Generalizing Sample Tree Information with Semiparametric and Parametric Models

Annika Kangas and Kari T. Korhonen

Kangas, A. & Korhonen, K.T. 1995. Generalizing sample tree information with semiparametric and parametric models. Silva Fennica 29(2): 151–158.

Semiparametric models, ordinary regression models and mixed models were compared for modelling stem volume in National Forest Inventory data. MSE was lowest for the mixed model. Examination of spatial distribution of residuals showed that spatial correlation of residuals is lower for semiparametric and mixed models than for parametric models with fixed regressors. Mixed models and semiparametric models can both be used for describing the effect of geographic location on stem form.

Keywords mixed models, models, nonparametric models, semiparametric models, volume, forest inventories.
Authors' addresses *Kangas*, The Finnish Forest Research Institute, Kannus Research Station, P.O. Box 44, FIN-69101 Kannus, Finland; *Korhonen*, The Finnish Forest Research Institute, Joensuu Research Station, P.O. Box 68, FIN-80101 Joensuu, Finland
Fax to *Kangas* +358 68 871 164 **E-mail** annika.kangas@metla.fi
Accepted July 15, 1995

1 Introduction

Volume or biomass of growing stock is one of the parameters to be estimated in most forest inventories. Both these parameters are difficult to measure directly if the population is large. Therefore, two- or multi-phase sampling is applied in most forest inventory systems. The first phase sample consists of a large number of trees for which diameter and other easily measurable characteristics are measured. The second phase sample consists of trees measured more thoroughly. Trees in the first phase sample are called tally trees, and the trees in the second phase are called sample trees. Height, age and additional diameters are the most typical characteristics measured for the sample trees. A third phase sample may be collected to derive volumes or biomass from sample tree characteristics (Cunia 1986).

If two phase sampling is applied, the sample tree information has to be generalized for tally trees. This means that for every tally tree an expected value of each sample tree characteristic with respect to measured characteristics is given. Most methods used are based on regression techniques (Korhonen 1993, Korhonen 1992, Cunia 1986, Kilkki 1979). The advantage of using regression models is that unbiased estimates for the population parameters are easily obtained. In some cases, however, it is difficult to formulate the model so that it is both robust to deviations about the assumed shape and structure and precise enough.

One example of a problem of this kind is describing the effect of geographic location on stem form. In the paper of Korhonen (1993) it was demonstrated that, in Finland, the stem form of Scots pine (*Pinus sylvestris*) depends on the geographic location. A quadratic trend surface, for example, estimated with Ordinary Least Squares (OLS) can be used for describing the effect, but the fit is far from complete.

The Kriging method can be used for modelling spatial relationships (e.g. Ripley 1981). This method gives the best linear unbiased predictors for unobserved values and also provides an estimate of the accuracy of the predictions. Unfortunately, the Kriging method may be impractical for large data sets due to problems related to the inversion of the covariance matrix of observations (e.g. Henttonen 1991).

Nonparametric models can offer a more flexible solution than parametric models. In contrast to parametric models, nonparametric models do not require any assumptions on the shape of the model. Nonparametric models can, however, be difficult to apply in multi-dimensional cases (i.e. when several regressors are necessary). Therefore, in this paper a semiparametric approach (see below) was tested for generalizing sample tree information.

Ojansuu and Henttonen (1983) have applied semiparametric models for predicting the local values of climatic variables. They estimated the climatic variables first with an ordinary regression model. The residuals of the model were then smoothed with moving averages in the neighbourhood of each observation. The main problem in this approach was that the correlations in the observations located in clusters could not be taken into account and the prediction error could not be estimated analytically. In this study semiparametric models are used in a similar manner as in the study of Ojansuu and Henttonen (1983) except that the residuals are smoothed with a kernel method.

The goal of this paper is to test if semiparametric estimators can be applied in large area forest inventories such as the National Forest Inventory (NFI) of Finland. Locally calibrated volume functions are necessary because NFI data are also used for estimating parameters of forests for small areas with help of remote sensing techniques and geographic information systems (Tomppo 1992). The reliability of semiparametric models is compared with parametric estimators: regression models with fixed regressors and mixed models.

2 Material

The pine sample trees measured in the 7th National Forest Inventory (NFI7) of Finland were used in the study. The data consist of 28 575 pines measured from 8472 sample plots during 1977–1983. The following characteristics were used in this study:

- diameter at breast height,
- height of the tree from ground level to top of the tree, and
- upper diameter at the height of 6 meters from the ground.

Sample trees were selected with a relascope (basal area factor 2). From each plot several characteristics describing the site and growing stock were registered (see e.g. Valtakunnan metsien... 1977).

3 Methods

Semiparametric approach

In this paper semiparametric and mixed models were tested for modelling stem volume in large

data sets. Semiparametric model is a combination of an ordinary regression model and nonparametric model. A nonparametric estimate of a variable y at a point i is the weighted average of the measured values of y. The weight of a sample point depends on the differences in the values of the independent variables between the point of interest and the other sample points.

In this study a nonparametric regression model (Nadaraya 1964)

$$\hat{y}(x_j) = \frac{\sum_{i=1}^{n} y_i K\left[\frac{(x_j - x_i)}{h}\right]}{\sum_{i=1}^{n} K\left[\frac{(x_j - x_i)}{h}\right]} \quad (1)$$

was used,

where

- y is the dependent variable,
- x_j (x_i) is a vector containing the values of the independent variables at point j (i),
- K is the kernel function, and
- h is the window-parameter.

The kernel function used was a multivariate normal density function (Silverman 1986):

$$K(x_j) = (2\pi)^{-\left(\frac{d}{2}\right)} \exp\left(-\frac{1}{2} x_j' x_j\right) \quad (2)$$

where d is the dimension of the distribution (number of elements in vector x_j in Formula 1).

Other estimators for nonparametric models using kernel functions have been presented by Gasser and Müller (1979) and Priestley and Chao (1972). Also, many other kinds of kernel functions have been presented (see e.g. Härdle 1989). Other kinds of nonparametric models can also be found in the literature, for example, models based on spline functions (e.g. Silverman 1985), nearest neighbour regression (Altman 1992) or local regression functions (Müller 1987).

One problem in applying the kernel method is to select the window-parameter, h. The window-parameter determines the degree of smoothing: the larger the window, the smoother the nonparametric model. For choosing the window-parameter, mean square error is often used as a criteria. The MSE for a given x is defined as

$$MSE(x,h,n) = E\left(\hat{f}_{h,n}(x) - f(x)\right)^2 \quad (3)$$

where f(x) is the value of true function (Altman 1990). The value of true function is unknown except for the design points. The optimal value of h is often considered to be the one that minimizes the average (or total) MSE over the design points.

The use of nonparametric methods in the case of several independent variables is difficult. When the number of independent variables increases, the data set may be surprisingly sparsely distributed in a high-dimensional Euclidean space. Thus, the window-parameter values obtained by minimizing the average MSE are too large to describe the relationships between the dependent and independent variables properly. Further, the model becomes difficult to interpret and impossible to demonstrate in a graphic form (Härdle 1989). The model may remain as a 'black box' because no simple numeric presentation for a nonparametric model is available.

Several methods for overcoming this problem have been presented, for example, by considering the linear combinations of the independent variables (see Härdle 1989). One obvious solution for the problem of several independent variables is to use semiparametric methods, i.e. combination of parametric and nonparametric methods. Semiparametric approaches have been used, for instance, by Carroll and Härdle (1989) and Engle et al. (1986).

Mixed model approach

Random parameter models and mixed models have been successfully applied for obtaining reliable 'locally calibrated' estimates (Lappi 1986, Lappi 1991, Ojansuu 1993). A general formulation of a mixed linear model is

$$y = Xb + Zc + e \quad (4)$$

where

- y is data vector of the dependent variable,
- X and Z are data matrices related to fixed and ran-

dom regressors, respectively,
e is random error (vector) with E(e) = 0 and var(e) = R,
b is a fixed parameter vector, and
c is the random parameter vector with E(c) = 0 and var(c) = D. The random parameters (c) and random error (e) are mutually independent.

In this study a variance component model (5) was applied.

$$y_{ij} = x'_{ij}b + c_i + e_{ij} \quad (5)$$

where
y_{ij} is the sample tree characteristics of interest,
e_{ij} is a random tree effect with $var(e_{ij}) = \sigma_e^2$ and $E(e_{ij}) = 0$, and
c_i is a random plot effect with $var(c_i) = \sigma_c^2$ and $E(c_i) = 0$, and c_i and e_{ij} are independent.

This means that a within-plot correlation

$$r = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2}$$

is assumed. In the application stage plot effects of the mixed model can be estimated with

$$\hat{c}_i = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2}\left(y_{ij} - \hat{y}_{ij}\right)$$

if one sample tree from each plot is measured. The last term is the residual of the fixed part of model (5) for the measured tree.

Reliability of the estimators

The fixed part of the mixed model and its variance components and the parametric part of the semiparametric model were estimated using the whole data set. At the application stage the reliability of different estimators was tested by dividing the data into two parts. One tree per plot was used for estimating the plot effect (c_i) of the mixed model. In the semiparametric approach the residuals of the fixed part were smoothed with (1) using one tree per plot as a data set. The rest of the data were used for studying the reliability of the model.

The reliability of the models was studied with the help of two criteria. Visual examination of the spatial distribution of residuals was the main criteria. Residual maps were plotted for studying the residuals (see Korhonen 1993). For plotting the maps, mean values of residuals was calculated for each tract (cluster of 21 sample plots) of the NFI data. A weighted mean value of the tract mean and neighbouring tracts were calculated to smooth the data (Korhonen 1993). The purpose of the smoothing was only to make the possible spatial correlation more visible.

Another criteria used for studying the reliability of the models was the MSE values. MSE was estimated as the mean value of the squared differences between estimated and observed volumes of trees not used as data points for the model calibration (i.e. estimation of the nonparametric component of the semiparametric model and estimation of plotwise random effect of the mixed model).

4 Results

Model (6) was used as the parametric part of the semiparametric model. The OLS (ordinary least squares) estimates of the parameter values from the paper of Korhonen (1993) were used.

$$v_{ij} / d_{ij}^2 = b_0 + b_1 d_{ij} + b_2 d_{ij}^2 + b_3 \ln(G_j) + e_{ij} \quad (6)$$

where
v_{ij} = volume of tree i on plot j, dm^3
d_{ij} = diameter at 1.3 m height, cm
$\ln(G_j)$ = natural logarithm of the basal area of the growing stock of the plot j, m^2/ha.

The only independent variables in the nonparametric part of the model were the coordinates. 4 km was found to be the optimal value of the window parameter when one tree per plot were used for estimating the nonparametric part of the model. The optimal value was obtained with cross validation (e.g. Altman 1990).

Model (7) was used as the fixed part of the

Table 1. GLS-estimates of parameters and variance components for model (7).

Regressor	Parameter estimate
Constant	0.0582829
d	0.0218561
d^2	–0.0002608
ln(G)	0.0629807
RDIST	–0.0562567
YC	–0.0822138
YC^2	0.0451772
XC	0.3012367
XC^2	–0.1332506
XC · YC	–0.2176091
Variance of random plot effect (σ_c^2)	0.0048178
Variance of random tree effect (σ_e^2)	0.0038560

mixed model. The values of the parameters were estimated with the GLS (generalized least squares) method (Table 1).

$$v_{ij} / d_{ij}^2 = b_0 + b_1 d_{ij} + b_2 d_{ij}^2 + b_3 \ln(G_j) + b_4 RDIST_j + b_5 YC_j + b_6 YC_j^2 + b_7 XC_j + b_8 XC_j^2 + b_9 XC_j \cdot YC_j + c_j + e_{ij} \quad (7)$$

where
RDISTj = 0 if $DIST_j > 2$
1 / ($DIST_j$ + 0.2), otherwise
$DIST_j$ = distance of the plot j from the coast (of Gulf Botnia or Finnish Gulf), km
YC_j = $(Y_j - 6620) / 1000$
Y_j = y-coordinate of the plot (distance from the Equator), km
XC_j = $(X - 60) / 1000$, and
X_j = x-coordinate of the plot (distance from the Greenwich meridian), km.

When the semiparametric model with one tree from each plot as data points were tested, MSE was 0.00779 $(dm^3/cm)^2$ for the rest of the trees in the data. The map of the residuals of the semiparametric model is in Fig. 3.

In Fig. 2 is the map of residuals of the fixed part of the mixed model (7). In this model, coordinates have been used for forming a second order trend surface. The MSE for the fixed part of the model was 0.01036 $(dm^3/cm)^2$.

Fig. 1. Residuals of the calibrated mixed model in the NFI7 data. Blue colours indicate negative residuals and red colours positive residuals (the darker the colour, the greater the absolute value). Yellow colour indicates residuals close to zero. White colour is for missing values.

An MSE value of 0.00644 $(dm^3/cm)^2$ was obtained when mixed model (7) and one tree per plot for estimating the random plot effect were applied. The map of residuals for the mixed model is in Fig. 1.

5 Discussion

MSE is lowest for the mixed model and highest for the parametric model with fixed regressors without plotwise calibration. The negligible differences in MSE values of different models are meaningless when considering the application of the models. Comparison of residual maps shows

Fig. 2. Residuals of the parametric model with fixed regressors in the NFI7 data. Blue colours indicate negative residuals and red colours positive residuals. Yellow colour indicates residuals close to zero. White colour is for missing values.

Fig. 3. Residuals of the semiparametric model in the NFI7 data. Blue colours indicate negative residuals and red colours positive residuals. Yellow colour indicates residuals close to zero. White colour is for missing values.

that there are, however, significant differences in the estimators. There is clear geographic correlation in the residuals of fixed model (Fig. 2). Spatial correlation is clearly lower for the semiparametric and mixed estimators (Figs. 1 and 3). The semiparametric model seems to be more efficient than mixed model in local calibration – the difference is, however, quite small.

Both mixed model and semiparametric approach can be considered as simplified versions of the Kriging method. The differences between the 'real' Kriging and the methods used in this study are that

1) in the mixed model approach the covariances between data points are restricted to a single sample plot whereas in the Kriging method the covariance depends on distance and is continuous;
2) in the semiparametric approach the covariance structure of observations is described implicitly in a nonparametric model whereas in the Kriging the covariance structure is modeled with a parametric function.

In this study, 4 km was found optimal value for the window parameter of the nonparametric part of the semiparametric model. Since the distance between clusters in the NFI data were 8 km, the window-parameter value 4 km means that sample trees measured at neighbouring clusters have only negligible weight when the residual of the parametric part of the semiparametric model is estimated for a tree. With the multivariate normal kernel function with window width 4, for example, the weight of a neighbouring tree from a same sample plot is 0.63 and from an adjacent cluster it is 0.0002.

With small window parameter values the nonparametric model practically interpolates through the data points. In the studied application, however, this was not a problem, because the goal was to generalize the sample tree characteristics for the tally trees. It may even be desirable to use quite small window parameter, since in this way the natural variation of sample trees remains in the generalization process (see e.g. Holm et al. 1979).

It would be possible, and in some cases even desirable, to estimate the parametric and nonparametric parts of the semiparametric model with the same optimization procedure. In this study, however, the parametric part was estimated first and the nonparametric approach was used to smooth the residuals of the parametric model. This guarantees that the parametric part is unbiased and it can be used also independently of the nonparametric smoothing. Also, the relationship between tree volume and breast height diameter can be described quite well with parametric functions whereas the spatial relationships are difficult to describe.

In this study it was assumed that the random effect of the mixed model is estimated by measuring one sample tree for each plot. In practical applications this would require many more sample tree measurements than are currently measured in the NFI of Finland. The semiparametric approach offers a statistically sound method to utilize sample trees measured at neighbouring plots. Therefore, it is not necessary to measure sample trees on each plot when semiparametric models are applied. Semiparametric models require massive computations. It was not found to be too big problem in this NFI application. The results in this paper show that semiparametric models can be recommended for smoothing of geographic trends in large data, such as NFI data.

References

Altman, N.S. 1990. Kernel smoothing of data with correlated errors. Journal of American Statistical Association 85: 749–759.

— 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46: 175–185.

Carrol, R.J. & Härdle, W. 1989. Second order effects in semiparametric weighted least squares regression. Statistics 20: 176–186.

Cunia, T. 1986. Error of forest inventory estimates: its main components. In: Estimating tree biomass regressions and their error. Proceedings of the Workshop on Tree Biomass Regression Functions and their Contribution to the Error of Forest Inventory Estimates. May 26–30, 1986. Syracuse, New York. p. 1–13.

Engle, R.F., Granger, G.W., Rice, J. & Weiss, A. 1986. Semiparametric estimates of the relation between weather and electricity sales. Journal of American Statistical Association 81(394): 310–320.

Gasser, T. & Müller, H.G. 1979. Kernel estimation of regression functions. In: Gasser, T. & Rosenblatt, M. (eds.). Smoothing techniques for curve estimation. Springer-Verlag, New York. p. 23–68.

Härdle, W. 1989. Applied nonparametric regression. Cambridge University Press. 323 p.

Henttonen, H. 1991. Kriging in interpolating July mean temperatures and precipitation sums. Reports from the Department of Statistics, University of Jyväskylä, Finland, no. 12. 42 p.

Holm, S., Hägglund, B & Mårtensson A. 1979. En metod för generalisering av Riksskogstaxeringens provträdsdata. Summary: A method for generalisation of sample tree data from the Swedish National Forest Survey. Swedish University of Agricultural Sciences, Department of Forest Survey, Report 26. 94 p.

Kilkki, P. 1979. An outline for a data processing system in forest mensuration. Silva Fennica 13(4): 368–384.

Korhonen, K.T. 1992. Calibration of upper diameter models in large scale forest inventory. Silva Fennica 26(4): 231–239.

— 1993. Mixed estimation in calibration of volume functions of Scots pine. Silva Fennica 27(4): 269–276.

Lappi, J. 1986. Mixed linear models for analyzing and predicting stem form variation of Scots pine. Communicationes Instituti Forestalis Fenniae 134. 69 p.

— 1991. Calibration of height and volume equations with random parameters. Forest Science 37(3): 781–801.

Müller, H.G. 1987. Weighted local regression and kernel methods for nonparametric curve fitting. Journal of American Statistical Association 82:

231–238.
Nadaraya, E.A. 1964. On estimating regression. Theory of Probability Application 9: 141–142.
Ojansuu, R. 1993. Prediction of Scots pine increment using a multivariate variance component model. Acta Forestalia Fennica 239. 72 p.
— & Henttonen, H. 1983. Kuukauden keskilämpötilan, lämpösumman ja sademäärän paikallisten arvojen johtaminen Ilmatieteen laitoksen mittaustiedoista. Summary: Estimation of the local values of monthly mean temperature, effective temperature sum and precipitation sum form the measurements made by the Finnish Metereological Office. Silva Fennica 17(2): 143–157.
Priestley, M.B. & Chao, M.T. 1972. Non-parametric function fitting. Journal of Royal Statistical Society, Series B 34: 385–392.
Ripley, B.D. 1981. Spatial statistics. John Wiley & Sons Inc., New York. 252 p.
Silverman, B.W. 1985. Some aspects of the spline smoothing approach to nonparametric curve fitting. Journal of Royal Statistical Society, Series B 47: 1–52.
— 1986. Density estimation for statistics and data analysis. Chapman & Hall, London.
Tomppo, E. 1992. Multi-source National Forest Inventory of Finland. In: Nyyssönen, A., Poso, S. & Rautala, J. (eds.). Proceedings of Ilvessalo Symposium on National Forest Inventories. Finland 17–21 August 1992. The Finnish Forest Research Institute, Research Papers 444: 52–59.
Valtakunnan metsien 8. inventointi. Kenttätyön ohjeet. Pohjois-Karjalan versio. [Field instructions for the 8th National Forest Inventory of Finland. In Finnish.] The Finnish Forest Research Institute. Helsinki. 96 p.

Total of 24 references