Sampling Schemes for Policy Analyses Using Computer Simulation Experiments

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

Working Paper 96-WP 153
February 1996

Sampling Schemes for Policy Analyses Using Computer Simulation Experiments

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

Working Paper 96-WP 153
February 1996

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

Alicia L. Carriquiry is an associate professor of statistics, CARD and the Department of Statistics, Iowa State University; F. Jay Breidt is an assistant professor, Department of Statistics, ISU; and P.G. Lakshminarayan is an associate scientist, CARD.

Corresponding author: Alicia L. Carriquiry, Iowa State University, 222 Snedecor Hall, Ames, Iowa 50011. Phone: (515) 294-3440; email: alicia@iastate.edu.

ABSTRACT

Evaluating the environmental and economic impacts of agricultural policies is not a simple task. A systematic approach to policy analysis would include investigating the effect of factors such as tillage practices, crop rotations, type and amount of chemical used, weather, topography, and other soil attributes, on observables such as amount of soil eroded or chemical leached into the groundwater. For comparison purposes, the effects of those factors on the response variable would have to be assessed under alternative policy scenarios. Because the number of factor levels is alarmingly high in most problems, and because policies to be evaluated are often not in place at the time of the study, practitioners have resorted to simulation experiments to generate data. In this paper, we discuss the problem of designing computer simulation experiments, and propose an approach that is based on subsampling the 1992 National Resources Inventory (NRI) points. We apply the procedure to the problem of assessing soil erosion under different policy scenarios.

Keywords: computer experiment, sampling, policy analysis.

Sampling Schemes for Policy Analyses Using Computer Simulation Experiments

1. Introduction

Agricultural practices may have significant impacts on the environment. One important example is the water quality problems caused by agricultural nonpoint source (NPS) pollution (EPA 1992). Another example is soil erosion resulting from the application of certain management practices in combination with cropping systems (Ribaudo, 1986). There are several policy alternatives to guide agricultural activity to have minimum undesirable impact on the environment, while not compromising on economic efficiency. For some policy alternatives, the trade-off between environmental impact and economic efficiency is significant. 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 et al., 1995).

Evaluating the environmental and economic impacts of agricultural policies is not a simple task. Consider, for example, the issue of soil erosion. Researchers interested in assessing the impact of agricultural activity on soil erosion would have to consider the effect of several factors. Among those are not only soil and weather characteristics, but also management practices and cropping systems and rotations, since it is well known that all these factors have a significant effect on erosion (Lal and Elliot, 1994). Thus, all of these factors need to be accounted for when assessing the economic and environmental consequences of the application 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, as well as other chemicals used in agriculture, depends not only on the amount of the chemical applied, but also on soil and weather conditions, time of application, and other management practices such as tillage. Furthermore, 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, Bouzaher, and Shogren, 1996).

Ideally, evaluation of sustainable agricultural practices would proceed via the "traditional" field experimental approach. A typical example is the Man-

agement Systems Evaluation Area (MSEA) experiments (Ward et al., 1994). An experiment would be planned where differences among the relevant factor combinations would be tested by measuring their effect on the chosen observables. In the case of policy analysis, factors would include not only those that are subject to variation as a consequence of the application of a certain policy (e.g., tillage practices, type and amount of chemical used, crops and crop rotations), but also those covariates that are likely to affect the observable outcome (e.g., weather conditions, topography, and soil attributes). For assessing environmental impacts from agricultural policies, observables might include the amount of a chemical that leaches into the groundwater, or the amount of a chemical that enters surface water through run-off, or the amount of soil eroded and the sediment damage caused by erosion.

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

1. The number of factor 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 such policy evaluations are required at large regional levels. Researchers and policy makers cannot therefore rely on the usual experimental approach to gather data, and must resort to other approaches when attempting to answer "what if?" type questions. An approach that has been gaining in popularity consists in simulating the measurements that would have been collected in an experiment had a field experiment been possible.

Biogeophysical process models to simulate, for example, the fate and transport of chemicals in different media, have been in use for some time

(Taub and Burns, 1991; Wagenet and Hutson, 1991). Soil erosion and leaching of chemicals and nutrients into the groundwater and transport in surface run-off and air are processes that can be simulated by using programs such as PRZM (Pesticide Root Zone Model), EPIC (Erosion Productivity Impact Calculator), SWAT (Surface Water Assessment Tool), and BLAYER (Boundary Layer), for a variety of inputs (soil and chemical attributes, weather conditions, and management practices, among them). The simulated observations generated from these programs have been used to draw conclusions regarding soil erosion and the fate and transport of chemicals in the environment.

When the problem of interest consists of assessing the impacts of different agricultural policies, then it becomes necessary to draw inferences about a region or a collection of regions (e.g., states, major land resource areas (MLRAs), etc.). If the study area is large, then the number of possible combinations of soils, crops, chemicals, and other factors such as tillage practices becomes unmanageable. In this case, it is necessary to design a *computer simulation experiment* using methods similar to those used when designing a traditional experiment (Dillaha and Gale 1992, Bouzaher et al. 1993).

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

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

One method of reducing the number of factor combinations is by sampling from the set of all possible factor combinations. In this report, we review one such sampling scheme, which draws a sample of soils from a soils series data base. Soils are stratified by soil attributes, then randomly selected with probability proportional to size.

We then propose an alternative factor-reducing sampling scheme, which draws a subsample from the 1992 National Resources Inventory (NRI) points in the Northcentral region of 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 MLRA levels.

2. Methods

We begin this discussion with an overview of the methods of probability sampling, then turn to the design of computer experiments for assessing environmental impacts of agricultural policies.

2.1 Probability Sampling

Consider a population of elements denoted by a set U of labels k; k might be thought of as the spatial location of a segment of land in what follows. A *probability sample* is a randomized selection of a subset of labels, $s \subset U$, where the *inclusion probability*

$$\mathrm{P}[k \in s] = \pi_k > 0$$

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,

Gregoire and Wood (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 makes the design flexible, since different probability sampling schemes can be used in different strata. 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 substantially 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. This may be because no adequate listing (*sampling frame*) of the elements is available, but a listing of the clusters is available, or it may be 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. Multistage sampling can also be used, in which case we have primary sampling units, secondary sampling units, ..., ultimate sampling units.

For special studies (i.e., those with more expensive measurements), it is fairly 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, of course, is that data from the large sample s may be used in the design and estimation for the subsample s^*.

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,

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

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:

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

Multi-stage sampling is handled analogously.

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,

$$\hat{t}_{reg} = \mathbf{t}_x' \hat{\mathbf{B}} + \sum_{k \in s} \frac{y_k - \mathbf{x}_k' \hat{\mathbf{B}}}{\pi_k},$$

where $\mathbf{x}_k$ is a $J \times 1$ column vector of auxiliary variables observable for element k, $\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$ is a column vector of known totals of the auxiliary variables, and

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

is a vector of estimated regression coefficients, obtained via weighted least squares regression of the sample y_k's on the sample $\mathbf{x}_k$'s. Also,

$$\sigma_k^2 = \sigma^2 \mathbf{x}_k' \boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is a vector of known constants. Special cases of the regression estimator $\hat{t}_{reg}$ include the poststratified estimator, simple ratio estimator,

separate and combined ratio estimators, simple regression estimator, and so on (e.g., Särndal, Swensson and Wretman 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 uncorrelated random variables with mean $\mu_k = \mathbf{x}_k'\beta$ and variance $\sigma_k^2 = \sigma^2 \mathbf{x}_k'\boldsymbol{\lambda}$, the approximate unbiasedness of the estimator still holds *even if the model is completely misspecified.* The better the specification of the model, the better the efficiency of the regression estimator relative to that of the Horvitz-Thompson estimator.

The regression estimator can be rewritten as

$$\hat{t}_{reg} = \sum_{k \in s} \left[1 + \left(\mathbf{t}_x - \sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} \right)' \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k \sigma_k^2} \right)^{-1} \mathbf{x}_k \right] \frac{y_k}{\pi_k} = \sum_{k \in s} w_{ks} y_k,$$

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 $\mathbf{x}_k$. The modified weights w_{ks} are called *regression weights* or *expansion factors.*

In addition to increasing the precision of estimates for characteristics correlated with $\mathbf{x}_k$, regression estimation makes certain sample estimates consistent with the known control totals, $\mathbf{t}_x$; specifically,

$$\sum_{k \in s} w_{ks} \mathbf{x}_k = \mathbf{t}_x.$$

The weighted estimators can be used to estimate population totals (hence means and proportions as well) and linear combinations of population totals. Many non-linear functions of interest take the form of functions of population totals,

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

which can be effectively estimated by substituting the weighted estimators:

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

This is exactly the principle employed in the estimated regression coefficient vector $\hat{\mathbf{B}}$ above; the population quantity

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

a nonlinear function of the population totals $\sum_{k \in U} x_{1k}^2$, $\sum_{k \in U} x_{1k} x_{2k}$, ..., $\sum_{k \in U} x_{Jk}^2$, $\sum_{k \in U} x_{1k} y_k$, ..., and $\sum_{k \in U} x_{Jk} y_k$, is estimated by plugging in the corresponding Horvitz-Thompson estimators.

Since the estimators in survey sampling often have the form of totals or functions of totals, versions of the central limit theorem are available for approximating the relevant sampling distributions and forming confidence intervals. Provided $v(\hat{\theta})$ is a consistent estimator of the variance of $\hat{\theta}$, the interval

$$\hat{\theta} \pm 1.96 \left\{ v(\hat{\theta}) \right\}^{1/2}$$

will cover the true population θ in approximately 95% of all samples.

For a complicated design and/or estimation procedure, the problem of variance estimation becomes quite complex. Many variance estimation techniques for data from complex surveys have been developed; an excellent summary is given by Wolter (1985).

2.2 Computer Simulation Experiments

Computer experiments have been in use for some time to assess environmental effects of alternative agricultural policies (Bouzaher et al. 1993, Bernardo et al. 1993). Researchers at CARD, Iowa State University, have proposed a novel approach to the systematic evaluation of the economic and environmental impacts of agricultural policies. The environmental component of this approach is a two-step process.

In the first step, simulation outcomes for chosen environmental indicators are generated from program runs for a sample of inputs from the set of all possible factor combinations. Typically, a simulated observation, for example on the amount soil lost to erosion, 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.

Other inputs to the program need to be determined as well. For the case of soil lost to erosion, the value of the simulation outcome depends on soil characteristics such as clay, sand and organic matter content, pH, and bulk density; on soil cover such as crop and crop rotation and weed populations; soil topography; and on several other factors including tillage practices and weather. In order to measure the erosion potential of a policy to be applied to a large region, say the Northcentral United States, it is therefore necessary to consider a very large number of factor combinations to be used as inputs for computer runs. The universe of possible inputs becomes unmanageable as the number of candidate policies and the size of the region increase.

Since it may be desired to "observe" the effect of a policy on every possible factor combination or to observe the effect of additional policies, the second step of the two-step procedure is to estimate metamodels from the set of simulation outcomes in step one. These metamodels can then be used for predicting the value of the environmental indicator for those factor combinations that were not included in the sample or policy scenarios that were not considered (Bouzaher et al., 1993; Lakshminarayan, Johnson, and Bouzaher, 1995); hence the environmental impacts of different policies at different sites can be assessed.

3. A Sampling Scheme Based on the 1992 NRI

The statistical reliability of the scheme outlined in the previous section depends in great measure on the sample of factor combinations used as input for generating the simulation outcomes in the first step. 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 factors to be used as inputs in the computer simulation experiment.

In the SOILS 5 data base[1], 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 factor combinations for which to generate a pseudo observation (assuming all factor 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 factor combinations expands alarmingly depending on the number of soil layers.

It is therefore necessary to draw a sample of these factor combinations. 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 indeed be extrapolations, and produce misleading results. An alternative procedure consists in incorporating information about the actual set of factors found in combination with each soil in the population, and using 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

[1]The SOILS 5 data base is a layered soil series developed by the Natural Resources Conservation Service (NRCS), formerly known as the Soil Conservation Service (SCS), as part of their Soil Interpretation Record System (SIRS).

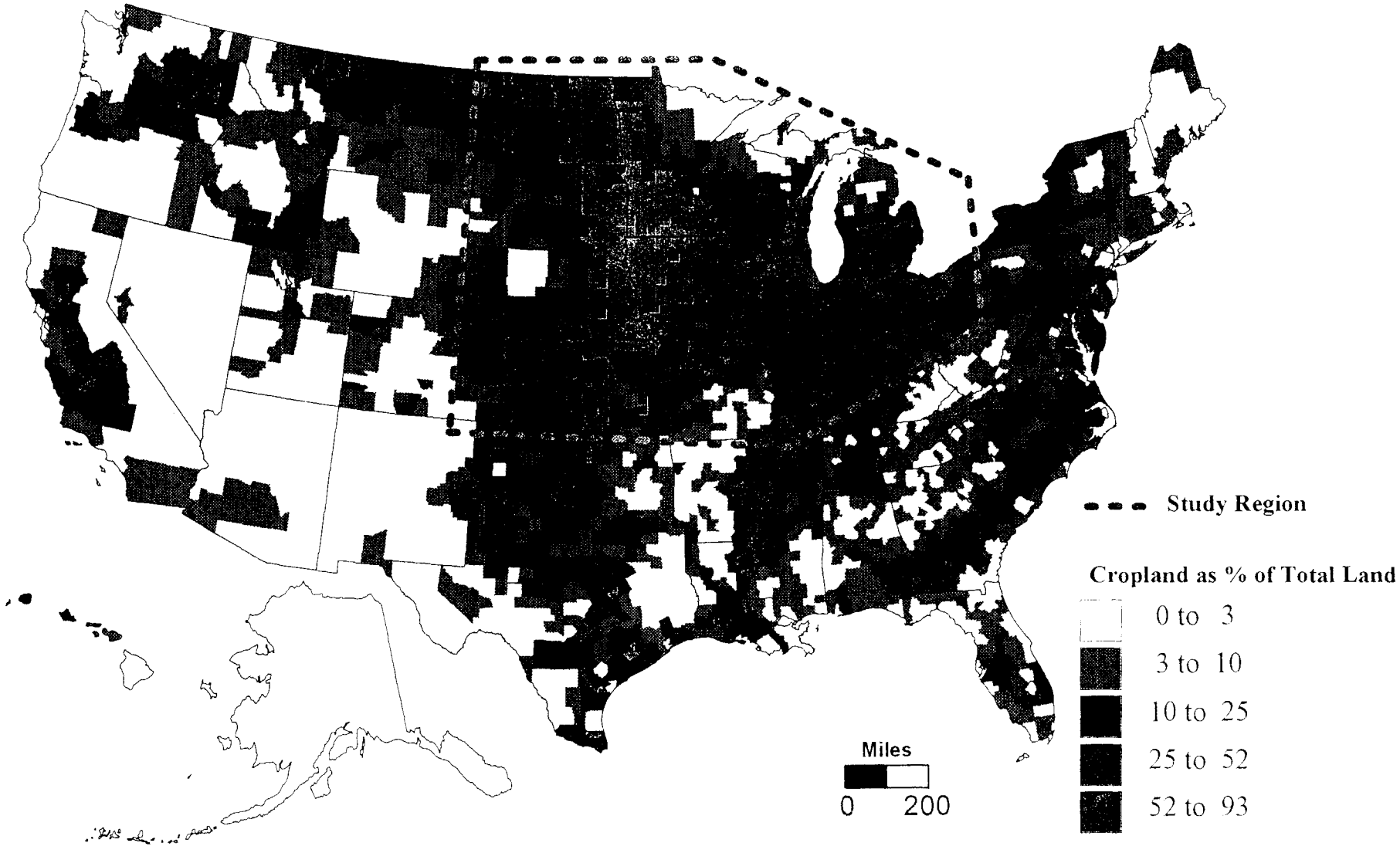


Figure 1. The study region: The north central United States

sampling scheme based on linked NRI 92 and SOILS 5 data bases.

The 1992 NRI is a multi-resource inventory collected by the Natural Resource Conservation Service (formerly the Soil Conservation Service) of the United States Department of Agriculture (USDA). 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 (e.g., Kellogg et al. 1994). An alternative design for a computer experiment then would use the most recent NRI as the sampling frame, and would select NRI points, rather than soils, to build the sample. This approach presents several advantages over the method outlined in Gassman (1994):

- The NRI points were selected according to a stratified two-stage sampling design that guarantees a good spreading of the points over the United States.

- Expansion factors have been estimated for each NRI point. These expansion factors are obtained by combining the sampling rates for each point and the post sampling weights computed so that relevant control variables add up to the correct totals when obtaining estimates at the regional, state, and substate (MLRA) levels. It is therefore possible to draw inferences at those levels with given statistical reliability.

- The information available for each NRI point is extensive. It includes not only soil characteristics, but also land use, cropping history, conservation practices, and so on. It is therefore unnecessary to link with other data bases in order to access information on agricultural activities at each point.

- Confidentiality issues determine that exact 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, to within the boundaries of a spatially identifiable polygon (Kellogg et al. 1992). Simulated outcomes generated in the experiment and aggregated over these polygons can then be used for spatial analyses.

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.

3.1 The 1992 NRI

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.

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 1982 NRI sample is given in Goebel and Baker (1982). The 1992 NRI sample was constructed in a similar manner.

Data for the 1992 NRI were collected for more than 800,000 locations. The sampling design guarantees that inferences at the national, regional, state, and substate (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 procedure closely related to the regression estimation procedure described above.

Data collected by the NRI can be organized into several general categories. Those categories that are relevant for policy analysis as described in this

document include

- Soil characteristics and interpretation.
- Earth cover.
- Land cover and use.
- Erosion.
- Land treatment (i.e., conservation tillage).
- Vegetative conditions.

The information provided by the NRI, therefore, can be used to design the computer experiments in which we are interested.

3.2 Subsampling the 1992 NRI Points

The sampling design we propose consists in drawing 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 in our example, 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.

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

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.

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 the subsample. The sample range for each attribute of interest can be expected to increase as the inclusion probability increases. In addition, since the 1992 was drawn as an area sample, those soils that occupy a larger surface are more likely to be selected in the subsample.

Estimates at the regional, state, and MLRA levels can be obtained in a

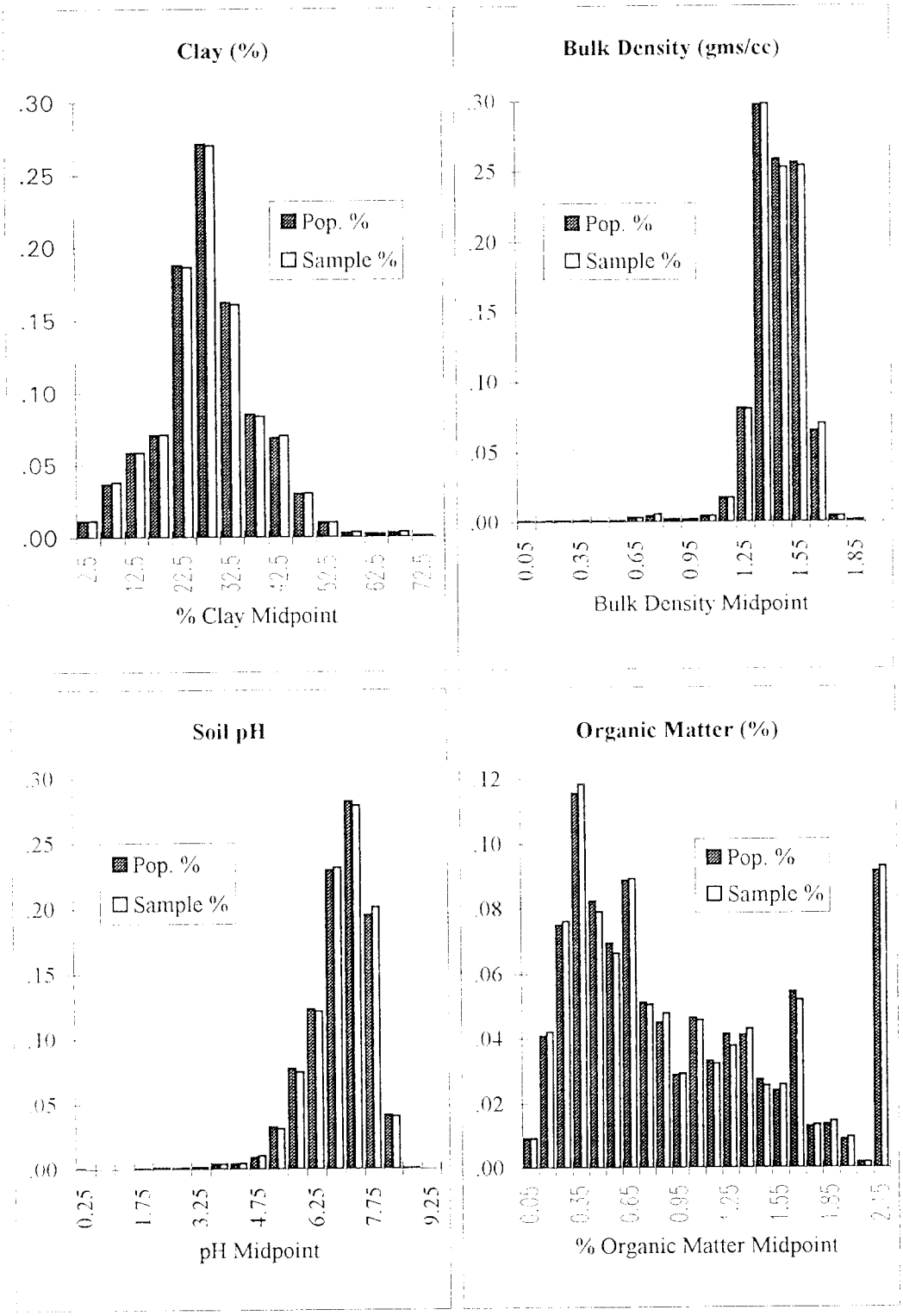


Figure 2. Frequency distributions of clay, bulk density, pH, and organic matter for the population of NRI points and for a 10 percent subsample.

statistically reliable manner by combining the 1992 NRI expansion factors with the sampling rates used to draw the subsample. 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 Goebel and Baker (1982).

3.3 Other Experimental Factors

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 and crop and crop rotation effects were brought into the experiment through the sampling scheme for NRI points described in the previous subsection. Other factor of potential importance is weather.

Weather variables associated to each NRI point are those obtained from historical observations from the nearest weather station. Consider, as an example, the region shown in Figure 1. There are 359 weather stations in this region, with records that go back at least 30 years. Since the approximate geographic location of each point in the subsample can be determined (as in Kellogg et al. 1992), it is possible to assign each point to its closest weather station, so that "real" weather conditions can be used when running the simulations. Given that the number of subsampled points can be expected to be much larger than the number of weather stations in the region, the full set of weather conditions in the study area gets replicated multiple times in the study, without an increase in the number of experimental points.

4. 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 Northcentral region. 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. A statistical response function summarizing the input-output relationship captured in the EPIC simulated outcome for sheet and rill erosion was estimated for the subsampled NRI points. This response function, a simple regression model explaining the complex simulation program outcomes by a subset of input parameters, is a useful tool for summarizing EPIC simulated output. Because of their ease of use, estimated response functions are widely used in agricultural nonpoint source pollution assessments using integrated modeling systems (Kleijnen, 1987; Bouzaher et al., 1993).

An EPIC simulation experiment is a set of executions of the simulation model intended to approximate the values of a response variable y (in this case, sheet and rill erosion) associated to a vector of site-specific physical and management variables v. For statistical purposes, it would be preferable to experiment with the real-life system itself rather than a simulation model of the system. However, this would mean incurring the cost and delay of waiting, in this case for 30 years of weather to present itself to the real-life system. Let g be the unknown, true function relating the response variable y to the input vector v

$$y = g(v).$$

Given the EPIC output, we can specify an analytical response function f, that is an approximation to g, with relatively few inputs. Letting $\{x, s, r, u\}$ represent management inputs, soil characteristics, topography and hydrological characteristics, and random error, respectively, the response function can be written as

$$y = f(x, s, r, u).$$

Standard statistical regression procedures were employed to identify and estimate the function f which approximates the true, unknown function g. The estimated function was a linear model, and regressors in the model included location effects, variables related to soil properties and topography, and to management practices, and indicator variables to represent various crop rotations.

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

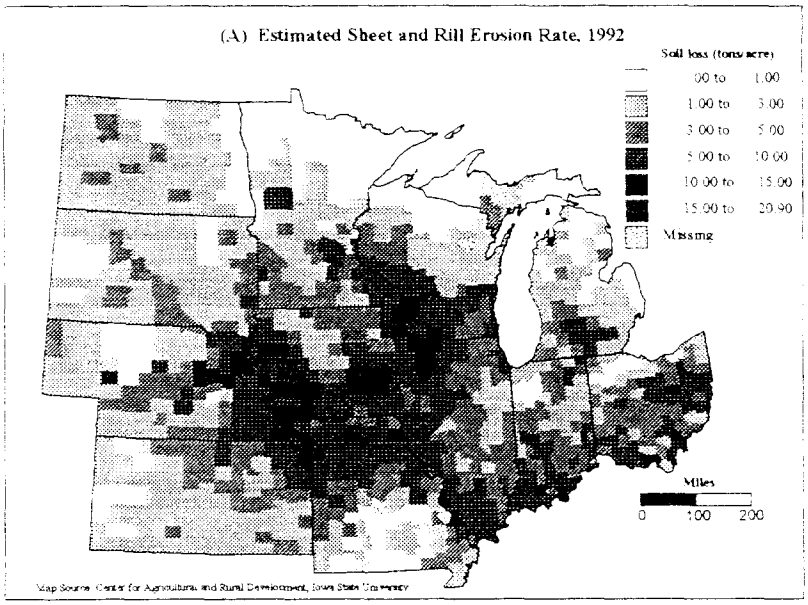
	Estimated	NRCS Reported
Statistic		
Mean erosion rate (tons/acre)	4.10	3.36
Standard deviation	6.06	5.53
Correlation	0.82	0.82
Frequency Distribution		
0 to 3	63%	72%
3 to 5	13%	12%
5 to 10	14%	9%
10 to 15	5%	3%
15 to 20	2%	1%
> 20	3%	2%

Using the regression model estimated for the EPIC simulated sheet and rill erosion values, we predicted erosion rates for all the NRI cropped points in the study region. In Figure 3, erosion rates for the population predicted using the estimated response function fitted to the sampled points 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.

5. Conclusions

There is an increasing need for regional scale agricultural NPS pollution



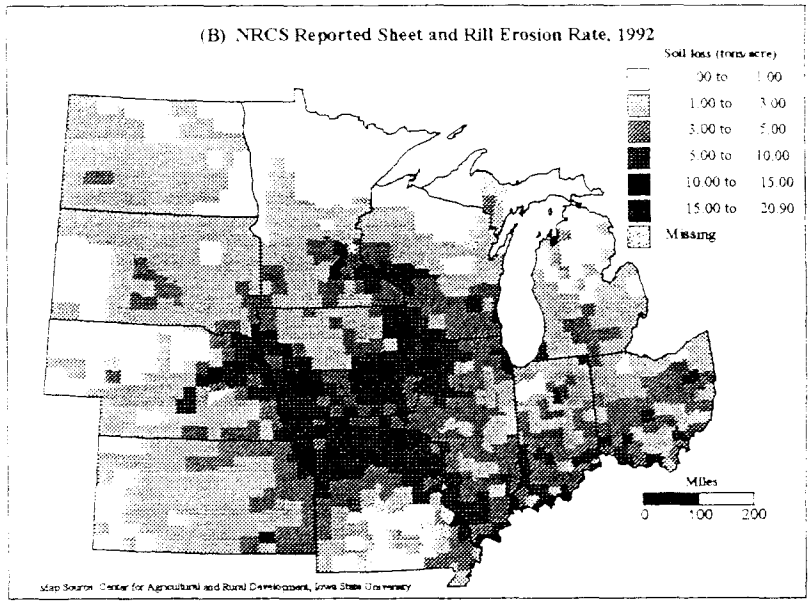


Figure 3. County-level maps of regression model predicted sheet and rill erosion rates (top map) and average erosion rates reported by the NRCS, USDA (bottom map).

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 make it impractical if not impossible to evaluate all factor combinations. This paper outlined spatial sampling schemes for diverse agricultural NPS pollution assessment using biogeophysical model simulation experiments. A unique feature of the sampling design is that it is based on a state-of-the-art geophysical database called NRI.

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. L. Carriquiry, Gassman, P. and J. Shogren. 1993. "Metamodels and Nonpoint Pollution Policy in Agriculture," *Water Resources Research* **29**, 1579–1587.

Bouzaher,A., J. Shogren, D. Holtkamp, A.L. Carriquiry, P. Gassman, P. G. Lakshminarayan, D. Archer, and R. Reese, ISU; W. Furtan, University of Saskatchewan; R. C. Izaurralde, University of Alberta; J. Kiniry, USDA/ARS. "Agricultural Policies and Soil Degradation in Western Canada: An Agro-ecological Economic Assessment. Report 3: The Integration of the Environmental and the Economic Component." CARD Staff report 95-SR 75, 1995.

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

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.

Gassman, P., R.C. Izaurralde, A. Bouzaher, 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.

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.

Horvitz, D.G. and D.J. Thompson 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statis-*

tical Association **47**, 663–685.

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

Kellogg, R.L. M.S. Maizel and D.W. Goss. 1992. Agricultural Chemical Use and Ground Water Quality: Where are the Potential Problem Areas? USDA.

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

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.

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

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

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

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

Schreuder, H.T., T.G. Gregoire 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*, ed. Helmet Freshe, 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 Pollution: 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*, ed. Helmet Freshe, VCH Publishers, New York.

Ward, A. D., J. L. Hatfield, J. A. Lamb, E. E. Albers, T. J. Logan, 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.