Available online at www.sciencedirect.com
ScienceDirect
Remote Sensing of Environment 112 (2008) 862–871
Remote Sensing of Environment
www.elsevier.com/locate/rse

# Smoothing methodology for predicting regional averages in multi-source forest inventory

Petri Koistinen [a,*], Lasse Holmström [b], Erkki Tomppo [c]

[a] Department of Mathematics and Statistics, P.O. Box 68, FIN-00014, University of Helsinki, Finland
[b] Department of Mathematical Sciences, P.O. Box 3000, FIN-90014, University of Oulu, Finland
[c] Finnish Forest Research Institute, Unioninkatu 40 A, FIN-00170 Helsinki, Finland

Received 9 June 2004; received in revised form 28 June 2007; accepted 29 June 2007

**Abstract**

The paper examines alternative non-parametric estimation methods or smoothing methods in the context of the Finnish multi-source forest inventory. It uses satellite images in addition to field data to produce forest variable predictions for regions ranging from the single pixel level up to the national level. With the help of the bias-variance decomposition, the influence of the smoothing parameters on prediction accuracy is considered when the smoother's pixel-level predictions are averaged in order to produce predictions for larger areas. A novel variation of cross-validation, called region-wise cross-validation, is proposed for selecting the smoothing parameters. Experimental results are presented using local linear ridge regression (LLRR), which is a variant of the better known local linear regression method.


*Keywords:* Non-parametric regression; Smoothing parameter selection; Cross-validation; Local linear ridge regression; *k*-nearest neighbor method; Satellite images

## 1. Introduction

The national forest inventories collect country level and large area information about the forests of the country. The number of variables is high, typically 100–400. Traditionally, this task has been carried out using measurements from field plots, but now the forest inventories increasingly have moved towards multi-source forest inventory, in which field data is supplemented by remote sensing data, e.g., from multi-spectral satellite images.

One of the most successful prediction methods in forest inventory is the non-parametric *k*-nearest neighbor (*k*-NN) method. In the Finnish Multi-Source National Forest Inventory it has been in operational use since 1990 (Tomppo, 1991). For the present study, the *k*-nearest neighbor method is regarded as just one out of several possible smoothing or non-parametric regression methods, any of which could, in principle, be applied in forest inventory. In addition to the nearest neighbor estimates, smoothing methods include, e.g., orthogonal series estimators, spline smoothers and local linear regression (e.g., Fan & Gijbels, 1996; Hastie et al., 2001). One related recent development is the non-parametric Bayesian regression method proposed by Taskinen and Heikkinen (2004), which is able to produce predictions and their error estimates by utilizing Markov Chain Monte Carlo (MCMC) techniques.

In the present paper, the smoother predicts a value for the forest variable based on the spectral values associated with the corresponding pixel. In addition to these pixel-level predictions, predictions for the average value of a forest variable over a region larger than a pixel are needed. We compute such predictions straightforwardly by averaging the predictions obtained for all pixels in the region. This manner of combining pixel-level predictions to produce region-level predictions is likely not the optimal way of aggregating them due to the spatial dependence of forest variables. Nevertheless, we use averaging because of its simplicity. There have been attempts, based e.g. on kriging, to exploit the spatial dependency of the forest variables in order to improve the prediction accuracy, see Wallerman (2003) for a review. These attempts have met with

* Corresponding author.
*E-mail addresses:* petri.koistinen@helsinki.fi (P. Koistinen), lasse.l.holmstrom@oulu.fi (L. Holmström), erkki.tomppo@metla.fi (E. Tomppo).

0034-4257/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2007.06.019

only partial success due to the difficulty of modeling the spatial statistics of a real forest.

All smoothing methods depend on one or several smoothing parameters, whose values need to be selected carefully. In doing so one must keep in mind the task for which the smoother is eventually used. Cross-validation is often used for smoothing parameter selection. Its standard version, leave-one-out cross-validation, is only appropriate for assessing pixel-level prediction errors (such as the mean squared error, MSE) and, if blindly used, may suggest quite misleading smoothing parameter values when one predicts mean values within a larger region. One should pay close attention to the prediction bias, too, as has been done in practice (Franco-Lopez et al., 2001; Katila & Tomppo, 2001). In the Appendix, we discuss the relationship between (1) the pixel-level MSE and pixel-level bias and (2) the MSE and bias for region-level prediction, when the region-level prediction is calculated by averaging pixel-level predictions. Unfortunately, that relationship depends both on the spatial statistics of the forest variables and the properties of the smoother. Therefore those results cannot be easily used for selecting a value for the smoothing parameter. Instead, we propose selecting the smoothing parameter using a variant of cross-validation we call region-wise cross-validation, where data belonging to regions whose size approximates the target region are left out.

As for the smoothing methods used, local linear regression has a number of theoretical advantages over nearest neighbor methods. However, preliminary experiments in our application showed that it was not possible to obtain both low pixel-level MSE and low bias with the same smoothing parameter, and therefore the method is not appropriate in this application. Instead, we present results using a variant of local linear regression called local linear ridge regression (LLRR) (Seifert & Gasser, 1996, 2000), where the idea is to introduce an additional smoothing parameter to penalize for too steep slopes in the fitted predictor.

## 2. Materials

The field plot data (from the 9th National Forest Inventory of Finland, NFI9) and remote sensing data have been previously analyzed in Halme and Tomppo (2001). The study area is located in the southwest part of Finland. The main tree species are Scots pine (*Pinus sylvestris* L.), Norway spruce (*Picea abies* (L.) Karst.), birch (*Betula* spp.) and other deciduous species. We used data from 4964 field plots lying entirely on forestry land. The sampling design of NFI9 (Tomppo, 2006) groups the field plots systematically into clusters with 10 or 14 field plots per cluster. Within one cluster, the field plots are located along a rectangular or an L-shaped tract. On the main part of the study area, the distances between the clusters is 6 km × 6 km and, within a cluster, the field plot distance is 250 m. However, on one subarea (with 490 field plots) the cluster distance is 7 km × 7 km and the field plot distance within a cluster is 300 m. A Bitterlich plot with a basal area factor of 2 and with a maximum distance of 12.52 m was employed. Using the measured trees, the mean volumes by tree species ($m^3$/ha) were predicted for the plot.

The remote sensing data consist of spectral channels 1–5 and 7 from two adjacent Landsat 5 TM images from the same date of the summer of 1999. Only cloud-free parts of the images were used. The pixel size in the satellite data is 30 m × 30 m. The original data were rectified to the national coordinate system and re-sampled to a pixel size of 25 m × 25 m. Image data consist of six-dimensional feature vectors attached to each pixel belonging to the forestry land. Each field plot is associated with the image pixel that contains its center point.

The spectral information of a pixel is theoretically taken from an area of about 30 m × 30 m. However, due to the scattering of the light in the atmosphere the information actually comes from a larger area. Hence, the field data is measured from a smaller area than the spectral data. Another source of measurement error arises from the fact that the locations of the field plots are subject to error due to uncertainties in the preprocessing. Halme and Tomppo (2001) reduce the influence of this error source by relocating the field plots using multi-criteria optimization, but in this study we have used the pixels corresponding to the original field plot coordinates.

## 3. Methods

### 3.1. Notation

Let $p$ index pixels and let $x_p$ be the feature vector and $y_p$ the forest variable of interest at pixel $p$. In our examples $x_p$ consists of the six spectral values recorded at pixel $p$, but in principle $x_p$ could include other variables as well, such as the pixel's geographic coordinates, as in Taskinen and Heikkinen (2004) or ancillary information about the large scale variation of key forest variables as in the current operational method (Tomppo & Halme, 2004). While $x_p$ is observed for each pixel in the study area, $y_p$ is observed only at field plot locations, which constitute a sparse subset of all the pixels in the study area. We denote the set of field plot pixels by $F$. When the average value of the forest variable in a region $A$ which lies inside the study area is predicted, the true (but unobserved) value is

$$y(A) = \frac{1}{|A|}\sum_{p\in A} y_p, \tag{1}$$

where $|A|$ is the number of pixels in the region $A$.

The value returned by a smoother $g$ for feature vector $x$ and smoothing parameter $h$ is denoted by $g(x, h)$. The smoothing parameter $h$ is different in different smoothing methods, e.g., in $k$-NN it is $k$, but in LLRR $h$ is the pair $(k, \lambda)$, see Section 3.3. In addition to the arguments $x$ and $h$, the smoother also depends on the training data $(x_p, y_p)$, $p \in F$ although this is suppressed in the notation. As our prediction of the regional average $y(A)$, we use the average of the pixel-level predictions,

$$\hat{y}(A) = \frac{1}{|A|}\sum_{p\in A} g(x_p, h). \tag{2}$$

Implementations of the smoothers used in this paper are available through the web page http://www.rni.helsinki.fi/~pek/software.html.