Remote Sensing of Environment 152 (2014) 29–37

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Mapping by spatial predictors exploiting remotely sensed and ground data: A comparative design-based perspective

Piermaria Corona [a,*], Lorenzo Fattorini [b], Sara Franceschi [b], Gherardo Chirici [c], Fabio Maselli [d], Luca Secondi [e]

[a] Consiglio per la ricerca e la sperimentazione in agricoltura, Forestry Research Centre (CRA-SEL), Arezzo, Italy
[b] Department of Economics and Statistics, University of Siena, Siena, Italy
[c] DiBT, University of Molise, Pesche, Italy
[d] IBIMET-CNR, Firenze, Italy
[e] Department for Innovation in Biological, Agro-food and Forest Systems (DIBAF), University of Tuscia, Viterbo, Italy

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 16 May 2014
Accepted 17 May 2014
Available online xxxx

Keywords:
Forest mapping
Tessellation stratified sampling
Universal kriging
Cokriging
Locally weighted regression
k-Nearest neighbor predictor
Simulation

ABSTRACT

This study was designed to compare the performance – in terms of bias and accuracy – of four different parametric, semiparametric and nonparametric methods in spatially predicting a forest response variable using auxiliary information from remote sensing. The comparison was carried out in simulated and real populations where the value of response variable was known for each pixel of the study region. Sampling was simulated through a tessellation stratified design. Universal kriging and cokriging were considered among parametric methods based on the spatial autocorrelation of the forest response variable. Locally weighted regression and *k*-nearest neighbor predictors were considered among semiparametric and nonparametric methods based on the information from neighboring sites in the auxiliary variable space. The study was performed from a design-based perspective, taking the populations as fixed and replicating the sampling procedure with 1000 Monte Carlo simulation runs. On the basis of the empirical values of relative bias and relative root mean squared error it was concluded that universal kriging and cokriging were more suitable in the presence of strong spatial autocorrelation of the forest variable, while locally weighted regression and *k*-nearest neighbors were more suitable when the auxiliary variables were well correlated with the response variable. Results of the study advise that attention should be paid when mapping forest variables characterized by highly heterogeneous structures. The guidelines of this study can be adopted even for mapping environmental attributes beside forestry.

1. Introduction

Sound management of forest areas requires accurate information regarding the extent, condition and productivity of natural resources. Estimation of these variables is an agreed objective of environmental monitoring programs at a variety of spatial scales. Distinctively, wall-to-wall forest variable maps are important information sources to guide forest and environmental management decisions (e.g., Corona, Chirici, McRoberts, Winter, & Barbati, 2011; Davis, Johnson, Bettinger, & Howard, 2001).

Traditionally, data on forests and forestry are collected by means of user-driven sampling procedures, in the context of forest inventories (Corona & Marchetti, 2007). In turn, forest inventories are usually performed using probabilistic sampling schemes in which a set of plots is selected from the investigated area in accordance with some spatial design. It should be noticed that the purely random scheme is likely to produce unsuitable voids in the study area. To ensure that plots are evenly spread over the region, stratified schemes are preferred. One such scheme involves tessellation of the study area by means of regular polygons of equal size (e.g., the national forest inventories of USA and Italy), each of them containing at least a portion of the study area, and then the random selection of one point in each of these polygons. This scheme is usually referred to as the tessellation stratified sampling (TSS).

Remote sensing data acquired from aerial and space platforms can support modern forest inventories as *i*) auxiliary information to improve the precision of inventory estimates, and *ii*) auxiliary information for forest variable mapping (Corona, 2010; McRoberts & Tomppo, 2007). In most situations, remotely sensed data from multispectral optical or radar images or metrics calculated from airborne laser scanning (ALS) are available for all the pixels of the investigated area, in contrast with the values of forest variables which are known only for the sampled portion of the area (McRoberts, Cohen, Naesset, Stehman, & Tomppo, 2010).

Several statistical techniques employing information coupled from remotely sensed imagery and ground data are available for predicting forest variables for the unsampled portion of the area and thus deriving a wall-to-wall map. All these techniques involve the use of spatial predictors based on a set of assumptions generating the probability distribution of the forest variable over the study area and, at the same time,

* Corresponding author.

http://dx.doi.org/10.1016/j.rse.2014.05.011
0034-4257/

determining the statistical properties of the predictors (model-based inference). Among them, the nonparametric technique usually referred to as the *k*-nearest neighbor (*k*-NN) method is likely the most utilized in forest operational cases (Corona et al., 2011; McRoberts & Tomppo, 2007). Such a method joins a conceptual simplicity with an easy and efficient applicability, but is sub-optimal in exploiting the spatial autocorrelation which generally characterizes both ground and remote sensing data. Parametric and semiparametric techniques such as ordinary kriging (OK), universal kriging (UK), cokriging (CK) and locally weighted regression (LWR) should theoretically overcome this drawback, but their use in forest surveys is still relatively limited.

In this framework it is worth noting that traditional spatial predictors assume sampling locations either fixed or stochastically independent from the forest variable under study. In the latter situation, the sampling design is said to be ignorable (Rubin, 1976) and the sampling locations can be nicely treated as if they were fixed. On the other hand, under preferential sampling schemes, e.g. when a forest variable is more intensively sampled near sites where it is likely to be greater, traditional spatial predictors adopted by ignoring preferential sampling can lead to misleading, highly-biased inferential conclusions (Diggle, Menezes, & Su, 2010; McArthur, 1987). In these cases, the traditional techniques should be modified adopting the proposal by Diggle et al. (2010), which is computationally intensive and not easy to use if one has to perform predictions for all the unobserved locations in order to construct a map. Fortunately, the customary use of TSS in forest inventories ensures sampling locations independent from forest variables in such a way that traditional prediction procedures can be straightforwardly applied for mapping as if locations were purposively selected. Independence from sampling locations is also ensured by systematic sampling schemes which in forest inventories constitute the most common alternatives to TSS.

The goal of this study is to compare the performance of UK, CK, LWR and k-NN in mapping a forest variable when remotely sensed data are used as auxiliary information and TSS is performed to achieve ground data. The statistical properties of these methods are empirically evaluated from a design-based perspective by means of a simulation study in which a set of artificial forest stands and a real stand are considered and left fixed and TSS is replicated at each simulation run. The results are presented to give advice in the selection of the most suitable prediction methods in relationship with the level of spatial autocorrelation of the forest variable and the level of correlation with auxiliary variables from remote sensing.

2. Previous investigations

While k-NN has been widely tested for the prediction of forest attributes both in Europe and North America (see for ex. Franco-Lopez, Ek, & Bauer, 2001; Katila & Tomppo, 2001), this is less the case for the other prediction methods considered. In particular, few studies have been devoted to inter-comparing the performances of these methods for forest applications.

Hudak, Lefsky, Cohen, and Berterretche (2002) found that ordinary least squares regression worked better than OK and CK for mapping forest canopy height on the basis of Lidar and Landsat 7 ETM + imagery. Wallerman, Joyce, Vencatasawmy, and Olsson (2002) found that the presence of spatial discontinuities was decisive for the reliability of kriging in mapping forest standing volume.

A comparative investigation of the predictors of forest standing volume in Central Italy was carried out by Maselli and Chiesi (2006) using ground data from a regional forest inventory and Landsat ETM + imagery: this experiment showed that LWR can provide an accuracy comparable to that of *k*-NN and kriging. Freeman and Moisen (2007) also found that the improvement brought by kriging over remote sensing based nonparametric prediction methods was not clearly defined.

Meng, Cieszewski, and Madden (2009) inter-compared the performance of univariate kriging (OK, and UK) and multivariable kriging (CK and regression kriging) to predict basal area of pine stands using Landsat ETM + images as auxiliary data. Regression kriging resulted in the smallest errors and the highest R^2. Similar results were obtained by Viana, Aranha, Lopes, and Cohen (2012), who inter-compared spatial predictions of forest above ground biomass obtained using various methods, among which OK and RK, using remotely sensed data as auxiliary information: as the tested forest variables showed low spatial autocorrelation, OK was less effective than RK.

Chen, Zhao, McDermid, and Hay (2012) investigated the influence of sampling density on the interpolation of canopy height data using optical satellite imagery as auxiliary variables: LWR outperformed OK and CK and preserved patterns of geographic features better than these techniques at most sampling densities.

Tsui, Coops, Wulder, and Marshall (2013) integrated airborne Lidar and space-borne radar to predict above-ground biomass finding that UK works better than CK. Torresan, Strunk, Zald, Zhiqiang, and Cohen (2014) found that *k*-NN worked better than parametric regression on the basis of Lidar metrics.

At least to our knowledge, very few investigations have been devoted to the comparative assessment of spatial predictors in study areas where the response variable is known in each pixel.

Recently, Ver Hoef and Temesgen (2013) carried out the comparison of spatial linear predictors including OK and UK and *k*-NN: using a wide variety of artificial populations and from re-sampling from real forestry data, the authors demonstrated the superiority of spatial linear predictors over *k*-NN predictors in terms of empirical RRMSEs.

This brief review of existing literature indicates that the matter is still controversial and no definitive consensus has been reached on the actual advantages and limitations of these prediction methods.

3. Materials and methods

3.1. Preliminaries and notations

Consider a delineated study area partitioned into a population U of N pixels. y_j denotes the value of the *response variable* Y at pixel j. Frequently used response variables in forestry are the proportion of forest area, the growing stock volume (m^3/ha), the basal area (m^2/ha) or the stem density (tree count/ha) (e.g. McRoberts, Tomppo, Finley, & Heikkinen, 2007). For a sample S selected from U, the forest response variable is available because it was collected in the field by a sample survey. A very familiar scheme is the so-called TSS, which involves tessellation of the study area by means of rectangles, quadrats, or other regular figures of equal size (each of them containing at least a portion of the study area) and then selecting at random one point in each of these polygons. A wall-to-wall map of the response variable Y is the result of the prediction of y_j for any unsampled pixel $j \in$ U-S.

$\mathbf{x}_j$ denotes the q-vector containing the values of q auxiliary variables $X_1, \ldots, X_q$ which are known for any $j \in$ U. Auxiliary variables in the remote sensing context are the multispectral bands from optical imagery, metrics calculated on the basis of ALS data or other geospatial dataset (for instance, from a Digital Elevation Model). In some studies, these auxiliary variables are known as *feature variables*, while in other cases they are known as *predictive variables* or *covariates*. For any $h \neq j \in$ U, $d_X(j, h)$ denotes the distance between pixels j and h induced by a distance criterion adopted in the space of auxiliary variables. The common criteria are Euclidean distance or some weighted Euclidean distances such as Mahalanobis distance and canonical correlation or canonical correspondence distances.

By means of remotely sensed digital imagery, each plot visited in a forest inventory can be assigned to the pixel containing the plot center. Accordingly, S ⊂ U denotes the sample of the n pixels containing the plot centers. As to the recording of forest variable y_j for any pixel $j \in$ S, it is apparent that plots do not generally coincide with pixels. In many situations, pixels are somewhat larger than plots (for instance, 900 m^2 of Landsat multispectral imagery against 530 m^2 of the standard plot of