NRC Research Press

ARTICLE

Comparison of classical, kernel-based, and nearest neighbors regression estimators using the design-based Monte Carlo approach for two-phase forest inventories

Alexander Massey and Daniel Mandallaz

Abstract: This paper compares design-based properties of the classical two-phase regression estimator with several nonparametric kernel-based estimators of which *k* nearest neighbors (*k*NN) is a special case. Metrics are based on the Euclidean distance applied to either a multidimensional space of explanatory variables or to a one-dimensional space of predictions obtained from a linear model. The main concepts of kernel-based regression estimators are reformulated in the design-based Monte Carlo approach to forest inventory. The results, based on a case study of a forest inventory in Switzerland and extensive simulations, suggest that the commonly used analytical external variance formula may systematically underestimate the true variance for a variety of kernel-based estimators including *k*NN but that it is still adequate for the classical regression estimator. Although using a bootstrap variance can help to correct this underestimation, it was also found that the bootstrap variance estimates could be unstable if the optimal bandwidth is recalculated in each bootstrap sample. These findings suggest that if the model captures the main features of the underlying process, then it is advisable to use the classical regression estimator, because it performs at least as well as the other techniques and is by far simpler to implement.

Key words: two-phase forest inventory, *k*NN, regression estimator, kernel regression, design based.

Résumé : Cet article compare plusieurs estimateurs non paramétriques basés soit sur des noyaux, soit sur la méthode des proches voisins (*k*NN), en utilisant des métriques euclidiennes dans l'espace multidimensionel des variables explicatives ou l'espace unidimensionel des prédictions obtenues par un modèle linéaire. En particulier on considère de nouveaux estimateurs obtenus par l'estimation de l'espérance conditionnelle de la variable réponse donné sa prédiction, et ce avec plans de sondage à deux phases pour inventaires forestiers, en utilisant le formalisme Monte Carlo et d'inférence statistique uniquement sous le plan (design-based model-assisted). Une étude de cas et une simulation montrent que la formule usuelle de la variance externe sous-estime, parfois largement, la variance empirique pour tous les estimateurs excepté l'estimateur classique par régression. La technique d'estimation de la variance par ré-échantillonage permet de remédier à ce défaut mais elle requiert un choix judicieux entre diverses options possibles. Les résultats suggèrent que si le modèle linéaire est adéquat l'estimateur classique est au moins aussi performant que les autres et bien plus simple à mettre en oeuvre. Il semblerait que ce soit le cas pour l'estimation du matériel sur pied en utilisant comme variables explicatives les caractéristiques d'écrivant la canopée accessibles aujourd'hui par les techniques de télédétection comme LiDAR.

Mots-clés : inventaire forestier à deux phases, méthode des *k* plus proches voisins, estimateur par régression, estimateur par noyau, inférence sous le plan de sondage.

1. Introduction

Design-based inference has long been a standard for estimation in forest inventory programs. This is largely due to the fact that one does not need to believe that a model, and all of its underlying assumptions, is true but rather that the sampling design was appropriately carried out via a realization of a well-defined probability sampling scheme. Traditionally, one of the most common models used in design-based estimation is based on linear regression and yields the classical regression estimator. In recent years, there has been a rise in the popularity of nonparametric methods for forest mapping and estimation applications, the most common being some variant of the so-called *k* nearest neighbor (*k*NN) estimator (for examples, see Magnussen et al. (2010) and McRoberts et al. (2007)).

A common method for incorporating these nonparametric methods into the design-based framework is to average model predictions at points where auxiliary information was observed and correcting by a mean residual across points where both the variable of interest and auxiliary information were observed. Estimators of this form are often referred to in the literature as generalized difference estimators (Wu and Sitter 2001). For any estimator of this form, an exactly design-unbiased variance estimator can be derived if one can assume that the model predictions and its residuals are independent and identically distributed (i.i.d.) with respect to the sampling design. This condition is clearly fulfilled if the chosen model is external, that is to say, that it is fitted using a training set (referred to as a reference set in *k*NN literature) that is independent of the sample realized by the current inventory's design, e.g., the training set is a past inventory or a separate inventory conducted in another region (see Massey et al. (2014) for a general explanation of external vs. internal models). However, as it is impractical and economically inefficient to fit using a separate inventory, a common work-around is to fit the

Received 21 April 2015. Accepted 29 June 2015.

A. Massey. WSL, Swiss Federal Research Institute, CH 8903 Birmensdorf, Switzerland.
D. Mandallaz. Chair of Land Use Engineering, ETH Zurich, CH 8092 Zurich, Switzerland.
Corresponding author: Alexander Massey (e-mail: alexander.massey@wsl.ch).

model internally from the current inventory at hand and use the variance estimator derived for external models as if this was not the case. It should be emphasized that in some cases, it can be acceptable to treat internal models as external, as with, for example, the classical regression estimator, provided that the sample size is large enough. In the classical regression case, we are mathematically reassured by asymptotic equivalency to a known consistent design-based variance estimator, derived using the *g*-weight technique initially proposed by Särndal et al. (2003) and adapted to forest inventory by Mandallaz (2008). The Monte Carlo version of the calibration estimators of Wu and Sitter (2001) shows that the external variance is asymptotically valid also for nonlinear regression models (see appendix C in Mandallaz and Massey (2015)). However, analogous reassurances do not exist for the family of nonparametric kernel-based methods, to which *k*NN is a special case, because no practical analytical design-based variance estimator is currently known in the literature for internally fitted models. At present, we are limited to evaluating the design-based properties of kernel-based regression estimators by empirical investigation. However, statistically rigorous evaluations based on both real and simulated data are rare.

The scope of this paper is to test the mathematical validity of treating internal models as external under a moderate sample size for a variety of kernel-based regression estimators, using empirical examples arising from a case study of a Swiss forest inventory in the Canton of Grisons, as well as in a rigorous simulation example. A variety of bandwidth selection algorithms and two main metrics are considered: one based on the Euclidean distance in a multidimensional space of explanatory variables and the other based on the one-dimensional Euclidean distance between predictions from a linear regression model. New nonparametric estimators are proposed that are based on the conditional expectation of the response variable, given its linear regression prediction, in an effort to address a known drawback of many kernel-regression estimators, namely the curse of dimensionality and that the predictions are bounded by the range of observed response variables. The underlying theoretical concepts of kernel-based estimators are summarized and reformulated in the design-based Monte Carlo approach to forest inventory. Finally, variance estimates derived under the external model assumption are compared with a nonparametric bootstrap, and a simulation example is given in which all variance estimates can be compared with the true design-based variance of the estimator for that simulation scenario.

2. Mathematical terminology

We use the Monte Carlo approach (infinite population model), as described chapter 5 in Mandallaz (2008). We consider the forested area F of surface area $\lambda(F)$. For a well-defined population $\mathcal{P}$ of N trees, we have two population parameters of interest: the sum $Y_F = \sum_{i=1}^{N} Y_i$ of the response variable Y_i, e.g., the total volume of the N trees, and the spatial mean $\bar{Y}_F = (1/\lambda(F)) \sum_{i=1}^{N} Y_i$. For the sampling design, we have a sample s_1 of n_1 points independently and uniformly distributed in F, out of which n_2 are selected by simple random sampling (SRS) without replacement. Note that $x \in s_2$ inherit i.i.d. uniformness in F. At all points $x \in s_1$, we have the vector of auxiliary information $\boldsymbol{Z}(x) \in \mathbb{R}^p$. For each point $x \in s_2$, trees are drawn from the population $\mathcal{P}$ with probabilities π_i, e.g., with concentric circles or angle count techniques. The set of trees selected at point x is denoted by $s_2(x)$. From each of the selected trees $i \in s_2(x)$, one determines Y_i. The indicator variable $I_i(x)$ is defined as $I_i(x) = 1$ if $i \in s_2(x)$ and 0 otherwise. At each point $x \in s_2$, the terrestrial inventory provides the local density $Y(x)$, defined by

$$Y(x) = \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x) Y_i}{\pi_i} = \frac{1}{\lambda(F)} \sum_{i \in s_2(x)} \frac{Y_i}{\pi_i} \tag{1}$$

The term $1/(\lambda(F)\pi_i)$ is the tree extrapolation factor f_i (ha^{-1}). Because of possible boundary adjustments, $\lambda(F)\pi_i = \lambda(F \cap K_i)$, where K_i is the inclusion circle of the i th tree. In the Monte Carlo approach, one samples the function $Y(x)$, and we have for the expectation with respect to a random point x that is uniformly distributed in F the fundamental relationship

$$\mathbb{E}_x(Y(x)) = (1/\lambda(F)) \int_F Y(x) \mathrm{d}x = (1/\lambda(F)) \sum_{i=1}^{N} Y_i$$

At any point $x \in s_1$, a prediction $\hat{Y}_m(x)$ is constructed with $\boldsymbol{Z}(x)$. In the external model approach, the subscript m identifies any model used to calculate the predictions, provided that one can assume that the prediction $\hat{Y}_m(x)$ depends only on $\boldsymbol{Z}(x)$ because it is fitted (at least in theory) from a training set that is independent to the sample selection of the current inventory (i.e., s_1 and s_2). For illustration of external vs. internal models, consider a linear regression model where we set $\hat{Y}_m(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}$ for a fixed given $\hat{\boldsymbol{\beta}}$ (presumably based on past or similar inventory data) as opposed to $\hat{Y}_m(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}$ for a $\hat{\boldsymbol{\beta}}$ dependent on s_2. We shall consider generalized difference estimators of the form

$$\hat{Y}_m = \frac{1}{n_1} \sum_{x \in s_1} \hat{Y}_m(x) + \frac{1}{n_2} \sum_{x \in s_2} R_m(x) \tag{2}$$

where $R_m(x) = Y(x) - \hat{Y}_m(x)$ are the theoretical residuals, which generally do not have zero mean for external models. For linear and nonlinear prediction models, eq. 2 can be viewed in the Monte Carlo approach as a special case of the calibration estimator discussed by Wu and Sitter (2001) (see appendix of Mandallaz and Massey (2015) for an adaptation to the infinite population approach and proof). Using the law of total variance, it can be shown that under the external model approach, $\hat{Y}_m$ is exactly design unbiased for $\bar{Y}_F$, with the following theoretical variance (see Mandallaz (2008)):

$$\mathbb{V}(\hat{Y}_m) = \frac{1}{n_1} \mathbb{V}(Y(x)) + \left(1 - \frac{n_2}{n_1}\right) \frac{1}{n_2} \mathbb{V}(R_m(x)) \tag{3}$$

Note that if one assumes that the first phase is exhaustive, as many practitioners do, then eq. 3 simplifies to $\mathbb{V}(\hat{Y}_m) = (1/n_2)\mathbb{V}(R_m(x))$. We now define the external variance estimator using the following sample-copy variance estimators:

$$\hat{\mathbb{V}}(\hat{Y}_m) = \frac{1}{n_1} \frac{\sum_{x \in s_2} (Y(x) - \bar{Y}_{s_2})^2}{n_2 - 1} + \left(1 - \frac{n_2}{n_1}\right) \frac{1}{n_2} \frac{\sum_{x \in s_2} (R_m(x) - \bar{R}_{s_2})^2}{n_2 - 1} \tag{4}$$

with $\bar{Y}_{s_2} = (1/n_2) \sum_{x \in s_2} Y(x)$ and $\bar{R}_{s_2} = (1/n_2) \sum_{x \in s_2} R_m(x)$.

We emphasize that the approach used throughout this article is purely design based and not model dependent, which may be a source of confusion for some readers. A detailed explanation of the philosophical differences between these approaches can be found in Gregoire (1998) in the finite population context and in chapters 4, 6, and 7 in the Monte Carlo setup in Mandallaz (2008). In the model-dependent approach, at a fixed point x, $Y(x)$ is viewed

as the realization of a stochastic process. For example, a common model-dependent distribution assumption would be that $Y(x)$ has mean $\boldsymbol{Z}^t(x)\boldsymbol{\beta}$ and $R_m(x)$ has zero mean and a spatial covariance structure such as in the geostatistical kriging procedure (see chapter 7 in Mandallaz (2008)). In the design-based framework, x is random, and the sample must be the result of a stochastic selection characterized by the sampling design. We are free of any distribution assumptions on $Y(x)$, $\boldsymbol{Z}(x)$, and $R_m(x)$, which can only be regarded as random because x is random (and therefore are fixed for a given x). Inference is based on the theoretical distribution of estimates generated by the sampling design. According to the terminology of Särndal et al. (2003), this approach is model assisted, i.e., we use models to reduce the variance but do not assume that they are correct. Some authors use the term model based instead of model dependent, which is, in our opinion, a source of confusion, because the inference is valid only if the model is true.

In practice, the external variance estimator cannot be expected to be valid sensu stricto, because the prediction $\hat{Y}_m(x_0)$ at an arbitrary point x_0 is almost always fitted internally using the current inventory data $\{\boldsymbol{Z}(x), Y(x), x \in s_2\}$. For this reason, we shall denote the empirical residual $Y(x) - \hat{Y}_m(x)$ by $\hat{R}_m(x)$, which should be used in eq. 4 in place of $R_m(x)$. So far, design-based asymptotic analytical results are available only for linear models, e.g., the classical regression estimator discussed in Section 3, and for nonlinear models via a Monte Carlo model calibration approach (see appendix C in Mandallaz and Massey (2015)). In these cases, asymptotic unbiasedness and consistency of the external variances estimates (eq. 4) can be proven. This is not the case for the other nonparametric techniques used to obtain the predictions $\hat{Y}_m(x)$ that are presented in subsequent sections of this paper. Intuition suggests that using external variance estimators calculated with eq. 4 with internal models will underestimate the theoretical variance, because we are essentially ignoring variation in the value of $\hat{Y}_m(x_0)$ across different sample realizations of s_2. As the design-based variances for these estimators are currently not very well understood from an analytical point of view, one must rely on finding an adequate resampling method such as the bootstrap routine.

The bootstrap routine should be selected in such a way that it mimics the sampling process that produced the original sample. The simplest formulation for the given design here is to create B independent bootstrap samples each of size n_1 using SRS with replacement from the observed $x \in s_1$. From each bootstrap sample, we calculate a bootstrap replicate version of eq. 2 denoted $\hat{Y}_m^b$, where $b = 1, 2, \ldots, B$. The bootstrap population estimate is then

$$\hat{Y}_m^* = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_m^b \tag{5}$$

and the bootstrap estimate of variance is

$$\hat{\mathbb{V}}^*(\hat{Y}_m) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{Y}_m^b - \hat{Y}_m^*\right)^2 \tag{6}$$

Note that each bootstrap sample will have a random number of points $x \in s_2$, which could lead to bootstrap samples with insufficient second-phase sample size to (adequately) fit the desired model internally. This situation, albeit extremely rare for the sample sizes considered here, can be handled by either throwing out the replicate or by considering a stratified bootstrap in which n_2 points are resampled from s_2 and n_1–n_2 are resampled independently from s_1 (see chapter 5.6 in Wolter (2007) for details). This formulation is slightly more computationally expensive but ensures a fixed sampling fraction across all bootstrap replicates.

3. Classical regression estimator

We now present different choices for $\hat{Y}_m(x)$. Consider the linear model

$$Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta} + R(x) \tag{7}$$

In the Monte Carlo approach, it is important to understand that we do not assume that $Y(x)$ has been generated by eq. 7. The real forest, and hence $Y(x)$, is likely to have been generated by a far more complex mechanism, probably with variables not contained in $\boldsymbol{Z}(x)$ and with some nonlinear relationships. We shall view eq. 7 as the working or fitted model. In Section 6.2, we discuss a simulation example, in which we are in the lucky situation to know the true model, i.e., the model used to generate $Y(x)$, so that we can assess the impact of using a working model that is not the true model generating $Y(x)$. To put it in simple terms, the (model-assisted) Monte Carlo objective is not to model the mechanism generating the forest but, rather, to provide estimators that have a much smaller variance than the ordinary sample mean $\bar{Y}_{s_2} = (1/n_2)\sum_{x\in s_2} Y(x)$.

Within the framework of the working model eq. 7, the true regression coefficient $\boldsymbol{\beta}$ is, by definition, the theoretical least squares estimate minimizing

$$\int_F R^2(x)\mathrm{d}x = \int_F \left(Y(x) - \boldsymbol{Z}^t(x)\boldsymbol{\beta}\right)^2\mathrm{d}x \tag{8}$$

It satisfies the normal equation

$$\left(\int_F \boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\mathrm{d}x\right)\boldsymbol{\beta} = \int_F Y(x)\boldsymbol{Z}(x)\mathrm{d}x \tag{9}$$

and the orthogonality relationship

$$\int_F R(x)\boldsymbol{Z}(x)\mathrm{d}x = 0 \tag{10}$$

We shall assume that $\boldsymbol{Z}(x)$ contains the intercept term 1, or more generally, we shall assume that the intercept can be expressed as a linear combination of the components of $\boldsymbol{Z}(x)$, which then ensures that the mean residual is zero, i.e., $\int_F R(x)\mathrm{d}x = 0$. The least squares estimate of the regression vector parameters is

$$\hat{\boldsymbol{\beta}}_{s_2} := \boldsymbol{A}_{s_2}^{-1}\boldsymbol{U}_{s_2} \tag{11}$$

where $\boldsymbol{A}_{s_2} := (1/n_2)\sum_{x\in s_2}\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)$ and $\boldsymbol{U}_{s_2} := (1/n_2)\sum_{x\in s_2} Y(x)\boldsymbol{Z}(x)$ is asymptotically design unbiased for $\boldsymbol{\beta}$. The empirical predictions and residuals are $\hat{Y}(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2}$ and $\hat{R}(x) = Y(x) - \hat{Y}(x)$. Plugging $\hat{Y}(x)$ into the general form described in eq. 2, we get the classical regression estimator for two phases

$$\hat{Y}_{\mathrm{reg}} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}(x) + \frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x) \tag{12}$$

For the variance, we now work under the external model assumption, which, in this case, is the equivalent of saying that we neglect the error in $\hat{\boldsymbol{\beta}}_{s_2}$ and formally set $\hat{\boldsymbol{\beta}}_{s_2} \equiv \boldsymbol{\beta}$. Using the sample copy of eq. 3 and the property that $\hat{Y}(x)$ and $\hat{R}(x)$ have zero empirical design-based covariance, we obtain the two asymptotically equivalent external variance estimates

$$\hat{\mathbb{V}}(\hat{Y}_{\text{reg}}) = \frac{1}{n_1}\frac{\sum_{x\in s_2}(Y(x) - \bar{Y}_{s_2})^2}{n_2 - 1} + \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\frac{\sum_{x\in s_2}(\hat{R}(x) - \bar{\hat{R}}_{s_2})^2}{n_2 - 1} \tag{13}$$

$$\hat{\mathbb{V}}(\hat{Y}_{\text{reg}}) \approx \frac{1}{n_1}\frac{\sum_{x\in s_1}(\hat{Y}(x) - \bar{\hat{Y}}_{s_1})^2}{n_1 - 1} + \frac{1}{n_2}\frac{\sum_{x\in s_2}(\hat{R}(x) - \bar{\hat{R}}_{s_2})^2}{n_2 - 1} \tag{14}$$

with $\bar{\hat{Y}}_{s_1} = (1/n_1)\sum_{x\in s_1}\hat{Y}(x)$ and $\bar{\hat{R}}_{s_2} = (1/n_2)\sum_{x\in s_2}\hat{R}(x)$, which in this case is zero, so that one could use n_2 instead of $n_2 - 1$ in the empirical variance of the residuals. This also implies that $\hat{Y}_{\text{reg}} = \bar{\boldsymbol{Z}}_{s_1}^t\hat{\boldsymbol{\beta}}_{s_2}$, with $\bar{Z}_{s_1} = (1/n_1)\sum_{x\in s_1}\boldsymbol{Z}(x)$.

We can compare eq. 14 with a variance estimate that incorporates the effect of internally fitting $\hat{Y}(x)$ with a linear model, e.g., the *g*-weight variance estimate. The asymptotic design-based variance–covariance matrix of $\hat{\boldsymbol{\beta}}_{s_2}$ is

$$\hat{\Sigma}_{\hat{\beta}_{s_2}} := \boldsymbol{A}_{s_2}^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}(x)^t\right)\boldsymbol{A}_{s_2}^{-1} \tag{15}$$

Using eq. 15 and a first-order Taylor expansion of $\bar{\boldsymbol{Z}}_{s_1}^t\hat{\boldsymbol{\beta}}_{s_2}$, we obtain the consistent *g*-weight variance estimate of $\hat{Y}_{\text{reg}}$, which can be shown to be asymptotically equivalent to eq. 14 (for details, see Mandallaz (2008), pp. 123–125 and Mandallaz (2013), eq. 23) and thus conclude that the external variance estimate is also consistent for internal linear models. Although the *g*-weight variance has slightly better statistical properties from a theoretical point of view, analogous versions are unfortunately not available for the nonparametric estimators discussed in the following sections. For this reason, we will restrict ourselves to external variance estimates (i.e., eq. 13) for comparability.

4. Kernel-based regression estimators

There is a huge amount of literature on kernel-based estimators but the overwhelming majority of the papers are in the model-dependent one-dimensional framework. In this paper, we only give the main results required for practical work. The interested reader is encouraged to consult Györfi et al. (2002) for general theory in nonparametric regression and the online technical report by Mandallaz and Massey (2015) for informal proofs and further examples with alternative numerical options relevant to this article.

We will use primarily one-dimensional kernels of the form $K: u \in \mathbb{R} \to K(u) > 0$. The kernel $K(\,)$ is a probability density function with zero mean and finite variance, i.e $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$, $\int_{\mathbb{R}} uk(u)\mathrm{d}u = 0$ and $\int_{\mathbb{R}} u^2K(u)\mathrm{d}u < \infty$. Popular choices are

$$\begin{aligned} K(u) &= 0.5I_{[-1,1]}(u) \quad \text{uniform kernel} \\ K(u) &= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-u^2}{2}\right) \quad \text{normal kernel} \\ K(u) &= \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right)I_{|u|\le\sqrt{5}}(u) \quad \text{Epanechnikov kernel} \end{aligned} \tag{16}$$

The first and third kernels have a finite support, whereas the normal kernel does not. The concept can be easily generalized to a multidimensional kernel $K(\boldsymbol{u})$ with $\boldsymbol{u} = (u_1, u_2, \ldots u_p) \in \mathbb{R}^p$, by setting $k(\boldsymbol{u}) = \prod_{l=1}^p K_l(u_l)$, where the $K_l(\cdot)$ are one-dimensional kernels.

A standard theoretical point of departure is to consider the conditional expectation of $Y(x)$ given the auxiliary information, i.e., $\mathbb{E}(Y(x)|\boldsymbol{Z}(x) \in \mathbb{R}^p)$, which requires the use of multivariate kernels to estimate multivariate densities and is subject to the curse of dimensionality problem when p is large (for a dramatic example, see Lehmann (1999), pp. 419–420). For simplicity and in an effort to escape the curse of dimensionality, we propose using $\hat{Y}(x) \in \mathbb{R}^1$ as a concise one-dimension summary of $\boldsymbol{Z}(x) \in \mathbb{R}^p$, which leads us to consider

$$\mathbb{E}(Y(x)\,|\,\hat{Y}(x)) = \mathbb{E}(\hat{Y}(x) + R(x)\,|\,\hat{Y}(x)) \tag{17}$$

$$\mathbb{E}(Y(x)\,|\,\hat{Y}(x)) = \hat{Y}(x) + \mathbb{E}(R(x)\,|\,\hat{Y}(x)) \tag{18}$$

To have an approximation of this conditional expectation, we need estimates of the joint bivariate density of $(Y(x), \hat{Y}(x))$ or $(\hat{R}(x), \hat{Y}(x))$ and of the marginal density of $\hat{Y}(x)$. This is done with bivariate and univariate Nadaraya–Watson density estimators (essentially smoothed versions of histograms) to obtain first an estimate of the conditional density and then, by integration, the estimate of the conditional expectation (see Mandallaz and Massey (2015) for details). Estimates will be defined point wise at any arbitrary point $x_0 \in s_1$, where for notational simplicity, we denote $\hat{y}_0 := \hat{Y}(x_0)$ when needed. Using the first form (i.e., eq. 17), we are led to the estimator

$$\hat{Y}_\epsilon^{(1)}(x_0) := \hat{\mathbb{E}}(Y(x_0)\,|\,\hat{Y}(x_0)) := \frac{\frac{1}{n_2\epsilon(n_2,\hat{y}_0)}\sum_{x\in s_2}Y(x)K\left(\frac{\hat{Y}(x_0) - \hat{Y}(x)}{\epsilon(n_2,\hat{y}_0)}\right)}{\frac{1}{n_2\epsilon(n_2,\hat{y}_0)}\sum_{x\in s_2}K\left(\frac{\hat{Y}(x_0) - \hat{Y}(x)}{\epsilon(n_2,\hat{y}_0)}\right)} \tag{19}$$

whereas the second form, $\hat{Y}(x) + \mathbb{E}(R(x)|\hat{Y}(x))$, leads to

$$\hat{Y}_\epsilon^{(2)}(x_0) = \hat{Y}(x_0) + \hat{R}_{\epsilon,\text{smooth}}^{(2)}(x_0) \tag{20}$$

and

$$\hat{R}_{\epsilon,\text{smooth}}^{(2)}(x_0) := \hat{\mathbb{E}}(\hat{R}(x_0)\,|\,\hat{Y}(x_0)) := \frac{\frac{1}{n_2\epsilon_{\hat{y}}(n_2,\hat{y}_0)}\sum_{x\in s_2}\hat{R}(x)K\left(\frac{\hat{Y}(x_0) - \hat{Y}(x)}{\epsilon_{\hat{y}}(n_2,\hat{y}_0)}\right)}{\frac{1}{n_2\epsilon_{\hat{y}}(n_2,\hat{y}_0)}\sum_{x\in s_2}K\left(\frac{\hat{Y}(x_0) - \hat{Y}(x)}{\epsilon_{\hat{y}}(n_2,\hat{y}_0)}\right)} \tag{21}$$

The Nadaraya–Watson regression estimator is simply a weighted average where the weights are calculated based on a kernel, and the fact that it can be derived using the conditional expectation lends it legitimacy because the conditional expectation is the best prediction, where "best" means minimum mean squared error. The choice of the kernel is of less importance than the bandwidth, which is denoted $\epsilon_{\hat{y}}(n_2, \hat{y}_0)$ (e.g., see Lehmann (1999)). The bandwidth is the tuning parameter that controls how smooth the predictions will be across various $\hat{Y}(x)$ (recall that $\hat{Y}(x)$ in this context is considered to be the auxiliary variable). Although the bandwidth can be arbitrarily fixed, it makes more sense to choose an optimal bandwidth based on some prediction criterion (e.g., asymptotic integrated mean-square error). Hence, in our notation, the bandwidth, $\epsilon_{\hat{y}}(n_2, \hat{y}_0)$, depends on n_2 and $\hat{y}_0$. Global bandwidths depend only on n_2, whereas locally varying bandwidths are allowed to expand for sparsely populated areas of the auxiliary space in the reference set and thus depend on both n_2 and $\hat{y}_0$. In

the context of this paper, we found that locally varying bandwidths lead to spurious behavior near the boundaries and very unstable bootstrap variance estimators (for examples, see Mandallaz and Massey (2015)). For this reason, only global bandwidth selection strategies will be considered here.

We can now define the two-phase kernel-based regression estimators as

$$\hat{Y}_{\epsilon}^{(l)} := \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}_{\epsilon}^{(l)}(x) + \frac{1}{n_2}\sum_{x\in s_2}\hat{R}_{\epsilon}^{(l)}(x), \quad l = 1, 2 \tag{22}$$

where $\hat{R}_{\epsilon}^{(l)}(x) = Y(x) - \hat{Y}_{\epsilon}^{(l)}(x)$ for $l = 1, 2$. Note that the sum of the residuals is no longer zero in general. In the absence of more tractable analytical methods to derive the true design-based expectation and variance of eq. 22, one conjectures that the design unbiasedness of eqs. 2 and 3, which is guaranteed for any external model, still holds at least asymptotically for internal $\hat{Y}_{\epsilon}^{(l)}$. In other words, one hopes that an analogy with the classical regression estimator exists where $\hat{Y}_{\epsilon}^{(l)}$ is asymptotically design unbiased with the asymptotically design-unbiased variance estimate

$$\hat{\mathbb{V}}\left(\hat{Y}_{\epsilon}^{(l)}\right) = \frac{1}{n_1}\hat{\mathbb{V}}(Y(x)) + \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2(n_2-1)}\sum_{x\in s_2}\left(\hat{R}_{\epsilon}^{(l)}(x) - \bar{\hat{R}}_{s_2}^{(l)}\right)^2, \quad l = 1, 2 \tag{23}$$

where $\hat{\mathbb{V}}(Y(x)) = (1/(n_2-1))\sum_{x\in s_2}(Y(x - \bar{Y}_{s_2})^2$ and $\bar{\hat{R}}_{s_2}^{(l)} = (1/n_2)\sum_{x\in s_2}\hat{R}_{\epsilon}^{(l)}(x)$.

Remarks:

- There are some hidden mathematical difficulties in the previous arguments. Under a true external approach with a fixed $\boldsymbol{\beta}$, the $\hat{Y}(x)$ are i.i.d. under the design, but the bivariate distribution of $(Y(x), \hat{Y}(x))$ or $(\hat{R}(x), \hat{Y}(x))$ is degenerate, as all the realizations lie on a one-dimensional curve in $\mathbb{R}^2$. However, with internal models, we use $\hat{\boldsymbol{\beta}}_{s_2}$, and this is no longer true. The $\hat{Y}(x)$ values have a correlation of order n_2^{-1}. Note that in contrast to the standard notation in the literature on nonparametric regression, $\hat{Y}(x)$ plays formally the role of the independent variable, assumed to be fixed or i.i.d. distributed on $\mathbb{R}$, and $Y(x)$ plays the role of the response variable. In a purely pragmatic approach, we dismiss the correlation of the $\hat{Y}(x)$'s as being asymptotically negligible and compare empirically the estimators based on eqs. 19 or 20. The true theoretical mathematical properties are certainly much more difficult to derive than in the standard setup.
- Note that to ensure consistency at x_0, one must have $\epsilon(n_2) \to 0$ and $n_2\epsilon(n_2) \to \infty$ as $n_2 \to \infty$ in the standard setup and intuitively also in our case, as the correlation of the $\hat{Y}(x)$ is asymptotically negligible (see Mandallaz and Massey (2015)).
- If the linear model is given by a poststratification (i.e., $\boldsymbol{Z}(x)$ is a vector of strata indicator variables), then $\hat{Y}(x)$ has a discrete distribution taking only the values of the strata means. One can check that both estimators $\hat{Y}_{\epsilon}^{(1)}(x_0)$ and $\hat{Y}_{\epsilon}^{(2)}(x_0)$ will tend to $\hat{Y}(x_0)$, the stratum mean, for $\epsilon_{\hat{y}}(n_2, \hat{y}_0)$ small enough. If the $\hat{Y}(x)$ have a continuous density function, then for $x_0 \in s_2$, both estimators will tend to $Y(x_0)$ (little smoothing and wiggled regression curve). If the bandwidth goes to infinity (strong smoothing), then $\hat{Y}_{\epsilon}^{(1)}(x_0)$ tends to the sample mean (the model is completely ignored) and $\hat{Y}_{\epsilon}^{(2)}(x_0)$ goes to $\hat{Y}(x_0)$ and nothing is gained compared with the classical regression estimator.
- $\hat{Y}_{\epsilon}^{(1)}(x_0)$ is a convex weighted mean (positive weights summing to 1), so that it cannot escape the range of observations $Y(x)$, whereas $\hat{Y}_{\epsilon}^{(2)}(x_0)$ can escape the range of observations $Y(x)$. There exists in the literature other nonparametric estimators using differently defined kernels that may lead to nonconvex linear combinations of the observations and thus extrapolations. This is the case for the frequently used estimators proposed and discussed in Jennen-Steinmetz and Gasser (1988) and Gasser and Mueller (1984). Allowing the predictions to escape the range of observations can be considered an advantage, because difficulty in prediction at the boundary of the auxiliary space (a.k.a. feature space) is a known drawback of many kernel-regression estimators.

5. Design-based *k*NN

As for kernel-based estimators, the literature on *k*NN is immense and largely model dependent. In the forest inventory context, Bafetta et al. (2009, 2011) are key references in the design-based context, whereas Tomppo et al. (2008), Moeur and Stage (1996), McRoberts et al. (2007), McRoberts et al. (2011), Magnussen and Tomppo (2014), McRoberts (2012), and Breidenbach and Nothdurft (2010) are key references in the model dependent context. Also, most authors work in the finite population framework of pixels in which the first phase is exhaustive and unequal inclusion probabilities on the on the plot level are possible. In the Monte Carlo approach, we must assume uniform i.i.d. sampling for all $x \in F$, because the notion of an inclusion probability of a given point x_0 is meaningless.

We shall use here the procedure described by Hechenbichler and Schliep (2004). For an arbitrary point $x_0 \in s_1$, the nearest neighbor in s_2, with respect to a distance $d(\cdot,\cdot)$ in $\mathbb{R}^p$, is the point $x_{(1)} \in s_2$ such that $d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(1)})) = \min_{x\in s_2} d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x))$. The second nearest neighbor $x_{(2)}$ is defined by $d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(2)})) = \min_{x\in s_2\setminus x_{(1)}} d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x))$ and so on until we have obtained the $k+1$ nearest neighbors. We emphasize the fact that when using an internal *k*NN model, the reference set is s_2. Thus, if $x_0 \in s_2$, then we set $x_{(1)} = x_0$ (i.e., x_0 counts as its own nearest neighbor). The simulations and case study presented in Section 6 contain only continuous variables so that distance ties among neighbors do not occur. The *k*NN methods with the uniform kernel simply take the mean, e.g., $\hat{Y}_{\text{knn}}(x_0) = (1/k)\sum_{i=1}^{k} Y(x_{(i)})$. In this particular case, it is possible, in principle, to calculate the theoretical design-based variance of the corresponding regression estimator, with an exhaustive first phase for finite populations (see Bafetta et al. (2009)). However, the uniform kernel is usually slightly less efficient than it is to give larger weights to the closer neighbors. As explanatory variables may have different scales, each auxiliary component should be standardized. Furthermore, we should standardize the distances themselves to avoid them becoming smaller as the sample size increases. Here, this is done according to

$$D(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)})) = \frac{d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)}))}{d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(k+1)}))} \tag{24}$$

and the weighted *k*NN estimator is given by

$$\hat{Y}_{\text{knn}}(x_0) = \frac{\sum_{i=1}^{k} K(D(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)})))Y(x_{(i)})}{\sum_{i=1}^{k} K(D(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)})))} \tag{25}$$

for an arbitrary one-dimensional kernel $K(\cdot)$ (note that the kernel is one dimensional, whereas the distance is calculated in the multidimensional feature space). This leads to the two-phase *k*NN regression estimator

$$\hat{Y}_{\text{knn}} = \left(\frac{1}{n_1}\right)\sum_{x\in s_1}\hat{Y}_{\text{knn}}(x) + \left(\frac{1}{n_2}\right)\sum_{x\in s_2}(Y(x) - \hat{Y}_{\text{knn}}(x)) \tag{26}$$

Note that the predictions, $\hat{Y}_{knn}(x)$, are a weighted mean of the observations and therefore constrained to be in the observed range of the observations. Thus *k*NN is also a purely interpolation technique and never extrapolates, which, as already mentioned, can be a disadvantage, especially when n_1 is large and n_2 is relatively small (the typical case).

As was the case for the other kernel-based regression estimators, one can use the one-dimensional metric based on the predictions alone, i.e., $d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)})) = |\hat{\boldsymbol{\beta}}^t(\boldsymbol{Z}(x_0) - \boldsymbol{Z}(x_{(i)}))|$. The resulting estimator is denoted by $\hat{Y}_{pred,knn}$ and should give us some insight into how much information is lost by using $\hat{Y}(x)$ to reduce the dimension of $\boldsymbol{Z}(x)$. Recall that for the other kernel-regression estimators, the dimension reduction was necessitated by the curse of dimensionality in conjunction with a multidimensional kernel. By construction for *k*NN, we only require a one-dimensional kernel based on distances to neighbors in the feature space. The external variance estimates are obtained as usual via eq. 4 where $R_m(x) = Y(x) - \hat{Y}_{knn}(x)$ and $R_m(x) = Y(x) - \hat{Y}_{pred,knn}(x)$ respectively.

In practice, one has to be careful about the tuning options available in software packages, in particular with respect to the definition of nearest neighbors when $x_0 \in s_2$. If the point itself is viewed as its nearest neighbor, the optimal choices $k = 1$ can result from cross-validation and should, of course, not be retained as it its leads to constant 0 residuals. We recommend at least $k \geq 3$ and to plot $\hat{Y}_{knn}$ and $\hat{\mathbb{V}}(\hat{Y}_{knn})$ as a function of *k*. In our experience, except for $\hat{Y}_{reg}$, the external variance estimate is likely too small and the bootstrap variance estimate should be preferred, as will be demonstrated in Section 6.

6. Examples

To illustrate the various techniques, we consider the real case study discussed in Mandallaz et al. (2013) and the artificial simulation example used in Mandallaz (2013). We have compared several kernel-based estimators. The Nadaraya–Watson kernel-based estimators (eqs. 19 and 20) were calculated using the npreg() function from the R package np (Hayfield and Racine (2008)), and the *k*NN estimators of Section 5 were calculated with kknn() from the R package kknn (Schliep and Hechenbichler (2014)). It should be noted that kknn() standardizes all feature space variables by default. For the purpose of comparison, estimators discussed by Jennen-Steinmetz and Gasser (1988) and Gasser and Mueller (1984) are also considered due to their popularity and general availability for a variety of statistical software. Thus, the widely used R functions glkern() and lokern() from the package lokern were used (for documentation, see Herrmann and Maechler (2014)). We shall not discuss here the function lokern(), because it was found that the local optimal bandwidth led to very spurious results near the boundaries, particularly at the lower end, and highly unstable bootstrap estimates (for details, see Mandallaz and Massey (2015)).

$\hat{Y}^{(1)}_{glkern}$ and $\hat{Y}^{(1)}_{npreg}$ were calculated using glkern() and npreg(), respectively, to calculate $Y(x)$ based on $\hat{Y}(x)$ while allowing the imbedded cross-validation procedures select the bandwidths. For $\hat{Y}^{(2)}_{glkern}$, glkern() was applied to the residual part only, and the optimal bandwidth was obtained via the imbedded cross-validation procedure. For $\hat{Y}^{(2)}_{npreg}$, the optimal bandwidth obtained via the imbedded leave-one-out cross-validation procedure and applied to only the residual part led essentially to a flat horizontal line close to 0 (i.e., making $\hat{Y}^{(2)}_{npreg} \approx \hat{Y}_{reg}$). For this reason, we wrote our own cross-validation procedure based on the mean squared error criteria by sequentially deleting each $x_0 \in s_2$ and refitting both $\hat{Y}(x_0)$ using the linear model, as well as the smoothed residual part (using npreg() on the second term) to get the optimal bandwidth out of a small set of feasible values. The resulting $\hat{Y}^{(2)}_{npreg}$ then differs from $\hat{Y}_{reg}$. The same procedure was used for the simulation presented in Section 6.2.

The bootstrap was implemented as explained in Section 2 in which each bootstrap sample was obtained using SRS with replacement over the s_1 from the original sample. The regression coefficients used to calculate $\hat{Y}(x)$ were recalculated for each bootstrap replicate. As for bandwidth selection, it was found that when the optimal bandwidth was recalculated in each bootstrap sample, the resulting bootstrap variance estimates were very unstable due to unacceptably extreme outliers, sometimes above 1000 times the mean (examples found in Mandallaz and Massey (2015)). This was also the case, but not as extreme, for the case study. As a result, to stabilize the bootstrap variances, the optimal global bandwidth obtained from the original full sample was used in all the bootstrap samples. Likewise, the optimal choices of *k* in the *k*NN estimators were only calculated from the original samples. This was not due to instability in the bootstrap variance estimates but rather because the imbedded cross-validation routine in kknn() allowed for $k = 1$ for a significant proportion of the replicates, which, as already discussed, leads to constant zero residuals in s_2. The double sampling bootstrap routine that fixes n_1 and n_2 as discussed in Wolter (2007) was also implemented for the case study, but because it seemed to produce bootstrap variance estimates that only differed negligibly and was somewhat more involved computationally to implement, it was dropped in favor of the simpler routine that only fixed n_1. A sample of the R code used for the simulation can be found in appendix E of Mandallaz and Massey (2015).

Table 1. Global estimation of timber volume ($m^3 \cdot ha^{-1}$) in the Canton of Grisons ($n_1 = 306$, $n_2 = 67$).

Estimator	Point estimate	External SE	Bootstrap SE
$\bar{Y}_F$	399.43	23.82	23.90
$\hat{Y}_{reg}$	384.95	16.52	18.34
$\hat{Y}^{(1)}_{glkern}$	389.89	16.08	19.36
$\hat{Y}^{(1)}_{npreg}$	388.58	16.08	17.79
$\hat{Y}^{(2)}_{glkern}$	389.79	16.08	19.41
$\hat{Y}^{(2)}_{npreg}$	387.51	16.28	18.55
$\hat{Y}_{knn}$	390.53	16.35	17.83
$\hat{Y}_{pred,knn}$	389.22	15.69	17.42

Note: SE, standard error. $\bar{Y}_F$, simple sample mean of plots in *F*. Bootstrap, 1000 replications.

6.1. Case study

The auxiliary vector $\boldsymbol{Z}(x)$ has seven components: $Z_0(x) \equiv 1$, mean canopy height $Z_1(x)$, maximal canopy height $Z_2(x)$, 75% quantile of the canopy height, $Z_3(x)$, standard deviation of canopy height $Z_4(x)$, the LiDAR-estimated volume density $Z_5(x)$ and the LiDAR-estimated density of stems $Z_6(x)$. The sample sizes were $n_1 = 306$ and $n_2 = 67$. The fitted model is

$$\hat{Y}(x) = 322.57 + 52.55Z_1(x) - 19.24Z_2(x) - 33.04Z_3(x) + 71.06Z_4(x) + 0.19Z_5(x) - 0.09Z_6(x)$$

and the coefficient of determination is $R^2 = 0.64$.

$\hat{Y}_{knn}$ is the *k*NN estimator using kknn() with the Euclidean distance applied to a six-dimensional feature space and a Gaussian kernel. We chose $k = 7$, which is the optimal value obtained from the leave-one-out cross-validation procedure (further justifications for this choice are given in Mandallaz and Massey (2015)). $\hat{Y}_{pred,knn}$ is defined likewise, except that its support is one dimensional, because it is based on the Euclidean distance between the linear predictions, $\hat{Y}(x)$. With the Gaussian kernel, it was decided to set $k = 9$ based on cross-validation.

Table 1 displays the results. The differences between the point estimates are not statistically significant from the sample mean, which is reassuring. However, the bootstrap standard errors are roughly 15% larger than the external standard errors, which is our first indication that the external variances could be underestimat-

ing their respective theoretical variances. However, for the case study, this is only conjecture, as the truth is not known. It should be noted that in a similar case study using the same data set (see Mandallaz et al. (2013)), the *g*-weight variance for $\hat{Y}_{reg}$ was found to be between the external and bootstrap standard errors calculated here, as intuition would suggest.

For completeness, we consider a small-area estimation problem. We have a partition of the entire domain F, in four small areas G_k, i.e., $F = \cup_{k=1}^{4} G_k$, with approximately equal surface areas $\lambda(G_k)$. Under the simple external-model approach, we apply all estimators to G_k exactly as we would to F except that the regression coefficients, predictions, and nearest neighbors are always based on the entire domain F. To apply a design-based small-area estimator in this way, one is implicitly assuming that there is adequate sample in the final phase of the small area (for examples in which n_2 as low as six was still adequate, see Mandallaz (2013)). For the bootstrap calculations, the resampling takes place in F.

Table 2 displays the results for a small area. All points estimates are close to each other and not significantly different. The external variance estimates and bootstrap variances of all estimators are comparable. There is no clear trend in underestimation of the bootstrap variance by the external variance as we saw in the global case, but it should be noted that the sample sizes are relatively low. One may obtain better bootstrap results for the small areas by using a modified balanced replication, as suggested in Magnussen et al. (2010). However, this was not the case here.

6.2. Simulations

In the case study, we observed some evidence that the external variance formula underestimates the theoretical design-based variance, because it seemed to systematically underestimate bootstrap variance, especially in the global case. However, the truth was not known, so it is difficult to determine the adequacy of the implemented bootstrap. Now we present a purely artificial example in which the truth is actually known, and we can illustrate the theory by empirically checking the validity external variance formula, as well as our implementation of the bootstrap.

We present the simulation example already used in Mandallaz (2013). The local density $Y(x)$ is defined according to the following procedure: at point $x = (x_1, x_2)^t \in \mathbb{R}^2$, the auxiliary vector is defined as $\boldsymbol{Z}(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^t \in \mathbb{R}^p$ ($p = 6$ in this example). The true parameter is $\boldsymbol{\beta}_0 = (30, 13, -6, -4, 3, 2)^t \in \mathbb{R}^6$, and the local density over the domain $F = [0, 2] \times [0, 3]$ is given by the function

$$Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta}_0 + 6\cos(\pi x_1)\sin(2\pi x_2) \coloneqq \hat{Y}_0(x) + R(x) \tag{27}$$

As we have an analytical representation of $Y(x)$ for the entire domain, F, we can iteratively draw samples according to our two-phase sampling design while calculating estimates, external variance estimates, and bootstrap estimates for any arbitrary estimator. The empirical variance of the point estimates across all performed iterations should offer excellent insight to the theoretical design-based variance of an estimator. Likewise, the empirical mean allows us to observe the design-based expectation. The mean of the variance estimates, either external or bootstrap, across all iterations tells us the expectation of the respective variance estimator, which we can then compare with the empirical variance.

The two sample size scenarios performed were $n_1 = 400$ and $n_2 = 100$, as well as $n_1 = 200$ and $n_2 = 50$. For each of these scenarios, two model choices were applied: the true model based on x_1, x_2, x_1^2, x_1x_2, and x_2^2, and a working model based on the subset of the true model, using only x_2, x_1^2, and x_2^2. The number of neighbors for $\hat{Y}_{knn}$ was fixed at $k = 3$ (details about this decision are given in Mandallaz and Massey (2015)). For the estimator $\hat{Y}_{pred,knn}$, the optimal values of k were obtained from restricted simulations with 500 runs for each sample size and model scenario. The median of these optimal k's obtained via cross-validation were used as the fixed k for each of the scenarios.

Table 2. Small-area estimation of timber volume ($m^3 \cdot ha^{-1}$) in the Canton of Grisons.

	Small areas, n_1:n_2			
Estimator	G_1, 94:19	G_2, 81:17	G_3, 66:15	G_4, 65:16
$\bar{Y}_G$	410.40	461.44	318.00	396.85
	(44.58)	(56.35)	(34.36)	(47.86)
	[44.17]	[55.37]	[33.93]	[47.88]
$\hat{Y}_{reg}$	397.27	426.81	327.64	366.44
	(28.80)	(35.01)	(31.64)	(36.01)
	[31.61]	[34.93]	[32.06]	[35.53]
$\hat{Y}^{(1)}_{glkern}$	405.60	427.16	340.50	364.63
	(28.82)	(34.63)	(29.65)	(34.65)
	[34.40]	[34.30]	[30.80]	[34.76]
$\hat{Y}^{(1)}_{npreg}$	397.36	428.37	341.08	369.36
	(28.65)	(35.11)	(29.17)	(34.76)
	[30.72]	[33.27]	[29.21]	[33.67]
$\hat{Y}^{(2)}_{glkern}$	405.46	426.94	340.58	364.59
	(28.81)	(34.62)	(29.67)	(34.64)
	[34.42]	[34.25]	[30.86]	[34.70]
$\hat{Y}^{(2)}_{npreg}$	400.43	430.39	334.59	363.70
	(29.61)	(34.73)	(30.09)	(35.01)
	[32.72]	[34.23]	[30.91]	[34.98]
$\hat{Y}_{knn}$	382.46	464.14	319.42	376.96
	(27.07)	(36.64)	(32.59)	(32.14)
	[28.28]	[34.77]	[28.04]	[30.02]
$\hat{Y}_{pred,knn}$	398.60	424.02	353.21	364.48
	(28.87)	(34.09)	(29.96)	(32.14)
	[29.36]	[32.42]	[28.41]	[30.62]

Note: The standard error based on the external variance estimator (i.e., eq. 4) are given in parentheses, and the standard errors based on the bootstrap variance estimates (1000 replications) are given in brackets. $\bar{Y}_G$ is the simple sample mean of plots in G_k, $k = 1, 2, 3, 4$.

Table 3. $n_1 = 400$, $n_2 = 100$. The fitted model is the true model.

Estimator	$\mathbb{E}^*()$	$\mathbb{V}^*()$	$\mathbb{E}^*(\hat{\mathbb{V}}())$	$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$	$\mathbb{E}^*(\hat{P})$	$\mathbb{E}^*(\hat{P}_{boot})$
$\hat{Y}_{reg}$	39.17	0.19	0.19	0.20	0.95	0.95
$\hat{Y}^{(1)}_{glkern}$	39.15	0.20	0.19	0.20	0.94	0.95
$\hat{Y}^{(1)}_{npreg}$	39.15	0.20	0.18	0.20	0.94	0.94
$\hat{Y}^{(2)}_{glkern}$	39.15	0.20	0.19	0.20	0.94	0.95
$\hat{Y}^{(2)}_{npreg}$	39.17	0.20	0.18	0.20	0.94	0.95
$\bar{Y}_{pred,knn,k=11}$	39.16	0.20	0.18	0.20	0.93	0.94
$\bar{Y}_{knn,k=3}$	39.15	0.18	0.14	0.19	0.92	0.95

Note: The true mean value is $\bar{Y} = 39.17$, and the true coefficient of determination is $R^2 = 0.83$. $\mathbb{E}^*(\hat{P})$, the empirical coverage probability of the 95% confidence interval based on the estimated variance and the normal approximation; $\mathbb{E}^*(\hat{P}_{boot})$, the empirical coverage probability based on the bootstrap confidence intervals; $\mathbb{E}^*()$ and $\mathbb{V}^*()$, the empirical mean and variances over 10 000 runs, respectively; $\mathbb{E}^*(\hat{\mathbb{V}}())$ is the empirical mean of the external variance formula; $\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$ is the empirical mean of the bootstrap variances obtained from 1000 bootstrap replicates for each simulation run.

Tables 3 and 4 give the results for global estimation when the fitted model is indeed the true model, and Tables 5 and 6 give the results for global estimation when the fitted model is the working model. All simulations are based on 10 000 runs. The 95% bootstrap confidence intervals are obtained via the well-known formula $[2\hat{Y} - q^*_{0.975}, 2\hat{Y} - q^*_{0.025}]$, where q^*_α is the α bootstrap quantile of the bootstrap estimate $\hat{Y}^*$, defined as the mean of the point estimates of the 1000 bootstrap samples obtained in each of the 10 000 runs (see Davison and Hinkley (1997)).

Table 4. $n_1 = 200$, $n_2 = 50$. The fitted model is the true model.

Estimator	$\mathbb{E}^*()$	$\mathbb{V}^*()$	$\mathbb{E}^*(\hat{\mathbb{V}}())$	$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$	$\mathbb{E}^*(\hat{P})$	$\mathbb{E}^*(\hat{P}_{boot})$
$\hat{Y}_{reg}$	39.18	0.41	0.38	0.42	0.95	0.95
$\hat{Y}^{(1)}_{glkern}$	39.17	0.42	0.36	0.45	0.92	0.95
$\hat{Y}^{(1)}_{npreg}$	39.15	0.42	0.35	0.43	0.92	0.94
$\hat{Y}^{(2)}_{glkern}$	39.17	0.42	0.36	0.45	0.92	0.95
$\hat{Y}^{(2)}_{npreg}$	39.17	0.43	0.36	0.42	0.93	0.94
$\bar{Y}_{pred,knn,k=6}$	39.15	0.43	0.35	0.43	0.92	0.94
$\bar{Y}_{knn,k=3}$	39.15	0.43	0.31	0.43	0.91	0.94

Note: Please see the Note under Table 3 for variable definitions.

Table 5. $n_1 = 400$, $n_2 = 100$. The fitted model is not the true model.

Estimator	$\mathbb{E}^*()$	$\mathbb{V}^*()$	$\mathbb{E}^*(\hat{\mathbb{V}}())$	$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$	$\mathbb{E}^*(\hat{P})$	$\mathbb{E}^*(\hat{P}_{boot})$
$\hat{Y}_{reg}$	39.18	0.28	0.26	0.27	0.95	0.95
$\hat{Y}^{(1)}_{glkern}$	39.19	0.27	0.25	0.28	0.93	0.94
$\hat{Y}^{(1)}_{npreg}$	39.18	0.28	0.24	0.27	0.92	0.94
$\hat{Y}^{(2)}_{glkern}$	39.19	0.27	0.25	0.28	0.93	0.94
$\hat{Y}^{(2)}_{npreg}$	39.18	0.27	0.25	0.27	0.93	0.94
$\bar{Y}_{pred,knn,k=15}$	39.17	0.28	0.24	0.27	0.93	0.94
$\bar{Y}_{knn,k=3}$	39.12	0.21	0.12	0.22	0.91	0.96

Note: The columns headers are the same as in Tables 3 and 4. The true mean value is the same, i.e. $\bar{Y} = 39.17$. The fitted model did not include the explanatory variables x_1 and x_1x_2. The true coefficient of determination decreases from $R^2 = 0.83$ to $R^2 = 0.66$.

Table 6. $n_1 = 200$, $n_2 = 50$. The fitted model is not the true model.

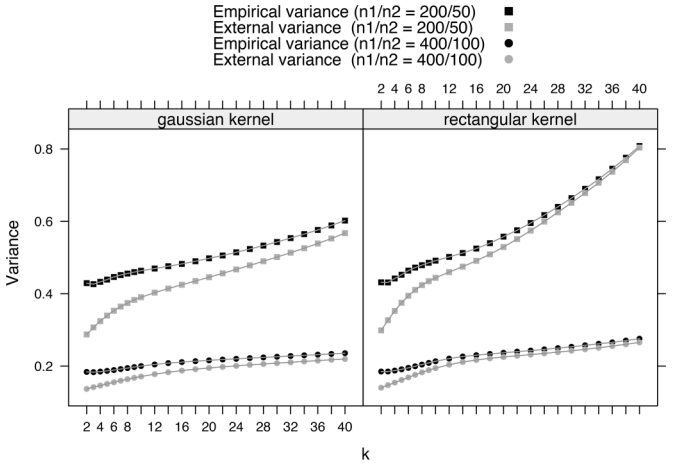
Estimator	$\mathbb{E}^*()$	$\mathbb{V}^*()$	$\mathbb{E}^*(\hat{\mathbb{V}}())$	$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$	$\mathbb{E}^*(\hat{P})$	$\mathbb{E}^*(\hat{P}_{boot})$
$\hat{Y}_{reg}$	39.21	0.56	0.50	0.56	0.94	0.95
$\hat{Y}^{(1)}_{glkern}$	39.21	0.57	0.48	0.62	0.92	0.94
$\hat{Y}^{(1)}_{npreg}$	39.18	0.57	0.46	0.56	0.92	0.93
$\hat{Y}^{(2)}_{glkern}$	39.21	0.57	0.48	0.62	0.92	0.94
$\hat{Y}^{(2)}_{npreg}$	39.19	0.57	0.48	0.56	0.92	0.94
$\bar{Y}_{pred,knn,k=9}$	39.17	0.56	0.46	0.56	0.92	0.94
$\bar{Y}_{knn,k=3}$	39.12	0.50	0.32	0.51	0.89	0.94

Note: Please see the Note under Table 5.

Discussion

1. All estimators are practically design unbiased. The empirical variances, which closely approximate the normally unobservable theoretical variance, are smaller for the true model than for the working model, as expected. The external variances always underestimate the empirical variance, as mathematical intuition and the case study suggest. The underestimation appears to be much more pronounced for the nonparametric estimators than for $\hat{Y}_{reg}$. For the *k*NN estimator, the underestimation occurs regardless of the choice of the kernel or k (see Fig. 1).
2. Although these results were not presented in Tables 3, 4, 5, or 6, the estimators were also calculated using a true external approach in which the models were fitted using an independent sample selection based on n_2 observations from eq. 27. As mathematically expected, the empirical variances closely matched the empirical means of the external variance estimates in all scenarios, and the coverage probabilities were equal to 95%.
3. There is strong evidence that the implemented bootstrap procedure was adequate. The empirical mean of the bootstrap variance estimates are close to their associated empirical variances. Furthermore, the empirical coverage probabilities of the bootstrap confidence intervals were all close to the nominal 95% probability. With respect to the nominal 95% confidence intervals, the classical regression estimator $\hat{Y}_{reg}$ performs well in all investigated cases. For all other estimators, one must use the bootstrap confidence intervals, because their external variances clearly underestimates the empirical variances. The underestimation was less but still present in the scenario with higher sample sizes.
4. The better performance of the classical regression estimator $\hat{Y}_{reg}$ is probably due to the fact that the coefficients of determination are rather high under the true model and under the working models. On the whole, the kernel-based or *k*NN estimators did not perform better than than the classical regression estimator $\hat{Y}_{reg}$, with the exception being the nearest neighbor estimator $\hat{Y}_{knn}$ with the choice $k = 3$. This estimator yielded a slightly smaller empirical variance when the working model was not the true model, and its variance could be correctly estimated with the bootstrap. This may suggest that $\hat{Y}_{knn}$ is better as the goodness of fit decreases. That being said, the R^2 of the working model was 0.66, which is not outstanding in the context of forest inventories, e.g., with timber volume as response variable and with LiDAR canopy measurements as explanatory variables.
5. A further theoretical advantage of $\hat{Y}_{reg}$ is that it allows for correct analytical expressions of the asymptotic variances, which currently seems intractable to derive for the other estimators. Analytical variance estimates are exactly reproducible given the same data set.

Fig. 1. Influence of choice of k for two different kernels using *k*NN estimator with multidimensional support based on the complete set of explanatory variables from the true model.



Empirical and external variance estimates based on 10000 simulations

7. Conclusions

Intuitively, treating internally fit regression estimators as external via the application of the external variance formula can lead to systematic underestimation of the true design-based variance because one implicitly ignores the influence of the random sample selection on the realization of the model prediction. This intuition was empirically confirmed for kernel-regression estimators by simulation results that demonstrated that the observed coverage rates for kernel-based estimators fell consistently below the nominal coverage probability. The underestimation can be corrected by bootstrapping, but special attention should be given to the choice of the tuning options within the resampling procedures, because they can lead to unstable variance estimates. An acceptable way of addressing this instability is to select the optimal bandwidth on the original sample and then fix the bandwidth to that value across all bootstrap replicates. The external variance formula for the classical two-phase regression estimator, $\hat{Y}_{reg}$, on the other hand, remained acceptable even for internal models, as the underestimation appears to be asymptotically negligible.

These preliminary comparisons between the various estimators show that the classical two-phase regression estimator with the

external variance estimates performs on the whole at least as well as the kernel-based regression estimators with bootstrap variance. The only exception was $\hat{Y}_{knn}$, the *k*NN estimator with distance defined in the multidimensional feature space. $\hat{Y}_{knn}$ appears to do better when the goodness of fit of the auxiliary variables is low, but the variance estimate should be bootstrapped to avoid potentially dramatic underestimation. It should be noted that different versions of $\hat{Y}_{reg}$ are available for more complex situations such as three-phase cluster sampling with two-types of auxiliary information and with two-stage sampling of trees at the plot level, as well as its extensions to small-area estimation (Mandallaz 2013, 2014; Mandallaz et al. 2013), where it is not yet very clear how to proceed using the kernel-based approach. It can reasonably be expected that these conclusions will hold in general if the model is adequate, i.e., if it incorporates the most important explanatory variables (say with an $R^2 \geq 0.6$), which is certainly the case for timber volume with the standard LiDAR explanatory variables (mean canopy height being the most important).

In the context of forest inventories, the main advantage of $\hat{Y}_{reg}$ is its flexibility and the existence of an analytical design-based variance estimator. Being constructed using a standard linear model, it is relatively easy to implement and to obtain highly reproducible results that are independent of the choice of kernels, metrics, and further tuning options. $\hat{Y}_{knn}$ may perform a bit better in situations where the auxiliary information leads to a decreased goodness of fit, but the variance should be obtained via bootstrap, which leads to more potential arrangements of tuning parameters. Practically speaking, this means that exact reproduction of estimates, even given identical data sets, becomes more problematic and less likely to occur.

References

Bafetta, F., Fattorini, L., Franceschi, S., and Corona, P. 2009. Design-based approach to *k*-nearest neighbours technique for coupling field and remote sensed data in forest inventory. Remote Sens. Environ. **113**: 463–479. doi:10.1016/j.rse.2008.06.014.

Bafetta, F., Corona, P., and Fatottorini, L. 2011. Design-based diagnostics for *k*-NN estimators of forest resources. Can. J. For. Res. **41**(1): 59–72. doi:10.1139/X10-157.

Breidenbach, J., and Nothdurft, A. 2010. Comparison of nearest neighbours approaches for small area estimation of tree species specific forest inventory attributes in central Europe using airborne laser scanner data. Eur. J. For. Res. **129**: 833–846. doi:10.1007/s10342-010-0384-1.

Davison, A., and Hinkley, D. 1997. Bootstrap methods and their application. Cambridge Series in Statistical and Probabilistic Mathematics.

Gasser, T., and Mueller, H. 1984. Estimating regression functions and their derivatives by the kernel method. Scand. J. Stat. **11**: 171–185.

Gregoire, T. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. **28**(10): 1429–1447. doi:10.1139/x98-166.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. 2002. A distribution-free theory of nonparametric regression. Springer, New York.

Hayfield, T., and Racine, J.S. 2008. Nonparametric econometrics: the np package. Journal of Statistical Software, **27**(5).

Hechenbichler, K., and Schliep, K. 2004. Weighted *k*-nearest neighbor techniques and ordinal classification. Technical Report, University of Munich and Massey University.

Herrmann, E., and Maechler, M. 2014. lokern: kernel regression smoothing with local or global plug-in bandwidth. R package, version 1.1-6.

Jennen-Steinmetz, C., and Gasser, T. 1988. A unifying approach to nonparametric regression estimation. J. Am. Stat. Assoc. **83**: 1084–1089. doi:10.1080/01621459.1988.10478705.

Lehmann, E. 1999. Elements of large-sample theory. Springer Texts in Statistics, New York.

Magnussen, S., and Tomppo, E. 2014. The *k*-nearest neighbor technique with local linear regression. Scand. J. For. Res. **29**: 120–131. doi:10.1080/02827581.2013.878744.

Magnussen, S., McRoberts, R.E., and Tompo, E. 2010. A resampling variance estimator for the *k* nearest neighbours technique. Can. J. For. Res. **40**(4): 648–658. doi:10.1139/X10-020.

Mandallaz, D. 2008. Sampling techniques for forest inventories. Chapman and Hall, Boca Raton Florida.

Mandallaz, D. 2013. Design-based properties of some small-area estimators in forest inventory with two-phase sampling. Can. J. For. Res. **43**(5): 441–449. doi:10.1139/cjfr-2012-0381.

Mandallaz, D. 2014. A three-phase sampling extension of the generalized regression estimator with partially exhaustive information. Can. J. For. Res. **44**(4): 383–388. doi:10.1139/cjfr-2013-0449.

Mandallaz, D., and Massey, A. 2015. Regression and non-parametric estimators for two-phase forest inventories in the design-based Monte-Carlo approach. Technical Report, ETH Zurich, Department of Environmental Systems Science. Available from http://e-collection.library.ethz.ch.

Mandallaz, D., Breschan, J., and Hill, A. 2013. New regression estimators in forest inventory with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation. Can. J. For. Res. **43**(11): 1023–1031. doi:10.1139/cjfr-2013-0181.

Massey, A., Mandallaz, D., and Lanz, A. 2014. Integrating remote sensing and past inventory data under the new annual design of the Swiss National Forest Inventory using three-phase design-based regression estimation. Can. J. For. Res. **44**(10): 1177–1186. doi:10.1139/cjfr-2014-0152.

McRoberts, R.E. 2012. Estimating forest attributes parameters for small areas using nearest neighbors techniques. For. Ecol. Manage. **272**: 3–12. doi:10.1016/j.foreco.2011.06.039.

McRoberts, R., Tomppo, E.O., Finley, A.O., and Heikkinen, J. 2007. Estimating areal means and variance of forest attributes using the *k*-nearest neighbors technique and satellite imagery. Remote Sens. Environ. **111**: 466–480. doi:10.1016/j.rse.2007.04.002.

McRoberts, R., Magnussen, S., Tomppo, E.O., and Chirici, G. 2011. Parametric bootstrap and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. Remote Sens. Environ. **115**: 3165–3174. doi:10.1016/j.rse.2011.07.002.

Moeur, M., and Stage, A. 1996. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For. Sci. **41**: 337–359.

Särndal, C., Swenson, B., and Wretman, J. 2003. Model assisted survey sampling. Springer Series in Statistics, New York.

Schliep, K., and Hechenbichler, K. 2014. kknn: weighted *k*-nearest neighbors. R package version 1.2-5.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., and Katila, M. 2008. Combining national forest inventory field plots and remote sensing data for forest databases. Remote Sens. Environ. **112**: 1982–1999. doi:10.1016/j.rse.2007.03.032.

Wolter, K. 2007. Introduction to variance Estimation. Second edition. Springer-Verlag, New York.

Wu, C., and Sitter, R. 2001. A model calibration approach to using complete auxiliary information from survey data. J. Am. Stat. Assoc. **96**: 185–193. doi:10.1198/016214501750333054.