2009

The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases

Bianca N. I. Eskelson
Oregon State University, bianca.eskelson@oregonstate.edu

Hailemariam Temesgen
Oregon State University, hailemariam.temesgen@oregonstate.edu

Valerie Lemay
University of British Columbia, Valerie.LeMay@ubc.ca

Tara M. Barrett
Pacific Northwest Research Station, tbarrett@fs.fed.us

Nicholas L. Crookston
Rocky Mountain Research Station, ncrookston@fs.fed.us

See next page for additional authors

Scandinavian Journal of Forest Research, 2009; 24: 235–246

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases

BIANCA N. I. ESKELSON[1], HAILEMARIAM TEMESGEN[1], VALERIE LEMAY[2], TARA M. BARRETT[3], NICHOLAS L. CROOKSTON[4] & ANDREW T. HUDAK[4]

[1]*Department of Forest Engineering, Resources and Management, Oregon State University, Corvallis, Oregon, USA,* [2]*Department of Forest Resources, University of British Columbia, Vancouver, Canada,* [3]*Pacific Northwest Research Station, USDA Forest Service, Anchorage, Alaska, USA,* [4]*Rocky Mountain Research Station, USDA Forest Service, Moscow, Idaho, USA*

Abstract

Almost universally, forest inventory and monitoring databases are incomplete, ranging from missing data for only a few records and a few variables, common for small land areas, to missing data for many observations and many variables, common for large land areas. For a wide variety of applications, nearest neighbor (NN) imputation methods have been developed to fill in observations of variables that are missing on some records (*Y*-variables), using related variables that are available for all records (*X*-variables). This review attempts to summarize the advantages and weaknesses of NN imputation methods and to give an overview of the NN approaches that have most commonly been used. It also discusses some of the challenges of NN imputation methods. The inclusion of NN imputation methods into standard software packages and the use of consistent notation may improve further development of NN imputation methods. Using *X*-variables from different data sources provides promising results, but raises the issue of spatial and temporal registration errors. Quantitative measures of the contribution of individual *X*-variables to the accuracy of imputing the *Y*-variables are needed. In addition, further research is warranted to verify statistical properties, modify methods to improve statistical properties, and provide variance estimators.

Keywords: *Consistent notation, forest measurements, input data for forest planning, nearest neighbor imputation, registration error, sources of X-variables.*

Introduction

Planning for sustainable forests has increased the demand for information. Management decisions are rarely based on single objectives, and hence, managing forested landscapes requires information to support several forest management goals such as timber production, wildlife habitat, fire hazard mitigation, biodiversity and carbon balance (Temesgen et al., 2007). Timely and accurate information about the entire forest resource is needed. In order to estimate forest characteristics of large areas at a more reasonable cost, nearest neighbor (NN) imputation approaches have been developed that use spatially comprehensive, inexpensive data that are available for all units, along with expensive, sparse data that are only available on a sample of units to provide detailed information for every unit in the forest management area.

Imputation is a procedure that is used to fill in missing values by using substitutes. These substitutes can be constructed with the aid of a statistical prediction mechanism such as a regression model, they can be values that have been observed for records that have similar characteristics as the missing records (e.g. NN imputation) or they can be values that were constructed by expert knowledge (Särndal & Lundström, 2005).

Regression is commonly used in forestry to fill in missing values (e.g. missing heights for some

Correspondence: H. Temesgen, Department of Forest Engineering, Resources and Management, Oregon State University, 237 Peavy Hall, Corvallis, OR 97331-5703, USA. E-mail: hailemariam.temesgen@oregonstate.edu

(Received 25 August 2008; accepted 4 March 2009)

ISSN 0282-7581 print/ISSN 1651-1891 online
DOI: 10.1080/02827580902870490

trees in the database). Mean imputation and ratio imputation are special cases of regression (Särndal & Lundström, 2005). Regression distorts the marginal distributions and measures of covariation of the completed data set, which is especially troubling when the tails of the distribution or standard errors of the estimates are being examined (Little & Rubin, 2002). Regression can make use of many continuous and categorical variables. However, its performance is sensitive to model misspecifications. If the regression model is not accurate, the resulting estimates will be poor.

NN imputation approaches are donor-based methods where the imputed value is either a value that was actually measured for another record in a database or the average of measured values from more than one record. These donors can be determined in a variety of ways. In the context of forestry data, forest attributes that are measured on all units of the population are referred to as *X*-variables. The *Y*-variables are those forest attributes that are only measured on a sample of units. Usually, the *Y*-variables are more expensive to measure and sparse, whereas the *X*-variables are less expensive and spatially comprehensive. Database records with measured *X*- and *Y*-variables are called reference records and target records are those that only have *X*-variables measured. Missing values of *X*-variables are not allowed in the sets of target or reference records and missing values of *Y*-variables are not allowed in the set of reference records (Crookston & Finley, 2008). The idea that motivates NN imputation methods is that two records whose *X*-values are similar should also have *Y*-values that are similar (Särndal & Lundström, 2005).

For forest inventory applications, often a large number of units in a forest area are missing values for some variables that may be critical to management planning. The first forest inventory applications of NN methods based on remotely sensed data were presented by Kilkki and Päivinen (1987). Since then, NN methods using remotely sensed data have been widely used for forest inventory databases, most notably for the Finnish national forest inventory (Tomppo, 1991). In these applications, remotely sensed data were used to provide *X*-variables for every unit in the landbase. For a subset of units, the *Y*-variables, often measured on ground plots, are also available. NN imputation methods are used to impute vectors of *Y*-variables to database records missing these variables. The widespread availability of satellite and other remotely sensed imagery as a source of *X*-variables has increased the relevance of imputation methods.

NN imputation techniques use either one single neighbor ($k = 1$) as donor for the missing *Y*-variables of the target records (e.g. Moeur & Stage, 1995) or a simple or a weighted average of $k > 1$ near neighbors to fill in the missing *Y*-variables (e.g. Korhonen & Kangas, 1997; Maltamo & Kangas, 1998). The weights are chosen to reflect the degree of similarity in the *X*-variables between the *i*th target record and *j*th reference record. For example, the inverse of the distance metric that indicates similarity in the *X*-variables between target and reference records could be used to weight the averages (LeMay & Temesgen, 2005*a*). However, the choice of the weight function can also be guided by subject knowledge, prior beliefs, ecology, spatial distance and statistical considerations (Köhl et al., 2006).

NN imputation methods are non-parametric or distribution-free in that they do not rely on any underlying probability distribution for estimation (Everitt, 1998). Since NN imputation techniques can be used to estimate more than one *Y*-variable at once, they are multivariate methods. For example, LeMay and Temesgen (2005*a*) estimated basal area and stems per hectare using aerial auxiliary variables. In another study, Temesgen et al. (2003) imputed trees sizes and stems per hectare for seven tree species from aerial attributes of complex stands in south-eastern British Columbia. NN imputation methods have also been used to estimate the type and frequency of regeneration (Hassani et al., 2004), the number of snags and cavity trees (Temesgen et al., 2008; Eskelson, 2008), and status and change of forest attributes from paneled inventory data (Eskelson, 2008).

While NN imputation methods are an active area of research, a comprehensive review of the approaches used in forestry for filling in missing data and a discussion of their advantages and weaknesses are lacking. Potential users of NN imputation methods might prefer the use of regression techniques to NN imputation methods owing to voluminous literature on regression techniques and their ease of application. The intention of this review is to point out the advantages and weaknesses of NN imputation methods for filling forestry databases and to motivate additional research and improvements of these methods. The main objectives are to provide an overview of the variety of methods that have been used for this purpose, and list obstacles that await NN imputation researchers and users.

Advantages of nearest neighbor methods for imputation

In a forest inventory context, missing data can occur where a few inventory plots out of the sample could not be measured owing to, for example, hazardous or environmental conditions, shortage of resources and

time, or lack of access to privately owned lands (McRoberts, 2003). The easiest way to deal with missing data is to delete all observations with missing data and analyze the remaining data as a complete data set. This is known as complete-case analysis (Little & Rubin, 2002). Simplicity and comparability of univariate statistics are the advantages of complete-case analysis. However, this approach reduces the sample size and results in loss of information due to discarding incomplete cases. This can cause loss of precision and potentially bias when the complete cases are not a random sample of the population (Little & Rubin, 2002). Instead, imputation methods can be used to supply missing observations to complete a data set. When the resulting data set is used for analysis, all available information is used and no loss of information occurs.

Another common use of imputation methods in forest inventory occurs when variables of critical interest to forest management are very expensive to measure and are only available for a subset of units in the forest land areas. For example, volume or biomass per hectare, tree-size distributions and other variables are measured via very expensive ground sampling. However, remotely sensed information that is related to the variables of interest can be less costly to acquire and may be available at all locations. Imputation is therefore used to associate expensive but sparse data with inexpensive and spatially comprehensive data to obtain more accurate estimates of critical information.

One of the major advantages of NN imputation methods is that they retain the complex variance–covariance structure and natural variation of the *Y*-variables as long as $k=1$ (Moeur & Stage, 1995; Ek et al., 1997; Korhonen & Kangas, 1997; Holmström & Fransson, 2003; McRoberts, 2009). In regression approaches, in contrast, *Y*-variables are often estimated separately, which may lead to estimates with unreasonable relationships and a variance–covariance structure that differs greatly from the original field data (Moeur & Stage, 1995; Tuominen et al., 2003). When variables are estimated separately, the dependence structure among the *Y*-variables is generally lost (Tomppo et al., 2008). Hence, the multivariate aspect of the NN methods is crucial, especially for inventory applications where information on multiple stand attributes is frequently required for stand management decisions (McRoberts, 2008). As long as a single neighbor ($k=1$) is used as a donor, illogical relationships among imputed attributes are impossible, and the relationships of the imputed *Y*-variables will always be within the bounds of biological reality (Moeur & Stage, 1995; LeMay & Temesgen, 2005*a*). For example, density (trees per unit area) and average tree size are related, and there are certain combinations of density and average tree size that do not occur in nature. NN imputation using a single neighbor will always result in values for imputed records that retain the logical relationship between density and tree size, because these values are imputed from another record with observed values.

Misspecified regression models or the use of models outside the range of the modeling data may result in unreasonable estimates. In NN imputation methods, the magnitude of the most extreme estimate is limited to the most extreme reference observation. Therefore, NN imputation methods using $k=1$ do not extrapolate outside the range of sampled conditions (e.g. no high elevation stand in the field sample) (Moeur & Stage, 1995). NN imputation methods behave more like regression as k increases, however. In NN imputation, the only assumption is that the *X*-variables have a strong relationship to the *Y*-variables and can therefore be used to impute missing *Y*-variables. NN imputation can employ *X*-variables without a complete knowledge of the complicated relationships between *X*- and *Y*-variables (Fehrmann et al., 2008).

As noted, NN imputation methods are non-parametric. Temesgen et al. (2003) asserted that non-parametric NN imputation methods may provide better estimates of tree-lists for complex stands with multiple species and a wide variety of tree sizes. The diameter distributions for these stands tend to be multimodal, and are not easily represented by probability distributions.

Types of nearest neighbor imputation methods used in forestry

Distance metrics

NN imputation methods use different distance metrics to determine the similarity between target and reference records. Typically, the distance metrics are based on absolute differences, Euclidean or Mahalanobis distance functions (Maltamo et al., 2003). Absolute differences are calculated as:

$$d_{ij} = \sum_{l=1}^{p} c_l |x_{il} - x_{jl}| \quad (1)$$

where x_{il} is the value of the *X*-variable l for target record i, x_{jl} is the value of the *X*-variable l for reference record j, p is the number of *X*-variables, and c_l is the coefficient for variable x_l. The distance metrics most widely used for NN imputation are of the quadratic form (Stage & Crookston, 2007):

$$d_{ij}^2 = (x_i - x_j)W(x_i - x_j)' \quad (2)$$

where x_i is the $(1 \times p)$ vector of x-variables for the ith target record, x_j is the $(1 \times p)$ vector of x-variables for the jth reference record, and W is a $(p \times p)$ symmetric matrix of weights.

For the squared Euclidean distance the weight matrix, W, is the diagonal identity matrix, giving equal weight to each X-variable. The squared Euclidean distance gives more emphasis to larger differences than the absolute difference distance (eq. 1) because the differences are squared (LeMay & Temesgen, 2005*a*). The Mahalanobis distance is produced by using the inverse covariance matrix of the X-variables for W (Stage & Crookston, 2007). In the most similar neighbor (MSN) procedure (Moeur & Stage, 1995), W is derived from canonical correlation analysis. The relationships between X- and Y-variables are used and stronger correlations result in higher weights for a particular X (LeMay & Temesgen, 2005*a*). Moeur and Stage (1995) derived W from canonical correlation analysis, while Ohmann and Gregory (2002) derived W from canonical correspondence analysis for their gradient nearest neighbor (GNN) procedure. Some other distance metrics used are a modified Minkowski distance (Fehrmann et al., 2008), a regression transform distance (Holmström et al., 2001), fuzzy distance (Maselli, 2001; Chirici et al., 2008), a distance modified by a multiple regression method (Maselli et al., 2005; Chirici et al., 2008) and a distance modified by the use of non-parametric weights (Maselli et al., 2005; Chirici et al., 2008). In addition to these distance metrics, Crookston and Finley (2008) used a proximity matrix obtained from multiple classification and regression trees (see e.g. Breiman, 2001, for details) in their "randomForest" method to determine the similarity between target and reference records.

Stage and Crookston (2002) found that the addition of the linear correlations between the Y- and X-variables does not always alter the selection of neighbors and, therefore, may not improve the precision of imputed values. Including linear correlations in the imputation process when there is a perfect unknown, but non-linear, relationship between X- and Y-variables would degrade the matches. However, a good match on the X-variables results in a good match on the Y-variables if the relationship between X- and Y-variables is strong. The results depend on the strength of the relationship between X- and Y-variables, but may be confounded by the choice of the distance metric and the proportion of reference records with full information (LeMay & Temesgen, 2005*a*; Temesgen et al., 2008). The choice of a particular distance metric may depend on the relation of the Y-variables to the X-variables (Stage & Crookston, 2002, 2007).

Although many applications of NN imputation focus on the use of continuous variables, categorical variables can also be used as X-variables. Crookston et al. (2002) and Maltamo et al. (2006) created dummy variables for categorical data. For imputations using categorical variables or a mixture of continuous and categorical variables, LeMay and Temesgen (2005*a*) suggested using the City Block distance (Dillon & Goldstein, 1984) by enumerating the number of matches for class data or the generalized distance for discrete variables (Kurczynski, 1970). In the "randomForest" method (Crookston & Finley, 2008), the variables can be a mixture of continuous and categorical variables.

The distance metric used in the MSN procedure can be locally adapted to improve regional and local imputation results. Local adaption can be performed by first using the distance metric of the MSN procedure to select the local neighborhood and then using this local neighborhood to calculate a new weight matrix W. The final imputation is then performed by using the local W and local reference data. Another way to perform local adaption is to select a combination of neighbors from the neighborhood where the average of the X-variables is closest to the target record X-variables (Maltamo et al., 2003; Malinen, 2003).

Number of neighbors (k)

LeMay and Temesgen (2005*a*) compared the use of the nearest neighbor, the average of three near neighbors and the distance-weighted average of three near neighbors. They found that the estimates may not be within the bounds of reality if more than one neighbor is used. All variability that exists in the observations is preserved when $k = 1$, whereas $k > 1$ results in smoothing, since estimates are based on averages of multiple observations (McRoberts et al., 2002).

With small k values, NN methods may produce results that are less accurate than using the mean over all observations for every prediction (McRoberts et al., 2002). The accuracy of the estimates improves with increasing k to an optimal choice of k. When a large number of reference records is available in the database, larger values of k can be applied (Tuominen et al., 2003; LeMay & Temesgen, 2005*a*). However, the estimation precision for extreme values of Y-variables increases with an increase in k (McRoberts et al., 2002). This is known as the classic bias/variance dilemma, which complicates the use of non-parametric methods (Malinen, 2003).

The optimal choice of k, the distance metric including weights, and X-variables is difficult to determine (LeMay & Temesgen, 2005*a*). Muinonen et al. (2001) found that increasing the number of similar neighbors beyond $k=3$ did not improve the accuracy. In other studies for imputation of tree-level variables, the optimal k was found to be larger than 10 (Sironen et al., 2001, 2003), because of a large number of available reference records. The best combination depends on the problem and the available data (Malinen, 2003) and the optimal value for k is a trade-off between the accuracy of the estimates and the variation that is retained in the estimates (McRoberts et al., 2002; Tuominen et al., 2003). The strength of the relationship between the X- and Y-variables inversely affects the optimal value of k, with weaker relationships resulting in larger k values. McRoberts et al. (2002) suggested using an objective criterion for choosing k. Malinen (2003) found the optimal value of k that minimizes the root mean square error of certain characteristics could be determined, and Tomppo and Halme (2004) developed an algorithm to determine the optimal weights.

Potential sources and choice of X-variables

The X-variables can come from easily measured ground variables (e.g. Ek et al., 1997; Korhonen & Kangas, 1997; Hanus et al., 1998; Hassani et al., 2004), remotely sensed data (e.g. McRoberts et al., 2002, 2006; Holmström & Fransson, 2003; Tomppo et al., 2008; McRoberts, 2008), existing stand records such as age, site index, silvicultural stand history data, terrain data (i.e. slope, aspect, elevation) (e.g. Temesgen et al., 2003), environmental data (Ohmann & Gregory, 2002; Holmström & Fransson, 2003) or combinations of data sources (e.g. Hudak et al., 2002, 2008*a*; LeMay et al., 2008; Packalén & Maltamo, 2008). The use of visually interpreted aerial photograph data was found to be superior to the use of digital aerial photograph features by Tuominen et al. (2003).

The resolution of the X-variables in terms of spatial extent varies for each medium. Photographs are often very detailed, but then frequently are reduced to polygons (e.g. stands) via interpretation of the images. For other remotely sensed media, often the data are gathered by pixel, and pixel size varies with type of remotely sensed imagery and with wavelength. Ground data are often gathered in plots, where the spatial extent is the plot size. The use of these different sizes of reference records in imputation affects the spatial resolutions of the imputed data. For example, imputing field data to each pixel of Landsat data provides a spatially continuous set of grid data (e.g. McRoberts et al., 2002; Ohmann & Gregory, 2002). Alternatively, using interpreted photographs, field data are imputed to polygons to provide a spatially continuous set of vector data (e.g. Moeur & Stage, 1995; LeMay & Temesgen, 2005*a*). Detailed reference plot information can be imputed to target plots lacking detailed information (e.g. McRoberts, 2001; Hassani et al., 2004) which, if spatially represented, would most typically be shown as discontinuous areas.

A fairly recent remote sensing technology with rapidly emerging utility for forestry applications is light detection and ranging (lidar) (Næsset et al., 2004; Reutebuch et al., 2005). Lidar systems have the ability to measure directly the three-dimensional structure of imaged areas. Subsequent processing of the three-dimensional lidar point clouds can be used to separate biophysical data (measurements of aboveground vegetation) from geophysical data (measurements of the terrain surface) (Evans & Hudak, 2007). Thus, accurate measures of both ground height and canopy height can be derived, as well as useful information on the intervening canopy layers (Reutebuch et al., 2005; Hudak et al., 2006). The potential of lidar data for predicting fundamental forest attributes such as plot-level basal area and tree density has been demonstrated using multiple linear regression (Hudak et al., 2006) and imputation approaches (Hudak et al., 2008*b*). Volume of forest stands has successfully been estimated with lidar-assisted ratio estimation (Corona & Fattorini, 2008) and NN imputation methods (Maltamo et al., 2006). Using independent stand inventory data, Hudak et al. (2008*a*) reported that imputation methods resulted in smaller average differences between observed and imputed values than those found using regression models.

Recognizing that different remotely sensed technologies sense different aspects of forest structure and that no single technology can provide all useful and relevant information, the integration of data from different remote sensors is worthwhile (Hudak et al., 2002; LeMay et al., 2008). Landsat imagery is useful for characterizing the spatial extent and seasonal phenology of forest stands across a landscape, but is less sensitive to canopy height variation. Lidar accurately measures canopy height, but usually has much more limited coverage and is relatively insensitive to vegetation phenology. Polygon imputation has been commonly applied in forest management, in part owing to the inability of 30 m Landsat image pixels to capture canopy structure variation at a finer scale. The high spatial density of lidar data allows the variable structure of forest canopies to be mapped within polygons, improving estimates of within-stand heterogeneity.

As well as a variety of *X*-variables and their transformations (Temesgen et al., 2008), the ranges for the *X*- and *Y*-variables will affect imputation accuracy. The reference records must be well distributed over the ranges of variability in *X*-variables for efficient and unbiased NN imputation. Because NN imputation methods neither extrapolate values outside the range of the reference data (Moeur & Stage, 1995; Holmström & Fransson, 2003) nor interpolate when $k=1$ (Crookston et al., 2002, p. 24), the set of reference records needs to consist of a representative sample that covers the complete joint ranges of values of the *X*-variables without large gaps. For details see McRoberts (2009). If there are several *Y*-variables or "rare" target records that are not represented in the reference records, then a good match will not be possible (McRoberts et al., 2002; Temesgen et al., 2003). The required sampling proportion differs based on the complexity of stands (Hassani et al., 2004; LeMay & Temesgen, 2005*a*).

Canonical correlation analysis, used in the MSN procedure, requires that the relationships between *Y*- and *X*-variables collectively can be described by a linear combination and correlations among the linear combinations of *X*- and *Y*-variables need to be known. Hence, the choice of variables and adequate transformations are important. Maltamo et al. (2003) used second powers of some independent variables that resulted in more linear relationships to improve the results. This may give biased results if transformations only create a small window with a linear relationship (Korhonen & Kangas, 1997). The *X*-variable selection algorithms presented by Maltamo et al. (2006) and Packalén and Maltamo (2007) include tests of each *X*-variable as well as the transformations $\ln(x)$, $\mathrm{sqrt}(x)$, x^2 and $\mathrm{inv}(x)$ to find transformations that best improve the relationship between *X*- and *Y*-variables.

The choice of *X*-variables depends on the information that is available and on the variables related to the *Y*-variables (LeMay & Temesgen, 2005*a*). Increasing the number of *X*-variables does not guarantee improvement in the estimation results (McRoberts et al., 2002). As the number of *X*-variables increases, it becomes increasingly difficult to find relevant neighbors (Maltamo et al., 2006). The selection of an appropriate set of *X*-variables has been found to be a very laborious and time-consuming task and should therefore be automated (Maltamo et al., 2006; Packalén & Maltamo, 2006). Packalén and Maltamo (2007) presented a heuristic *X*-variable selection algorithm that minimizes the weighted average of relative root mean square errors. The weight matrix W (see eq. 2) defines the number and choice of *X*-variables. Tomppo and Halme (2004) used a genetic algorithm to select optimal weights of the *X*-variables for predicting continuous forest attribute variables. Tomppo et al. (2009) modified the genetic algorithm to optimize the weights of the *X*-variables for predicting categorical variables. Walter et al. (2008) presented a non-linear optimization routine that converges on values for W that minimizes the root mean square error.

Critical challenges for imputation methods

Forest resource managers in all parts of the world are faced with a myriad of increasingly complex decision problems. The intensity of these problems is compounded by missing or inadequate data. As a result, developing, testing and improving NN methods are active areas of current research. In the authors' view, the most critical challenges for imputation methods and areas that warrant further research or need to be clarified to improve and facilitate NN applications in forest planning and management include:

- developing consistent notation
- evaluating statistical properties and recommending new estimators, including variance estimators
- evaluating and improving imputation accuracy
- combining data sources.

This list of challenges is not exhaustive. For example, the use of NN imputation techniques for either design-based or model-based inference, the need for efficient techniques, small area applications, and the need for developing flexible and comprehensive tools to visualize imputation results are not specifically discussed in this article.

Consistent notation

The choice of *X*- and *Y*-variables, the distance metric and k contribute to the imputation error (Stage & Crookston, 2007). Differences in data structure, selection of *Y*-variables and availability of *X*-variables suggest that no single choice of distance metric, *X*- and *Y*-variables and k gives the best results for all applications. Hence, these choices must be determined on a case-by-case basis. Consistent notation and methods to evaluate results of the imputation would help in making these choices.

Currently, the notation for imputation is not consistent among scientists and practitioners. For example, NN imputation methods are referred to as "near-neighbor" methods (Stage & Crookston, 2007), "nearest-neighbor" methods (e.g. Fehrmann et al., 2008; Sironen et al., 2008), "non-parametric regression" (Altman, 1992) and "k-NN regression"

or "NN regression" (Korhonen & Kangas, 1997; Tommola et al., 1999; Maltamo & Eerikäinen, 2001). The reference data set is also called training data set (e.g. Fehrmann et al., 2008) and the distance metric is sometimes referred to as the similarity function (e.g. Malinen, 2003). The *X*-variables are also called predictor variables (Hudak et al., 2008*b*), explanatory variables (Ohmann & Gregory, 2002; Fehrmann et al., 2008), independent variables (Korhonen & Kangas, 1997; Maltamo et al., 2003), carrier data (Holmström et al., 2001; Barth et al., 2009) or indicator attributes (Moeur & Stage, 1995). Chirici et al. (2008) referred to *X*-variables derived from remotely sensed data as feature space variables and to those *X*-variables that were not derived from remotely sensed data as ancillary variables. The *Y*-variables are also referred to as dependent variables (Korhonen & Kangas, 1997; Maltamo et al., 2003) and design attributes (Moeur & Stage, 1995).

While some of the mentioned inconsistencies appear minor, they can result in confusion in communicating and comparing methods and results. Some terminology, for example the use of independent and dependent variables or the term "NN regression", may make it difficult to distinguish between regression and NN imputation methods. To advance imputation methods and communicate effectively, especially with practitioners who might not be very familiar with the ongoing research and terminology, a common vocabulary for different imputation methods and approaches is needed.

Statistical properties and new estimators

The statistical foundation for imputation methods is not well developed. In general, estimation techniques are chosen based on statistical properties such as unbiasedness, consistency and efficiency. These properties are not well understood for NN imputation. In addition, new estimators are being proposed that will alter these properties in the future.

The biasedness of the NN estimators has been considered as the most serious drawback of NN methods by some authors (Korhonen & Kangas, 1997), which may make it hard to justify the use of NN imputation over traditional regression techniques. Non-parametric methods tend to be highly biased at the edge of the data cloud because targets will likely be paired with a more central point owing to the asymmetric neighborhood (McRoberts et al., 2002). Extremely small values and extremely high values will be overestimated and underestimated, respectively, if the reference data do not cover the whole range of variability (Packalén & Maltamo, 2007). Bias can also be a problem in the interior of the data cloud if the *X*-variables are non-uniformly distributed (Maltamo et al., 2003; Stage & Crookston, 2007). Also, since the estimates of parameters do not necessarily approach their true values with an increase in size of the reference data set, NN methods are not statistically consistent (Maltamo & Kangas, 1998).

For forest inventory applications, it is important to be familiar with the mechanism that led to missing data since this may affect the range of variability in the reference data set. If the probability of missing records is unrelated to any measured or unmeasured characteristic, then the data are missing completely at random (Little & Rubin, 2002). Reference data are likely to include wide ranges of *Y*- and *X*-variables. However, in the case of missing ground plot information, missing data may be a result of access issues, such as steep terrain. Since *Y*- and *X*-variables for those ground plots may be quite different, they may be well outside the ranges of *Y*- and *X*-variables in the reference data set, resulting in poorer imputation results.

Some very recent papers have recommended new estimators. McRoberts et al. (2007) suggested a variance estimator for area of interest estimates obtained from NN imputation that incorporates spatial correlation. Magnussen et al. (2009) presented model-based estimators of the uncertainty of pixel-level and areal NN predictions, while Baffetta et al. (2009) recommended the use of a design-based approach to derive the statistical properties of the NN estimators. These represent the first attempts to derive estimates of precision for NN methods and, hence, further investigations are warranted.

Accuracy evaluation and improvements

One of the most critical challenges with imputation methods in forestry is that they often have been used to develop data sets that are of interest to resource specialists not directly involved in filling the missing data. In these situations, there is a risk that users of imputed data will not understand the sources and level of error in the data. This problem is exacerbated because the imputed data are often of high resolution and detail, potentially leading third parties to misunderstand appropriate uses for the data. Estimation of uncertainty associated with imputation is necessary for understanding appropriate uses.

For users, it is important to know that the errors from imputation differ from those of regression-based estimates, in that imputation includes different error components (Stage & Crookston, 2007). To maximize imputation accuracy, it is crucial to

understand the sources of errors in imputation, which include:

- measurement errors in the *Y*-variables (e.g. species identification errors in ground plots) as with regression analysis, which is controllable and should be minimized;
- pure error as with regression analysis, which depends on the choice of *X*- and *Y*-variables as well as the choice of useful transformations that can improve the representation of the relationship between *X*- and *Y*-variables (Temesgen et al., 2008). Pure error arises, for example, when *X*-variables that would improve the imputation are omitted or when there is a lack of accurate registration between the locations of *Y*-variables and *X*-variables;
- the availability and similarity of reference records to target records, affecting their applicability as donor records;
- the choice of k and their relative weights (Stage & Crookston, 2007).

Other important sources of imputation errors in forestry applications are temporal registration errors resulting from differences in times of measurement of the *X*- and *Y*-variables, spatial registration errors resulting from inaccurate spatial matching of measures for *Y*- and *X*-variables, and spatial resolution errors error due to different spatial extents for measures of the *X*- and *Y*-variables. For example, *X*-variables may be measured on a pixel that does not match in size to the ground plot on which the *Y*-variables are measured.

To determine whether a given imputation method provides satisfactory results in filling databases, information is needed concerning how accurate the imputation needs to be, how well the dependencies of *Y*-variables need to be maintained, and how well key aspects of the environment need to be captured. However, these are among the most poorly quantified issues in using NN imputation in forestry, and the required accuracy may differ between users.

Despite the need to quantify the uncertainty of predicted values, a good measure of uncertainty (goodness of imputation estimate) is still lacking. Stage and Crookston (2007) used root mean square error, which they termed mean square difference to emphasize the unique error properties of this statistic, to measure how well the imputations match for reference records. McRoberts (2009) pointed out that root mean squared error may not be a good measure of accuracy when *Y*-variables have heteroscedacity variances around the *X*-variables. However, for a single, continuous *Y*-variable, he proposed graphical tools to evaluate issues of bias, homoscedasticity, influential observations, outliers and extrapolations. The development of diagnostic tools for multiple continuous variables and for categorical variables for NN techniques is still warranted. An approach that uses a model of the *X*-variable space variogram to quantify prediction uncertainty was proposed by Kim and Tomppo (2006), but is computationally demanding. Relevant accuracy statistics for assessing the quality of predictions of categorical variables are still lacking (Tomppo et al., 2009).

The exploration of alternative imputation approaches is made easier by providing consistent measures of the quality of imputation (Stage & Crookston, 2002). Useful techniques for diagnosing whether one distance metric performs better than another are discussed in Crookston and Finley (2008). The inclusion of NN imputation methods into standard software packages could facilitate the comparison of different NN approaches. The recently developed yaImpute R package (Crookston & Finley, 2008) is an example of such an endeavor.

One possibility to enhance imputation performance is to use a locally adaptable MSN method (Malinen, 2003; Maltamo et al., 2003). Localization can also be achieved by using spatial coordinates as *X*-variables or by restricting the selection of neighbors to a circular area around the target unit (Sironen et al., 2008). Barth et al. (2009) developed a method that maintains what they termed "spatial consistency", where natural variability within a local area is maintained. They argued that spatial consistency has become more important, since comparisons among alternative forest management scenarios have become more spatially explicit.

Combining data sources

Integration of multiple data sources and advanced technology is critical in meeting contemporary requirements for monitoring, assessment and resource analysis. This integration needs to include spatial and temporal information to describe and interpret vegetation layers, to detect changes and trends.

More often, NN imputation methods combine multiple sources of *X*-variables that match variables at multiple scales. Ground data from both overstory and understory vegetation need to be connected to remote sensing data such as aerial photography, lidar or satellite data as well as other sources of *X*-variables. Ground location errors between the paired *Y*- and *X*-variable records contribute to the pure error that is part of the imputation error (Stage & Crookston, 2007). Since matching of ground-measured and remotely sensed data is complicated by difficulties

in obtaining accurate locations on each data source and errors in spatial positioning (LeMay & Temesgen, 2005*b*), the spatial registration errors between data sources will be increased as more and more data sources are combined. Care should be taken that the different data sources are obtained at approximately the same time (Packalén & Maltamo, 2007) in order to reduce temporal error. The need to quantify the thematic and spatial accuracy of imputation techniques at various spatial and temporal scales will persist, towards the goal of minimizing co-registration errors between independent data sets.

Estimation accuracy has been found to improve when *X*-variables from a number of different sources have been used (e.g. Holmström & Fransson, 2003; Tuominen et al., 2003). Combining *X*-variables derived from lidar and aerial photographs improved the estimation of species-specific stand attributes in terms of accuracy when evaluated at the plot level (Maltamo et al., 2006; Packalén & Maltamo, 2007). However, a straightforward way of relating the contribution of an individual *X*-variable or a group of *X*-variables to the accuracy of the outcome is still lacking (Packalén & Maltamo, 2007).

The spatial extent of each record for the *X*-variables can be plot, pixel or polygon. Where the *X*-variables represent measures of spatially contiguous pixels or polygons for complete coverage of the forest area, imputation of the *Y*-variables for all records results in a spatially comprehensive data set that can be used to create maps of any attribute that can be created from either the *X*- or *Y*-variables. In forestry applications, field plot measures (i.e. ground measures) are often used to obtain the *Y*-variables. Conversely, where *X*-variables represent plots or only a subset of polygons and not a spatially contiguous data set, imputation is not intended for mapping purposes. Instead, plot-level imputation is used to fill in missing values for some variables in plots or to update inventory information to a common temporal reference and does not result in a spatially comprehensive data set. Spatial mismatches between *X*- and *Y*-variables are a problem in imputation, regardless of whether or not the *X*-variables represent a spatially contiguous set of data, as the spatial extents represented by *X*-variables often differ substantially in shape and size from field plots that are often used to provide the *Y*-variables.

A detailed comparison of imputation using *X*-variables at the pixel, polygon or plot spatial extent is still lacking. One important question is whether each of the imputation error sources (listed above, in the section *Accuracy evaluation and improvements*) contributes the same amount of error depending upon this spatial extent. Spatial registration errors between *Y*- and *X*-variables increase the pure error (Stage & Crookston, 2007), and may differ.

Conclusions

The problem of missing data is ubiquitous in forest inventory, monitoring and planning. NN imputation methods are increasingly being used for a wide variety of applications by combining spatially comprehensive data for the entire forested area with detailed information from a sample of stands, and are also being used to fill in missing variables at a plot or polygon level. When the purpose of imputation is to evaluate management options, it is important to preserve the complex relationships between the forest attributes being imputed. It is also important that the range of variability in each forest attribute of interest be represented across the management region.

The NN imputation methods currently applied in forestry practice differ in their choice of distance metrics, the number of nearest neighbors and their relative weights, potential sources of *X*-variables, and the level and scope of imputation. Automated approaches for choosing the number of nearest neighbors and the most appropriate set of *X*-variables need to be improved.

To take advantage of different technologies, the current trend is to use *X*-variables that were derived from different sources, e.g. aerial photographs, satellite data, lidar and stand records. This can cause additional error in spatial and temporal registration. Methods to minimize registration error and to relate the contribution of individual *X*-variables or groups of *X*-variables need to be developed.

NN imputation methods have a role in improving stochastic approaches and in validating assumptions used in forest planning, and help to meet forest management challenges at a range of spatial scales. Increasing the understanding of their strengths and weaknesses will help to ensure appropriate use. Further development and use of NN methods call for inclusion of these methods in standard software packages that make available common methods for defining and measuring the accuracy of the imputation results. Moreover, further research is warranted to mitigate the bias associated with NN methods, and to develop sound variance estimation procedures with specifications of the conditions under which they can be applied.

In selecting NN imputation methods, one needs to consider accuracy, objectivity, feasibility, robustness and simplicity. To have general utility, a selected approach needs to be transparent and reproducible.

Acknowledgements

We gratefully acknowledge the support provided by the Forest Inventory and Analysis program, Pacific Northwest Research Station, United States Forest Service. We also thank Drs Robert Monserud, Jeff Brandt, Helge Dzierzon, Hampus Holmström, and two anonymous reviewers for their insights and comments on an early draft.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*, 175–185.

Baffetta, F., Fattorini, L., Franceschi, S. & Corona, P. (2009). Design-based approach to *k*-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment*, *113*, 463–475.

Barth, A., Wallerman, J. & Ståhl, G. (2009). Spatially consistent nearest neighbor imputation of forest stand data. *Remote Sensing of Environment*, *113*, 546–553.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D., Maselli, F. & Bertini, R. (2008). Non-parametric and parametric methods using satellite images for estimating growing stock volume in alpine and Mediterranean forest ecosystems. *Remote Sensing of Environment*, *112*, 2686–2700.

Corona, P. & Fattorini, P. (2008). Area-base lidar-assisted estimation of forest standing volume. *Canadian Journal of Forest Research*, *38*, 2911–2916.

Crookston, N. L. & Finley, A. O. (2008). yaImpute: An R package for kNN imputation. *Journal of Statistical Software*, *23*(10), 1–16.

Crookston, N. L., Moeur, M. & Renner, D. (2002). *User's guide to the most similar neighbour imputation program version 2* (Gen. Tech. Rep. RMRS-96). US Department of Agriculture, Forest Service.

Dillon, W. R. & Goldstein, M. (1984). *Multivariate analysis*. Toronto: John Wiley & Sons.

Ek, A. R., Robinson, A. P., Ratdke, P. J. & Walters, D. K. (1997). Development and testing of regeneration imputation models for forests in Minnesota. *Forest Ecology and Management*, *94*, 129–140.

Eskelson, B. N. I. (2008). *Examination of imputation methods to estimate status and change of forest attributes from paneled inventory data* (Doctoral dissertation, Oregon State University). Retrieved January 1, 2009, from Oregon State University ScholarsArchive@OSU website: http://hdl.handle.net/1957/10021

Evans, J. S. & Hudak, A. T. (2007). A multiscale curvature algorithm for classifying discrete return lidar in forested environments. *IEEE Transactions on Geoscience and Remote Sensing*, *45*, 1029–1038.

Everitt, B. S. (1998). *The Cambridge dictionary of statistics*. Cambridge: Cambridge University Press.

Fehrmann, L., Lehtonen, A., Kleinn, C. & Tomppo, E. (2008). Comparison of linear and mixed-effect regression models and a *k*-nearest neighbour approach for estimation of single-tree biomass. *Canadian Journal Forest Research*, *38*, 1–9.

Hanus, M. L., Hann, D. W. & Marshall, D. D. (1998). Reconstructing the spatial patterns of trees from routine stand examination measurements. *Forest Science*, *44*, 125–133.

Hassani, B. T., LeMay, V., Marshall, P. L., Temesgen, H. & Zumrawi, A.-A. (2004). Regeneration imputation models for complex stands of southeastern British Columbia. *The Forestry Chronicle*, *80*, 271–278.

Holmström, H. & Fransson, J. E. S. (2003). Combining remotely sensed optical and radar data in kNN-estimation of forest variables. *Forest Science*, *49*, 409–418.

Holmström, H., Nilsson, M. & Ståhl, G. (2001). Simultaneous estimations of forest parameters using aerial photograph interpreted data and the k nearest neighbour method. *Scandinavian Journal of Forest Research*, *16*, 67–78.

Hudak, A. T., Lefsky, M. A., Cohen, W. B. & Berterretche, M. (2002). Integration of lidar and Landsat ETM+ data for estimating and mapping forest canopy height. *Remote Sensing of Environment*, *82*, 397–416.

Hudak, A. T., Crookston, N. L., Evans, J. S., Falkowski, M. J., Smith, A. M. S., Gessler, P. E. & Morgan, P. (2006). Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. *Canadian Journal of Remote Sensing*, *32*, 126–138.

Hudak, A. T., Evans, J. S., Crookston, N. L., Falkowski, M. J., Steigers, B., Taylor, R. & Hemingway, H. (2008*a*). Aggregating pixel-level basal area predictions derived from LiDAR data to industrial forest stands in Idaho. In R. N. Havis & N. L. Crookston (Eds.), *3rd Forest Vegetation Simulator Conference*: Proceedings RMRS-P-54 (pp. 133–146). Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Hudak, A. T., Crookston, N. L., Evans, J. S., Hall, D. E. & Falkowski, M. J. (2008*b*). Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment*, *112*, 2232–2245.

Kilkki, P. & Päivinen, R. (1987). Reference sample plots to compare field measurements and satellite data in forest inventory. In *Proceedings of Seminar on remote sensing-aided forest inventory*, Hyytiälä, Finland. Vol. 19. Department of Forest Mensuration and Management, Research Notes (pp. 209–215). Helsinki: University of Helsinki.

Kim, H.-J. & Tomppo, E. (2006). Model-based prediction error uncertainty estimation for *k*-nn method. *Remote Sensing of Environment*, *104*, 257–263.

Köhl, M., Magnussen, S. S. & Marchetti, M. (2006). *Sampling methods, remote sensing and GIS multiresource forest inventory.* Berlin: Springer.

Korhonen, K. T. & Kangas, A. (1997). Application of nearest-neighbour regression for generalizing sample tree information. *Scandinavian Journal of Forest Research*, *12*, 97–101.

Kurczynski, T. W. (1970). Generalized distance and discrete variables. *Biometrics*, *26*, 525–534.

LeMay, V. & Temesgen, H. (2005*a*). Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. *Forest Science*, *51*, 109–199.

LeMay, V. & Temesgen, H. (2005*b*). Connecting inventory information sources for landscape level analyses. Forest Biometry. *Modelling and Information Sciences*, *1*, 37–49.

LeMay, V., Maedel, J. & Coops, N. (2008). Estimating stand structural details using variable-space nearest neighbour analyses to link ground data, forest cover maps, and Landsat imagery. *Remote Sensing of Environment*, *118*, 2578–2591.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: John Wiley & Sons.

McRoberts, R. E. (2001). Imputation and model-based updating techniques for annual forest inventories. *Forest Science*, *47*, 322–330.

McRoberts, R. E. (2003). Compensating for missing plot observations in forest inventory estimation. *Canadian Journal of Forest Research*, *33*, 1990–1997.

McRoberts, R. E. (2008). Using satellite imagery and the *k*-nearest neighbors technique as a bridge between strategic and management forest inventories. *Remote Sensing of Environment*, *112*, 2212–2221.

McRoberts, R. E. (2009). Diagnostic tools for nearest neighbors techniques when used with satellite imagery. *Remote Sensing of Environment*, *113*, 489–499.

McRoberts, R. E., Nelson, M. D. & Wendt, D. G. (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the *k*-nearest neighbors technique. *Remote Sensing of Environment*, *82*, 457–468.

McRoberts, R. E., Holden, G. R., Nelson, M. D., Liknes, G. C. & Gormanson, D. D. (2006). Using satellite imagery as ancillary data for increasing the precision of estimates for the forest inventory and analysis program of the USDA Forest Service. *Canadian Journal of Forest Research*, *36*, 2968–2980.

McRoberts, R. E., Tomppo, E. O., Finley, A. O. & Heikkinen, J. (2007). Estimating areal means and variances of forest attributes using the *k*-Nearest Neighbors technique and satellite imagery. *Remote Sensing of Environment*, *111*, 466–480.

Magnussen, S., McRoberts, R. & Tomppo, E. (2009). Model-based mean square error estimators for *k*-nearest neighbour predictions and applications using remotely sensed data for forest inventories. *Remote Sensing of Environment*, *113*, 476–488.

Malinen, J. (2003). Locally adaptable non-parametric methods for estimating stand characteristics for wood procurement planning. *Silva Fennica*, *37*, 109–120.

Maltamo, M. & Eerikäinen, K. (2002). The most similar neighbour reference in the yield prediction of *Pinus kesiya* stands in Zambia. *Silva Fennica*, *35*, 437–451.

Maltamo, M. & Kangas, A. (1998). Methods based on *k*-nearest neighbor regression in the prediction of basal area diameter distribution. *Canadian Journal of Forest Research*, *28*, 1107–1115.

Maltamo, M., Malinen, J., Kangas, A., Härkönen, S. & Pasanan, A.-M. (2003). Most similar neighbour-based stand variable estimation for use in inventory by compartments in Finland. *Forestry*, *76*, 449–463.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A. & Kangas, J. (2006). Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. *Canadian Journal of Forest Research*, *36*, 426–436.

Maselli, F. (2001). Extension of environmental parameters over the land surface by improved fuzzy classification of remotely sensed data. *International Journal of Remote Sensing*, *17*, 3597–3610.

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+ images. *International Journal of Remote Sensing*, *26*, 3781–3796.

Moeur, M. & Stage, A. R. (1995). Most similar neighbor: An improved sampling inference procedure for natural resource planning. *Forest Science*, *41*, 337–359.

Muinonen, E., Maltamo, M., Hyppänen, H. & Vainikainen, V. (2001). Forest stand characteristics estimation using a most similar neighbor approach and image spatial structure information. *Remote Sensing of Environment*, *78*, 223–228.

Næsset, E., Gobakken, T., Holmgren, J., Hyyppä, H., Hyyppä, J., Maltamo, M., et al. (2004). Laser scanning of forest resources: The Nordic experience. *Scandinavian Journal of Forest Research*, *19*, 482–499.

Ohmann, J. L. & Gregory, M. J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. *Canadian Journal of Forest Research*, *32*, 725–741.

Packalén, P. & Maltamo, M. (2006). Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. *Forest Science*, *52*, 611–622.

Packalén, P. & Maltamo, M. (2007). The *k*-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. *Remote Sensing of Environment*, *109*, 328–341.

Packalén, P. & Maltamo, M. (2008). Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. *Canadian Journal of Forest Research*, *38*, 1750–1760.

Reutebuch, S. E., Andersen, H.-E. & McGaughey, R. J. (2005). Light detection and ranging (LIDAR): An emerging tool for multiple resource inventory. *Journal of Forestry*, *103*, 286–292.

Särndal, C.-E. & Lundström, S. (2005). *Estimation in surveys with nonresponse*. The Atrium: John Wiley & Sons.

Sironen, S., Kangas, A., Maltamo, M. & Kangas, J. (2001). Estimating individual tree growth with *k*-nearest neighbour and *k*-most similar neighbour methods. *Silva Fennica*, *35*, 453–467.

Sironen, S., Kangas, A., Maltamo, M. & Kangas, J. (2003). Estimating individual tree growth with nonparametric methods. *Canadian Journal of Forest Research*, *33*, 444–449.

Sironen, S., Kangas, A., Maltamo, M. & Kalliovirta, J. (2008). Localization of growth estimates using non-parametric imputation methods. *Forest Ecology and Management*, *256*, 674–684.

Stage, A. R. & Crookston, N. L. (2002). Measuring similarity in nearest neighbor imputation: Some new alternatives. In *Proceedings of Symposium on Statistics and Information Technology in Forestry* (pp. 91–96). Blackburg, VA: Virginia Polytechnic Institute and State University. Retrieved January 4, 2007, from www.forestry.ubc.ca/prognosis/documents/MSN_StageCrookston.pdf

Stage, A. R. & Crookston, N. L. (2007). Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. *Forest Science*, *53*, 62–72.

Temesgen, H., LeMay, V., Froese, K. L. & Marshall, P. L. (2003). Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. *Forest Ecology and Management*, *177*, 277–285.

Temesgen, H., Goerndt, M. E., Johnson, G., Adams, D. & Monseurd, R. A. (2007). Forest measurement and biometrics in the pacific Northwest USA: Status and future needs of the Pacific Northwest USA. *Journal of Forestry*, *105*, 233–238.

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

Tommola, M., Tynkkynen, M., Lemmetty, J., Harstela, P. & Sikanen, L. (1999). Estimating the characteristics of a marked stand using *k*-nearest-neighbour regression. *Journal of Forest Engineering*, *10*, 75–81.

Tomppo., E. (1991). Satellite image based national forest inventory of Finland. *International Archives of Photogrammetry and Remote Sensing*, *28*(7-1), 419–424.

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 Sensing of Environment*, *92*, 1–20.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O. & Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment*, *112*, 1982–1999.

Tomppo, E., Gagliano, C., De Natale, F., Katila, M. & McRoberts, R. E. (2009). Predicting categorical forest variables using an improved *k*-nearest neighbour estimator and Landsat imagery. *Remote Sensing of Environment*, *113*, 500–517.

Tuominen, S., Fish, S. & Poso, S. (2003). Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventory. *Canadian Journal of Forest Research*, *33*, 624–634.

Walter, B. F., Finley, A. O. & McRoberts, R. E. (2008, October). *Estimating optimal model parameter values for k-nearest neighbor prediction.* Paper presented at the FIA Symposium 2008, Park City, UT.