Silva Fennica 1993, Vol. 27 N:o 4: 269–276

Mixed estimation in calibration of volume functions of Scots pine

Kari T. Korhonen

TIIVISTELMÄ: SEKAESTIMOINTI MÄNNYN TILAVUUSYHTÄLÖIDEN KALIBROINNISSA

Korhonen, K.T. 1993. Mixed estimation in calibration of volume functions of Scots pine. Tiivistelmä: Sekaestimointi männyn tilavuusyhtälöiden kalibroinnissa. Silva Fennica 27(4): 269–276.

Regression models for estimating stem volume of Scots pine were constructed using sample tree data measured in the 7th and 8th National Forest Inventory of Finland. Stem volume were regressed on diameter, basal area of growing stock, and geographic location. The results of the study show that using second order trend surface to describe the geographic variation of the residuals gives satisfactory results. Using mixed estimation for combining old and new sample tree data improves the efficiency of an inventory. The weight of the prior information must be low, because remarkable difference in stem form was found in the two inventories.

Tutkimuksessa laadittiin valtakunnan metsien 7:nnen ja 8:nnen inventoinnin koepuuaineistosta regressiomallit männyn tilavuuden estimoimiseksi. Työssä tarkasteltiin sekaestimointia kahden inventointiaineiston koepuumittausten yhdistämisessä. Tutkimus osoittaa, että tilavuusmallin alueellista jäännösvaihtelua voidaan kuvata tyydyttävästi toisen asteen trendipinnalla. Vanhan koepuuaineiston hyödyntämisellä voidaan tehostaa inventoinnissa käytettäviä malleja. Koska inventointien välissä näyttää tapahtuneen merkittävä muutos mäntyjen runkomuodossa, ennakkoinformaatiolle on annettava pieni paino. Lisätutkimuksia tarvitaan mallien kalibroimisessa pienille alueille ja puun runkomuodon muutoksista inventointien välillä.

Keywords: forest inventories, volume, models, estimation, *Pinus sylvestris*.
FDC 524

Author's address: The Finnish Forest Research Institute, Joensuu Research Station. P.O. Box 68, FIN-80101 Joensuu, Finland.

Accepted January 21, 1994

Notation

[recording units in brackets]

d	= diameter at breast height (1.3 meters from the ground level) [cm]
d_6	= diameter measured at the height of 6 meters [cm]
h	= height of a tree [dm]
v	= volume of a tree [dm^3]
G	= basal area of the growing stock [m^2/ha]
DIST	= distance from the coast (of Gulf Botnia or Finnish Gulf) [km]
RDIST	= 0, if DIST > 20 1 / DIST – 0.05, otherwise
Y	= p-coordinate of the plot (distance from the Equator) [km]
YC	= (Y – 6620) / 1000
X	= i-coordinate of the plot (distance from the Greenwich meridian) [km]
XC	= (X – 60) / 1000
ln(X)	is natural logarithm of variable X

1 Introduction

Two phase sampling is generally applied in forest inventories. At the first phase a large number of trees, so called tally trees, is measured for diameter and other easily measurable variables. At the second phase, a relatively small number of trees, so-called sample trees, are selected for further measurements. The goal of the measuments at the second phase is to estimate the relationship between the easily measurable tally tree variables (e.g diameter) and the variables of final interest (e.g. stem volume) (see e.g. Cunia 1986, Kilkki 1979).

When inventories are repeated it is customarily to select new sample trees in each inventory occasion. Both theoretical studies on methods involved (e.g. Theil and Goldberger 1961, Johnston 1972, Teräsvirta 1981, Pekkonen 1983, Meng et al. 1990, Lappi 1991) and empirical results on the use of different methods in combining prior information with new measurements has been published (e.g. Pekkonen 1983, Green and Strawdeman 1985, Burk and Ek 1982, Meng et al. 1990, Korhonen 1992). However, there is only little knowledge of how applicable different methods are for example in the National Forest Inventory (NFI) of Finland.

The purpose of this study is to investigate mixed estimation in the use of sample tree data in the NFI of Finland. This study is limited to Scots pine (*Pinus sylvestris* L.).

2 Material and methods

2.1 Study material

Two data sets were used in the study. The first data set was the pine sample trees measured in the 7th National Forest Inventory (NFI7) of Finland. The second data set was the pine sample trees measured in the 8th National Forest Inventory (NFI8) in National Board Districts Eteläkarjala and Pirkka-Häme (see Fig. 1). The sample tree data of NFI7 consists of 28575 pines measured during 1977–1983. The sample tree data of NFI8 consists of 1157 pines in Eteläkarjala and 1226 pines in Pirkka-Häme. The NFI8 data were measured in years 1986–1987.

From each sample tree in both data sets the following measured dimensions were used in this study (recording units in brackets):

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

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... 1988).

2.2 Construction of volume functions

For each sample tree the stem volume was estimated with volume functions of Laasasenaho (1982) using d, d_6 (for trees higher than 8 meters, only), and h as independent variables. Estimated volumes were used as 'true' values when volume functions for tally trees were established.

Models for estimating stem volume as a function of diameter and variables describing the growing stock and location were constructed at two stages. At the first stage the sample tree data of NFI7 were used for finding the form of the model and for calculating first-level estimates of the parameters. At the second stage second-level (localized) estimates of the model parameters were estimated using the data of NFI8.

At Stage 1 ordinary least squares technique was applied for choosing the independent variables and the form of the volume function. At Stage 2 the volume function was re-estimated using mixed estimation (Theil and Goldberger 1961, Johnston 1972). In this method the first-level information (sample tree data of NFI7) and second-level information (data of NFI8) are combined as follows. Let us assume a regression model:

$$y = X\beta \quad (1)$$

where
y = vector of dependent variables,
X = matrice containing independent variables, and
β = parameter vector.

Let us note the first-level information on y and X with r and R, respectively. Correspondingly, the second-level information on y and X is noted with s and S. The parameters of the model are estimated with Formula (2) using both data sets (Theil and Goldberger 1961, Teräsvirta 1981, Korhonen 1992).

$$\hat{\beta} = (S'S + kR'R)^{-1} (S's + kR'r) \quad (2)$$

where
k = weight of the prior information.

If some of the parameters of the model are estimated using only the first-level data and some of the parameters are estimated with both data sets the model (1) must be re-written as follows.

$$r = R_1\beta_1 + R_2\beta_2 \quad (3)$$

$$s = S_1\beta_1 + S_2\beta_2 \quad (4)$$

where
R_1 and S_1 contain regressors whose parameters are estimated using the first-level information only,
R_2 and S_2 contain regressors whose final parameter estimates are obtained using both first- and second-level information, and
β_1 and β_2 are parameter vectors, respectively.

Then, Formula (5) gives the parameter estimates (Korhonen 1992).

$$\hat{\beta}_2 = (S_2'S_2 + k\,m / n\,R_2'R_2)^{-1} (S_2'u + k\,m / n\,R_2'v) \quad (5)$$

where
m = number of observations in the second-level data,
n = number of observations in the first-level data,
$u = s - S_1\hat{\beta}_1$,
$v = r - R_1\hat{\beta}_1$, and
$\hat{\beta}_1$ = first-level estimate for β_1.

2.3 Analysis of residuals

The geographic variation of residuals was studied by plotting the residuals on a map. The mean value of relative residuals $(v - \hat{v} / \hat{v})$ were calculated for each cluster of sample plots in the NFI7 data. Clusterwise means were smoothed using moving averages -method with formula (6) (Ripley 1981, page 36).

$$\hat{Z}(x) = \sum_{i=1}^{L} W_i \cdot Z(x_i), \quad (6)$$

where
$\hat{Z}(x)$ = smoothed surface,
L = number of points (clusters),
$Z(x_i)$ = observed value at point x_i (mean of residuals for cluster i),
W_i = weight of observation i.

In this study,
$W_i = j_i / J_i$, for the cluster itself and all the 8 neighbouring clusters,
= 0, for all the other clusters,
j_i = number of sample trees at cluster i,
J_i = number of sample trees at cluster i and all the 8 neighbouring clusters.

2.4 Sampling simulations

Use of mixed estimation in calibration of volume functions was studied by sampling simulations in the NFI8 data. In the simulations, height and upper diameter ('true' volume) was assumed to be measured from varying number of trees. These data were used to estimate volume functions for tally trees (trees with d as only measured dimension). Mixed estimation was compared with OLS in estimating the functions. Different values for weight parameter k (see Formula (5)) were tested. Selection of sample trees was done with systematic sampling with random starting point. Sampling with each sample size was repeated 100 times varying the selected sample trees (see Korhonen 1992).

In each sampling simulation RMSE of mean volume estimates of the inventory area and RMSE of volume estimates for single trees were calculated. RMSE's of treewise volume estimates were calculated for relative errors $(= (v - \hat{v}) / \hat{v})$. It should be noticed that the mean volumes estimated in this study are the mean volumes of pines on those sample plots were at least one pine was measured. Therefore, results can not be interpreted as mean volumes of pines on forest land.

3 Volume functions

3.1 First-level estimates

Equations (7) and (8) were chosen to estimate the stem volume of pine.

$$v / d^2 = b_0 + b_1 \cdot d + b_2 \cdot d_2 + b_3 \cdot \ln(G) + b_4 \cdot RDIST + b_5 \cdot YC + b_6 \cdot YC^2 + b_7 \cdot XC + b_8 \cdot XC^2 + b_9 \cdot YC \cdot XC \quad (7)$$

$$v / d^2 = b_{10} + b_{11} \cdot d + b_{12} \cdot d^2 + b_{13} \cdot \ln(G) \quad (8)$$

where

$b_0,...,b_{13}$ are parameters,

v = volume of a tree [dm^3]

d = diameter at breast height (1.3 meters from the ground level) [cm],

G = basal area of the growing stock [m^2/ha],

RDIST = 0, if DIST > 20
1 / DIST – 0.05, otherwise

DIST = distance from the coast (of Gulf Botnia or Finnish Gulf) [km],

YC = (Y – 6620) / 1000,

Y = p-coordinate of the plot (distance from the Equator) [km],

XC = (X – 60) / 1000,

X = i-coordinate of the plot (distance from the Greenwich meridian) [km].

The first-level estimates for the parameters were estimated using the NFI7 data with OLS-technique. Parameter estimates and their t-values are in Table 1.

In Model (8) the effect of geographic location on the volume is described using distance from the coast and a second order trend surface as

Table 1. First-level parameter estimates, t-values and RMSE for volume functions (7) and (8).

	Function (7)		Function (8)	
Variable	Parameter estimate	t-value	Parameter estimate	t-value
constant	–0.055623	9.72	–0.0000578	0.02
d	0.025963	119.41	0.0257744	114.38
d^2	–0.000284	60.31	–0.0002817	57.55
ln(G)	0.052251	68.18	0.0591519	76.55
RDIST	–0.072047	19.58		
YC	0.024861	1.72		
YC^2	–0.066980	1.48		
XC	0.372593	14.60		
XC^2	–0.219698	6.21		
YC · XC	–0.247642	8.24		
	RMSE = 0.093		RMSE = 0.097	

Table 2. Comparison of mean volumes estimated with different models. Notation: measured = measured volumes of trees were used; Model 7a = Model (7) with parameters estimated in NFI7 data from whole country were used; Model 7b = Model (7) with parameters estimated in NFI7 data only from the studied district were used; Model 8 = Model (8) with parameters estimated in NFI7 data from whole country.

District	Measured	Model 7a	Model 7b	Model 8
Etelä-Karjala	72.15	70.58	74.29	75.37
Pirkka-Häme	73.67	71.30	72.48	71.10

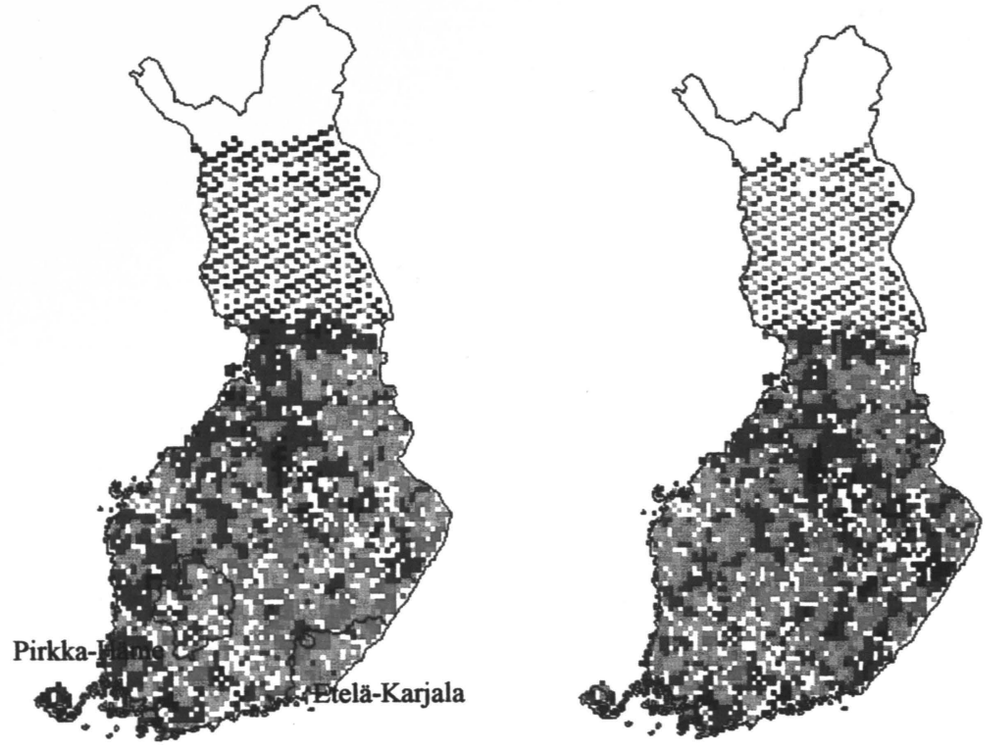

Fig. 1. Residuals of Model (7) in the NFI7 data. Blue colours indicate negative residuals and red colours positive residuals. Yellow colour indicates residuals close to zero. White colour is for missing values.

Fig. 2. Residuals of Model (8) 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.

regressors. Both of these regressors are statistically significant. The usefulness of these variables was further studied by comparing the geographic variation of residuals of Models (7) and (8) (see Figs. 1 and 2). Residual maps show that using geographic variables as regressors clearly diminishes large scale spatial correlation of residuals. For Model (7) the residuals are mostly negative for west and north parts of Finland and mostly positive in east and south-east parts of the country. For Model (8) no clear differences between different parts of the country are found.

In order to study if there are differences in the stem form of NFI7 and NFI8 data mean volumes for Etelä-Karjala and Pirkka-Häme districts were calculated using Models (7) and (8) and NFI7 data. Mean volumes were calculated in 4 different ways:

1. Using measured volumes of trees (= volumes obtained using measured d and h).
2. Using Model (7) with parameters estimated in NFI7 data from whole country.
3. Using Model (7) with parameters estimated in NFI7 data from the studied district, only.
4. Using Model (8) with parameters estimated in NFI7 data from whole country.

Results in Table 2 show that both Models (7) and (8) are remarkably biased in NFI8 data if the parameters are estimated using NFI7 data. When Model (7) with parameters estimated for the whole country is used (7a in Table 2) for Etelä-Karjala district, the bias evidently consists of two components: differencies as a function of location and differencies as a function of time. When Model (8) is used the first component of bias is reduced and results are close to those obtained with Model (7) with parameters estimated for the study area (7b in Table 2). In Pirkka-Häme district the difference between Models (7) and (8) is neglible but the difference

Table 3. Parameter estimates for functions (7) and (8) in the NFI8 data from Etelä-Karjala and Pirkka-Häme districts.

	Function (7)		Function (8)	
	Etelä-Karjala	Pirkka-Häme	Etelä-Karjala	Pirkka-Häme
With prior information				
constant	–0.1707758	–0.1304988	–0.07471644	–0.0603038
d	0.0299622	0.0291630	0.02964140	0.0291997
d^2	–0.0003405	–0.0003033	–0.00033479	–0.0003018
ln(G)	0.0655989	0.0634001	0.07094529	0.0644148
RDIST	–0.0720467	–0.0720467		
YC	0.0248618	0.0248618		
YC^2	–0.0166980	–0.0166980		
XC	0.3725925	0.3725925		
XC^2	–0.2196978	–0.2196978		
YC · XC	–0.2476425	–0.2476425		
Without prior information				
constant			–0.0893162	–0.0695003
d			0.0300710	0.0298306
d^2			–0.0003416	–0.0003105
ln(G)			0.0743722	0.0648874

between the data sets from NFI7 and NFI8 is clear.

3.2 Second-level estimates

Second-level estimates for the parameters of Functions (7) and (8) were calculated using the NFI8 data measured in National Board districts Etelä-Karjala and Pirkka-Häme. Only constant and parameters of d, d^2, and ln(G) were re-estimated. Parameters of Function (8) were estimated also without using NFI7 data as prior information. The parameter estimates are in Table 3.

3.3 Comparison of mixed estimation and OLS in sampling situation

Reliability of the volume functions and effect of sampling error were studied by sampling simulations using the the NFI8 data measured in National Board districts Etelä-Karjala and Pirkka-Häme. Results obtained by calibration of Function (8) with mixed estimation were compared with results obtained by using Function (7) without prior information (= first-level parameter estimates). Different values for weight parameter k were tested. Results using values 0.1 and 0.5 for k are presented in Figs. 3–6.

Figs. 3 and 4 show the expected fact that mixed estimation improves reliability of mean volume estimates most effectively when the number of sample trees is low. In most cases it is more effective to use relatively small weight for prior information. When several hundreds of sample trees are measured the reliability of OLS estimator improves more than the reliability of mixed estimators. Figs. 3 and 4 indicate that if more than 400–500 sample trees are measured, the effect of sampling error (on OLS estimator) is smaller than the effect of the bias in the prior information (on the mixed estimators).

Comparison of RMSE's of treewise volume estimates (Figs. 5 and 6) shows that mixed estimators are superior to OLS in all studied cases. Weight 0.5 gives better results than weight 0.1.

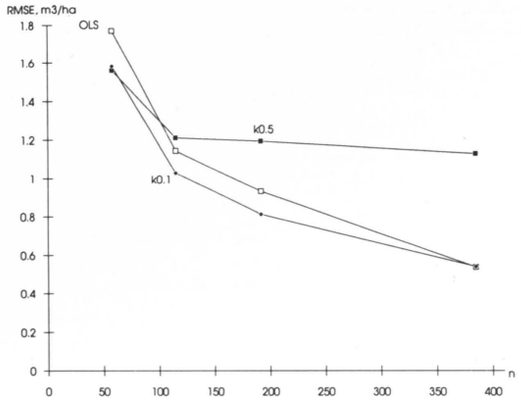

Fig. 3. Root mean square error (RMSE) of mean volume estimate of pines in Etelä-Karjala district as a function of number of sample trees. OLS = ordinary least squares estimator, k0.1 = mixed estimator with k = 0.1, k0.5 = mixed estimator with k = 0.5.

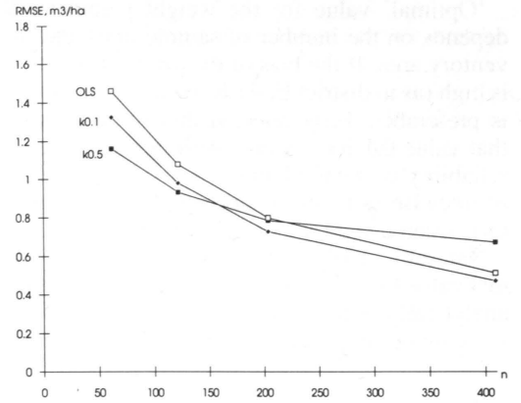

Fig. 4. Root mean square error (RMSE) of mean volume estimate of pines in Pirkka-Häme district as a function of number of sample trees. OLS = ordinary least squares estimator, k0.1 = mixed estimator with k = 0.1, k0.5 = mixed estimator with k = 0.5.

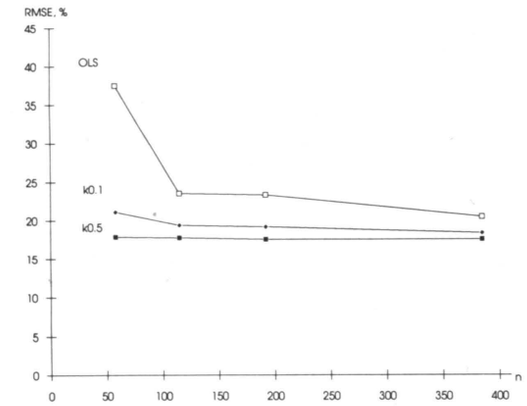

Fig. 5. Root mean square error (RMSE) of treewise volume estimates in Etelä-Karjala district as a function of number of sample trees. OLS = ordinary least squares estimator, k0.1 = mixed estimator with k = 0.1, k0.5 = mixed estimator with k = 0.5.

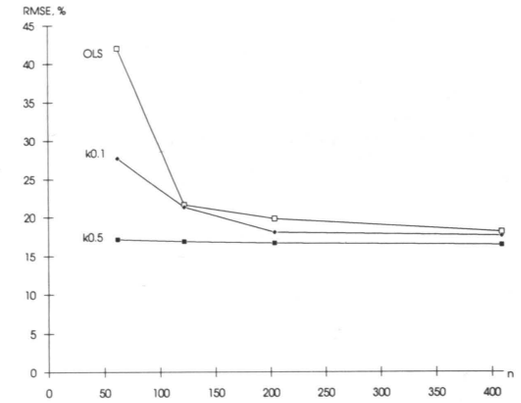

Fig. 6. Root mean square error (RMSE) of treewise volume estimates in Pirkka-Häme district as a function of number of sample trees. OLS = ordinary least squares estimator, k0.1 = mixed estimator with k = 0.1, k0.5 = mixed estimator with k = 0.5.

4 Conclusions

Comparison of Figs. 1 and 2 shows that using second order trend surface as a regressor improves the reliability of the model for different parts of the country. However, Fig. 2 shows that geographic variables can not totally remove the spatial correlation of residuals. Second order trend surfaces can not handle small scale variation – "...trend surface analysis is best at showing broad features of the data" (Ripley 1981, page 35). Moving averages or kriging -methods would be useful in describing small scale geographic variation.

'Optimal' value for the weight parameter k depends on the number of sample trees and inventory area. If the bias of the prior information is high (as in district Etelä-Karjala) small weight is preferable. Tests done in this study showed that value 0.1 for k works well in most cases – reliability of mean volumes is good and RMSE of treewise estimates is not significantly higher than when using for example value 0.5 for k.

Naturally, it is possible to determine the optimal value for the weight of the prior information analytically, also (see e.g. Pekkonen 1983). Formula for optimal value of k would require determination of variance of residuals between sample tree areas. These sample tree areas are not fixed from one inventory to another. Furthermore, analytical formula for optimal k would work only if variation between sample tree areas would not change as a function of time. Therefore, empirical optimisation for k was regarded as a better choice.

Acknowledgements: The author wishes to thank Dr. Risto Ojansuu for his critical comments and suggestions to improve the manuscript.

References

Cunia, T. 1986. Error of forest inventory estimates: its main 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.

Green, E.J. & Strawderman, W.E. 1985. The use of Bayes/empirical Bayes estimation in individual tree volume equation development. Forest Science 31: 975–990.

Burk, T.E. & Ek, A.R. 1982. Application of empirical Bayes/James Stein procedures to simultaneous estimation problems in forest inventory. Forest Science 28: 753–771.

Johnston, J. 1972. Econometric methods. 2nd edition. McGraw-Hill Kogakusha, Ltd., Tokyo. 437 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. Tiivistelmä: Yläläpimittamallien kalibrointi suuralueen metsäninventoinnissa. Silva Fennica 26(4): 231–239.

Laasasenaho, J. 1982. Taper curve and volume functions for pine, spruce and birch. Communicationes Instituti Forestalis Fenniae 108. 74 p.

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

Meng, C.H., Tang, S.Z. & Burk, T.E. 1990. A stochastic restrictions regression model approach to volume equation estimation. Forest Science 36(1): 54–65.

Pekkonen, T. 1983. Leimikon puuston tilavuuden arviointi regressioennustinta käyttäen. [Regression estimators in predicting the volume of the stand marked for cutting. In Finnish]. Metsäntutkimuslaitoksen tiedonantoja 86. 63 p.

Ripley, B.D. 1981. Spatial statistics. John Wiley & Sons, New York. 252 p.

Teräsvirta, T. 1981. Some results on improving the least squares estimation of linear models by mixed estimation. Scandinavian Journal of Statistics 8: 33–38.

Theil, H. & Goldberger, A.S. 1961. On pure and mixed statistical estimation in economics. International Econom. Rev. 2: 65–78.

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 14 references