Nonparametric Survey Regression Estimation in Two-Stage Spatial Sampling

Siobhan Everson-Stewart, F. Jay Breidt, Jean D. Opsomer

January 20, 2004

Key Words: auxiliary information, environmental surveys, kernel regression, smoothing

Abstract

A nonparametric model-assisted survey estimator for status estimation based on local polynomial regression is extended to incorporate spatial auxiliary information. Under mild assumptions, this estimator is design-unbiased and consistent. Simulation studies show that the nonparametric regression estimator is competitive with standard parametric techniques when a parametric specification is correct, and outperforms those techniques when the parametric specification is incorrect. The methodology is applied to water chemistry data from the EMAP Northeastern Lakes Survey.

1 Introduction

1.1 Motivation

Many surveys contain auxiliary information at the population level in addition to sample data. This auxiliary information can come from satellite images, GPS data, aerial photographs, or other sources. In a common survey situation, a statistical agency, often a federal, state, or tribal, collects data and auxiliary information. A data set is then created and released to the users. This data set reflects knowledge of both the design and auxiliary information. The end users are then responsible for estimating the status of many study variables. The U.S. EPA's EMAP (Environmental Monitoring and Assessment Program) Northeastern Lakes survey is a good example of this. The EPA had some information about all the lakes in the region of interest, such as longitude, latitude, and elevation. A sample of lakes was then taken. Each of the sampled lakes was visited, and additional measurements were taken. The final data set includes both this sample information and the auxiliary information previously available.

In this situation, it is of interest to develop a model relating sample and auxiliary information so that inferences can be made about non-sampled population values. However,

in this modeling environment, time and other resources are often limited. The agency using the data may need quick answers to questions about population status. Their statistical resources may also be limited. Additionally, there may be controversy among the end users. The preferred model should not be influenced by user biases. This ideal model should be capable of handling many study variables without requiring modeling efforts for each. It should also be efficient if the model is correct, but not fail if the model is wrong. In this situation, that means the estimator should have small variance compared to other available estimators when the model is specified correctly, but still give reasonable answers if this specification is incorrect. That is, the estimator should be approximately unbiased and consistent.

Breidt and Opsomer (2000) developed a local polynomial regression estimator for direct element sampling with scalar auxiliary information available for every item in the population. This estimator is asymptotically unbiased, and more efficient than regression models when the model regression function is not correctly specified, while being competitive with the regression estimator when the regression model is correct. In this paper, the method developed by Breidt and Opsomer is extended to bivariate auxiliary information. This extension has numerous practical applications, as many data sets include more than one auxiliary variable. Frequently, for example, the spatial location of every population element is known. For concreteness, here we consider the auxiliary vector to be spatial coordinates of the first stage sampling unit, though any other auxiliary vector would be handled the same way.

1.2 Overview of Paper

In this paper, we first define some of the notation used for two stages of sampling. Model-assisted estimation is discuess and the nonparametric regression estimator is then presented. The nonparametric regression technique is examined through a simulation study, and then applied to data from the EMAP Northeastern Lakes study in the empirical example. The paper concludes with a discussion of results and further topics of research.

2 Methods

2.1 Notation

Consider a finite population of elements $U = \{1, \ldots, k, \ldots, N\}$. At stage one, a probability sample s is drawn from U according to a fixed size design $p_I(\cdot)$, where $p_I(s)$ is the probability of drawing the sample s from U. Let n be the size of s. The 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. The study variable z_k is observed for $k \in s$. The parameter of interest is the the population total $t_z = \sum_{k\in U} z_k$. Let $I_i = 1$ if $i \in s$ and $I_i = 0$ otherwise. Using this notation, an estimator $\hat{t}$ of t is said to be design-unbiased if $\mathrm{E}_p\left[\hat{t}\right] = t_z$. For each population element, i, an auxiliary vector, $\boldsymbol{x}_i$, is available. Here,

we consider $\boldsymbol{x}_i$ to be the spatial location of i, $\boldsymbol{x}_i = (x_i, y_i)$.

2.2 Model-Assisted Estimation

The response for the ith population unit, z_i, is assumed to be a smooth function of $\boldsymbol{x}$ plus random error, i.e.

$$z_i = g_t(\boldsymbol{x}_i) + \epsilon_i$$

where $g_t(\boldsymbol{x}_i)$ is a smooth function of $\boldsymbol{x}_i$, ϵ_i are independent random variables with zero mean and variance $v(\boldsymbol{x}_i)$, a smooth, strictly positive function of $\boldsymbol{x}_i$.

Model-assisted estimators are formed from a model-based prediction, plus a design bias adjustment:

$$\hat{t} = \left\{ \sum_{i \in U} \hat{\mu}_i + \sum_{i \in s} \frac{z_i - \hat{\mu}_i}{\pi_i} \right\}$$

where $\hat{\mu}_i$ is the model fit for the ith population unit.

These estimators are approximately design-unbiased, with small variance if the model is correct. For the popular Horvitz-Thompson estimator, $\hat{\mu}_i \equiv 0$ since no auxiliary information is used. In generalized regression, $\hat{\mu}_i = x_i' \hat{\beta}$, while $\hat{\mu}_i$ comes from a kernel smooth in local polynomial regression estimation.

2.3 Two-Stage Estimator for Spatial Sampling

Using local polynomial regression, the form of the estimator for the two dimensional case is the basic model assisted estimator, where the model based prediction of $\hat{\mu}_i$ is

$$\hat{\mu}_i = \boldsymbol{e}_1' \left(\boldsymbol{X}_{si}' \boldsymbol{W}_{si} \boldsymbol{X}_{si} \right)^{-1} \boldsymbol{X}_{si}' \boldsymbol{W}_{si} \boldsymbol{z}_s = \boldsymbol{w}_{si}' \boldsymbol{z}_s$$

Here, $\boldsymbol{e}_1$ is a vector with a 1 in the first position and 0 for the rest of the entries; $\boldsymbol{z}_s$ is the vector of sampled responses. The local design matrix, $\boldsymbol{X}_{si}$, is defined as

$$\boldsymbol{X}_{si} = \left[\begin{array}{cccccc} 1 & x_j - x_i & \cdots & (x_j - x_i)^p & y_j - y_i & \cdots & (y_j - y_i)^q \end{array} \right]_{j \in s},$$

where p and q are the degrees of the local polynomials being fitted in the x and y directions, respectively. The local weighting matrix is

$$\boldsymbol{W}_{si} = \mathrm{Diag} \left\{ \frac{1}{\pi_j h_x h_y} K \left(\frac{x_j - x_i}{h_x}, \frac{y_j - y_i}{h_y} \right) \right\}_{j \in s}.$$

Here, K is a two-dimensional kernel function. The bandwidth in the x direction is h_x, while h_y is the bandwidth in the y direction. This form assumes a diagonal bandwidth matrix, $\boldsymbol{H}$. If $\boldsymbol{H}$ is not diagonal, then the expression above becomes

$$\boldsymbol{W}_{si} = \mathrm{Diag} \left\{ \frac{1}{\pi_j |\boldsymbol{H}|} K \left(\boldsymbol{H}^{-1} \boldsymbol{x}_{ij} \right) \right\}_{j \in s},$$

where

$$\boldsymbol{x}_{ij} = \begin{bmatrix} x_j - x_i \\ y_j - y_i \end{bmatrix}.$$

The estimator of $\hat{t}$ can now be expressed as

$$\hat{t} = \sum_{i \in U} \hat{\mu}_i + \sum_{i \in s} \frac{z_i - \hat{\mu}_i}{\pi_i}.$$

2.3.1 Weights

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

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

Thus, $\hat{t}$ is a linear combination of the z_i's in s, with weights $\{\omega_{is}\}$ that reflect both the design and the auxiliary information. Because these weights are independent of the study variables, they can be applied to any study variable of interest.

3 Simulation

We performed simulations to examine the performance of the local polynomial regression estimator, and to compare it to that of other estimators. The estimators used are the Horvitz-Thompson (HT), linear regression (REG) as in Särndal, Swensson, and Wretman (1992), and local polynomial regression, with $p = q = 0$ (LPR0) and $p = q = 1$ (LPR).

In this simulation, we sample from a continuous, spatially correlated surface, or from a noisy version of this surface. The surface is considered fixed, as we are looking at properties of estimators under repeated sampling of a population. Note that it is not possible to compute and store a continuous surface in a simulation study without first parameterizing it. In order to create the continuous surface, we laid down an 11×11 grid of points over the unit square. The response at each of these grid points is a deterministic trend plus a spatially correlated stochastic process:

$$g(x_i, y_i) = \mu(x_i, y_i) + G(x_i, y_i),$$

where

$$\mathrm{Cov}\left(G(x_i, y_i), G(x_j, y_j)\right) = \sigma^2 \exp\left\{ \log(\rho) \left((x_i - x_j)^2 + (y_i - y_j)^2 \right)^{\frac{1}{2}} \right\}.$$

and the deterministic trend used was the plane

$$\mu(i) = \beta_0 + \beta_x x_i + \beta_y y_i.$$

Between the grid points, an interpolating thin plate spline was used to create a continuous surface. Aldworth and Cressie did a similar study where they created a grid of points with a planar trend and spatially correlated stochastic process (1999). However, in their study, rather than using a spline to create a continuum, they used the grid points as the population of interest.

The simulation was run for 48 different parameter settings. The sample size was either $m = 100$ or $m = 200$. Two different deterministic trends were considered: $\beta_0 = 0.5, \beta_x = \beta_y = 1$ and $\beta_0 = 0.5, \beta_x = \beta_y = 0$. We also varied the correlation between the points, ρ. Runs were completed with values of $\rho = 0.3487, 9.7656e-4$, and 0, which gave correlations of 0.9, 0.5, and 0, respectively, between points that were one grid unit apart (a distance of 0.1). A correlation of zero adds random noise to the grid points, while those of 0.5 and 0.9 create stochastic trends of varying smoothness. The observed z values were then

$$z(x_i, y_i) = \beta_0 + \beta_x x_i + \beta_y y_i + G(x_i, y_i) + \epsilon_i$$

where ϵ_i are independent, identically distributed normal random variables with $\mathrm{Var}(\epsilon_i) = \nu^2$, and with $G(x_i, y_i)$ the spatially correlated stochastic process defined above. Both the case where $\nu^2 = 0$ and that when $\nu^2 > 0$ were considered. We chose different values of σ^2 so that the spatially correlated stochastic process accounted for 5% and 15% of the total variation in the response values at the grid points in the case with planar trend. That is, for $\mathrm{Var}(G(x_i, y_i)) = \sigma^2$,

$$\frac{\sigma^2}{\sigma^2 + \int_0^1 \int_0^1 (\mu(x,y))^2 dxdy - \left(\int_0^1 \int_0^1 \mu(x,y) dxdy\right)^2 + \nu^2} = 0.05 \text{ or } 0.15$$

The calculations of total variability were done for the planar response, and the same values of the variance parameters were used for the flat response. These two levels of variance were chosen to create an overall smooth that is nearly deterministic, and one that has a more substantial stochastic component. In the case of $\nu^2 = 0$, $\sigma^2 = 0.00877$ for the 5% case and 0.02632 for 15%. With $\nu^2 > 0$ and 5% of the total variation coming from the noise, these values of σ^2 become 0.008795 and 0.03214. The noise variance, ν^2, is 0.008795 with $\sigma^2 = .008795$ and $\nu^2 = 0.01033$ when $\sigma^2 = .03214$.

For one run of the simulation at a particular setting, the 11×11 grid of points was generated. The spline was created, and the volume under the spline calculated by numerical integration. This value was later used in bias calculations. Then, for each replication, we took a simple random sample over the unit square. The interpolating spline was evaluated at each sampled point.

For each sample, the four estimates and their estimated variances were calculated. A sample was taken 100 times for each setting. Then, for each estimator, the average estimate, the average estimated variance, the variance of the estimates, the percent bias of

Settings	Estimator	Estimate	Var(Est)	Est.Var.	%Bias	Efficiency	%Bias Var.
$\rho = 0$	HT	1.5064	0.000686	0.000859	0.546	16.092	5.281
PSSV=5%	REG	1.4987	0.000021	0.000020	0.032	0.493	-3.810
Noise=0%	LPR	1.4992	0.000043	0.000016	0.065	1.000	-61.502
True=1.4982	LPR0	1.5004	0.000067	0.000051	0.147	1.577	-24.851
$\rho = 0.5$	HT	1.5083	0.000729	0.000895	0.549	19.703	22.771
PSSV=5%	REG	1.5005	1.58E-05	1.75E-05	0.025	0.427	10.759
Noise=0%	LPR	1.4995	0.000043	9.98E-06	-0.041	1.000	-73.015
True=1.4982	LPR0	1.5013	6.51E-05	4.39E-05	0.079	1.759	-32.606
$\rho = 0.9$	HT	1.4708	8.28E-04	9.94E-04	0.585	26.938	19.995
PSSV=5%	REG	1.4623	3.90E-06	4.47E-06	0.009	0.127	14.662
Noise=0%	LPR	1.4624	3.08E-05	2.04E-06	0.012	1.000	-93.362
True=1.4622	LPR0	1.4640	6.22E-05	3.59E-05	0.125	2.022	-42.185
$\rho = 0$	HT	1.5053	6.96E-04	8.98E-04	0.565	8.423	29.121
PSSV=15%	REG	1.4977	6.30E-05	6.05E-05	0.055	0.763	-4.030
Noise=0%	LPR	1.4982	8.26E-05	4.91E-05	0.088	1.000	-40.533
True=1.4969	LPR0	1.4991	1.02E-04	8.86E-05	0.149	1.237	-13.251
$\rho = 0.5$	HT	1.5087	7.66E-04	9.60E-04	0.570	12.543	25.348
PSSV=15%	REG	1.5008	4.73E-05	5.26E-05	0.044	0.774	11.409
Noise=0%	LPR	1.4987	6.11E-05	3.00E-05	-0.097	1.000	-50.938
True=1.5002	LPR0	1.5006	8.84E-05	6.95E-05	0.030	1.448	-21.423
$\rho = 0.9$	HT	1.4436	9.28E-04	1.12E-03	0.634	24.692	20.698
PSSV=15%	REG	1.4348	1.17E-05	1.34E-05	0.015	0.311	14.665
Noise=0%	LPR	1.4345	3.76E-05	6.12E-06	-0.006	1.000	-83.711
True=1.4345	LPR0	1.4361	7.05E-05	4.38E-05	0.108	1.877	-37.869

Table 1: For all settings, 100 replications were run with a 11×11 population grid, and a 15×15 grid for the interpolation within the LPR estimator. PSSV is the proportion of small-scale variation. This table presents the results for $m = 200$, $\mu_i = 0.5 + x_i + y_i$, and no noise.

the estimator, the relative efficiency of the estimator, as compared to the LPR estimator, and the percent bias of the variance estimate were calculated. For the local polynomial regression estimators, we used the product Epanechnikov kernel function with $h = 0.35$ in both directions. This corresponds to approximately 1/8 of the total surface being included in each smooth. Because the response surface is continuous, an estimate of the smooth could not made for every population element. Rather, the smooth was evaluated over a 15×15 grid, and then a piecewise constant interpolation was used to extend the smooth over the entire sampling space.

A sample output table is given in Table 1. From the simulation results, it can be seen that the local polynomial regression estimator performs well, especially in the cases where the parametric models are misspecified. In the case of planar trend, the regression

estimator dominates all others. It should be noted that, for these parameter setting, the LPR estimator does best in the case with a correlation of 0.5. This makes sense, since when $\rho = 0$, the truth is close to a plane with random noise, and for the highest value of ρ, the high correlation forces a flatter response surface. The 0.5 correlation creates a locally smooth surface with peaks and valleys. Additionally, the LPR's performance relative to the regression estimator improves with the introduction of random error. For the simulation runs with a flat plane, the LPR0 estimator outperforms all other estimators except when there is no correlation in the data, and, even then, is competitive and is sometimes the most efficient estimator. Here, the response surface is a smoothed spatially correlated stochastic process, a situation in which local polynomial regression would be expected to perform well. And because the underlying structure is flat, not planar, regression of degree zero not surprisingly outperforms local planar regression. It should also be noted that in limited simulations with higher levels of variance in the stochastic trend, the LPR estimator dominated the other estimator for all values of the trend and ρ. The relative efficiencies of the HT and REG estimators as compared to the LPR estimator are plotted in Figures 1 and 2 for a sample size of $m = 100$.

From these simulation results, it is apparent that there is a bias problem with the variance estimation of the LPR estimators. For almost every parameter setting, both variances estimates are negatively biased, often greatly so. We suspect much of this problem can be attributed to under-smoothing. If the bandwidth is too small, the sampled kernel smooth will follow the data points more closely than does the population level smooth. This results in small residuals and underestimation of variance.

A few additional runs of the simulation program were made to see if increasing the bandwidth would decrease the bias of the estimated variance of the local polynomial regression estimator. Here, the original bandwidth would be used to estimate the parameter of interest, and then the over-smoothing bandwidth would be used to obtain an estimate of the variance. For the LPR estimator, this oversmoothing yields good estimates of the variance in the case with no planar trend, but is still negatively biased for the planar case. Higher bandwidths might solve that problem. For the LPR0 estimator, this bandwidth over-corrects. The variance estimates are now positively biased, often highly so.

As previously mentioned, these simulations were done on a continuous population. In the empirical example to follow, the data are discrete, made up of individual lakes. To verify that these simulation results are applicable to a discrete population, a short additional simulation study was performed. In this study, the 11×11 grid of points was set up, and the spline calculated the same way as in the original study. Then, $10,000$ random points over the unit square were selected. The interpolating spline was used to evaluate the response surface at each of these points. These $10,000$ points then served as the population of interest. For each repetition of the simulation, a simple random sample of these points was selected. The estimators were evaluated as before, and summary statistics were computed at the end of each run. The results, are very similar to those for the continuous population. This suggests that the earlier simulation results are applicable to both continuous and discrete resources.

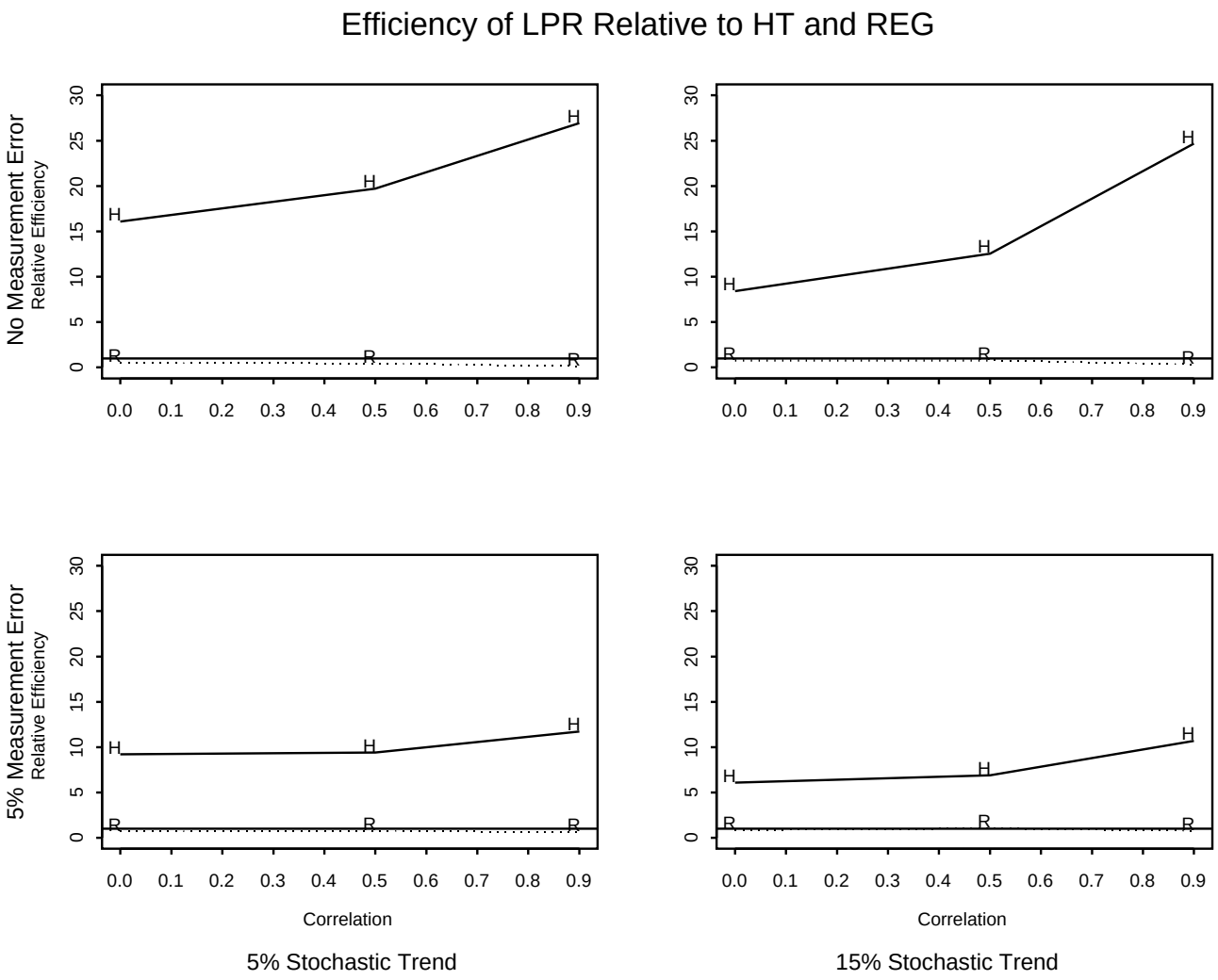

Figure 1: Simulation results for data with a planar trend. H= efficiency of HT relative to LPR, R= efficiency of REG relative to LPR.

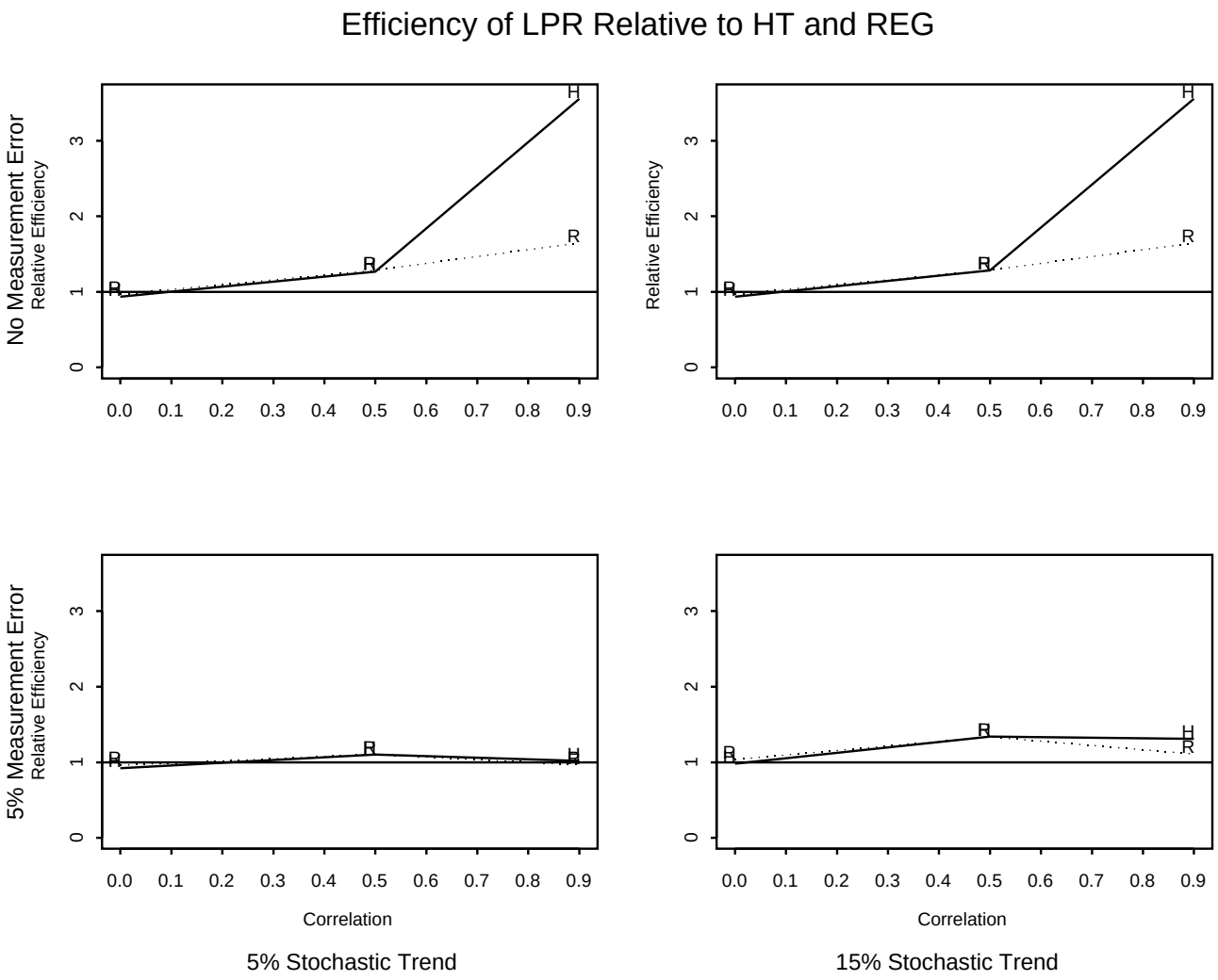

Figure 2: Simulation results for data with no planar trend. H= efficiency of HT relative to LPR, R= efficiency of REG relative to LPR.

4 Empirical Example

We applied the two-dimensional nonparametric regression technique to the EMAP Northeastern lakes survey. The sampling frame for this study consisted of over 20,000 lakes in 8 northeastern states. Over the course of 6 years, from 1991 to 1996, over 330 individual lakes were visited, each from one to six times. Each individual lake was treated as a cluster, and water samples within lakes (obtained on repeat visits) were treated as secondary sampling units. During a visit, many variables of interest were measured and recorded. For the purpose of this example, we examine the lake chemistry measures. Auxiliary information is available for each lake. For this example, we worked with Universal Transverse Mercator (UTM) coordinates, a form of spatial coordinates. For UTM coordinates, the world is divided into 60 zones. Each zone has an origin point. The UTM coordinates of a spatial location within the zone are then the meters to the north and east of this origin. The range of the UTM easting coordinates in the Northeastern lakes data set is 108095 to 1127690, and that of the UTM northing coordinates is 4311620 to 5263630. These location measures can serve as a proxy for other spatial covariates, or account for spatial correlation. To account for multiple visits to a single lake, we applied the technique to the lake averages. Because of the type of variables used, population totals are not of interest; instead the average was computed. We calculated nonparametric regression estimates of the average lake mean for several lake chemistry measures.

Table 2 displays the average mean for four chemistry measures calculated using the Horvitz-Thompson (HT), linear regression (REG), local linear regression (LPR), and local constant (LPR0) estimators. The estimated coefficient of variation is also given. The model assisted procedures are estimated to be much more efficient than the Horvitz-Thompson estimator. Additionally, the local regression estimators are estimated to make small gains over the linear regression estimator in each case. However, it was apparent from the simulation study that the LPR variance estimator is negatively biased. Based on those results, it is reasonable to conclude, for this example, that the LPR estimator improves upon the Horvitz-Thompson estimator and is competitive with the regression estimator. It should be noted that the trends in the variables over space are noisy, with weak linear trends.

To minimize numerical errors, all UTM coordinates were divided by 1,000,000. For both LPR and LPR0, a bandwidth of $h = 0.22$ was used in both the north-south (y) and east-west (x) directions. This bandwidth is approximately 1/4 of the range of the standardized UTM coordinates in each direction.

Variance estimation was done using a stratified with replacement approximation, where the lakes were stratified by a clustering variable used in the original sampling process. The original sampling plan contained built-in spatial control to ensure a well-dispersed sample in space, with some similarity to systematic sampling (Larsen, et.al. 1993). The stratification employed here is analogous to the standard collapsed-stratum variance estimator for systematic sampling (Särndal, Swensson, and Wretman, p.423). The original sampling frame consisted of all lakes with their centroids in a grid of hexagons. These hexagons were spatially clustered. All clusters of hexagons were listed in random order, with the hexagons

Chemistry Measure	HT	REG	LPR	LPR0
Log K	2.845	2.916	2.921	2.890
	(7.00%)	(2.26%)	(2.20%)	(2.23%)
Log SO_4	4.828	4.785	4.791	4.779
	(6.94%)	(1.00%)	(0.95%)	(1.01%)
Log Ca	5.835	5.723	5.742	5.760
	(6.76%)	(1.50%)	(1.21%)	(1.38%)
Log Cl	4.531	4.754	4.772	4.690
	(6.97%)	(2.27%)	(1.91%)	(2.10%)

Table 2: Estimated average mean of lake chemistry measures using Horvitz-Thompson (HT), linear regression (REG), local linear regression (LPR) and local linear regression fitting a local constant (LPR0). Numbers in parentheses are estimated coefficients of variation.

within a cluster randomly ordered and the lakes within a hexagon also listed randomly. A systematic sample was then taken (Larsen et al. 1993). This design forced spatial control in the sample.

To form the strata for the variance estimation, these spatial clusters were grouped together into 14 larger clusters. These clusters were created by combining the old clusters with their nearest neighbors. Each of the new clusters consists of between one and seven of the original clusters, with each of the new clusters containing at least 19 but not more than 122 lakes.

For many of the lakes, only one sample was taken. In order to estimate the variability in the second stage of sampling, we assumed the variance within each lake was constant from lake to lake. We then took the sample variance of all lakes with more than one sample. The average of these sample variances was then taken to be the within lake variance for all lakes.

5 Discussion and Conclusions

In this paper, we have introduced the local polynomial regression estimator in the context of spatial sampling. This estimator incorporates known population auxiliary information to improve status estimates. Here, the estimator was described in the situation where, in addition to the sample data, the spatial location of every cluster in the population is known. The estimator was given, then the performance of the local polynomial regression estimator was examined through a simulation study. When the parametric model was incorrectly specified, the LPR estimator outperformed standard parametric techniques. The local polynomial regression technique was also applied to data from the EMAP Northeastern lakes data set. The estimator also performed well in this setting, giving comparable estimates, but smaller estimated standard errors than the other techniques. It should be noted, however, that the simulation study showed that the variance estimator for local

polynomial regression is negatively biased, greatly so if the data are under-smoothed.

Based on the results of the simulation study and the other work done here, the local polynomial regression estimator is a reasonable modeling choice in spatial sampling. Unless a population is known to have a planar trend, the local polynomial regression estimator should perform better than other estimators, especially if the variable of interest is spatially correlated or is highly variable. The local polynomial regression estimator also has the advantage of requiring little modeling efforts. Because the estimate can be expressed in terms of a weighted sum of sampled observations, with the weights remaining constant from one variable to the next, the technique can easily and quickly be applied to many study variables. In sampling with auxiliary information available for the entire population, the local polynomial regression estimator should improve the precision of estimation in populations with spatial correlation or other non-linear trends.

References

Aldworth, J. and Cressie, N. (1999). Sampling Designs and Prediction Methods for Gaussian Spatial Processes. *Multivariate Analysis, Design of Experiments, and Survey Sampling*, edited by S. Ghosh. Dekker, New York, 1-54.

Breidt, F.J. and Opsomer, J.D. (2000). Local polynomial regression estimation in survey sampling. *Annals of Statistics* **28**, 1026–1053.

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.

Green, P.J., and Silverman, B.W.. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.

Larsen, D.P., Thornton, K.W., Urquhart, N.S., Paulsen, S.G. (1993). Overview of Survey Design and Lake Selection. *EMAP - Surface Waters 1991 Pilot Report*, edited by Larsen, D.P. and Christie, S.J. EPA/620/R-93/003.

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.