LOCAL POLYNOMIAL REGRESSION ESTIMATION IN TWO-STAGE SAMPLING

Ji-Yeon Kim, F. Jay Breidt, Jean D. Opsomer, Iowa State University
Jean D. Opsomer, Iowa State University, Ames, IA 50011, U.S.A.

Key Words: Calibration, cluster sampling, National Resources Inventory, nonparametric regression

Abstract:

We consider local polynomial regression estimation for finite population totals in two-stage element sampling. The estimators are linear combinations of estimators of cluster totals with weights that are calibrated to known control totals. The estimators are asymptotically design-unbiased and consistent under mild assumptions. We provide a consistent estimator for the design mean squared error of the local polynomial regression estimators. Simulation results show that the estimators are more efficient than Horvitz-Thompson and linear regression estimators when the mean function of the superpopulation model is non-linear while being nearly as efficient when the model is linear. The estimation approach performs well in an example using data from a 1995 study associated with the National Resources Inventory.

1 Introduction

To improve the efficiency of surveys, auxiliary information may be used in the sampling design or the estimation of parameters. In this paper we consider use of auxiliary information in estimation.

Many kinds of estimators have been proposed for estimating a finite population total under a superpopulation model ξ describing the relationship between the variable of interest and the auxiliary variables. Often, a linear model is selected as a superpopulation model. Ratio estimators, regression estimators, and poststratification estimators can be derived from an assumed linear model.

Estimators are sought which have good efficiency if the model is true, but maintain desirable properties like asymptotic design unbiasedness and design consistency if the model is false. Because of concerns about the performance of the estimators under model misspecification, some researchers have considered nonparametric models for ξ. Dorfman (1992) and Chambers, Dorfman, and Wehrly (1993) developed model-based nonparametric estimators using this approach. Breidt and Opsomer (2000) proposed a type of model-assisted nonparametric regression estimator for the finite population total, based on local polynomial smoothing. The local polynomial regression estimator has the form of the generalized regression estimator, but is based on a nonparametric superpopulation model applicable to a much larger class of functions.

The local polynomial regression estimator introduced in Breidt and Opsomer (2000) applies only to direct element sampling designs with auxiliary information available for all elements of the population. In many large-scale sample surveys, however, more complex survey designs such as multistage sampling designs or multiphase sampling designs with various types of auxiliary information are commonly used. In this paper, we consider the extension of local polynomial 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.

Often, two-stage sampling is used because an adequate frame of elements is not available, but a listing of clusters is available. In this case, it is not likely that detailed auxiliary information would be available for all population elements. Therefore, we consider local polynomial regression estimation in two-stage element sampling with auxiliary information available for all clusters. Results for single-stage cluster sampling, in which each sampled cluster is completely enumerated, are obtained as a special case.

In Section 2 we propose the local polynomial regression estimator in two-stage element sampling. Desirable design properties of the estimator are described in Section 3. Section 3.1 shows that the estimator is a linear combination of estimators of cluster totals with weights that are calibrated to known control totals. Section 3.2 provides asymptotic design unbiasedness and consistency of the estimator, an approximation to the estimator's mean squared error, and a consistent estimator of the mean squared error. Section 4 gives simulation results for the esti-

mator, comparing its performance with that of the Horvitz-Thompson and the linear regression estimators. We apply the estimator to data from a 1995 study of erosion, using National Resources Inventory (NRI) data as frame materials, in Section 5.

2 Local Polynomial Regression Estimator

Consider a finite population of elements $U = \{1, \ldots, k, \ldots, N\}$ partitioned into M clusters, $U_1, \ldots, U_i, \ldots, U_M$. The population of clusters is denoted $C = \{1, \ldots, i, \ldots, M\}$. The number of elements in the ith 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 $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{Gi})'$ 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.

For every 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 ith 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. *Invariance* of the second-stage design means that for every i, and for every $s \ni i$, $p_i(\cdot|s) = p_i(\cdot)$. That is, the same within-cluster design is used whenever the ith cluster is selected, regardless of what other clusters are selected. *Independence* of the second-stage design means that subsampling in a given cluster is independent of subsampling in any other cluster.

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 ith cluster total.

Let $I_i = 1$ if $i \in s$ and $I_i = 0$ otherwise. Note that $\mathrm{E}_p[I_i] = \mathrm{E}_I[\mathrm{E}_{II}[I_i]] = \mathrm{E}_I[I_i] = \pi_i$, where $\mathrm{E}_p[\cdot]$ denotes expectation with respect to the sampling design, $\mathrm{E}_I[\cdot]$ denotes expectation with respect to stage one, and $\mathrm{E}_{II}[\cdot]$ denotes conditional expectation with respect to stage two given s. Also, $\mathrm{V}_I(\cdot)$ and $\mathrm{V}_{II}(\cdot)$ 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 $\mathrm{E}_p[\hat{t}] = t$.

The Horvitz-Thompson (1952) estimator of t_y in two-stage element sampling is given by

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

where

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

is the Horvitz-Thompson estimator of t_i with respect to stage two. Since $\hat{t}_i$ is design-unbiased for t_i, the Horvitz-Thompson 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 Horvitz-Thompson estimator $\hat{t}_y$ under the sampling design can be written as the sum of two components,

$$\begin{aligned} \mathrm{Var}_p(\hat{t}_y) &= \mathrm{V}_I(\mathrm{E}_{II}[\hat{t}_y]) + \mathrm{E}_I[\mathrm{V}_{II}(\hat{t}_y)] \\ &= \sum_{i,j \in C} (\pi_{ij} - \pi_i \pi_j) \frac{t_i}{\pi_i} \frac{t_j}{\pi_j} + \sum_{i \in C} \frac{V_i}{\pi_i} \end{aligned} \tag{2}$$

where

$$\begin{aligned} V_i &= \mathrm{V}_{II}(\hat{t}_i) \\ &= \sum_{k,l \in U_i} (\pi_{kl|i} - \pi_{k|i}\pi_{l|i}) \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \end{aligned}$$

is the variance of $\hat{t}_i$ with respect to stage two. 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 $V_i = 0$ for all $i \in C$.

The local polynomial regression estimator is motivated by modeling the M points (x_i, t_i) as a realization from an infinite superpopulation model ξ in which

$$t_i = \mu(x_i) + \varepsilon_i,$$

where the ε_i are independent random variables with mean zero and variance $\nu(x_i)$, $\mu(x)$ is a smooth function of x, and $\nu(x)$ is smooth and strictly positive.

Let K denote the kernel function and h_M denote the bandwidth. Let $\boldsymbol{t}_C = [t_i]_{i \in C}$ be the vector of t_i's in the population of clusters. Define the $M \times (q+1)$ matrix

$$\begin{aligned} \boldsymbol{X}_{Ci} &= \begin{bmatrix} 1 & x_1 - x_i & \cdots & (x_1 - x_i)^q \\ \vdots & \vdots & & \vdots \\ 1 & x_M - x_i & \cdots & (x_M - x_i)^q \end{bmatrix} \\ &= \begin{bmatrix} 1 & x_j - x_i & \cdots & (x_j - x_i)^q \end{bmatrix}_{j \in C}, \end{aligned}$$

and define the $M \times M$ matrix

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

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

$$\mu_i = \boldsymbol{e}_1' \left(\boldsymbol{X}_{Ci}' \boldsymbol{W}_{Ci} \boldsymbol{X}_{Ci}\right)^{-1} \boldsymbol{X}_{Ci}' \boldsymbol{W}_{Ci} \boldsymbol{t}_C = \boldsymbol{w}_{Ci}' \boldsymbol{t}_C, \tag{3}$$

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

If these μ_i's were known, then a design-unbiased estimator of t_y would be the generalized difference estimator

$$t_y^* = \sum_{i \in C} \mu_i + \sum_{i \in s} \frac{t_i - \mu_i}{\pi_i} \tag{4}$$

(Särndal, Swensson, and Wretman, 1992, p. 222). The design variance of the estimator,

$$\text{Var}_p\left(t_y^*\right) = \sum_{i,j \in C} (\pi_{ij} - \pi_i \pi_j) \frac{t_i - \mu_i}{\pi_i} \frac{t_j - \mu_j}{\pi_j}, \tag{5}$$

depends on residuals from the nonparametric regression and hence is expected to be smaller than (2).

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

$$\boldsymbol{X}_{si} = \begin{bmatrix} 1 & x_j - x_i & \cdots & (x_j - x_i)^q \end{bmatrix}_{j \in s},$$

and define the $m \times m$ matrix

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

A design-based sample estimator of μ_i is then given by

$$\hat{\mu}_i = \boldsymbol{e}_1' \left(\boldsymbol{X}_{si}' \boldsymbol{W}_{si} \boldsymbol{X}_{si}\right)^{-1} \boldsymbol{X}_{si}' \boldsymbol{W}_{si} \hat{\boldsymbol{t}}_s = \boldsymbol{w}_{si}' \hat{\boldsymbol{t}}_s, \tag{6}$$

as long as $\boldsymbol{X}_{si}' \boldsymbol{W}_{si} \boldsymbol{X}_{si}$ is invertible. Breidt and Opsomer (2000) discuss finite sample adjustments to this estimator that guarantee its existence for any sample $s \subset C$, as long as (3) is well-defined. Substituting $\hat{t}_i$ and $\hat{\mu}_i$ respectively for t_i and μ_i in (4), we have the local polynomial regression estimator for the population total of y,

$$\tilde{t}_y = \sum_{i \in C} \hat{\mu}_i + \sum_{i \in s} \frac{\hat{t}_i - \hat{\mu}_i}{\pi_i}. \tag{7}$$

The estimator for single-stage cluster sampling is obtained if we set $\hat{t}_i = t_i$ for all $i \in C$.

3 Main Results

3.1 Weighting and Calibration

Note from (7) that

$$\begin{aligned} \tilde{t}_y &= \sum_{i \in s} \frac{\hat{t}_i}{\pi_i} + \sum_{j \in C} \left(1 - \frac{I_j}{\pi_j}\right) \boldsymbol{w}_{sj}' \hat{\boldsymbol{t}}_s \\ &= \sum_{i \in s} \left\{\frac{1}{\pi_i} + \sum_{j \in C} \left(1 - \frac{I_j}{\pi_j}\right) \boldsymbol{w}_{sj}' \boldsymbol{e}_i\right\} \hat{t}_i \\ &= \sum_{i \in s} \omega_{is} \hat{t}_i. \end{aligned} \tag{8}$$

Thus, $\tilde{t}_y$ is a linear combination of $\hat{t}_i$'s in s, with weights ω_{is} that are the sampling weights of clusters, suitably modified to reflect the auxiliary information $[x_i]_{i \in C}$. Because the weights are independent of y_k's, they can be applied to any study variable of interest. In particular, they give perfect estimates when applied to the auxiliary variables. It is straightforward to verify that for the local polynomial regression weights ω_{is},

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

for $\ell = 0, 1, \ldots, q$. That is, the weights are exactly *calibrated* to the $q+1$ known control totals $N, t_x, \ldots, t_{x^q}$. If $\mu(x_i)$ is exactly a qth degree polynomial, then the unconditional expectation (with respect to design and model) of $\tilde{t}_y - t_y$ is exactly zero.

3.2 Asymptotic Design Properties

In this section we state without proof some theorems concerning asymptotic design properties of the local polynomial regression estimator. Proofs will be provided elsewhere. We begin with some assumptions. Let the first-stage sample rate $mM^{-1} \to \pi \in (0,1)$, the bandwidth $h_M \to 0$ and $Mh_M^2 \to \infty$ as the population number of clusters $M \to \infty$. The assumptions on $\mu(\cdot)$, $\nu(\cdot)$, and the kernel K are the usual ones in local polynomial kernel smoothing (Wand and Jones, 1994, Chapter 5). For cluster inclusion probabilities π_i and π_{ij} at stage one, we assume that for all M, $\min_{i \in C} \pi_i \geq \lambda > 0$, $\min_{i,j \in C} \pi_{ij} \geq \lambda^* > 0$, and $\limsup_{M \to \infty} m \max_{i,j \in C: i \neq j} |\pi_{ij} - \pi_i \pi_j| < \infty$, with additional assumptions on higher-order inclusion probabilities. We also assume that $\limsup_{M \to \infty} M^{-1} \sum_{i \in C} \text{E}_{II}[(\hat{t}_i - t_i)^4] < \infty$ and $\limsup_{M \to \infty} M^{-1} \sum_{i \in C} \text{E}_{II}[\hat{V}_i^2] < \infty$. These assumptions are reasonable for many two-stage element sampling designs.

In general, the local polynomial regression estimator $\tilde{t}_y$ is not design unbiased because the $\hat{\mu}_i$'s are nonlinear functions of unbiased estimators. However, $\tilde{t}_y$ is asymptotically design unbiased and design consistent.

Theorem 1 *In two-stage element sampling under the above assumptions, 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

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

and is design consistent in the sense that

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

for all $\eta > 0$.

Under the same conditions as in Theorem 1, we obtain the asymptotic mean squared error of the local polynomial regression estimator $\tilde{t}_y$ in two-stage element sampling. The asymptotic mean squared error consists of first and second stage variance components. The first stage variance component is equivalent to the variance of the generalized difference estimator, while the second stage variance is unaffected by the regression estimation.

Theorem 2 *In two-stage element sampling under the above assumptions,*

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

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

Theorem 3 *In two-stage element sampling under the above assumptions,*

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

where

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

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

and

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

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

Analogous results for the parametric (linear) regression estimator are given in Result 8.4.1 of Särndal, Swensson, and Wretman (1992).

4 Simulation Studies

We performed some simulation experiments in order to compare the performance of the local polynomial regression estimator in two-stage element sampling with the Horvitz-Thompson estimator in equation (1) and the linear regression estimator (Särndal, Swensson, and Wretman, 1992, p. 309).

We consider four mean functions for the cluster totals:

$$\begin{aligned} \text{linear:} \quad & \mu_1(x) = 1 + 2(x - 0.5), \\ \text{quadratic:} \quad & \mu_2(x) = 1 + 2(x - 0.5)^2, \\ \text{bump:} \quad & \mu_3(x) = 1 + 2(x - 0.5) \\ & \qquad + \exp(-200(x - 0.5)^2), \\ \text{jump:} \quad & \mu_4(x) = \{1 + 2(x - 0.5) I_{\{x \le 0.65\}}\} \\ & \qquad + 0.65 I_{\{x > 0.65\}}, \end{aligned}$$

with $x \in [0, 1]$. For μ_1, the linear regression estimator is expected to perform best because the model is correctly specified. The quadratic function is smooth but far from linear, bump is smooth and nearly linear, and jump is not smooth.

The population consists of $M = 1000$ clusters. The x_i are generated as independent and identically distributed (iid) uniform(0,1) random variables. For

Population	σ	h	HT	REG
	0.1	0.10	21.66	0.91
linear	0.1	0.25	22.36	0.94
	0.4	0.10	2.27	0.91
	0.4	0.25	2.34	0.94
	0.1	0.10	2.06	2.16
quadratic	0.1	0.25	2.09	2.19
	0.4	0.10	0.99	1.01
	0.4	0.25	1.02	1.04
	0.1	0.10	22.63	5.31
bump	0.1	0.25	9.20	2.16
	0.4	0.10	2.44	1.13
	0.4	0.25	2.38	1.10
	0.1	0.10	3.53	2.68
jump	0.1	0.25	2.81	2.14
	0.4	0.10	1.11	1.04
	0.4	0.25	1.10	1.03

Table 1: Ratio of design MSE of Horvitz-Thompson (HT) and linear regression (REG) estimators to local linear regression (LPR1) estimator.

each generated value x_i and each study variable ($j = 1, 2, 3, 4$), N_i element values are generated as

$$y_{jk} = \frac{\mu_j(x_i)}{N_i} + \frac{\varepsilon_{jk}}{N_i^{1/2}}, \quad \{\varepsilon_{jk}\} \text{ iid } \mathrm{N}(0, \sigma^2)$$

where $k \in U_i$. Two values for the standard deviation of the errors are used: $\sigma = 0.1$ and 0.4. At stage one, a sample of clusters is first generated by simple random sampling with sample size $m = 100$ and then samples of elements within each selected cluster at stage two are generated by simple random sampling using sample size n_i.

We have considered three cases with different second-stage sampling rates: constant cluster size $N_i = 100$ with $n_i = 10$, constant cluster size $N_i = 100$ with $n_i = 100$, and random cluster size N_i distributed as Poisson(3) + 1 with $n_i = \lfloor 0.5 N_i \rfloor + 1$, where $\lfloor a \rfloor$ denotes the integer part of a. As the second-stage sampling rate increases, the local linear regression estimator gains more improvement in efficiency over the other estimators. Here, we only report on the experiment with the random cluster sizes. Such clusters of moderate and variable size might be encountered in a household survey.

The Epanechnikov kernel,

$$K(t) = \frac{3}{4}(1 - t^2) I_{\{|t| \le 1\}},$$

and two bandwidth values ($h = 0.1$ and 0.25) are used for the local linear regression estimator. For each combination of mean function, standard deviation and bandwidth, 100 replicate two-stage element samples from the four fixed populations are selected and then the estimators are calculated.

Table 1 shows the ratios of design mean squared errors (MSEs) of the Horvitz-Thompson (HT) and the linear regression (REG) estimators to that of the local linear regression (LPR1) estimator. In all populations, the LPR1 estimator performs better than the HT estimator. The LPR1 estimator loses a small amount in efficiency over the REG estimator for the linear population, but is better for other populations. At small values of σ, the LPR1 estimator is much better than the other estimators. We also considered the case with $h = 0.5$ to see how the performance of the LPR1 estimator changes with increasing bandwidth, but do not display those reports here. For all bandwidths, the LPR1 estimator is better than the REG estimator for all but the linear population. As the bandwidth becomes large, the performance of the LPR1 estimator becomes similar to that of the REG estimator.

5 Example: National Resources Inventory data

In this section, 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 agricultural lands in the United States conducted by the Natural Resources Conservation Service of the U.S. Department of Agriculture. 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, a size measure of land with erosion potential. The variables of interest were two kinds of erosion measurements, roughly characterized as wind erosion (WEQ) and water erosion (USLE). 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. At stage two, samples of NRI segments within the selected counties were chosen by stratified unequal probability sampling. In total, 1900 segments were selected.

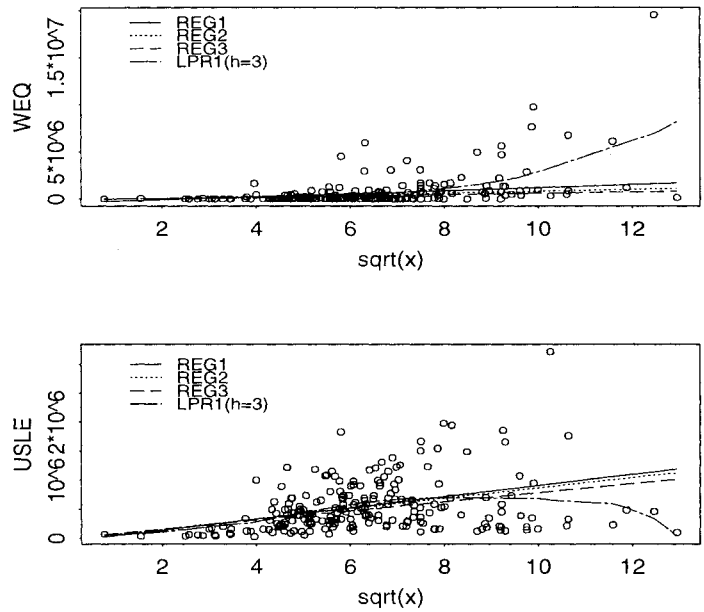

Figure 1: Plots of the relationship between square root of size measure ($x_i^{1/2}$) and estimated county total ($\hat{t}_i$) in selected counties at stage one for wind erosion (WEQ) and water erosion (USLE). Linear regression (REG1, REG2, REG3) and local linear regression (LPR1 with $h = 3$) fits are added in the plots.

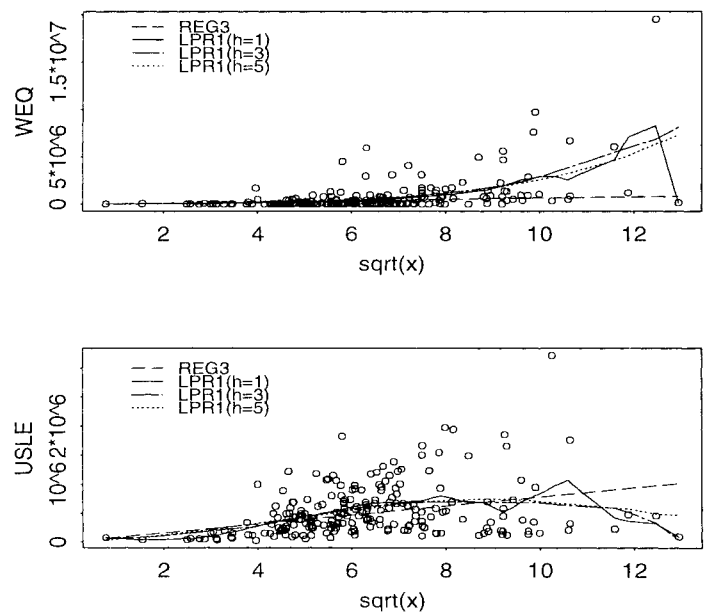

Figure 2: Plots of the relationship between square root of size measure ($x_i^{1/2}$) and estimated county total ($\hat{t}_i$) in selected counties at stage one for wind erosion (WEQ) and water erosion (USLE). Linear regression (REG3) and local linear regression (LPR1 with $h = 1, 3, 5$) fits are added in the plots.

		WEQ	USLE
HT		443.6	551.5
		(49.4)	(31.8)
REG1	$\nu(x) \propto 1$	485.9	544.4
		(92.1)	(29.3)
REG2	$\nu(x) \propto x$	448.7	540.2
		(52.7)	(27.1)
REG3	$\nu(x) \propto x^2$	442.5	537.8
		(50.7)	(26.5)
LPR1	h=1	434.1	529.0
		(47.5)	(24.4)
LPR1	h=3	427.4	532.3
		(48.9)	(25.3)
LPR1	h=5	430.5	541.2
		(48.7)	(27.6)

Table 2: Horvitz-Thompson (HT), linear regression (REG1, REG2, REG3), 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. We used the square root of the size measure $x_i^{1/2}$ instead of x_i to reduce the sparseness of points in the regressor space. We calculated REG estimates with three different variances of the errors ($\nu(x) \propto 1, x$, and x^2), denoted by REG1, REG2, and REG3 respectively. This was done because the data displayed large amounts of heteroskedasticity (See Figure 1), affecting the parametric fit. Three bandwidths ($h = 1$, 3, 5) were used for LPR1 (the smallest possible bandwidth to the nearest tenth for these data was $h = 1$).

Figure 1 shows the relationship between square root of size measure ($x_i^{1/2}$) and estimated county total ($\hat{t}_i$) in counties selected at stage one for WEQ and USLE. Linear regression with three different error variances (REG1, REG2, REG3) and local linear regression with bandwidth $h = 3$ (LPR1($h = 3$)) fits are added in the plots. In REG estimates, REG3, the best performing among them, has the smallest slope for both WEQ and USLE. The behavior of LPR1 is quite different from that of REG estimates in the sparse part of x_i.

Figure 2 shows the linear regression with the variance of the errors proportional to x^2 (REG3) and local linear regression with three different bandwidths: LPR1($h = 1$), LPR1($h = 3$), and LPR1($h = 5$).

Table 2 shows HT, REG and LPR1 estimates of WEQ and USLE totals and estimated standard errors. Using the estimated standard error as a guide, LPR1 with $h = 1$ performs best among all estimates and REG3 (REG with the variance of the errors proportional to x^2) is best among REG estimates. Overall, LPR1 estimates except of the largest bandwidth are better than HT and REG estimates on the basis of estimated standard errors for both WEQ and USLE. In WEQ, the estimated standard error of REG is unexpectedly large, compared to that of HT. This seems to be due to the presence of a few strata with zero estimated variance for the HT estimator.

Acknowledgements

This work was supported in part by cooperative agreement 68–3A75–43 between the USDA Natural Resources Conservation Service and Iowa State University. Computing for the research was done with equipment purchased with funds provided by an NSF SCREMS grant award DMS 9707740.

References

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.

Breidt, F.J. and Opsomer, J.D. (2000) Local polynomial regression estimators in survey sampling. *Annals of Statistics*, to appear.

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.

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

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.

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.

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

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