Remote Sensing of Environment 176 (2016) 282–294

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Review

A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data

Gherardo Chirici [a,*], Matteo Mura [a], Daniel McInerney [b], Nicolas Py [c], Erkki O. Tomppo [d], Lars T. Waser [e], Davide Travaglini [a], Ronald E. McRoberts [f]

[a] Università degli Studi di Firenze, Department of Agricultural, Food and Forestry Systems, Via San Bonaventura, 13-50145, Firenze, Italy
[b] Coillte Teoranta, Ireland
[c] Institut National de l'Information Géographique et Forestière, France
[d] The Natural Resources Institute, Finland
[e] Swiss Federal Institute for Forest, Snow and Landscape Research, WSL, Birmensdorf, Switzerland
[f] United States Department of Agriculture, Forest Service, Northern Research Station, United States of America

ARTICLE INFO

Article history:
Received 24 March 2015
Received in revised form 14 November 2015
Accepted 2 February 2016
Available online xxxx

Keywords:
k-Nearest Neighbors
Forestry applications
Review
Meta-analysis

ABSTRACT

The k-Nearest Neighbors (k-NN) technique is a popular method for producing spatially contiguous predictions of forest attributes by combining field and remotely sensed data. In the framework of Working Group 2 of COST Action FP1001, we reviewed the scientific literature for forestry applications of k-NN. Information available in scientific publications on this topic was used to populate a database that was then used as the basis for a meta-analysis. We extracted qualitative and quantitative information from 260 experimental tests described in 148 scientific papers. The papers represented a geographic range of 26 countries and a temporal range from 1981 to 2013. Firstly, we describe the literature search and the information extracted and analyzed. Secondly, we report the results of the meta-analysis, especially with respect to estimation accuracies reported for k-NN applications for different configurations, different forest environments, and different input information. We also provide a summary of results that may reasonably be expected for those planning a k-NN application using remotely sensed data from different sensors and for different forest attributes. Finally, we identify some methodological publications that have advanced the state of the science with respect to k-NN.

Contents

* Corresponding author.
E-mail address: gherardo.chirici@unifi.it (G. Chirici).

http://dx.doi.org/10.1016/j.rse.2016.02.001
0034-4257/

1. Introduction

Nearest neighbors techniques are a class of multivariate, non-parametric approaches to continuous or categorical prediction. The multivariate property of these techniques has made them particularly popular for use with remotely sensed and national forest inventory (NFI) data. With these techniques, predictions are calculated as linear combinations of observations for population units in a sample that are similar or nearest in a space of auxiliary variables to population units requiring predictions. Nearest neighbors techniques are appealing because they can be used for both univariate and multivariate prediction; they are non-parametric in the sense that no assumptions regarding the distributions of response or auxiliary variables are necessary; they are synthetic in the sense that they can readily use information external to the geographic area of interest; and they can be used with a wide variety of data sets. When used with remotely sensed and spatially referenced NFI field data, nearest neighbors techniques can produce spatially continuous predictions (maps) of forest variables rather than just large area aggregations of plot data. These finer resolution map products add a new and useful dimension to NFIs by facilitating small area estimation, increased precision for large area estimation, and support for forest management, planning and monitoring.

Nearest neighbors techniques were first introduced in an unpublished U.S. Air Force report by Fix and Hodges (1951) as a non-parametric discriminant technique for classification into populations whose distributions are unknown. Much of the early foundational work on nearest neighbors techniques for classification purposes appears in the pattern recognition and machine learning literature. Within the natural resources area, these techniques were developed for the Finnish NFI in seminal papers by Tomppo (1990, 1991) and Tomppo, Haakana, Katila, and Peräsaari (2008) based on earlier proposals by Kilkki and Päivinen (1987) and the ideas used with aerial photos by Poso (1972). McRoberts (2012) documented the broad international extent of the technique's use for a wide range of forestry applications including imputation of missing values for forest inventory and monitoring databases, mapping, small area estimation, and support for statistical inference. Commonly estimated forest response variables include growing stock volume, forest/non-forest, and forest type, and commonly used remotely sensed feature variables include Landsat spectral bands and increasingly airborne laser scanning metrics. Recent forestry investigations have begun to emphasize foundational work on diagnostics (McRoberts, 2009), efficiency (e.g., Finley & McRoberts, 2008), optimization (e.g., Tomppo & Halme, 2004), and inference (e.g., Baffetta, Fattorini, Franeschi, & Corona, 2009; McRoberts, Tomppo, Finley, & Heikkinen, 2007).

Variations of nearest neighbors techniques have been used operationally in both Europe and North America. In Finland, the first operational implementation of k-Nearest Neighbors (k-NN) was based on NFI, satellite and digital map data in 1990 (Tomppo, 1990, 1991). The primary initial purpose was forest resource estimation for small administrative units. The basic technique has since been enhanced using digital map data for stratification and genetic algorithms to weight feature variables as a means of increasing prediction accuracy (Tomppo & Halme, 2004). The resulting municipality-level estimates are included in the official NFI statistics in Finland (Metinfo, 2007; Metla, 2013). In Sweden, the k-NN technique has been used to map forest variables such as wood volume, age, and height using NFI, satellite and digital map data (Reese, Granqvist-Pahlén, Egberth, Nilsson, & Olsson, 2005). Basic end products include raster datasets for age, height, total wood volume, and volume by common species (SLU Forest Map, 2013). Additional products include dominant tree species, stand delineation, and base information for property taxation.

The k-NN technique has also been used operationally in North America. In Canada, Beaudoin et al. (2014) used the k-NN technique to produce continuous maps of 127 forest attributes to support regional policy and management issues. Reference data consisted of standardized observations from NFI photo plots, and feature variables were obtained from geospatial data layers that included MODIS spectral data, climatic and topographic variables. The map products provide unique baseline information for strategic analyses of Canadian forests (https://nfi.nfis.org). For the United States of America (USA), Wilson, Lister, and Riemann (2012) and Wilson, Woodall, and Griffith (2013b) used nearest neighbors techniques with NFI plot data and vegetation phenology derived from multi-temporal MODIS imagery and other auxiliary variables to map live tree basal area for individual species across the eastern United States and to map individual carbon stocks for all of the contiguous states of the USA (Wilson, Lister, & Riemann, 2013a, 2013c). In the Pacific Northwest region of the USA, Ohmann and Gregory (2002) used nearest neighbors techniques to map and assess biodiversity, wildland fuels, and species composition and to monitor change in older forests, biomass and carbon. The maps have been widely used for research, land management, forest monitoring, and conservation planning applications (http://lemma.forestry.oregonstate.edu/). Thus, the widespread popularity of nearest neighbors techniques for both research and operational purposes justifies a review of the literature on the topic and identification of important methodological advances along with issues regarding practical and scientific applications.

COST (European Cooperation on Science and Technology) is a European framework for promoting and facilitating scientific cooperation among scientists and researchers (COST, 2014). COST Action FP1001 focused on European approaches for using multi-source NFIs to improve information on the potential supply of wood resources. Within COST Action FP1001, Working Group 1 focused on NFI sampling designs and estimation techniques with an emphasis on harmonization; Working Group 2 focused on methods for combining remotely sensed and NFI field data to improve estimates of wood resources; and Working Group 3 focused on the exchange of inventory volume and consumption information with emphasis on wood markets (COST FP1001, 2014).

The popularity of k-NN for use with forest inventory and remotely sensed data motivated Working Group 2 of COST Action FP1001 to conduct a comprehensive literature review of forestry applications. The review was implemented as a meta-analysis of the most relevant studies published in peer-review journals, book chapters and conference proceedings. A meta-analysis is a quantitative analysis based on sound and reliable approaches aimed at providing an objective summary of results that may be helpful for other researchers in support of future applications. The usefulness of this kind of investigation, if compared to narrative or qualitative reviews, has been demonstrated for both ecological studies (Arnqvist & Wooster, 1995) and more recently for remote sensing applications in forestry (Garbulsky, Peñuelas, Gamon, Inoue, & Filella, 2011, Zolkos, Goetz, & Dubayah, 2013).

The study objectives were fourfold: (1) to document development and application of nearest neighbors techniques with respect to multiple factors including response and feature variables, distance metrics, algorithm characteristics, geographical regions of applications, accuracy and uncertainty measures, and results achieved in terms of prediction accuracy; (2) to provide a range of benchmark accuracies that may reasonably be expected for combinations of factors such as response variable and forest type; (3) to provide guidelines for prospective users; and (4) to identify and briefly summarize methodological papers that have advanced the state of the science. Thus, the paper complements and expands upon the support for practical implementations of nearest neighbors techniques documented in Eskelson et al. (2009) and McRoberts, Cohen, Næsset, Stehman, and Tomppo (2010), and the literature review section of McRoberts (2012), all of which focused more on specific forestry applications of k-NN than quantitative reviews of the different configurations actually used.

2. The k-Nearest Neighbors technique

For notational purposes, **Y** is commonly used to denote a possibly multivariate vector of response variables with observations for a sample of size n from a finite population of size N, and **X** is used to denote a vector of auxiliary variables with observations for all population units. In the terminology of nearest neighbors techniques, the auxiliary variables are designated *feature variables* and the space defined by the feature variables is designated the *feature space*; the set of sample population units for which observations of both response and feature variables are available is designated the *reference set*; and the set of population units for which predictions of response variables are desired is designated the *target set*. All population units for both the reference and target sets are assumed to have complete sets of observations for all feature variables.

For continuous response variables such as biomass or growing stock volume, the nearest neighbors prediction, $\hat{y}_i$, for the ith target set unit is calculated as,

$$\hat{y}_i = \sum_{j=1}^{k} w_{ij} y_j^i \tag{1}$$

where $\{y_j^i, j = 1, 2, \ldots, k\}$ is the set of response variable observations for the k reference set units that are nearest or most similar to the ith target set unit in feature space with respect to a distance metric, d, and w_{ij} is the weight assigned to the jth nearest neighbor with $\sum_{j=1}^{k} w_{ij} = 1$. Weights for neighbors are often of the form $w_{ij} \propto d_{ij}^{-t}$ with $0 \le t \le 2$ where d_{ij} is the distance in feature space between the ith target unit and the jth nearest neighbor. For categorical variables such as forest/non-forest or forest type, the predicted class of the ith target set unit is the most heavily weighted class among the k nearest neighbors, a weighted median or mode in case of ordinal scale variables, or a mode in the case of nominal variables.

Implementation of nearest neighbors techniques requires three selections: (i) the distance metric, d, to assess similarity, (ii) the number, k, of nearest neighbors to be used when calculating predictions, and (iii) a scheme to weight individual neighbors when calculating predictions. Multiple distance metrics have been proposed ranging from simple unweighted Euclidean distance to more complex metrics that attempt to optimize the selection and/or weighting of the feature variables. Many familiar metrics can be expressed in matrix form as,

$$d_{ij}^2 = (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{M} (\mathbf{X}_i - \mathbf{X}_j), \tag{2}$$

where i denotes a target set unit for which a prediction is sought, j denotes a reference set unit, $\mathbf{X}_i$ and $\mathbf{X}_j$ are vectors of observations of feature variables for the ith and jth units, respectively, and **M** is a square, positive definite matrix. When **M** is the identity matrix, Euclidean distance results; when **M** is a non-identity diagonal matrix, weighted Euclidean distance results; and when **M** is the inverse of the covariance matrix of the feature variables, Mahalanobis distance results. Metrics based on canonical correlation and canonical correspondence analyses can also be expressed in matrix form.

The value of k is often selected as a small number in the range 1–10, although some approaches attempt to optimize the selection with respect to criteria such as classification accuracy or root mean square error. Less attention has been paid to neighbor weighting schemes. The term *k-Nearest Neighbors* (k-NN) is generic and refers to any nearest neighbors technique regardless of the distance metric, value of k, or neighbor weighting scheme.

3. Methods

Bibliographic resources for this review were obtained from systematic searches of the most important scientific databases and search engines: Scopus (http://www.scopus.com/home.url), Thomson Reuters Web of Science, Science Direct (http://www.sciencedirect.com/), IEEE Xplore (http://ieeexplore.ieee.org/Xplore/home.jsp), and Google Scholar (http://scholar.google.com). The search was conducted in 2013 using English keywords that refer to the integrated use of forest inventory and remotely sensed data for prediction using the k-NN technique. The main keywords used were: k-nearest neighbor (or neighbor), connected using the logical "or" operator with the following keywords estimation, imputation, forest management, forest inventory, and remote sensing. The search returned 148 peer-reviewed contributions from scientific peer-reviewed journals, conference proceedings, and book chapters that reported 260 experimental applications and evaluations.

We then constructed a database with 24 fields and populated it with the qualitative and quantitative information extracted from the literature review that described the 260 experiments (Table 1). The resulting matrix, after deletion of duplicated studies, had 24 fields/columns and 260 records/rows with information for the different k-NN configurations extracted from the 148 peer-reviewed documents and served as the information source for our meta-analysis.

4. Results

4.1. General characteristics of studies

The main source of information was articles published in scientific journals (81.1%) with only 16.2% from conference proceedings. Nearly one-third of papers (31.1%) were published in Remote Sensing of Environment with the majority of the remainder published in Forest Ecology and Management, the Canadian Journal of Forest Research, the Scandinavian Journal of Forest Research, and the International Journal of Remote Sensing. Most papers (82.4%) reported applications of well-documented existing methods, but 7.4% reported more methodological contributions which were generally supported with practical applications or simulations.

The first paper identified from the literature search dates back to 1981 (Short & Fukunaga, 1981). Papers published in the 1980s often focused on the theoretical advantages of k-NN, while papers from the 90s and into the 21st century increasingly reported applications using forest inventory and remotely sensed data. The number of publications per year increased continuously until 2009 (Fig. 1).

The geographical coverage of published papers was mainly Europe and North America (Fig. 2). The total number of countries for which applications have been reported is 26 and includes papers from six continents (Austria, Brazil, Canada, Chile, China, Costa Rica, Ecuador, Estonia, Finland, Germany, Ghana, Ireland, Italy, Japan, Korea, Lithuania, Mexico, Namibia, New Zealand, Norway, Portugal, Russia, Scotland, South Korea, Sweden, and USA).

The forest environments most frequently investigated were boreal coniferous forests (36.5%) mainly located in Europe, and temperate continental (18.8%) and temperate mountain (11.5%) forests, mainly in the USA. Some European studies focused on temperate oceanic forests (6.9%) and Mediterranean forests (5%), with additional contributions for boreal mountain forests in the Nordic region (5.4%) (Fig. 3). A few studies were classified as "continental" because they covered very large areas such as the entirety of Canada (Beaudoin et al., 2014), the entire USA (Wilson et al., 2013a), and the eastern part of the USA (Wilson et al., 2012).

The most frequently investigated response variables, representing 55.4% of studies, were the closely related growing stock volume, biomass and carbon stock variables, with other continuous response variables such as basal area (9.2%) and tree density (3.8%) also reported. Categorical response variables such as land use/land cover and forest type were investigated in 13.1% of studies.

Among feature variables, Landsat imagery was undoubtedly the most frequent source of remotely sensed data, being used in 48.8% of studies. Airborne laser scanning (ALS) metrics were used in almost 7% of the studies, digital aerial imagery in 6.5%, and SPOT imagery in 5.4%.

Table 1
Short description and data type in the database of information extracted from the analyzed papers.

ID	Field	Definition	Type	Categories adopted
1	Id paper	Identification number of the paper (internal)	Numeric	
2	Paper	Paper metadata. The full citation of the resource.	Free text	
3	Author/s	Name/s of author/s	Free text	
4	Title	Title of the paper	Free text	
5	Year	Year of publication	Free text	
6	Source title	Name of the journal or conference where the paper was published or presented	Free text	
7	Document type	Type of resource	Classes	Published article Conference paper/poster Technical report Book Paper in press
8	Study type	This information is related to the main dominant content type of the resource	Classes	Application (well known methods with no or very limited innovation applied to real cases) Theoretical (new methods presented without applications or with simulated data only) Theoretical with practical application (new methods presented applied to real cases) Review Software description
9	Method	Method used in estimation/imputation. Even if the review is oriented to analyze papers based on the k-NN approach, many of them present the comparison of k-NN with other imputation methods that for this reason were included in this review.	Classes	k-NN Random Forest Regression Decision Trees Others
10	Predicted variable/s	Predicted/response variables to be estimated	Classes	Growing stock volume/biomass LULC (including estimation of forest area) Basal area Stem/tree density (count/ha) Others
11	Main feature variables	Describe the remotely sensed source of imagery/layers upon which the estimates are based	Classes	Landsat Data fusion, (e.g. ALS + Landsat) ALS metrics Digital aerial imagery SPOT

(continued on next page)

Table 1 (*continued*)

ID	Field	Definition	Type	Categories adopted
				SAR IRS ASTER MODIS Quickbird Hyperspectral Others
12	Pre-processing	Account if pre-processing and/or band transformation were performed	Classes	Yes No NA
13	Auxiliary variables	Account if LULC, soil or climate -related layer were used (not remotely sensed data)	Classes	Yes No NA
14	Country	Country where the test was performed or dominant country in terms of forest area in the case of cross-boundaries study	Free text	
15	Geographic area	More precise location in the country (e.g. region, province, municipality)	Free text	
16	Area	Area covered by the estimations, in hectares.	Numeric	
17	GEZ	Dominant Global Ecological Zones (FAO, 2012) of the "Geographical Area"	Classes	See FAO classes (FAO, 2012) Global, when two or more GEZ were included
18	Numerosity of reference dataset	Number of units of the reference dataset.	Numeric	
19	Number of k's	Number of k nearest neighbors used	Numeric	
20	Distance type	Multivariate distance type used	Classes	Euclidean Mahalanobis Random Forest Canonical correlation analysis Canonical correspondence analysis Others
21	Accuracy measure	Accuracy measure used to assess the performance of the estimation	Classes	RMSE/SE or similar approaches Overall accuracy, K-index of Agreement or similar approaches Regression-oriented approaches (R^2, R^2-adj, R-Pearsons) Mean deviation (e.g. bias) Others NA
22	Accuracy value	Value of the accuracy measure	Numeric	–
23	Accuracy assessment level	Level used to assess the performance	Classes	Pixel Area
24	Confidence interval	If confidence intervals were estimated or not	Classes	Yes No NA

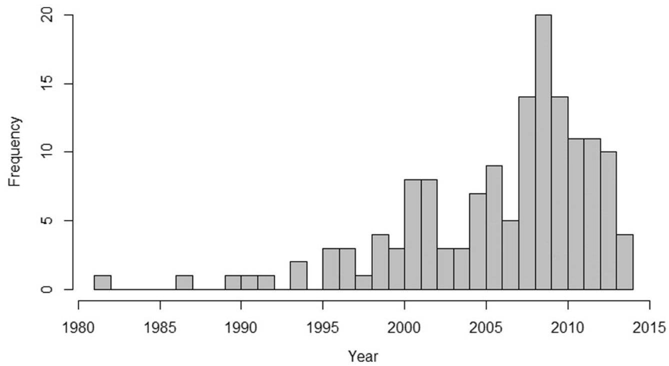

Fig. 1. Number of publications per year.

More than 16% of the studies integrated remotely sensed data from multiple sources, while 25% of studies used data from non-remote sensing-based digital maps (e.g., digital elevation models) as feature variables or as auxiliary information for stratification purposes.

All studies used field data, mostly from local inventories or NFIs. For the 235 studies that reported the number of plots, sample sizes ranged from a minimum of 9 to a maximum of 190,888, with an average of 2575. For 85% of these studies, reference set (sample) sizes for the field data were less than 2000, and for 50% of the studies, the sizes were less than 400 (Fig. 4).

4.2. k-NN configurations used

Among the k-NN distance metrics used, the Euclidean metric was most often reported (70%), followed by the Mahalanobis metric (3.5%), and the canonical correlation analysis metric (1.9%), all three of which can be readily expressed in matrix form. Additional metrics reported that cannot be so readily expressed in matrix form include the Random Forest (Lin & Jeon, 2002), Fuzzy, and non-parametric metrics (Chirici et al., 2008; Maselli, Chirici, Bottai, Corona, & Marchetti, 2005). Values of k generally ranged between 1 and 10 with $k = 5$ being used most often and $k > 10$ used only rarely (Fig. 5). Selection of the k-value appeared to be independent of forest environment and the choice of a distance metric, although $k = 1$ is often selected for use with the canonical correlation analysis and canonical correspondence metrics.

4.3. Accuracy assessment

For continuous response variables, the most commonly used measures for assessing the accuracy of k-NN predictions were root mean square error (RMSE) and the standard error (SE) of estimates, in both absolute and relative terms (72%). Other measures included the correlation coefficient (r), coefficient of determination (R^2), and both absolute and relative mean deviation between observations and predictions (3.3%). For categorical response variables, overall accuracy or the Kappa index of agreement were most often used (9.2%). From a spatial scale perspective, accuracy assessments were conducted for aggregations of pixel-level estimates (e.g., plots, compartments) for almost 65% of studies, whereas more than the 26% of accuracy assessments were at the pixel-level. A few papers (3.4%) reported confidence intervals for estimates of population parameters.

A subset of 110 of the 260 studies was used to assess the accuracy of k-NN predictions using RMSE% which expresses RMSE as a percentage of the estimated mean, or occasionally the observed mean. For these studies, RMSE% varied between 0.3% and 170.6%, with an average of approximately 33% (standard deviation of ±29.3%). Because average RMSE% at pixel-level scale and aggregations of pixels to larger scales were generally comparable (39% and 31%, respectively), we do not distinguish between them in the following analyses.

Independently of other factors such as response variable, value of k, and distance metric, the most accurate results were obtained when the source of feature variables was fine resolution aerial imagery (15 studies with average RMSE% of 26% ± 18.7%) or fused multi-sensor data (27 studies with average RMSE% of 27% ± 19.6%), followed by synthetic aperture radar (SAR) and lidar-based airborne laser scanning (ALS). Landsat imagery, the mostly commonly used source of remotely sensed data as demonstrated by its use in almost 50% of the studies, produced an average RMSE% of 40.4% (±29.6%). Independently of other factors, the average RMSE% for estimating basal area (9 studies) was 23% (±22.4%), for tree density (2 studies) was almost 25% (±12.7%), while for growing stock volume/biomass/carbon (81 studies) the average RMSE% was 37% (±31.6%).

The choice of distance metric did not substantially influence estimation performance, even though four studies using the canonical correlation analysis metric and nine studies using the Mahalanobis metric reported smaller RMSE% results (18.8% ± 11.4% and 21.5% ± 22.9% respectively) than the average for studies using the Euclidean metric (RMSE% of almost 35% ± 31.4%). Of importance, however, studies using the Euclidean metric are more numerous (77) and represent a much broader range of forest types, response variables and feature variables.

To better understand the relationship between k-NN prediction accuracies, feature variables and reference set size, we selected a homogeneous and comparable subset of 53 studies that used growing stock volume/biomass/carbon as the response variable, a single type of feature variable, and RMSE% as the measure of prediction accuracy. For studies that used only feature variables based only on aerial photography, the median reference set size of 332 produced median RMSE% = 48.5; for studies that used only satellite-based spectral feature variables, reference set sizes less than 1000 (median 280) produced median RMSE% = 40.0, whereas reference set sizes of 1000 or greater (median 3117) produced median RMSE% = 15.8; for studies that used only lidar-based feature variables, the median reference set size of 124 produced median RMSE% = 31.3; and finally for studies that used only radar-based feature variables, median reference set size of 300 produced median RMSE% = 44.8. This subset of studies confirms that large reference sets are necessary to obtain small RMSE% when using satellite spectral

Fig. 2. Locations of the studies according to the sub-country geographic area reported in the paper (the size of the dots are proportional to the area covered by the study). Base map from Esri ArcGIS online.

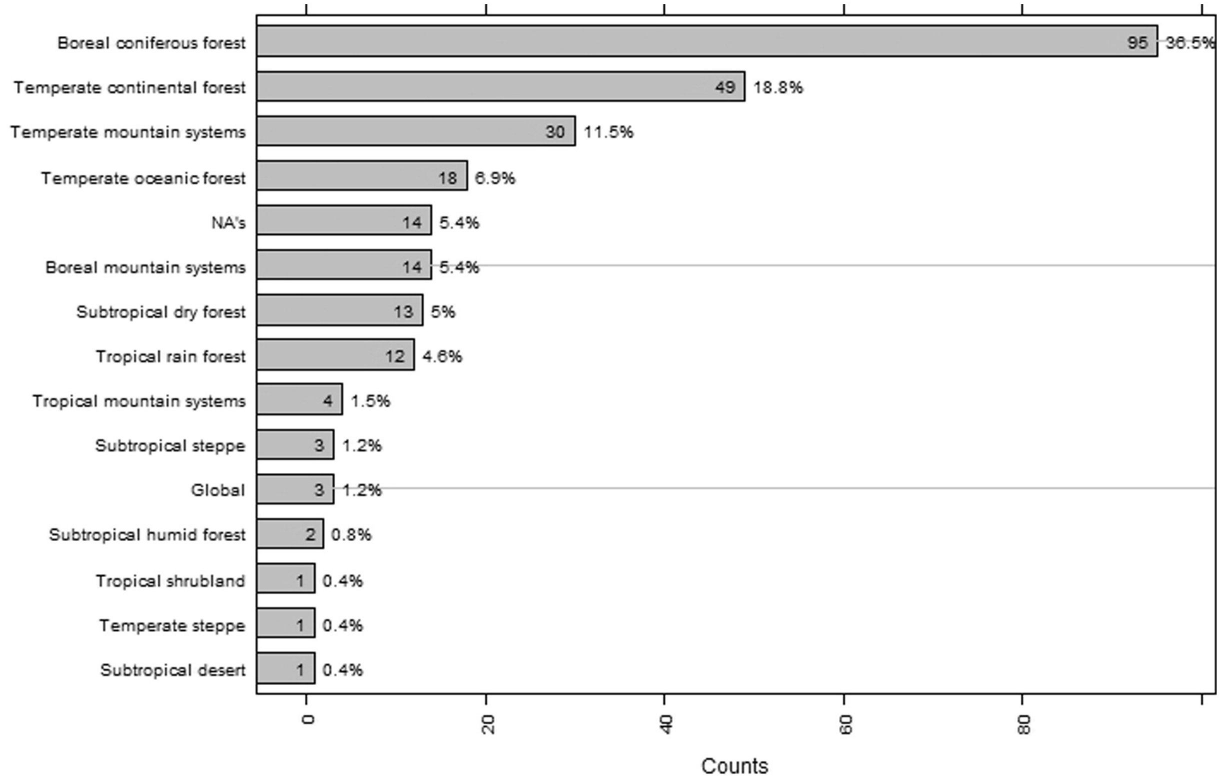

Fig. 3. Number of publications per Global Ecological Zones (FAO, 2012). NA for not available.

data. However, these results also depend to some degree on the complexity and diversity of the forests; in particular, most of the studies using satellite spectral feature variables were for somewhat less complex boreal and northern temperate forests. These studies also confirm the growing popularity of lidar-based feature variables for obtaining small RMSE%, even with small reference sets.

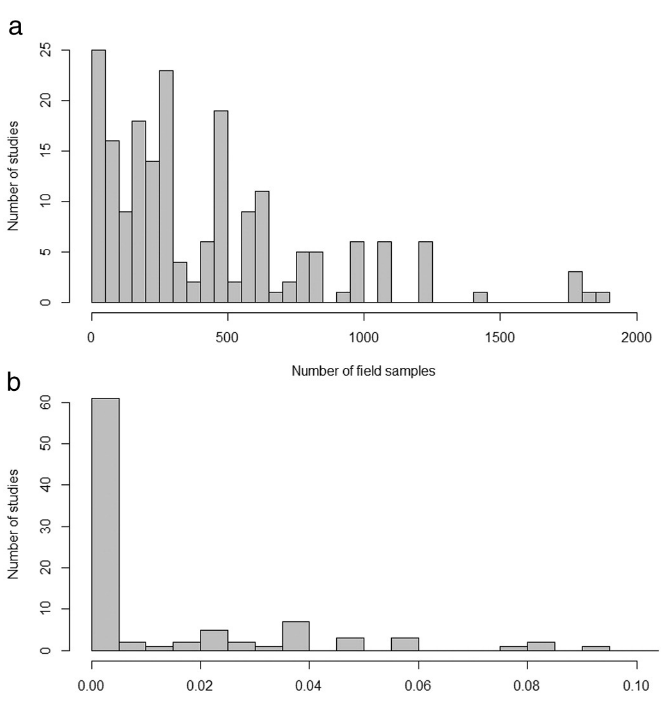

Fig. 4. a) Number of field samples used in studies. The maximum value was limited to 2000 for display reasons. b) Plot density (number of plots per hectares) of the studies. The maximum value was limited to a density of 0.10 for display reasons.

Over all 260 studies, transforming raw remotely sensed input feature data using dimension reduction techniques such as principal component analysis (PCA) or band ratio indices such as the normalized difference vegetation index (NDVI) had no effect on accuracy: the average RMSE% for the 80 studies using transformations was 41.5% ($\pm$32.7%) contrasted with average RMSE% of 36.8% ($\pm$32.8%) for the 33 studies that used no transformations.

Although for some studies, accuracy tended to decrease with greater values of k, generally we found no relationship between the final accuracy and the value of k selected as optimal.

4.4. Summary of meta-analysis

The k-NN technique is well-affirmed for integrating forest inventory field data and remotely sensed data. In this bibliographic meta-analysis we analyzed 260 experimental tests published in 148 papers since 1981, all dealing with application of k-NN for a variety of forest environments across the world. In the 1990s and the 21st century most studies were from Finland and the northern parts of the USA; since then, the technique has been successfully tested and applied in 26 countries on six continents.

The year 2009 had the largest number of published k-NN papers. Subsequently, the number of publications decreased, perhaps because the technique had become well-known and well-documented. Nowadays, the technique is widely adopted, but it is less frequently a direct research objective. However, other papers not analyzed for this study because they did not provide details on methods or the k-NN configuration, reported using the k-NN technique as a standard procedure for constructing spatially contiguous forest information to be used separately or integrated with other information for further geographical analysis.

More than half the k-NN studies had growing stock volume, wood biomass or carbon content as the response variable. However, k-NN

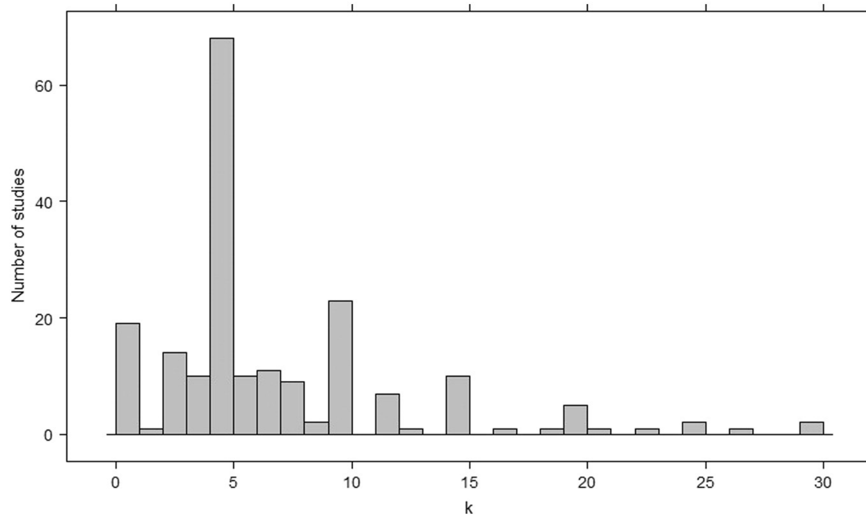

Fig. 5. Distribution of numbers of k's in the investigated studies.

was also applied for estimating a very large number of other continuous and categorical response variables related to functional attributes such as leaf area index, structural attributes such as basal area or trees per unit area, or levels of ecosystem disturbance such as defoliation.

5. Expected results

Because the meta-analysis did not reveal a particular k-NN configuration that could be considered optimal for all the cases, a reasonable conclusion is that for any application, an optimization phase using available data should be considered for calibrating the particular configuration. However, some guidelines may be proposed. Firstly, the selection of feature variables is related to the geographical resolution of the analysis (pixel size) and the availability of remotely sensed data. Almost half the studies used Landsat imagery alone, but the majority of the studies augmented Landsat data with data from other similar multispectral satellites such as SPOT, IRS and ASTER. Studies with finer geometric resolution were frequently based on aerial photography, whereas those at coarser resolutions were frequently based on MODIS images. The advent of ALS technology since 2008 has changed this trend with the number of studies using ALS metrics alone or integrated with optical imagery increasing rapidly.

Secondly, the simplest distance metric, Euclidean distance, was most commonly used (70%), presumably because of its simpler coding implementation and because more complex metrics did not produce better results. For values of k, the majority of the studies adopted a value of approximately 5 as a compromise between larger values that may produce greater accuracy and smaller values that better retain variability in the reference set and are computationally less intensive.

Thirdly, in terms of accuracy, the k-NN technique may be expected to produce RMSE% in the range of 20–40% for common response variables (69% of cases RMSE% < 40%). However, despite the great variety of k-NN configurations and local environmental conditions, only 22% of studies reported successful use of k-NN with RMSE% less than 10%, while multiple studies reported unsuccessful uses with RMSE% as great as 100% or more (Fig. 6).

Of interest, relative to the results of experimental tests conducted in the 1990s, the scientific community in the period 2000–2014 has achieved a better tuning of the k-NN method that then produced increasingly more accurate results than in prior years (Fig. 7). This result can probably be attributed to greater visibility of k-NN studies via articles published in the scientific literature and presentations at international conferences (e.g., the ForestSAT series of conferences and the k-NN workshops). In addition, greater prediction accuracy can also be at least partially attributed to the recent availability of remotely sensed data of greater quality from both passive and active sensors and to improved global navigation satellite systems for the geolocation of field inventory plots.

Fourthly, the vegetation characteristics of the study areas where k-NN was used did not substantially affect estimation performance. For boreal study areas which had the greatest number of studies (59), the average RMSE% was 35%, while the average RMSE% for other vegetation zones ranged from 6% for temperate mountain systems (only 5 studies) to 42% for temperate oceanic forests (15 studies) (Fig. 8).

Fifthly, the accuracy of k-NN predictions was not greatly influenced by the selection of the feature variables. Accuracies obtained using feature variables from sensors other than Landsat were all within the range of accuracies using Landsat-based feature variables (Fig. 9). However, Landsat-based studies produced a wide range of results, presumably because they were the earliest and consequentially the largest in number. The recent advent of ALS feature variables may substantially alter this finding; in particular, ALS metrics, alone or in conjunction with optical

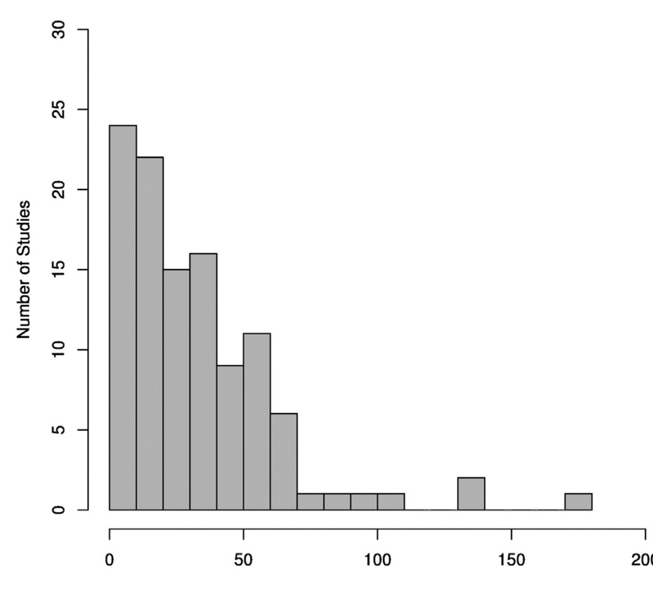

Fig. 6. Distribution of the accuracies of 110 studies where the RMSE% of the k-NN estimates was reported. The average is 33%.

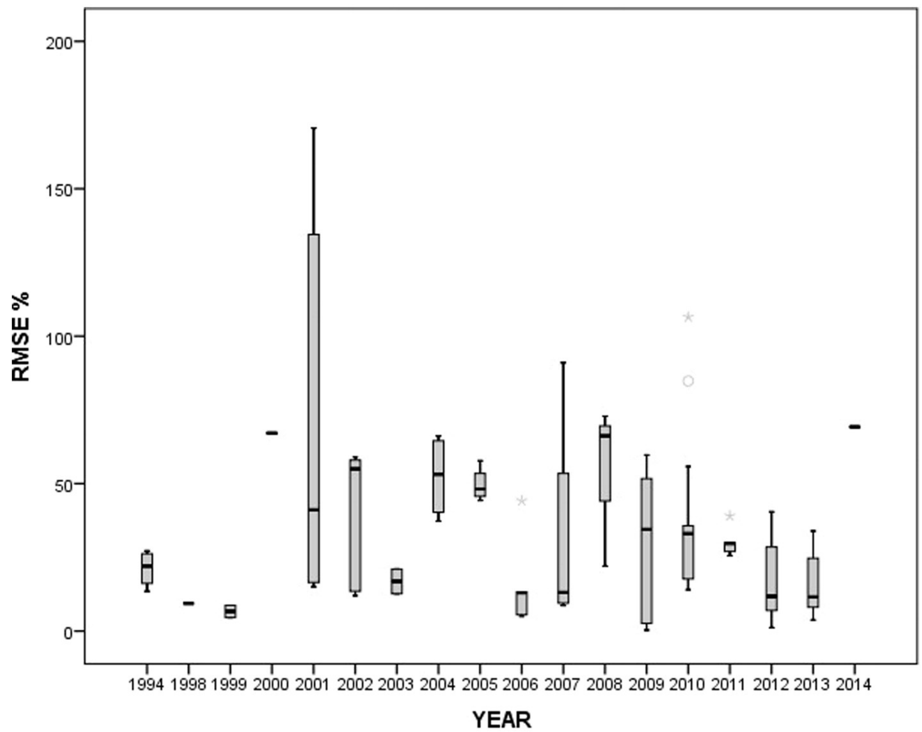

Fig. 7. RMSE% distribution of the accuracies for k-NN studies for the different years.

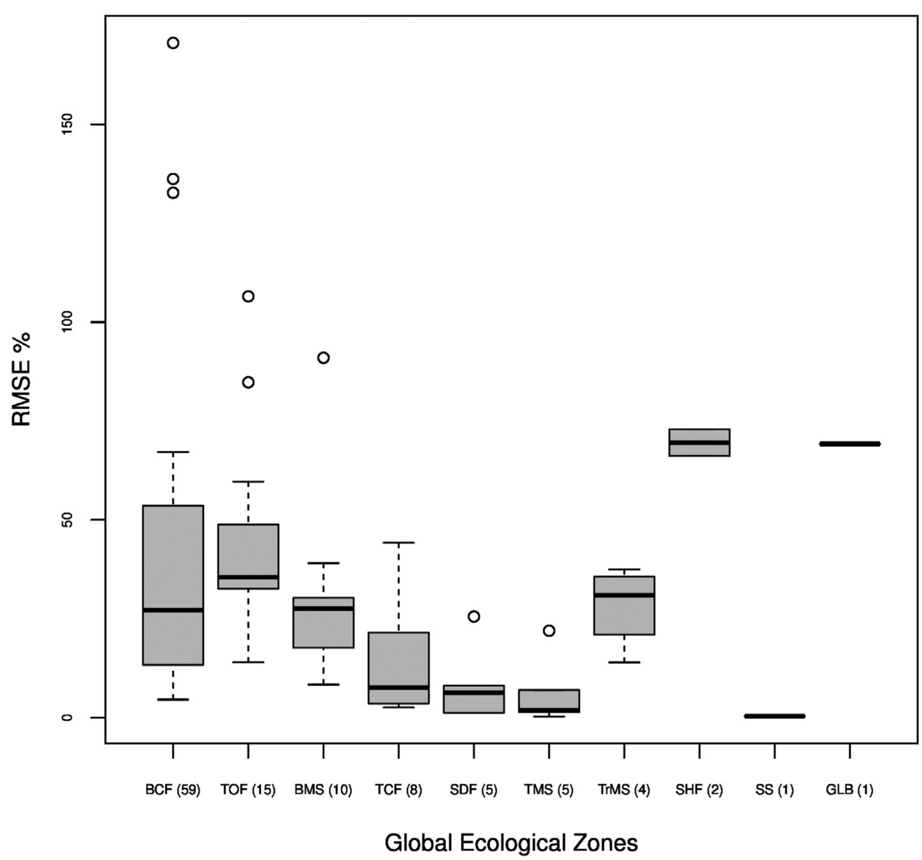

Fig. 8. RMSE% distribution of the accuracies for k-NN studies in different Global Ecological Zones (FAO, 2012). SS: subtropical steppe, TMS: temperate mountain systems, SDF: subtropical dry forest, TCF: temperate continental forests, TrMS: tropical mountain systems, BMS: boreal mountain systems, BCF: boreal coniferous forests, TOF: temperate oceanic forests, GLB: global, SHF: subtropical humid forests.

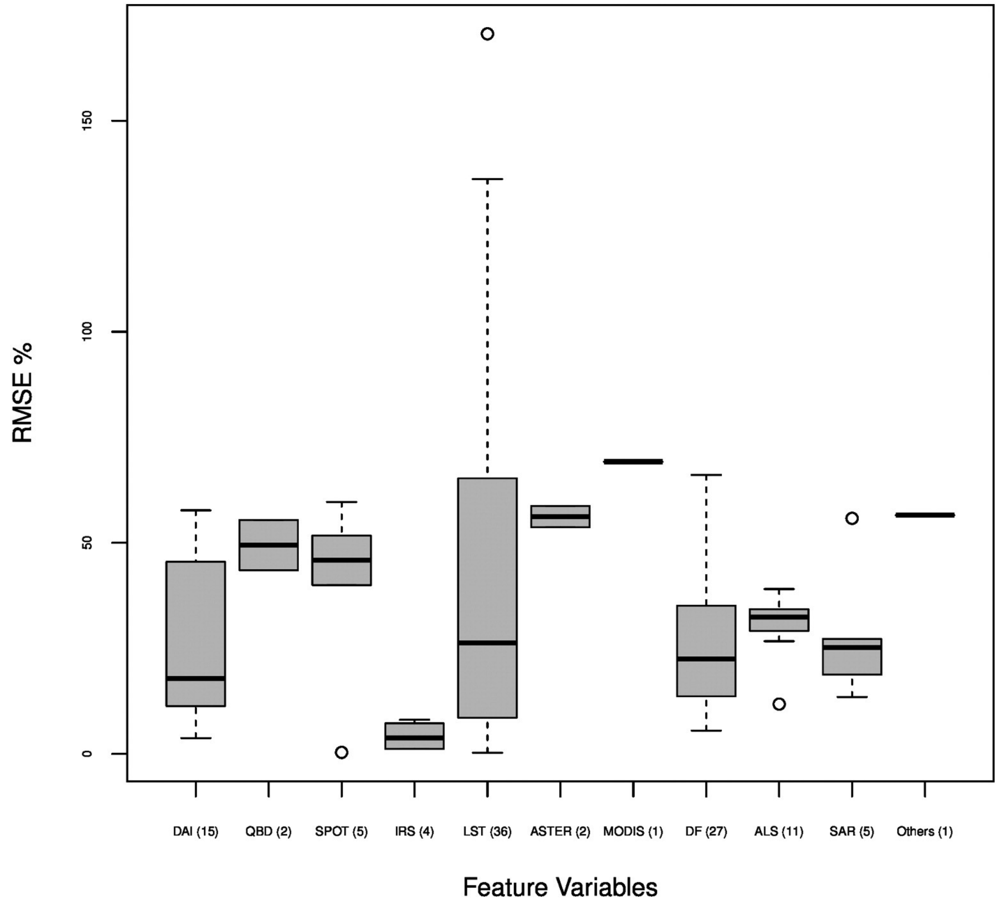

Fig. 9. RMSE% distribution of the accuracies for k-NN studies based on feature variables from different remotely sensed data. IRS: Indian for Remote Sensing, DAI: digital aerial imagery, DF: data fusion from different sensors, SAR: synthetic aperture radar, SPOT: *Satellite Pour l'Observation de la Terre*, LST: Landsat, QBD: Quick Bird, ASTER: ASTER, MODIS: MODIS Terra and Aqua.

imagery, are among the most promising feature variables for producing k-NN predictions.

Sixthly, on the basis of the meta-analysis over all 260 studies, we found that field reference set sizes did not affect the accuracy of k-NN predictions. Thus, the k-NN technique can be successfully used from local scales through to continental-level investigations.

6. Methodological advances

Citations and meta-analyses are useful for documenting the scope, range, geographic distribution, and history of k-NN forest applications. Further, they are useful for determining the most frequently cited papers as a measure of information sources that others have found useful and relevant. Table 2 reports the most frequently cited papers published during or before 2013 based on the number of citations standardized to reflect elapsed time since publication.

As noted in Section 4.1, the vast majority of published papers focused on specific applications whereas only a much smaller number focused on methodological issues. For the purposes of assessing methodological advances, a limited e-mail survey was conducted to augment the literature review and meta-analysis. The survey solicited opinions from k-NN users and researchers regarding methodological publications that have advanced the state of the science for forest applications. The following summary represents an admittedly subjective consensus of the results of the survey.

6.1. Optimization

Optimization of the k-NN technique entails selecting a distance metric and values for k and t. Nearly all optimization efforts have focused on selecting or formulating a distance metric with much less effort focused on optimizing k and t. Although many metrics have been proposed, the unweighted Euclidean distance and the metric based on canonical correlation analysis have been used most widely. The unweighted Euclidean distance metric, which can be expressed using an identity matrix in Eq. (2), is the simplest, most intuitive, and easiest to implement. The only degree of optimization for this metric pertains to the particular feature variables to be used. Selection of feature variables can be accomplished by comparing all combinations of all numbers of

Table 2
Nearest neighbors articles ranked by number of citations normalized by years[a].

Rank	Average number of citations per year	Publication
1	16	Ohmann and Gregory (2002)
2	15	Franco-Lopez, Ek, and Bauer (2001)
3	14	Crookston and Finley (2008)
4	11	McRoberts & Tomppo (2007)
5	11	McRoberts et al. (2007)
6	9	Tomppo and Halme (2004)
7	8	Katila & Tomppo (2002)
7	8	McRoberts, Nelson, and Wendt (2002)
8	7	Tomppo, Nilsson, Rosengren, Aalto, and Kennedy (2002)
9	6	Tokola, Pitkanen, Partinen, and Muinonen (1996)
10	6	Trotter, Drymond, and Goulding (1997)
11	5	Pretsch (1997)
12	5	Fazakas, Nilsson, and Olsson (1999)
13	4	Maltamo and Kangas (1998)

[a] Ranked by Scopus citations.

feature variables (McRoberts, 2012), stepwise selection (Chirici et al., 2016), or use of genetic algorithms (Tomppo & Halme, 2004; Tomppo, Gagliano, De, Katila, & McRoberts, 2009).

The canonical correlation analysis metric was proposed by Moeur and Stage (1995) and has been used fairly widely (e.g., LeMay & Temesgen, 2005; Maltamo & Kangas, 1998; Maltamo, Malinen, Kangas, Härkönen, & Pasanen, 2003). The metric is based on an optimal relationship between a linear combination of response variables and a linear combination of feature variables. The configuration consisting of this metric and $k = 1$ has been characterized as the *Most Similar Neighbor* approach (Moeur & Stage, 1995). The metric based on canonical correspondence analysis was proposed by Ohmann and Gregory (2002). The configuration consisting of this metric and $k = 1$ has been characterized as the *Gradient Nearest Neighbor* approach. However, both metrics have also been used with $k > 1$ (e.g., Maltamo et al., 2003; Ohmann et al., 2014).

The weighted Euclidean distance metric, which can be expressed using a non-identity diagonal matrix in Eq. (2), is similar to the unweighted metric, albeit with unequal values on the diagonal of the matrix (Tomppo & Halme, 2004). Because optimization of this metric is computationally intensive, its use has not been reported sufficiently frequently to appear in lists of most cited publications; nevertheless, recent technological advances suggest that it has considerable optimization potential. Optimization entails selection of the diagonal values, one for each feature variable. For a large number of feature variables, optimization can be excessively computationally intensive. A two-step alternative is to first select a small subset of feature variables and then to optimally select the corresponding diagonal values of the matrix corresponding to the selected feature variables. The first step can be accomplished using the same methods noted for the unweighted Euclidean metric, while the second step would typically be accomplished using a genetic algorithm (Tomppo & Halme, 2004) (Section 6.3).

Little attention has been devoted to optimizing the k-NN technique with respect to selection of the k and t values. Small values of k, typically in the range $1 \leq k \leq 10$, have been common along with selections of $t = 0$, $t = 1$, and occasionally $t = 2$. Optimization of the distance metric followed by non-optimized selections of k and t may be self-defeating; specifically, the beneficial effects of optimizing the distance metric may be mitigated by the adverse effects of non-optimized selections of k and t. McRoberts (2012) and McRoberts, Næsset, and Gobakken (2015) are the only known attempts to optimize the distance metric, k, and t simultaneously.

6.2. Inference

Multiple variations of nearest neighbors techniques have been shown to be useful and effective for both prediction and mapping. However, the ultimate judgment is whether these techniques can contribute to the construction of an inference in the form of a confidence interval for a population parameter. Two modes of inference are common, design-based inference and model-based inference. McRoberts et al. (2002) demonstrated the utility of nearest neighbors techniques as a key component of design-based, stratified estimation, and Baffetta et al. (2009) and Baffetta, Corona, and Fattorini (2011) demonstrated its utility for design-based, model-assisted estimation. For model-based inference, McRoberts et al. (2007) derived parametric estimators for means and variances and demonstrated their inferential utility. To circumvent the complexity and computational intensity associated with parametric estimators, McRoberts, Magnussen, Tomppo, and Chirici (2011) proposed and demonstrated a bootstrapping approach to model-based variance estimation. Magnussen, McRoberts, and Tomppo (2009) noted and McRoberts et al. (2015) confirmed that parametric estimates may be unreliable for $k \leq 7$. In the latter case, the bootstrapping approach should be used.

6.3. Enhancements

Little attention has been devoted to developing or using comprehensive sets of diagnostics specific to k-NN applications. Common diagnostics include comparisons of RMSEs for different combinations of feature variables and graphs of observations versus predictions for assessing quality of fit. McRoberts (2009, 2012) and McRoberts et al. (2015) proposed additional diagnostics for identifying influential outliers in the reference set and for assessing the degree to which gains achieved by optimization in the reference set are realized in the target set.

Several approaches have been proposed to improve the accuracy of k-NN predictions. Katila and Tomppo (2002) introduced constraints on the selection of neighbors as a means of increasing accuracy. The constraints include a maximum distance in geographic space between target and reference units and restriction of neighbors to the same land use stratum as the stratum of the target unit. Tomppo and Halme (2004) augmented feature space with additional variables whose values were constructed as interpolations of field plot observations over the target space.

Beyond simply reducing the dimension of feature space by deleting some auxiliary variables, selecting optimal diagonal values for the weighted Euclidean distance matrix can be an extremely computationally intensive task, even for small numbers of feature variables. Tomppo and Halme (2004) proposed genetic algorithms for this purpose. Genetic algorithms are heuristic search procedures that theoretically converge to optimal or near optimal solutions. The search mechanism mimics natural selection using techniques inspired by natural evolution such as inheritance, mutation, selection, and crossover.

A unique feature of k-NN techniques is that no prediction can be smaller than the smallest reference set observation nor larger than the largest reference set observation. Therefore, unlike regression models, k-NN techniques cannot extrapolate predictions beyond the range of the response variable in the reference set, even if ranges of feature variables in the target set may be greater than ranges in the reference set. Magnussen, Tomppo, and McRoberts (2010) developed an approach that uses a local linear model to extrapolate predictions beyond the range of reference set feature variables.

7. Conclusions

The study was motivated by the popularity of the k-NN technique for use with forest inventory and remotely sensed data. The analyses were conducted within the framework of Working Group 2 of COST Action FP1001 and focused on a review of the scientific literature for forestry applications. Information available in the scientific publications was used to populate a database that served as the basis for a meta-analysis.

Multiple conclusions drawn from previous experiences are useful as background for future forest implementations and to stimulate future research. Nowadays, k-NN can be considered a useful, well-affirmed technique for a broad scope of international forest inventory applications using remotely sensed data. Previous experiences have demonstrated that k-NN can be used in all vegetation zones and at spatial scales ranging from local applications based on a limited number of field observations to large national and continental applications. The k-NN technique is frequently used to estimate growing stock volume, tree biomass, and carbon stock, but it can be used also with a large variety of response variables.

Feature variables can be derived from the outputs of traditional multi-spectral optical sensors as well as active radar and ALS systems. Performance of the k-NN technique is dependent on an optimization phase aimed at selecting the set of feature variables and their weights, the value of k, neighbor weighting, and the multidimensional distance metric.

The meta-analysis revealed that the simplest k-NN configuration frequently produced excellent results. For this reason we suggest a first tentative implementation of k-NN using the Euclidean distance metric, values of k ranging between 3 and 10, and t = 0, t = 1, and t = 2. If the pixel-level or area-level k-NN predictions produce RMSE% less than 30%, the result should be considered in line with results reported in the literature. Starting from this simple k-NN configuration, multiple diagnostic tests and investigations of variations of the parameters and feature variables can be conducted to obtain more accurate results. Examples include transformations of feature variables, more sophisticated distance metrics, and optimization of k and t. There is no consensus regarding the order of the optimization steps that produce the greatest gains in accuracy. A reasonable order would be first to select the feature variables that reduce the effects of the curse of dimensionality by eliminating redundant feature variables. Then, depending on the nature of the feature variable set, more sophisticated distance metrics such as the weighted Euclidean or canonical correspondence analysis metrics can be tested. Finally the k value can be optimized on the basis of the leave-one-out approach.

Integration of remotely sensed data from different sensors, particularly both passive and active sensors, and non-remote sensing variables (such as elevation or climatic maps) can be considered to facilitate more accurate prediction of both structural and functional forest attributes.

Finally, we strongly encourage scientists using the k-NN technique to report results obtained for all of the multiple configurations considered in the optimization phase. Doing so will facilitate future meta-studies and benefit both researchers and practitioners involved in the operational implementation of the k-NN technique.

Acknowledgements

This paper was prepared within the activities of COST Action FP1001 "USEWOOD — Improving Data and Information on the Potential Supply of Wood Resources: A European Approach from Multisource National Forest Inventories", Working Group 2.

Part of the work was supported by the French National Research Agency (project ANR-2010-BIOE-008 FORESEE) and by the Italian Ministry of Research and Education (project PRIN2012 "Development of innovative methods for forest ecosystems monitoring based on remote sensing", grant 2012EWEY2S, project coordinator: G. Chirici).

We acknowledge the contributions of two anonymous reviewers whose suggestions greatly improved the manuscript.

References

Arnqvist, G., & Wooster, D. (1995). Meta-analysis: Synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution*, *10*(6), 236–240.

Baffetta, F., Corona, P., & Fattorini, L. (2011). Design-based diagnostics for k-NN estimators of forest resources. *Canadian Journal of Forest Research*, *40*(1), 59–72.

Baffetta, F., Fattorini, L., Franeschi, 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*(3), 463–475.

Beaudoin, A., Bernier, P. Y., Guindon, L., Villemaire, P., Guo, X. J., Stinson, G., ... Hall, R. J. (2014). Mapping attributes of Canada's forests at moderate resolution through kNN and MODIS imagery. *Canadian Journal of Forest Research*, *44*, 521–532.

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

Chirici, G., McRoberts, R. E., Fattorini, L., Mura, M., & Marchetti, M. (2016). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sensing of Environment*, *174*, 1–9. http://dx.doi.org/10.1016/j.rse.2015.11.010.

COST (2014). European Cooperation in Science and Technology. Available at: http://www.cost.eu/ last accessed 5 October 2015

COST FP1001 (2014). USEWOOD. Improving data and information on potential supply of wood resources, a European approach from multisource national forest inventories. Available at: https://sites.google.com/site/costactionfp1001/ (Last accessed October 2015)

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

Eskelson, B. N. I., Temesgen, H., Lemay, V., Barrett, T. M., Crookston, N. L., & Hudak, A. T. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research*, *24*, 235–246.

FAO (2012). Global ecological zones for FAO forest reporting: 2010 Update. *Forest Resources Assessment Working Paper*, *179*, Rome: FAO.

Fazakas, Z., Nilsson, M., & Olsson, H. (1999). Regional forest biomass and wood volume estimation using satellite data and ancillary data. *Agricultural and Forest Meteorology*, *98-99*, 417–425.

Finley, A. O., & McRoberts, R. E. (2008). Efficient k-nearest neighbor searches for multi-source forest attribute mapping. *Remote Sensing of Environment*, *112*, 2203–2211.

Fix, E., & Hodges, J. L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Report Number 4, Project 21-49-004*. Randolph Field, Texas: USAF School of Aviation Medicine (Reprinted as: Fix, E., and Hodges, J.L. (1989). Discriminatory analysis — Nonparametric discrimination: Consistency properties. International Statistical Review, 57, 238–247).

Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment*, *77*, 251–274.

Garbulsky, M. F., Peñuelas, J., Gamon, J., Inoue, Y., & Filella, I. (2011). The photochemical reflectance index (PRI) and the remote sensing of leaf, canopy and ecosystem radiation use efficiencies. A review and meta-analysis. *Remote Sensing of Environment*, *115*(2), 281–297.

Katila, M., & Tomppo, E. (2002). Stratification by ancillary data in multisource forest inventories employing k-nearest neighbor estimation. *Canadian Journal of Forest Research*, *32*(9), 1548–1561.

Kilkki, P., & Päivinen, R. (1987). Reference sample plots to combine field measurements and satellite data in forest inventory. Department of Forest Mensuration and Management, University of Helsinki. *Research Notes*, *19*, 210–215.

LeMay, V., & Temesgen, H. (2005). Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. *Forest Science*, *51*(2), 109–119.

Lin, Y., & Jeon, Y. (2002). *Random forests and adaptive nearest neighbors. Technical Report No. 1055*. University of Wisconsin.

Magnussen, S., McRoberts, R. E., & Tomppo, E. O. (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.

Magnussen, S., Tomppo, E., & McRoberts, R. E. (2010). A model-assisted kNN approach to remove extrapolation bias. *Scandinavian Journal of Forest Research*, *25*, 174–184.

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*(8), 1107–1115.

Maltamo, M., Malinen, J., Kangas, A., Härkönen, S., & Pasanen, A. -M. (2003). Most similar neighbour based stand variable estimation for use in inventory by compartments in Finland. *Forestry*, *76*, 449–464.

Maselli, F., Chirici, G., Bottai, L., Corona, P., & Marchetti, M. (2005). Estimation of Mediterranean forest attributes by the application of k-NN procedure to multitemporal Landsat ETM+ images. *International Journal of Remote Sensing*, *26*(17), 3781–3796.

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. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. *Forest Ecology and Management*, *272*, 3–12.

McRoberts, R. E., Cohen, W. B., Næsset, E., Stehman, S. V., & Tomppo, E. O. (2010). Using remotely sensed data to construct and assess forest attribute maps and related spatial products. *Scandinavian Journal of Forest Research*, *25*, 340–367.

McRoberts, R. E., Magnussen, S., Tomppo, E. O., & Chirici, G. (2011). Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. *Remote Sensing of Environment*, *115*, 3165–3174.

McRoberts, R. E., Næsset, E., & Gobakken, E. (2015). Optimizing the k-Nearest Neighbors technique for estimation of forest aboveground biomass using airborne laser scanning data. *Remote Sensing of Environment*, *168*, 252–264.

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., & Tomppo, E. O. (2007). Remote sensing support for national forest inventories. *Remote Sensing of Environment*, *110*, 412–419.

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

Metinfo (2007). Metinfo-forest information services, Finnish Forest Research Institute. Available at: http://www.metla.fi/metinfo/index-en.htm (Last accessed, October 2015)

Metla (2013). File service for publicly available data. Available at: http://kartta.metla.fi/index-en.html (Last accessed: October 2015)

Moeur, M., & Stage, A. R. (1995). Most similar neighbor — An improved sampling inference procedure for natural resource planning. *Forest Science*, *41*(2), 337–359.

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, U.S.A. *Canadian Journal of Forest Research*, *32*, 725–741.

Ohmann, J. L., Gregory, M. J., & Roberts, H. M. (2014). Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping. *Remote Sensing of Environment*, *151*, 3–15.

Poso, S. (1972). A method of combining photo and field samples in forest inventory. *Communicationes Instituti Forestalis Fenniae*, *76*, 1–133.

Pretsch, H. (1997). Analysis and modeling of spatial stand structures. Methodological considerations based on mixed beech–larch stands in Lower Saxony. *Forest Ecology and Management*, *97*(3), 237–253.

Reese, H., Granqvist-Pahlén, T., Egberth, M., Nilsson, M., & Olsson, H. (2005). Automated estimation of forest parameters for Sweden using Landsat data and the *k*NN algorithm. *Proceedings of the 31st International Symposium on Remote Sensing. June 20–24, 2005. St. Petersburg, Russia* (Available at: http://www.isprs.org/proceedings/2005/ISRSE/html/papers/652.pdf. Last accessed: October 2015).

Short, R., & Fukunaga, K. (1981). The optimal distance measure for nearest neighbor classification. *IEEE Transactions on Information Theory*, *27*(5), 622–627.

SLU Forest Map (2013). Available at: http://www.slu.se/en/collaborative-centres-and-projects/swedish-national-forest-inventory/services-and-products/interactive-services/knn-sweden/ (Last accessed: October 2015)

Tokola, T., Pitkanen, J., Partinen, S., & Muinonen, E. (1996). Point accuracy of a non-parametric method in estimation of forest characteristics with different satellite materials. *International Journal of Remote Sensing*, *17*, 2333–2351.

Tomppo, E. (1990). Designing a satellite image-aided national forest survey in Finland. *Proceedings: The usability of remote sensing for forest inventory and planning. Report*, *4*. (pp. 43–47). Swedish University of Agricultural Sciences, Remote Sensing Laboratory.

Tomppo, E. (1991). Satellite image-based national forest inventory of Finland. *Proceedings of the Symposium on Global and Environmental Monitoring, Techniques and Impacts. 17–21 September 1990. Victoria, British Columbia, Canada. International Archives of Photogrammetry and Remote Sensing*, *28*. (pp. 419–424).

Tomppo, E., & Halme, M. (2004). Using coarse scale forest variables as ancillary information and weighting of k-NN estimation: A genetic algorithm approach. *Remote Sensing of Environment*, *92*, 1–20.

Tomppo, E., Gagliano, C., De, Natale, Katila, F. 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.

Tomppo, E., Haakana, M., Katila, M., & Peräsaari, J. (2008). *Multi-source national forest inventory — Methods and applications. Managing Forest Ecosystems*, *18*, Springer (374 p).

Tomppo, E., Nilsson, M., Rosengren, M., Aalto, P., & Kennedy, P. (2002). Simultaneous use of Landsat-TM and IRS-1C WiFS data in estimating large area tree stem volume and aboveground biomass. *Remote Sensing of Environment*, *82*(1), 156–171.

Trotter, C. M., Drymond, J. R., & Goulding, C. J. (1997). Estimation of timber volume in a coniferous plantation forest using Landsat TM. *International Journal of Remote Sensing*, *18*, 2209–2223.

Wilson, B. T., Lister, A. J., & Riemann, R. I. (2013a). *Live tree species basal area of the contiguous United States (2000–2009)*. Newtown Square, PA: USDA Forest Service, Northern Research Station (Available at: http://www.fs.usda.gov.rds/archive/Product/RDS-2013-0013. Last accessed: October 2015).

Wilson, B. T., Woodall, C. W., & Griffith, D. M. (2013b). Imputing forest carbon stock estimates from inventory plots to a nationally continuous coverage. *Carbon Balance and Management*, *2013*(8), 1.

Wilson, B. T., Lister, A. J., & Riemann, R. I. (2013c). *Forest carbon stocks of the contiguous United States (2000–2009)*. Newtown Square, PA: USDA Forest Service, Northern Research Station (Available at: www.fs.usda.gov/rds/archive/Product/RDS-2013-0004. Last accessed: October 2015).

Wilson, B. T., Lister, A. J., & Riemann, R. I. (2012). A nearest neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. *Forest Ecology and Management*, *271*, 182–198.

Zolkos, S. G., Goetz, S. J., & Dubayah, R. (2013). A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sensing of Environment*, *128*, 289–298.