Model-based nonparametric variance estimation for systematic sampling *

J.D. Opsomer
Colorado State University†

M. Francisco-Fernández
Universidad de A Coruña‡

X. Li
Pfizer, Inc.§

January 3, 2010

Abstract

Systematic sampling is frequently used in surveys, because of its ease of implementation and its design efficiency. An important drawback of systematic sampling, however, is that no direct estimator of the design variance is available. We describe a new estimator of the model-based expectation of the design variance, under a nonparametric model for the population. The nonparametric model is sufficiently flexible that it can be expected to hold at least approximately for many practical situations. We prove the model consistency of the estimator for both the anticipated variance and the design variance. We compare the nonparametric variance estimators with several design-based estimators through a simulation study and on data from a forestry survey.

Key Words: local polynomial regression, two-per-stratum variance approximation, smoothing.

*Short title: Nonparametric variance estimation.

†Department of Statistics, Fort Collins, CO 80523, USA. Email: jopsomer@stat.colostate.edu.
‡Departamento de Matemáticas, Facultad de Informática, La Coruña,15071, Spain.
§Pfizer Global Research & Development, New London, CT 06320, USA.

1 Introduction

Systematic sampling and its variants such as fractional systematic sampling are commonly used sampling designs in finite population surveys. These designs are easy to implement and result in estimators that are highly efficient for any survey variables that are correlated with the variable(s) used to sort the population prior to sampling. A well-known and long-standing issue in surveys that follow a systematic sampling design is the lack of a theoretically justified, generally applicable design-based variance estimator. Because of this, there have been attempts to adjust the sampling design itself to allow for variance estimation, including for example the drawing of multiple systematic samples (Törnqvist, 1963) and the drawing of a partial systematic sample supplemented with a simple random sample (Zinger, 1980). This type of solutions are only rarely used, however, and the majority of applications continue to apply "pure" systematic sampling combined with a not fully satisfactory variance estimator.

A whole chapter of the recently reissued classic monograph by Wolter (2007) is devoted to this issue, and a number of possible estimation approaches are evaluated there. In particular, it considers a set of eight "model-free" estimators, some of which we will discuss further below, and outlines a model-based estimation approach. For the set of eight estimators, their statistical properties are evaluated for several model scenarios and through simulation experiments. None of these estimators is best overall, and there is a clear interaction between the behavior of the estimators and the underlying data model. Despite this implicit model dependence, the two estimators based on averages of pairwise differences (see next section) are found to be the best compromise between good performance and general applicability among this set of eight estimators. They are also widely used in practice.

In the model-based estimation approach described in Wolter (2007), the model dependence is explicitly recognized and a Rao-Blackwell type estimator is proposed, which minimizes the model mean squared error in estimating the sampling variance. The models considered in Wolter (2007) are parametric, and the Rao-Blackwell estimator therefore

depends on unknown parameters that must be estimated from the sample data. An example of this approach is Montanari and Bartolucci (1998), who proposed an unbiased model-based variance estimator when the population follows a linear regression model.

In practice, despite its potential efficiency, wide applicability of the model-based method is viewed as being hampered by lack of robustness. Wolter (2007, p.305) noted that:

> "Since [the model] is never known exactly, the practicing statistician must make a professional judgment about the form of the model and then derive [the variance estimator] based on the chosen form. The 'practical' variance estimator [with estimated parameters] is then subject not only to errors of estimation [...] but also to errors of model misspecification."

However, this lack of robustness can be at least partly offset by the use of a nonparametric model specification. Compared to parametric models, this class of models makes much less restrictive assumptions on the shape of the relationship between variables, typically only requiring that the relationship be continuous and smooth, i.e. possessing a prespecified number of derivatives. Hence, the risk of model misspecification is significantly reduced. This is particularly important in the survey context, because the same variance estimation method often needs to be applied to many survey variables collected in the same survey, and a single parametric model is much less likely to be correct for all these variables. Bartolucci and Montanari (2006) discussed the use of nonparametric estimation as a way to "robustify" the model-based approach. They evaluate the bias properties of the resulting estimator under the linear population model, and then consider the behavior under nonlinear population models through simulation. The latter results suggest that the nonparametric approach remains effective in estimating the variance under a range of population model specifications.

In the current article, we will consider a broadly applicable model for the data, in which both the mean and the variance are left unspecified subject only to smoothness assumptions. We propose a model-based nonparametric variance estimator, in which

both the mean and the variance functions of the data are estimated nonparametrically. The smoothing method we will use is local polynomial regression (see Wand and Jones (1995) for an overview). We will show that the proposed estimator is model consistent for the design variance of the survey estimator, subject only to the population smoothness assumptions. The article will focus on the case of estimating the finite population mean using the sample mean for a systematic sample, but there is no inherent difficulty in extending the method to estimate the (approximate) variance of more complicated estimators such as model-assisted estimators.

The rest of the article is organized as follows. In Section 2, we describe the systematic sampling estimation context and the main variance estimators in use today. In Section 3, we introduce the nonparametric variance estimator and describe its statistical properties. Section 4 evaluates the practical properties of the estimator in a simulation study. Section 5 illustrates the applicability of the methodology on a real forestry dataset.

2 Systematic sampling and design-based variance estimation

We will be sampling from a finite population U of size N. For now, we consider a single study variable $Y_j \in \mathbb{R}$, $j = 1, 2, \cdots, N$ with population mean

$$\bar{Y}_N = \frac{1}{N} \sum_{j=1}^{N} Y_j.$$

Let n denote the sample size and $k = N/n$ denote the *sampling interval*. For simplicity, we assume throughout this article that N is an integral multiple of n, i.e. k is an integer. The variable Y will only be observed on the sampled elements only.

Let $\mathbf{x}_j \in \mathbb{R}^p$ $(j = 1, 2, \cdots, N)$ be vectors of auxiliary variables available for all the elements in the population. To draw a systematic sample, the population is first sorted by some appropriate criterion. For example, we can sort by one or several of the auxiliary variables in $\mathbf{x}_j$. If the study variable Y and auxiliary variables $\mathbf{x}$ are related to each other,

sorting by $\mathbf{x}$ and then drawing a systematic sample has been long known to reduce the variance of the sample mean. Conversely, if the population is sorted by a criterion that is not related to Y, for instance, by a variable in $\mathbf{x}$ which is independent of Y, then we will have a random permutation of the population. In this case, systematic sampling is equivalent to simple random sampling without replacement. After sorting the population, drawing a systematic sample is done by randomly choosing an element among the first k with equal probability, say the bth one, after which the systematic sample, denoted by S_b, consists of the observations with labels $\{b,\ b+k,\ \ldots,\ b+(n-1)k\}$. The random sample S can therefore only take on k values on the set of possible samples $\{S_1, \ldots, S_k\}$.

The sample mean,

$$\bar{Y}_S = \frac{1}{n}\sum_{j\in S} Y_j,$$

is the Horvitz-Thompson estimator for the finite population mean. Its design-based variance was first derived by Madow and Madow (1944) and is equal to

$$\mathrm{Var}_p(\bar{Y}_S) = \frac{1}{k}\sum_{b=1}^{k}(\bar{Y}_{S_b} - \bar{Y}_N)^2. \tag{1}$$

It should be clear that, if only a single systematic sample is drawn and hence only one of the $\bar{Y}_{S_b}$ is observed, no unbiased design-based estimator of $\mathrm{Var}_p(\bar{Y}_S)$ exists for general variable Y. A more formal way to state this is that the systematic sampling design is *not measurable* (Särndal et al. 1992, p.33).

We describe the three main methods used in practice to estimate $\mathrm{Var}_p(\bar{Y}_S)$, all of which are part of the eight estimators evaluated by Wolter (2007) and mentioned in Section 1. The simplest estimator is to treat the systematic sample as if it had been obtained by simple random sampling. This estimator is defined as

$$\hat{V}_{SRS} = \frac{1-f}{n}\frac{1}{n-1}\sum_{j\in S}(Y_j - \bar{Y}_S)^2, \tag{2}$$

where $f = n/N$. The two remaining estimators are based on pairwise differences and are recommended in Wolter (2007) as being the best general-purpose estimators of $\mathrm{Var}_p(\bar{Y}_S)$).

They are defined as

$$\hat{V}_{OL} = \frac{1-f}{n}\frac{1}{2(n-1)}\sum_{j=2}^{n}(Y_j - Y_{j-1})^2, \tag{3}$$

which uses all successive pairwise differences (and hence uses overlapping differences, OL), and

$$\hat{V}_{NO} = \frac{1-f}{n}\frac{1}{n}\sum_{j=1}^{n/2}(Y_{2j} - Y_{2j-1})^2. \tag{4}$$

which takes successive non-overlapping differences (NO). Additional estimators based on higher-order differences are described in Wolter (2007) but will not be further considered here.

All three estimators just described are design biased for $\mathrm{Var}_p(\bar{Y}_S)$ in general. The estimator $\hat{V}_{SRS}$ is viewed as suitable when the ordering of the population is thought to have no effect on $\bar{Y}_S$, or is considered as a conservative estimator when the ordering is related to the variable Y. However, as discussed in Opsomer et al. (2007), the unbiasedness of $\hat{V}_{SRS}$ for uninformative ordering only holds if one averages over samples *and* over orderings of the population (see Cochran, 1977, Thm 8.5), so this is not, strictly speaking, design unbiasedness. The design bias of $\hat{V}_{SRS}$ for a fixed ordering of the population can be large and either positive or negative, so that relying on its conservativeness can be potentially misleading. This was clearly seen in the synthetic estimator approach used in Opsomer et al. (2007), for instance. The remaining two estimators tended to have smaller bias in the simulation experiments reported in Wolter (2007), but their formal statistical properties as estimators of $\mathrm{Var}_p(\bar{Y}_S)$ are not generally available.

3 Variance estimation under a nonparametric model

In the model-based context, the finite population is regarded as a random realization from a superpopulation model. A simple approach consists of assuming a parametric model for this superpopulation model. Under the assumption of linearity for the model, Bartolucci and Montanari (2006) proposed a model unbiased estimator for the *anticipated*

variance $\mathrm{E}[\mathrm{Var}_p(\bar{Y}_S)]$, using a least squares estimator for the regression parameters and a model unbiased estimator for the variance of the errors. In this section, we propose a model consistent variance estimator under a nonparametric model, using local polynomial regression as the estimation method. For simplicity, we will describe the nonparametric method and obtain its statistical properties for a univariate auxiliary variable x. The extension of the approach to higher dimensions is straightforward, except for the fact that the curse of dimensionality will make local polynomial regression less suitable for dimensions above two or three, and hence should be replaced by a more appropriate method such as additive modeling. We do not explore this further here.

The nonparametric superpopulation model considered here is

$$Y_j = m(x_j) + v(x_j)^{1/2}\, e_j \quad 1 \le j \le N, \tag{5}$$

where $m(\cdot)$ and $v(\cdot)$ are continuous and bounded functions. The errors e_j, $1 \le j \le N$, are independent random variables with model mean 0 and variance 1. Define $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)^T$, $\mathbf{m} = (m(x_1), \ldots, m(x_N))^T$ and $\mathbf{\Sigma} = \mathrm{diag}\{v(x_1), v(x_2), \cdots, v(x_N)\}$.

The design variance of $\bar{Y}_S$ can be written as

$$\mathrm{Var}_p(\bar{Y}_S) = \frac{1}{k}\sum_{b=1}^{k}(\bar{Y}_{S_b} - \bar{Y}_N)^2 = \frac{1}{kn^2}\mathbf{Y}^T\mathbf{D}\mathbf{Y}, \tag{6}$$

where $\mathbf{D} = \mathbf{M}^T\mathbf{H}\mathbf{M}$, with $\mathbf{M} = \mathbf{1}_n^T \otimes \mathbf{I}_k$ and $\mathbf{H} = \mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T$, with $\otimes$ denoting Kronecker product and $\mathbf{1}_r$ a vector of 1's of length r. Stated more explicitly, $\mathbf{H}$ is a $k \times k$ matrix with diagonal elements being $1 - \frac{1}{k}$ and off-diagonal element being $-\frac{1}{k}$, and $\mathbf{D}$ is a $N \times N$ matrix composed of $n \times n$ $\mathbf{H}$s. Then, the model anticipated variance of $\bar{Y}_S$ under model (5) is

$$\mathrm{E}[\mathrm{Var}_p(\bar{Y}_S)] = \frac{1}{kn^2}\mathbf{m}^T\mathbf{D}\mathbf{m} + \frac{1}{kn^2}\mathrm{tr}(\mathbf{D}\mathbf{\Sigma}). \tag{7}$$

To estimate $\mathrm{E}[\mathrm{Var}_p(\bar{Y}_S)]$, we propose the following estimator

$$\hat{V}_{NP} = \frac{1}{kn^2}(\hat{\mathbf{m}}^T\mathbf{D}\hat{\mathbf{m}}) + \frac{1}{kn^2}\mathrm{tr}(\mathbf{D}\hat{\mathbf{\Sigma}}), \tag{8}$$

where $\hat{\mathbf{m}} = (\hat{m}(x_1), \cdots, \hat{m}(x_N))^T$, with $\hat{m}(x_j)$ the local polynomial regression (LPR) estimator of $m(x_j)$ computed on the observations in the sample S and $\hat{\mathbf{\Sigma}} =$

diag$\{\hat{v}(x_1), \hat{v}(x_2), \cdots, \hat{v}(x_N)\}$, with $\hat{v}(x_j)$ the LPR estimator of $v(x_j)$. We briefly describe the two LPR estimators, and for simplicity we will assume that the degree of the two local polynomials is equal to p. Note that there is no restriction that the x_j should or should not be related to the sorting variable used to draw the systematic sample.

For the estimator of the mean function,

$$\hat{m}(x_j) = \mathbf{e}_1^T (\mathbf{X}_{Sj}^T \mathbf{W}_{Sj} \mathbf{X}_{Sj})^{-1} \mathbf{X}_{Sj}^T \mathbf{W}_{Sj} \mathbf{Y}_S,$$

with $\mathbf{e}_1$ a vector of length $(p+1)$ having 1 in the first entry and all other entries 0, $\mathbf{Y}_S$ a vector containing the $Y_j \in S$, $\mathbf{X}_{Sj}$ a matrix with ith row equal to $(1, (x_i - x_j), \ldots, (x_i - x_j)^p)$, $i \in S$, and

$$\mathbf{W}_{Sj} = \text{diag}\left\{ K\left(\frac{x_i - x_j}{h_m}\right), \quad i \in S \right\},$$

where h_m is the bandwidth and K is a kernel function. For the estimator of the variance function, the expression is completely analogous, except that $\mathbf{Y}_S$ is replaced by the vector of squared residuals $\hat{\mathbf{r}}_S$ with elements $\hat{r}_j = (Y_j - \hat{m}(x_j))^2, j \in S$, and a different bandwidth h_v is used instead of h_m in the weight matrix $\mathbf{W}_{Sj}$. This variance estimator was previously used in Fan and Yao (1998) in a different context and does not include a "degrees of freedom" correction term as in Ruppert et al. (1997). While the latter estimator can certainly be used here, we found little difference between both in this setting, so that we chose the simpler estimator.

Under suitable regularity conditions on the population and the nonparametric estimator, which are stated in the Appendix, we obtain the following results on the asymptotic properties of $\hat{V}_{NP}$. An outline of the proof is given in the Appendix. The theorem shows that $\hat{V}_{NP}$ is a model consistent estimator for $\mathrm{E}[\mathrm{Var}_p(\bar{Y}_S)]$ and a model consistent predictor for $\mathrm{Var}_p(\bar{Y}_S)$.

Theorem 3.1 *Assume that the degree p of the local polynomials is odd. Using superpopulation model* (5) *and under assumptions A.1–A.6 in the Appendix, the design variance is model consistent for the anticipated variance, in the sense that*

$$Var_p(\bar{Y}_S) - E[\,Var_p(\bar{Y}_S)] = O_p\left(\frac{1}{\sqrt{N}}\right), \tag{9}$$

and the nonparametric variance estimator is model consistent for the anticipated variance and for the design variance, in the sense that

$$\hat{V}_{NP} - E[\,Var_p(\bar{Y}_S)] = O_p(h_m^{p+1}) + O_p\left(\frac{1}{\sqrt{nh_m}}\right) \tag{10}$$

and

$$\hat{V}_{NP} - \,Var_p(\bar{Y}_S) = O_p(h_m^{p+1}) + O_p\left(\frac{1}{\sqrt{nh_m}}\right). \tag{11}$$

The best bandwidth h_m should satisfy the condition $h_m^{p+1} = O\left(\frac{1}{\sqrt{nh_m}}\right)$, which leads to $h_m = cn^{-1/(2p+3)}$, the usual optimal rate for local polynomial regression (see e.g. Fan and Gijbels, 1996, p.67). Hence, it is expected that the usual bandwidth selection methods such as (generalized) cross-validation or a plug-in method could be applied in this context as well. We do not further explore bandwidth selection in this article.

In these results, the effect of estimating the variance function is asymptotically negligible, because of assumption A.4 on the relationship between h_m and h_v. Without that assumption, model consistency of $\hat{V}_{NP}$ would continue to hold but a more complicated expression for the convergence rates would apply. Similarly, the restriction that p be odd simplifies the expressions for the rates but does not affect the overall consistency.

In Li (2006), a simpler nonparametric estimator is defined as

$$\hat{V}_{NP}^{ho} = \frac{1}{kn^2}(\hat{\mathbf{m}}_S^T \mathbf{D} \hat{\mathbf{m}}_S) + \frac{1}{kn^2}\mathrm{tr}(\mathbf{D})\hat{\sigma}_S^2 \tag{12}$$

with

$$\hat{\sigma}_S^2 = \frac{1}{n}\sum_{j \in S}(Y_j - \hat{m}(x_j))^2, \tag{13}$$

and its properties were studied under the special case of superpopulation model (5) with homoscedastic errors, i.e. when $v(x_j) \equiv \sigma^2, j = 1 \ldots, N$. Under this model, Li (2006) obtained the same results for $\hat{V}_{NP}^{ho}$ as in Theorem 3.1. Because this estimator does not require the additional smoothing step on the residuals, it is easier to compute.

4 Simulation Study

The practical behavior of the proposed nonparametric estimator is evaluated in a simulation study in this section, and in an example on real data in the next. The covariate x_j is uniformly distributed in the interval [0,1], and the errors e_j are generated as an independent and identically distributed (*iid*) sample from a standard normal distribution. Superpopulations of size $N = 2,000$ are generated according to model (5) with two different mean functions

$$\begin{aligned} \text{“linear”:} \quad & m(x_j) = \; 5 + 2x_j \\ \text{“quadratic”:} \quad & m(x_j) = \; 5 + 2x_j - 2x_j^2 \end{aligned}$$

and three different shapes for the variance functions

$$\begin{aligned} \text{“constant”:} \quad & v(x_j) = \; \beta \\ \text{“linear”:} \quad & v(x_j) = \; \beta x_j \\ \text{“quadratic”:} \quad & v(x_j) = \; \beta(1 - 4\,(x_j - 0.5)^2). \end{aligned}$$

The values of β for the three variance functions are selected to achieve two levels for the population coefficient of determination (R^2), equal to $R^2 = 0.75$ (the “precise model”) and $R^2 = 0.25$ (the “diffuse model”).

Several of the estimators are sensitive to the relationship between the modeling covariate and the sorting variable used in generating the systematic samples. We therefore investigate three sorting scenarios, based on the strength of the association between the sorting variable z_j and the x_j. We construct the z_j as $z_j = x_j + \sigma_z \eta_j$ with the η_j *iid* standard normal, and we select the value of σ_z to achieve $R^2 = 1$ (i.e. sorting by x), $R^2 = 0.75$ (“z strongly associated with x”) and $R^2 = 0.25$ (“z weakly associated with x”). We consider samples of sizes $n = 500$ and $n = 100$.

For simplicity, we begin by comparing the performance of the estimators $\hat{V}_{NP}^{ho}$, $\hat{V}_{OL}$, $\hat{V}_{NO}$ and $\hat{V}_{SRS}$ described in Sections 2 and 3 for the homoscedastic error scenarios. This will allow us to evaluate the performance of the nonparametric approach relative to the

commonly used design-based estimators. For the nonparametric estimator, we used local linear ($p = 1$) regression and the Epanechnikov kernel equal to $K(t) = (1-t^2)$ if $|t| \leq 1$ and 0 otherwise. We consider three different values for the bandwidth, $h_m = 0.10, 0.25, 0.50$. The stratification-based estimators $\hat{V}_{OL}$, $\hat{V}_{NO}$ construct pairs of observations based on the sorting variable z_j.

In each simulation run, we keep the population x_j and the z_j fixed but generate new population errors e_j, and draw a systematic sample according to the sorted z values (corresponding to the model-based setting we are considering in this article). Each simulation setting is repeated $B = 10000$ times and the results are obtained by averaging over the B replicates. We consider $\mathrm{E}(\mathrm{Var}_p(\bar{Y}_S))$ as target for the estimators, with $\mathrm{Var}_p(\bar{Y}_S)$ computed exactly for each replicate. Letting $\hat{V}$ denote one of the estimators above, we calculate the relative bias (RB) and mean squared error (MSE), where

$$\begin{aligned} \mathrm{RB} &= \frac{\mathrm{E}^*(\hat{V}) - \mathrm{E}^*[\mathrm{Var}_p(\bar{Y}_S)]}{\mathrm{E}^*[\mathrm{Var}_p(\bar{Y}_S)]}, \\ \mathrm{MSE} &= \mathrm{E}^*(\hat{V} - \mathrm{E}^*[\mathrm{Var}_p(\bar{Y}_S)])^2, \end{aligned}$$

with E^* indicating which expectations are obtained by averaging across the replicates. We also performed simulations in which $\mathrm{Var}_p(\bar{Y}_S)$ is targeted and the prediction mean squared error $\mathrm{MSPE} = \mathrm{E}^*(\hat{V} - \mathrm{Var}_p(\bar{Y}_S))^2$ is computed, but these results are not reported here. Complete simulation results for the homoscedastic case are available in Opsomer et al. (2009). Here, we show the main results for some of the scenarios and briefly summarize the findings of the overall simulation study.

Tables 1 and 2 report the relative bias (in percent) and MSE of $\hat{V}_{NP}^{ho}$, $\hat{V}_{OL}$, $\hat{V}_{NO}$ and $\hat{V}_{SRS}$, for the sample size $n = 500$ and the diffuse ($R^2 = 0.25$) regression model with homoscedastic errors. The MSE results in Table 2 for the different estimators are normalized by dividing by the MSE of $\hat{V}_{NP}^{ho}$ with bandwidth $h_m = 0.10$, to facilitate comparison. The results show that the nonparametric estimator performs well overall, with most biases below 2% in magnitude and most (relative) MSEs smaller than those of the other approaches. It appears that some care is needed in selecting the bandwidth, because in this experiment, larger bandwidth values resulted in significant bias in some of

Mean function	Linear			Quadratic		
Sorting variable R^2	1	0.75	0.25	1	0.75	0.25
$\hat{V}_{NP}^{ho}$: $h_m = 0.10$	-1.98	-0.99	-0.32	-1.96	-1.03	-0.97
$h_m = 0.25$	-0.92	-0.70	-0.16	-0.33	-1.00	-6.98
$h_m = 0.50$	-0.56	-0.54	-0.27	4.93	1.52	-22.5
$\hat{V}_{OL}$	0.05	2.42	16.1	0.13	11.3	-33.8
$\hat{V}_{NO}$	0.01	2.50	15.9	0.09	11.4	-32.7
$\hat{V}_{SRS}$	33.1	26.4	23.9	30.0	22.4	-32.7

Table 1: Relative bias (in percent) for $\hat{V}_{NP}^{ho}$ (with bandwidth $h_m = 0.10, 0.25, 0.50$), $\hat{V}_{OL}$, $\hat{V}_{NO}$ and $\hat{V}_{SRS}$ with $n = 500$, regression model $R^2 = 0.25$ and homoscedastic errors.

Mean function	Linear			Quadratic		
Sorting variable R^2	1	0.75	0.25	1	0.75	0.25
$\hat{V}_{NP}^{ho}$: $h_m = 0.25$	0.94	0.91	0.88	0.92	0.80	1.16
$h_m = 0.50$	0.93	0.89	0.74	1.51	0.84	5.48
$\hat{V}_{OL}$	1.40	1.70	5.83	1.40	4.37	11.9
$\hat{V}_{NO}$	1.88	2.22	6.19	1.88	4.95	11.3
$\hat{V}_{SRS}$	27.0	18.6	10.9	22.4	12.0	11.1

Table 2: MSE of $\hat{V}_{NP}^{ho}$ (with bandwidth $h_m = 0.25, 0.50$), $\hat{V}_{OL}$, $\hat{V}_{NO}$ and $\hat{V}_{SRS}$ divided by MSE of $\hat{V}_{NP}^{ho}$ with bandwidth $h_m = 0.10$, with $n = 500$, regression model $R^2 = 0.25$ and homoscedastic errors.

the cases, as well as larger MSE. Even in those cases, however, the results were generally better than those of the competing methods.

The estimator based on simple random sampling is very inaccurate, resulting in both large biases and MSE values. As noted earlier, the bias of this estimator can be both negative and positive unless the sorting is itself randomized over repeated systematic sampling draws. The two stratification-based estimators perform similarly and are essentially unbiased and efficient when the sorting variable is the same as the model covariate. Performance decreases substantially when the relationship between the sorting variable and the model covariate becomes weaker, resulting in substantial bias and large MSE values. The interpretation of this result is that the stratification-based variance estimators work well only when the "implicit model" under which it is constructed is correct. This implicit model assumes that the relationship between z_j and y_j is well approximated by a piecewise constant function. This is true when $R^2 = 1$, but not in the remaining cases. In contrast, the nonparametric estimator is able to use the correct modeling variable x_j in all cases and is able to capture any unknown but smooth trend.

In the remaining homoscedastic simulations settings that are not displayed here (varying the sample size and R^2 level for the regression model), the results of these estimators are similar to those displayed here, with the relative performance of the different estimators remaining unchanged. Similarly, when MSPE is computed instead of MSE (and hence $\mathrm{Var}_p(\bar{Y}_S)$ is targeted instead of the anticipated variance), the conclusions just stated continue to hold. The main difference is that because of the randomness of $\mathrm{Var}_p(\bar{Y}_S)$, the mean squared differences between the estimators and the target are larger than the differences between the estimators and $\mathrm{E}(\mathrm{Var}_p(\bar{Y}_S))$, and hence the normalized MSPEs are all closer to 1. Full results are in Opsomer et al. (2009).

The anticipated variance in (7) contains two components, with the first related to the difference in the sample means and the second to the model variance. While the first component can in principle be decreased by choosing an appropriate sorting variable (this is a major theme in the systematic sampling literature), the second component is independent of the sorting. Because the proposed model-based variance estimator targets

Mean function	Linear			Quadratic		
Variance function	Constant	Linear	Quadratic	Constant	Linear	Quadratic
$\hat{V}_{NP}^{ho}$: $h_m = 0.10$	-0.99	-1.19	-0.93	-1.03	-1.23	-0.97
$h_m = 0.25$	-0.70	-0.82	-0.59	-1.00	-1.12	-0.89
$h_m = 0.50$	-0.54	-0.59	-0.41	1.52	1.48	1.64
$\hat{V}_{NP}$: $h_m = 0.10$, $h_v = 0.10$	-1.12	-1.32	-1.93	-1.16	-1.36	-1.97
$h_m = 0.10$, $h_v = 0.25$	-1.11	-1.31	-5.95	-1.15	-1.35	-5.96
$h_m = 0.10$, $h_v = 0.50$	-1.14	-1.34	-13.2	-1.18	-1.39	-13.1
$\hat{V}_{NP}$: $h_m = 0.25$, $h_v = 0.10$	-0.77	-0.89	-1.58	-1.05	-1.18	-1.85
$h_m = 0.25$, $h_v = 0.25$	-0.80	-0.91	-5.65	-1.14	-1.25	-5.95
$h_m = 0.25$, $h_v = 0.50$	-0.84	-0.96	-13.0	-1.21	-1.34	-13.3
$\hat{V}_{NP}$: $h_m = 0.50$, $h_v = 0.10$	-0.57	-0.63	-1.39	1.57	1.53	0.75
$h_m = 0.50$, $h_v = 0.25$	-0.60	-0.65	-5.47	1.37	1.34	-3.47
$h_m = 0.50$, $h_v = 0.50$	-0.65	-0.71	-12.8	0.24	0.181	-11.8
$\hat{V}_{NO}$	2.50	2.49	2.48	11.4	11.5	11.4

Table 3: Relative bias (in percent) for $\hat{V}_{NP}^{ho}$ (with bandwidths $h_m = 0.10, 0.25, 0.50$), $\hat{V}_{NP}$ (with bandwidths $h_m, h_v = 0.10, 0.25, 0.50$) and $\hat{V}_{NO}$ when the sorting variable is strongly associated with the model covariate ($R^2 = 0.75$), with $n = 500$, regression model $R^2 = 0.25$.

the anticipated variance, it would therefore appear critical to capture the model variance correctly in order to obtain a good variance estimator. This was the main reason for the nonparametric specification of the function $v(\cdot)$.

In order to investigate the effect of the model variance specification and the ability of the nonparametric estimator to account for heteroscedasticity, Tables 3 and 4 display the relative bias and normalized MSE of the nonparametric estimators for different model variance function specifications, for a range of bandwidths for the mean and variance smoothing steps, for sample size $n = 500$, sorting variable strongly associated with x

Mean function	Linear			Quadratic		
Variance function	Constant	Linear	Quadratic	Constant	Linear	Quadratic
$\hat{V}_{NP}^{ho}$: $h_m = 0.25$	0.91	0.91	0.96	0.80	0.86	0.82
$h_m = 0.50$	0.89	0.89	0.96	0.84	0.89	0.85
$\hat{V}_{NP}$: $h_m = 0.10, h_v = 0.10$	1.03	1.03	1.05	1.02	1.02	1.04
$h_m = 0.10, h_v = 0.25$	1.04	1.08	1.66	1.03	1.07	1.55
$h_m = 0.10, h_v = 0.50$	1.09	1.19	4.53	1.08	1.18	3.93
$\hat{V}_{NP}$: $h_m = 0.25, h_v = 0.10$	0.94	0.94	1.0	0.82	0.88	0.85
$h_m = 0.25, h_v = 0.25$	0.95	0.99	1.55	0.84	0.93	1.36
$h_m = 0.25, h_v = 0.50$	1.00	1.11	4.38	0.89	1.04	3.81
$\hat{V}_{NP}$: $h_m = 0.50, h_v = 0.10$	0.91	0.92	0.98	0.87	0.91	0.80
$h_m = 0.50, h_v = 0.25$	0.93	0.97	1.51	0.86	0.95	0.94
$h_m = 0.50, h_v = 0.50$	0.98	1.09	4.29	0.86	1.02	3.15
$\hat{V}_{NO}$	2.22	2.04	2.26	4.95	4.40	4.37

Table 4: MSE of $\hat{V}_{NP}^{ho}$ (with bandwidths $h_m = 0.25, 0.50$), $\hat{V}_{NP}$ (with bandwidths $h_m, h_v = 0.10, 0.25, 0.50$) and $\hat{V}_{NO}$ divided by MSE of $\hat{V}_{NP}^{ho}$ with bandwdith $h_m = 0.10$, when the sorting variable is strongly associated with the model covariate ($R^2 = 0.75$), with $n = 500$, regression model $R^2 = 0.25$.

($R^2 = 0.75$) and diffuse regression model ($R^2 = 0.25$). For comparison, we include $\hat{V}_{NO}$, which was the best among the alternative variance estimation methods. The results show that the variance function specification generally has only a modest effect on both the bias and the MSE of the estimators. The only exception to this is when the variance function is modeled nonparametrically and too large a bandwidth is used ($h_v = 0.50$). An interesting result is that the estimator $\hat{V}_{NP}^{ho}$, which uses the mean squared residuals, appears to perform better than many of the more complicated $\hat{V}_{NP}$ even when the errors were heteroscedastic (this was also true in the other scenarios not displayed here). We conjecture that this is due to the fact that under heteroscedasticity, the variance com-

ponent of the anticipated variance is of the form $c\sum_{j=1}^{N} v(x_j)/N$, and the mean of the squared residuals is a very good estimator for the mean of the variance over the population for approximately balanced samples, such as those obtained by systematic sampling.

5 Application

The applicability of the method to real surveys is illustrated on a forestry data, which was previously analyzed in Opsomer et al. (2007). The data were collected during the 1990's by the U.S. Forest Service within a 2.5 million ha ecological province in northern Utah, USA. The field sample plots located on a regular spatial grid and are supplemented by remotely sensed data available on a much finer spatial grid. Figure 1 displays the study region and sample locations for the survey data and additional remote sensing data. The latter can be used as auxiliary information to improve the precision of survey estimators, as previously done by Opsomer et al. (2007) who explored nonparametric model-assisted estimation methods. In the current article, we will use the auxiliary information to construct an estimator for the variance of survey estimators.

Data are available for 24,980 remote sensing points and 968 field-visited points. The remote sensing data are available at essentially any desirable resolution, so this grid of points is somewhat arbitrary and can be used as an approximation for the underlying continuous population. We therefore treat these as the population of interest and field-visited points as a sample drawn from that population, corresponding to a 1-in-25 systematic sample. At the "population" level, we have auxiliary information such as location (**LOC**, bivariate scaled longitude and latitude) and elevation (ELEV). At the sample level, information is available for the field-collected forestry variables in addition to the population-level variables.

The sampling design is spatial systematic sampling, as seen from Figure 1. While the discussion in the previous sections only addressed the univariate case, the methods readily carry over for higher dimensions, as will be illustrated here. We consider here the following forestry variables:

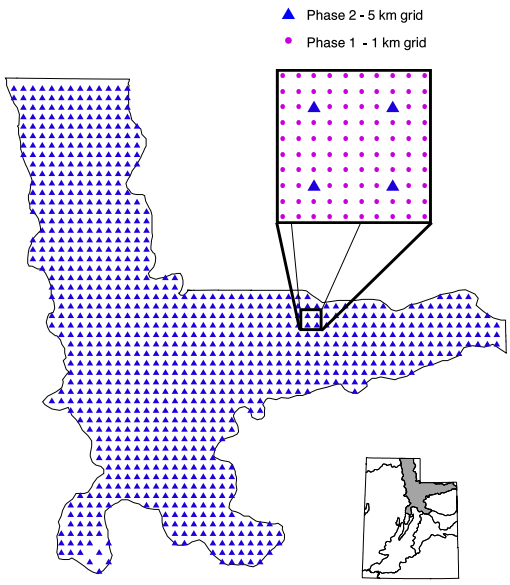

Figure 1: Map of the study region in northern Utah. Each triangle represents a field-visited sample point. Each dot in the magnified section represents a remote sensing point.

- BIOMASS - total wood biomass per acre in tons
- CRCOV - percent crown cover
- BA - tree basal area per acre
- NVOLTOT - total volume in cubic feet per acre
- FOREST - forest/nonforest indicator.

We are interested in estimating the population mean for these variables using the systematic sample mean $\bar{Y}_S$, and estimating its design-based variance $\mathrm{Var}_p(\bar{Y}_S)$. We will consider two traditional design-based variance estimators, $\hat{V}_{SRS}$ as in (2) and $\hat{V}_{ST}$ (see below), and the model-based nonparametric variance estimator $\hat{V}^{ho}_{NP}$. The stratified sampling variance estimator $\hat{V}_{ST}$ is similar to the nonoverlapping differences estimator $\hat{V}_{NO}$ in (4), generalized to a spatial setting by considering an approximate 4-per-stratum design obtained by

overlaying a grid of equal-sized "cells" over the study region. This estimator is defined as

$$\hat{V}_{ST} = \frac{1-f}{n}\frac{1}{n}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{j\in S_h}(Y_j - \bar{Y}_{S_h})^2,$$

where S_h denotes the sample in cell h and n_h the corresponding cell sample size. For points near the edge of the map, there may be more or less than four points per cell, because we collapsed all cells that contained less than two points with their closest neighbor.

For the purpose of constructing the model-based nonparametric variance estimator $\hat{V}_{NP}$, we consider the following model with location (**LOC**) as bivariate auxiliary variables:

$$Y_j = m(\mathbf{LOC}_j) + \varepsilon_j. \tag{14}$$

We considered both homoscedastic and heteroscedastic versions of this model. Because the homoscedastic version of the nonparametric estimator appeared to behave at least as well as the more complicated estimator that captures heteroscedasticity, we will assume here that the errors are independent with homogeneous variance. Full results for other model specifications are shown in Opsomer et al. (2009).

Under model (14), we implemented the nonparametric variance estimator $\hat{V}_{NP}^{ho}$ given in (12) and (13) with x_j replaced by $\mathbf{LOC}_j$. Here $m(\cdot)$ is estimated by bivariate local linear regression, and the estimator $\hat{m}(\cdot)$ is obtained using loess() in R. In loess(), the bandwidth parameter h is replaced by the *span*, the fraction of the sample observations that have non-zero weight in the computation of $\hat{m}(\mathbf{LOC}_j)$. Since the samples points are approximately equally spaced (5×5 km grid), using loess() will produce similar results to those obtained using a fixed bandwidth in the interior of the estimation region. At the boundaries of the region, it will tend to select larger bandwidths and hence reduce some of the increased variability often experienced close to boundaries in fixed-bandwidth smoothing. This results in improved overall stability of the fits. In order to evaluate the sensitivity of the results to the choice of the smoothing parameters, we choose three spans: 0.1, 0.2 and 0.5. After obtaining $\hat{m}(\cdot)$, we can calculate the nonparametric variance estimator $\hat{V}_{NP}^{ho}$ for each response variable. Table 5 presents the sample means and the estimated variances using $\hat{V}_{SRS}$, $\hat{V}_{ST}$ and $\hat{V}_{NP}^{ho}$.

	$\bar{Y}_S$	$\hat{V}_{SRS}$	$\hat{V}_{ST}$	$\hat{V}^{ho}_{NP0.5}$	$\hat{V}^{ho}_{NP0.2}$	$\hat{V}^{ho}_{NP0.1}$
BIOMASS	14.5	0.46	0.36	0.40	0.38	0.37
CRCOV	22.5	0.71	0.62	0.64	0.62	0.59
BA	48.5	3.87	3.19	3.40	3.30	3.12
NVOLTOT	906.9	1886	1538	1645	1584	1511
FOREST (%)	54.8	2.46	1.89	2.16	2.05	1.91

Table 5: Mean and variance estimates for the five response variables for forestry data, using estimators $\hat{V}_{SRS}$, $\hat{V}_{ST}$ and $\hat{V}^{ho}_{NP}$ under model (15) with span = 0.5, 0.2 and 0.1.

While we do not know the true variance, the estimator $\hat{V}_{ST}$ is likely to be a reasonable approximation as long as the Y_j can be modeled as a spatial trend plus random errors. The naive estimator $\hat{V}_{SRS}$ produces the largest values among the five variance estimators for all response variables and so is likely to be biased upwards for this survey. In contrast, the nonparametric variance estimator $\hat{V}^{ho}_{NP}$ results in estimates that are close to those of $\hat{V}_{ST}$, with smaller spans leading to slightly smaller estimates.

As already discussed in Section 4, an important advantage of the model-based nonparametric method is that one is not restricted to using only the sorting variable (**LOC** in this case) in the construction of the estimator, if other variables are thought to be good predictors of the survey variables. We illustrate this here by considering a more sophisticated model that also includes elevation (ELEV) in additive to **LOC**:

$$Y_j = m_1(\mathbf{LOC}_j) + m_2(ELEV_j) + \varepsilon_j. \tag{15}$$

We fit model (15) in R using the Generalized Additive Models (gam) package. For simplicity, we use the same span for both **LOC** and ELEV. Table 6 shows that, relative to the simpler model without elevation, the estimated variances all decreased, by 8-14%. This decrease is due primarily to a reduction in the $\hat{\sigma}^2_S$ component in (12), which accounts for the fact that the extended mean model in (15) captures more of the observed behavior of these forestry variables. This allows for a more precise estimator of the design-based variance of systematic sampling estimators for these data. For the variables CRCOV, BA

	$\hat{V}^{ho}_{NP0.5}$	$\hat{V}^{ho}_{NP0.2}$	$\hat{V}^{ho}_{NP0.1}$
BIOMASS	0.36	0.34	0.33
CRCOV	0.59	0.55	0.53
BA	3.11	2.96	2.78
NVOLTOT	1487	1417	1342
FOREST (%)	1.92	1.77	1.65

Table 6: Variance estimates for five response variables for forestry data, using nonparametric estimator for additive model 15 with span = 0.5, 0.2 and 0.1. Same span used for both variables.

and NVOLTOT, the estimated variances based on model (15) are smaller than $\hat{V}_{SRS}$ and $\hat{V}_{ST}$ for all considered bandwidth choices, while those for BIOMASS and FOREST are similar. Hence, this application illustrates the potential for obtaining variance estimates that are both sharper and more theoretically justified than using traditional design-based approaches. We refer to Opsomer et al. (2009) for a more complete discussion of the results using the extended model.

Acknowledgments

This work was partially supported by MEC Grant MTM2008-00166 (ERDF included) and by Xunta de Galicia Grant PGIDIT07PXIB105259PR.

References

Bartolucci, F. and G. E. Montanari (2006). A new class of unbiased estimators for the variance of the systematic sample mean. *Journal of Statistical Planning and Inference 136*, 1512–1525.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*, 645–660.

Li, X. (2006). *Application of Nonparametric Regression in Survey Statistics*. Ph. D. thesis, Department of Statistics, Iowa State University.

Madow, W. G. and L. G. Madow (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics 15*, 1–24.

Montanari, G. E. and F. Bartolucci (1998). On estimating the variance of the systematic sample mean. *Journal of the Italian Statistical Society 7*, 185–196.

Opsomer, J., M. Francisco-Fernández, and X. Li (2009). Additional results for model-based nonparametric variance estimation for systematic sampling in a forestry survey. Technical report, Department of Statistics, Colorado State University.

Opsomer, J. D., F. J. Breidt, G. G. Moisen, and G. Kauermann (2007). Model-assisted estimation of forest resources with generalized additive models (with discussion). *Journal of the American Statistical Association 102*, 400–416.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *Annals of Statistics 22*, 1346–1370.

Ruppert, D., M. P. Wand, U. Holst, and O. Hössjer (1997). Local polynomial variance function estimation. *Technometrics 39*, 262–273.

Särndal, C. E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Törnqvist, L. (1963). The theory of replicated systematic cluster sampling with random start. *Revue de l'Institut International de Statistique 31*, 11–23.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman and Hall.

Wolter, K. M. (2007). *Introduction to Variance Estimation* (2 ed.). New York: Springer-Verlag Inc.

Zinger, A. (1980). Variance estimation in partially systematic sampling. *Journal of the American Statistical Association 75*, 206–211.

A Assumptions

A.1 *The x_j's are treated as fixed with respect to superpopulation model* (5). *The x_j's are independent and identically distributed with $F(x) = \int_{-\infty}^{x} f(t)dt$, where $f(\cdot)$ is a density function with compact support $[a_x, b_x]$ and $f(x) > 0$ for all $x \in [a_x, b_x]$. The first derivative of f exists for all $x \in [a_x, b_x]$.*

A.2 *The third and fourth moments of e_j exist and are bounded.*

A.3 *The sample size n and sampling interval k are positive integers with nk=N. We assume that $n, N \to \infty$ and allow $k = O(1)$ or $k \to \infty$.*

A.4 *As $n \to \infty$, we assume $h^* \to 0$ and $nh^* \to \infty$, where h^* is h_m or h_v. Additionally, $\left\{h_m^{2(p+1)} + (nh_m)^{-1}\right\} = o\left(h_v^{p+1}\right)$, $h_v^{p+1} = o(nh_m^{p+1})$ and $\frac{1}{n^2 h_v^{1/2}} = o\left(\frac{1}{n^{1/2}h_m^{1/2}}\right)$.*

A.5 *The kernel function $K(\cdot)$ is a compactly supported, bounded, symmetric kernel with $\int u^{q+1}K(u)du = \mu_{q+1}(K)$. Assume that $\mu_{p+1}(K) \neq 0$.*

A.6 *The $(p+1)$th derivatives of the mean function $m(\cdot)$ and the variance function $v(\cdot)$ exist and are bounded on $[a_x, b_x]$.*

B Outline of Proof of Theorem 3.1

Statement (9) is obtained by showing that $\mathrm{Var}[\mathrm{Var}_p(\bar{Y}_S)] = O(1/N)$. This is done by computing the variance of the quadratic form in (6) under model (5), and then bounding each of the terms using assumptions A.1-A.3 and A.6. See Theorem 1.1 in Li (2006) for details.

In order to prove (10), we write

$$\begin{aligned}\hat{V}_{NP} - \mathrm{E}[\mathrm{Var}_p(\bar{Y}_S)] &= \frac{1}{Nn}(\hat{\mathbf{m}}^T\mathbf{D}\hat{\mathbf{m}} - \mathbf{m}^T\mathbf{D}\mathbf{m}) + \left(1 - \frac{n}{N}\right)\frac{1}{Nn}\sum_{j\in U}(\hat{v}(x_j) - v(x_j)) \\ &= A + B. \end{aligned} \tag{16}$$

The term A in (16) can be broken down into several components that are functions of $\frac{1}{n}\sum_{j\in S_b}(\hat{m}(x_j)-m(x_j))^l$ for $b = 1,\ldots,k$ and $\frac{1}{N}\sum_{j\in U}(\hat{m}(x_j)-m(x_j))^l$ with $l = 1, 2$. Using the same approach as in the proof of Theorem 4.1 in Ruppert and Wand (1994) except that we are treating the x_j as fixed, and applying assumptions A.1-A.6, we approximate the required moments of these quantities. Bounding arguments for the expectation and variance of each of the components of A show that $A = O_p(h_m^{p+1} + (nh_m)^{-1/2})$. Theorem 1.2 in Li (2006) provides a complete description.

For the term B in (16), the squared residuals are decomposed into $\hat{r}_j = v(x_j)e_j^2 + (\hat{m}(x_j)-m(x_j))^2 - 2\sqrt{v(x_j)}e_j(\hat{m}(x_j)-m(x_j)) = r_j + b_{1j} + b_{2j}$, with corresponding sample vectors $\hat{\mathbf{r}}_S = \mathbf{r}_S + \mathbf{b}_{1S} + \mathbf{b}_{2S}$. Hence, $\mathbf{r}_S$ contains the true model errors for the sample observations and does not depend on the first nonparametric regression. Letting $\tilde{v}$ denote the local polynomial regression fit using $\mathbf{r}_S$ instead of $\hat{\mathbf{r}}_S$, straightforward moment approximations and bounding arguments show that

$$\left(1 - \frac{n}{N}\right)\frac{1}{Nn}\sum_{j\in U}(\tilde{v}(x_j) - v(x_j)) = O_p\left(\frac{h_v^{p+1}}{n} + \frac{1}{n^2 h_v^{1/2}}\right) = o_p\left(h_m^{p+1} + \frac{1}{\sqrt{nh_m}}\right)$$

by assumption A.4. Using the fact that $\mathrm{E}(b_{1j}) = O(h_m^{2p+2} + (nh_m)^{-1})$ and A.4 again, we can show that the local polynomial regression for $\mathbf{b}_{1S}$ is $o_p(h_v^{p+1})$. Similarly, the local polynomial regression for $\mathbf{b}_{2S}$ leads to terms that are of the same or smaller order. Hence, we conclude that $B = o_p\left(h_m^{p+1} + \frac{1}{\sqrt{nh_m}}\right)$.

Finally, statement (11) follows directly from (9) and (10).

■