The efficiency of post-stratification compared to model-assisted estimation

Mari Myllymäki[1], Terje Gobakken[2], Erik Næsset[2] and Annika Kangas [2,3,*]

[1] Natural Resources Institute Finland (Luke), Economics and Society Unit, P.O. Box 18, FI-01301 Vantaa, Finland

[2] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432, Ås, Norway

[3] Natural Resources Institute Finland (Luke), Economics and Society Unit, P.O. Box 68, FI-80101 Joensuu, Finland

[*]corresponding author

Abstract

Survey sampling with model-assisted estimation has gained popularity in forest inventory recently. Another option for utilizing the auxiliary information is to use post-stratification, which is a special case of model-assisted estimation with class variables as explanatory variables. In this study, we compared the efficiency of post-stratification with increasing number of strata to model-assisted estimation. We carried out a study based on a simulated population. We considered four different types of post-stratifications, namely (i) stratification based on predictions of a linear model, (ii) stratification based on a regression tree model, (iii) stratification based on the first principal component of the explanatory variables, and (iv) stratification based on the regression tree model with the first principal component as the only explanatory variable. Furthermore, we examined both the traditional post-stratification mean and variance estimators and the difference estimator and its variance estimator for post-stratification. Within the recommended range of number of strata, the model-assisted approach was more efficient than post-stratification. With a large number of strata, post-stratification produced smaller standard error of estimates, but problems such as empty strata were encountered with small sample sizes. Using the first principal component directly for stratification or as an explanatory variable was the most efficient approach.

Keywords: copula, difference estimator, linear model, regression tree, principal component

1. Introduction

Utilizing remotely sensed data as auxiliary information in forest inventory can markedly improve the accuracy and precision of the estimates. Although the model-assisted (MA) framework for estimation (Särndal et al. 1992) has gained popularity also in forest inventory in recent years (e.g. Gregoire et al. 2011), in practice post-stratification (PS), stratification carried out after sampling, may seem more attractive. One reason for this is that the number of variables of interest in forest inventory is usually very high. In both MA estimation and PS, it is possible either to model each variable of interest separately or to utilize one generic model for many variables of interest. The latter approach may seem more attractive, as modelling all the variables may be impractical (Opsomer et al. 2007). In PS, using different stratum borders for different variables may cause practical problems if results need to be calculated for different domains (McRoberts et al. 2014).

PS cannot be used for allocating the sample optimally, but in the case of known stratum sizes and approximately proportional allocation, PS is almost as efficient as pre-stratification (Särndal et al. 1992, p. 265). If the true stratum sizes are unknown, an additional (unknown) error component related to the error in the stratum size will be introduced to the estimates (Cochran 1977, p. 118).

For a single variable y the best characteristics for PS would be the distribution of y itself, or another variable x highly correlated with it (Cochran 1977 p. 127). When remotely sensed data are used as auxiliary information, the number of potential explanatory variables is usually very high. There are two options available: 1) the auxiliary information is condensed to one variable that is used to define the strata; or 2) the explanatory variables are directly used to classify the data to strata using some classification algorithm such as a regression tree (RT). If the first option is used, PS can be based, for instance, on the predictions $\hat{y}$ from a (linear or non-linear) model using some explanatory variables x (e.g. Magnussen et al. 2015) or the first principal component (PC1) of those variables. It should be noted that in the former approach a model is constructed, but it is only used as a basis for stratification. An attractive feature in using PC1 instead is that no models are needed.

PS is in fact a special case of MA estimation, where the stratum identifier is used as a sole explanatory variable (Breidt and Opsomer 2000). If the strata are obtained using predictions from a model, it means that the original model is simplified to a step model. Instead of using the original predictions $\hat{y}_i$ for MA estimation, the within-stratum mean $\bar{\hat{y}}_{hi}$ is used as a prediction for all units i within stratum h. Therefore, such PS estimation can be expected to have a higher variance than MA estimation using the predictions from the original model. It also means that it is possible to use the estimators designed for MA or regression estimation in connection with PS (see e.g. Magnussen et al. 2015).

Several ways for dividing the range of predictions, $\hat{y}_1 ... \hat{y}_N$, into fixed intervals have been proposed (Magnussen et al. 2015). The division may, for example, be based on (1) the quantiles of the predicted values $\hat{y}_i$ producing equal strata weights (e.g. Breidt and Opsomer 2008), (2) the quantiles of the square roots of $\hat{y}_i$ (Baillargeon and Rivest 2011), (3) the square roots of the relative frequency of $\hat{y}_i$ (Dalenius and Hodges 1959), (4) the range of $\hat{y}_i$ (McRoberts et al. 2012) or many other criteria (Magnussen et al. 2015). Each of these approaches can obviously be used to divide also the range of PC1 to strata. In our study, we employ the first, "equal strata weights" option only.

The prediction error in $\hat{y}_i$ is usually considered problematic, as PS requires that the sampling units are assigned to the strata without error (Tipton et al. 2013, Dahlke et al. 2013). With PC1 we do not face this problem. It should be noted that when an external model is used, the $\hat{y}$:s are sums of known explanatory variables weighted by known coefficients and could also be interpreted as known.

Using classification algorithms to define the stratification has been seen as problematic, because the number of resulting strata may be large and their sizes small. There may, for instance, be post-strata without any sample units, or without any variation (Czaplewski 2010). While the number of strata in many classification algorithms can be restricted, restrictions may result in a less efficient classification. The RT approach differs from many other classification algorithms in the sense that it produces at the same time a model that can be directly used in MA estimation in the same way as a linear model (LM), and a classification which can be used as stratification in PS. Therefore, PS and MA estimators can be used equally well.

If the model used in PS is constructed from the sample (i.e. internal), it is called endogenous post-stratification (EPS, Breidt and Opsomer 2008). Such approach has been very popular in forestry in recent years (McRoberts et. al 2012, Dahlke et al. 2013, Tipton et al. 2013). However, Magnussen et al. (2015) showed in a simulation study that such an approach may lead to serious underestimation of variances. Later, Kangas et al. (2016) showed also in a simulation study that using an internal model in MA estimation may lead to serious underestimation of variances. In both cases, the underestimation was more pronounced the more the model was optimized to the sample. Therefore, in this study, we included only external models.

The aims of the current study were to compare accuracy and precision of PS and MA estimation. We considered different types of post-stratifications, either based on linear model predictions (LM), first principal component (PC1) or a classification algorithm (RT). We examined two different sets of estimators for the PS approach, namely the traditional PS mean and variance estimators and the difference estimator and its variance estimator (Särndal et al. 1992 chapter 6.3).

A C Vine copula population similar to that used in Kangas et al. (2016) was utilized for the analyses. From this population, simple random samples were drawn, which were then post-stratified. Estimated means and variances were compared to simulated means and variances.

2. Material

The study area (altogether 853 ha) is located in a boreal forest region in Våler Municipality in southeastern Norway. The forest is actively managed, with Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.) as the dominant species. The study area was delineated into forest stands belonging to four classes related to stand age and species dominance: (1) recently regenerated forest, (2) young forest, (3) mature, spruce dominated forest, and (4) mature, pine dominated forest. A sample survey was conducted with sampling intensities approximately equal for the first three strata, but for the fourth stratum the intensity was only approximately one third of that in the other three strata (Næsset et al., 2013).

Measurements were obtained for 178 systematically distributed, circular, 200-m^2 (radius 7.98 m) forest inventory plots measured in 1999 and 2010. Five plots were discarded from the analysis due to missing values in 1999 and three in 2010. The 1999 data were used for fitting the external models and the 2010 data for copula construction.

Tree-level aboveground biomass was predicted for all trees within the plots using allometric models (Marklund 1988) based on field observations of species and measurements of diameter at breast height (1.3 m) and height. Plot-level aboveground biomass (AGB) was then estimated as the sum of individual tree biomass predictions, scaled to per hectare values (Mg/ha) and denoted ground reference AGB. The uncertainty in the allometric model predictions was assumed negligible (McRoberts and Westfall 2016).

Wall-to-wall airborne laser scanning (ALS) data were acquired for the study area in 1999 and 2010. Pulse density was approximately 1.2 pulses per m^2 in 1999 and 7.3 pulses per m^2 in 2010. Empirical distributions of first echo heights were constructed for the 200-m^2 circular plots. A threshold of 1.3 m above the ground surface was used to remove the effects of echoes from ground vegetation whose biomass is not included in tree-level biomass. For each plot, heights corresponding to the 0th, 10th, 20th, …, 90th percentiles (p0, p10, p20,…, p90) of the ALS height distributions were calculated. Furthermore, several measures of canopy density were derived. The range between 1.3 m above ground and the 95 percentile was divided into 10 vertical fractions of equal height. Canopy densities were then calculated as the proportions of echoes with heights above fraction 0 (>1.3 m), 1, …, 9 to total number of echoes (d0, d1,…,d9). Maximum value (*hmax*), mean value (*hmean*), and coefficient of variation (*hcv*) were also computed. Thus, 23 ALS metrics were available as explanatory variables. Næsset et al. (2013) provide more details for the study area and the dataset.

3. Methods

First, the copula population on which the simulation study is based is explained (Section 3.1). Second, the post-stratified and difference estimators to be compared are presented (Section 3.2) and different stratifications to be considered are introduced (Section 3.3). Finally, Section 3.4 explains the setup for the simulation study.

3.1. The copula population

We used the same approach as Kangas et al. (2016) for the copula construction. That is, we calculated the empirical marginal distributions for the variables AGB, p0, p20, p40, p60, p80, *hmax*, d2, d4, d6 and d8 from the 2010 data using the *logspline* package in R (Kooperberg 2015) and estimated the C vine copula using the *VineCopula* package in R (Schepsmeier et al. 2015). In the current study, we restricted the variables p0, p20, p40, p60, p80, *hmax* to be larger than 1.3 m and the variables d2, d4, d6 and d8 to obtain values in the interval from 0 to 1, mimicking the range of these variables in the data. In the copula construction, we ignored the strata of the Våler data (see also Kangas et al. 2016).

The copula model was used to simulate 22000 (reflecting the population size in the original Våler data) uniformly distributed observations with the modelled (pairwise) dependencies. These 22000 observations can be interpreted as 200 m^2 grid cells mimicking the original laser scanning (Næsset et al. 2013). The copula population was then obtained by calculating the quantiles of the empirical distributions at those simulated uniformly distributed values. The properties of the resulting population are presented in Table 1 and the correlation structure in Table 2.

We assumed that simple random sampling (SRS) was used in the sample selection. Thus, there was no need to simulate geographical locations for the population units.

3.2. The estimators

3.2.1 Post-stratified estimators

Let us assume that we have H strata, N_h is the size of stratum h $(h = 1, \ldots, H)$ and $N = \sum_{h=1}^{H} N_h$ is the size of the population. Then the PS estimator for population mean is

$$\hat{\bar{y}}_{PS} = \sum_{h=1}^{H} W_h \hat{\bar{y}}_h \qquad (1)$$

where $W_h = N_h/N$ is the proportion of stratum h and $\hat{\bar{y}}_h$ is the estimated stratum mean. In PS, the sample size in each stratum h, n_h, is a random variable, as opposed to pre-stratification in which the sample size is fixed a priori (see, however, discussion in Gregoire & Valentine 2008 p. 155). Due to the variation of n_h, the approximate PS variance estimator has an additional element when compared to the pre-stratified estimator (Cochran 1977 p. 135, Särndal et al. 1992 p. 267):

$$\mathrm{var}(\hat{\bar{y}}_{PS}) = \frac{1-f}{n}\sum_{h=1}^{H} W_h s_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{H}(1-W_h)s_h^2 \qquad (2)$$

where s_h^2 is the within-stratum variance. The first term in the estimator is the variance of the stratified estimate under proportional allocation, $f=n/N$, and the second term represents the increase in variance due to the deviation from proportional allocation.

3.2.2 Difference estimators

The difference estimator for the mean AGB is

$$\hat{\bar{y}}_d = \frac{1}{A}\left(\sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{n} \frac{e_i}{\pi_i}\right), \qquad (3)$$

where $\hat{y}_i$ is the model prediction of AGB in cell *i, A* is the total area ($A = N \cdot a$, where a is cell area), $e_i = y_i - \hat{y}_i$ and π_i is the inclusion probability for cell *i*. Its variance estimator (the simplified estimator assuming g-weights to be 1 for all *i*, Särndal et al 1992 p. 362) is

$$\mathrm{var}(\hat{\bar{y}}_d) = \frac{1}{A^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{e_i}{\pi_i} \frac{e_j}{\pi_j} \qquad (4)$$

where π_{ij} is the joint inclusion probability of cells *i* and *j*. Under SRS without replacement, when *i*=*j*, this joint probability is π_i, otherwise it is $n(n-1)/N(N-1)$ (Särndal et al. 1992 p. 31-32). If the model is linear, it is possible to account for the estimation errors of the model coefficients by using the g-weighted variance estimator (Särndal et al. 1992 p. 232, Mandallaz 2008 p. 45). Moreover, the g-weighted sample mean of each explanatory variable is equal to the respective population mean, which is expected to improve the efficiency of the estimator (Särndal et al. p. 234 remark 6.5.1). However, the g-weights have not been defined for other types of models (Massey and Mandallaz 2015), so we ignored them in the current study.

3.2.3 Estimators for simulations

For the simulated copula population, the true mean $\left(\bar{Y}\right)$ is known and biases as well as empirical standard errors of the mean estimators (Eqs. 1 and 3) can be estimated for samples drawn from the population. The bias of a mean estimator was estimated as the difference between the mean of the sample means and the true mean. The mean of the sample means was

$$\mu = \frac{1}{s}\sum_{j=1}^{s}\hat{\bar{y}}_j \,, \qquad (5)$$

where $\hat{\bar{y}}_j$ is either the PS (Eq. 1) or the difference estimator (Eq. 3) calculated for the *j*th sample and *s* is the number of simulated samples.

The estimates obtained by the analytical variance estimators (Eqs. 2 and 4), were compared to the empirical standard errors of the mean estimators, called *simulated standard errors* in what follows, which were calculated as the standard deviation between the *s* sample means as

$$\sigma(\hat{\bar{y}}) = \sqrt{\sum_{j=1}^{s} \frac{(\hat{\bar{y}}_i - \mu)^2}{s-1}} . \tag{6}$$

We further calculated the relative bias (*bias%*) for the mean estimators with respect to the true mean (i.e. $100(\mu - \bar{Y})/\bar{Y}$) and assessed its significance by its Monte Carlo error (*MCE*),

$$MCE\, bias\% = \frac{100}{\bar{Y}} \frac{\sigma(\hat{\bar{y}})}{\sqrt{s}} . \tag{7}$$

3.3. Studied stratifications and estimators

We considered four different types of post-stratifications (Table 4), (i) stratification based on predictions of a LM, (ii) stratification based on a RT model with the original explanatory variables, (iii) stratification based on the PC1 of the explanatory variables and (iv) stratification based on the RT model with the PC1 as the sole explanatory variable. For both LM and RT, we applied both the PS and difference estimators to the stratified data.

To employ the difference estimator (Eq. 3) in connection with post-strata, stratum identifier models (SMs) were used for the predictions and errors of this fitted model. In a SM, the only explanatory variables were the stratum identifiers specified by discretized predictions of the original LM. The PS and difference estimators based on stratified data were compared to the difference estimator based on the LM directly (the MA approach, Eqs. 3-4) and to the simulated estimators (Eqs. 5-6). The models considered were external models that were estimated based on the 1999 data.

3.3.1 The linear model and strata identifier models

Based on the results of our previous study (Kangas et al. 2016), we chose a LM for the MA estimation and as a basis for stratification (case (i, Table 4)). The external model chosen based on the 1999 data included the explanatory variables p40, p60, p80 and d6. The other variables were discarded as they did not statistically improve the model. The residual standard error of the model was 29.91 Mg/ha, R^2 was 0.8022, and adjusted R^2 was 0.7975. The predicted AGB and the residuals of the predictions in the 1999 data are presented in Figure 1.

We predicted $\hat{y}$ by the LM for the whole copula population and used the predictions to define strata boundaries for 2, 4, 6, …, 14 and 16 equally sized classes by selecting the respective quantiles from the empirical distribution of $\hat{y}$ (the PS method "Equal Strata Weights" of Magnussen et al. 2015).

The respective quantiles were also used to define the strata for the external Våler 1999 data. Then, a SM where the stratum identifier was the sole explanatory class variable was fitted and used for prediction (Figure 2 for a case with 16 strata with R^2 0.83 and standard error 28.59 Mg/ha). As the quantiles of the distribution of $\hat{y}$ in the external data and copula population did not necessarily coincide, the stratum borders (means) underlying in SM possibly also differed slightly from the stratum borders (means) used in the PS estimator. Another option would have been to fix the stratum borders also in the copula population to those defined by the $\hat{y}$:s for the external data. That approach would have produced strata with unequal weights in the copula population, however.

3.3.2 The regression tree model

An RT model classifies data to leaves of the tree, which can be interpreted as strata (case ii, Table 4). The number of leaves, and thus strata, can be controlled by restricting the depth of the tree: the maximum number of strata is the depth to the power of two. Thus, the leaves are used as stratum identifiers. In the RT approach, the stratum borders used for the external 1999 data and copula population coincide exactly, as they are defined using fixed values of the explanatory variables (Figure 3).

The mean of each leaf is a model prediction in the difference estimator. When using an external model in the difference estimator, the stratum means in the 1999 data were thus used to predict AGB in the respective strata in the copula population. In the PS estimators (Eqs. 1 and 2), the observed sample mean and variance within the stratum (or leaf) were used. If an internal RT model were used, the mean (variance) within each leaf would also coincide with the observed sample mean (variance).

We used the *rpart* package in R (Breiman et al. 1984) for estimating RT models. We fitted five different regression trees to the 1999 data, with depth varying from 1 to 5, i.e. the maximum number of strata varying from 2 to 25. With 1, the number of splits was 1 (corresponding to 2 strata) and relative error 0.522. With increasing maximum depth the number of splits increased to 8 (9 strata) and the relative error was reduced to 0.172 (Figure 3).

3.3.3 Principal component

We constructed PC1 for the copula population and defined the strata boundaries for 2, 4, 6, …,16 equally sized classes by selecting the respective quantiles from the empirical distribution of PC1 (case iii, Table 4). The PC1 explained about 66% of the variation. Note that PCs can be calculated using the population values. Thus, for the PS estimator (Eqs. 1-2), no model is needed. To apply the difference estimator (Eqs. 3-4) to the stratified data, a SM with the stratum identifier as the explanatory variable was fitted to the external 1999 data. This model was based on PC1 constructed for the 1999 data.

We further employed the RT approach using PC1 as the sole explanatory variable (case iv, Table 4). With a maximum depth of 5, this model fitted to the 1999 data used 7 splits and the relative error was 0.174, i.e. this model was nearly as accurate as the RT with the original explanatory variables.

We further fitted the external LM where PC1 was the only explanatory variable (Figure 4) and considered the difference estimator for this LM (the MA approach).

3.4 The simulation study setup

We generated s = 5000 samples of size n = 100, 200, 500, 1000 from the copula population (N=22000). For each of these samples we employed the mean and variance estimators specified above, and calculated the average of the obtained estimates over all the samples.

We calculated the proportion of samples with at least one empty post-stratum (i.e. cases where stratum mean cannot be estimated with the PS estimator (Eq. 1) without collapsing two strata) and the proportion of samples with only one observation (i.e. cases where the variance cannot be estimated with the PS estimator (Eq. 2)). In the simulation study, we did not collapse the strata, however, but used zero variance and mean estimates for such strata. This was done to illustrate the difference between the PS estimator (collapsing is needed) and difference estimator (collapsing is not needed). Reducing the resulting bias using e.g. sample mean is possible, but beyond the scope of this paper.

4. Results

4.1. Comparison of post-stratification and model-assisted estimation

Figure 5 shows the estimated standard errors of the PS and difference estimators for the LM and RT models (cases i and ii, Table 4). Both estimators with strata based on the LM predictions led to smaller estimated standard errors than the MA approach when the number of strata $H \geq 8$ and $n \geq 200$. Thus, the LM was less accurate than the SMs with a large number of strata. This is likely due to slight nonlinearity between AGB and explanatory variables. In this situation SM models were more flexible than LM, thus providing better predictions for the dependent variable.

The estimated standard errors of the estimators based on the RT model were comprehensivelylarger than those based on the LM predictions. A probable reason for this is that with each split, RT used only one independent variable. Therefore, with two strata the stratification was based on one variable and with 4 strata at most three variables. In the LM predictions, all the four explanatory variables were included also with two strata.

In all cases, the estimated standard errors were very close to the simulated ones, except for the PS estimator for $n = 100$ (Figure 5). This was at least partly due to strata with less than two observations in the simulation experiment, which caused underestimation of variance with large number of strata. There were 0, …,0,12,96,387,987 simulations (out of 5000) for the 2-16 strata and 0,0,69,583,584 simulations for the five RT models, respectively, that led to strata with less than two observations. There were also a few samples that led to such strata for n = 200, but the effect of these was negligible. In the difference estimator, simulated and estimated values of standard errors were fairly similar.

4.2. Results for the estimators based on the PC1

Figure 6 shows the estimated standard errors of the PS and difference estimators for the PC1 and RT with PC1 as the only explanatory variable (cases iii and iv, Table 4). Figure 7 further shows the difference between the stratifications based on the original variables (cases i and ii) and the stratifications based on the PC1 (cases iii, iv). The use of PC1 led to smaller standard errors compared to the stratification based on the LM predictions and the difference increased with increasing number of strata in the case of PS estimators. PC1 was able to stratify the data more efficiently than the predicted $\hat{y}$ from the external LM. Dividing the population into equally sized strata obviously did not minimize the variation of y within the strata as well as did the division based on PC1. For instance, with 16 strata the mean within-stratum variance for the strata based on predictions was 2381, while for PC1-based strata it was 1771.

Again, the empty strata affected the results for $n = 100$ such that the mean simulated and estimated standard errors differed from each other. There were 0, …, 0,13,101,350,977 simulations for the 2-16 strata and 0,876,921,962,962 simulations for the five RT models, respectively, producing a sample with zero or only one observation at least in one stratum.

4.3. Relative biases of post-stratified and difference estimators

All the external models (LM, SM, RT) and both estimators (Eqs. 1 and 3) gave empirically unbiased mean estimates for the sample sizes n =200, 500, 1000 (Figure 8). For n =100, the PS estimator produced statistically significantly biased results with 16 strata, while the difference estimator did not. This was due to the strata with less than two observations. If the simulations that led to such strata were left out, the results showed no bias.

Likewise, when PC1 was used for stratification, in all other cases, except for the case n =100, the estimators were unbiased (Figure 9). For n =100, the PS estimator for the 16 strata using the LM model predictions and RT stratification with $H > 2$ was statistically significantly biased. Also here, this was due to the empty strata.

4.4. Comparison of post-stratified and difference estimators

The difference estimator yielded up to 3 % larger standard errors than the PS estimator for $n \geq$ 500 when the strata were based on the LM predictions. For the strata based on RT models and for $n \leq 200$ the difference between the two estimators was smaller. The difference was due to the use of SM models where the strata borders underlying the stratum identifiers were not exactly the same as those used by the PS estimator, relying on the stratification of the copula population. Moreover, in the difference estimator, the external model was used to estimate the mean in each stratum while in the PS estimator the observed sample mean was used. The difference was smaller with RT models, as for RT, the PS and difference estimators utilized the same stratum borders ($\hat{y} : s$) defined by stratification of the external data. However, the difference estimator still used the mean estimated from the 1999 data as a prediction for each stratum, while the PS estimator relied on the observations from the current sample.

5. Discussion

In our simulation study, the MA approach, i.e. the difference estimator based on the original LM with continuous explanatory variables, was clearly more efficient than the PS or difference estimators based on data stratified by the LM predictions or RT models when the number of strata was within the recommended range ($H < 6$). However, in this study, the estimators based on the stratified data with $H \geq 8$ produced more accurate results than the MA approach. In the case where the stratifications were based on PC1, the estimators based on the stratified data produced more accurate results than the MA approach in some cases even with smaller H. A possible explanation for this is that the relationships between the AGB and the explanatory variables were not exactly linear, leading to a nonlinear relationship between the observed and predicted AGB (see Figure 1 left). Thus, the stratum means could describe the relationship more accurately, provided the number of strata was large enough to make the model more flexible than the LM (see Figure 2). We note that the number of observations in the external data used in this study was only 173 and the relationship between the AGB and explanatory variables estimated from that data may not describe the true relationship.

McRoberts et al. (2014) compared the MA approach to PS for two variables of interest, proportion of forests and mean volume, with 4 strata. McRoberts et al. (2014) stratified the data directly according to the range of the sole explanatory variable to equal size strata. The models used in their study were nonlinear. In their results, PS was more accurate for the proportion of forests, while the MA approach was slightly more accurate for mean volume. Apparently the nonlinear model was not sufficiently flexible to adequately describe the proportion of forests. On the other hand, in the study of Magnussen et al. (2015), the regression estimator was always more accurate than the PS estimator, but they only tested 4-6 strata and stem volume was the sole variable of interest.

In the simulation study by Breidt & Opsomer (2008), the regression estimator was better than the PS estimator when the true model was linear or close to linear, but the PS estimator was better when the model was seriously misspecified. Indeed, if the original model is correctly specified, the MA approach should always be more efficient than PS. Misspecifications can be expected, e.g. when one generic regression model is used for several variables of interest (Breidt and Opsomer 2008, Dahlke et al. 2013). In our case, the LM was slightly misspecified (the residuals show a quadratic pattern), while the stratum means captured this trend.

It should be noted, that within the design-based framework it is not possible to select the best estimator (Godambe 1955, Mandallaz 2008 chapter 3.2), but the best estimator is case specific. Therefore, while our study gives evidence that model misspecification will introduce uncertainty in MA estimates, it does not give evidence that PS with a large number of strata would be more efficient than MA also in other cases. Using the difference estimator for post-stratified data emphasises the fact that PS is a special case of MA estimation with class variables as predictors. Using PS based on LM predictions means that a SM is used in MA rather than the original LM, i.e. while MA estimation is in fact used, the best available model (i.e. the LM) is not. Therefore, we find it more recommendable to always use MA rather than using the estimated LM just for defining the strata. It remains to be studied, however, if the PS approach is more practical than MA with a large number of variables of interest, i.e. if the same stratification can be used for all of them.

In the current study, the stratification based on PC1 was more efficient than the stratification based on predictions of a LM with the four most important explanatory variables. PC1 is a linear combination of all explanatory variables and can also be interpreted as a LM, even though it has not been optimized for predicting y. Instead, it is optimized to capture as much of the variation among the explanatory variables as possible. In our study, the stratification based on PC1 contained more information on the variation of AGB within the strata than that based on the LM.

The good results obtained when using PC1 as the basis for stratification are important for several reasons. PC1 is based on a linear combination of measured values, and therefore there are no residual errors that would affect the results as when the stratification is based on a model. The observations can correctly be assigned to the strata and correct strata sizes can be calculated. It also removes the need to explicitly model the dependency between auxiliary remotely sensed variables and variables of interest. Consequently, no external data are required in PS based on PC1.

In this study, the PS variance estimator (Eq. 2) typically gave smaller estimates than the difference estimator (Eq. 4) based on the SM. This can be explained by the fact that the PS estimator used observed sample values, whereas the difference estimator based on external LM or SM models relied on predictions from the external model. The difference was especially evident with the PC1 approach. Obviously, values at arbitrarily selected quantiles of the external data may be poor predictors of the same quantiles in a differently distributed population. In addition, while the stratification used in the SM model and the stratification of the copula population were based on predictions of the same model, the quantiles that defined the strata borders were not exactly the same in the external 1999 data and the copula population. Thus, for SM and PS to give equal results, internal models or fixed stratum borders (in terms of $\hat{y}$:s) are needed. Using the borders from the external data obviously reduced the efficiency of the difference estimator.

One argument for using the difference estimator instead of the classical PS estimator in the PS approach is that the difference estimator can be used also if there are empty strata (provided an external model is used for which there is information for those strata). It means that the prediction is used for that stratum, but no corrections from observations are available (second part in Eq. 3). Thus, this approach is not as prone to problems caused by empty strata, and the external mean may be a better estimator for the empty strata than e.g. the sample mean. From the point of view of traditional PS, this approach would mean using model-based or synthetic estimator for the empty strata. On the other hand, from the MA point of view, predictions for the empty strata are just ordinary model predictions. Especially the RT approach can equally well be seen from both perspectives, it is both a model and a stratification at the same time. In the future, however, it may be wise to test also other versions of the difference estimator (e.g. Baffetta et al. 2009, Wu & Sitter 2001).

The usefulness of the external prediction can be seen also from Figures 8 and 9, which show larger biases for the PS estimator than for the difference estimator. In real life cases, empty strata would be merged with neighbouring strata. This may cause problems in calculating results over more than one region, if the merging process differs in neighbouring regions (McRoberts et al. 2014). However, combinations of small sample sizes ($100 < n < 200$) and large number of strata ($H > 6$) would most likely not be used for stratification in real life applications. We tested the methods also for N = 200000, with n = 1000, 2000, and 5000, and in these simulations no empty or one-observation strata were observed. Otherwise, the results were similar.

With a small number of strata, the PS based on the predictions of the LM was more efficient than the PS based on the RT, as in the latter case the classification was based only on a small number of the potential explanatory variables. On the other hand, when the stratification was based on PC1 rather than the original explanatory variables, the RT appeared to be an attractive alternative. Already with six strata, the RT based on PC1 produced as accurate results as the stratification based on PC1 with ten strata. However, external data are needed for the RT stratification, but not for the PS approach based on PC1.

6. Conclusion

Basing stratification on PC1 calculated from the actual population seems an attractive approach as then no external data or models are needed. Using PC1 as an explanatory variable in a RT also led to efficient stratifications, but estimating a RT still requires external data.

Using the difference estimator in calculating the variance instead of the traditional formulas in PS was not useful in our study. This was because the stratum indicators had different information content in the external data and the population. In the case of more natural class variables (like site types etc.), the difference estimator should work better, and reduce the problems with empty strata. However, the traditional PS estimator has the advantage that it demands no external data, whilst the difference estimator relies on a SM estimated from external data. All in all, it can be recommended to use MA estimation rather than PS based on model predictions, as the MA approach is both efficient and practical, even though the PS produced more accurate results with a large number of strata in our experiments.

References

Baffetta, F., Fattorini, L., Franceschi, S., & Corona, P. 2009. Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sensing of Environment 113: 463–475.

Breidt, F.J. and Opsomer, J.D. 2000. Local polynomial regression estimators in survey sampling. Annals of statistics 28:1026-1053.

Breidt, F.J. and Opsomer, J.D. 2008. Endogenous post-stratification in surveys: classification with a sample fitted model. The annals of statistics 36:403-427.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. 1984. Classification and regression trees. Wadsworth. 358 p.

Cochran, W.G. 1977. Sampling techniques. John Wiley and Sons. 428 p.

Czaplewski, R.L. 2010. Recursive restriction estimation: An alternative to post-stratification in surveys of land and forest cover. Res. Pap. RMRSRP-81. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 32 p.

Dahlke, M., Breidt, F.J., Opsomer, J. and Van Keilegom I. 2013. Nonparametric endogenous post-stratification estimation. Statistica Sinica 23:189-211.

Dalenius, T. and Hodges, J.L. Jr. 1959. Minimum variance stratification. Journal of American Statistical Association 54:88-101.

Ene, L.T., Næsset E., Gobakken, T., Gregoire, T.G., Ståhl, G., and Nelson, R. 2012. Assessing the accuracy of regional LiDAR-based biomass estimation using a simulation approach. Remote Sensing of Environment 123:579–592.

Godambe V. 1955. A unified theory of sampling from finite population. J.R. Statistical Society B, 17:269-278.

Gregoire, T. G., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., and Holm, S. (2011). Model-assisted estimation of biomass in a lidar sample survey in Hedmark county, Norway. Canadian Journal of Forest Research, 41: 83-95.

Gregoire, T.G. and Valentine, H.T. 2008. Sampling strategies for natural resources and the environment. Chapman & Hall/CRC. 474 p.

Magnussen, S., Andersen, H-E and Mundhenk, P. 2015. A Second Look at Endogenous Poststratification. Forest Science 61:624–634.

Mandallaz, D. 2008. Sampling techniques for forest inventories. Chapman & Hall / CRC. 256 p.

Massey, A. and Mandallaz, D. 2015. Comparison of classical, kernel-based and nearest neighbors regression estimators using the design-based Monte Carlo approach for two-phase forest inventories. Canadian Journal of Forest Research 45:1480–1488.

Marklund, L.G. 1988. Biomassafunktioner för tall, gran och björk i Sverige. Sveriges lantbruksuniversitet, Institutionen för skogstaxering. Rapport 45. 73 s.

Massey, A., Mandallaz, D. and Lanz, A. 2014. Integrating remote sensing and past inventory data under the annual design of the Swiss National Forest Inventory using three-phase design-based regression estimator. Canadian Journal of Forest Research 44:1177-1186.

McRoberts R.E., Gobakken, T and Næsset E. 2012. Post-stratified estimation of forest area and growing stock volume using lidar-based stratifications. Remote sensing of environment 125:157–166.

McRoberts R.E., Liknes, G.C. and Domke G.M. 2014. Using a remote sensing-based, percent tree cover map to enhance forest inventory estimation. Forest Ecology and Management 331: 12–18.

Næsset, E., Bollandsås, O. M., Gobakken, T., Gregoire, T., Ståhl, G. 2013. Model-assisted estimation of change in forest biomass over an 11 year period in a sample survey supported by airborne LIDAR: A case study with post-stratification to provide “activity data”. Remote sensing of environment 128:299-314.

Opsomer, J.D., Breidt, F.J., Moisen, G.G., Kauermann, G. 2007. Model-assisted estimation of forest resources with generalized additive models. Journal of American Statistical Association 102:400-409.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Saarela, S., Grafström, A., Ståhl, G., Kangas, A., Holopainen, M., Tuominen, S., Nordkvist, K. and Hyyppä, J. 2015. Model-assisted estimation of forest resources using different combinations of LiDAR and Landsat data as auxiliary information. Remote Sensing of Environment 158:431-440.

Särndal, C-E., Swensson, B. and Wretman, J. 1992. Model assisted survey sampling. Springer-Verlag. 694 p.

Schepsmeier, U., Stoeber, J., Brechmann, E. C., and Graeler, B. 2015. VineCopula: Statistical inference of vine copulas. R package version 1.6. http://CRAN.R-project.org/package=VineCopula

Tipton, J, . Opsomer J, Moisen, G.2013. Properties of Endogenous Post-Stratified Estimation using remote sensing data. Remote sensing of environment. 139:130–137.

Wu C. & Sitter, R.R. 2001. A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association 96:185-193.

Table 1. The properties of variables in the copula population

	AGB	p0	p20	p40	p60	p80	*hmax*	d2	d4	d6	d8
Min	0.0002	1.3	1.3	1.3	1.3	1.3	1.31	0.000	0.000	0.000	0.000
1st Q	74.9	1.332	6.444	8.772	10.56	11.97	16.58	0.607	0.507	0.352	0.170
Median	119.315	1.508	8.29	11.467	12.94	14.99	19.57	0.765	0.671	0.534	0.253
Mean	128.607	1.863	8.699	11.126	12.84	14.68	18.95	0.670	0.604	0.484	0.250
3rd Q	172.1	2.122	10.575	13.792	15.38	17.45	22.46	0.855	0.796	0.671	0.328
Max	710.859	8.885	36.854	28.844	37.45	38.99	42.74	0.999	1.000	0.999	0.963

Table 2. The lower triangular of the correlation matrix of the variables in the copula population

	AGB	p0	p20	p40	p60	p80	*hmax*	d2	d4	d6	d8
AGB	1.00										
p0	0.38	1.00									
p20	0.78	0.44	1.00								
p40	0.77	0.34	0.91	1.00							
p60	0.77	0.31	0.86	0.97	1.00						
p80	0.69	0.26	0.76	0.86	0.90	1.00					
hmax	0.59	0.16	0.62	0.79	0.85	0.83	1.00				
d2	0.65	0.18	0.53	0.58	0.50	0.34	0.42	1.00			
d4	0.67	0.22	0.61	0.64	0.55	0.38	0.43	0.96	1.00		
d6	0.71	0.27	0.71	0.71	0.61	0.44	0.42	0.88	0.95	1.00	
d8	0.73	0.29	0.75	0.72	0.63	0.48	0.39	0.74	0.82	0.91	1.00

Table 3. The coefficients of the linear model and their standard errors and t-values for the external model estimated from Våler 1999 data.

Variable	Estimate	Std.Error	t-value	Pr(>|t|)
Intercept	-76.826	8.660	-8.871	1.04e-15
p40	6.913	3.190	2.167	0.0316
p60	-11.941	4.751	-2.513	0.0129
p80	13.733	2.852	4.815	3.27e-06
d6	172.045	19.515	8.816	1.45e-15

Table 4. Cases (i)-(iv): the different combinations of models, explanatory variables and estimators tested for stratification.

Case	i		ii		iii	iv	
Model	LM		RT		no model	RT	
Explanatory variables	p40, p60, p80 and d6	Stratum identifier	p40, d2, p20, *hmax*	Stratum identifier	PC1	PC1	PC1
Estimators	PS	MA	PS	MA	PS	PS	MA
Mean	Eq. 1	Eq. 3	Eq. 1	Eq. 3	Eq. 1	Eq. 1	Eq. 3
Variance	Eq. 2	Eq. 4	Eq. 2	Eq. 4	Eq. 2	Eq. 2	Eq. 4

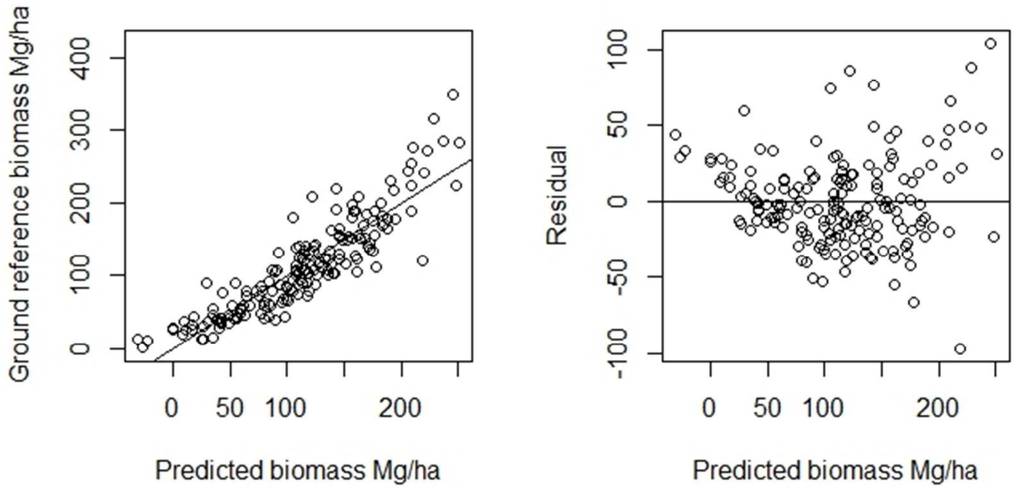


Figure 1. Scatterplot of predicted versus ground reference aboveground biomass and residual plot.

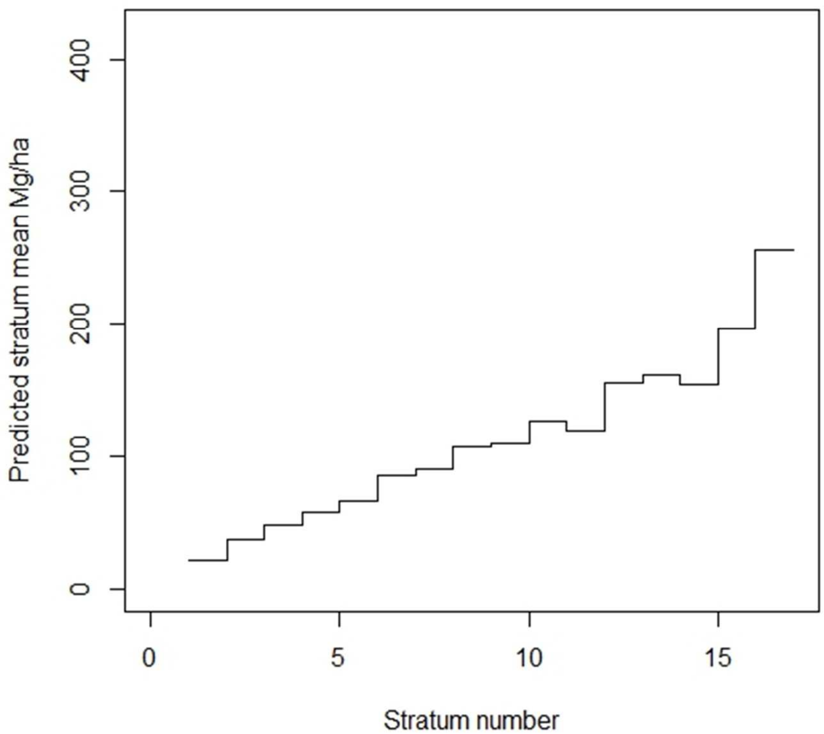
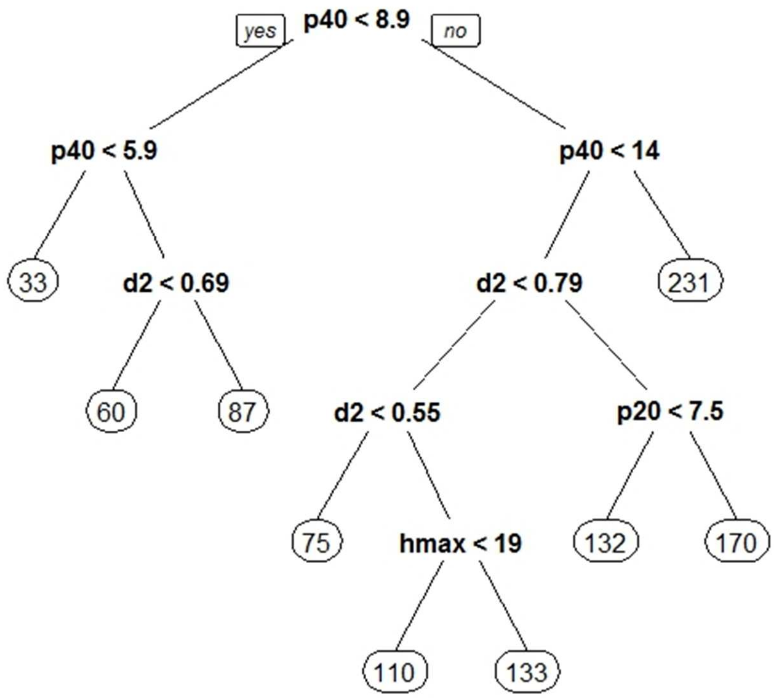


Figure 2. Step function of predicted aboveground biomass using stratum identifier (16 strata) as the sole predictor.



611

612 Figure 3. Regression tree with maximum depth set at five.

613

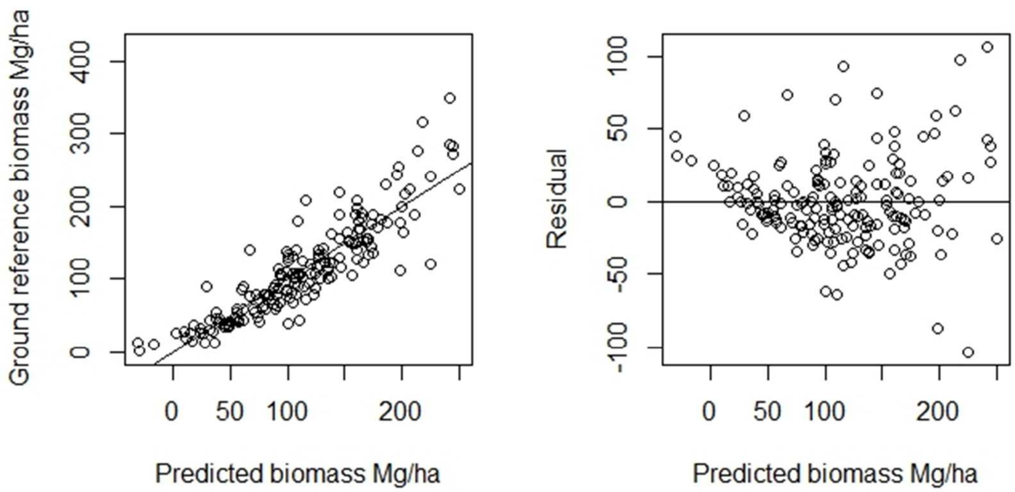


Figure 4. Scatterplot of predicted versus ground reference aboveground biomass and residual plot based on PC1 as sole explanatory variable.

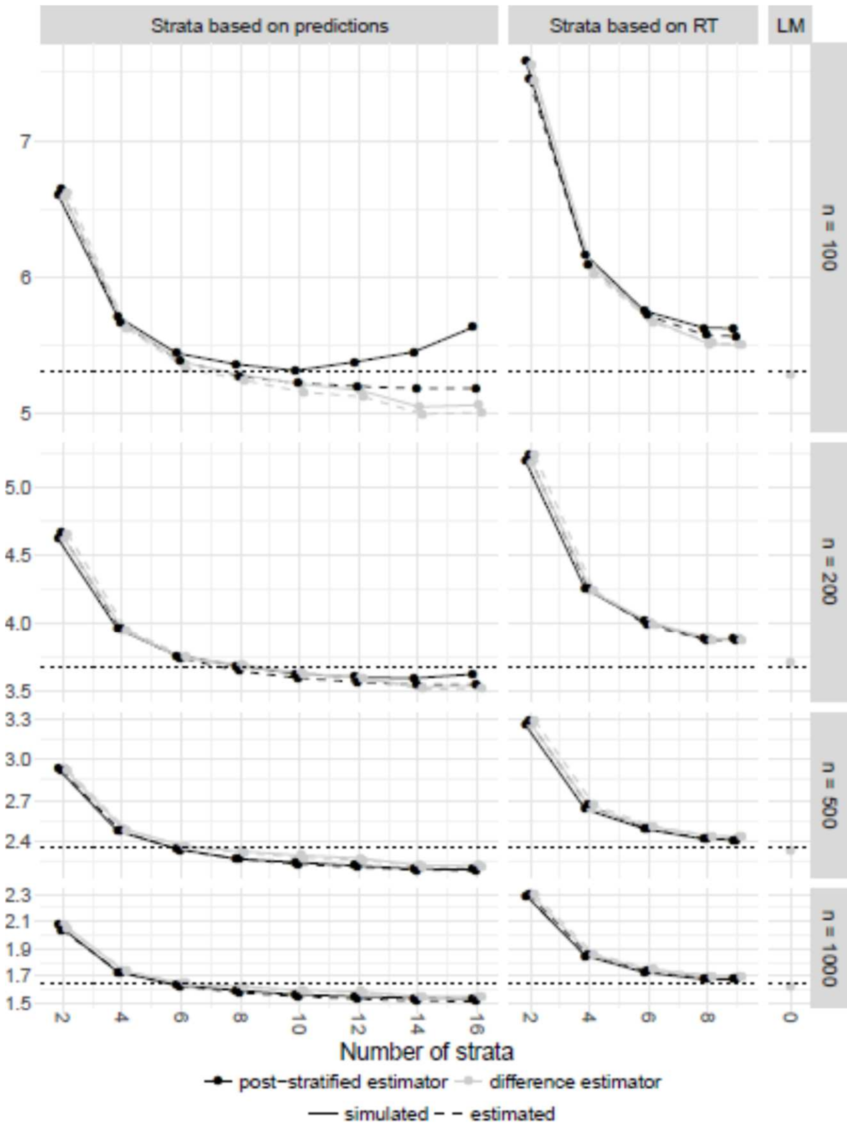


Figure 5. Simulated and estimated standard errors estimated by the post-stratified (eq 2.) and difference estimators (eq. 4). The results for the strata based on the linear model (LM) predictions are presented in the left column, those based on RT models in the middle, and the estimated standard error of the MA approach based on the LM model in the right column. The horizontal dashed lines give the simulated standard errors of the MA approach. The results were calculated from s = 5000 samples of size n = 100, 200, 500, 1000.

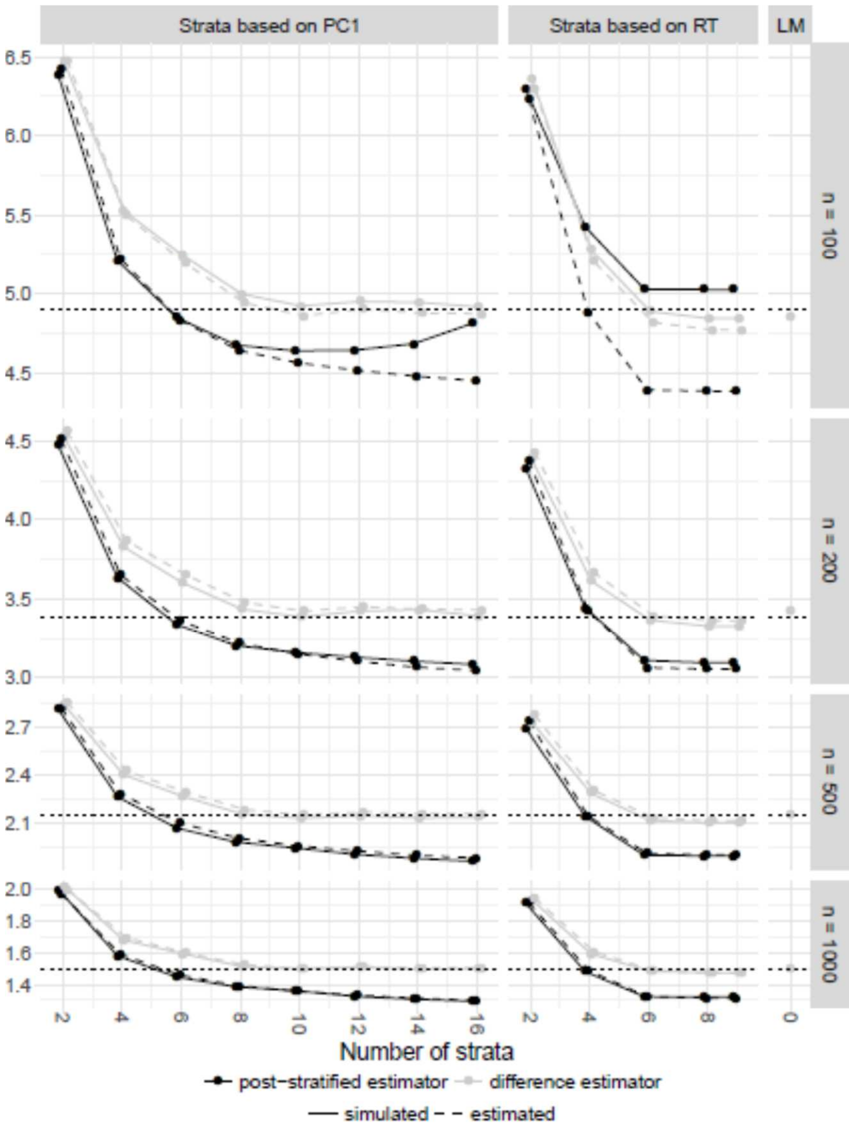


Figure 6. Simulated and estimated standard errors estimated by the post-stratified (eq 2.) and difference estimators (eq. 4). The results for the strata based on the PC1 directly are presented in the left column, those based on RT models (based on the PC1) in the middle, and the estimated standard error of the MA approach based on the LM model (with PC1 as explanatory variable) in the right column. The horizontal lines give the simulated standard errors of the MA approach. The results were calculated from $s = 5000$ samples of size $n = 100, 200, 500, 1000$.

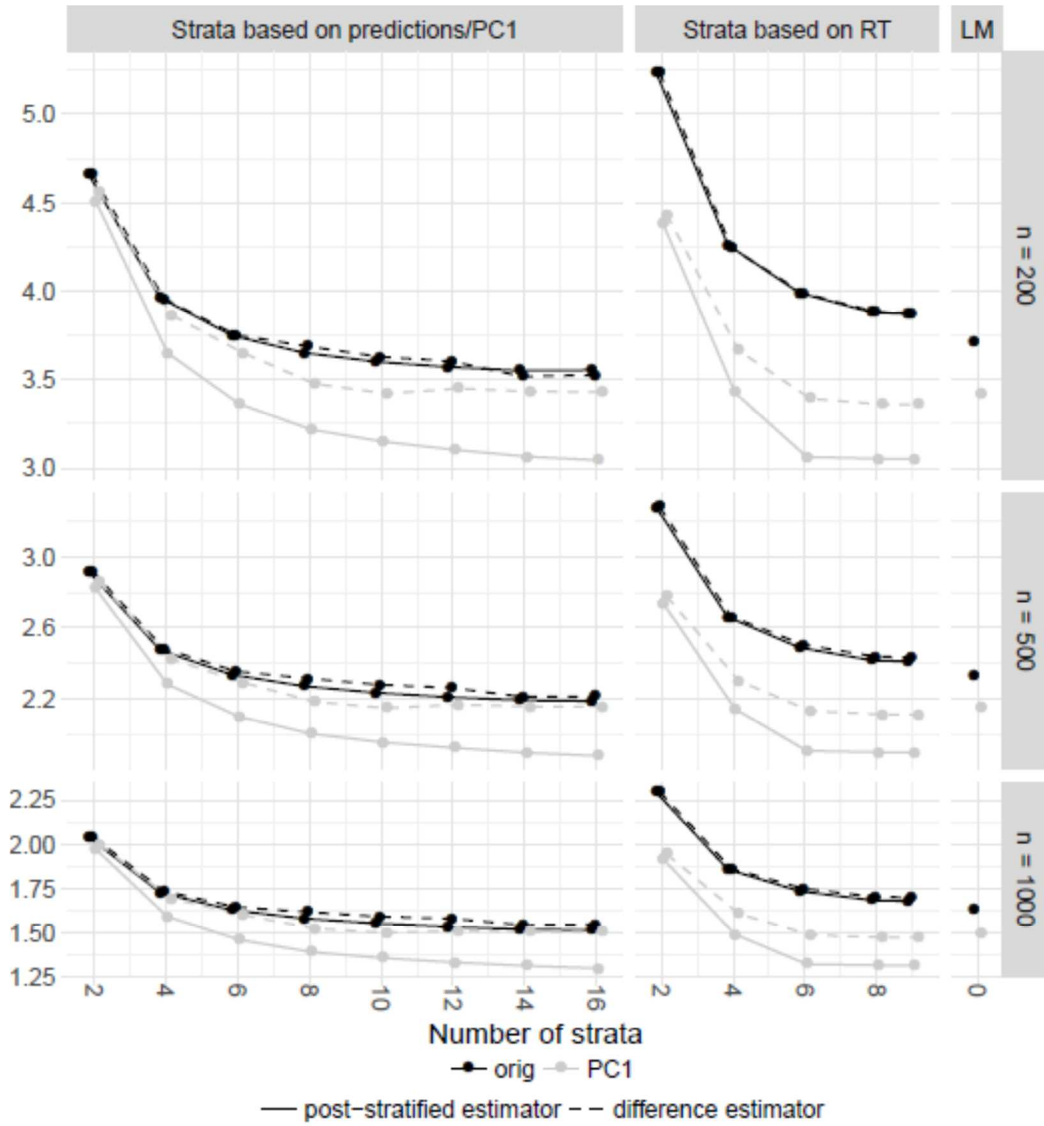


Figure 7. Estimated standard errors estimated by the post-stratified (eq 2.) and difference estimators (eq. 4) for models/strata based on the original explanatory variables ("orig") and for those based on PC1 ("PC1"). The results for the strata based on the model predictions or PC1 are presented in the left column, those based on RT models in the middle, and the estimated standard error of the MA approach based on the LM model in the right column. The results were calculated from *s* = 5000 samples of size *n* = 200, 500, 1000.

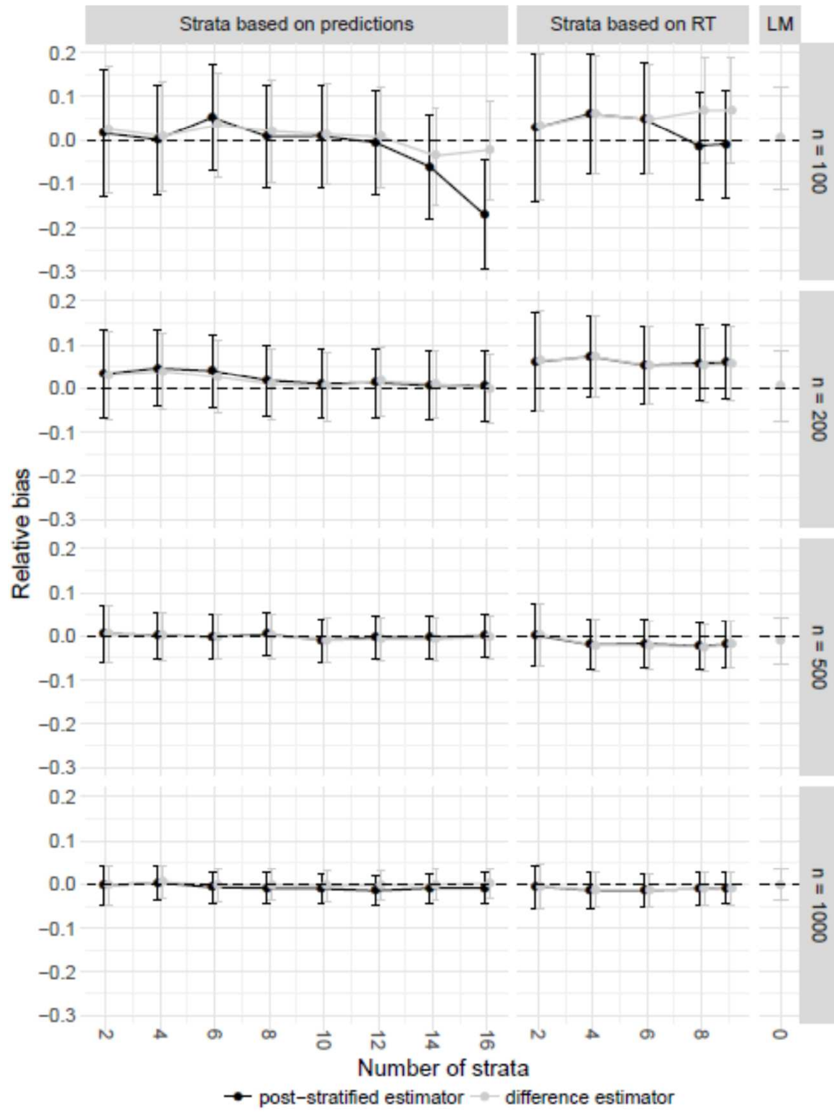


Figure 8. Relative biases +/- two times their MCE for the post-stratified (Eq. 1) and difference (Eq. 3) estimators for the population mean. The strata based on the linear model (LM) predictions are presented in the left column, those based on RT models in the middle, and the result of the MA approach based on the LM model in the right column. The results were calculated from $s = 5000$ samples of size $n = 100, 200, 500, 1000$. The true mean in the copula population was 128.41.

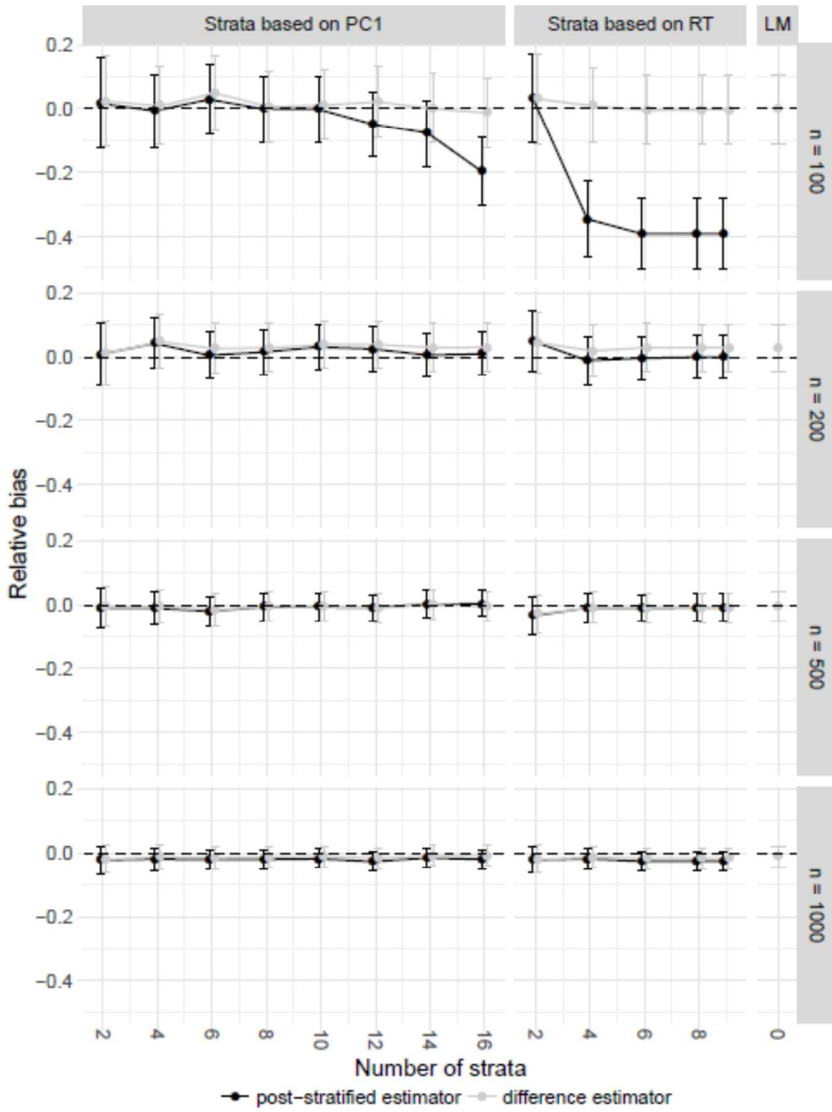


Figure 9. Relative biases +/- two times their MCE for the post-stratified (Eq. 1) and difference (Eq. 3) estimators for the population mean. The strata defined based on the PC1 is presented in the left column, strata defined by the RT models in the middle, and the difference estimator based on the linear model (LM) with the PC1 in the right column. The results were calculated from s = 5000 samples of size n = 100, 200, 500, 1000.