and data analysis. London: Chapman & Hall.

Srinivas S, Menon D, Prasad AM. 2006. Multivariate simulation and multimodal dependence modeling of vehicle axle weights with copulas. J Trans Eng. 132:945–955.

Stage AR, Crookston NL. 2007. Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. For Sci 53:62–72.

Stern H. 1990. Models for distribution on permutations. J Am Stat Assoc. 85:558–564.

Temesgen H, Barrett T, Latta G. 2008. Estimating cavity tree abundance using nearest neighbor imputation methods for western Oregon and Washington forests. Silva Fenn. 42:337–354.

Temesgen H, LeMay V, Froese KL, Marshall PL. 2003. Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. For Ecol Manage. 177:285.

Thiébaux HJ, Zwiers FW. 1984. The interpretation and estimation of effective sample size. J Clim Appl Meteor. 23:800–811.

Tomppo E. 1991. Satellite image-based national forest inventory of Finland. In: International Archives of Photogrammetry and Remote Sensing; Vol. 28, Part 7-1. Proceedings of the symposium on global and environmental monitoring, techniques and impacts; Victoria (BC); p. 419–424.

Tomppo E. 2006. The Finnish multi-source national forest inventory – small area estimation and map production. In: Kangas A, Maltamo M, editors. Forest inventory – methodology and applications. Dordrecht: Springer; p. 195–224.

Tomppo E, Halme M. 2004. Using coarse scale forest variables as ancillary information and weighting of variables in *k*-NN estimation: a genetic algorithm approach. Remote Sens Environ 92:1–20.

Tomppo E, Heikkinen J, Henttonen HM, Ihalainen A, Katila M, Mäkelä H, Tuomainen T, Vainikainen N. 2011. Designing and conducting a forest inventory – case: 9th National Forest Inventory of Finland. New York: Springer; 272 p.

Ver Hoef JM, Temesgen H. 2013. A comparison of the spatial linear model to nearest neighbor (*k*-NN) methods for forestry applications. PLoS ONE. 8:e59129.

Wolter KM. 2007. Introduction to variance estimation. 2nd ed. New York: Springer.

Yim JS, Kim YH, Kim SH, Jeong JH, Shin MY. 2010. Comparison of the *k*-nearest neighbor technique with geographical calibration for estimating forest growing stock volume. Can J For Res. 41:73–82.

Zhang L, Shi H. 2004. Local modeling of tree growth by geographically weighted regression. For Sci. 50:225–244.

Appendix 1. Notation and definitions in order of appearance in the text.

Symbol	Definition	Equations
kNN	k-nearest neighbor technique of imputation	
K	The number of neighbors used in a kNN imputation	
$\mathbf{X}$	The population matrix of auxiliary variables, $\mathbf{X}$ defines the feature space	
Y ($\mathbf{Y}$)	The population vector (matrix) of the target variable(s)	
$\tilde{\mathbf{Y}}_{knn}^{\text{est}}$	A kNN estimate of Y, est is short for *estimator*. est = {st, cal, lr}. st = standard kNN estimator, cal = calibrated kNN estimator, lr = local regression kNN estimator	
$\tilde{y}_i^{k,\text{est}}$	A kNN estimate for the ith population unit (pixel) obtained with estimator est and k-nearest neighbors	(1)–(3)
N	Population size	
N	Sample size (number of elements)	
n_c	Sample size (number of clusters of m elements)	
f	A generic function used to transform $\mathbf{X}$ to $\mathbf{Y}$	
s	Sample, used as a subscript to identify sample elements	
U	Population	
μ_y	Population mean of Y	
$\hat{\mu}_y^{k,\text{st}}$	kNN estimator (est) of the population mean of Y	
p	The number of auxiliary variables	
SRS	Simple random sampling	
CLU	Cluster sampling	
m	Size of a cluster (number of elements)	
M	The number of clusters in the population	
$\Gamma_k^{U_s}(i)$	The set of k-nearest neighbors in the reference set (sample) to the ith population units	
$\Gamma_k^{U}(i)$	The set of k-nearest neighbors in the population to the ith population units	
w_{ij}	Weight assigned to the jth nearest neighbor to the ith population element (j = 1, …., k)	(3)–(4)
$\hat{\Delta}_i'$	Bias adjustment (estimated linear effect of substituting the k-nearest reference units with the k-nearest population units	
$\mathbf{1}_k$	Row-vector of k ones	
$\hat{\boldsymbol{\beta}}_p$	Row-vector of p-weighted linear regression coefficients	(3)–(4)
$\mathbf{Z}_s$	An (n m) × (p + 1) matrix resulting from a left-concatenation of $\mathbf{X}_s$ with a vector of ones (intercept)	(4)
$\mathbf{W}_j$	An (n m) × (n m) diagonal matrix with elements $w_{11}, \ldots, w_{n\times m}$	(4)
$K(u)$	Kernel density estimator of a standardized random variable (u)	(5)
ϑ_v	Kernel bandwidth for variable v	(5)–(6)
DM	Distribution matching	
$\tilde{F}^{k,\text{est}}(y)$	Distribution function of $\tilde{y}_i^{k,\text{est}}$	
$\hat{F}_s$	A sample-based estimator of the distribution function of Y	
$\hat{y}_{(j)}^{k,\text{est}}$	A DM estimate of $\tilde{y}_{(j)}^{k,\text{est}}$, ($j$) is the order of j, j = 1, …, N	
$q_p^{t,\nu}$	The (100p)th sample percentile of a t-distribution with ν degrees of freedom	
locreg	A locally weighted linear least squares regression estimator	
RMSE$^{k,\text{est}}$	Root mean squared error of a kNN estimator (est) with k-nearest neighbor imputation	
T^2	Hotelling's T-squared statistics	
Bias	The difference between an estimate and its true value	

Appendix 2. Populations (POP1, POP2 and POP3)

In POP1, *Y1, Y2*, and *Y3* were marginally distributed as a 25:50:25 mixture of three two-parameter distributions. Two-parameter gamma parameters with parameters (10, 8), (30, 12), and (50, 16) were used for *Y1*. Three non-central chi-squared distributions with parameters (4.5, 0.2), (9, 0.5), and (14, 1) for *Y2* and three gamma distributions with parameters (8, 200), (4, 200), and (2, 200) for *Y3* were used. The marginal distributions of *X1*, *X2*, and *X3* were 50:50 mixtures of two triangular distributions with parameters (min, max, mode) of (10, 50, 30) and (20, 60, 50) for *X1*, (40, 100, 70) and (50, 110, 100) for *X2*, and (80, 120, 110) and (90, 130, 120) for *X3*.

In POP2, *Y1, Y2*, and *Y3* were marginally distributed as left-truncated skew-normal distributions (Azzalini 1985) with parameters (300, 400, 0.2), (25, 30, 0.1), and (600, 500, 1), respectively. The left-truncation was fixed at y_{trunc} so that $P(y \leq y_{\text{trunc}}) = 0.10$ in the non-truncated skew-normal distribution. Marginal distributions of the four *X* variables in POP2 were PERT-distributions (a scaled beta distribution, Fazar 1959) on the interval [0, 256] with parameters (175, 2) for *X1*, (125, 3) for *X2*, (75, 2) for *X3*, and (25, 3) for *X4*.

In POP3, *Y1, Y2*, and *Y3* had marginally uniform distributions on the intervals (0, 80), (0, 40), and (0, 4000). The *X* variables were marginally distributed as triangular distributions on the interval (0, 256) with modes at 175 (*X1*), 125 (*X2*), 75 (*X3*), and 25 (*X4*).

The target pair-wise correlation coefficients among the variables in the three populations are in Table 1. Generation of the 8000 multivariate correlated random variables was done using the copula technique with a multivariate Gaussian copula defined by the target correlation structure (Srinivas et al. 2006; Fischer 2010). Achieved correlations may deviate by as much as ≤0.02 from their target value as the target correlations may have violated the Fréchet bounds.

A cluster structure with clusters of size (m) was imposed on the three populations by (1) adding a uniform-distributed [0,1] random variable (u) to each populations; (2) specifying a target correlation ρ between u and the X and Y variables in a population; (3) sorting the population elements on their u values; (4) adding an element identifier variable ω (ω = 1, …, N) to the sorted population values; and (5) adding a cluster identifier γ (γ = 1, …, M) defined as $\lceil \omega \times m^{-1} \rceil$ where $\lceil x \rceil$ is the smallest integer larger than or equal to x. In POP1 ρ was fixed at 0.4, resulting in an intra-cluster correlation coefficient (ρ_{clu}) (Cochran 1977, p. 209) that varied between 0.12 (*Y1*) and 0.14 (*Y2*). In POP2 ρ was 0.5, which generated a ρ_{clu} of 0.24 (*Y1*), 0.25 (*Y2*), and 0.26 (*Y3*). A weaker ρ of 0.22 was set for POP3. It gave a ρ_{clu} between 0.03 (*Y1*) and 0.05 (*Y3*). The achieved values of ρ_{clu} are in line with reported values for forest inventory cluster plots Magnussen and Köhl (2006).