Silva Fennica 1991, Vol. 25 N:o 3: 181–191

Updated measurement data as prior information in forest inventory

Annika Kangas

TIIVISTELMÄ: PÄIVITETYN MITTAUSTIEDON KÄYTTÖ ENNAKKOINFORMAATIONA METSÄVAROJEN INVENTOINNISSA

Kangas, A. 1991. Updated inventory data as prior information in forest inventory. Tiivistelmä: Päivitetyn mittaustiedon käyttö ennakkoinformaationa metsävarojen inventoinnissa. Silva Fennica 25(3): 181–191.

Old inventory data have been widely used as prior information in forest inventory using the method of sampling with partial replacement (SPR). In this method knowledge about forest growth has not been utilized. However, the accuracy of the inventory results can be improved if this knowledge is utilized. The usability of the inventory results can be improved if the prior information is updated by treewise growth models. In this paper a statistical basis is presented for a method in which such information can be used. The applicability of the method is also discussed. An example is given to demonstrate the method.

Vanhaa inventointitietoa on yleisesti hyödynnetty ennakkoinformaationa metsävarojen inventoinnissa käyttäen otantaa osittaisella palutuksella. Tässä menetelmässä ei tietoa metsien kasvusta käytetä hyväksi. Kuitenkin inventointitulosten tarkkuutta voitaisiin parantaa, mikäli myös tämä informaatio käytettäisiin hyväksi. Tulosten käyttökelpoisuutta taas voitaisiin parantaa käyttämällä päivityksessä puukohtaisia malleja. Tässä työssä esitetetään tilastolliset perusteet menetelmälle, jonka avulla tällaista prioritietoa voidaan käyttää hyväksi. Myös menetelmän käyttökelpoisuutta käsitellään. Lopuksi esitetään esimerkki menetelmän toiminnasta.

Keywords: forest inventories, mixed estimator, prior information, model based inference.
FDC 524

Correspondence: *Annika Kangas*, University of Joensuu, Faculty of Forestry, P.O. Box 111, SF-80101 Joensuu Finland. Fax + 358 73 151 4567.

Accepted December 13, 1991

1 Introduction

There are at least two good reasons for using prior information in forest inventory. We can either improve the accuracy of the results and keep the costs the same as before, or, reduce the costs and keep the accuracy of the results the same as before. In forest inventory many kinds of prior information can be used: satellite imagery or aerial photographs, models estimated from data of previous samples, statistics or data measured in previous inventories.

The most commonly used method for using data from previous inventories in forest inventory is the method of sampling with partial replacement (SPR, Ware and Cunia 1962, see also Jessen 1942). For estimating the current population mean two independent estimates are combined to form a single linear unbiased estimator. The weight placed on the two estimates is dependent on the correlation between the timber volumes of the remeasured plots on first and second occasions and also on the population variances on these two occasions.

In the SPR method knowledge about the forest growth is not utilized. When information on forest growth is utilized the estimation is more efficient (Dixon and Howitt 1979).

In the approach of Dixon and Howitt (1979) the Kalman filter was used (see Kalman 1960). The sampling error was handled using the model based inference theory and updated data from previous inventories were used as prior information. The growth model used in their study is crude: it simply gives the proportional change of the state vector over time. It may be assumed that the efficiency and usability of the method can be improved, when more detailed models are incorporated.

Often forest inventory and forest management are treated as separate problems. The data measured should be used as efficiently as possible. Thus, the same data should be applicable for both objectives. In the approach of Dixon and Howitt (1979) information about all control actions (harvests, cultural treatments etc.) is used in the forest inventory. However, the use of these inventory results in forest management systems is limited, because only means and sums are known. For most management systems more detailed information is needed. The required information can be obtained when treewise growth models are used for updating the data.

In Finland the MELA-system is used both for forest management and for updating the inventory data (Siitonen 1983). The inventory data are updated by treewise growth models, and the control actions are taken into account using the values obtained from statistics as restrictions in a linear optimization model.

In Finland, updated data have so far only been used until new data were measured, because methods for using both updated and measured data have not been available. Nowadays such methods are more and more important. The information used in decision making must be as recent as possible. Thus, new measurements are needed more often than before. Due to this the number of sample plots measured in one area at a time will decrease if the costs are not allowed to increase. Consequently, the accuracy of the results will also decrease if only the latest measurements are used.

In this paper a method for using old inventory data as prior information in forest inventory is presented. Using this method, data updated by treewise growth models can be used as prior information. The method by which the precision of the results can be estimated, when both the measured and updated data are used, is also presented. Finally, the usefulness of biased prior information is discussed.

2 The methods based on superpopulation models

In the approach of Dixon and Howitt the growth model can be written as

$$X_{t+1}=AX_t+Bu_t+e_t \qquad (1)$$

The error terms, e_t, are normally distributed with mean zero and covariance matrixΩ. Vector x is the vector of the state variables, vector u is the vector of control actions, matrix A is the state transition matrix and matrix B gives the impact of control actions. The model of the sampling system can be written as

$$y_t=Cx_t+v_t \qquad (2)$$

Matrix C is needed if the observed variables (vector y) are not the same as the state variables (vector x). The error terms, v, are also normally distributed with mean zero and covariance matrix Θ. The error term describes the sampling error (see e.g. Cassel et al. 1977).

The Kalman estimator of the state vector can be calculated by the following procedure. The Kalman filter has a prediction step and an update step that follow each other in sequence.

The predicted conditional mean given all the data through time t is

$$X_{t+1|t}=AX_{t|t}+Bu_t \qquad (3)$$

and the conditional covariance P of this estimator is

$$P_{t+1|t}=AP_{t|t}A' + \Omega \qquad (4)$$

when P_1 is Θ.

A sample is then taken to obtain y_{t+1}. The predicted value will almost always not be the same as the observed value, so a residual vector can be defined as

$$\eta_{t+1}=y_{t+1}-CX_{t+1|t} \qquad (5)$$

The prior information, $x_{t+1|t}$, and the sample information, η_{t+1}, are then combined in the update cycle to yield

$$X_{t+1|t+1}=X_{t+1|t}+K_{t+1}\eta_{t+1} \qquad (6)$$

where

$$K_{t+1}=(P^{-1}_{t+1|t}+C'\Theta_{t+1}C)^{-1}C'\Theta^{-1} \qquad (7)$$

The conditional covariance of this estimator is

$$P_{t+1|t+1}=(P^{-1}_{t+1|t}+C'\Theta^{-1}_{t+1}C)^{-1} \qquad (8)$$

In this paper the approach is, in principle, the same as that of Dixon and Howitt (1979); only the state transition matrix is more complex. The state transition matrix is assumed to consist of many models. The system consists of treewise growth models and plotwise cutting, mortality and ingrowth models. The plotwise models are probability models. These components could not be estimated in a treewise manner, but they were transformed into treewise form with the aid of other models.

The simplest possible model of the sampling system was used in this study. The model used can be written as

$$y_i=\mu+e_i \qquad (9)$$

where y_i is the value observed from sample plot i, μ is the unknown population parameter and e_i is the error term. The error terms are assumed to be normally distributed with mean zero and covariance matrix Σ. In this study, y_i is the timber volume (per hectare) of the plot and μ is the mean timber volume in the area considered.

In this study the mixed estimator of Theil (1971) was used instead of the Kalman estimator in calculating the inventory results. The Kalman estimator and the mixed estimator are, however, equivalent if certain assumptions hold true (see e.g. Dixon and Howitt 1979). The mixed estimator was selected because the growth models were used to update the initial treewise data, and not the state vector as in the Kalman filter approach. In each step the volumes of the plots were calculated from the updated data for the mixed estimator, but they were not

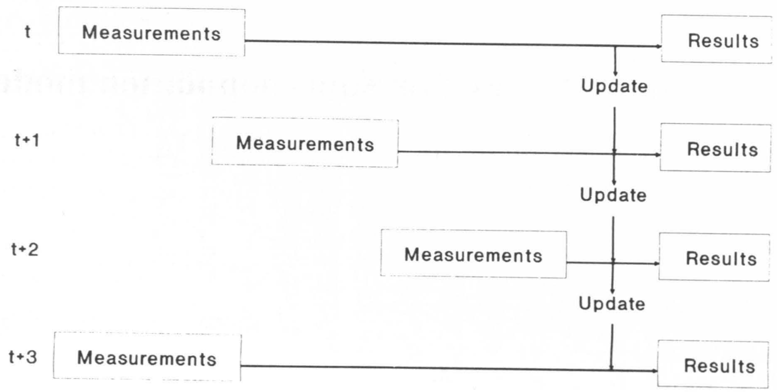

Fig. 1. Formulation of the problem.

used in updating. The formulation of the problem is shown in Fig. 1. The Kalman filter approach was, however, used in calculating the updating errors.

The estimator of the unknown population parameter, μ, can be calculated by the GLS-method. The model can be written as

$$Y=D\mu+E \tag{10}$$

where Y is the 1 (n+m) vector of observed values, D is the 1 (n+m) design vector that consists of ones and E is the 1 (n+m) vector of error terms. Thus, n new plots are measured and m are updated. When the covariance matrix of the error terms, Σ, is known, the GLS estimator of is

$$\mu=(D'\Sigma^{-1}D)^{-1}D'\Sigma^{-1}Y \tag{11}$$

When the number of sample plots is large, it may be difficult to invert the covariance matrix Σ. Often it may, however, be assumed that the sample plots are independent of each other if the distance between the plots is great. This means that the covariance matrix is block-diagonal, which is easier to invert than full matrix.

Occasionally, it may be further assumed that the covariance matrix is diagonal, i.e., the sample plots are independent of each other. If it is assumed that the covariance matrix of the sample data is σ^2 and that of the prior information is k times σ^2, where k is a constant relating the variances, and if the prior information is assumed independent of sample data, the model can be written as (Teräsvirta 1981, Theil 1971)

$$\begin{bmatrix} Y_p \\ Y_u \end{bmatrix} = \begin{bmatrix} D_p \\ D_u \end{bmatrix}\mu+\begin{bmatrix} E_p \\ E_u \end{bmatrix} \tag{12}$$

where Y_p is the vector of new plots and Y_u is the vector of updated plots, and the covariance matrix Σ as

$$\Sigma=\sigma^2 \begin{bmatrix} I & 0 \\ 0 & kI \end{bmatrix} \tag{13}$$

and the mixed estimator of as

$$\mu=(D_p'D_p+D_u'D_u/k)^{-1}(D_p'Y_p+D_u'Y_u/k)=$$

$$\frac{\left(\sum_{i=1}^{n} Y_i+\frac{\sum_{i=n+1}^{m+n} Y_i}{k}\right)}{n+\frac{m}{k}} \tag{14}$$

which is simply a weighted mean of the measured and updated data, where the weight is the inverse of the error variance involved.

3 Estimation of the MSE of the updated data

If formula (14) is going to be used, the error variances of sample data and updated data must first be estimated. The error variance of sample data, σ^2, can be estimated by fitting model (9) to the sample. Estimating the error variance of the updated data is not as simple, because it contains both sampling error and error caused by updating.

The error caused by updating in one five-year updating period was estimated by simulations. The state of the sample plots in the beginning of the period was estimated using increment measurements. The state of these plots was then updated by the models. The error variance was estimated by comparing the true state and the updated state in each of the plots at the end of the period.

If the data is updated for more than one five-year period, the error will increase exponentially as the number of updating periods increase. The error variance after t updating periods can be calculated recursively by Kalman filter method using formula (4) of conditional covariances, if the updating errors are not correlated over subsequent periods.

Thus, t is the number of updating periods, Ω is the updating error in one period, P_1 is the sampling error (σ^2), A is the proportional change of state vector in one period and P_{t+1} is the total error variance of updated data after t updating periods. Coefficient A is estimated from the updated data. Coefficient k can then be written as a function of conditional variances for the two updating periods

$$k(t)=\frac{P_t}{P_0} \tag{15}$$

If the updating errors are, however, correlated over subsequent updating periods, the formula used for calculating the conditional covariances can be written as

$$P_{t+1}=A^2P_t+\Omega_t+2A\,cov(E_{t+1},E_t) \tag{16}$$

instead of (4). If sampling error, P_1, is large when compared with the updating error Ω, or the number of updating periods is small, the effect of correlation is small even when the correlation is large; the greater the updating error the greater effect the correlation has. In Fig. 2 coefficient k is shown as a function of the number of updating periods with four different assumptions of correlation.

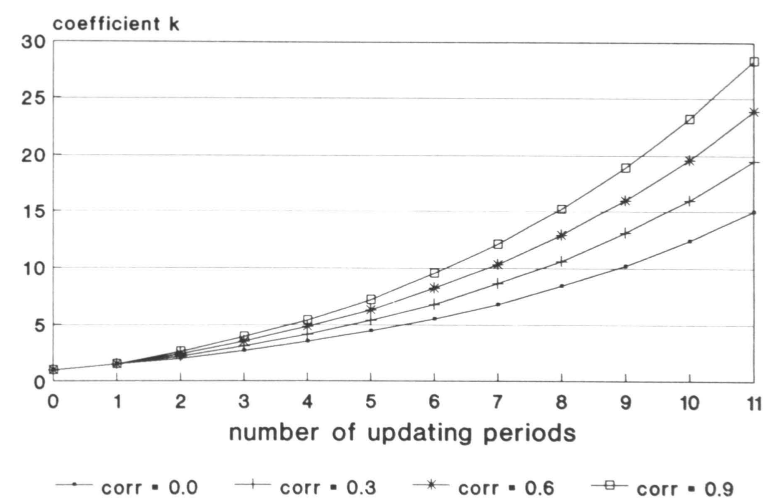

Fig. 2. Coefficient k as a function of the number of updating periods with three different assumptions of correlation. Here P_1 is 94.1^2, Ω is 49.9^2 and A is 1.089.

4 The precision of the mixed estimator

The precision of the estimate of μ can be calculated from the formula of the variance of the GLS-estimator when estimator (11) is used. The variance can be written as

$$var(\mu)=(D'\Sigma^{-1}D)^{-1} \tag{17}$$

When the mixed estimator (14) is used the variance is

$$var(\mu)=\sigma^2 (D'_pD_p+D'_uD_u/k)^{-1}=\sigma^2(n+\frac{m}{k})^{-1} \tag{18}$$

The precision of the mixed estimator as a function of k compared with the accuracy of the basic estimator (in which no prior information has been used) is shown in Fig. 3. The variance of the mixed estimator is always less than that of the basic estimator. The reduction in variance diminishes as the value of k increases.

If the prior information is not unbiased, as is often the case, the true but unknown model of prior information is not (12) but

$$X_u=D_u\mu+g+E_u \tag{19}$$

where g is the unknown bias (Teräsvirta 1980). If the prior information is biased the mixed estimator of μ is also biased. The bias is (Toutenburg 1982)

$$b(\mu)=k^{-1}(D'pDp+k^{-1}D'_uD_u)^{-1}D'_ug=\frac{mg}{k(n+\frac{m}{k})} \tag{20}$$

If the prior information is biased, the accuracy and the precision of the estimators must be studied with the aid of the MSE instead of the variance. Toutenburg (1982) has shown that the MSE of the mixed estimator is smaller than that of the basic estimator if (21) holds true.

$$\sigma^{-2}g'(kI+D_u(D'_pD_p)^{-1}D'_u)^{-1}D'_ug<1 \Leftrightarrow \frac{mg^2}{\sigma^2(k+\frac{m}{n})}<1 \tag{21}$$

The value of coefficient k has an impact on the usefulness of the biased prior information: the smaller the value of k, the smaller the bias that can be tolerated in the prior information. Furthermore, the greater the error variance σ^2, the greater the bias which can be tolerated. This tolerated amount of bias as a function of k is shown in Fig. 4, with three different assumptions of error variance. However, if the mixed estimator must be unbiased, the prior information must also be unbiased.

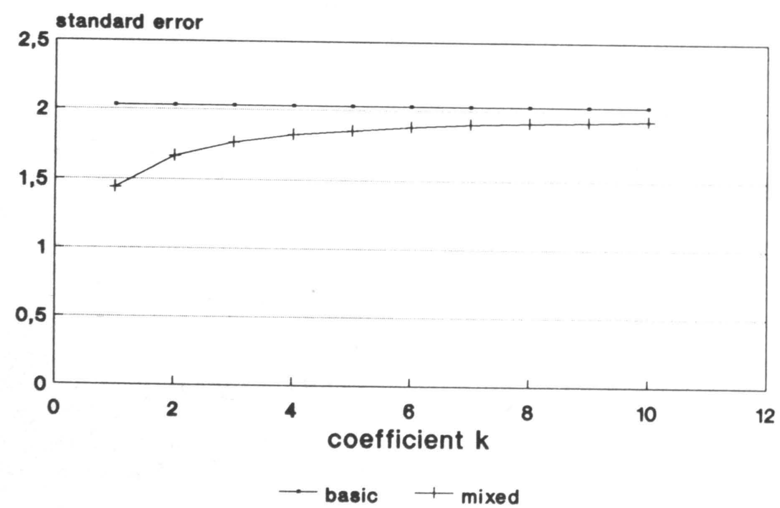

Fig. 3. Standard error of the mixed estimator as a function of coefficient k compared with the standard error of the basic estimator. The number of plots measured/updated is 2150 and σ is 94.1.

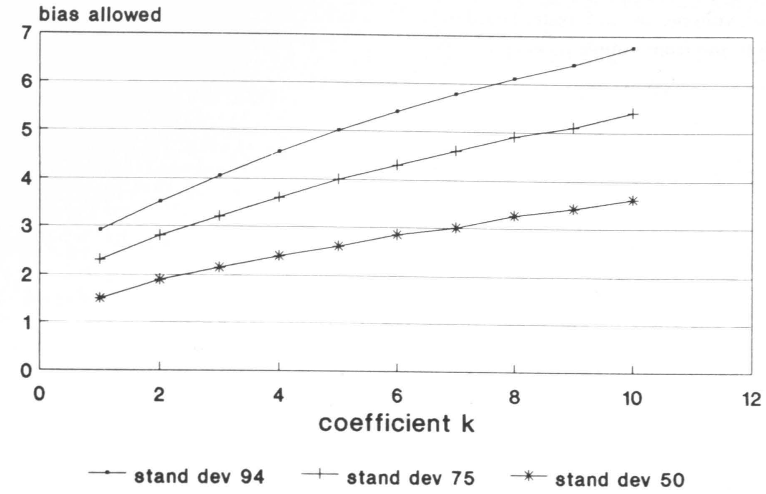

Fig. 4. The amount of bias that can be tolerated in prior information as a function of k with three different assumptions of error variance σ^2. The number of plots measured/updated is 2150.

5 An example

In the example, data from 7th Finnish National Forest Inventory (measured in 1977-1978) were updated for two five-year periods to match with data of 8th NFI (measured in 1986-1987). Both the updated data and data from 8th NFI were then used in a mixed estimator in order to obtain the best linear unbiased estimator of the mean volume.

The data used were from southern Finland from six forestry board districts. Only one seventh of the measured sample plots were used in the example: the plots from which only tally trees were measured were not used in this study. The values estimated from all sample plots were assumed to be the true values. However, these values also contain sampling error. Table 1 contains the mean volumes by district in 7th and 8th National Forest Inventory, estimated from all sample plots and from the plots used in this example.

There are great differences between the true and the estimated values, which are caused by sampling error. A great difference also exists between the true values of 7th and 8th NFI. It is therefore questionable, if the methods based on the correlation coefficient between the two inventories are of much help in this situation.

The updating was done with very simple models in order to demonstrate the method. The MELA-system, which is usually used for updating purposes, was not used due to practical problems. The models used in this study were estimated from the data measured during 8th NFI. Details of the models used can be found in Kangas (1991).

The sampling error was estimated by fitting model (9) to the sample plots. The estimated value of the error variance was 94.1^2. The updating error, Ω, was estimated by si-

Table 1. The mean volumes in six forestry board districts in 7th and 8th National Forest Inventory, estimated from all sample plots and from sample plots used in the example.

District	7th inv true m^3/ha	Sample	8th inv true m^3/ha	Sample
1	112	119.6	139	140.2
2	113	114.0	128	129.8
3	100	95.3	114	102.1
4	126	121.8	150	154.6
5	115	114.1	127	134.5
8	113	117.2	121	118.1
Mean	113	112.8	128	128.2

mulations and the estimated value of this error variance was 49.9^2. The conditional error variance of the updated data on the occasion of 8th NFI – after two updating periods – can be calculated using equation (4) two times as

$$1.089^2 (1.089^2 \times 94.1^2 + 49.9^2) + 49.9^2 = 133.8^2,$$

when coefficient A was 1.089. Thus coefficient k was

$$133.8^2/94.1^2 = 2.02$$

from equation (15). Table 2 shows the updated mean volumes and the true volumes, calculated from the sample plots, by districts at the end of the period.

From Table 2 it can be seen that the updated total mean is unbiased, but there are great differences within districts. This is due to the fact that the models used for updating did not take the regional differences into account.

Table 2. Updated and true volume by districts calculated from the sample plots.

District	True value	Updated
1	140.2	142.9
2	129.8	129.7
3	102.1	105.5
4	154.6	148.1
5	134.5	131.4
8	118.1	120.2
Mean	128.2	128.0

The performance of the models does not, however, affect to the method presented.

In the empirical examples it is assumed that the sample plots are independent of each other. This means that the sample is assumed to be taken by simple random sampling. In fact the sample was taken by systematic cluster sampling. Therefore, there is positive spatial autocorrelation between the sample plots. Simple random sampling is assumed in order to simplify the approach, but the spatial autocorrelation can, however, be taken into account with correlation functions (see e.g. Matérn 1960).

It is also assumed that the mean square error of the updated data is independent of the sample plot, that no measurement errors occur and that the variables, such as height and age of the trees, are known without error.

Table 3 contains the true mean volumes, the estimate obtained from sample plots, the estimate obtained from the updated data of 7th NFI and the mixed estimator. The updated volumes deviate more from the true values than the sample volumes, which is quite natural. In three districts the updated value is, however, nearer to the true value than the sample value. The mixed estimator is the best of these estimators, in the sense that the deviations from true values are small in all districts: the probability of great differences is smaller when prior information is used.

If the mixed estimator is sensitive to the value of coefficient k, it is not useful in practise. Coefficient k used in the calculations will not be precisely optimal for more than one variable at a time. Nevertheless, the same coefficient must be used for all variables

Table 3. The true mean volume from 8th National Forest Inventory, volumes estimated from sample plots and from updated sample plots of 7th NFI and the mixed estimator when k = 2.02.

District	True	Sample	Updated	Mixed
1	139	140.2	142.8	140.5
2	128	129.8	132.9	131.4
3	114	102.1	120.5	108.9
4	150	154.6	148.0	152.6
5	127	134.5	133.2	134.9
8	121	118.1	136.2	125.1
Mean	128	128.2	134.4	130.8

Table 4. True value and mixed estimator with four different assumptions of coefficient k.

District	True	k = 1.75	k = 2.02	k = 2.33	k = 2.57
1	139	140.5	140.5	140.4	140.4
2	128	131.5	131.4	131.2	131.1
3	114	109.6	108.9	108.3	107.9
4	150	152.4	152.6	152.8	152.9
5	127	135.0	134.9	134.9	134.9
8	121	125.8	125.1	124.4	124.4
Mean	128	131.0	130.8	130.6	130.4

Table 5. The number of sample plots updated/measured and the standard error of the basic estimator and the mixed estimator with three different assumptions of coefficient k.

District	Plots	Basic	k = 2.02	k = 2.33	k = 2.57
1	224/229	6.2	5.1	5.2	5.3
2	315/298	5.4	4.4	4.5	4.6
3	411/360	5.0	4.0	4.1	4.1
4	309/269	5.7	4.6	4.7	4.8
5	496/420	4.6	3.6	3.7	3.8
8	392/347	5.0	4.0	4.1	4.2
Mean	2147/1923	2.15	1.72	1.76	1.79

of interest, otherwise there may be inconsistencies in the estimated state vector. Sensitivity calculations made with four different assumptions of coefficient k are shown in Table 4.

The last two values of coefficient k come from the assumption that the correlation of the updating errors over the two updating periods is 0.5 and 0.9. If the correlation is 0.5 and the updating variance on each period is assumed to be constant over time, the conditional error variance of prior information, P_t, can be calculated as

$$133.8^2 + 2(0.5 \times 49.9^2) = 142.8^2$$

and the coefficient k as

$$142.8^2/94.1^2 = 2.33$$

Sensitivity of the mixed estimators is greatest in the districts where the deviation between the sample estimate and the updated estimate is greatest (districts 3 and 8). Yet the changes are quite small in these districts also. When the relative change in coefficient k is about 15 per cent, the change in the mean volume in all districts is less than one per cent. Thus, the mixed estimator is quite robust.

Table 5 shows the standard error of the basic estimator and the standard error of the mixed estimator with three different assumptions of correlation between the updating errors in the two periods. The reduction in the standard error, when the mixed estimator is used, is almost 20 per cent when compared with the basic estimator for the total mean. Thus, the accuracy of the inventory results can be greatly improved even with simple updating models.

The sensitivity of the estimator of the standard error is, however, greater than that of the mixed estimator. The relative change in the estimate of the standard error of the total mean is 2.3 per cent when the relative change in coefficient k is 15 per cent. Therefore, coefficient k must be carefully estimated so as not to make too optimistic assumptions about the standard error.

The updating did not seem to cause bias to the data (see Table 2), but the possibility of bias must be taken into account. The largest bias that gives the MSE of the mixed estimator smaller than that of the basic estimator can be calculated from equation (21) as

$$g < \sqrt{\frac{(2.02+\frac{2147}{1923})\times 94.1^2}{1923}}$$

In this example a bias as great as 3.6 m^3/ha can be allowed. The usefulness of biased prior information also depends on the amount of bias in the mixed estimator that is judged tolerable. If, for example, a 1 m^3/ha bias in the estimator of mean volume is tolerable, the bias in the prior information may be 2.8 m^3/ha.

6 Discussion

In this study the data updated by treewise growth models are used as prior information in forest inventory. Many simplifying assumptions, which may affect the results, have been made about the covariance structure of the sampling and updating errors. The measurement errors and the errors of models used for predicting, for instance, volume or height of trees have not been taken into account. Considering this, the results obtained may seem too optimistic.

In a real situation models used for updating will probably be much more accurate than the models used here (e.g. MELA-system) and also the updating period will often be shorter. These factors can make the benefit, which is obtained using prior information, even bigger.

If the method is to be applicable in real situations, other forms of prior information must also be compatible. This is apparent when model based inference is used. For example, the information obtained from satellite imagery can be used by giving areal weights to the plots, in addition to the weights obtained from error variances (see also Mandallaz 1991). The information obtained from satellite imagery or statistics could also be used in a model form, as stochastic restrictions, like the updated data in equation (14).

The method presented is not yet applicable in practice. The model describing the sampling system must be improved so that the effects of systematic cluster sampling can be taken into account. The measurement and model errors must also be considered. The estimation of the updating error and coefficient k need further study, as does the sensitivity of variables other than volume to coefficient k. The calculations made so far have, however, shown that the presented method is promising way to use updated prior information in forest inventory.

References

Cassel, C-M., Särndal, C-E. & Wretman, J.H. 1977. Foundations on inference in survey sampling. John Wiley & Sons, New York. 192 p.

Dixon, B. & Howitt, R. 1979. Continuous forest inventory using a linear filter. Forest Science 25: 675–698.

Jessen, R. 1942. Statistical investigations of a sample survey for obtaining farm facts. Iowa Agr. Expt. Sta. Res. Bul. 304.

Kalman, R.E. 1960. A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering. 82: 34–45.

Kangas, A. 1991. Päivitetyn mittaustiedon käyttö prioritietona metsävarojen inventoinnissa. Thesis for the degree of Licenciate of Forestry. University of Joensuu. 68 p.

Mandallaz, D. 1991. A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models. Doctoral thesis. Chair of Forest Inventory and Planning, Department of Forest and Wood Sciences, ETH-Zentrum. CH-8092. Zürich. 256 p.

Matérn, B. 1960. Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations. Meddelande från Statens Skogsforsknings-institut. Band 49(5). 144 p.

Siitonen, M. 1983. A long term forestry planning system based on data from the Finnish National Forest Inventory. In: Forest inventory for improved management. University of Helsinki, Department of Forest Mensuration and Management, Res. Notes No. 17: 195–207.

Teräsvirta, T. 1980. Some results on improving the least squares estimation of linear models by mixed estimation. Scandinavian Journal of Statistics 8: 33–38.

Theil, H. 1971. Principles of econometrics. John Wiley & Sons, New York. 736 p.

Toutenburg, H. 1982. Prior information in linear models. John Wiley & Sons. 215 p.

Ware, K. & Cunia, T. 1962. Continuous forest inventory with partial replacement of samples. Forest Science Monograph 3. 40 p.

Total of 12 references