

Spatial prediction of forest stand variables

Arne Nothdurft · Joachim Saborowski ·
Johannes Breidenbach

Received: 18 December 2007 / Revised: 6 December 2008 / Accepted: 22 January 2009 / Published online: 21 February 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract This study aims at the development of a model to predict forest stand variables in management units (stands) from sample plot inventory data. For this purpose we apply a non-parametric most similar neighbour (MSN) approach. The study area is the municipal forest of Waldkirch, 13 km north-east of Freiburg, Germany, which comprises 328 forest stands and 834 sample plots. Low-resolution laser scanning data, classification variables as well rough estimations from the forest management planning serve as auxiliary variables. In order to avoid common problems of k -NN-approaches caused by asymmetry at the boundaries of the regression spaces and distorted distributions, forest stands are tessellated into subunits with an area approximately equivalent to an inventory sample plot. For each subunit only the one nearest neighbour is consulted. Predictions for target variables in stands are obtained by averaging the predictions for all subunits. After formulating a random parameter model with variance components, we calibrate the prior predictions by means of sample plot data within the forest stands via BLUPs (best linear unbiased predictors). Based on bootstrap simulations, prediction errors for most management units finally prove to be smaller than the design-based sampling error of

the mean. The calibration approach shows superiority compared with pure non-parametric MSN predictions.

Keywords k -NN · Calibration method · Imputation · BLUP · Forest inventory · Lidar · Laser data

Introduction

Forest inventories have been conducted in forest enterprises of the German federal state of Baden-Württemberg since 1986. The sampling design consists of concentric fixed radius plots in regular sampling grids. A forest stand in Baden-Württemberg represents one single planning and control unit. Information regarding the individual stands is necessary for efficient forest planning and operational management. This need is the motivation for the specific task assignment of this study. For each forest stand, estimations of the following target variables should be rendered separately for each tree species: area proportion; volume and number of trees per ha in 5 cm DBH classes; volume per ha of timber assortments; mean age, height and DBH; basal area per ha; damages; and proportion of regenerated area. As we aim at predictions for polygons (forest stands) by means of data from circular plots (sample plots of forest inventory), this undertaking is subject to spatial statistics and regionalisation methods.

So far this challenge has been met by deriving estimations for the target variables from the average of values observed on the sample plots in the respective stand. The area of forest stands usually ranges from 0.5 to 20 ha. The regular sampling grid is 100 m × 200 m. Therefore, a single stand mostly comprises only few sample plots and frequently even none. In the case that no sample plots are available from a stand, estimations are obtained by

Communicated by D. Mandallaz.

A. Nothdurft (✉) · J. Breidenbach
Department of Biometry and Informatics,
Forest Research Institute Baden-Württemberg,
Postbox 708, 79007 Freiburg, Germany
e-mail: arne.nothdurft@forst.bwl.de

J. Saborowski
Chair of Ecoinformatics, Biometrics and Forest Growth,
University of Göttingen, Büsingenweg 4,
37073 Göttingen, Germany

synthetic estimators from heuristically classified post-strata. However, this approach has several weaknesses and could result in the following three problems:

1. Due to low sample size in stands, estimations of the mean are highly uncertain and are therefore in most cases unreliable.
2. The application of synthetic estimators derived from heuristically classified strata can cause serious bias for stands without sample plots. Further, predictions could be too smooth and only describe a small amount of variation.
3. The area estimations by means of the sampling grid are very imprecise, because of its wide meshes. Consequently, it cannot be guaranteed that the sum of area-weighted mean estimations would yield the same result as the unbiased (Gregoire and Valentine 2007) Horvitz–Thompson estimator for the population (forest enterprise).

In this paper we present a new 3-stage procedure for precise and reliable predictions of forest stand variables from inventory-based sample plot data. The three stages are:

Stage 1: preliminary predictions based on a non-parametric most similar neighbour (MSN) approach (Moeur and Stage 1995).

Stage 2: random parameter calibrations of the preliminary predictions by means of sample plot observations. We call this calibrated nearest neighbour (CNN) approach.

Stage 3 (optional): global bias corrections.

Stage 1: Most similar neighbour predictions

Because of the small area of forest stands (problem 1) and the stands without sample plots (problem 2), we use off-site sample plot data (Lappi 2001) for predictions of target variables. In comparison with the large number of response variables, only a small number of auxiliary predictor variables are available. As a result, the application of a parametric regression model would leave a large proportion of total variance unexplained.

For the purpose of local predictions of stem density and basal area from double sampling inventory schemes geostatistical approaches were successfully applied by Mandallaz (1993, 2008). Geostatistical approaches follow a stringent mathematical theory and provide closed form expressions for error variance estimates. Nieschulze (2003) also examined diverse geostatistical models for regionalisation of forest variables from inventory data. However, in that study he proved the intrinsic stationarity hypothesis (Diggle and Ribeiro 2007) to be untenable. In Germany,

forests are managed in very small and mostly even-aged stands. Therefore, it is not assured that nearby locations must be more strongly correlated than far-off locations. Then, empirical variograms cannot be reasonably fitted by monotonously increasing covariance functions. Geostatistical approaches, especially ordinary kriging, revealed these weaknesses in studies by Nieschulze (2003) and Nieschulze and Saborowski (2002). The application of external trend functions in universal kriging, as tested by Nieschulze (2003), led to similar problems. External trend functions are usually regression models, e.g. yield tables, and the residual deviation from the trend function may be assumed stationary. Such an approach would also result in the dilemma of estimating plenty of response variables from only a few covariates in many models.

To solve the problems named above, we will use a non-parametric MSN approach in our study following Moeur and Stage (1995) and Nieschulze et al. (2005). The neighbour distances are expressed by the similarity between auxiliary variables in a forest stand and those observed on the sample plot. This method has proven to be the most promising approach in Nieschulze's study. In contrast to Nieschulze (2003), who used colour infrared images, we employ airborne laser scanning data as auxiliary information.

Maltamo et al. (2006) detected that the application of laser scanning data in the k -nearest-neighbour (k -NN) approach is superior to aerial photographs or the combination of class variables and old inventory data. They also concluded that the combination of laser data with additional information from other data sources produces even better results on plot level. However, for stand-level predictions the usage of additional information failed to be beneficial. In this study we operate low-density laser data in a canopy height distribution approach (Maltamo et al. 2006).

A main disadvantage of k -NN approaches is that the neighbourhood on the boundaries of the regression space is asymmetric. In the models, this results in a bias towards the mean (Malinen 2003). Besides, distorted distributions of the reference data over the regression space can cause serious problems. Malinen (2003) developed a locally adapted non-parametric MSN approach, which enables the search for a combination of nearest neighbours that have a minimal distance to the target stand. In order to solve the problem associated with the regression space, we propose a tessellation of the target objects (forest stands) into subunits, which will have an area approximately equal to that of the reference objects (sample plots) with 12 m radius. Thereby, we assign to each of these subunits the observations of only one, namely the nearest, sample plot. The estimates on the stand level are obtained by averaging over the assigned nearest-neighbour observations of all subunits.

As one forest stand comprises a lot of subunits, a variable number of sample plots (nearest neighbours) are applied for the stand level estimates.

Stage 2: Random parameter calibration

In our study region sample plots are located in approximately 77% of the forest stands. Therefore, additional information on prior observations is available for these forest stands. In order to enhance the precision of the predictions we seek to strike a new path in the present study: we treat the nearest-neighbour estimates as preliminary and apply sample plot observations for calibrations of the nearest-neighbour predictions obtained from stage 1.

Stage 3 (optional): Global bias correction

However, there is still no guarantee that adding up the nearest-neighbour estimates will lead to unbiased estimates for the entire region (forest enterprise) (Lappi 2001). Hence, a nearest-neighbour approach will generally not solve problem (3), which results from the comprehensible demand in practice for consistent estimators to be established.

In contrast, Horvitz–Thompson estimations by means of design-based weights deliver unbiased results for the entire region. Deville and Särndal (1992) present regression estimators with weights for the auxiliary variables being calibrated, so that weights are as close to the design-based weights as possible. Lappi (2001) derives prior weights from a spatial variogram model and calibrates them via BLUPs (best linear unbiased predictors) in order to assure equality between the weighted sum and the average of the auxiliary variables.

In the present study we compare subsequently our accumulated results with the Horvitz–Thompson estimates for the entire forest enterprise. For this purpose we apply simple proportional multipliers for bias corrections. This heuristic procedure can definitely not be viewed as a local bias adjustment nor does it improve precision of the predictions on stand level. Rather, it should merely provide consistency on the global level. Therefore, we opt to provide the bias correction method in stage 3 only in the case that practitioners insist on complete consistency between the calculated Horvitz–Thompson estimates from inventory software application, on the one hand, and the added up results of the non-parametric prediction, on the other.

A main drawback of nearest neighbour approaches is that closed form expressions for unbiased error variance estimates do not exist. Stage and Crookston (2007) developed closed-form expressions for approximations of the nearest neighbour prediction variance, with the variance being approximately partitioned into several components.

Approximation of certain components is done via extrapolation of the regression space. This implies strong correlation between response and predictor variables, which in general cannot be presumed due to the small set of auxiliary variables. Therefore, we preferred to approximate prediction variance by bootstrap resampling. It has to be mentioned clearly, that the prediction variance obtained by resampling is of limited evidence because of the unknown amount of bias in nearest neighbour estimates. Therefore, bootstrap variance can only provide approximations for the variability of predictions.

Data

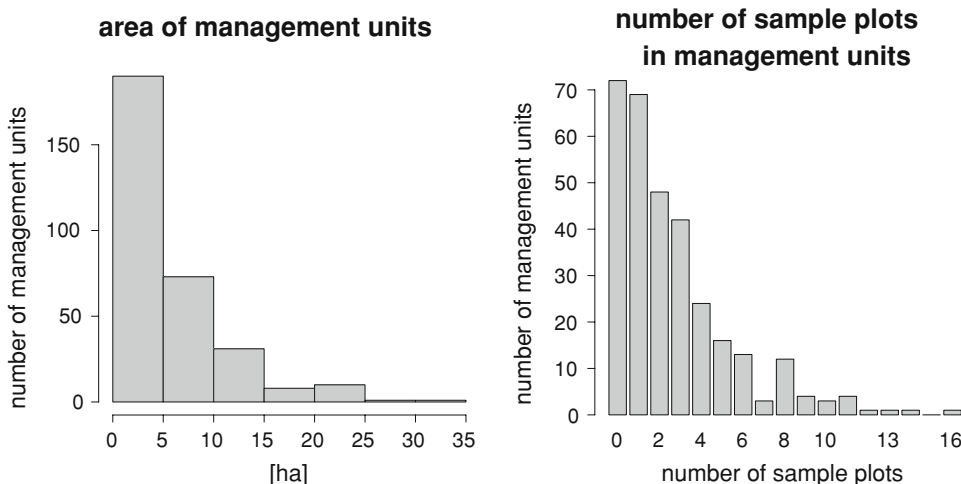
The study area is the municipal forest of Waldkirch, 13 km north-east of Freiburg. The total forest area is 1,775.6 ha and comprises 328 management units (stands). Delineation of the forest stands was conducted in 2002 during forest management planning. In the context of terrestrial surveys, forest stands were stratified into eight stand types (ST) and four treatment classes (TC) (young growth tending, thinning I, thinning II, final harvesting). Each stand was assigned a mean age recorded in 10-year age classes, except for stand type “permanent forest”. In addition, the proportions of observed tree species were estimated. No further features were recorded during forest management planning.

In 2002 an airborne laser scanning was accomplished, originally for the purpose of constructing a digital terrain model (DTM). The lidar (light detection and ranging) vegetation height has been determined by calculating the difference between the elevation of the lidar vegetation data (raw data) and the corresponding DTM raster bin elevation (Breidenbach et al. 2007). For 1,741.9 ha (comprising 314 stands), equalling nearly the entire forest enterprise, laser data are available. The mean area of the forest stands is 5.5 ha (min: 0.2, 25%: 1.5, median: 3.9, 75%: 7.9, max: 30.6). The left graph of Fig. 1 displays a histogram of the stand areas.

Field measurements for the forest inventory were carried out between 1 September 2002 and 11 January 2003. The collected data contains measurements on 875 plots in a regular 200 m × 100 m grid. The sample plots consist of four concentric circles with radii of 2, 3, 6 and 12 m. Trees with DBH up to 10 cm are measured, when their distance to the plot centre is smaller than or equal 2 m; for DBH < 15 cm or DBH < 30 cm the maximum distances are 3 and 6 m, respectively. Trees with DBH ≥ 30 cm are measured in the circle of 12 m radius.

Laser data are available for 834 of the sample plots. For 72 forest stands with laser information no sample plots are obtainable (see Fig. 1, right graph). Approximately 60% of

Fig. 1 *Left* Histogram of stand areas. *Right* Histogram of number of sample plots in forest stands



forest stands contain two or fewer sample plots. Since the aim of this study is to test the general applicability of laser data for an integral resource information system (IRIS), we only consider stands and plots with laser information.

Methods

Forest stands are generally much larger than sample plots and therefore contain more laser pulses. Accordingly, auxiliary variables based on laser pulses will have different variance decompositions of between- and within-units variance components on the stand level, compared with the plot level. By tessellating the forest stands into squared subunits of nearly the same area (subunit area 452.1 m²) as the sample plots, we are able to solve that compatibility problem. Each subunit in each stand is only assigned the observations for the target variables of one plot, namely those from the nearest neighbour. In order to estimate the stand values of the target variables for a certain stand, their subunit values are averaged over all subunits.

According to Härdle et al. (2004, p. 86), we formulate the non-parametric regression model as follows:

$$\begin{aligned}
 y_{hnr} &= f(x_{hnr}) + \varepsilon_{hnr}, & E[y|x] &= f(x) \\
 h &= 1, \dots, H & \text{strata: stand-type (ST)/} & \\
 & & \text{treatment-class (TC)} & \\
 n &= 1, \dots, N_h & \text{forest stands in stratum } h & \\
 r &= 1, \dots, R_n & \text{subunits in stand } n &
 \end{aligned}
 \tag{1}$$

The MSN approach (Moeur and Stage 1995) is based on similarity measure between the auxiliary variables observed in the target stands and on the reference sample plots. As regressor covariates we apply:

- mean laser-derived vegetation height
- variance of the vegetation height
- mean age if stand type is not “permanent forest”
- proportion of species with highest timber volume

In order to obtain predictions the observations of the nearest sample plot are assigned to a certain subunit:

$$y_{hnr} = y_{hj} \quad \text{with} \quad d_{hn,r,j} = \min_{k=1, \dots, K_h} (d_{hn,r,k}) \tag{2}$$

Then all target variables in stand *n* of stratum *h* are predicted by the area-weighted average:

$$\hat{y}_{hn} = \sum_{r=1}^{R_n} g_{hnr} \cdot y_{hnr} \tag{3}$$

Since the area of subunits at stand borders can be smaller than 452.1 m², the weight of the *r*th subunit *g_{hnr}* is the proportion of its area *a_{hnr}* related to the total stand area *F_{hn}*:

$$g_{hnr} = \frac{a_{hnr}}{\sum_{r=1}^{R_n} a_{hnr}} = \frac{a_{hnr}}{F_{hn}} \tag{4}$$

For clearness, in our proceeding the stand level estimate is not only based on one nearest neighbour, rather on several sample plots, namely those with minimum distance to each subunit obtained by the tessellation.

A detailed description of the MSN approach (Moeur and Stage 1995) is provided in the appendix.

Sample plots are located in approximately 77% of the forest stands. Therefore, additional information on prior observations is available for these forest stands. This information can then be used for calibrations of nearest-neighbour estimates.

The predictions of the non-parametric regression model in Eq. 1 can be calibrated by means of prior observations and the a posteriori knowledge about the variances within and between the forest stands. For this purpose, we formulate a random parameter model for the observed response variables on sample plot *q* = 1, ..., *q_{hn}* in stand *n*:

$$y_{hnq} = f(X_{hn}) + b_{hn} + \epsilon_{hnq} \tag{5}$$

Hereof *f* (*X_{hn}*) is assumed to be an approximately unbiased non-parametric regression model. *b_{hn}* is a random parameter on stand level comprising the observed mean deviation on the sample plots from the prior non-

parametric predictions. Its prediction \hat{b}_{hn} is obtained by BLUPs according to Henderson (1963) and Harville (1976) as referred to by Vonesh and Chinchilli (1996, p 252).

With

$$\hat{y}_{hn}^* = \hat{y}_{hn} + \hat{b}_{hn} \quad (6)$$

we receive the final calibrated nearest neighbour predictions for the attributes in stand n . For methodical details see Appendix.

By now the prediction error of \hat{y}_{hn}^* was not considered. For both, the statistician and the practitioner the knowledge about possible prediction error is important. Unfortunately, no closed-form expressions for variance estimation of nearest-neighbour models exist. Therefore, the regression space would have to be extrapolated for approximation via the partitioning approach of Stage and Crookston (2007). Yet, having only a moderate number of auxiliary variables, this procedure might be problematic in our case. Thus, we apply a bootstrap resampling in order to approximate the prediction variance. By running a loop of 200 simulations, we draw subsamples without replacement, which amount to a proportion $1 - \exp(-1) = 0.632$ (Harrell 2001, p 88) of all sample plots. As mentioned above nearest neighbour estimates are generally biased. Thus, the estimated prediction variance might be optimistic because of the unknown amount of bias. The bootstrap variance is rather treated as a measure of the variability of the predictions.

First, the impact of the calibration approach in stage 2 on the prediction error shall be assessed. After this we evaluate the range of the confidence intervals approximately obtained by the bootstrap resampling. For that purpose we compare the bootstrap errors with design-based sampling errors of mean estimations. We use the estimations of the within-stand variance γ [Eq. (11), Appendix] from the linear mixed model in Eq. 5 for constructions of reference levels $t[1 - \frac{\alpha}{2}, df = K - N] \cdot \sqrt{\frac{\gamma^2}{q_{hn}}}$. Due to the mere approximative unbiasedness of the MSN predictions, the confidence intervals should be interpreted with caution.

The software R (R Development Core Team 2007) was used for computations.

In order to achieve consistency between the added up results from the non-parametric CNN predictions and the Horvitz–Thompson estimates for the user of the forest inventory calculation software we provide a global bias correction in the optional stage 3. The method is based on the derivation of heuristic global multipliers. The predictions by nearest-neighbour models are generally biased, whereas the Horvitz–Thompson estimates are unbiased, but much less precise at the stand level. Still, for the purpose of mean predictions for the entire forest enterprise, the Horvitz–Thompson estimates are not only unbiased but also of very high precision. Therefore, it is promising to use the

ratio of these global Horvitz–Thompson estimates and the respective non-parametric predictions, which are obtained by stand area-weighted averages over all stands, as a bias correction factor for the presumable global bias of nearest neighbour predictions. For details of this optional procedure see Eqs. 22–25 in Appendix.

Results

The goal of our study was to estimate data for polygons (forest stands) based on sample plot data. Therefore, results regarding the estimations of target variables for the individual forest stands are presented in the form of maps, which are easily accessible for the forest planning service and management. Figure 2 displays the estimations for selected target variables. Before planning timber harvests and their selling, it is crucial to know the standing timber volume (a), belonging to each species (c, d) and each assortment (e, f). In order to save expenses in reforestation, the forester needs information on the extent of regeneration area under shelter of mature woods (b).

The amount of prediction variance is exposed by way of bootstrap resampling. In order to assess the benefit of the calibration method in stage 2, we compare the empirical distributions of confidence intervals for standwise predictions after calibration (stage 2) with those for pure MSN predictions (stage 1). As exemplified by the results in Fig. 3, calibration by means of sample plot observations in stage 2 significantly reduces the prediction limits. Even calibrations by the use of only one or two sample plots per stand decrease prediction variance. As expected, the precision improves with the increasing number of sample plots within the forest stands. Applying only the pure non-parametric MSN predictor in stage 1, we achieve a mean confidence interval range for total timber volume predictions to the amount of 124.9 m³/ha. Stage 2 calibration (CNN prediction) reduces this figure down to 103.6 m³/ha. This means an improvement by 17%. If prediction errors of stands comprising only sample plots are considered the mean prediction error on 10%-niveau decreases from 119.1 to 92.2 m³/ha (23%). Furthermore, this calibration approach proves to be beneficial for other target variables, e.g. the number of stems per ha (19% reduction of the mean error for all stands) and the fraction of regenerated area under the shelter of old stands (17% reduction).

Because of high within-stands variance components, the prediction variance itself can only be a rough indicator for assessing the prediction quality. Nieschulze (2003) carried out extensive field measurements for collecting evaluation data. In some stands he created more than 20 large area sample plots (radius = 15 m). Nevertheless, a 20% target precision for mean estimations (on $\alpha = 0.05$ -level) could not be hold. In mixed stands, consisting of beeches and

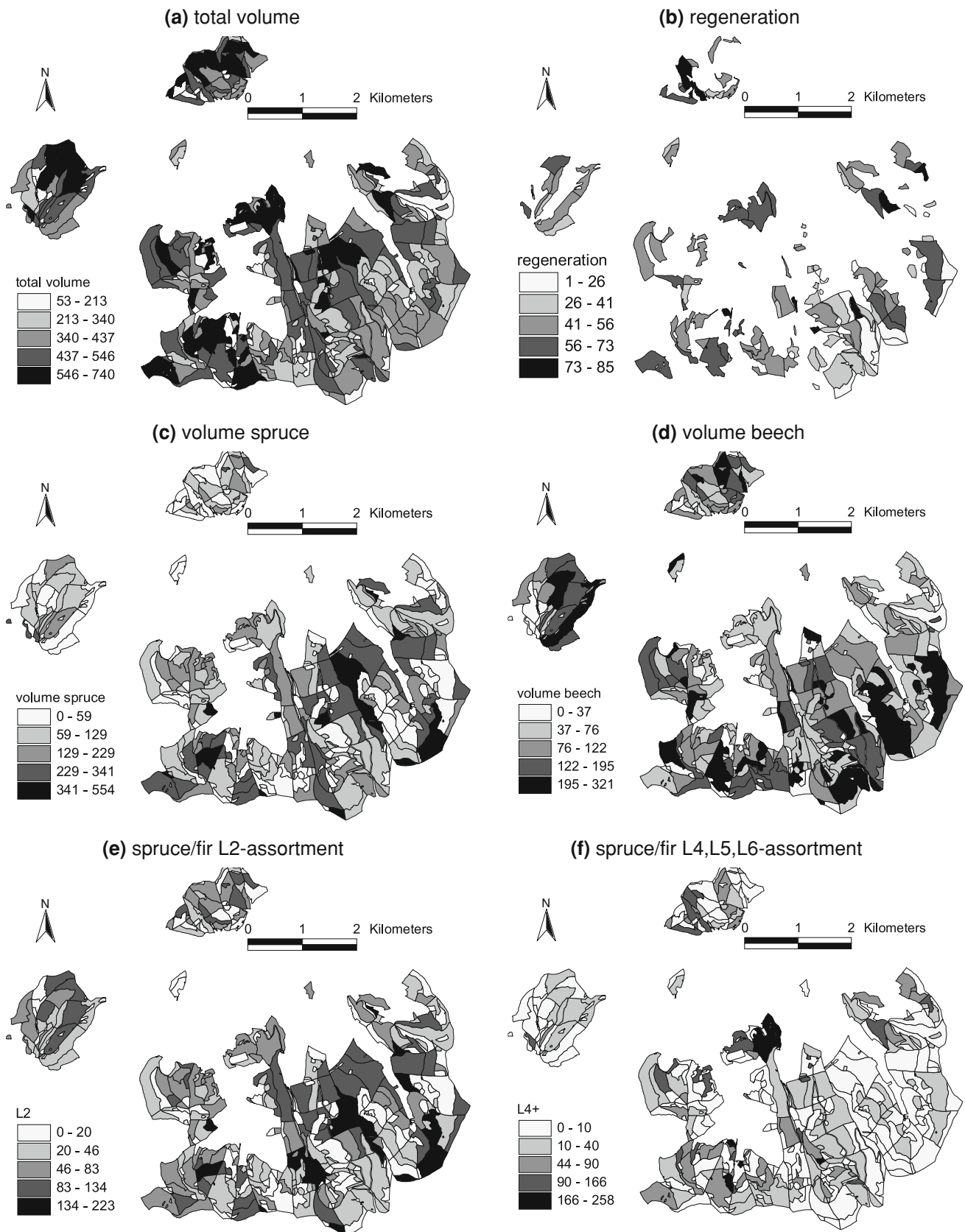
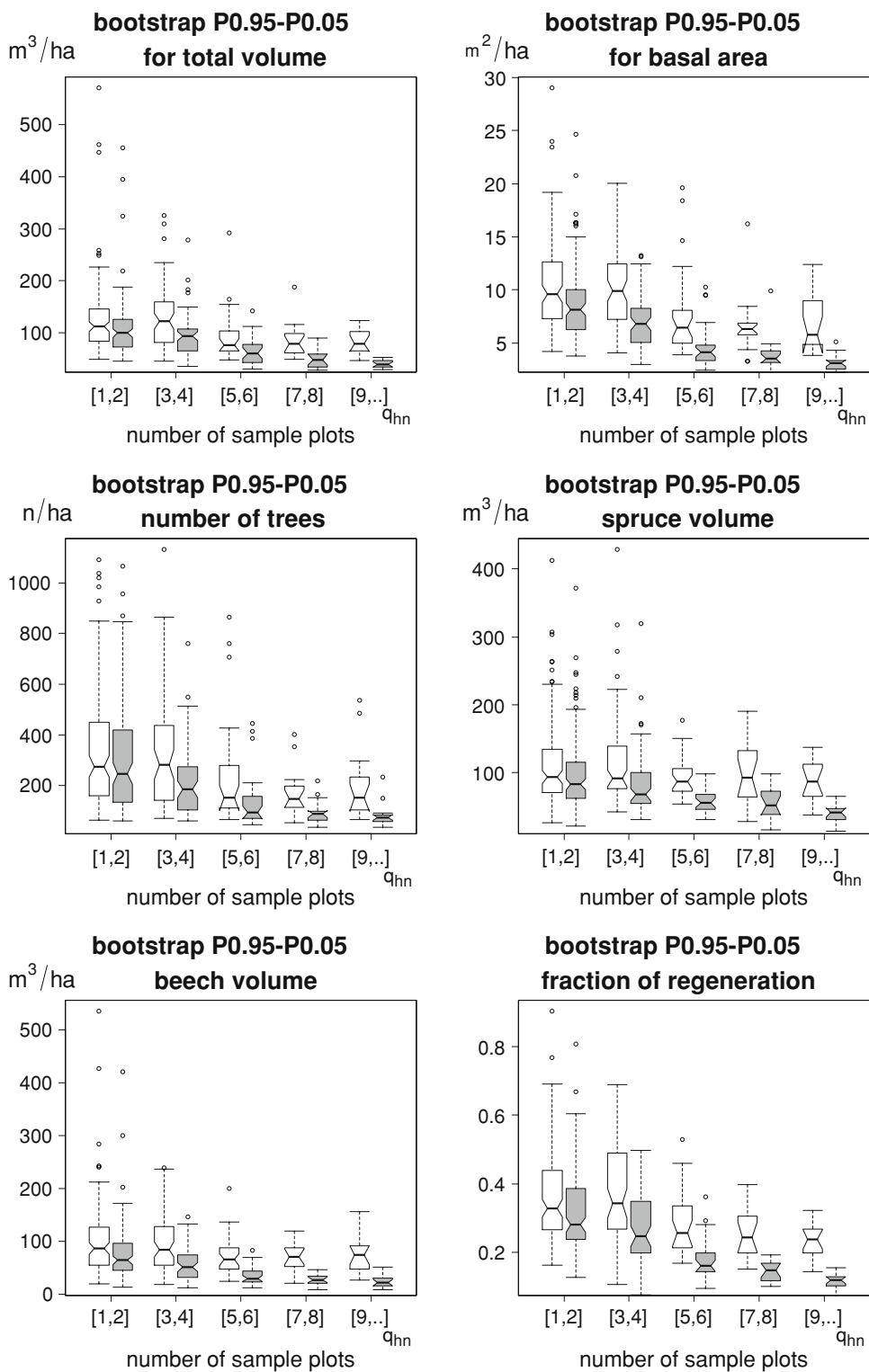


Fig. 2 Predictions of selected target variables in management units. **a** total volume (m^3/ha), **b** regeneration (%), **c** volume spruce (m^3/ha), **d** volume beech (m^3/ha), **e** spruce/fir L2-assortment (m^3/ha), **f** spruce/fir L4, L5, L6-assortment (m^3/ha)

Fig. 3 Prediction errors from bootstrap resampling as *box-plots* of inter-percentile-ranges for selected variables. *White boxes* errors for pure MSN predictions. *Grey boxes* errors for random parameter calibrated MSN predictions (CNN)



spruces, Nieschulze (2003) observed variation coefficients as high as between 50 and 250%.

With this in mind, a claim for 10% precision must be discarded a priori as utopic. It is not surprising that bootstrap resampling based on subsamples from inventory plots also shows high absolute variation. Rather, it is more

interesting to ask whether the simulation variance is substantially larger than the estimated variance within forest stands. In response to this question, Fig. 4 displays inter-percentiles ranges from simulations against the number of sample plots in forest stands for the target variable total timber volume per ha. The design-based sampling error of

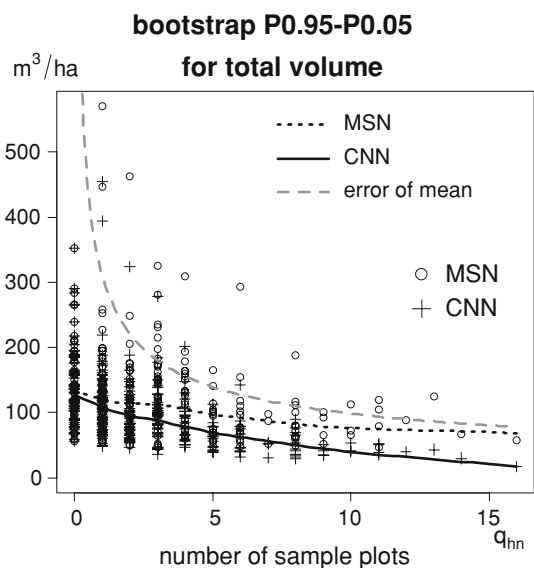


Fig. 4 Prediction errors from bootstrap resampling. Circles prediction errors for the pure MSN predictions. Dotted black line smoother based on robust locally linear fits for prediction errors of the pure MSN predictions (Venables and Ripley, 2002). Crosses errors for random parameter calibrated MSN predictions (CNN). Solid black line smoothed errors of CNN predictions. Dashed grey line design-based error of mean estimation

mean estimation is shown as grey dashed reference curve. It becomes clear that prediction errors for most stands, as derived by bootstrap simulations, are smaller than the design-based error of the mean estimate. The more sample plots are available for calibration, the more precise the CNN predictions become compared to the pure non-parametric predictions. In addition to the absolute prediction error, also the relative error because of its higher practical expressiveness must be acknowledged. The mean relative prediction error is calculated as ratio of the half 90% confidence-interval range to the mean estimate. The mean relative error for total volume per ha achieved by the pure MSN prediction in stage 1 averages 18.7%, and is reduced to 16.6% by the calibration. The gainful impact of the calibration approach in stage 2 proves to be even stronger for further target variables, e.g. the number of trees per ha, the area of regeneration and the amount of specific timber assortments (Fig. 5). As pointed out above, the confidence limits might be not exact because of the biased non-parametric predictions. Due to the unknown statistical properties of the heuristic global bias correction in the optional stage 3, the above-described statistical analysis has been carried out only for stage 1 and stage 2 predictions.

Discussion and conclusions

The model developed in our study enables simultaneous predictions of several forest variables. Target variables

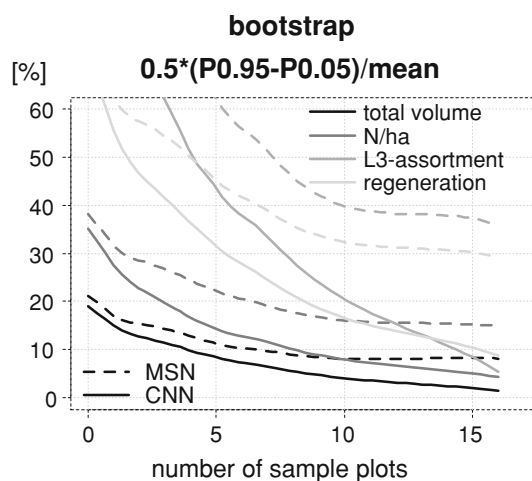


Fig. 5 Smoothed mean relative prediction errors from bootstrap resampling against the number of sample plots in stands. Dashed lines pure MSN prediction. Solid lines CNN predictions (calibrated MSN predictions)

have been predicted for forest stands as polygons, on the basis of inventory data from sample plots. Instead of applying design-based estimators, the non-parametric CNN approach has been used in order to obtain reliable predictions, even with a small number of sample plots in a forest stand or even none.

By constructing a random parameter model, we have been able to calibrate the prior predictions by means of sample plot observations in each stand. However, sample weights of the nearest neighbours have not been calibrated as suggested by Deville and Särndal (1992) as well as Lappi (2001). Instead, the response variables have been calibrated directly with a posteriori knowledge about variances within and between the forest stands. Neglecting the potential bias of nearest neighbour estimates, the results of the bootstrap resampling show, for most forest stands, prediction errors to be smaller than the error of mean estimates. The subsequent calibration of non-parametric predictions by means of sample plot observations in each stand proves to be superior to pure MSN predictions.

The nearest-neighbour approach does not necessarily lead to unbiased predictions. However, the comparison of results for the entire area based on the Horvitz–Thompson estimators and the non-parametric nearest-neighbour estimators yields reliable estimations for the bias on the global level of the entire forest enterprise. The consistency of aggregated estimation results, being an important demand of practitioners, can easily be established by the application of global multipliers in the optional last data-processing step.

Unfortunately only a sparse set of auxiliary variables was available for our study. Due to high costs, terrestrial measurement of additional variables is impracticable.

Therefore, we have used estimations of species proportions obtained in the cruising process of the forest management planning. Persson et al. (2004) and Ørka et al. (2007) present solutions for tree species classification from laser data in Scandinavia. The forest service of Baden–Württemberg plans to collect high-density laser scanning data. Thus, it can be expected that in future more accurate data will be available.

The presented model approach allows for predictions of forest variables in management units without the need to conduct additional and expensive field measurements. Recent aerial surveys with low-resolution airborne laser scanning generate costs to the amount of 1€ per ha. The systematic random sampling scheme of the underlying forest inventory in the study region of Waldkirch is not optimal with regard to costs and precision. In our future research we will aim at reducing the total survey costs. To this end, the construction of double sampling for stratification would be a convenient approach rendering higher efficiency (Nothdurft et al. 2009).

Anttila (2002) shows opportunities to reduce costs by using old inventory data for k -NN estimators of timber volume in small private forests. Going even a step further, it is also possible to use off-site inventory data. In this manner, additional costs induced by laser scanning can be compensated by reducing expenditure for field measurements.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix

Stage 1: Most similar neighbour predictions

Similarity is defined by a quadratic distance function (Moeur and Stage 1995). The distance of subunit r in stand n of stratum h to sample plot k is given by:

$$d_{hn,r,k} = \sqrt{(\mathbf{x}_{hnr} - \mathbf{x}_{hk})^T V_h (\mathbf{x}_{hnr} - \mathbf{x}_{hk})} \tag{7}$$

$h = 1, \dots, H$ strata(ST/TC)
 $k = 1, \dots, K_h$ sample plots in stratum h .

Hereof \mathbf{x}_{hnr} and \mathbf{x}_{hk} are column vectors with $p = 4$ elements for the regressor covariates.

The predictions of the $t = 2$ design attributes total timber volume per ha (m^3/ha) and number of trees per ha require the highest accuracy. The weighting matrix V_h in Eq. 2 is derived by canonical correlation analysis, which seeks for $s = \min(p, t)$ linear combinations

$$\begin{aligned} \mathbf{u}_1 &= X_h \boldsymbol{\alpha}_1, & \mathbf{v}_1 &= Y_h \boldsymbol{\beta}_1 \\ & \vdots & & \vdots \\ \mathbf{u}_s &= X_h \boldsymbol{\alpha}_s, & \mathbf{v}_s &= Y_h \boldsymbol{\beta}_s \end{aligned} \tag{8}$$

for the $K_h \times p$ -matrix X_h of indicator attributes and the $K_h \times t$ -matrix Y_h of design attributes by maximising the coefficient of correlation between the u and the v under the constraint of s uncorrelated linear combinations. In our case the number of possible pairs is $s = 2$. According to Moeur and Stage (1995) we fill the weighting matrix V_h with the product of the squared canonical coefficients Γ and their canonical correlation coefficients Λ

$$V_h = \begin{matrix} & \Gamma & \Lambda^2 & \Gamma^T \\ p \times p & p \times s & s \times s & s \times p \end{matrix} \tag{9}$$

Stage 2: Random parameter calibration

The random parameter model for the observed response variables on sample plot $q = 1, \dots, q_{hn}$ in stand n is formulated by:

$$y_{hnq} = f(X_{hn}) + b_{hn} + \epsilon_{hnq} \tag{10}$$

Therein $f(X_{hn})$ is a non-parametric regression model at X_{hn} , which contains x_{hnr} for all R_n subunits. b_{hn} is a random-effected deviation in stand n and ϵ_{hnq} is a random deviation on plot q in stand n . For the random parameters we assume Gaussian distributions and independence by:

$$\begin{aligned} b_{hn} &\sim N(0, \delta), & \epsilon_{hnq} &\sim N(0, \gamma), & \text{Cov}[b_{hn}, \epsilon_{hnq}] &= 0 \\ y_{hnq} &\sim N[f(X_{hn}), \delta + \gamma] \end{aligned} \tag{11}$$

In matrix notation we formulate the model as

$$\mathbf{y}_{hn} = \mathbf{f}_n(X_{hn}) + Z_{hn} b_{hn} + \boldsymbol{\epsilon}_{hn} \tag{12}$$

where $Z_{hn} = \mathbf{1}_{q_{hn}}$ is a $q_{hn} \times 1$ -column vector containing 1-values, $\mathbf{f}_n(X_{hn}) = f(X_{hn}) \cdot \mathbf{1}_{q_{hn}}$ and b_{hn} is a scalar with the random parameter on plot level. The expected value given the random parameter is

$$E[\mathbf{y}_{hn} | b_{hn}] = \mathbf{f}_n(X_{hn}) + Z_{hn} b_{hn} \tag{13}$$

with variance

$$\text{Var}[\mathbf{y}_{hn} | b_{hn}] = R_{hn} = \text{diag} \{ \gamma, \dots, \gamma \} \tag{14}$$

The expected value of the response variable is

$$\begin{aligned} E[\mathbf{y}_{hn}] &= E[E[\mathbf{y}_{hn} | b_{hn}]] = E[\mathbf{f}_n(X_{hn}) + Z_{hn} b_{hn}] \\ &= \mathbf{f}_n(X_{hn}) = \boldsymbol{\mu}_{hn} \end{aligned} \tag{15}$$

with variance

$$\text{Var}[\mathbf{y}_{hn}] = V_{hn} = Z_{hn} D Z_{hn}^T + R_{hn} \tag{16}$$

In our case $D = \delta$ is a scalar. The matrix Z_{hn} is the design matrix of Eq. 12, and $R_{hn} = \gamma I$ is a $q_{hn} \times q_{hn}$ diagonal matrix.

According to Vonesh and Chinchilli (1996, p. 362), the random parameter value given y_{hn} can be expected:

$$E[b_{hn}|y_{hn}] = DZ_{hn}^T(Z_{hn}DZ_{hn}^T + R_{hn})^{-1}e_{hn} \tag{17}$$

with

$$e_{hn} = (y_{hn} - \mu_{hn}) = y_{hn} - f_n(X_{hn}) \\ = (y_{hn1}, \dots, y_{hnq_{hn}})^T - f_n(X_{hn}) = Z_{hn}b_{hn} + \epsilon_{hn} \tag{18}$$

being a column vector comprising the entire deviation from the mean vector of the non-parametric nearest-neighbour estimates. We use the prior predictions in Eq. 3 as an estimator for $f(X_{hn})$:

$$f(\widehat{X}_{hn}) = \hat{y}_{hn} \tag{19}$$

The variance parameters δ and γ are estimated by using restricted maximum likelihood (REML) techniques.

By means of observations $y_{hn} = (y_{hn1}, \dots, y_{hnq_{hn}})^T$ on q_{hn} sample plots in stand n , the stand-level random parameters can be predicted via BLUPs according to Henderson (1963) and Harville (1976) as referred to by Vonesh and Chinchilli (1996, p 252):

$$\hat{b}_{hn} = \hat{D}Z_{hn}^T(Z_{hn}\hat{D}Z_{hn}^T + \hat{R}_{hn})^{-1}\hat{e}_{hn} \\ = \hat{D}Z_{hn}^T(Z_{hn}\hat{D}Z_{hn}^T + \hat{R}_{hn})^{-1}(y_{hn} - \hat{y}_{hn} \cdot \mathbf{1}_{q_{hn}}) \\ = \left(\frac{q_{hn}}{\hat{\gamma}} + \frac{1}{\hat{\delta}}\right)^{-1} \frac{\sum_{q=1}^{q_{hn}} (y_{hnq} - \hat{y}_{hn})}{\hat{\gamma}} \\ = \frac{\hat{\delta}}{\hat{\delta} + \frac{\hat{\gamma}}{q_{hn}}}(y_{hn} - \hat{y}_{hn} \cdot \mathbf{1}_{q_{hn}}) \tag{20}$$

where $\bar{y}_{hn} = \frac{1}{q_{hn}} \sum_{q=1}^{q_{hn}} y_{hnq}$

with

$$\hat{y}_{hn}^* = \hat{y}_{hn} + \hat{b}_{hn} \tag{21}$$

we receive the final predictions for the attributes in stand n .

Stage 3: Global bias correction

For the entire forest enterprise we obtain with

$$\hat{y}^* = \frac{1}{\sum_{h=1}^H \sum_{n=1}^{N_h} F_{hn}} \sum_{h=1}^H \sum_{n=1}^{N_h} F_{hn} \hat{y}_{hn}^* \tag{22}$$

the mean prediction, weighted by stand area (F_{hn}), for the target variables by the calibrated non-parametric regression model.

Based on the sample plot data we also receive the respective mean prediction by the Horvitz–Thompson estimator:

$$\hat{y}^f = \frac{1}{\sum_{h=1}^H K_h} \sum_{h=1}^H \sum_{k=1}^{K_h} y_{hk} \tag{23}$$

For each attribute we derive multipliers

$$c = \frac{\hat{y}^f}{\hat{y}^*} \tag{24}$$

for bias corrections

$$\hat{y}_{hn}^{*(c)} = c\hat{y}_{hn}^* \tag{25}$$

References

Anttila P (2002) Nonparametric estimation of stand volume using spectral an spatial features of aerial photographs and old inventory data. *Can J For Res* 32(10):1849–1857

Breidenbach J, McGaughey RJ, Andersen H-E, Kändler G, Reutebuch SE (2007) A mixed-effects model to estimate stand volume by means of small footprint airborne lidar data for an American and a German study site. In: ISPRS workshop on laser scanning2007 and silvilaser2007, Espoo, 12–14 September

Deville J-C, Särndal C-E (1992) Calibration estimators in survey sampling. *J Am Stat Assoc* 87(418):376–382

Diggle PJ, Ribeiro PJ (2007) Model-based geostatistics. Springer, New York

Gregoire TG, Valentine HT (2007) Sampling techniques for natural and environmental resources. Chapman & Hall, Boca Raton

Härdle W, Müller M, Sperlich S, Werwatz A (2004) Nonparametric and semiparametric models. Springer, New York

Harrell FE (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, New York

Harville DA (1976) Extension of the Gauss–Markov theorem to include the estimation of random effects. *Ann Stat* 4:384–395

Henderson CR (1963) Selection index and expected genetic advance. *Statistical genetics and plant breeding*, vol 982. National Academy of Sciences-National Research Council, pp 141–163

Lappi J (2001) Forest inventory of small areas combining the calibration estimator and a spatial model. *Can J For Res* 31(9):1551–1560

Malinen J (2003) Locally adaptable non-parametric methods for estimating stand characteristics for wood procurement planning. *Silva Fennica* 37(1):109–120

Maltamo M, Malinen J, Packalén P, Suvanto A, Kangas J (2006) Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. *Can J For Res* 36(2):426–436

Mandallaz D (1993) Geostatistical methods for double sampling schemes: application to combined forest inventories. ETH Zürich, Habilitation thesis

Mandallaz D (2008) Sampling techniques for forest inventories. Chapman & Hall, Boca Raton

Moeur M, Stage AR (1995) Most similar neighbor: an improved sampling inference procedure for natural resource planning. *For Sci* 41(2):337–359

Nieschulze J (2003) Regionalization of variables of sample based forest inventories at the district level. Ph.D. thesis, University of Göttingen, Faculty of Forestry and Forest Ecology, Göttingen

Nieschulze J, Saborowski J (2002) Monitoring of forests under continuous cover system management—tools for the regionalization of forest inventories. In: Gadow K, Nagel J, Saborowski J (eds) *Continuous cover forestry—assessment, analysis, scenarios*. Kluwer, Dordrecht, pp 53–66

Nieschulze J, Böckmann T, Nagel J, Saborowski J (2005) Herleitung von einzelbestandesweisen Informationen aus Betriebsinventuren für die Zwecke der Forsteinrichtung. *Allgemeine Forst- und Jagdzeitung* 9/10:169–176 (in German)

Nothdurft A, Borchers J, Niggemeyer P, Saborowski J, Kändler G (2009) Eine Folgeaufnahme einer Betriebsinventur als

- zweiphasige Stichprobe zur Stratifizierung/A repeated forest inventory based on double sampling for stratification. Allgemeine Forst- und Jagdzeitung (in press) (in German)
- Ørka HO, Næsset E, Bollandsås OM (2007) Utilizing airborne laser intensity for tree species classification. In: ISPRS workshop on laser scanning2007 and silvilaser2007, Espoo, 12–14 September
- Persson Å, Holmgren J, Söderman U, Olsson H (2004) Tree species classification of individual trees in Sweden by combining high-resolution laser data with high-resolution near-infrared digital images. In: laser-scanners for forest and landscape assessment, Freiburg, 3–6 October
- Stage AR, Crookston NL (2007) Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. For Sci 53(1):62–72
- R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. <http://www.R-project.org>
- Venables WN, Ripley BD (2002) Modern applied statistics with S-PLUS. Springer, New York
- Vonesh EF, Chinchilli VM (1996) Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker, New York