Canadian Journal of Forest Research
Revue canadienne de recherche forestière

Model-assisted forest inventory with parametric, semi-parametric and non-parametric models

Journal:	*Canadian Journal of Forest Research*
Manuscript ID	cjfr-2015-0504.R3
Manuscript Type:	Article
Date Submitted by the Author:	06-Apr-2016
Complete List of Authors:	Kangas, Annika; Natural Resources Institute Finland (Luke), Economics and society Myllymäki, Mari; Natural Resources Institute Finland (Luke), Economics and society Gobakken, Terje; Norwegian University of Life Sciences, Department of Ecology and Natural Resource Management Næsset, Erik; Norwegian University of Life Sciences, Department of Ecology and Natural Resource Management
Keyword:	kernel regression, generalized additive model, internal model, external model, difference estimator

Model-assisted forest inventory with parametric, semi-parametric and non-parametric models

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

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

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

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

Abstract

Survey sampling with model-assisted estimation has been gaining popularity in forest inventory recently, as the availability of cheap good-quality remotely sensed data that can be used as auxiliary information has improved. Most of the studies have been carried out using parametric (linear or non-linear) models. However, non-parametric and semi-parametric models such as k-nn, kernel, and GAM are widely used models in forest inventory. The results are usually calculated using the difference estimator (i.e. assuming an external model), even though the models used are based on the sample (i.e. an internal model). In that case, variances will likely be underestimated. In this study, we analyze how well the difference estimator works for different types of models, both internal and external. The study is based on simulated populations produced using a C vine copula model with empirical marginals. The external model is based on real data, and the internal models are estimated from samples from the simulated population. The results show that the analytical variance estimates for a difference estimator based on an overfitted kernel model can seriously underestimate the true variance.

Keywords: kernel, penalized spline, internal model, external model, copula, difference estimator

1. Introduction

Remotely sensed data have been utilized in forest inventory for years (e.g. McRoberts & Tomppo 2007). Remotely sensed data can be used for making observations, for estimating areal results and for mapping. Calculating the results using non-parametric k-nearest neighbors method has been very popular (e.g. McRoberts et al. 2007). However, making inferences concerning the accuracy of the results has been a problem with the non-parametric methods. Some analytical estimators have been developed (e.g. Baffetta et al. 2009), but the theory is not yet fully developed. Therefore, resampling methods have often been used for inferences (e.g. Magnussen et al. 2010).

In recent years, the model-assisted framework for inferences (Särndal et al. 1992) has gained popularity also in forest inventory (e.g. Gregoire et al. 2011). While single explanatory variable and a strictly linear model are assumed in classical design-based regression estimation (Cochran 1977), multiple regression is utilized in the model-assisted framework and analytical estimators for estimating the mean and variance exist. Although the model-assisted framework is gaining interest in forest inventory, applications in practical scale (regional or national) are rarer. One reason for this is that the number of variables of interest (dependent variables) is usually very high, and modelling each of these variables is not practical (Opsomer et al. 2007). Using one model for all variables of interest is possible, but it may not be very efficient. Another option would be to use a model-based framework, in which multiple regression is widely used (e.g. Ståhl et al. 2011). The model-assisted approach is generally assumed safer, as utilizing the design as the basis for inference guarantees (at least almost) unbiased results while in a model-based framework such guarantee does not exist (Massey & Mandallaz 2015).

There exist already quite many applications of the model-assisted framework in forest inventory, typically utilizing parametric (linear or non-linear) models (e.g. Gregoire et al. 2011). The model-assisted framework can, however, also utilize non-parametric or semiparametric models such as k-nn, kernel (Massey & Mandallaz 2015), local polynomials (Breidt & Opsomer 2000) and penalized splines (Breidt et al. 2005) for inference. Opsomer et al. (2007) utilized six generalized additive models (GAM) in a forest inventory application.

One possible problem with the model-assisted framework is that the estimators are developed for a case where the relationships are linear (Saarela et al. 2015a) while in most forest applications they are not. In the model-assisted framework the model does not have to be correct, but using generalized regression estimation assuming linear relationships when the relationships are truly non-linear may cause inefficiency (Wu & Sitter 2001). However, Breidt and Opsomer (2000) proved that asymptotically the mean squared error of the local polynomial model is equivalent to the variance of the difference estimator, provided the model is a smooth function of explanatory variables.

Another complication is that the model used for model-assisted inference is typically estimated from the sample, i.e. the model is *internal*. Yet, difference estimator is designed for cases where the model is estimated from independent data, i.e. the model is *external*. When an internal model is used instead of an external in the difference estimator, the variance will be underestimated, i.e. the precision of the results will be overly optimistic (Massey et al. 2014). However, the point estimators should still be approximately unbiased.

79

80 The possible problems due to non-linearity and external and internal models, and the validity of
81 the difference estimator can most easily be studied by simulations. Thus, many recent studies on
82 the applicability of certain estimators have been carried out in simulated populations. Such
83 populations can be easily generated using copula techniques (e.g. Nelsen 2006). For instance,
84 Ene et al. (2012) tested estimators based on simple random sampling (SRS) for a systematic strip
85 sampling with auxiliary information in a Gaussian copula population. Later on, Saarela et al.
86 (2015b) utilized C vine copulas to test if model form had an effect on the results. They used a
87 model-based framework, in which the inference is based solely on the model, and therefore
88 model form has an important effect. Magnussen et al. (2015) tested if endogenous post-
89 stratification (EPS) (i.e. post-stratification based on a model estimated from the sample)
90 produces reliable results in a vine copula population.

91

92 The aims of the current study were to analyze 1) the effect of using non-linear models in the
93 difference estimator and 2) the effect of using internal rather than external models. The tested
94 model types were a) linear, b) smooth non-linear function based on GAM and penalized splines
95 and c) a highly non-linear model based on kernel regression. We first explored the models in a
96 real data. To analyze the performance of the difference estimator, we generated two populations
97 related to a real population using C vine copula and empirical marginal distributions. From these
98 populations, simple random samples were drawn. Analytically estimated means and variances
99 were compared to simulated means and variances. The simulated variances were calculated from
100 the variation between the simulated samples. To analyze the effect of using internal models, the
101 difference estimator was calculated both with an external and internal model. The models fitted

to the real data were used as external models and the models estimated from the simulated samples as internal models.

2. Material

The study area (altogether 853 ha) is located in a boreal forest region in Våler Municipality in southeastern Norway. It 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 of 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 excluded from the 1999 data and three from the 2010 data due to missing values. The 1999 data were utilized for estimating the external models and the 2010 data for constructing the copula population.

Tree-level aboveground biomass was predicted for all trees within the plots using allometric models based on field observations of species and measurements of diameter at breast height (1.3 m) and height (Marklund 1988). Plot-level aboveground biomass (AGB) was then estimated as the sum of individual tree biomass predictions, scaled to per hectare values (Mg ha^{-1}). The

uncertainty in the allometric model predictions assumed negligible (McRoberts & Westfall 2016).

Wall-to-wall airborne laser scanning (ALS) data were acquired for the study area in 1999 and 2010. Pulse densities were approximately 1.2 pulses per m^2 for 1999 data and 7.3 pulses per m^2 for 2010 data. Empirical distributions of first echo heights were constructed for the 200-m^2 circular plots and 200-m^2 square cells that tessellated the study area. The number of square cells included into the study area in 2010 was 22858 (agricultural areas etc. were excluded).

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 and cell, heights corresponding to the 0^{th}, 10^{th}, 20^{th}, …, 90^{th} percentiles (p0, p10, p20,…, p90) of the 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.

147 3. Methods

148

149 *3.1 The models used*

150 We did not attempt to optimize the selected explanatory variables for the modelling task in any
151 way, as the aim was to test the modelling approaches. For the first copula population, the density
152 d6 corresponding to proportion of echoes above fraction 6 to the total number of echoes and
153 percentiles 10, 40 and 70 of the ALS height distributions (p10, p40 and p70) were selected. The
154 dependent variable was the plot-level AGB. For this simple copula, we show the details of
155 copula construction. We tested for a strata effect in the original field sample for all the modelling
156 approaches described below, but the strata did not have a statistically significant effect, given the
157 selected explanatory variables. Thus, the explanatory variables accounted for the differences
158 between the original strata adequately. For a second copula population, we used a systematic
159 selection of variables: p0, p20, p40, p60, p80, hmax, d2, d4, d6 and d8. This copula population
160 was constructed in order to illustrate the effect of variable selection.

161

162 We tested three different modelling approaches to estimate the AGB: (1) linear model (denoted
163 LM), (2) local constant kernel model (LCK) and (3) a generalized additive model (GAM) with
164 penalized splines as smoothing factors. These three approaches were selected to represent strictly
165 linear (LM), smooth non-linear (GAM) and non-monotonic non-linear (LCK) models to be
166 examined in the model-assisted inference. Note that the k-nn, which is often used in forestry, is
167 actually a special case of the kernel model (Massey & Mandallaz 2015), and therefore the kernel
168 model was selected. The GAM approach with penalized splines was selected as it can describe
169 non-linear relationships without a pre-defined assumption of the model shape, which would be
170 the case with non-linear parametric models. The penalized splines ensure a smooth model.

171

172 In the first model type (simple linear model), no transformations were carried out to linearize the
173 relationship of AGB and explanatory variables or to homogenize the variance. We did not
174 attempt to find a "true" model but just a working model which is correlated with the AGB and
175 reduces variance in the estimation. The residual standard error of the model fitted to the 1999
176 data was 31.34 Mg ha^{-1}, coefficient of determination R^2 was 0.783 and Adjusted R^2 was 0.778.
177 The estimated coefficients of the model are presented in Table 1. Graphs of predicted AGB and
178 the residuals of the model are presented in Fig 1. The model predicts negative AGB for some
179 plots, but this illogical behavior was ignored here. The fact that neither p10 nor p40 improved the
180 model statistically significantly was ignored as well: also for all the internal models fitted in the
181 simulation study we kept all the four available explanatory variables. Thus, the variables in the
182 external and internal models were the same, but their significance and therefore their effect on
183 predictions varied from a model to another.

184

185 The second modelling approach was a non-parametric kernel smoothing method. The model was
186 fit using *np* package in R (Hayfield & Racine 2008). The model option of local constant, i.e.
187 Nadaraja-Watson type model (Nadaraja 1964) with Epanechnikov kernel function
188 (Epanechnikov 1969) was utilized. The optimal bandwidths were selected with least squares
189 cross validation. The residual standard error of this model was 22.31 Mg ha^{-1}, and R^2 was 0.889.
190 The bandwidths were set as fixed in the estimation (i.e. not varying as a function of the values of
191 the explanatory variables). The estimated bandwidths were d6 = 0.0285, p10 = 1.317, p40 =
192 3.608 and p70 = 2.648. Graphs of predicted values and residuals of this model are presented in
193 Figure 2. The regression functions for each of the explanatory variables and the confidence

intervals are shown in Figure 3. Also in this case, the importance of p10 and p40 appeared quite small. This modelling approach was sensitive to local optima. For example, giving the the explanatory variables in different order to the *np* package could produce different optimal bandwidths. We used eight restarts (twice the default) of the process of finding extrema in the bandwidth optimization in order to find the global optimum.

The local linear model option was also tested. This model fitted to the data had much smaller variability in the predictions (i.e. it was smoother or less wiggly) than the fitted local constant kernel model. However, this option was not included into the simulation study, as in the copula population the internal local linear models produced highly extreme, illogical values (e.g. AGB > 70 000 Mg ha^{-1}). The linear model also produced illogical (negative) values occasionally, but not nearly as extreme as the local linear model. The local constant model only produced positive values within the observed range.

The third model type was a generalized additive model with identity link function and Gaussian distribution family. The model was estimated using *mgcv* package in R (Wood 2006). The smooth terms used were cubic splines, which are penalized, and the only parametric term was the intercept. Graphs of predicted values and residuals of this model are presented in Figure 4. The model was estimated using generalized cross validation. According to the approximated F ratio distribution and estimated degree of freedom (edf) (see Wood 2006, p. 191), neither p40 nor p10 were significant in this model either (Table 2). This can also be seen in the relation between them and AGB: the smooths for these two variables are almost horizontal (Figure 5). The adjusted R^2 was 0.81 and the deviance explained 82.2%.

Finally, we similarly estimated from the 1999 data the three models (LM, LCK, GAM) using the ten explanatory variables selected for the larger copula population. These models served as external models in the larger population (the model properties not reported here).

3.2 Copula population

The empirical marginal distributions for the variables AGB, d6, p10, p40 and p70 in the 2010 data were calculated with *logspline* package in R (Kooperberg 2015). The distributions were bounded from below (at zero) to ensure realistic values. The resulting marginal distributions are presented in Figure 6. Then, uniformly distributed new variables were formed from the cumulative probabilities of the original observations. The original stratification was not accounted for in the copula construction for the sake of simplicity.

Then *VineCopula* package in R (Schepsmeier et al. 2015) was used to fit the C vine copula to the uniformly distributed new data. In the C vine copula, a multivariate distribution of the variables is formed from pair copulas that describe dependencies between each pair of the variables (see Aas et al. 2009) and a specific tree structure of the variables. In the *VineCopula* package, the tree structure is selected sequentially (see Dißmann et al. 2013). It is based on Kendall's tau, so that the variable *i* that maximizes the sum of absolute Kendall's taus $\tau_{i,j}$ is selected as the root node of the first tree. Then, the copula family is selected for each of the pairs *i,j,* $j \neq i$. After that, Kendall's tau for each pair of remaining variables, conditioned on the previously selected root variable D, $\tau_{i,j|D}$ is calculated. The second tree is constructed based on maximizing the sum of

240 those absolute Kendall's taus and so on. The metacontour plots describing the bivariate
241 dependencies for AGB and d6, p10, p40 and p70 are presented in Figure 7. The copula families
242 used were Survival Joe-Frank, Rotated Clayton, Tawn and Survival Joe-Frank, respectively
243 (Schepsmeier et al. 2015).
244
245 The copula model was used to simulate 22000 (reflecting the number of cells in the real Våler
246 data) uniformly distributed observations with the modelled dependencies. The copula population
247 was then obtained by calculating the quantiles of the empirical distributions at those simulated
248 values. The quantiles are found from the inverse of the joint copula distribution function. The
249 QQ plots between the original sample (real 2010 data) and simulated distributions for each
250 variable are presented in Figure 8. The distributions all follow the distributions in the original
251 sample fairly well, except for single observations in the extreme tails. The correlations between
252 the variables within the C vine copula population are listed in Table 3. These also represent the
253 true data fairly well: in the 2010 data the largest correlation was 0.95 between p40 and p70,
254 while in C vine copula population it was 0.92. The smallest correlation in the data was 0.60
255 between d6 and p70, while in the simulated population it was 0.52.
256
257 The simulated population units mimic cells from the original data (Table 4, Figure 9). The
258 shapes of the pairwise dependencies have been well captured by the C vine copula population
259 (Figure 9). All the variables have zero as lower bound. The maximum values are a little larger
260 than in the 2010 data (e.g. AGB = 859 Mg ha $^{-1}$ in the copula population, 407 in the data), but the
261 3rd quantiles were almost the same (AGB = 170 Mg ha $^{-1}$ in the copula population, 171 in the
262 data). Having larger maxima is to be expected as the C vine copula population is much larger.
263

As SRS was used in the sampling, we did not need to simulate geographical locations for the population units. For the purposes of this study the simulated population was deemed appropriate.

Finally, we made a larger copula population using the selected 11 variables, namely AGB, p0, p20, p40, p60, p80, hmax, d2, d4, d6 and d8. The properties of the larger copula population are not presented here, but the population was deemed appropriate for the study based on similar inspections as those described for the smaller population above.

3.3 Sampling simulation

The simulations in the copula population were first carried out to determine the number of simulated samples s for which the results stabilized. In this initial phase, the number of simulated samples were s = 100, 200, 500, 1000, 5000, 10 000 and the sample size was n = 100 observations. Then, with the selected s, the simulations were carried out with different sample sizes n = 100, 200, 500, 1000. The external model was the one estimated from the real data from 1999, and internal models were estimated from the samples selected from the simulated copula population.

The Horvitz-Thompson (HT) estimator for the total AGB is

$$\hat{t}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}, \qquad (1)$$

where y_i is the AGB of cell *i* and π_i is the inclusion probability of cell *i*. Assuming a simple random sampling without replacement this inclusion probability is n / N, where *N* is the size of the population. The estimator of the mean is

$$\hat{\bar{y}}_{HT} = \frac{1}{A}\hat{t}_{HT}, \quad (2)$$

where *A* is the total area. Its variance estimator is

$$\mathrm{var}(\hat{\bar{y}}_{HT}) = \frac{1}{A^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}, \quad (3)$$

where π_{ij} is the joint inclusion probability of cells *i* and *j*. When *i*=*j*, this joint probability is π_i, otherwise it is $n(n-1)/N(N-1)$.

The difference estimator for the total AGB is

$$\hat{t}_d = \sum_{i=1}^{N}\hat{y}_i + \sum_{i=1}^{n}\frac{e_i}{\pi_i}, \quad (4)$$

where $\hat{y}_i$ is the model prediction for AGB in cell *i,* and $e_i = y_i - \hat{y}_i$. In case of an internal model, the second term is zero for a linear model with an intercept term. For an external model, the first term of model predictions is the same for all samples and the variation comes from the second term. The mean estimator ($\hat{\bar{y}}_d$) is obtained by dividing the total estimator (4) by *A* as in Eq. 2. Its variance estimator (the simplified estimator assuming g-weights to be 1 for all *i*, Särndal et al 1992 p. 362) is

$$\mathrm{var}(\hat{\bar{y}}_d) = \frac{1}{A^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{e_i}{\pi_i}\frac{e_j}{\pi_j} \quad . \quad (5)$$

In the simulation study, variance estimates were calculated for the *s* samples using the analytical estimator (5) and the mean of these estimates is called the *mean analytical estimate* in what follows. In addition, the *simulated standard error* of the mean was calculated as the standard deviation between the simulated *s* sample means as

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

where

$$\mu = \frac{1}{s}\sum_{i=1}^{s} \hat{\bar{y}}_i \, . \qquad (7)$$

Bias was estimated as the difference between the mean of sample means (μ) and the true mean in the simulated population ($\bar{Y}$). The relative bias (*bias%*) was calculated related to the true mean. Its significance was assessed by the Monte Carlo error (MCE) of the bias estimate, i.e.

$$\text{MCE bias\%} = \frac{100}{\bar{Y}} \frac{\sigma(\hat{\bar{y}})}{\sqrt{s}} \qquad (8)$$

4. Results

When the number of simulated samples *s* was 100 and sample size was *n* = 100, the mean analytical standard error estimate for HT was 2.74% larger than the simulated standard error. For *s* = 500 the difference between these two standard errors turned from positive to negative, 1.90% (Table 5). For the different models the difference between the simulated and analytical standard

323 errors did not markedly change with *s* larger than 500 (see e.g. GAM or LM with external
324 model). Therefore, in the next simulations, *s* = 500 was used.
325
326 All the external models gave unbiased point estimates for all tested sample sizes *n* (Table 6) in
327 the smaller copula population. However, for some internal models bias was observed. The
328 relative biases were statistically significant (i.e. the bias% was greater than twice the MCE
329 bias%) with internal GAM models with all sample sizes, and with the internal LCK model with
330 smallest sample size. The largest relative bias was 0.765%, observed with GAM model with *n* =
331 100.
332
333 For a SRS it is possible to calculate the true sampling variance for varying sample size *n*. The
334 mean analytical standard error calculated with HT estimator (Equation 3) was a good estimate
335 for the true sampling variance for the smallest sample sizes (Table 7). For the largest sample size
336 the simulated standard error (calculated as variation between the simulated samples, see (6))
337 underestimated the true standard error by 7.66%. In this particular case, the mean analytical
338 standard error was closer to the true standard error than the simulated standard error. The
339 simulated standard error varied from the mean analytical error at most by 5.80% (Table 7). The
340 results with HT are shown to reflect the effect of chance in the results.
341
342 The local constant kernel model (LCK) gave the smallest simulated standard errors for the
343 internal model for all sample sizes (Table 7), 13-18% smaller than LM and 3-10% smaller than
344 GAM. Thus, the internal LCK model provided the most precise estimates for the mean AGB in
345 the simulated population. On the other hand, the external LCK model gave larger simulated

346 standard errors than LM and GAM. The internal LCK model had 16-20% smaller simulated
347 standard errors of mean than the external LCK model.
348
349 The variance estimator (5) consistently produced underestimations with the internal LCK model
350 when compared to the simulated variances. The resulting underestimation of standard error was
351 about 33% with the smallest sample size, but reduced to about 14% with increasing sample size.
352 This difference is large enough to be of practical importance. On the other hand, with the
353 external LCK model, the mean analytical variance estimates were well in line with the simulated
354 estimates.
355
356 The generalized additive model (GAM) with penalized splines produced consistently smaller
357 simulated standard errors than LM, 7-13% smaller for internal and 6-11% for external model.
358 For the external model, GAM thus gave the smallest simulated standard errors. The external and
359 internal models gave quite similar simulated standard errors for both models: for GAM the
360 differences were from about -8% to 7% and for LM from -3% to 2 %. The variance estimator (5)
361 produced clearly less severe underestimations of variance for the linear model and the GAM
362 model than for the LCK model. Still, using the internal model with the smallest sample size n =
363 100 provided serious underestimate (24.91%) of the variance for the GAM model with penalized
364 spline smooths. With larger sample size, the underestimation observed with the internal model
365 was reduced. The external GAM model gave so accurate variance estimates that they would most
366 likely be adequate for most practical purposes and with all sample sizes. For the linear model,
367 both the internal and external model worked fairly well for all sample sizes.
368

For the larger copula population with 11 variables, all the internal models for LCK gave biased means (Table 8). Also for the internal GAM models, all means were biased except for the largest sample size n=1000. Highest relative bias was 0.6% for GAM. All the external models again gave unbiased results. For the internal models, the variance was underestimated by the estimator (5) even more than in the smaller copula population (Table 9). Even the internal linear model led to 8-16% underestimation in the standard error with $n \leq 500$, and the internal LCK and GAM even over 50%. This is due to the larger possibilities for optimizing the internal model when more explanatory variables are available. On the other hand, increasing sample size reduced the underestimation clearly, for instance for GAM from 55% to 8% when n increased from 100 to 1000. External models also here gave analytical standard errors well in line with the simulated standard errors.

In order to analyze how much the internal models varied between the samples, we calculated the proportion of LM and GAM models where each of the four explanatory variables were significant (Table 10). It is notable that while p10 and p40 were not significant in the external models, they were still significant in quite many of the internal models. In general, the number of significant variables increased with increasing sample size n. For the LCK models we calculated the proportion of models where the estimated optimal bandwidth was shorter than a given boundary (Figure 10). Thus, also the internal LCK models varied. In the larger population, this variation was higher.

391 **5. Discussion**

392

393 The difference estimator based on local constant kernel model (Watson-Nadaraja type) provided
394 the smallest simulated standard errors when an internal model was used. This can be explained
395 by the fact that the model fitted to the 1999 Våler data had a small error variance and large R^2,
396 and the same can be expected for the internal models fitted to samples. From the external
397 models, the GAM model produced the smallest simulated standard error of mean AGB. The
398 external GAM model was smooth and had smaller error variance and larger R^2 than the external
399 LM. For all the considered model types, the mean analytical standard error estimates (obtained
400 by (5)) were within about 4% of the simulated standard errors in the smaller population and
401 within about 7% in the larger copula population if an external model was used. Thus, a key point
402 in an external model is that the predictions are fixed. Due to that, the wiggliness of external LCK
403 models is less problematic than with internal models.

404

405 For the internal local constant kernel model, the analytical estimator of standard error (see (5))
406 underestimated the simulated standard error 14-35% in the smaller population and 31-56% in the
407 larger population. For the internal linear model, the analytical estimator underestimated the
408 standard errors as well, but the underestimation was less than 7% with all sample sizes in the
409 smaller population and below 17% in the larger population. It remains to be studied, if a sample-
410 dependent g-weighting would improve the results (Särndal et al. 1992 p. 234). For the largest
411 sample size and smaller population, the GAM approach also worked quite well. In the larger
412 population, the underestimations were clear for all n (for $n = 1000$ only 8%). The GAM model,
413 which was quite smooth and close to linear, behaved quite similarly to the linear model in the

smaller population, but more like the LCK model in the larger population. It still remains to be analyzed, how a model that is smooth and monotonic but clearly non-linear would perform.

While the variance estimator (5) using the internal local constant kernel model did not work properly in this analysis, the problem is not necessarily in the kernel method per se, but the properties of the model used. The optimal bandwidths selected using cross validation were quite narrow, with the result that the LCK model was quite wiggly compared to the other models. Increasing the sample size did not reduce the wiggliness, which may be due to overfitting rather than describing a true relationship. In GAM, wiggliness was penalized in the optimization.

We tested also the expected Kullback-Leibler information as the bandwidth selection criterion in the LCK model, and cross-validation based on this criterion produced a clearly smoother (or less wiggly) model than the least squares cross-validation. The model was still less smooth than the GAM model, but the standard error estimator for the internal model was fairly well in line with that of GAM (about 23% underestimation with $s = 500$ and $n = 100$). It seems probable that the wiggliness and non-monotonicity are important factors for the greater underestimation of the standard error with the internal LCK model than with the other internal models.

The problem can also be in the estimation method used; according to Hayfield & Racine (2008), the *np* package can select a local minima. To avoid local minima, we increased the number of restarts from the default four to eight. This change did not have large effect in the field data, however. It is difficult to see if local optima were an important factor for the results here, but the overfitting most likely was.

437

438 Most likely the wiggliness of the LCK model can be reduced with good modelling practices. In
439 the simulations, we needed to make the calculations automatic, but in real modelling situation it
440 would be possible, for instance, to reduce the number of correlated explanatory variables using
441 principal components. We compared the performance of different models in a copula population
442 that was built for the mean AGB and four principal components that were constructed from all
443 the explanatory variables available (see Section 2). However, the results obtained with the
444 internal models did not improve in comparison to using the originally chosen four variables. The
445 standard errors were still underestimated for the internal LCK more than for the other models. It
446 can be concluded that the kernel model is more sensitive to deficiencies in the modelling process
447 than the other methods.

448

449 The local linear kernel model did not work adequately in this study. Some of the random samples
450 from the copula population were possibly quite unbalanced (i.e. with the sample mean for one or
451 more variables far from the population mean). In such a situation the fitted models may produce
452 highly illogical results when the results are calculated for such parts of the data with only few
453 observations or even extrapolated to an area with no observations at all, especially when the
454 model is estimated in an automatic fashion. With non-linear models such extrapolation can in
455 fact be problematic. If the samples from the copula population had been stratified in order to
456 produce a more balanced sample, both the local constant and linear kernel models could have
457 performed better. The importance of balanced data when dealing with non-linear dependencies
458 also remains to be studied.

459

460 In our study, the external model was based on the real Våler field data collected at 1999, whereas
461 the copula population was constructed based on the 2010 data. The external model was thus
462 independent from the copula population as much as a model from a previous inventory can be. In
463 this study, we first pre-selected four (or ten) explanatory variables among the 23 possibilities for
464 copula construction. All these variables were then used in the models, so that the explanatory
465 variables involved did not change from sample to sample. In fact, in the linear model, only the
466 coefficients changed from sample to sample. In the LCK and GAM models, the significance of
467 explanatory variables as well as the model shape could change. However, Li and Racine (2007)
468 and Wood (2006) state that both the kernel model and the GAM model can find the irrelevant
469 explanatory variables without any pre-selection phase. Our case resembles this situation, albeit
470 with a small number of possible explanatory variables compared with most real life situations.
471 The differences in the results between the two copula populations reflect the effect of variables
472 selection.
473
474 The more possibilities there were to optimize the internal model based on a sample, the worse
475 was the underestimation of the standard error. Thus, optimizing the internal models too much
476 may enhance the underestimation. For instance, in the smaller copula population, the residual
477 error of the LM model estimated from the whole copula population with 22000 observations was
478 47.1 mg ha^{-1} and R^2 0.71, while the external model estimated from the relatively small 1999 data
479 was seemingly much more accurate. In this model all four variables were statistically significant,
480 while in the external model they were not.
481

Similar results were obtained by Magnussen et al (2015). By using a copula population mimicking a real population, they concluded that the variance estimators for EPS did not work well when an attempt to minimize the EPS variance was made. Our interpretation here is that optimizing in EPS and in the local constant kernel model easily lead to overlearning (or undersmoothing).

When Dahlke et al. (2013) examined the EPS (which can be interpreted as model-assisted estimation with class variables), the internal models worked fairly well, but they only simulated models that were monotone. Dahlke et al. (2013), Tipton et al. (2013) and Breidt & Opsomer (2008) also made other assumptions that are not realistic in a real forest inventory setting, like the assumption of uniformly distributed auxiliary data. Thus, if used cautiously in suitable conditions the internal models may work well, but not in all conditions.

6. Conclusion

Overfitting and non-monotonicity are possible problems with non-parametric models in model-assisted inference. Using the analytical variance estimator (Equation 5) of the difference estimator for these models appears to lead to serious underestimation of the variance with the internal model. On the other hand, the variance estimator worked well with all the studied external models. Thus, an external model is always recommended. If an internal model is used anyway, e.g. in the absence of an external model, it is recommendable to utilize smooth models in the inference, even when such models are less accurate than a less smooth non-monotonic model. Using explanatory variables selected based on theory should reduce the risk of too much

505 optimization. The copula population approach provided a useful tool for analyzing the
506 performance of the estimators. It was fairly easy to model the population even with 11 variables.
507 It can be recommended also for future studies.

508

509

References

Aas, K., Czado, C., Frigessi, A & Bakken, H. 2009. Pair-Copula constructions of multiple dependence. Insurance: Mathematics and Economics 44:182-198.

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. & Opsomer, J.D. 2000. Local polynomial regression estimators in survey sampling. Annals of Statistics 28:1026-1053.

Breidt, F.J., Claeskens, G. & Opsomer, J.D. 2005. Model-assisted estimation for complex surveys using penalized splines. Biometrika 92:831-846.

Cochran, W.G. 1977. Sampling techniques. John Wiley & Sons.

Dißmann, J., Brechmann, E.C., Czado, C. & Kurowicka, D. 2013. Selecting and estimating regular vine copulae and application to financial returns. Computational Statistics & Data Analysis 59: 52–69.

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

Epanechnikov, V. A. 1969. Non-parametric estimation of a multivariate probability density. Theory Probab. Appl. 14:153–158.

Gregoire, T. G., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., & 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.

533 Hayfield, T. & Racine, J.S. 2008. Nonparametric economics: The np package. Journal of
534 Statistical Software 27:1-32.
535 Kooperberg, C. 2015. logspline: Logspline Density Estimation. Routines. R package version
536 2.1.8. http://CRAN.R-project.org/package=logspline
537 Li, Q & Racine, J.C. 2007. Nonparametric econometrics. Princeton University Press. 746 p.
538 Magnussen, S., McRoberts, R. E., & Tomppo, E. O. 2010. A resampling variance estimator for
539 the k nearest neighbours technique . Canadian Journal of Forest Research 40: 648-
540 658.
541 Magnussen, S., Andersen, H-E & Mundhenk, P. 2015. A second look at endogenous post -
542 stratification. Forest Science 61:624–634.
543 Marklund, L. G. 1988. Biomass functions for pine, spruce and birch in Sweden. Umeå: Swedish
544 University of Agricultural Sciences, Department of Forest Survey (In Swedish.).
545 Massey, A. & Mandallaz, D. 2015. Comparison of classical, kernel-based and nearest neighbors
546 regression estimators using the design-based Monte Carlo approach for two-phase
547 forest inventories. Canadian Journal of Forest Research 45: 1480–1488.
548 Massey, A., Mandallaz, D. & Lanz, A. 2014. Integrating remote sensing and past inventory data
549 under the annual design of the Swiss National Forest Inventory using three-phase
550 design-based regression estimator. Canadian Journal of Forest Research 44: 1177-
551 1186.
552 McRoberts, R.E., & Tomppo, R.O. 2007. Remote sensing support for national forest
553 inventories. Remote Sensing of Environment 110: 412–419.

554 McRoberts, R.E., Tomppo, E.O., Finley, A. O. & Heikkinen, J. 2007. Estimating areal means
555 and variances of forest attributes using the k-Nearest Neighbors technique
556 and satellite imagery. Remote Sensing of Environment 111: 466-480.
557 McRoberts, R.E. & Westfall, J.A. 2015. Propagating uncertainty through individual tree volume
558 model predictions to large-area volume estimates. Annals of Forest Science. DOI
559 10.1007/s13595-015-0473-x
560 Nadaraya, E. A. 1964. On estimating regression. Theory of Probability and its Applications 9:
561 141–2.
562 Næsset, E., Bollandsås, O. M., Gobakken, T., Gregoire, T., Ståhl, G. 2013. Model-assisted
563 estimation of change in forest biomass over an 11 year period in a sample survey
564 supported by airborne LIDAR: A case study with post-stratification to provide
565 "activity data". Remote Sensing of Environment 128: 299-314.
566 Nelsen, R.B. 2006. An introduction to copulas (2nd ed). New York: Springer.
567 Opsomer, J.D., Breidt, F.J., Moisen, G.G., Kauermann, G. 2007. Model-assisted estimation of
568 forest resources with generalized additive models. Journal of American Statistical
569 Association 102:400-409.
570 R Core Team (2014). R: A language and environment for statistical computing. R Foundation
571 for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
572 Saarela, S., Grafström, A., Ståhl, G., Kangas, A., Holopainen, M., Tuominen, S., Nordkvist, K.
573 & Hyyppä, J. 2015a. Model-assisted estimation of forest resources using different
574 combinations of LiDAR and Landsat data as auxiliary information. Remote
575 Sensing of Environment 158:431-440.
576 Saarela, S., Schnell, S., Grafström, A., Tuominen, S., Hyyppä, J., Nordkvist, K., Kangas, A. &
577 Ståhl, G., 2015b. Effects of sample size and model form on the accuracy of model-

based estimators of growing stock volume. Canadian Journal of Forest Research 45:1524-1534.

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

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

Ståhl, G. Holm, S., Gregoire, T.G. Gobakken, T. Næsset, E & Nelson, R. 2011. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. Canadian Journal of Forest Research 41: 96–107.

Wood, S.N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Wu, C. & Sitter, R.R. 2001. A Model-calibration approach to using complete auxiliary information from survey data. Journal of American statistical association 96:185-193.

Table 1. The estimates of the coefficients and their standard error and t-value for the linear model.

	Estimate	Std. Error	t- value	Pr(>|t|)
(Intercept)	-66.508	8.818	-7.542	0.000
d6	146.875	21.887	6.711	0.000
p10	0.842	1.922	0.438	0.662
p40	1.282	3.102	0.413	0.680
p70	7.875	2.145	3.672	0.000

d6 canopy density corresponding to the proportion of echoes above fraction 6 to the total number of echoes and p10, p40 and p70 are percentiles of the ALS height distribution

599 Table 2. The estimate of intercept and its standard error and t-value (first row) and the estimated
600 degrees of freedom and the approximated F test statistics for the penalized spline smooths (the
601 last four rows).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	112.654	2.205	51.09	0.000
	edf	Ref.df	F	p-value
s(d6)	3.205	4.036	16.718	0.000
s(p10)	4.439	5.313	1.198	0.310
s(p40)	1.000	1.000	0.002	0.963
s(p70)	2.638	3.332	7.459	0.000

602
603

604 Table 3. Correlation matrix for the smaller C vine copula population

	AGB	d6	p10	p40	p70
AGB	1.00	0.73	0.76	0.79	0.73
d6	0.73	1.00	0.74	0.70	0.52
p10	0.76	0.74	1.00	0.85	0.73
p40	0.79	0.70	0.85	1.00	0.92
p70	0.73	0.52	0.73	0.92	1.00

605

606

607

608

609 Table 4. Summary of the smaller simulated C vine copula population of 22000 observations

	AGB	d6	p10	p40	p70
Min	0.00	0.00	0.00	0.00	0.00
1st Qu	73.69	0.35	4.59	8.58	10.95
Median	119.64	0.53	6.37	11.40	14.00
Mean	128.15	0.48	6.81	11.08	13.65
3rd Qu	170.95	0.67	8.66	13.78	16.66
Max	859.71	1.12	30.96	31.47	38.20

610

611

612 Table 5. Simulated standard errors (variation between the *s* = 100, 200, 500, 1000, 5000 and
613 10000 simulated samples with sample size *n* = 100) and mean analytical standard error using the
614 HT and difference estimators (Equations 3 and 5) in C vine copula population with *N* = 22000.

Method	Simulated samples *s*	Simulated standard error Mg ha^{-1}		Mean analytical standard error Mg ha^{-1}		Relative difference %	
		Internal	External	Internal	External	Internal	External
HT	100	8.56		8.80		2.74	
	200	8.49		8.73		2.77	
	500	8.92		8.75		-1.90	
	1 000	8.87		8.75		-1.28	
	5 000	8.83		8.78		-0.50	
	10 000	8.87		8.79		-0.90	
LM	100	4.53	4.53	4.56	4.82	0.75	6.41
	200	4.75	4.72	4.50	4.77	-5.18	1.00
	500	4.82	4.73	4.52	4.79	-6.18	1.26
	1000	4.85	4.83	4.51	4.79	-7.03	-0.76
	5000	4.88	4.84	4.51	4.81	-7.54	-0.79
	10000	4.90	4.85	4.51	4.80	-7.87	-0.92
LCK	100	4.32	4.22	2.70	4.96	-37.49	17.53
	200	4.20	4.85	2.76	4.82	-34.34	-0.60
	500	4.16	5.04	2.79	4.88	-33.08	-3.10
	1000	4.17	5.03	2.81	4.89	-32.62	-2.71
	5000	4.22	4.93	2.80	4.90	-33.70	-0.62
	10000	4.21	4.93	2.80	4.90	-33.43	-0.62
GAM	100	5.69	4.13	3.33	4.41	-41.48	6.81
	200	4.42	4.17	3.34	4.38	-24.54	5.04
	500	4.48	4.19	3.36	4.37	-24.91	4.23
	1000	4.50	4.33	3.35	4.34	-25.52	0.12
	5000	4.60	4.37	3.35	4.34	-27.19	-0.68
	10000	4.63	4.38	3.35	4.33	-27.77	-1.19

615 Table 6. Estimated mean AGB, relative bias and the Monte Carlo error (MCE) of relative bias in
616 the smaller C copula population with N = 22000 and s = 500 using the HT and difference
617 estimators (Equations 2 and 4). Population true mean is 128.15 (Mg ha^{-1}).

	Sample size	Estimate of mean Mg ha^{-1}		Bias %		MCE Bias %	
Method		Internal	External	Internal	External	Internal	External
HT	100	128.30		0.121		0.311	
	200	128.17		0.019		0.218	
	500	128.25		0.077		0.134	
	1000	128.12		-0.022		0.090	
LM	100	128.11	128.05	-0.028	-0.078	0.168	0.165
	200	128.07	128.07	-0.058	-0.057	0.116	0.115
	500	128.23	128.21	0.068	0.048	0.072	0.074
	1000	128.18	128.15	0.028	0.005	0.050	0.050
LCK	100	128.58	128.17	0.339	0.014	0.145	0.176
	200	128.23	128.28	0.062	0.105	0.096	0.119
	500	128.27	128.20	0.097	0.040	0.059	0.074
	1000	128.17	128.13	0.015	-0.015	0.042	0.052
GAM	100	129.13	128.07	0.765	-0.062	0.156	0.146
	200	128.67	128.03	0.406	-0.088	0.107	0.105
	500	128.53	128.24	0.298	0.075	0.064	0.069
	1000	128.30	128.15	0.117	0.003	0.043	0.047

618

619

620 Table 7. Simulated standard errors (variation between the s = 500 simulated samples) and mean
621 analytical standard errors obtained using the variance estimator (Equation 5) of the difference
622 estimator (Equation 4) in the smaller C vine copula population with N = 22000.

	Sample size n	Simulated standard error Mg ha^{-1}		Mean analytical standard error Mg ha^{-1}		Relative difference %	
method		Internal	External	Internal	External	Internal	External
HT	100	8.92		8.75 (true 8.81)		-1.90	
	200	6.23		6.23 (true 6.23)		0.00	
	500	3.83		3.89 (true 3.94)		1.68	
	1000	2.57		2.72 (true 2.79)		5.80	
LM	100	4.82	4.73	4.52	4.79	-6.18	1.26
	200	3.32	3.30	3.26	3.41	-1.77	3.35
	500	2.07	2.13	2.07	2.13	-0.03	-0.08
	1000	1.42	1.44	1.45	1.49	1.80	3.57
LCK	100	4.16	5.04	2.79	4.88	-33.08	-3.10
	200	2.74	3.42	2.15	3.49	-21.79	1.91
	500	1.69	2.12	1.43	2.17	-15.47	2.38
	1000	1.20	1.48	1.03	1.52	-14.14	2.89
GAM	100	4.48	4.19	3.36	4.37	-24.91	4.23
	200	3.06	3.00	2.57	3.07	-16.01	2.25
	500	1.84	1.96	1.69	1.91	-8.38	-2.56
	1000	1.24	1.34	1.20	1.34	-2.66	-0.39

623
624
625
626

627 Table 8. Estimated mean AGB, relative bias and the Monte Carlo error (MCE)of relative bias in
628 the larger C copula population with $N = 22000$ and $s = 500$ using the differences estimator
629 (Equation 4). Population true mean is 128.88 (Mg ha^{-1}).

	Sample size	Estimate of mean Mg ha^{-1}		Bias %		MCE Bias %	
Method		Internal	External	Internal	External	Internal	External
LM	100	128.597	128.868	-0.223	-0.013	0.165	-0.013
	200	128.598	128.785	-0.222	-0.077	0.111	-0.077
	500	128.764	128.746	-0.093	-0.107	0.072	-0.107
	1000	128.806	128.847	-0.061	-0.029	0.047	0.053
LCK	100	128.37	128.778	-0.399	-0.083	0.14	-0.083
	200	128.376	128.571	-0.394	-0.243	0.099	-0.243
	500	128.544	128.92	-0.264	0.028	0.059	0.028
	1000	128.623	128.828	-0.203	-0.044	0.038	0.057
GAM	100	129.709	128.73	0.64	-0.119	0.18	-0.119
	200	129.272	128.74	0.301	-0.112	0.106	-0.112
	500	129.052	128.753	0.13	-0.102	0.063	-0.102
	1000	128.94	128.807	0.043	-0.06	0.041	0.063

630

631

632 Table 9. Simulated standard errors (variation between the s = 500 simulated samples) and mean
633 analytical standard errors obtained using the variance estimator (Equation 5) of the difference
634 estimator (Equation 4) in the larger C vine copula population with N=22000.

	Sample size n	Simulated standard error Mg ha^{-1}		Mean analytical standard error Mg ha^{-1}		Relative difference %	
method		Internal	External	Internal	External	Internal	External
LM	100	4.769	4.999	3.989	4.947	-16.355	-1.034
	200	3.208	3.515	2.94	3.495	-8.369	-0.579
	500	2.078	2.331	1.881	2.187	-9.474	-6.16
	1000	1.353	1.53	1.33	1.532	-1.715	0.162
LCK	100	4.021	5.356	1.765	5.222	-56.1	-2.502
	200	2.85	3.728	1.569	3.695	-44.963	-0.869
	500	1.703	2.459	1.166	2.311	-31.563	-6.016
	1000	1.109	1.652	0.865	1.622	-22.004	-1.805
GAM	100	5.185	5.856	2.33	5.714	-55.06	-2.435
	200	3.051	4.026	2.122	4.024	-30.449	-0.054
	500	1.828	2.714	1.481	2.518	-18.939	-7.233
	1000	1.177	1.811	1.078	1.764	-8.391	-2.619

635

636

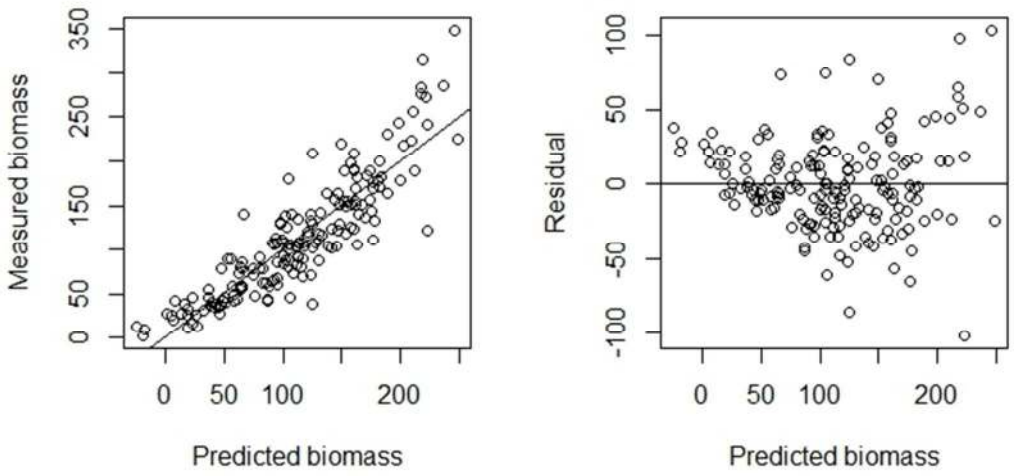
637 Table 10. The proportion of internal models where each of the variables was significant

638 explanatory variable for AGB with linear model and GAM in the smaller C copula population.

	s	*n*	intercept	d6	p10	p40	p70
LM	10000	100	99.85	95.71	48.82	13.13	81.90
	500	200	100	100	72.20	16.60	97.80
	500	500	100	100	95.40	21.80	100
	500	1000	100	100	31.00	100	100
GAM	10000	100	100	99.99	35.06	60.09	82.12
	500	200	100	100	45.20	70.60	96.40
	500	500	100	100	71.10	91.12	100
	500	1000	100	100	95.00	98.20	100

639

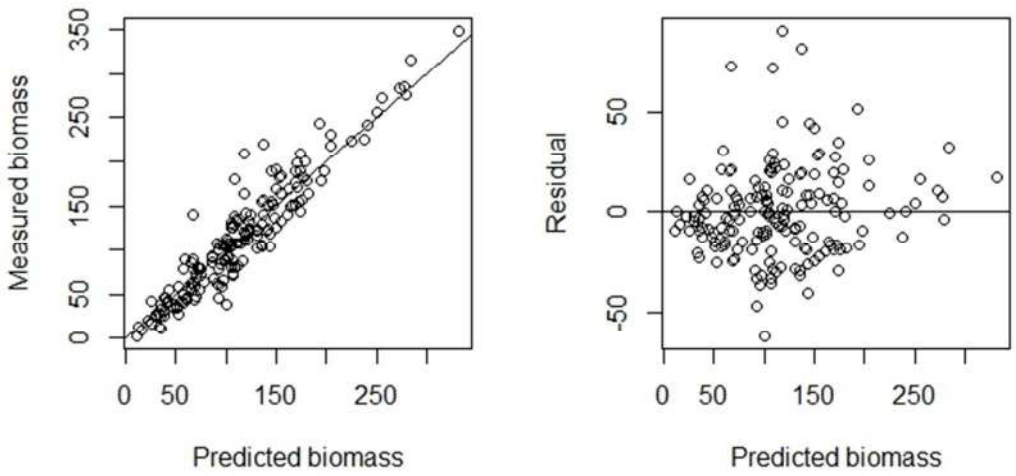
640



641

642 Figure 1. Predicted values and residuals of the linear model estimated from the 1999 data.

643



644

645 Figure 2. Predicted values and residuals for the local constant kernel model estimated from 1999

646 data.

647

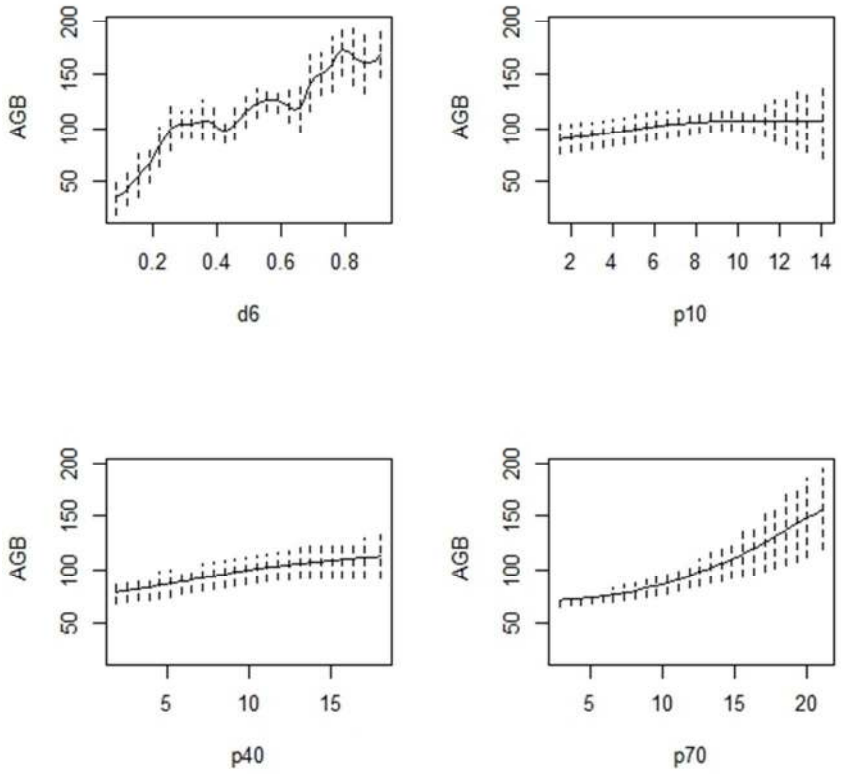
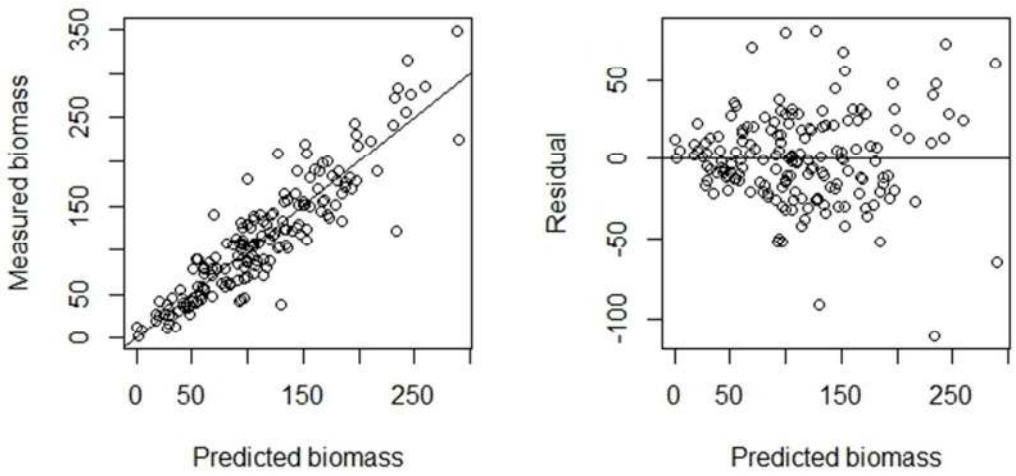


Figure 3. Regression functions of the local constant kernel model for AGB versus d6, (proportion of echoes above fraction 6 to the total number of echoes) and p10, p40, p70 (percentiles of the ALS height distribution) in the local constant kernel model from the 1999 data. The confidence interval based on bootstrapping is given by the dashed lines.



654

655 Figure 4. Predicted values and residuals for the GAM model with penalized splines estimated

656 from the 1999 data.

657

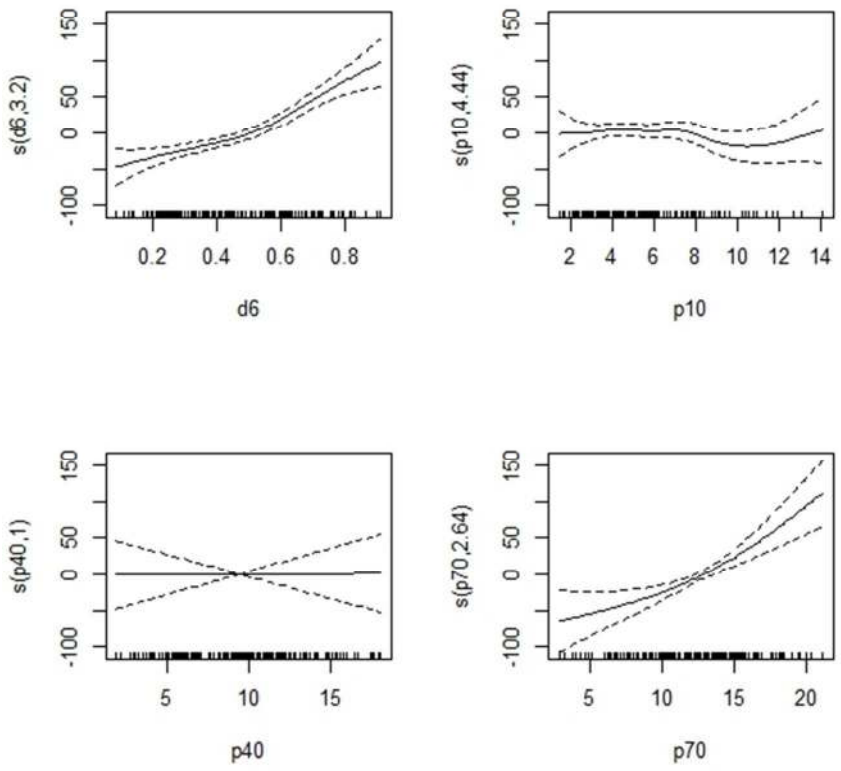
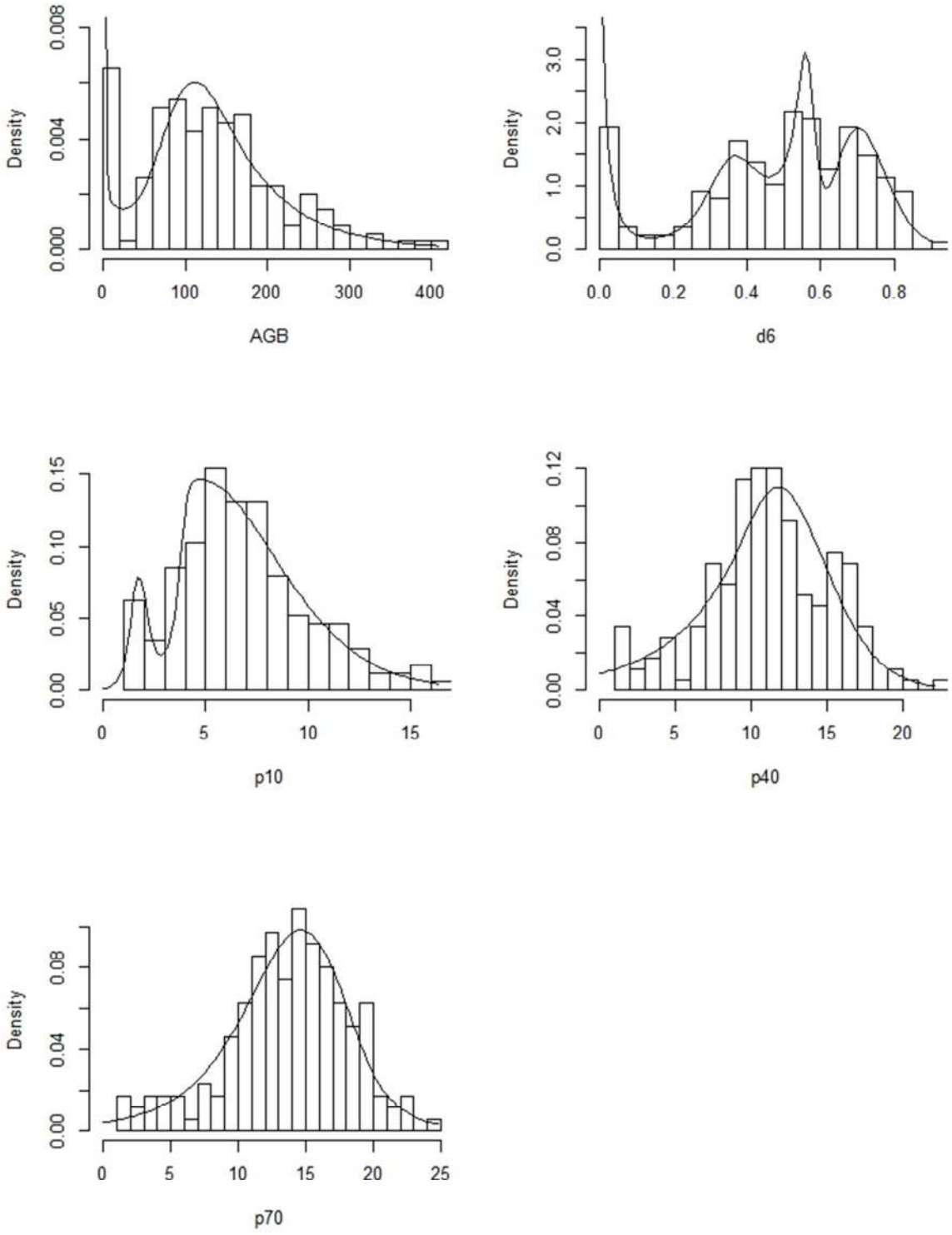
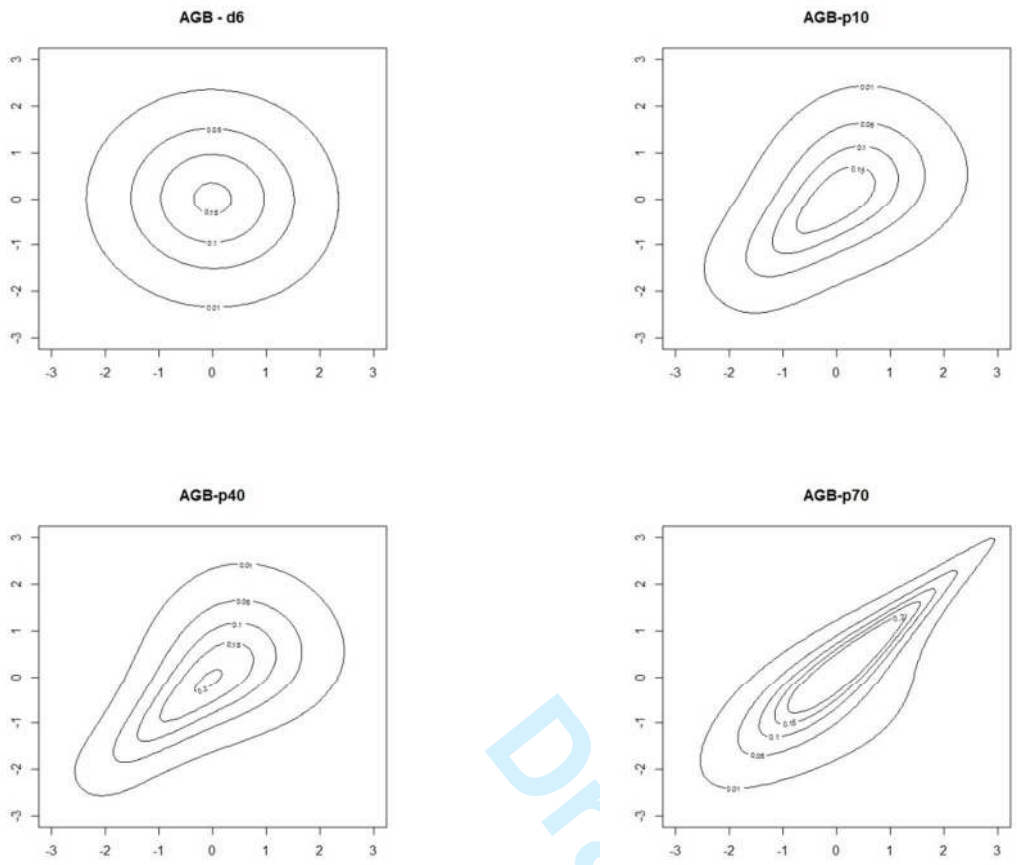


Figure 5. Spline smooths of the GAM model for d6, p10, p40, and p70 estimated from the 1999 data. The dashed lines describe the confidence intervals of the smooths.



662
663 Figure 6. Empirical marginal distributions for the variables AGB, d6, p10, p40 and p70 for C

664 vine copula.

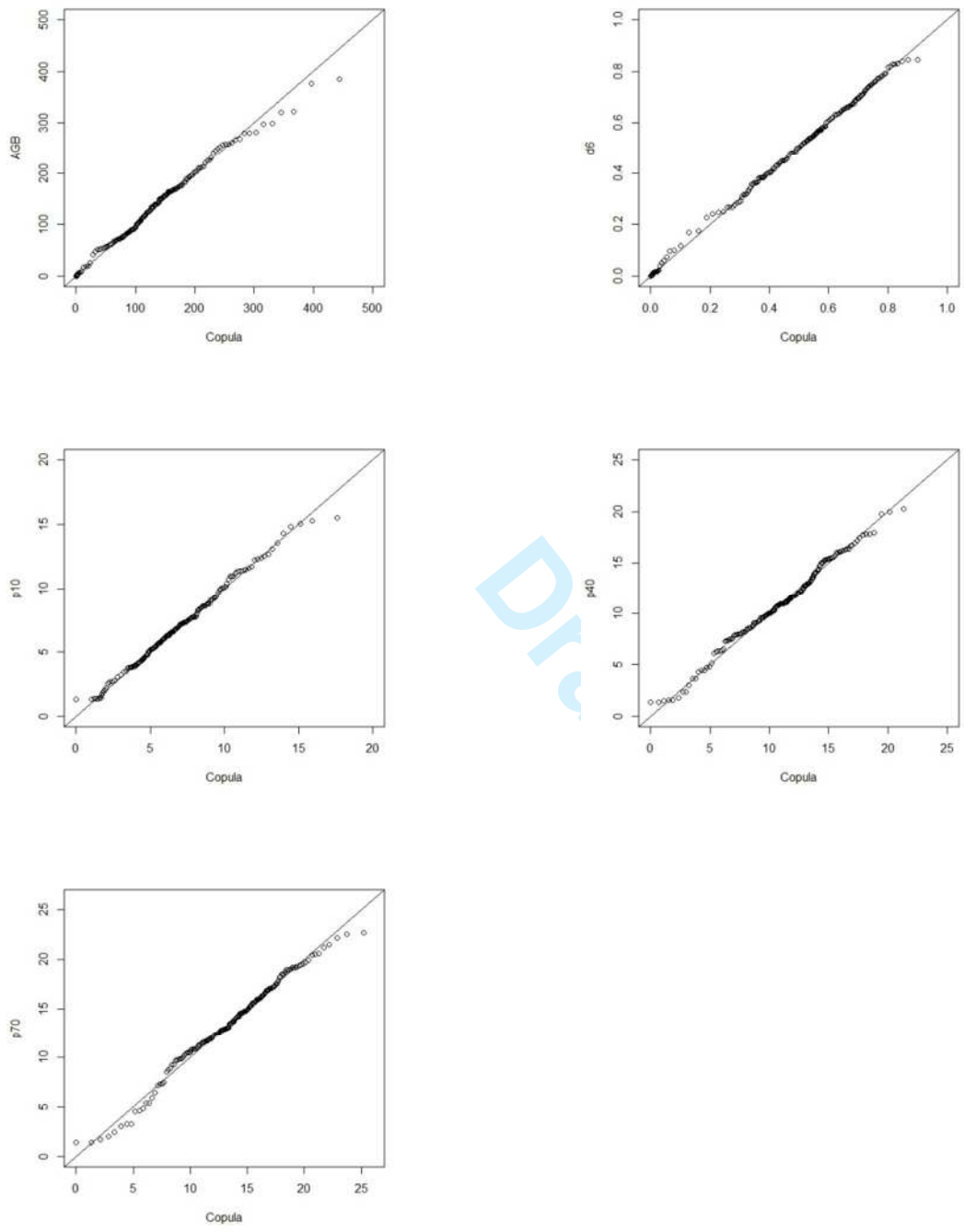


665 Figure 7. Pairwise metacontours for AGB and the variables d6, p10, p40 and p70 in the C vine

666 copula population

667

668

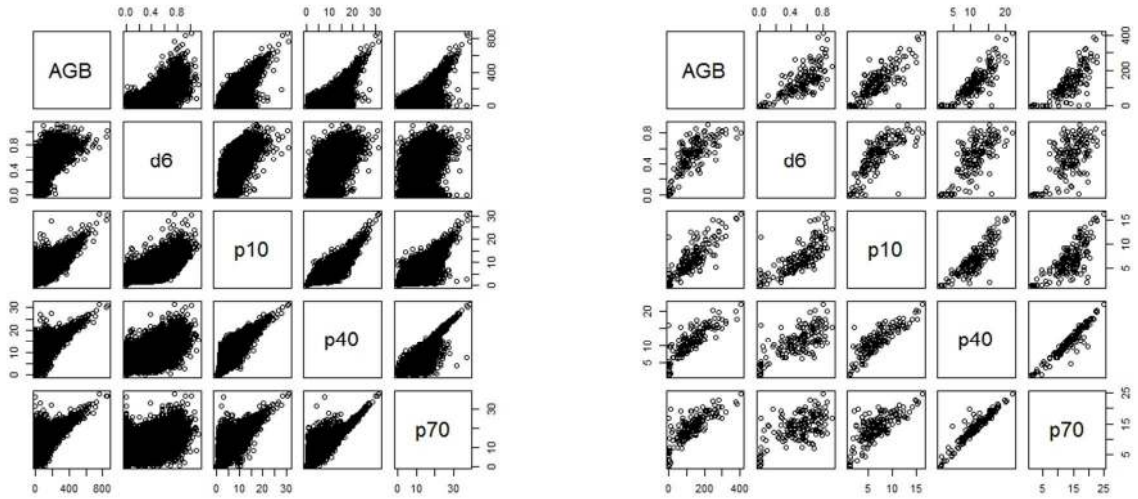


669 Figure 8. QQplots for each variable (AGB, d6, p10, p40 and p70) in the C vine copula

670 population and the original distributions

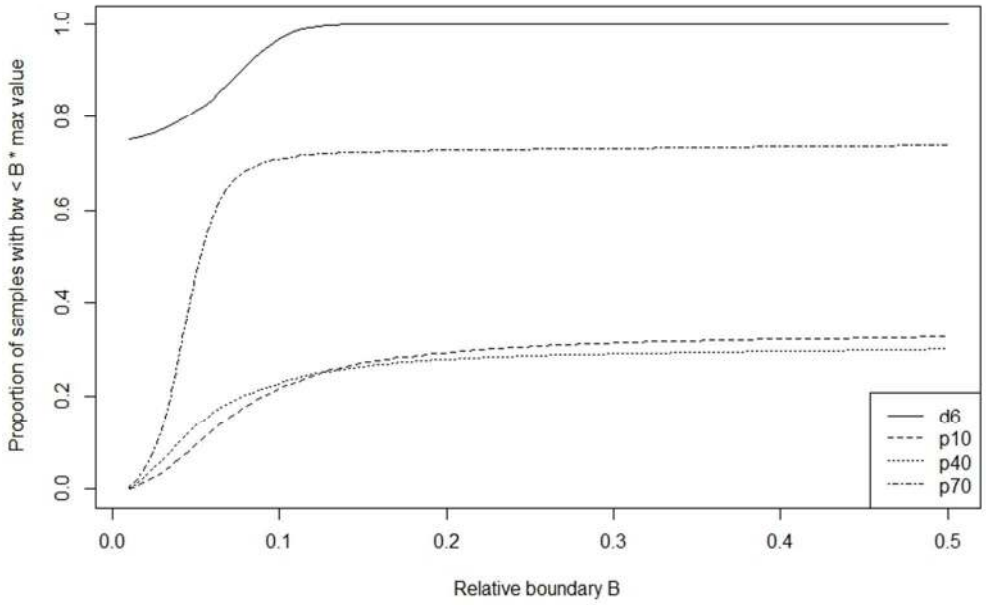
671

672



673 Figure 9. Simulated and sample variable pair scatterplots.

674



675

676 Figure 10. The proportions of the internal LCK models with the bandwidths of the four variables

677 (d6, p10, p40, p70) smaller than the relative boundary B multiplied by the maximum value of the

678 variable in the copula population.

679