Sampling Schemes for Policy Analyses Using Computer Simulation Experiments

Alicia L. Carriquiry, F. Jay Breidt, and P.G. Lakshminarayan

Working Paper 97-WP 184
November 1997

Sampling Schemes for Policy Analyses Using Computer Simulation Experiments

Alicia L. Carriquiry, F. Jay Breidt,
and P.G. Lakshminarayan

Working Paper 97-WP 184
November 1997

Center for Agricultural and Rural Development
Iowa State University
Ames, IA 50011-1070

Alicia L. Carriquiry is an associate professor in the Department of Statistics and the Center for Agricultural and Rural Development (CARD), Iowa State University; F. Jay Breidt is an associate professor in the Department of Statistics, Iowa State University; and P.G. Lakshminarayan is an economist with American Express TRS, Phoenix, Arizona.

This paper is a Journal Paper No. J-17632 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 813186, and supported by Hatch Act and State of Iowa Funds.

For more information, please correspond with Dr. Alicia Carriquiry, Department of Statistics, Iowa State University, Snedecor Hall, Ames, IA 50011. E-mail: alicia@iastate.edu.

Sampling Schemes for Policy Analyses Using Computer Simulation Experiments

Abstract

Evaluating the environmental and economic impacts of agricultural policies is not a simple task. A systematic approach to evaluation would include the effect of policy-dependent factors (such as tillage practices, crop rotations, and chemical use) as well as the effect of policy-independent covariates (such as weather, topography, and soil attributes) on response variables (such as amount of soil eroded or chemical leached into the groundwater). For comparison purposes, the effects of these input combinations on the response variable would have to be assessed under competing policy scenarios. Because the number of input combinations is high in most problems, and because policies to be evaluated are often not in use at the time of the study, practitioners have resorted to simulation experiments to generate data. But generating data from simulation models is often costly and time consuming; thus, the number of input combinations in a study may be limiting even in simulation experiments. In this paper, we discuss the problem of designing computer simulation experiments that require generating data for just a fraction of the possible input combinations. We propose an approach that is based on subsampling the 1992 National Resources Inventory (NRI) points. We illustrate the procedure by assessing soil erosion in a situation where there are "observed" data (reported by the Natural Resources Conservation Service (NRCS)) for comparison. Estimates for soil erosion obtained using the procedure we propose are in good agreement with NRCS reported values.

Sampling Schemes for Policy Analyses Using Computer Simulation Experiments

We begin with an overview of simulation-based inferential methods for agricultural economic policy analyses, in which real data of interest such as long-run averages of soil lost to erosion or chemicals leached to groundwater are not available due to the infeasibility of field experiments and the high costs of monitoring. Briefly, biogeophysical process 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 *sample* of inputs in what is known as a computer simulation experiment. Researchers then typically fit a regression "metamodel" to the results of this computer simulation experiment, and finally the fitted metamodel is used to make projections about environmental impacts at different spatial scales and under different policy scenarios. All of this is described in more detail in the following subsections.

Our goal in this paper is to improve the design of computer simulation experiments; that is, the sampling schemes used for selecting the input combinations on which the biogeophysical process model is run.

Agricultural Policy Analysis

Agricultural practices have significant impacts on the environment. One important example is water quality problems caused by agricultural nonpoint source (NPS) pollution (US EPA, 1992). Another example is soil erosion resulting from the application of certain management practices in combination with cropping systems (Ribaudo, 1986). Good policy would guide agricultural activity to have minimum undesirable impact on the environment, while not compromising on economic efficiency. Therefore, for informed policy decisions, it is important to evaluate different policy alternatives in a systematic fashion, accounting not only for the economics of agricultural activities, but also for their impact on the environment (Bouzaher and others, 1995).

Evaluating the environmental and economic impacts of agricultural policies is not a simple task. Suppose that environmental impact is quantified via some *response variable* such as chemical leaching into the groundwater, soil erosion, or nitrogen runoff. Inputs that affect the response variable may be of at least two different types: (1) *factors* that are subject to change via policy, and (2) *covariates* that, while not susceptible to policy, still have an effect on the response (soil characteristics, weather, and topography). We denote by $\mathbf{x}_k$ the combined vector of factors and covariates.

For example, researchers interested in assessing the impact of agricultural activity on soil erosion would have to consider the effect of factors such as management practices and cropping systems and rotations, as well as covariates such as soil and weather characteristics, since it is well known that all of these inputs have a significant effect on erosion (Lal and Elliot, 1994) and must be accounted for when assessing the

economic and environmental consequences of a given policy.

A similar case is that of regulating the use of triazines on corn and soybeans, which is currently under debate. The potential environmental damage of triazines depends on soil and weather covariates, but also on factors such as the amount of the chemical applied, time of application, tillage systems, and other management practices. These inputs interact in complex ways, and introduction of new policies may introduce new complications. For example, farmers faced with regulations on the use of triazines are likely to substitute the chemical by choosing from a variety of alternative products, any of which may also have an environmental impact (Lakshminarayan and others, 1996).

Ideally, evaluation of sustainable agricultural practices would proceed via the "traditional" field experimental approach. A typical example is the Management Systems Evaluation Area (MSEA) experiments (Ward and others, 1994). An experiment would be planned where differences among the relevant factor combinations would be tested by measuring their effect on the chosen response variables, while controlling for the effects of the covariates.

From an experimental viewpoint, consideration of all feasible input combinations results in three major problems:

1. The number of input combinations (soil types, chemicals, weather, management practices) that arise when evaluating environmental impacts of alternative agricultural policies can be dizzyingly large.

2. "Candidate" policies are often not in place at the time of the study, and therefore data cannot be collected in the traditional way.

3. Even for those policies currently in place, data on environmental indicators may be expensive or even impossible to collect. Consider, for example, the long-run average amount of soil lost to erosion for a given set of management practices on a farm.

These problems become much more pronounced if policy evaluations are required at large regional levels.

Biogeophysical Process Models

Because environmental monitoring and field experimentation are very expensive and time consuming, policy makers interested in evaluating potential environmental impacts of different agricultural policies are increasingly relying on data generated from biogeophysical process models which simulate, for example, the fate and transport of chemicals in different media (Taub and Burns, 1991; Wagenet and Hutson, 1991).

Specific simulation models for physical processes include the *Erosion Productivity Impact Calculator* (EPIC) model (USDA, 1990), the *Risk of Unsaturated/Saturated Transport and Transformation of Chemical Concentrations* (RUSTIC) system (Dean

and others, 1989), and the *Surface Transport and Agricultural Runoff of Pesticides for Exposure Assessment* (STREAM) model (Donigian and others, 1986). For a site k with input variables $\mathbf{x}_k$, these physical process models produce a simulated response denoted by y_k; i.e., $y_k = \varphi(\mathbf{x}_k)$. Typically, a simulated observation is generated as the average over a large number of "yearly" outputs from a long term (15–30 years) simulation run of a process model, using historical weather data. In this context, a "yearly" output is defined as the amount of soil eroded or chemical that leaches as a result of one cycle of agricultural activity under specified weather conditions.

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. While much more economical than monitoring, these simulation models are practical for site-specific problems only (Evans and Myers, 1990), because they are often costly and time consuming to operate, and thus generating data for a large number of input combinations may become impractical. 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 and others, 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.

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* (Dillaha and Gale, 1992; Bouzaher and others, 1993; Lakshminarayan and others, 1995). A computer simulation experiment in this context consists of running the simulation model at a probability sample of input combinations representative of the region of interest, in which every input combination has a known, positive probability of inclusion in the sample (e.g., Särndal and others, 1992). Under probability sampling, unbiased estimators of population parameters such as means and totals can be formed without appealing to any assumed statistical model. (A brief review of probability sampling concepts used in this paper is provided in Section 2.)

For the kinds of evaluations in which we are interested, the simulation experiment must accomplish two important goals:

- Reduce the number of input combinations used in the simulation program(s) to a manageable number.
- Generate output that is representative of the study area, so that inferences for the area can be drawn with acceptable statistical reliability.

One method of reducing the number of input combinations is by sampling directly from the set of all possible input combinations. In Section 3, we review one such

sampling scheme which draws a sample of soils from a soil series data base. Soils are stratified by soil attributes, then randomly selected with probability proportional to size.

We then propose an alternative in which the sample of inputs is obtained as 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. The sampling design is a multi-stage stratified approach that uses the 1992 NRI points as a basis and further stratifies according to crop and crop rotations. Sampling rates from this study can then be combined with the 1992 NRI expansion factors to obtain statistically reliable estimates at the regional, state, and Major Land Resource Area (MLRA) levels.

Metamodels

For environmental impact assessment at the regional level, the use of complex process simulation models such as EPIC, 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 biogeophysical process model (Kleijnen, 1987; Bouzaher and others, 1993). A metamodel is fitted to the data from a computer simulation experiment using standard statistical techniques. A metamodel approximates the output of the simulation model, but has a simpler functional form and, typically, fewer inputs. For example, a metamodel may be a linear multiple regression model, with output denoted by

$$\hat{y}_k = \mathbf{x}_k' \hat{\beta}, \tag{1}$$

where $\hat{\beta}$ is a vector of coefficients estimated by regressing the y_k's from the computer simulation experiment on the input variables $\mathbf{x}_k$. (Some components of $\hat{\beta}$ might be zero if only a subset of the input variables are used in the metamodel.)

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 and others (1993), Bouzaher and others (1993), and Lakshminarayan and others (1995).

The remainder of this paper is organized as follows. In Section 2, we review the methods of probability sampling useful in understanding the design and estimation procedures we describe. A sampling scheme for soils is briefly reviewed in Section 3,

and the new NRI-based sampling scheme we propose is presented in Section 3.2 after a description of the NRI in Section 3.1. For illustrative purposes, we apply the proposed method to the problem of estimating sheet and rill erosion in Section 4. This is an artificial example which is useful because erosion values reported by the Natural Resources Conservation Service (NRCS; formerly the Soil Conservation Service (SCS)) are available for comparison. Section 5 contains a summary.

Review of Probability Sampling

Consider a population of elements denoted by a set U of labels k. For example, k might index the spatial location of a segment of land, with U a given region (e.g., a state county, or watershed), or k might index a particular combination of simulation model inputs, with U the set of all possible combinations of inputs. A *probability sample* is a randomized selection of a subset of labels, $s \subset U$, where the *inclusion probability*

$$\Pr[k \in s] = \pi_k > 0$$

is known for all $k \in U$; that is, all elements of the population have a known, positive probability of being included in the sample, s. Measurements y_k are then obtained on the selected elements, $k \in s$. This process of drawing the sample and obtaining the measurements is a kind of designed observational study called a *survey sample.* Cochran (1977) is a standard reference.

Probability sampling, as opposed to purposive selection of "representative" elements or haphazard selection of convenient elements, is now a standard scientific tool since it guards against selection biases and it leads to objective inferences. In particular, inferences can be drawn about the population without appealing to any assumed statistical model. If a model is desired, probability sampling is a good method for collecting the data used to fit the model, because observations are likely to be well distributed throughout the design space. Probability sampling has a particularly long history in resource studies, such as soil mapping, forest inventories, and crop surveys. See Schreuder and others (1993), Chapter 1, for an overview of the history of probability sampling.

An important feature of many probability sampling designs is *stratification.* In stratification, the population is deliberately divided into disjoint, homogeneous subpopulations, called *strata*, and independent samples are drawn from each stratum. Stratification can allow for precise stratum-level estimation and stratum-to-stratum comparisons since the sample size for each stratum can be determined in advance. Further, a well-chosen stratified sampling design can yield estimators with less variability than the corresponding estimators from an unstratified design. For these reasons, stratification is nearly always employed in surveys of real populations.

Real populations are often naturally subdivided into disjoint groups of elements called *clusters* ; e.g., people live in households. Often, it is more convenient or less

costly to sample clusters of elements than to sample the elements themselves. In resource studies, this is typically because sampling elements will lead to a widely-scattered sample with high logistical costs. There is usually some loss of efficiency in drawing a probability sample of clusters and observing all elements in each selected cluster instead of drawing a sample of elements directly since elements within a cluster are often positively correlated. Because of this positive correlation, there is often little loss in efficiency, and substantial reduction in cost, if a probability sample of elements within each selected cluster is drawn, a procedure known as *two-stage sampling*. In two-stage sampling, the clusters are sampled first and hence are called *primary sampling units* or PSUs.

The basic principle of estimation in probability sampling is to compute the inclusion probability π_k of element k, weight the measurements on k inversely proportional to π_k, and sum the weighted measurements over the sample,

$$\sum_{k \in s} y_k / \pi_k. \tag{2}$$

This is the famous Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and it has the desirable property that, provided $\pi_k > 0$ for all k in the population, it is *unbiased*: that is, its average value over all possible probability samples is the true population total. For two-stage sampling, element inclusion probabilities are computed via the multiplication rule of elementary probability:

$$\Pr[k \in s] = \Pr[k \in s | k\text{'s cluster selected}] \Pr[k\text{'s cluster selected}].$$

Though a well-designed probability sample can yield precise estimates, further efficiency is often gained by the effective use of auxiliary information. Often this takes the form of a regression-type estimator,

$$\hat{t}_{reg} = \sum_{k \in U} \mathbf{z}_k' \hat{\gamma} + \sum_{k \in s} \frac{y_k - \mathbf{z}_k' \hat{\gamma}}{\pi_k}, \tag{3}$$

where $\mathbf{z}_k$ is a column vector of auxiliary variables observable for element k, $\sum_{k \in U} \mathbf{z}_k$ is a column vector of *known* totals of the auxiliary variables, and

$$\hat{\gamma} = \left(\sum_{k \in s} q_k \mathbf{z}_k \mathbf{z}_k' \right)^{-1} \sum_{k \in s} q_k \mathbf{z}_k y_k$$

is a vector of estimated regression coefficients, obtained via least squares regression (weighted by q_k) of the sample y_k's on the sample $\mathbf{z}_k$'s. Special cases of the regression estimator $\hat{t}_{reg}$ include the poststratified estimator, simple ratio estimator, separate and combined ratio estimators, and the simple regression estimator (e.g., Särndal, et al., 1992). The regression estimator is, to a very good approximation, unbiased for the population total.

It is important to note that though the regression estimator can be motivated by modeling the measurements $\{y_k\}$ as possibly heteroskedastic, uncorrelated random variables with means $\mu_k = \mathbf{z}_k'\gamma$, the approximate unbiasedness of the estimator still holds *even if the model is misspecified.* The better the specification of the model, the better the efficiency of the regression estimator relative to that of the Horvitz-Thompson estimator of (2).

After some algebra, the regression estimator can be rewritten as

$$\hat{t}_{reg} = \sum_{k\in s}\left[1 + \left(\sum_{k\in U}\mathbf{z}_k - \sum_{k\in s}\frac{\mathbf{z}_k}{\pi_k}\right)'\left(\sum_{k\in s}q_k\mathbf{z}_k\mathbf{z}_k'\right)^{-1}\mathbf{z}_k\right]\frac{y_k}{\pi_k} = \sum_{k\in s}w_{ks}y_k, \quad (4)$$

from which it is apparent that $\hat{t}_{reg}$ has the weighted-sum form of the Horvitz-Thompson estimator, but the inverse inclusion probability weights are modified to take account of the auxiliary information in $\mathbf{z}_k$. The modified weights w_{ks} are called *regression weights* or *expansion factors.* All 1992 NRI points have associated expansion factors estimated as in (4). We revisit these expansion factors in the next section.

In addition to increasing the precision of estimates for characteristics correlated with $\mathbf{z}_k$, regression estimation makes certain sample estimates consistent with the known control totals, $\sum_{k\in U}\mathbf{z}_k$; specifically,

$$\sum_{k\in s}w_{ks}\mathbf{z}_k = \sum_{k\in U}\mathbf{z}_k.$$

The weighted estimators can be used to estimate population totals (hence means and proportions as well) and linear combinations of population totals. Subpopulation totals are computed by summing the weighted observations over the sampled elements within the subpopulation, and subpopulation means and proportions are handled analogously. Non-linear functions which take the form of functions of population totals,

$$\theta = g(t_1, \ldots, t_q),$$

can be effectively estimated by substituting the weighted estimators:

$$\hat{\theta} = g(\hat{t}_1, \ldots, \hat{t}_q).$$

For special studies (i.e., those with more expensive measurements), it is common to select a subsample s^* of a large-scale probability sample s, a procedure known in the sampling literature as *two-phase sampling* or double sampling. The advantage of subsampling is that data from the large sample s may be used in the design and estimation for the subsample s^*. In particular, information from the large sample may be used at the design stage for efficient stratification of the subsample, and may be used at the estimation stage for regression estimation in the subsample.

Let π_k^* denote the probability of selecting element k for the subsample; in general this probability will depend on s. Then, a weighted estimator analogous to (2) is

$$\sum_{k \in s^*} \frac{w_{ks} y_k}{\pi_k^*}, \tag{5}$$

and a regression estimator analogous to (3) is

$$\hat{t}_{reg}^* = \sum_{k \in s} w_{ks} \mathbf{x}_k' \hat{\beta} + \sum_{k \in s^*} \frac{w_{ks}(y_k - \mathbf{x}_k' \hat{\beta})}{\pi_k^*}, \tag{6}$$

where $\mathbf{x}_k$ is a column vector of auxiliary variables observable for element k, $\sum_{k \in s} w_{ks} \mathbf{x}_k$ is a column vector of totals estimated from the large sample, and $\hat{\beta}$ is a vector of estimated regression coefficients, obtained via least squares regression (possibly weighted) of the subsample y_k's on the subsample $\mathbf{x}_k$'s.

For a complicated design and/or estimation procedure, the problem of producing standard errors for estimated characteristics is quite complex. Many estimation techniques for obtaining standard errors in complex surveys have been developed; an excellent summary is given by Wolter (1985).

A Sampling Scheme Based on the 1992 NRI

The statistical reliability of the analytic approach outlined in Section 1 depends in great measure on the sample of input combinations used for generating the simulation outcomes. Consider the problem of generating simulation outcomes for an environmental indicator in the major crop belt of the United States—the north-central region. This region is shown in Figure 1.

Suppose that soils, chemicals, tillage practices and crop rotations are relevant inputs for the computer simulation experiment. In the SOILS 5 data base[2], there are approximately 2,141 different soils that appear in the region in Figure 1. If 20 chemicals, 10 crop rotations, and four tillage practices are to be considered, then there are $2,141 \times 20 \times 10 \times 4 = 1.7$ million possible input combinations for which to generate a pseudo observation (assuming all combinations are viable). Furthermore, each soil unit is represented by more than one layer (profile), up to a maximum of six layers, so that the number of combinations expands alarmingly depending on the number of soil layers. It is therefore necessary to reduce the number of these input combinations.

[2]The SOILS 5 data base is a layered soil series developed by the NRCS as part of their Soil Interpretation Record System (SIRS).

One sampling method, described in Gassman et al. (1994), selects soils according to those soil properties that are deemed most relevant regarding their effect on the environmental indicator under study. As an example, let percent sand, clay, and organic matter, bulk density, and pH be the properties to be considered. Soils are stratified by their properties and selected at a fixed rate, guaranteeing a representative sample of soils.

Allocating a level of each of the additional factors (tillage, crop rotation, chemical, weather, etc.) to a sampled soil can be done in several ways. The simplest method is to allocate factors to soils at random. This method, while simple to carry out, has a serious drawback: the biogeophysical model may end up being run for sets of inputs that do not "exist," and the combined range of all inputs may be different than the population range. In this case, predictions based on metamodels may be extreme extrapolations, and produce misleading results. An alternative procedure incorporates information about the actual set of factors found in combination with each soil in the population, and uses this information to do a proportional allocation of factors to soils. This method has the advantage of producing inputs for process models that are not a product of the researcher's imagination, and thus predictions from metamodels will require no extrapolation. The disadvantage, however, is the amount of additional information that needs to be collected. To alleviate this problem while at the same time keeping the number of simulation runs manageable, without compromising on heterogeneity of production practices and chemical alternatives, we suggest an alternative sampling scheme based on linked NRI 92 and SOILS 5 data bases.

Selecting NRI points, rather than soils, to build the sample presents several advantages over the method outlined in Gassman and others (1994). In the remainder of the section, we give a brief description of the 1992 NRI and present a sampling method to construct a subsample from the 1992 NRI sample.

The 1992 NRI

The 1992 NRI is a multi-resource inventory collected by the NRCS of the United States Department of Agriculture. Information gathered includes the status, condition, and trends of land, soil, water, and related resources on nonfederal land in the 50 states, Puerto Rico, and the U.S. Virgin Islands.

The basic region used for constructing the NRI sample was the county (or analogous regions in some states). Samples were collected within counties, following a stratified two-stage area sampling procedure. Strata were formed geographically from parts of townships in regions covered by the Public Land Survey, and from analogous divisions elsewhere. This procedure guarantees a good spreading of the points over the United States.

In the first stage of sampling, land areas were the primary sampling units (PSUs). A typical PSU was a square area, 1/2-mile on a side, containing 160 acres, but PSU

sizes varied according to the heterogeneity of the area. Thus in heterogeneous areas such as those under irrigation, the PSU was as small as 40 acres, and in homogeneous areas such as range and forest lands, PSUs might be as large as 640 acres. The sampling rates for the PSUs varied from county to county, depending on factors such as the size of the county, the type of agricultural activity, and the number of counties in the state.

In the second stage, points were selected within each sampled PSU using a restricted random procedure that guaranteed that selected points were spread throughout the PSU. A detailed description of the stratified two-stage area sampling method used for drawing the NRI sample is given in Nusser and Goebel (1997).

Data for the 1992 NRI were collected for more than 800,000 locations. The sampling design guarantees that inferences at the national, state, and MLRA levels can be made in a statistically reliable manner. Each NRI point is accompanied by an expansion factor w_{ks} that assigns each point the appropriate weight under the design and available auxiliary information. Expansion factors were computed using a regression estimation procedure as described in equation (4), but with further constraints to guarantee non-negative weights. These weights have the property that relevant control variables add up to the correct totals when obtaining estimates at the regional, state, and MLRA levels.

Data collected by the NRI can be organized into several general categories, including soil characteristics and interpretation, land cover and use, erosion measures, cropping history, and conservation practices. This extensive information makes it unnecessary to link with other data bases to access point-level information on agricultural activities. The NRI, therefore, can be used to design the computer experiments in which we are interested.

Subsampling the 1992 NRI Points

The sampling design we propose draws a stratified sample from the NRI points in the region of interest. Points are stratified by crops and crop rotations within MLRAs; i.e., points within each MLRA are classified into strata according to crop and crop rotation information. As an example, an MLRA containing 50 NRI points may have some points under a corn/soybean rotation, and the remainder under continuous corn.

Points within each stratum are selected as follows:

1. Determine the sampling rate (or inclusion probability in the subsample). For the purposes of our example, let the sampling rate be equal to 10%.

2. Assign to each MLRA a sample size given by 10% of the number of NRI points within that MLRA. Compute the 1992 NRI estimated acreages for each crop or crop rotation classification within the MLRA, and proportionally allocate the MLRA sample size to the crop/crop rotation strata; e.g., if corn/soybean represents 80% of the acres in the MLRA and continuous corn represents 20%

of the acres, then four points would be randomly chosen from the corn/soybean stratum and one point would be selected from the continuous corn stratum, for a sample size of five out of a total of 50 points in the MLRA.

In this sampling scheme, soil properties are implicitly accounted for since the design for the 1992 NRI guarantees a good spatial spreading of the points. Since soil properties are geographically distributed, it is to be expected that the range for most soil attributes will be represented in the NRI (see Figure 2). Figure 2 compares the frequency distributions of clay, bulk density, pH, and organic matter for the population of NRI points and for a 10% subsample in the north-central United States.

Furthermore, varying inclusion probabilities for each point in the NRI partially reflect and account for varying degrees of heterogeneity in land and soil characteristics across the United States. This implies that those areas with high variability in land/soil characteristics were sampled at a higher rate. Thus, a sample drawn at random from the universe of 1992 NRI points, with constant inclusion probabilities for each point, should also be representative for soil attributes (see Table 1). Table 1 summarizes the mean of key soil properties for both the population of NRI points and a 10% subsample selected as described above. The sample range for each attribute of interest can be expected to increase as the inclusion probability increases. In addition, since the 1992 NRI was drawn as an area sample, those soils that occupy a larger surface are more likely to be selected in the subsample. Note that there is good agreement between the mean soil properties in the NRI and in the 10% subsample.

Estimates at the regional, state, and MLRA levels can be obtained in a statistically reliable manner by combining the 1992 NRI expansion factors with the sampling rates used to draw the subsample as in equation (5). For example, if inclusion probability in the subsample is set at 0.1, and the expansion factor for the kth point in the 1992 NRI was estimated to be w_{ks}, then the new expansion factor in the subsample is given by the ratio $w_{ks} \div 0.1 = w_{ks} \times 10$. For a discussion on the problem of estimation at different levels using the NRI points, the reader is referred to Nusser and Goebel (1997).

As mentioned earlier, computer experiments to assess environmental impacts from alternative policies must account for several factor effects. The use of the 1992 NRI points as the sampling frame partially solves the problem of allocating factors to points. Soil properties, crop, and crop rotation effects were brought into the experiment through the sampling scheme for NRI points described in the previous subsection.

Not all important factors and covariates are available from NRI; in particular, weather is not available. Instead, weather variables associated to each NRI point are those obtained from historical observations from the nearest weather station. Confidentiality issues determine that latitude and longitude information for each NRI point not be available to the general public. However, by combining information on county, hydrogeological group, and primary sampling unit identification numbers it is possible to determine an approximate location for each point, within the boundaries of a spatially identifiable polygon.

An Application of the NRI Based Sampling Scheme to Estimate Sheet and Rill Erosion

Using the sampling method described in Section 3.2, we obtained a 10% sample of points from the 1992 NRI database for the north-central region of the United States. On these sites we ran the EPIC program to simulate soil erosion measurements. EPIC is a field-based crop growth and physical process simulation program developed by the Soil and Water Research Laboratory, U.S. Department of Agriculture (USDA, 1990). This model has been extensively tested and calibrated to Midwest conditions. Because EPIC (like other biogeophysical process models) is costly to operate, it was used to generate a response y_k only on the sites selected by the sampling method described in Section 3.

The simulated sheet and rill erosion data from EPIC were then used to fit a metamodel. Standard statistical regression procedures were employed to identify and estimate the metamodel. In this example, the estimated function was a linear multiple regression model, and regressors $\mathbf{x}_k$ in the model included location effects (latitude and longitude), variables related to soil properties and topography (%sand, %clay, pH, bulk density, organic matter, slope, slope length), and management practices (tillage systems), and indicator variables to represent various crop rotations.

Equation (1) was then used to predict erosion rates for all the NRI cropped points in the study region, and finally these predictions were weighted up using the NRI expansion factors and the subsampling inclusion probability to yield county-level predictions. In Figure 3, predicted erosion rates from the metamodel are compared to the annual average sheet and rill erosion rates reported by the NRCS, USDA. County-level maps of predicted sheet and rill erosion rates (top map) and average erosion rates reported by the NRCS (bottom map) are given in Figure 3.

Summary statistics for the county-level aggregates shown in Figure 3 are displayed in Table 2. Note that in spite of the fact that the model fitted to the data generated at the sampled points appears to suffer mild lack of fit, the statistics calculated from the predicted values and from the values reported by NRCS are similar. The

correlation coefficient between estimated erosion values and "observed" values (as reported by NRCS) was 0.82, indicating reasonable agreement between the two. The frequency distribution given in Table 2 shows that estimates of erosion based on a metamodel fitted to a 10% subsample properly selected have a distribution similar to the "observed" distribution.

Summary

There is an increasing need for regional scale agricultural NPS pollution assessments and identification of best management practices to regulate and eliminate NPS pollution. A cost effective and scientific approach to make regional assessment of NPS pollution is to use biogeophysical model simulation experiments. Because NPS pollution is highly heterogeneous, any experimental approach should consider all diverse physical and management factors. On a regional scale there are millions of such factor combinations, which makes it impractical if not impossible to simulate a response variable for each of them.

This paper outlined spatial sampling schemes for agricultural NPS pollution assessment using biogeophysical model simulation experiments. A unique feature of the sampling design is that it uses the NRI, a state-of-the-art geophysical database, as the sampling frame. Thus, the sampling protocol we propose results in a two-phase sample, with immediately available expansion factors for each element in the sample.

In Section 2 we pointed out the importance of using the appropriate expansion factors when making inferences at different levels of geographic aggregation. Estimates obtained in this manner are unbiased (or nearly unbiased) *regardless* of the form of the metamodel ultimately used for prediction. This again points to the importance of drawing probability samples when conducting simulation experiments; since all metamodels are simple approximations to the more accurate biogeophysical process models, the property of unbiasedness of estimators under any metamodel becomes significant.

We have not discussed the problem of deriving standard errors (or other measures of uncertainties) for quantities estimated from the metamodels. Properly accounting for all uncertainties in this type of evaluation is not a trivial matter; we are currently conducting research in this area.

Acknowledgments

The authors gratefully acknowledge the constructive comments of the referees, Tommy Daniel, Dan Pote, and David Verbyla.

Literature Cited

Bernardo, D.J., Mapp, H.P., Sabbagh, G.J., Geleta, S., Watkins, K.B., Elliot, R.L., and J.F. Stone. 1993. Economic and Environmental Impacts of Water Quality Protection Policy in Agriculture. *Water Resources Research* 29:3081–3091.

Bouzaher, A., Lakshminarayan, P.G., Cabe, R., Carriquiry, A.L., Gassman, P. and J. Shogren. 1993. Metamodels and Nonpoint Pollution Policy in Agriculture. *Water Resources Research* 29:1579–1587.

Bouzaher,A., Shogren, J., Holtkamp, D., Carriquiry, A.L., Gassman, P., Lakshminarayan, P.G., Archer, D., Reese, R., Furtan, W., Izaurralde, R.C., and J. Kiniry. 1995. Agricultural Policies and Soil Degradation in Western Canada: An Agroecological Economic Assessment. Report 3: The Integration of the Environmental and the Economic Component. CARD Staff report 95-SR 75. Iowa State University.

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

Dean, J.D., Huyakorn, P.S., Donigian, Jr., A.S., Voos, K.A., Schanz, R.W., 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., Meyer D.W., 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.

Gassman, P., Izaurralde, R.C., Bouzaher, A. and P.G. Lakshminarayan. 1994. Assessing Soil Degradation in Western Canada with an Economic-Ecological Modeling System. In Proceedings of the International Conference on Modern Agriculture and the Environment, Revohot, Israel, October 2–6.

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.

Kellogg, R.L., TeSelle, G. and J. Goebel. 1994. Highlights from the 1992 NRI. *Journal of Soil and Water Conservation* 49:521–527.

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

Lakshminarayan, P.G., Johnson, S.R., and A. Bouzaher. 1995. A Multi-objective Approach to Integrating Agricultural, Economic, and Environmental Policies. *Journal of Environmental Management* 45:365-378.

Lakshminarayan, P.G., Bouzaher, A. and J. Shogren. 1996. Atrazine and Water Quality: An Evaluation of Alternative Policy Options. *Journal of Environmental Management.* In press.

Lal, R., and W. Elliot. 1994. Erodibility and Erosivity. In: *Soil Erosion Research Methods*, R. Lal (ed.). Soil and Water Conservation Society, Ankeny, Iowa.

Nusser, S.M. and J.J. Goebel. 1997. The National Resources Inventory: a Long-Term Multi-Resource Monitoring Program. *Environmental and Ecological Statistics.* In press.

Ribaudo, M.O. 1986. Reducing Soil Erosion: Offsite Benefit. Agricultural Economic Report 561, ERS, USDA, Washington, D. C.

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

Schreuder, H.T., Gregoire, T.G., and G.B. Wood. 1993. *Sampling Methods for Multiresource Forest Inventory.* Wiley, New York.

Taub, F.B. and L.A. Burns. 1991. Advances in Experimental Approaches to Estimate the Exposure of Ecosystems and Ground Water. In: *Pesticide Chemistry*, Helmet Freshe (ed.), VCH Publishers, New York.

U.S. Department of Agriculture. 1990. EPIC—Erosion Productivity Impact Calculator—Model Documentation. Technical Bulletin Number 1768, Agricultural Research Service, Washington, D. C.

U.S. Environmental Protection Agency. 1992. Managing Nonpoint Source Pol-

lution: Final Report to Congress on Section 319 of the Clean Water Act, 1989. EPA-506/9–90, Washington, D.C.

Wagenet, R.J. and J.L. Hutson. 1991. Modeling Pesticide Transport and Transformation in the Soil-Plant System. In: *Pesticide Chemistry*, Helmet Freshe (ed.), VCH Publishers, New York.

Ward, A.D., Hatfield, J.A., Lamb, J.A., Albers, E.E., Logan, T.J., and J.L. Anderson. 1994. The Management Systems Evaluation Areas Program: Tillage and Water Quality Research. *Soil and Tillage Research* 30:49-74.

Wolter, K.M. 1985. *Introduction to Variance Estimation.* Springer-Verlag, New York.

Table 1: *Summary statistics of key soil properties.*

Soil Property	Mean	
	Population	Sample
Clay (%)	27.746	27.765
K-Factor	0.366	0.366
AWC	0.167	0.166
Bulk Density (gms/cc)	1.422	1.422
Organic Matter (%)	1.124	1.165
Permeability (inch/hr)	1.652	1.630
pH	6.900	6.906
Slope (%)	2.471	2.465

Table 2: *Summary statistics for values of erosion estimated under the model, and for average values reported by NRCS.*

	Estimated	NRCS Reported
Statistic		
Mean erosion rate (tons/acre)	4.10	3.36
Standard deviation	6.06	5.53
Frequency Distribution (tons/acre)		
0 to 3	63%	72%
3 to 5	13%	12%
5 to 10	14%	9%
10 to 15	5%	3%
15 to 20	2%	1%
> 20	3%	2%

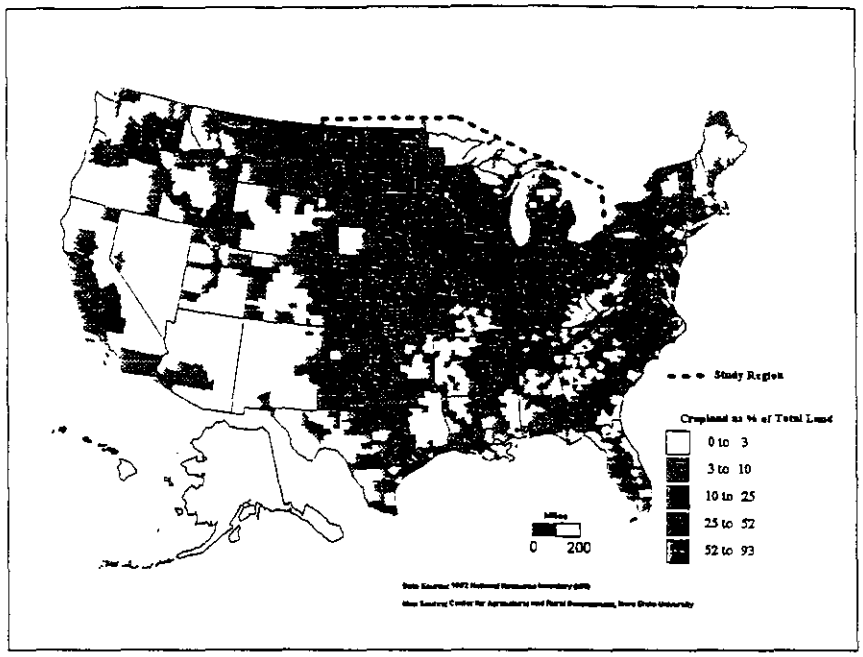


Figure 1. The study region: North Central United States.

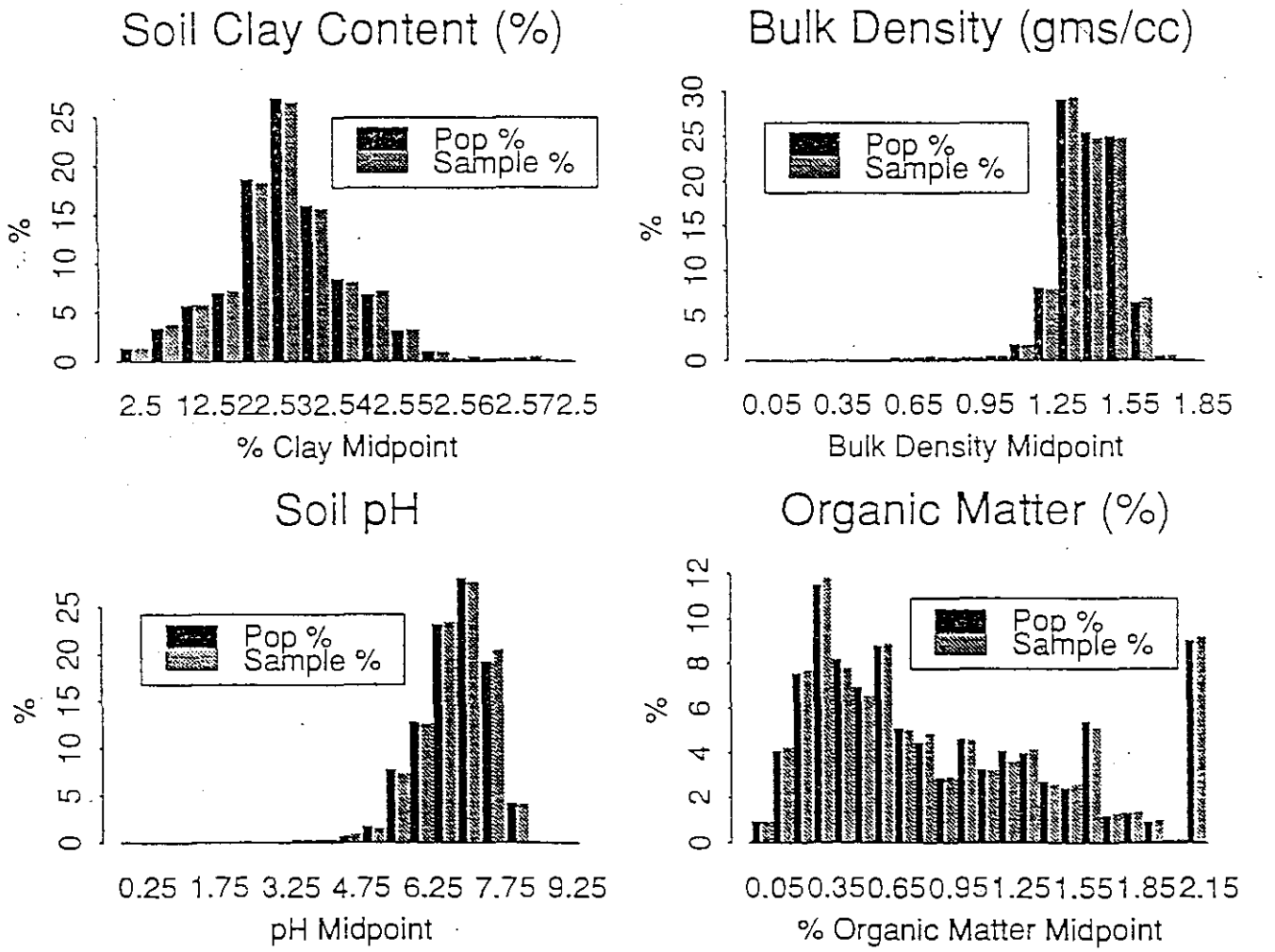
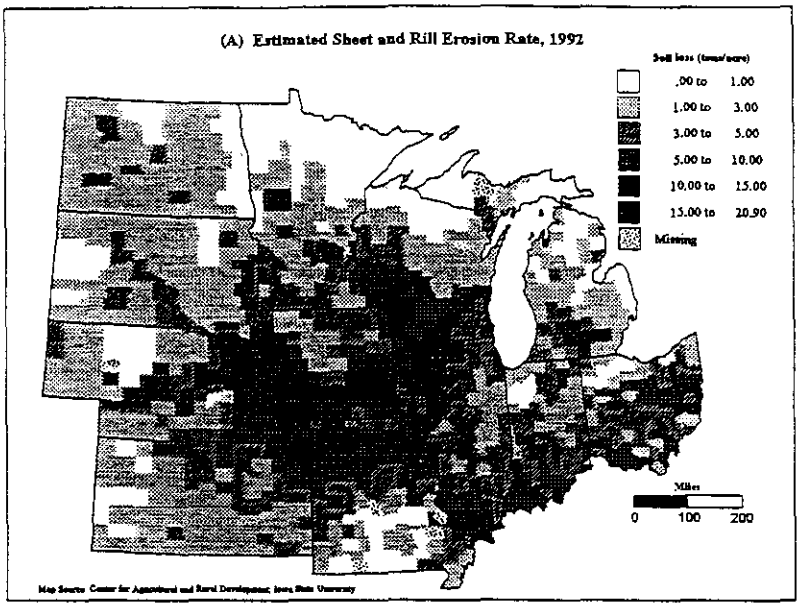


Figure 2: Frequency distributions of various soil properties in the population on NRI points and in the 10% subsample. (a) Clay (%); (b) Bulk density (gms/cc); (c) Soil pH; (d) Organic matter (%).



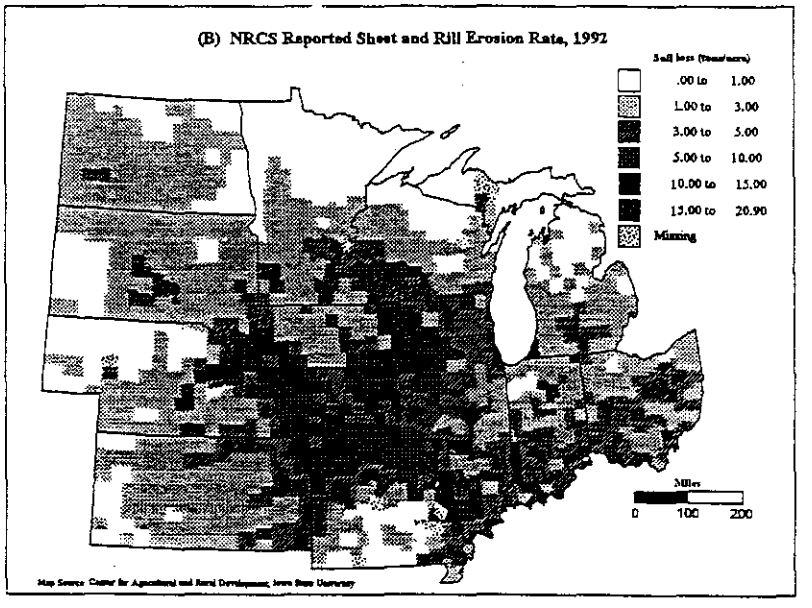


Figure 3: County-level maps of predicted sheet and rill erosion rates (top map); and average erosion rates reported by the NRCS (bottom map).