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 13 February 2016

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.