TWO-PHASE REGRESSION ESTIMATION FOR POLICY ANALYSIS USING COMPUTER SIMULATION EXPERIMENTS

H. Martin Axelson, University of Örebro,
F. Jay Breidt and Alicia L. Carriquiry, Iowa State University
F. Jay Breidt, Department of Statistics, Iowa State University, Ames, IA 50011 USA

KEY WORDS: double sampling, metamodel, National Resources Inventory, variance estimation.
ABSTRACT: In agricultural economic policy analyses, data of interest such as long-run averages of soil lost to erosion or chemicals leached to groundwater are not available due to high monitoring costs. Hence, computer simulation models are used to describe the corresponding physical processes of soil erosion and chemical movement in soils. These models are site-specific, as they depend on topography, soil properties, weather, management practices, etc. It is impractical to run a simulation model for all sites in a region of interest because input information is not available and computing resources are not adequate. Instead, the model is run for a subsample of the points drawn in the National Resources Inventory (NRI), a stratified two-stage area sample of the nonfederal lands in the United States. Researchers then typically fit a regression "metamodel" to the results of this computer simulation experiment, and finally the fitted metamodel is used on all NRI points to obtain estimates at regional levels (e.g., hydrologic regions, counties, etc.) We formulate this metamodel estimation problem in terms of two-phase regression estimation and develop a variance estimation strategy.

1 Introduction

1.1 Policy analysis

Agricultural activity has significant impacts on the environment. Control of nonpoint pollution from agricultural practices and source reduction of agricultural pollutants for water quality and ecosystem protection are increasingly debated policy goals (e.g., US EPA, 1992). These debates need to draw from informed evaluation of the environmental impact of different agricultural policies, so that policy makers can base their decisions on objective and reliable information.

Ideally, environmental monitoring would provide policy makers with the needed information. Evaluation of sustainable agricultural practices would, however, require environmental monitoring of such scale as to make it practically impossible. Suppose that environmental impact is quantified via some response (that may not be observable) such as chemical leaching into the groundwater, soil erosion, or nitrogen runoff. Denote the true value of the response under a given policy at site $k \in U$ by Φ_k, where U is some region of interest, such as a state. Monitoring of the response would need to be carried out at sites representative of the many possible combinations of inputs that affect the response. Inputs may be of at least two different types: (1) *factors* that are subject to change via policy (such as tillage practices, type and amount of chemical used, crops and crop rotations), and (2) *covariates* that, while not susceptible to policy, still have an effect on the response (soil characteristics, weather, and topography). We denote by $\mathbf{w}_{0k}$ and $\mathbf{z}_{0k}$, respectively, the vector of factors and the vector of covariates. Further, we can partition $\mathbf{z}_{0k}$ into $\mathbf{v}_{0k}$ and $\mathbf{u}_{0k}$, where $\mathbf{v}_{0k}$ represents a vector of observed inputs at site k and $\mathbf{u}_{0k}$ represents a vector of unobserved but imputed inputs (such as weather, which is recorded at a nearby monitoring station but not at every site k).

1.2 Computer simulation models

Because environmental monitoring is very expensive and time consuming, policy makers interested in evaluating potential environmental impacts from the application of different agricultural policies are increasingly relying on data generated by mathematical simulation models, such as those developed by the US Environmental Protection Agency (EPA) and the US Department of Agriculture (USDA). Examples of simulation models for physical processes include the *Water Quality and Erosion Productivity Impact Calculator* (EPIC-WQ) model, (Williams et al., 1988), the *Risk of Unsaturated/Saturated Transport and Transformation of Chemical Concentrations* (RUSTIC) system, (Dean et al., 1989), and the *Surface Transport and Agricultural Runoff of Pesti-*

cides for Exposure Assessment (STREAM) model (Donigian et al., 1986). For a site k with input variables $\mathbf{x}'_{0k} = \{\mathbf{w}'_{0k}, \mathbf{v}'_{0k}, \mathbf{u}'_{0k}\}$, these physical process models produce an estimated response denoted by φ_k.

While much more economical than monitoring, these simulation models are practical for site-specific problems only (Evans and Myers, 1990; Carriquiry et al., 1996). To use these field-scale models for regional assessments requires that the simulations be run for the area-wide distribution of soils, crop rotations, chemicals in use, and management practices. For example, almost 75,000 runs would be needed to cover a study area consisting of the Corn Belt and the Great Lakes states of the United States (Bouzaher et al., 1993). Furthermore, to compare different policies with regard to their potential environmental impacts, the simulation runs would need to be repeated for all combinations of factors used in the baseline evaluation.

1.3 Computer simulation experiments

Instead of running the process models for every possible site in the region of interest, a better approach is to design a computer simulation experiment (Bouzaher et al., 1993; Lakshminarayan et al., 1995; Carriquiry et al., 1996). A computer simulation experiment in this context consists of running the simulation model at a probability sample $s \subset U$ of sites in the region of interest, in which every site has a known, positive probability of inclusion in the sample (e.g., Särndal et al., 1992). Under probability sampling, unbiased estimators of population parameters such as means and totals can be formed without appealing to any assumed statistical model.

In the United States, examples of large-scale probability samples which may be of interest in policy evaluations include the National Resources Inventory (NRI), conducted by the Natural Resources Conservation Service of the USDA; the Forest Inventory and Analysis program (FIA) of the Forest Service of the USDA; agricultural chemical use and other surveys conducted by the National Agricultural Statistics Services of USDA; and the Environmental Monitoring and Assessment Program (EMAP) administered by the US Environmental Protection Agency.

We focus on the NRI in this paper because of its usefulness as input to process models such as EPIC-WQ, which forms the basis of our Monte Carlo experiment in section 3. The NRI is a stratified two-stage area sample used to collect detailed information on the status, condition, and trends of natural resources on the nonfederal lands in the United States. NRI data items, collected by a combination of remote sensing and ground observation, include soil characteristics, land use, agricultural practices, erosion measures, and so on. This survey, the largest of its kind, is based on approximately 300,000 primary sampling units (PSUs) and about 800,000 points, and is updated every five years (Goebel and Baker, 1982).

In the first stage of sampling, the region of interest, U, is divided into PSUs U_i $(i = 1, \ldots, N_I)$, which consist of tracts of land of varying sizes, but are usually 160-acre square quarter sections. The PSUs are further grouped into strata U_{Ih} $(h = 1, \ldots, H)$, which are sub-county-level geographic subdivisions. A simple random sample s_{Ih} of size n_{Ih} is drawn from the N_{Ih} PSUs in stratum h. The first- and second-order inclusion probabilities are denoted π_{Ii} and π_{Iij} for $i, j \in U_I$.

In the second stage, a sample of n_i points, s_i, is selected within PSU U_i $(i \in s_{Ih},\ h = 1, \ldots, H)$ according to a restricted randomization procedure with first- and second-order inclusion probabilities $\pi_{k|i}$, and $\pi_{kl|i}$ for $k, l \in U_i$. The point sample from stratum h is $s_h = \cup_{i \in s_{Ih}} s_i$ and the final point sample is $s = \cup_{h=1}^{H} s_h$, where n_s is the size of s. Combining the first-order inclusion probabilities over the two stages give $\pi_k = \pi_{Ii}\pi_{k|i}$, for $k \in U_i$.

1.4 Metamodels

For environmental impact assessment at the regional level, the use of complex process simulation models such as EPIC-WQ, RUSTIC, or STREAM presents at least two major drawbacks: (1) The models produce site-specific, deterministic point forecasts, and (2) Computations at each site require a significant amount of time and effort, and furthermore, need to be repeated for each different policy scenario under consideration. Both drawbacks can be simultaneously addressed by replacing the complex simulation model with a simpler metamodel.

A metamodel is a predictive model explaining the input-output relationship of the computer simulation model (Kleijnen, 1987; Bouzaher et al., 1993). These metamodels are used to "fill in the gaps"; that is, to predict a response value at those sites where the process model was not run. Also, "what if?" questions asked by policy makers can be easily and rapidly answered using the metamodels, by predicting the value of the response at any location when policy scenarios change. A partial list of applications of metamodels estimated from simulated

data includes Taub and Burns (1991), Dillaha and Gale (1992), Bernardo et al. (1993), Bouzaher et al. (1993), and Lakshminarayan et al. (1995).

2 Analysis

2.1 Two-phase sampling

A computer simulation experiment consists of runs of the computer simulation model at specified points $k \in U$ with inputs $\mathbf{x}_{0k}$. If auxiliary information $\mathbf{x}_{0k}$ is available from a probability sample s such as the NRI, then it is natural to choose $k \in s$. However, runs of the computer simulation model are expensive, so it is often impractical to run the model for all $k \in s$. Instead, the model is run for a subsample $r \subset s$, drawn according to a probability sampling design with first-order inclusion probabilities $\pi_{k|s} = \mathrm{P}[k \in r \mid s]$ and second-order inclusion probabilities $\pi_{kl|s} = \mathrm{P}[k \in r, l \in r \mid s]$.

It is worth noting that the design for the sample r may have strata and PSUs different from those of the original sample. In particular, the points in the original sample may be restratified on the basis of the auxiliary vector; e.g., into cropland and noncropland points. We will refer to the original strata as *design strata*.

2.2 Metamodel estimation

We describe the case of a linear metamodel. Let $\mathbf{x}_k' = (\mathbf{u}_k', \mathbf{v}_k', \mathbf{w}_k')$ denote the vector of inputs to the metamodel, where $\mathbf{u}_k$ is a subvector of $\mathbf{u}_{0k}$, $\mathbf{v}_k$ is a subvector of $\mathbf{v}_{0k}$, and $\mathbf{w}_k$ is a subvector of $\mathbf{w}_{0k}$. Denote the output of the metamodel by

$$f_k = f(\mathbf{x}_k) = \mathbf{x}_k'\boldsymbol{\beta},$$

for $\boldsymbol{\beta}$ an unknown vector of parameters. The prediction errors that arise from using the metamodel in place of the simulation model are $\varphi_k - f_k$. Since there is in fact a true model, the metamodel is almost certainly misspecified, and hence gives biased predictions. Nevertheless, we model the prediction errors $\varphi_k - f_k$ as uncorrelated zero-mean random variables with variance σ_k^2 to allow for possible heteroskedasticity in the fit; i.e., a model, ξ, that relates φ_k to $\mathbf{x}_k$ for all $k \in U$, is specified as

$$\mathrm{E}_\xi[\varphi_k] = \mathbf{x}_k'\boldsymbol{\beta},\ \mathrm{Var}_\xi(\varphi_k) = \sigma_k^2,\ \mathrm{Cov}_\xi(\varphi_k, \varphi_l) = 0,$$

where the functional form of σ_k^2 is known.

If φ_k were observed for all $k \in U$, we could compute the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$,

$$\mathbf{B}_U = \left(\sum_{k\in U} \frac{\mathbf{x}_k\mathbf{x}_k'}{\sigma_k^2}\right)^{-1} \sum_{k\in U} \frac{\mathbf{x}_k\varphi_k}{\sigma_k^2}.$$

If φ_k were observed for all $k \in s$, we could compute the weighted estimate of the BLUE

$$\mathbf{B}_s = \left(\sum_{k\in s} \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k\sigma_k^2}\right)^{-1} \sum_{k\in s} \frac{\mathbf{x}_k\varphi_k}{\pi_k\sigma_k^2},$$

and obtain the residuals $E_{ks} = \varphi_k - \mathbf{x}_k'\mathbf{B}_s,\quad k \in s$.

Since φ_k is observed only for $k \in r \subset s$, compute

$$\mathbf{B}_r = \left(\sum_{k\in r} \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k\pi_{k|s}\sigma_k^2}\right)^{-1} \sum_{k\in r} \frac{\mathbf{x}_k\varphi_k}{\pi_k\pi_{k|s}\sigma_k^2}$$

and residuals $e_{kr} = \varphi_k - \mathbf{x}_k'\mathbf{B}_r = \varphi_k - \hat{f}_k$, $k \in r$.

2.3 Two-phase regression estimation

We are interested in $\Theta = \sum_U \Phi_k$, but as this is unavailable we work with $\theta = \sum_U \varphi_k$. Through appropriate choice of φ_k, θ may be, for example, a spatial average, a domain total, or the proportion of sites in the region with values above some threshold.

Given the fitted metamodel, we use the two-phase regression estimator (e.g., Särndal et al., 1992, chapter 9):

$$\begin{aligned}
\hat{\theta}_r &= \sum_{k\in s} \frac{\mathbf{x}_k'\mathbf{B}_r}{\pi_k} + \sum_{k\in r} \frac{e_{kr}}{\pi_k\pi_{k|s}} \\
&= \left(\sum_{k\in s} \frac{\mathbf{x}_k'\mathbf{B}_s}{\pi_k} + \sum_{k\in r} \frac{E_{ks}}{\pi_k\pi_{k|s}}\right) \\
&\quad + \left(\sum_{k\in r} \frac{\mathbf{x}_k'}{\pi_k\pi_{k|s}} - \sum_{k\in s} \frac{\mathbf{x}_k'}{\pi_k}\right)(\mathbf{B}_s - \mathbf{B}_r) \\
&=: \hat{\theta}_* + \delta,
\end{aligned}$$

where $\hat{\theta}_*$ is a hypothetical estimator which could be computed given φ_k, $k \in s$. It follows that the estimation error is

$$\begin{aligned}
\hat{\theta}_r - \theta &= \hat{\theta}_* - \theta + \delta \\
&= \left(\sum_{k\in s} \frac{\varphi_k}{\pi_k} - \sum_{k\in U} \varphi_k\right) \\
&\quad + \left(\sum_{k\in r} \frac{E_{ks}}{\pi_k\pi_{k|s}} - \sum_{k\in s} \frac{E_{ks}}{\pi_k}\right) + \delta. \quad (1)
\end{aligned}$$

The first term is the phase one error, which would be incurred even if the simulation model was run for all $k \in s$. The phase one error is unbiased for

zero. The second term is the phase two error, incurred because the metamodel is used in place of the computer simulation model. The phase two error is exactly unbiased for zero, *even if the metamodel is completely misspecified.* If the metamodel has good predictive ability, then the $\{E_{ks}\}$ are small and the variance of the phase two error is small.

The remaining term, δ, is the product of two factors, one being an unbiased estimator of zero, the other being an approximately unbiased estimator of zero. Hence δ is close to zero with high probability and is of smaller order than $\hat{\theta}_*$, and so

$$\mathrm{E}\left[\hat{\theta}_r\right] = \mathrm{E}\left[\mathrm{E}\left[\hat{\theta}_r \mid s\right]\right] \doteq \mathrm{E}\left[\mathrm{E}\left[\hat{\theta}_* \mid s\right]\right] = \theta,$$

even under metamodel misspecification, an important consideration when results from environmental studies may be controversial.

2.4 Variance approximation

Ignoring δ in (1), the approximate variance of $\hat{\theta}_r$ is given by

$$\mathrm{AV}(\hat{\theta}_r) = \mathrm{Var}\left(\mathrm{E}\left[\hat{\theta}_* | s\right]\right) + \mathrm{E}\left[\mathrm{Var}\left(\hat{\theta}_* | s\right)\right] = V_1 + V_2$$

where V_1 is the variance due to the first phase sampling and V_2 is due to the second phase sampling. Let $t_i = \sum_{k \in U_i} \varphi_k$, $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$ and $\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}$. Then

$$V_1 = V_{1PSU} + V_{1SSU} = \sum_{h=1}^{H}\left(V_{Ih} + \sum\nolimits_{U_{Ih}} \frac{V_{IIi}}{\pi_{Ii}}\right)$$

where

$$V_{Ih} = \sum\sum\nolimits_{U_{Ih}} \Delta_{Iij} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}}$$

and

$$V_{IIi} = \sum\sum\nolimits_{U_i} \Delta_{kl|i} \frac{\varphi_k}{\pi_{k|i}} \frac{\varphi_l}{\pi_{l|i}}.$$

Also,

$$V_2 = \mathrm{E}\left[\sum\sum\nolimits_{s} \Delta_{kl|s} \frac{E_{ks}}{\pi_k \pi_{k|s}} \frac{E_{ls}}{\pi_l \pi_{l|s}}\right]$$

where $\Delta_{kl|s} = \pi_{kl|s} - \pi_{k|s}\pi_{l|s}$.

2.5 Variance estimation

Estimation of V_1

If φ_k was observed for $k \in s$, a possible estimator of V_1 (e.g., Särndal et al., 1992, p. 139) would be

$$\hat{V}_{1*} = \sum_{h=1}^{H} \hat{V}_{1h*}$$

where

$$\hat{V}_{1h*} = \frac{n_{Ih}}{n_{Ih}-1} \sum\nolimits_{s_{Ih}} \left(\frac{\hat{t}_{i\pi}}{\pi_{Ii}} - \frac{\hat{t}_{h\pi}}{n_{Ih}}\right)^2$$

with $\hat{t}_{i\pi} = \sum_{s_i} \varphi_k / \pi_{k|i}$ and $\hat{t}_{h\pi} = \sum_{s_{Ih}} \hat{t}_{i\pi} / \pi_{Ii}$. Expanding the square in the expression for $\hat{V}_{1h*}$,

$$\hat{V}_{1h*} = \frac{n_{Ih}}{n_{Ih}-1}\left(\sum\nolimits_{s_{Ih}} \sum\sum\nolimits_{s_i} \frac{\varphi_k}{\pi_k}\frac{\varphi_l}{\pi_l} - \frac{1}{n_{Ih}} \sum\sum\nolimits_{s_h} \frac{\varphi_k}{\pi_k}\frac{\varphi_l}{\pi_l}\right).$$

When simple random sampling is used in the first stage,

$$\mathrm{E}\left[\hat{V}_{1*}\right] = \sum_{h=1}^{H} \frac{V_{Ih}}{1 - n_{Ih}/N_{Ih}} + V_{1SSU}$$

and $\hat{V}_{1*}$ would thus be a conservative estimator of V_1. Since we only observe φ_k for $k \in r$ we cannot use $\hat{V}_{1*}$, but one suggested approach is to use an estimator $\hat{V}_{1r}$ such that $\mathrm{E}\left[\hat{V}_{1r} \mid s\right] \doteq \hat{V}_{1*}$.

If $\pi_{kl|s} > 0$ for all $k\&l \in s$, one estimator of V_1, based on familiar Horvitz-Thompson estimation principles, is

$$\hat{V}_{1,HT} = \sum_{h=1}^{H} \hat{V}_{1,HT,h}$$

where

$$\hat{V}_{1,HT,h} = \frac{n_{Ih}}{n_{Ih}-1}\left(\sum\nolimits_{s_{Ih}} \sum\sum\nolimits_{r_i} \frac{1}{\pi_{kl|s}} \frac{\varphi_k}{\pi_k}\frac{\varphi_l}{\pi_l} - \frac{1}{n_{Ih}} \sum\sum\nolimits_{r_h} \frac{1}{\pi_{kl|s}} \frac{\varphi_k}{\pi_k}\frac{\varphi_l}{\pi_l}\right)$$

with $r_i = r \cap s_i$ and $r_h = \cup_{i \in s_{Ih}} r_i$. Clearly

$$\mathrm{E}\left[\hat{V}_{1,HT} \mid s\right] = \hat{V}_{1*}.$$

Despite the fact that $\hat{V}_{1,HT}$ has the desired expected value conditioned upon the first-phase sample s, it can result in negative estimates, especially when H is large in combination with small sample sizes in the two stages of the first phase, as is the case for the 1992 NRI. Simulation studies indicate that the problem may be severe even after adding $\hat{V}_2$, the estimator of V_2, in order to obtain an estimate of the total variance $V_1 + V_2$. For these circumstances, we suggest an alternative estimator of V_1 which borrows strength from the regression estimators of the PSU and stratum totals. Let

$$\frac{\hat{t}_{ir}}{\pi_{Ii}} = \sum\nolimits_{s_i} \frac{\hat{f}_k}{\pi_k} + \sum\nolimits_{r_i} \frac{e_{kr}}{\pi_k \pi_{k|s}}$$

and

$$\hat{t}_{hr} = \sum_{s_{Ih}} \frac{\hat{t}_{ir}}{\pi_{Ii}}$$

where $\hat{f}_k = \mathbf{x}_k' \mathbf{B}_r$. The alternative estimator is

$$\hat{V}_{1,reg} = \sum_{h=1}^{H} \hat{V}_{1,reg,h}$$

where

$$\hat{V}_{1,reg,h} = \frac{n_{Ih}}{n_{Ih}-1}\Big[\sum_{s_{Ih}} \left(\frac{\hat{t}_{ir}}{\pi_{Ii}} - \frac{\hat{t}_{hr}}{n_{Ih}}\right)^2 - \sum_{s_{Ih}} \sum\sum_{r_i} \frac{\Delta_{kl|s}}{\pi_{kl|s}} \frac{e_{kr}}{\pi_k \pi_{k|s}} \frac{e_{lr}}{\pi_l \pi_{l|s}} + \frac{1}{n_{Ih}} \sum\sum_{r_h} \frac{\Delta_{kl|s}}{\pi_{kl|s}} \frac{e_{kr}}{\pi_k \pi_{k|s}} \frac{e_{lr}}{\pi_l \pi_{l|s}}\Big].$$

A related approach which does not rely on regression estimation is given in Kott (1990).

To justify $\hat{V}_{1,reg}$, note that $\mathrm{E}\left[\hat{t}_{hr} \mid s\right]^2 \doteq \hat{t}_{h\pi}^2$ and $\mathrm{E}\left[\frac{\hat{t}_{ir}}{\pi_{Ii}} \,\middle|\, s\right]^2 \doteq \left(\frac{\hat{t}_{i\pi}}{\pi_{Ii}}\right)^2$, while

$$\mathrm{E}\left[\hat{t}_{hr}^2 \mid s\right] \doteq \hat{t}_{h\pi}^2 + \mathrm{Var}\left(\hat{t}_{hr} \mid s\right)$$

and

$$\mathrm{E}\left[\left(\frac{\hat{t}_{ir}}{\pi_{Ii}}\right)^2 \,\middle|\, s\right] \doteq \left(\frac{\hat{t}_{i\pi}}{\pi_{Ii}}\right)^2 + \mathrm{Var}\left(\frac{\hat{t}_{ir}}{\pi_{Ii}} \,\middle|\, s\right).$$

Since

$$\mathrm{E}\left[\sum\sum_{r_h} \frac{\Delta_{kl|s}}{\pi_{kl|s}} \frac{e_{kr}}{\pi_k \pi_{k|s}} \frac{e_{lr}}{\pi_l \pi_{l|s}} \,\middle|\, s\right] \doteq \mathrm{Var}\left(\hat{t}_{hr} \mid s\right)$$

and

$$\mathrm{E}\left[\sum\sum_{r_i} \frac{\Delta_{kl|s}}{\pi_{kl|s}} \frac{e_{kr}}{\pi_k \pi_{k|s}} \frac{e_{lr}}{\pi_l \pi_{l|s}} \,\middle|\, s\right] \doteq \mathrm{Var}\left(\frac{\hat{t}_{ir}}{\pi_{Ii}} \,\middle|\, s\right),$$

it follows that

$$\mathrm{E}\left[\hat{V}_{1,reg} \mid s\right] \doteq \hat{V}_{1*}.$$

Under conditions for which a regression estimator outperforms the usual Horvitz-Thompson estimator in terms of lower variability, we expect $\hat{V}_{1,reg}$ to perform well; see section 3.

Remark: The variance estimators above are based on the assumption that $n_{Ih} > 1$ for all h. When the design is such that $n_{Ih} = 1$ for one or more h, the technique of collapsed strata may be used. In doing so, a positive bias is induced. In applications where the stratification is imposed mainly because of administrative reasons, this problem is of less concern than when a strong stratification effect is present.

Estimation of V_2

An approximately unbiased estimator of V_2, the component of the variance due to the second phase of randomization, is given by

$$\hat{V}_2 = \sum\sum_{r} \frac{\Delta_{kl|s}}{\pi_{kl|s}} \frac{e_{kr}}{\pi_k \pi_{k|s}} \frac{e_{lr}}{\pi_l \pi_{l|s}}.$$

3 Monte Carlo results

To assess the performance of the two-phase regression estimator and the associated variance estimators in the setting of a realistic computer simulation experiment, we selected as the first-phase sample s the set of all NRI points in Missouri which were classified as cropland in 1992. The first-phase sample remained fixed throughout this Monte Carlo experiment. The sample s was restratified on the basis of groups of Major Land Resource Areas (MLRAs), which are geographical units defined on the basis of soil and land cover characteristics. The values φ_k were simulated from a general linear model which approximates the erosion output of EPIC-WQ in tons/hectare/year (Lakshminarayan and Babcock, 1996); this single simulated realization of $\{\varphi_k\}$ remained fixed in the experiment. An additional study variable, $\varphi_k I_{k\in\{\text{Chariton County}\}}$, was used to estimate the total for a particular small domain.

For each replication $i = 1, \ldots, 1000$ of the experiment, the second-phase sample r was constructed by independently drawing 10% simple random samples without replacement from each MLRA group. A linear metamodel was fitted via weighted least squares using a subset of the available regressors. Predictors for each study variable were then constructed.

The two-phase regression estimator $\hat{\theta}_r$ and the variance estimators $\hat{V}_{1,HT}$, $\hat{V}_{1,reg}$, $\hat{V}_2$, $\hat{V}_{HT} = \hat{V}_{1,HT} + \hat{V}_2$ and $\hat{V}_{reg} = \hat{V}_{1,reg} + \hat{V}_2$ were calculated for each replication and each study variable. Table 1 reports simulation biases and root mean squared errors (RMSEs) for the above estimators, computed relative to the estimands $\mathrm{E}\left[\hat{\theta}_* \mid s\right]$, $\hat{V}_{1*}$, $\hat{V}_{1*}$, $\mathrm{Var}_{MC}(\hat{\theta}_r \mid s) \doteq \mathrm{Var}(\hat{\theta}_* \mid s)$, $\hat{V}_{1*} + \mathrm{Var}_{MC}(\hat{\theta}_r \mid s)$ and $\hat{V}_{1*} + \mathrm{Var}_{MC}(\hat{\theta}_r \mid s)$, respectively, where "MC" stands for "Monte Carlo." Thus, these simulation results are all conditional on s. Table 1 also reports the percentage of negative variance estimates.

For both study variables, the regression-type variance estimators dominate the Horvitz-Thompson variance estimators in terms of RMSE and percentage of negative estimates.

Estimator	Bias	RMSE	% < 0
Total, $\hat{\theta}_r$	-2.0e+3	5.5e+4	—
$\hat{V}_{1,HT}$	-2.3e+7	6.8e+8	1.6
$\hat{V}_{1,reg}$	4.1e+6	2.2e+8	0.0
$\hat{V}_2$	1.0e+7	3.9e+8	0.0
$\hat{V}_{HT}$	-1.3e+7	8.3e+8	0.0
$\hat{V}_{reg}$	1.4e+7	5.3e+8	0.0
Domain			
Total, $\hat{\theta}_r$	-3.7e+1	8.3e+3	—
$\hat{V}_{1,HT}$	-4.0e+6	4.4e+7	25.8
$\hat{V}_{1,reg}$	-8.7e+5	1.8e+7	8.7
$\hat{V}_2$	-7.1e+5	4.4e+7	0.0
$\hat{V}_{HT}$	-4.7e+6	5.9e+7	4.5
$\hat{V}_{reg}$	-1.6e+6	4.3e+7	0.0

Table 1: *Biases and root mean squared errors (RMSEs) for two-phase regression estimation of the regional total, a small domain total, and the associated variances. Also reported is the percentage of negative variance estimates. Values are based on 1000 simulated draws of a stratified simple random sample, r, from Missouri cropland points, s, in the 1992 NRI sample.*

Acknowledgements: The first author is grateful to the University of Örebro for supporting him as a visiting graduate student at Iowa State University, where this work was carried out. The work of the second author was partially supported by Cooperative Agreement 68-6114-6-73 with the National Resources Conservation Service, and the work of the second and third authors was partially supported by the Office of Naval Research Grant #N000149610279.

References

Bernardo, D. J., H. P. Mapp, G. J. Sabbagh, S. Geleta, K. B. Watkins, R. L. Elliot and J. F. Stone. (1993). Economic and environmental impacts of water quality protection policy in agriculture. *Water Resources Research* **29**, 3081–3091.

Bouzaher, A., P.G. Lakshminarayan, R. Cabe, A. Carriquiry, P. Gassman, and J. Shogren. (1993). Metamodels and nonpoint pollution policy in agriculture. *Water Resources Research* **29**, 1579–1587.

Carriquiry, A.L., F.J. Breidt and P.G. Lakshminarayan (1996). Sampling schemes for policy analyses using computer simulation experiments. CARD Working Paper 96–WP 153, Iowa State University.

Dean, J.D. P.S. Huyakorn, A.S. Donigian, Jr., K.A. Voos, R.W. Schanz, and R.F. Carsel. (1989). Risk of unsaturated/saturated transport and transformation of chemical concentrations (RUSTIC), vol II. User's guide. *Rep. US EPA/600/3-89/048B*, EPA, Washington, DC.

Dillaha, A. and J.A. Gale. (1992). Nonpoint source modeling for evaluating the effectiveness of best management practices. National Water Quality Evaluation Project Notes, No. 52, North Carolina Cooperative Extension Service, N.C. State University, Chapel Hill.

Donigian, A.S., Jr., D.W. Meyer and P.P. Jowise. (1986). Stream transport and agricultural runoff of pesticides for exposure: a methodology. *Rep. US EPA/600/3-86/011A*, EPA, Washington, DC.

Evans, B.M. and W.L. Meyers. (1990). A GIS-based approach to evaluating regional groundwater pollution potential with DRASTIC. *Journal of Soil Water Conservation* **45**, 242–245.

Goebel, J. J. and H. D. Baker. (1982). *The 1982 National Resources Inventory Sample Design and Estimation Procedures.* Survey Section, Statistical Laboratory, Iowa State University, Ames.

Kleijnen, J. P. C. (1987). *Statistical Tools for Simulation Practitioners.* Dekker, New York.

Kott, P.S. (1990). Variance estimation when a first phase area sample is restratified. *Survey Methodology* **16**, 99–103.

Lakshminarayan, P. G., and B. A. Babcock. (1996). Temporal and spatial evaluation of soil conservation policies. CARD Working Paper 96–WP 149, Iowa State University.

Lakshminarayan, P. G., S. R. Johnson, and A. Bouzaher. (1995). A multi-objective approach to integrating agricultural, economic, and environmental policies. *Journal of Environmental Management* **45**, 365-378.

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

Taub, F. B. and L. A. Burns. (1991). Advances in experimental approaches to estimate the exposure of ecosystems and ground water. In: Freshe, H. (ed.), *Pesticide Chemistry*, VCH Publishers, New York.

U.S. Environmental Protection Agency. (1992). Managing nonpoint source pollution: final report to Congress on section 319 of the Clean Water Act, 1989. EPA-506/9-90. Washington, D.C.

Williams, J.R., C.A. Jones, and P.T. Dyke. (1988). EPIC, the erosion productivity impact calculator. Technical report, USDA, Agricultural Research Service, Temple, TX, 1988.