

# Nonparametric Regression Estimation of Finite Population Totals under Two-Stage Sampling

Ji-Yeon Kim\*

Iowa State University

F. Jay Breidt†

Colorado State University

Jean D. Opsomer‡

Colorado State University

September 21, 2009

## Abstract

We consider nonparametric regression estimation for finite population totals for two-stage sampling, in which complete auxiliary information is available for first-stage sampling units. The estimators, based on local polynomial regression, are linear combinations of cluster total estimators, with weights that are calibrated to known control totals. The estimators are asymptotically design-unbiased and design consistent under mild assumptions. We provide a consistent estimator for the design mean squared error of the estimators. Simulation results indicate that the nonparametric estimator dominates standard parametric estimators when the model regression function is incorrectly specified, while being nearly as efficient when the parametric specification is correct. The methodology is illustrated using data from a study of land use and erosion.

---

\*Department of Statistics, Iowa State University, Ames IA 50011, USA.

†Department of Statistics, Colorado State University, Fort Collins CO 80523, USA, jbreidt@stat.colostate.edu.

‡Department of Statistics, Colorado State University, Fort Collins CO 80523, USA, jopsomer@stat.colostate.edu.

**Keywords:** Calibration, model-assisted estimation, local polynomial regression, cluster sampling.

## 1 Introduction

In many complex surveys, auxiliary information about the population of interest is available. One approach to using this auxiliary information in estimation is to assume a working model describing the relationship between the study variable of interest and the auxiliary variables. Estimators are then derived on the basis of this model. Estimators are sought which have good efficiency if the model is true, but maintain desirable properties like design consistency if the model is false.

Often, a linear model is selected as the working model. Generalized regression estimators (e.g., Cassel, Särndal, and Wretman, 1976; Särndal, 1980; Robinson and Särndal, 1983), including ratio estimators and linear regression estimators (Cochran, 1977), best linear unbiased estimators (Brewer, 1963; Royall, 1970), and poststratification estimators (Holt and Smith, 1979), are all derived from assumed linear models. In some situations, the linear model is not appropriate, and the resulting estimators do not achieve any efficiency gain over purely design-based estimators. Wu and Sitter (2001) propose a class of estimators for which the working models follow a nonlinear parametric shape. However, efficient use of any of these estimators requires a priori knowledge of the specific parametric structure of the population. This is especially problematic if that same model is to be used for many variables of interest, a common occurrence in surveys.

Because of these concerns, some researchers have considered nonparametric models for  $\xi$ . Dorfman (1992) and Chambers, Dorfman, and Wehrly (1993) developed model-based nonparametric estimators using kernel regression. More recently, Zheng and Little (2003) proposed a model-based estimator that uses penalized spline regression, and Zheng and Little (2004) extended this estimator to two-stage sampling designs. Breidt and Opsomer (2000) proposed a new type of model-assisted nonparametric regression estimator for the finite population total, based on local polynomial smoothing, a kernel-based method. The local polynomial regression estimator has the form of the generalized regression estimator, but is based on a nonparametric super-

population model applicable to a much larger class of functions. Breidt, Claeskens, and Opsomer (1995) consider a related nonparametric model-assisted regression estimator, replacing local polynomial smoothing with penalized splines.

The theory developed in Breidt and Opsomer (2000) for the local polynomial regression estimator applies only to direct element sampling designs with auxiliary information available for all elements of the population. In many large-scale surveys, however, more complex designs such as multistage or multiphase sampling designs with various types of auxiliary information are commonly used. In this paper, we extend local polynomial nonparametric regression estimation to two-stage sampling, in which a probability sample of clusters is selected, and then subsamples of elements within each selected cluster are obtained.

Two-stage sampling is frequently used because an adequate frame of elements is not available or would be prohibitively expensive to construct, but a listing of clusters is available. Särndal *et al.* (1992, p.304) identify three cases of auxiliary information available for two-stage sampling, depending on whether the information is available at the cluster level, element level for all elements, or element level for elements in selected clusters only. We consider the first of these cases, which is commonly encountered in practice and reflects the fact that an element-level frame is not available in many situations in which multi-stage sampling is applied.

When modeling at the cluster level, two possible options are to model element-level characteristics of the clusters, for instance by relating the cluster mean to the auxiliary variable(s), and to model cluster totals. This article will focus on the latter case, because cluster-level auxiliary variables most readily available are often related to the size of the cluster, which is usually correlated with the cluster totals of the survey variables. We will therefore construct a model-assisted estimator that uses a nonparametric regression model between the cluster totals and the auxiliary variables. Results for single-stage cluster sampling, in which each sampled cluster is completely enumerated, are obtained as a special case.

Situations in which this type of cluster-level modeling might be particularly useful include natural resource surveys, and we will illustrate our methodology on data from a 1995 study of erosion, using the National Resources Inventory (NRI) data as frame materials. Another example might be a survey with households as elements and, say,

cities as clusters, and the auxiliary variables are various measures related to the size of the cities.

The remainder of the article is structured as follows. In Section 1.1, we describe our two-stage sampling framework, and in Section 1.2, we adapt the local polynomial regression estimator of Breidt and Opsomer (2000) to two-stage sampling. Design properties of the estimator are described in Section 2. Section 2.1 shows that the estimator is a linear combination of estimators of cluster totals with weights that are calibrated to known control totals. Section 2.2 shows asymptotic design unbiasedness and design consistency of the estimator, approximates the estimator's design mean squared error, and provides a consistent estimator of the design mean squared error. Section 3 describes results of a simulation study. In Section 4, we apply the estimator to the NRI erosion data.

## 1.1 Two-stage design and estimation

Consider a finite population of elements  $U = \{1, \dots, k, \dots, N\}$  partitioned into  $M$  clusters, denoted  $U_1, \dots, U_i, \dots, U_M$ . The population of clusters is represented as  $C = \{1, \dots, i, \dots, M\}$ . The number of elements in the  $i$ th cluster  $U_i$  is denoted  $N_i$ . We have  $U = \cup_{i \in C} U_i$  and  $N = \sum_{i \in C} N_i$ . For all clusters  $i \in C$ , an auxiliary vector  $\mathbf{x}_i = (x_{1i}, \dots, x_{Gi})'$  is available. For the sake of simplicity we assume that  $G = 1$ ; that is, the  $x_i$  are scalars.

At stage one, a probability sample  $s$  of clusters is drawn from  $C$  according to a fixed size design  $p_I(\cdot)$ , where  $p_I(s)$  is the probability of drawing the sample  $s$  from  $C$ . Let  $m$  be the size of  $s$ . The cluster inclusion probabilities  $\pi_i = \Pr\{i \in s\} = \sum_{s:i \in s} p_I(s)$  and  $\pi_{ij} = \Pr\{i, j \in s\} = \sum_{s:i, j \in s} p_I(s)$  are assumed to be strictly positive, where  $p_I$  refers to first-stage design.

For every sampled cluster  $i \in s$ , a probability sample  $s_i$  of elements is drawn from  $U_i$  according to a fixed size design  $p_i(\cdot)$  with inclusion probabilities  $\pi_{k|i}$  and  $\pi_{kl|i}$ . That is,  $p_i(s_i)$  is the probability of drawing  $s_i$  from  $U_i$  given that the  $i$ th cluster is chosen at stage one. The size of  $s_i$  is denoted  $n_i$ . Assume that  $\pi_{k|i} = \Pr\{k \in s_i | s \ni i\} = \sum_{s_i:k \in s_i} p_i(s_i)$  and  $\pi_{kl|i} = \Pr\{k, l \in s_i | s \ni i\} = \sum_{s_i:k, l \in s_i} p_i(s_i)$  are strictly positive. As is customary for two-stage sampling, we assume invariance and independence of the second-stage design (Särndal *et al.*, 1992, p.134). The whole sample of elements

and its size are  $\cup_{i \in s} s_i$  and  $\sum_{i \in s} n_i$ , respectively. The study variable  $y_k$  is observed for  $k \in \cup_{i \in s} s_i$ . The parameter to estimate is the population total  $t_y = \sum_{k \in U} y_k = \sum_{i \in C} t_i$ , where  $t_i = \sum_{k \in U_i} y_k$  is the  $i$ th cluster total.

Let  $I_i = 1$  if  $i \in s$  and  $I_i = 0$  otherwise. Note that  $E_p [I_i] = E_I [E_{II} [I_i]] = E_I [I_i] = \pi_i$ , where  $E_p [\cdot]$  denotes expectation with respect to the two-stage sampling design,  $E_I [\cdot]$  denotes expectation with respect to stage one, and  $E_{II} [\cdot]$  denotes conditional expectation with respect to stage two given  $s$ . Also,  $V_I(\cdot)$  and  $V_{II}(\cdot)$  denote variances with respect to stage one and two, respectively. Using this notation, an estimator  $\hat{t}$  of  $t$  is said to be design-unbiased if  $E_p [\hat{t}] = t$ .

The simple expansion estimator of  $t_y$  in two-stage element sampling is given by

$$\hat{t}_y = \sum_{i \in s} \frac{\hat{t}_i}{\pi_i} = \sum_{i \in C} \frac{\hat{t}_i I_i}{\pi_i}, \quad (1)$$

where

$$\hat{t}_i = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}}$$

is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) of  $t_i$  with respect to stage two. We will refer to (1) as the Horvitz-Thompson (HT) estimator. Since  $\hat{t}_i$  is design-unbiased for  $t_i$ , the HT estimator  $\hat{t}_y$  is design-unbiased for  $t_y$ . Note that  $\hat{t}_y$  does not depend on the  $x_i$ . The variance of the HT estimator  $\hat{t}_y$  under the sampling design can be written as the sum of two components,

$$\begin{aligned} \text{Var}_p(\hat{t}_y) &= V_I(E_{II}[\hat{t}_y]) + E_I[V_{II}(\hat{t}_y)] \\ &= \sum_{i,j \in C} (\pi_{ij} - \pi_i \pi_j) \frac{t_i}{\pi_i} \frac{t_j}{\pi_j} + \sum_{i \in C} \frac{V_i}{\pi_i}, \end{aligned} \quad (2)$$

where

$$V_i = V_{II}(\hat{t}_i) = \sum_{k,l \in U_i} (\pi_{kl|i} - \pi_{k|i} \pi_{l|i}) \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}$$

is the variance of  $\hat{t}_i$  with respect to stage two. A design-unbiased estimator of  $V_i$  is given by

$$\hat{V}_i = \sum_{k,l \in s_i} \frac{\pi_{kl|i} - \pi_{k|i} \pi_{l|i}}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}.$$

Note that  $V_i$  is non-random due to invariance. Note also that the result for single-stage cluster sampling, in which all elements in each selected cluster are observed, is

obtained if we set  $\hat{t}_i = t_i$  and  $V_i = \hat{V}_i = 0$  for all  $i \in C$ . See, for example, Särndal *et al.* (1992, Result 4.3.1).

## 1.2 Local Polynomial Regression Estimator

The model-assisted approach to using the auxiliary information  $\{x_i\}_{i \in C}$  in a model for the cluster totals  $\{t_i\}_{i \in C}$  is to assume as a working model that the finite population point scatter  $\{(x_i, t_i)\}_{i \in C}$  is a realization from a superpopulation model  $\xi$ , in which

$$t_i = \mu(x_i) + \varepsilon_i, \quad i = 1, \dots, M \quad (3)$$

where the  $\varepsilon_i$  are independent random variables with  $E_\xi[\varepsilon_i] = 0$  and  $\text{Var}_\xi(\varepsilon_i) = \nu_i$ .

Typically, both  $\mu(x_i)$  and  $\nu_i$  are taken to be parametric functions of  $x_i$ , such as the linear specification  $\mu(x_i) = \sum_j \beta_j a_{\mu,j}(x_i)$  and  $\nu_i = \nu(x_i) = \sum_j \lambda_j a_{\nu,j}(x_i)$ , where the  $a_{\mu,j}$  and  $a_{\nu,j}$  are known functions and the  $\beta_j$  and  $\lambda_j$  are unknown parameters. A variety of heteroskedastic polynomial regression models could be specified in this way (e.g., Särndal *et al.*, 1992, Section 8.4).

As mentioned in Section 1, the model-assisted methodology offers efficiency gains if the working model describes the finite population point scatter reasonably well. The problem is that, in an actual survey, there is not a *single* point scatter, but *many*, corresponding to different study variables  $t_i$ . Standard survey practice is to use the working model to construct *one* set of weights that reflects the design and the auxiliary information in  $\{x_i\}$ , and apply this one set of weights to *all* study variables. Thus, it is critical to keep the model specification flexible, which motivates the nonparametric approach we employ. Rather than specify a parametric model, we assume only that  $\mu(x_i)$  is a smooth function of  $x_i$  and do not further specify the variance function  $\nu_i$ . This nonparametric working model has the potential to offer efficiency gains for a greater variety of variables measured on the population than the parametric model, while maintaining most of the efficiency of the parametric regression estimator if the parametric model is correct.

We now introduce some further notation used in the nonparametric regression, following the approach of Breidt and Opsomer (2000). Let  $K$  denote a kernel function and  $h_M$  denote its bandwidth. Let  $\mathbf{t}_C = [t_i]_{i \in C}$  be the vector of  $t_i$ 's in the population

of clusters. Define the  $M \times (q + 1)$  matrix

$$\mathbf{X}_{C_i} = \begin{bmatrix} 1 & x_1 - x_i & \cdots & (x_1 - x_i)^q \\ \vdots & \vdots & & \vdots \\ 1 & x_M - x_i & \cdots & (x_M - x_i)^q \end{bmatrix} = \left[ \mathbf{1} \quad x_j - x_i \quad \cdots \quad (x_j - x_i)^q \right]_{j \in C},$$

and define the  $M \times M$  matrix

$$\mathbf{W}_{C_i} = \text{diag} \left\{ \frac{1}{h_M} K \left( \frac{x_j - x_i}{h_M} \right) \right\}_{j \in C}.$$

Let  $\mathbf{e}_r$  represent the  $r$ th column of the identity matrix. The local polynomial regression estimator of  $\mu(x_i)$ , based on the entire population of clusters, is given by

$$\mu_i = \mathbf{e}'_1 (\mathbf{X}'_{C_i} \mathbf{W}_{C_i} \mathbf{X}_{C_i})^{-1} \mathbf{X}'_{C_i} \mathbf{W}_{C_i} \mathbf{t}_C = \mathbf{w}'_{C_i} \mathbf{t}_C, \quad (4)$$

which is well-defined as long as  $\mathbf{X}'_{C_i} \mathbf{W}_{C_i} \mathbf{X}_{C_i}$  is invertible.

The kernel function  $K$  determines “local neighborhoods” around each  $x_i$ , with the bandwidth  $h$  used to define the extent of those neighborhoods. Only observations in that neighborhood then enter into the regression fit at  $x_i$  to obtain  $\mu_i$ . This is the “traditional” local polynomial kernel estimator described in e.g. Wand and Jones (1995). Note that this estimator does not account for possible model heteroskedasticity, which could in principle be incorporated by including  $\nu_i$  into the weighted regression fit. This is not typically done in the local polynomial regression context, because the local nature of the kernel-weighted fits implies that, for each of the  $x_i$ , the observations in the relevant neighborhood are close to homoskedastic when  $\nu_i = \nu(x_i)$  is smooth. In our methodology, smoothness of  $\nu_i$  is not required. Any improvements in efficiency from incorporating  $\nu_i$  would come at the cost of a more complex estimation procedure and/or more restrictive modeling assumptions, and we do not pursue this further here.

If the population-level  $\mu_i$ 's were known, then a design-unbiased estimator of  $t_y$  would be the two-stage analogue of the generalized difference estimator (Särndal *et al.*, 1992, p. 222),

$$t_y^* = \sum_{i \in s} \frac{\hat{t}_i - \mu_i}{\pi_i} + \sum_{i \in C} \mu_i. \quad (5)$$

The design variance of (5) is

$$\text{Var}_p(t_y^*) = \sum_{i, j \in C} (\pi_{ij} - \pi_i \pi_j) \frac{t_i - \mu_i}{\pi_i} \frac{t_j - \mu_j}{\pi_j} + \sum_{i \in C} \frac{V_i}{\pi_i}, \quad (6)$$

which depends on residuals from the nonparametric regression and hence is expected to be smaller than (2). Note that, since a model is assumed for the cluster totals but not for the individual observations, only the variance component at the cluster level in (6) is affected by the model.

In the present context, the population estimator  $\mu_i$  cannot be calculated because only the  $y_k$  in  $\cup_{i \in s} s_i$  are known. Therefore, we will replace each  $\mu_i$  by a sample-based consistent estimator. Let  $\hat{\mathbf{t}}_s = [\hat{t}_i]_{i \in s}$  be the vector of  $\hat{t}_i$ 's obtained in the sample of clusters. Define the  $m \times (q+1)$  matrix

$$\mathbf{X}_{si} = \left[ \begin{array}{cccc} 1 & x_j - x_i & \cdots & (x_j - x_i)^q \end{array} \right]_{j \in s}, \quad (7)$$

and define the  $m \times m$  matrix

$$\mathbf{W}_{si} = \text{diag} \left\{ \frac{1}{\pi_j h_M} K \left( \frac{x_j - x_i}{h_M} \right) \right\}_{j \in s}. \quad (8)$$

A design-based sample estimator of  $\mu_i$  is then given by

$$\hat{\mu}_i^o = \mathbf{e}'_1 (\mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si})^{-1} \mathbf{X}'_{si} \mathbf{W}_{si} \hat{\mathbf{t}}_s = \mathbf{w}'_{si} \hat{\mathbf{t}}_s, \quad (9)$$

as long as  $\mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si}$  is invertible. This estimator differs from traditional local polynomial regression because of the inclusion of the sampling weights and the fact that the cluster totals are estimated, not observed. In a design-based context, these adjustments imply that  $\hat{\mu}_i^o$  is an estimator of  $\mu_i$ , the population fit, but not an estimator of  $\mu(x_i)$ , the model mean at  $x_i$ .

Substituting  $\hat{t}_i$  and  $\hat{\mu}_i^o$  respectively for  $t_i$  and  $\mu_i$  in (5), we have the proposed local polynomial regression estimator for the population total of  $y$

$$\tilde{t}_y^o = \sum_{i \in s} \frac{\hat{t}_i - \hat{\mu}_i^o}{\pi_i} + \sum_{i \in C} \hat{\mu}_i^o. \quad (10)$$

In theory, the estimator (9) can be undefined for some  $i \in C$  even if the population estimator in (4) is well-defined, so that estimator (10) cannot be computed. As in Breidt and Opsomer (2000), we will consider an adjusted sample estimator for the theoretical derivations in Section 2. The adjusted sample estimator for  $\mu_i$  is given by

$$\hat{\mu}_i = \mathbf{e}'_1 \left( \mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si} + \text{diag} \left\{ \frac{\delta}{M^2} \right\}_{j=1}^{q+1} \right)^{-1} \mathbf{X}'_{si} \mathbf{W}_{si} \hat{\mathbf{t}}_s = \mathbf{w}'_{si} \hat{\mathbf{t}}_s, \quad (11)$$

for some small  $\delta > 0$ . The value  $\delta M^{-2}$  in (11) is a small order adjustment that guarantees the estimator's existence for any  $s \subset C$ , as long as the population estimator in (4) is defined for all  $i \in C$ . This adjustment was also used by Fan (1992) in the study of the theoretical properties of local polynomial regression. We let

$$\tilde{t}_y = \sum_{i \in s} \frac{\hat{t}_i - \hat{\mu}_i}{\pi_i} + \sum_{i \in C} \hat{\mu}_i \quad (12)$$

denote the local polynomial regression estimator that uses the adjusted sample estimator in (11). The estimator for single-stage cluster sampling is obtained if we set  $\hat{t}_i = t_i$  for all  $i \in C$ .

As noted in Section 1, modeling the cluster totals as in (3) is often reasonable in practice, because many of the auxiliary variables  $x_i$  available for regression estimation are related to the size of the clusters, and hence are good predictors for the  $t_i$ . Another possible way to construct regression estimators is to model the cluster means as

$$N_i^{-1}t_i = \alpha(x_i) + \varepsilon_i, \quad (13)$$

where the  $\varepsilon_i$  are now independent random variables with mean zero and variance  $\nu_i$ ,  $\alpha(x)$  is smooth, and  $\nu_i$  is strictly positive. If the cluster sizes  $N_i$  are (approximately) equal, models (3) and (13) are essentially equivalent. If the  $N_i$  vary significantly between clusters, however, the interpretation of the two models differs.

It is useful to consider element-level models that lead to the two different specifications, noting that an explicit element-level model is *not* a requirement for the methodology we describe. To obtain cluster total model (3), we might consider the element-level model

$$y_k = \frac{\mu(x_i)}{N_i} + a_i + \eta_k,$$

where  $\{a_i\}$  are iid  $(0, \sigma_a^2)$  random effects for clusters and  $\{\eta_k\}$  are iid  $(0, \sigma^2)$  random errors, independent of the  $\{a_i\}$ . Then

$$t_i = \mu(x_i) + N_i a_i + \sum_{k \in U_i} \eta_k = \mu(x_i) + \epsilon_i,$$

where  $\nu_i = \text{Var}(\epsilon_i) = N_i^2 \sigma_a^2 + N_i \sigma^2$ . On the other hand, to obtain the cluster mean model (13), we might consider the element-level model

$$y_k = \alpha(x_i) + a_i + \eta_k,$$

leading to

$$t_i = N_i\alpha(x_i) + N_ia_i + \sum_{k \in U_i} \eta_k = N_i\alpha(x_i) + \epsilon_i.$$

In these two cases, the variance structure is the same, but the mean function depends explicitly on  $\{N_i\}$  in the cluster mean case.

If a regression estimator based on model (13) is preferred, it is still possible to construct a nonparametric regression estimator. The  $\hat{\mu}_i$  in (12) would be replaced by  $N_i\hat{\alpha}_i$ , where the  $\hat{\alpha}_i$  are obtained via nonparametric regression of  $N_i^{-1}\hat{t}_i$  on  $x_i$ , using the local design matrix (7) and local weighting matrix (8). Note that this approach requires that  $\{N_i\}$  are known for all  $i \in C$ . Much of the discussion in this paper carries over to this case, with suitable assumptions on the cluster sizes  $N_i$ .

To derive the statistical properties of  $\tilde{t}_y$ , we adopt the same asymptotic framework as in Breidt and Opsomer (2000) but extend it to two-stage designs. We let the population number of clusters  $M$  and the sample number of clusters  $m$  tend to infinity. The number of elements within each cluster,  $N_i$ , remains bounded, so that no cluster dominates the population. Subsampling within selected clusters is carried out as in Section 1.1. The remaining technical assumptions about the sampling design, the population and the nonparametric regression method are given in the Appendix.

## 2 Properties

### 2.1 Weighting and Calibration

The nonparametric regression estimator can be expressed as a linear combination of the study variables, with weights that do not depend on the study variables. These weights are extremely useful in practice. From (11) and (12), note that

$$\begin{aligned} \tilde{t}_y &= \sum_{i \in s} \left\{ \frac{1}{\pi_i} + \sum_{j \in C} \left( 1 - \frac{I_j}{\pi_j} \right) \mathbf{w}'_{sj} \mathbf{e}_i \right\} \hat{t}_i \\ &= \sum_{i \in s} \omega_{is} \hat{t}_i \\ &= \sum_{i \in s} \sum_{k \in s_i} \frac{\omega_{is}}{\pi_{k|i}} y_k. \end{aligned} \tag{14}$$

Thus,  $\tilde{t}_y$  is a linear combination of the  $\hat{t}_i$ 's in  $s$ , with cluster weights  $\{\omega_{is}\}$  that are the sampling weights of clusters, suitably modified to reflect auxiliary information  $[x_i]_{i \in C}$ .

Alternatively,  $\tilde{t}_y$  is a linear combination of the  $y_k$ 's in  $\cup_{i \in s} s_i$ , with element weights  $\{\omega_{is} \pi_{k|i}^{-1}\}$  which reflect both the design and the auxiliary information. Because both sets of weights are independent of the study variables, they can be applied to any study variable of interest. In particular, the weights  $\omega_{is}$  could be applied to  $x_i^\ell$ . If  $\delta = 0$ , then  $\omega_{is} = \omega_{is}^o$  and

$$\sum_{i \in s} \omega_{is}^o x_i^\ell = \sum_{i \in C} x_i^\ell$$

for  $\ell = 0, 1, \dots, q$ . That is, the weights are exactly *calibrated* to the  $q + 1$  known control totals  $M, t_x, \dots, t_{x^q}$ . If  $\mu(x_i)$  is exactly a  $q$ th degree polynomial, then the unconditional expectation (with respect to design and model) of  $\tilde{t}_y^o - t_y$  is exactly zero. If  $\delta \neq 0$ , then this calibration property holds approximately.

## 2.2 Asymptotic Properties

In general, the local polynomial regression estimator  $\tilde{t}_y$  is not design-unbiased because the  $\hat{\mu}_i$  are nonlinear functions of design-unbiased estimators. However,  $\tilde{t}_y$  is asymptotically design-unbiased and design consistent under mild conditions. We state the theorems without proofs here. The technical derivations are extensions of those in Breidt and Opsomer (2000), suitably modified to handle the additional stage of sampling. Complete details can be found in Kim (2004).

**Theorem 1** *In two-stage element sampling, and under A1–A9 given in the Appendix, the local polynomial regression estimator*

$$\tilde{t}_y = \sum_{i \in C} \left\{ (\hat{t}_i - \hat{\mu}_i) \frac{I_i}{\pi_i} + \hat{\mu}_i \right\}$$

*is asymptotically design-unbiased (ADU) in the sense that*

$$\lim_{M \rightarrow \infty} E_p \left[ \frac{\tilde{t}_y - t_y}{M} \right] = 0 \quad \text{with } \xi\text{-probability one,}$$

*and is design consistent in the sense that*

$$\lim_{M \rightarrow \infty} E_p \left[ I_{\{|\tilde{t}_y - t_y| > M\eta\}} \right] = 0 \quad \text{with } \xi\text{-probability one}$$

*for all  $\eta > 0$ .*

Under the same conditions as in Theorem 1, we obtain the asymptotic design mean squared error of the local polynomial regression estimator  $\tilde{t}_y$  in two-stage element sampling. The asymptotic design mean squared error consists of first- and second-stage variance components, and is equivalent to the variance of the generalized difference estimator, given in (6). As noted after equation (6) above, the second-stage variance is unaffected by the regression estimation at the cluster level.

**Theorem 2** *In two-stage element sampling, and under A1–A9 given in the Appendix,*

$$mE_p \left( \frac{\tilde{t}_y - t_y}{M} \right)^2 = \frac{m}{M^2} \sum_{i,j \in C} (t_i - \mu_i)(t_j - \mu_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} + \frac{m}{M^2} \sum_{i \in C} \frac{V_i}{\pi_i} + o(1)$$

with  $\xi$ -probability one.

The next result shows that the asymptotic design mean squared error can be estimated consistently under mild assumptions.

**Theorem 3** *In two-stage element sampling, and under A1–A9 given in the Appendix,*

$$\lim_{M \rightarrow \infty} mE_p \left| \hat{V}(M^{-1}\tilde{t}_y) - AMSE(M^{-1}\tilde{t}_y) \right| = 0$$

with  $\xi$ -probability one, where

$$\hat{V}(M^{-1}\tilde{t}_y) = \frac{1}{M^2} \sum_{i,j \in C} (\hat{t}_i - \hat{\mu}_i)(\hat{t}_j - \hat{\mu}_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}} + \frac{1}{M^2} \sum_{i \in C} \hat{V}_i \frac{I_i}{\pi_i}$$

and

$$AMSE(M^{-1}\tilde{t}_y) = \frac{1}{M^2} \sum_{i,j \in C} (t_i - \mu_i)(t_j - \mu_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} + \frac{1}{M^2} \sum_{i \in C} \frac{V_i}{\pi_i}.$$

Therefore,  $\hat{V}(M^{-1}\tilde{t}_y)$  is asymptotically design-unbiased and design consistent for  $AMSE(M^{-1}\tilde{t}_y)$ .

Using the weighted residual technique (Särndal *et al.*, 1989), we could construct an alternative variance estimator with the local polynomial regression weights  $\omega_{is}$ ,

$$\hat{V}_w(M^{-1}\tilde{t}_y) = \frac{1}{M^2} \sum_{i,j \in s} \omega_{is}(\hat{t}_i - \hat{\mu}_i)\omega_{js}(\hat{t}_j - \hat{\mu}_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} + \frac{1}{M^2} \sum_{i \in s} \pi_i \omega_{is}^2 \hat{V}_i.$$

Analogous results for the generalized regression estimator are given in Result 8.4.1 of Särndal *et al.* (1992).

### 3 Simulation Experiments

We performed simulation experiments following Breidt and Opsomer (2000) to compare the performance of the local linear model-assisted regression estimator in two-stage element sampling (equation (12) with  $q = 1$ , denoted in what follows as LLR) with that of other parametric and nonparametric estimators:

REG parametric regression Särndal et al., (1992, ch. 6)

PS poststratification Cochran (1977, p. 134)

DORF local linear model-based adapted from Dorfman (1992)

The first two estimators are parametric estimators (corresponding to a polynomial mean function for REG and a piecewise constant mean function for PS). The DORF estimator is a model-based nonparametric survey regression estimator, in contrast to the model-assisted LLR estimator. It is constructed by computing an unweighted local linear regression fit for the sample (say,  $\hat{m}_{i,\text{unw}}$ ), and then using this fitted model to predict the unobserved elements in the population:

$$\hat{t}_{y,\text{DORF}} = \sum_{i \in s} \hat{t}_i + \sum_{i \in U \setminus s} \hat{m}_{i,\text{unw}}.$$

In both LLR and DORF, we use the Epanechnikov kernel,  $K(t) = 0.75(1 - t^2)I_{\{|t| \leq 1\}}$ .

All of the estimators are chosen so that they have approximately the same degrees of freedom (df): the REG estimator is based on a polynomial of degree  $\text{df} - 1$ ; the PS estimator is based on a division of the  $x$ -range into  $\text{df}$  equally-sized strata; and the bandwidth parameters in the kernel regressions are chosen so that the traces of the population-level smoothing matrices for LLR and DORF are equal to  $\text{df}$  (see Hastie and Tibshirai, 1990, p.52). Degrees of freedom for the sample smoothing matrices in LLR and DORF will vary from sample to sample, but are approximately equal to  $\text{df}$ . Two levels of  $\text{df}$  are considered: 4 and 7.

In the simulation experiments discussed below, we have excluded the classical Horvitz-Thompson and linear regression estimators, as they are based on fewer  $\text{df}$  and hence not directly comparable. In results not reported here (see Kim 2004), both estimators perform poorly relative to the more flexible estimators considered above for most response variables.

We have also excluded the model-based estimator due to Zheng and Little (2004), which fits a penalized spline for cluster means with additional random effects to ac-

count for within-cluster correlation. The penalty parameter in the spline can be automatically selected by treating the coefficients of the spline basis functions as random effects in a linear mixed model fitted to the study variable of interest. Full implementation of the Zheng and Little estimator thus requires a separate restricted maximum likelihood (REML) fit for each study variable in each replication, to estimate variance components in the linear mixed model. Even if the variance components were fixed at some values not specific to any particular study variable, the degrees of freedom for the resulting fit would not be comparable to the methods considered here, since none of these estimators include df for cluster effects. For these reasons of complexity and lack of comparability, we have not included the Zheng and Little (2004) estimator in this simulation.

We consider eight mean functions  $m_k(x)$ . The first is a constant and the remaining functions are normalized to have minimum value of 0 and maximum value of 2:

$$m_k(x) = 2 \frac{g_k(x) - \min_{x \in [0,1]} g_k(x)}{\max_{x \in [0,1]} g_k(x) - \min_{x \in [0,1]} g_k(x)}$$

where  $x \in [0, 1]$  and

$$\begin{aligned} \text{constant: } & m_0(x) = 8, \\ \text{linear: } & g_1(x) = 1 + 2(x - 0.5), \\ \text{quadratic: } & g_2(x) = 1 + 2(x - 0.5)^2, \\ \text{bump: } & g_3(x) = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2), \\ \text{jump: } & g_4(x) = \{1 + 2(x - 0.5)I_{\{x \leq 0.65\}}\} + 0.65I_{\{x > 0.65\}}, \\ \text{exponential: } & g_5(x) = \exp(-8x), \\ \text{cycle1: } & g_6(x) = 2 + \sin(2\pi x), \\ \text{cycle4: } & g_7(x) = 2 + \sin(8\pi x). \end{aligned}$$

These represent a range of correct and incorrect model specifications for the various estimators considered. For  $m_0$ ,  $m_1$  and  $m_2$ , the models are polynomial; the remaining mean functions represent various departures from the polynomial model. The function  $m_3$  is linear over most of its range, except for a “bump” present for a small portion of the range of  $x_k$ . The mean function  $m_4$  is not smooth. The function  $m_5$  is an exponential curve. The function  $m_6$  is a sinusoid completing one full cycle on  $[0, 1]$ , while  $m_7$  completes four full cycles. The population  $x_k$  are generated as independent and identically distributed (iid) uniform(0,1) random variables.

The population values  $y_{hk}$  ( $h = 0, \dots, 7$ ) are generated from the mean functions by

$$y_{hk} = \frac{m_h(x_i)}{N_i} + a_i + \eta_k, \text{ for } k \in U_i,$$

so that

$$t_{hi} = m_h(x_i) + N_i a_i + \sum_{k \in U_i} \eta_k = m_h(x_i) + \epsilon_i,$$

where  $\{a_i\}$  are iid  $N(0, \rho\sigma^2)$  cluster effects and  $\{\eta_k\}$  are iid  $N(0, (1-\rho)\sigma^2)$  errors, independent of the  $\{a_i\}$ . Elements within clusters then have variance  $\sigma^2$  and correlation  $\rho$ . We consider  $\sigma = 0.08$  and  $\rho = 0$  or  $\rho = 0.5$ .

The population is of size  $M = 1000$  clusters, each consisting of a random number of elements ( $\{N_i - 2\}$  iid Poisson(3)). Samples are generated from one of two designs. Both use simple random sampling of two elements per cluster in stage two. One design is self-weighting, consisting of with-replacement sampling of 50 clusters with probability proportional to  $N_i$ . The other design is non-self-weighting, consisting of simple random sampling without replacement of 50 clusters.

For each combination of mean function, correlation, degrees of freedom, and design, 1000 replicate samples are selected and the estimators are calculated. For each sample, a single set of weights corresponding to each estimator (e.g., using (15) for LLR weights) is computed and applied to all eight study variables, as would be common practice in applications. As the population is kept fixed during these 1000 replicates, we are able to evaluate the design-averaged performance of the estimators, including the design bias, design variance and design mean squared error. For nearly all cases in this simulation, the absolute relative design biases  $|\mathbb{E}_p \hat{t}_y - t_y|/t_y$  were less than two percent for all estimators, and are not tabled; exceptions occurred for the DORF estimator, which is not design-unbiased, even asymptotically.

Table 1 shows the ratios of MSE's for the various estimators to the MSE for the local linear regression model-assisted estimator (LLR). The performance of all estimators is increasingly similar as the within-cluster correlation increases. (Similarly, in simulations not shown here, the performance of all estimators is increasingly similar as the model variance increases.)

In 160 of the 192 tabled cases, LLR is competitive or better than the other estimators (MSE ratio  $\geq 0.95$ ). LLR is always competitive and sometimes much better than PS by this criterion.

	df	$\rho$	Self-Weighting			Non-Self-Weighting		
			REG	PS	DORF	REG	PS	DORF
const	4	0.0	1.01	0.99	0.93	1.01	1.00	1.06
	4	0.5	1.01	1.01	0.93	1.01	1.01	1.05
	7	0.0	1.27	1.78	0.88	1.60	4.56	1.05
	7	0.5	1.16	1.55	0.94	1.16	3.60	1.04
ratio	4	0.0	1.01	1.22	0.97	1.01	1.24	1.08
	4	0.5	1.01	1.23	0.94	1.01	1.25	1.07
	7	0.0	1.34	0.98	0.92	1.65	1.11	1.04
	7	0.5	1.20	0.99	0.95	1.18	1.05	1.04
quad	4	0.0	0.94	2.08	3.92	0.93	2.15	4.81
	4	0.5	0.94	1.70	2.88	0.97	1.83	3.90
	7	0.0	1.30	1.19	1.03	1.63	1.36	1.27
	7	0.5	1.17	1.15	1.06	1.18	1.24	1.21
bump	4	0.0	1.17	1.33	0.89	1.16	1.35	1.01
	4	0.5	1.11	1.29	0.89	1.10	1.32	1.01
	7	0.0	1.47	0.96	0.88	1.81	1.16	1.01
	7	0.5	1.31	0.97	0.92	1.31	1.06	1.00
jump	4	0.0	1.16	1.26	1.17	1.18	1.22	1.34
	4	0.5	1.12	1.23	1.08	1.12	1.23	1.31
	7	0.0	1.87	1.15	0.91	2.93	1.22	1.01
	7	0.5	1.55	1.09	0.93	1.97	1.13	1.02
expo	4	0.0	0.99	1.52	1.94	0.95	1.59	2.28
	4	0.5	0.98	1.30	1.61	0.96	1.38	1.96
	7	0.0	1.29	1.09	1.05	1.62	1.28	1.31
	7	0.5	1.17	1.10	1.09	1.16	1.17	1.23
cycle1	4	0.0	0.72	1.66	0.79	0.77	1.53	0.89
	4	0.5	0.82	1.44	0.82	0.84	1.40	0.93
	7	0.0	1.28	1.24	0.86	1.57	1.33	0.98
	7	0.5	1.17	1.18	0.92	1.16	1.23	1.01
cycle4	4	0.0	0.99	1.04	0.85	0.95	1.07	0.94
	4	0.5	0.99	1.04	0.87	0.95	1.05	0.95
	7	0.0	2.44	1.22	0.83	4.00	1.25	0.94
	7	0.5	2.20	1.15	0.86	3.28	1.21	0.92

Table 1: Mean square error ratios greater than one favor the local linear regression model-assisted estimator (LLR). Based on 1000 replicate samples of a self-weighting design (with-replacement sampling of 50 clusters with probability proportional to  $N_i$ , followed by simple random sampling without-replacement of two elements per cluster) or a non-self-weighting design (simple random sampling without replacement of 50 clusters, followed by simple random sampling without-replacement of two elements per cluster) from a fixed population of size  $M = 1000$  clusters. All four estimators use the same degrees of freedom (df) for fitting. LLR and DORF estimators are computed with an Epanechnikov kernel and bandwidth  $h = 0.2943$  when  $df= 4$  and  $h = 0.135$  when  $df= 7$ .

LLR is clearly less efficient than REG with 4 df for `cycle1` under both designs: a one-cycle sinusoid is fitted extremely well with a cubic polynomial. In each of these cases, however, the LLR estimator with more df is better than REG. With the exception of `cycle1`, LLR is never much worse and sometimes much better than REG, and is the preferred estimator in this simulation.

The behavior of DORF is quite erratic in this simulation. It accounts for most (25/32) of the MSE ratios less than 0.95, with its best performance relative to LLR yielding an MSE ratio of 0.79. But its worst performance is in fact the worst overall MSE ratio (4.81), and it accounts for three of the worst five performances.

One common concern when using nonparametric regression techniques is how sensitive the results are to the choice of the smoothing parameter. Clearly, the df (and hence the bandwidth parameter) has an effect on the MSE of LLR, but Table 1 suggests that gains in efficiency over other estimators can be obtained for a variety of choices of df, even when df is chosen by default (without any input from the user or the data). That is, for many reasonable choices of df, the LLR estimator is expected to have performance as good or better than the estimators considered here, for a broad range of response variables and study designs. LLR can therefore be particularly useful in the context of large-scale survey sampling, in which the same set of regression weights (with a single choice for the smoothing parameter) is often used for a large number of different variables, as was done in the simulation experiment described above.

In any survey application using regression weights, negative weights are clearly undesirable. In the simulation above, there are  $(2 \text{ designs}) \times (2 \text{ df}) \times (1000 \text{ replications}) \times (50 \text{ weights}) = 200,000$  weights generated for each estimator. Among these, there were 1472 negative REG weights and 859 negative LLR weights. In practice, negative survey weights are often a result of inappropriate model selection. As nonparametric models are less likely to result in a severely misspecified population model, they tend to result in less frequent occurrences of negative weights, as noted in Breidt, Claeskens and Opsomer (2005). LLR thus tends to have better behavior than REG but, in regions with very low sampling density, LLR suffers from highly variable fits, which can occasionally lead to negative regression weights.

A problem shared by all the estimators considered here is the negative finite-

sample bias of the standard design-based variance estimator defined in Theorem 2.2. We found that the coverage of nominal 95% confidence intervals using the standard residual technique averages 90% for REG, PS, and LLR at the sample size of 50, for both the self-weighting and non-self-weighting designs. We do not report the details here.

## 4 Example: Erosion study

We apply local polynomial regression estimation to data from the 1995 National Resources Inventory Erosion Update Study (see Breidt and Fuller, 1999). The National Resources Inventory (NRI) is a stratified two-stage area sample of the agricultural lands in the United States conducted by the Natural Resources Conservation Service (NRCS) of the U.S. Department of Agriculture (Breidt, 2001). The 1995 Erosion Update Study was a smaller-scale study using NRI information as frame material.

In the 1995 study, first-stage sampling strata were 14 states in the Midwest and Great Plains regions and primary sampling units (PSUs) were counties within states. A categorical variable was used for within-county stratification in second-stage sampling. Second-stage sampling units (SSUs) were NRI segments of land, 160 acres in size. The auxiliary variable for each county was  $x_i$ , the square root of a size measure of land with erosion potential. (We used square root to reduce the sparseness of points in the regressor space.) The variables of interest were two kinds of erosion measurements, characterized as wind erosion (WEQ) and water erosion (USLE). In this application, model (3) used to construct regression estimators relates the total county erosion (of either type) to the sized-based auxiliary variable  $x_i$ , corresponding to the estimation situation targeted by our approach. As noted earlier, this model does not require the specification of a corresponding within-cluster (segment-level) model.

The auxiliary variable  $x_i$  was also used at the design stage. At stage one, a sample of 213 counties was selected by stratified sampling from the population of 1357 counties, with probability proportional to  $x_i^2$ . Subsamples of NRI segments within the selected counties were selected by stratified unequal probability sampling at stage two. In total, 1900 NRI segments were selected.

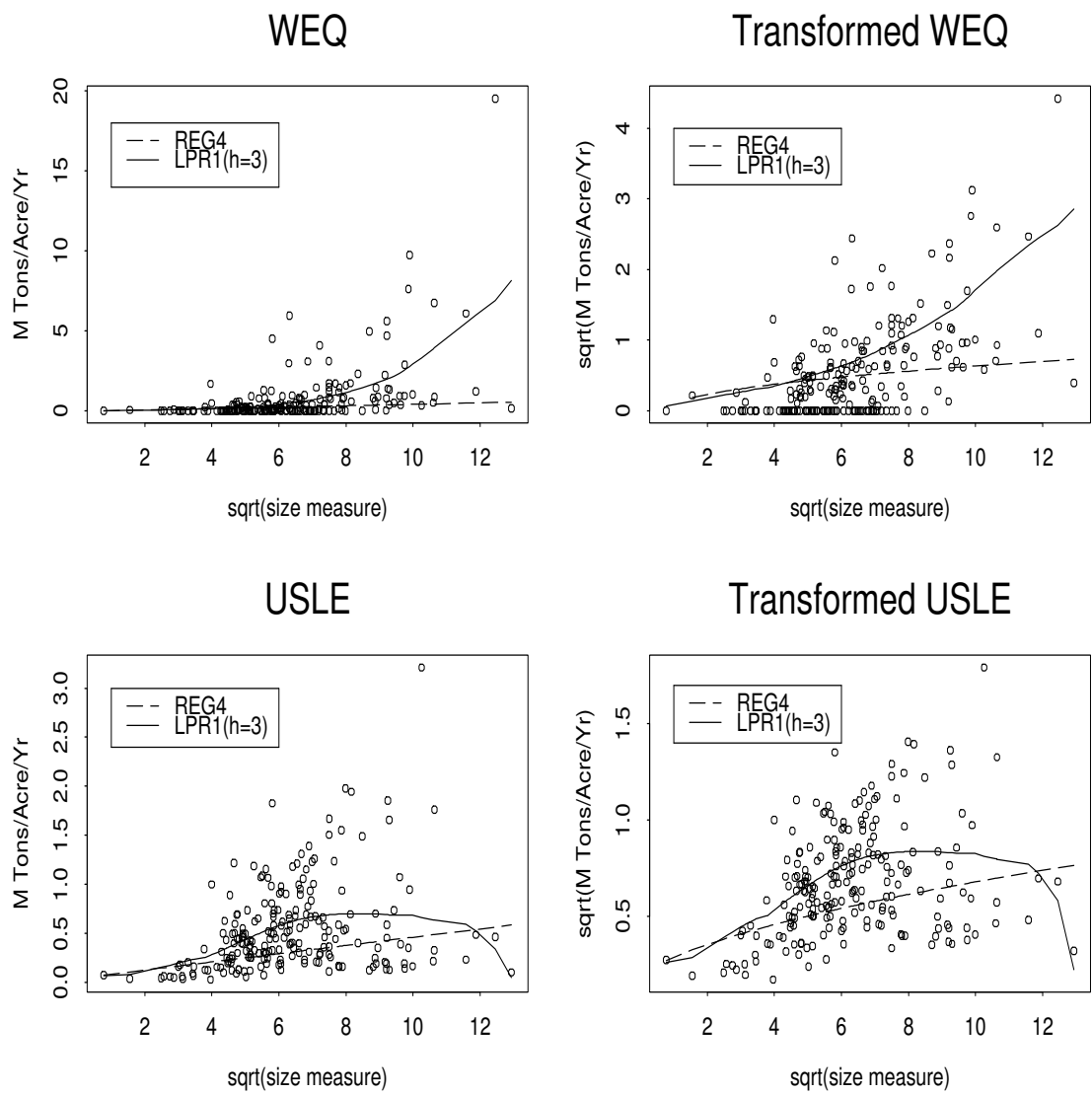


Figure 1: Relationship between  $x_i =$  square root of size measure of land with erosion potential and square root of estimated county total ( $\hat{t}_i$ ) in selected counties at stage one for wind erosion (WEQ) and water erosion (USLE), on both original (left column) and square root (right column) vertical scales. Dashed curve is weighted linear regression fit (REG4) and solid curve is local linear regression fit (LPR1 with  $h = 3$ ).

		WEQ	USLE
HT		443.6 (49.4)	551.5 (31.8)
REG2	$\nu(x) \propto x^2$	442.5 (50.7)	537.8 (26.5)
REG4	$\nu(x) \propto x^4$	442.1 (50.1)	537.7 (26.5)
REG8	$\nu(x) \propto x^8$	441.8 (50.3)	540.1 (27.6)
LPR1	h=1	434.1 (47.5)	529.0 (24.4)
LPR1	h=3	427.4 (48.9)	532.3 (25.3)
LPR1	h=5	430.5 (48.7)	541.2 (27.6)

Table 2: Horvitz-Thompson (HT), weighted linear regression (REG2, REG4, REG8), and local linear regression (LPR1 with  $h = 1, 3, 5$ ) estimates for wind erosion (WEQ) and water erosion (USLE) totals in millions of tons/acre/year. The numbers in parentheses are estimated standard errors.

The Horvitz-Thompson (HT), linear regression (REG), and local linear regression (LPR1) estimates for WEQ and USLE totals and the corresponding variance estimates were calculated from the sample. We calculated REG estimates with three different variances of the errors ( $\nu(x) \propto x^2, x^4, \text{ and } x^8$ ), denoted by REG2, REG4, and REG8 respectively. Weighted regressions were used because the data displayed large amounts of heteroskedasticity (see Figure 1), which can have an effect on the parametric fit. The Epanechnikov kernel with three different bandwidths ( $h = 1, 3, \text{ and } 5$ ) was used for the LPR1 estimator. The smallest allowable bandwidth for this example is (to the nearest tenth)  $h = 1$ .

Table 2 shows HT, REG and LPR1 estimates of WEQ and USLE totals and estimated standard errors. For each estimator, calibrated survey weights were constructed and applied to both study variables. Standard errors were estimated by assuming unequal-probability with-replacement sampling within design strata at stage one, and unequal-probability with-replacement sampling within clusters at stage two. Using the estimated standard errors as a guide, LPR1 with  $h = 1$  performs best among all estimates and REG4 is best among REG estimates. The local linear re-

gression fit with  $h = 1$  is quite rough because of data sparseness in some parts of the range of the  $x_i$ , so we use  $h = 3$  for further comparison. Except in relatively large bandwidths (e.g. LPR1 with  $h = 5$  here), LPR1 estimates provide a modest improvement over HT and REG estimates on the basis of estimated standard errors for both WEQ and USLE. Note that this efficiency gain is essentially “free,” since it is based on the same auxiliary information as the REG estimator and can be implemented in the same model-assisted framework.

To further evaluate the appropriateness of the superpopulation models, Figure 1 shows the relationship between  $x_i =$  square root of size measure of land with erosion potential and estimated county total ( $\hat{t}_i$ ) in selected counties at stage one for WEQ and USLE, on both the original and square-root transformed vertical scales. In all plots, the weighted linear regression fit with variance proportional to  $x^4$  (REG4) and the local linear regression fit with bandwidth  $h = 3$  (LPR1) are included. (The square root transformation in the figure is included to make differences in the fits more discernible.) The LPR1 fit appears quite sensible in those plots. It is at least competitive with the REG estimators, if not better, but requires neither mean nor variance function specification. The same weights used for WEQ and USLE could be applied to any other study variables obtained in the Erosion Update Study, with efficiency increases over HT if the county-level total for that variable is dependent on the erosion potential size measure, and with efficiency increases over REG if the dependence is non-linear.

## 5 Acknowledgments

The research reported here was supported in part by National Science Foundation grants DMS-0204642 and DMS-0204531 and by Cooperative Agreement 68-3A75-075 between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University.

## A Appendix: Assumptions

Assumptions A1–A7 below extend those in Breidt and Opsomer (2000) to the two-stage case. Assumptions A8 and A9 provide some additional regularity conditions on the cluster sizes and the second-stage design.

- **A1** *Distribution of the errors under  $\xi$ : the errors  $\varepsilon_i$  are independent and have mean zero, variance  $\nu_i > 0$ , and compact support, uniformly for all  $M$ .*
- **A2** *For each  $M$ , the  $x_i$  are considered fixed with respect to the superpopulation model  $\xi$ . The  $x_i$  are independent and identically distributed  $F(x) = \int_{-\infty}^x f(t)dt$ , where  $f(\cdot)$  is a density with compact support  $[a_x, b_x]$  and  $f(x) > 0$  for all  $x \in [a_x, b_x]$ .*

- **A3** *The mean function  $\mu$  is continuous on  $[a_x, b_x]$ .*

- **A4** *The kernel  $K(\cdot)$  has compact support  $[-1, 1]$ , is symmetric and continuous, and satisfies*

$$\int_{-1}^1 K(u) du = 1.$$

- **A5** *First-stage sampling rate  $mM^{-1}$ , bandwidth  $h_M$ : as  $M \rightarrow \infty$ ,  $mM^{-1} \rightarrow \pi \in (0, 1)$ ,  $h_M \rightarrow 0$ ,  $Mh_M^2/(\log \log M) \rightarrow \infty$ .*

- **A6** *First-stage (cluster) inclusion probabilities  $\pi_i$  and  $\pi_{ij}$ :*

*for all  $M$ ,  $\min_{i \in C} \pi_i \geq \lambda > 0$ ,  $\min_{i, j \in C} \pi_{ij} \geq \lambda^* > 0$  and*

$$\limsup_{M \rightarrow \infty} m \max_{i, j \in C: i \neq j} |\pi_{ij} - \pi_i \pi_j| < \infty.$$

- **A7** *Additional assumptions involving higher-order first-stage inclusion probabilities:*

$$\lim_{M \rightarrow \infty} m^2 \max_{(i_1, i_2, i_3, i_4) \in D_{4, M}} \left| E_I [(I_{i_1} - \pi_{i_1})(I_{i_2} - \pi_{i_2})(I_{i_3} - \pi_{i_3})(I_{i_4} - \pi_{i_4})] \right| < \infty,$$

where  $D_{t, M}$  denotes the set of all distinct  $t$ -tuples  $(i_1, i_2, \dots, i_t)$  from  $C$ ,

$$\lim_{M \rightarrow \infty} \max_{(i_1, i_2, i_3, i_4) \in D_{4, M}} \left| E_I [(I_{i_1} I_{i_2} - \pi_{i_1 i_2})(I_{i_3} I_{i_4} - \pi_{i_3 i_4})] \right| = 0,$$

$$\limsup_{M \rightarrow \infty} m \max_{(i_1, i_2, i_3) \in D_{3, M}} \left| E_I \left[ (I_{i_1} - \pi_{i_1})^2 (I_{i_2} - \pi_{i_2}) (I_{i_3} - \pi_{i_3}) \right] \right| < \infty,$$

and

$$\limsup_{M \rightarrow \infty} m^2 \max_{(i_1, i_2, i_3) \in D_{3, M}} \left| E_I \left[ (I_{i_1} - \pi_{i_1}) (I_{i_2} - \pi_{i_2}) (I_{i_3} - \pi_{i_3}) \right] \right| < \infty.$$

- **A8** The cluster sizes  $N_i$  are uniformly bounded above for all clusters and for all  $M$ .
- **A9** The second-stage design is invariant and independent, with  $n_i \geq 1$  for every  $i \in s$  and for every possible first-stage sample  $s$ . Further, the second-stage inclusion probabilities are uniformly bounded away from zero for all clusters and all  $M$ , and the second-stage joint inclusion probabilities are uniformly bounded away from zero for all clusters and all  $M$ .

## References

- [1] Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika* **92**, 831–846.
- [2] Breidt, F.J. and Fuller, W.A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological, and Environmental Statistics* **4**, 391–403.
- [3] Breidt, F.J. and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics* **28**, 1026–1053.
- [4] Breidt, F.J. (2001). The National Resources Inventory (NRI), US. *Encyclopedia of Environmetrics*, vol.3, pages 1353–1356. A.H. El-Shaarawi and W.W. Piegorsch, eds. Wiley.
- [5] Brewer, K.R.W. (1963). Ratio estimation in finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* **5**, 93–105.

- [6] Cassel, C.M., Särndal, C.E., and Wretman, J. H. (1976). Some results on generalized different estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–620.
- [7] Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* **88**, 268–277.
- [8] Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.
- [9] Dorfman, A.H. (1992). Nonparametric regression for estimating totals in finite populations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 622–625.
- [10] Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association* **87**, 998–1004.
- [11] Holt, D. and Smith, T.M. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A* **142**, 33–46.
- [12] Horvitz, D.G. and D.J. Thompson. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.
- [13] Kim, J.-Y. (2004). Nonparametric Regression Estimation in Survey Sampling. Ph.D. thesis, Iowa State University.
- [14] Robinson, P.M. and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimation in probability sampling. *Sankhyā: The Indian Journal of Statistics, Series B* **45**, 240–248.
- [15] Royall, R.M. (1970). On finite population sampling under certain linear regression models. *Biometrika* **57**, 377–387.
- [16] Särndal, C.-E., Swensson, B., and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika* **76**, 527–537.

- [17] Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer, New York.
- [18] Särndal, C.E. (1980). On  $\pi$ -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* **67**, 639–650.
- [19] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, Chapman and Hall, London.
- [20] Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* **96**, 185–193.
- [21] Zheng, H. and Little, R.J. (2003). Penalized Spline Model-Based Estimation of the Finite Population Total from Probability-Proportional-To-Size Samples. *Journal of Official Statistics* **19**, 99–117.
- [22] Zheng, H. and Little, R.J. (2004). Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples. *Survey Methodology* **30**, 209–218.