BAYESIAN HIERARCHICAL MODELING OF NRI SURVEY DATA

Michael D. Larsen, Iowa State Univeristy
Department of Statistics, 216-A Snedecor Hall, Ames, Iowa 50011-1210, larsen@iastate.edu

Keywords: Clustering. Gibbs sampling. Log normal distribution. Multilevel model. Stratification. Survey sampling. Survey weights.

Acknowledgement:

This work was supported in part by the grant "Statistical Methods for the National Resource Inventory" from the USDA NRCS. The author would like to thank Oz Zengin for his help accessing NRI data and colleagues including Dr. Zengin, Sarah Nusser, Wayne Fuller, and Jim Kienzler the Center Survey Statistics and Methodology (CSSM) at Iowa State University (ISU) for discussions concerning the NRI.

Disclaimer

The numbers reported in this article are not official results of the National Resource Inventory (NRI). They should not be considered official or approximate results in any sense for any characteristic monitored by the NRI. Comments in this article do not necessarily represent the opinion of USDA, NRCS, CSSM, ISU, or members of these organizations. The author solely is responsible for the content of this article.

Abstract:

The USDA Natural Resource Conservation Service (NRCS)in collaboration with Iowa State University conducted the National Resource Inventory (NRI) every five years from 1982 to 1997. The survey design was changed in 2000 to an annual supplemented panel design. The two-stage stratified area sample survey selects land segments as the primary sampling units, then points within the segments as the secondary sampling units. The land segments are highly stratified. Information is gathered on land use and the presence of water, trees, roads, and structures on nonfederal lands using a combination of fly-over photography and site visits. The results are used in erosion modeling and monitoring Conservation Reserve Program (CRP) lands and wetlands. Given the stratification and geographical organization of the segments and points, one could consider using models to incorporate the spatial and temporal relationships into estimation. One could also use subject matter knowledge to utilize associations between covariate and outcome variables in estimation. This work is the first step in developing hierarhical models for analyzing NRI survey data. Models are developed for the purpose of estimating average soil erosion per acre in 2003 due to wind in Iowa. Models are fit to data from the 2003 NRI and estimates are compared. Future work will use previous waves of data, more broad uses of land and outcomes, multiple states in the U.S., and covariates. It also will consider formally posterior predictive checks for model adequacy. Results from actual NRI data are not reported in this talk. Rather, artificial data indicating the nature of results are used for illustration.

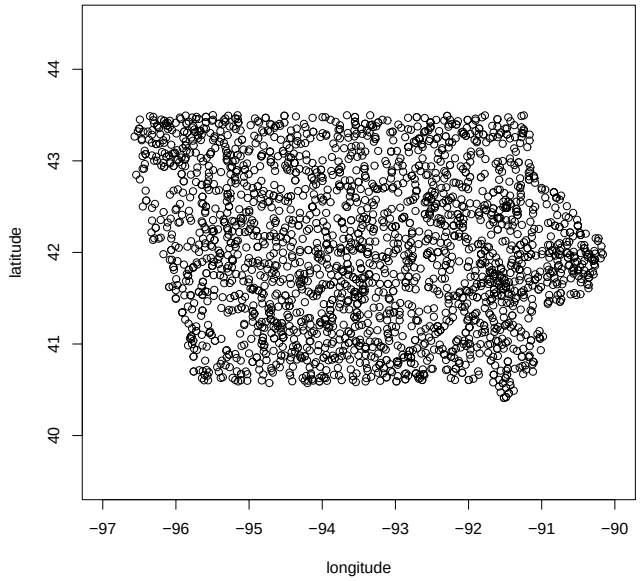
1 Introduction

The National Resource Inventory (NRI) survey (Nusser and Goebel 1997) has been conducted by Iowa State University and its survey laboratory, now called the Center for Survey Statistics and Methodology (CSSM) in collaboration with the National Resource Conservation Service (NRCS) of the U.S. Department of Agriculture (USDA) since 1956. The NRI, mandated by the U.S. Congress, is a national longitudinal survey of natural resources on nonfederal lands. From 1982 to 1997, the survey collected data on 300,000 primary sampling units (PSUs) every five years. In 1997, selected areas accounted for two to six percent of the land area by stratum. The level of coverage by stratum depends on the land use, soil type, irrigation practice, and current survey objectives and priorities. In 2000, the NRI became an annual survey program involving data collection on 42,000 PSUs in a core panel and 30,000 PSUs in a rotation or supplement every year. Nusser, Breidt, and Fuller (1998) provide design and estimation details.

The survey is used to measure the status and change in land cover and use, soil erosion due to water and wind, urban expansion and loss of prime farm land, and wetland composition. NRI data are used to model soil quality and other phenomena (see, e.g., Brejda *et al.* (2001, 2000a, 2000b). NRI data also are used in simulations, such as those of Mummey, Smith, and Bluhm (2000) and Kurkalova, Kling, and Zhao (2004), that evaluate polices and practices and estimate regional implications of environmental models.

Data are gathered by specialists who examine high-resolution digital ortho-rectified photography from low altitude airplane overflights. Administrative records are used to determine ownership (e.g., federal versus nonfederal) and Conservation Reserve Program (CRP) participation. Some field observations and special studies also are conducted. See http://cssm.iastate.edu/natlresinv for an overview and, for publicly available details on data, maps, reports, and tables, see http://www.nrcs.usda.gov/technical/NRI. The data are longitudinal with primary and secondary sampling units

Figure 1: Simulated ten percent of sample points in Iowa in 2003. The points are perturbed through added noise on both latitude and longitude. The locations do not represent the locations of actual NRI sample points.



tracked over time. Data are collected on units in the core panel every year. Data on rotation and supplemental units are collected every few years.

The 1997 design, which serves as a foundation behind the current annual survey program, is a two-stage area sample. A primary sampling unit (PSU) is an area segment comprising on average 160 acres. PSUs range in size from 40 to 640 acres. A stratum is defined as a collection of geographically grouped PSUs. Two PSUs per stratum are selected. Strata indicators were not available for this analysis. Instead, in this paper, the strata are taken to be counties. At the second stage of selection, points within the segments are selected. The random selection is restricted to encourage a geographic distribution of points within segments. Usually three points per segment are selected. Thus points (SSUs) are clustered within segments (PSUs) and segments are grouped into strata (or counties). A simulated ten percent sample (with noise added to both latitude and longitude) of 2003 NRI sample points in Iowa is represented in figure 1. The locations do not represent the locations of actual NRI sample points.

The goal of this paper is to examine the feasibility of modeling NRI data using hierarchical Bayesian models. Section 2 describes soil wind erosion data in Iowa from the 2003 NRI. Section 3 presents four models for these data. Algorithms for estimating the posterior distribution of model parameters are given in the appendix. Section 4 presents some results. Discussion of these initial efforts and plans for future work are given in section 5. This is an initial effort to specify and fit models to these data. Future work will consider additional models, larger subsets of data, more variables, and data from additional time periods.

As mentioned in the disclaimer above, the numbers reported in this article are not official results of the National Resource Inventory (NRI). They should not be considered official or approximate results in any sense for any characteristic monitored by the NRI. The author solely is responsible for the content of this article.

2 Wind Erosion Data in Iowa

The data considered in this paper concern points that fall on cultivated cropland in Iowa in the 2003 NRI. Of the several thousand points in Iowa, about one-quarter are on cultivated cropland and have soil wind erosion measurements in 2003. Of these, over sixty percent of the points have a wind erosion value of zero. A zero value could occur if the land is completely cultivated, is sheltered from the wind, or has a combination of other factors that eliminate wind as a source of erosion. Water is a treated as a different source of erosion. The remaining points have positive wind erosion values. As mentioned before, these figures are not official and are not the actual values.

The amount of soil lost due to wind erosion is measured in tons per acre per year. It is determined at a point by several factors including the point's knoll erodibility (slope), its soil ridge roughness, the unsheltered distance across the point, the amount of vegetative cover on the point, and the crop rotation schedule. These factors are recorded in the NRI data but are not used in the current analysis. If some of the contributing factors are zero, then the wind erosion value is zero. Thus, many zero values are structural in nature.

The positive wind erosion values are very skewed to the right. That is, the histogram of the positive values has a long tail to the right. About a quarter of the way between the minimum and maximum values, there is a sight bulge in the histogram. Otherwise the height of the bars on the histogram decreases steadily as values increase. A histogram of the log transformed values does not appear skew. Instead, on the log scale it appears that there is a mixture structure consisting of two underlying types of points. The clump on the left is much narrower and a little taller than the clump on the right. There is a deep trough, but not a gap, between the humps. The underlying reasons for this mixture struture are not currently known, but will be discussed in future work with NRCS subject matter experts.

3 Models

Four models for soil wind erosion data are presented. The model specifications include definitions of prior distributions and hyper-prior values. Model 1 is a single-level model for the points (the secondary sampling units) within the state of Iowa. A log normal mixture model is applied to the positive soil wind erosion values. The zero values are assumed to be truly zero. Models 2 and 3 are two-level models for the points within counties (strata) and segments (the primary sampling units), respectively. A hierarchical model is applied to the positive values in both cases. The impact of unequal variances within groups of points and of the log transformation are examined. Model 4 is a three-level model for the points (SSUs) within segments (PSUs) within counties (strata). Details of the Gibbs sampling algorithms used to simulate posterior distributions of model parameters are presented in the appendix.

3.1 Model 1: Points within states

The x_i's that are equal to zero are taken to be identically zero. Many such x_i's are zero by the fact that one or more of the factors contributing to the determination of the soil wind erosion factor are zero. This assumption is made throughout this paper. Future work will involve discussing this assumption with NRCS subject matter experts and postulating hypotheses about measurement error for both the positive and the zero values.

For positive soil wind erosion values ($x_i > 0$), let $y_i = \log x_i$, $i = 1, \ldots, n$. For $y_i, i = 1, \ldots, n$, the density in model 1 is a mixture of two normal densities: $f(y_i) = p\phi(y_i|\mu_1, \sigma_1^2) + (1 - p)\phi(y_i|\mu_2, \sigma_2^2)$. The variances σ_1^2 and σ_2^2 are not assumed to be equal, because the two humps in the histogram of the log values represent clearly of different variances. That is, it appears that the two standard deviations are different by more than a factor of two. The y_i's are independent of one another in this model.

The prior distribution for p is a Beta(α_p, β_p) distribution. The values of α_p and β_p are set equal to 0.5 for work reported in section 4. The prior distributions for (μ_1, σ_1^2) and (μ_2, σ_2^2) are Normal-Inverse Gamma distributions. That is, $\mu_g|\mu_{0g}, \sigma_g^2, k_g \sim$ Normal$(\mu_{0g}, \sigma_g^2/\kappa_g)$ and $\sigma_g^2 \sim$ InverseGamma(α_g, β_g), for $g = 1, 2$. The values of α_g and β_g for $g \in \{1, 2\}$ are set at 3.8. The prior means μ_{01} and μ_{02} are set equal to a value near the center of the y_i-values. The prior scale factors κ_1 and κ_2 are set equal to 0.5. The prior distributions on p and the two mean-variance pairs are independent of one another. Discussion of the choice of the Inverse Gamma prior distribution for the variance components is presented in sections 4 and 5.

The average soil loss due to wind erosion in tons per acre is $\sum_{i=1}^{N} x_i/N$, where N is the number of points in the entire state. Some of the x_i values are observed, but most are unobserved by design. The state average can be represented as $(\sum_{i\in\text{Obs}} x_i + \sum_{i\in\text{Mis}} x_i)/N$, where Obs is the set of indices of observed values of soil erosion due to wind and Mis is the set for unobserved values. Under a model with no stratification or clustering, the expected value of x_i when i is an element of Mis is the mean of the posterior distrubution of the mean of X. Since the mean of the log normal(μ, σ^2) distribution is $e^{\mu+\sigma^2/2}$, the expected value of an unobserved X given model parameters is $pe^{\mu_1+\sigma_1^2/2} + (1 - p)e^{\mu_2+\sigma_2^2/2}$.

The number of indices in Mis is much larger than Obs in this application. Of the thousands of points in Iowa in 2003, over sixty percent of them had values of zero for soil erosion due to wind. These points represented about two-thirds of the area in the state. Less than forty percent of the sample points had positive wind erosion values. These points represented about one third of the points in the state. Thus, the estimate of the mean wind erosion in the state is practically equal to the posterior mean of the positive values times the fraction of the state represented by non-zero points.

The posterior distribution of the parameters in model 1 is iteratively simulated as described in the appendix. The iterative simulation algorithm typicallly is run several hundred iterations independently from at least three starting values. In order estimate the mean wind erosion value in the state, one can proceed by estimating the finite population value at each simulation iteration. Given the data and one set of simulated parameters, say the parameter values at iteration t, the predictive distribution on the log scale for the unobserved data is a mixture of normal distributions. The probability that an unobserved observation has mean $\mu_1^{(t)}$ and variance $\sigma_1^{2(t)}$ is $p^{(t)}$. With probability $1 - p^{(t)}$, it has the other mean and variance at iteration t. In order to simulate the state average, one would have to generate tens of thousands of deviates for iteration t (one for each unobserved point) and transform them to the original scale. This would be computer intensive despite the fact that it would not be difficult. An alternative is to simply multiply posterior mean on the original scale by an appropirate weight factor. The weight factor in this case is the proportion of non-zero observations in the state. The law of large numbers justifies in this case simply multiplying the posterior mean on the original scale by the weight factor.

3.2 Model 2: Points within counties

Model 2 divides the points into strata defined by counties. As before, many of the x_{ij} values are zero, where i ($i = 1, \ldots, n_j$) indexes points and j ($j = 1, \ldots, J$) indicates counties. The non-zero values are transformed to the log scale: $y_{ij} = \log x_{ij}$. Within counties, the distribution of log positive wind erosion values do not all have a clear mixture structure. Actually choosing distributions based on looking at the data in each county is not feasible on a large scale. Figure 2 illustrates the variety of empirical distributions by county with modified data. Note that the means and variances for the log erosion values have been modified and can-

not be converted to actual values.

In the current analysis, it is assumed that y_{ij} is independently and normally distributed with county-level mean μ_j and variance σ^2. Future work will examine distributions within counties (strata) and segments (PSUs) in more detail. The distribution of μ_j for $j = 1, \ldots, J$ is normal with mean μ_0 and variance σ^2/κ_0. The prior distribution on σ^2 is an Inverse Gamma(α, β) distribution. The value of α is 3 and β is 4. Thus, the prior mean and variance of σ^2 are 2 and 4, respectively. Estimates computed with other hyperparameter values did not differ significantly from those produced with these choices of α and β.

The prior distribution on κ_0 is a Gamma(α_0, β_0) distribution. The values chosen for α_0 and β_0 are both 2, sot that the prior mean and variance of κ_0 are 1 and 1/2, respectively. The density function for μ_0 is *a priori* flat. That is, it is not an actual probability density but is expressed as $p(\mu_0) \propto 1$.

The estimator of the state-wide mean is analogous to the estimator used with model 1 except that the population is divided by counties and a mixture model is not used. The overall state average can be expressed as $\sum_{j=1}^{J}(\sum_{i\in \text{Obs}_j} x_{ij} + \sum_{i\in \text{Mis}_j} x_{ij})/N$, where Obs_j are the indices for the sample in county j and Mis_j are the indices for the unsampled points. The posterior mean in county j for the unobserved data is $e^{\mu_j+\sigma^2/2}$. Within counties, the number of sample points is reasonably large, so that using a weighted combination of means does not differ appreciably from simulating the unobserved data within counties.

An extension of model 2 is to allow unequal variances for measurements within counties. That is, $y_{ij} \sim N(\mu_j, \sigma_j^2)$, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$. The posterior mean in county j for the unobserved data for this case is $e^{\mu_j+\sigma_j^2/2}$.

The Gibbs sampling for model 2 is given in the appendix.

3.3 Model 3: Points within PSUs

Model 3 is the same as model 2 expect the points are grouped by PSUs rather than by counties. As a result, there are far fewer points per PSU than per county. There are 99 counties in Iowa, but many times this number of PSUs. About two thirds of the counties and one third of the PSUs contain points with positive soil wind erosion values. There were between two and three points on average per PSU in the state. Model 3 was fit with only the equal variance model due to the large number of PSUs.

Model 3 was also fit to the untransformed positive data (x_{ij}'s) for purposes of comparison to results using the log scale values (y_{ij}'s). The hyperparameter values of α and β were adjusted to reflect a more appropriate range of values for the data on the original scale. Discussion of the impact of the log normal distribution and the transformation are discussed in sections 4 and 5.

The Gibbs sampling algorithm for model 3 is the same as for model 2.

3.4 Model 4: Points within PSUs within counties

Model 4 is a hierarchical model with a distribution of points within PSUs, a distribution of PSUs within counties, and a prior distribution on county-level parameters. Due to the small sample size and concern about the impact of the log normal model, the positive soil wind erosion values are not transformed in this analysis. Future work will examine modeling assumptions and model adequacy.

It is assumed that x_{ijk} ($i = 1, \ldots, n_{jk}$, $j = 1, \ldots, J_k$, $k = 1, \ldots, K$) is independently and normally distributed with the mean for the PSU j within county k μ_{jk} and variance σ^2. One could consider allowing different variances within counties or within PSUs, but that is not undertaken here. The distribution of μ_{jk} ($j = 1, \ldots, J_k$) is normal with mean μ_k and variance σ^2/κ_1. The prior distribution on σ^2 is an Inverse Gamma(α, β) distribution. The value of α is 3 and β is 3. Thus, the prior mean and variance of σ^2 are 1.5 and 2.25, respectively. Given the number of parameters involved in this model, the prior distribution on σ^2 was made a little more concentrated. Future research will study sensitivity to choices of hyperparameter values. The prior distribution on κ_1 is a Gamma(α_1, β_1) distribution. The value chosen for α_1 is 6. For β_1 it is 4. The corresponding prior mean and variance of κ_1 are 1.5 and 0.375, respectively.

The prior distribution for μ_k ($k = 1, \ldots, K$) is normal with mean μ_0 and variance σ^2/κ_0. A flat improper prior distribution is placed on μ_0. The prior distribution on κ_0 is Gamma with $\alpha_0 = 8$ and $\beta_0 = 6$ (mean=1.33, variance=0.22).

The estimator of the state-wide mean is analogous to the estimators used with models 2 and 3. The overall state average can be expressed as $\sum_{j,k}(\sum_{i\in \text{Obs}_{jk}} x_{ijk} + \sum_{i\in \text{Mis}_{jk}} x_{ijk})/N$, where Obs_{jk} are the indices for the sample in PSU j in county k and Mis_{jk} are the indices for the unsampled points. The posterior mean in PSU k in county j for the unobserved data, when a log transformation is not used in the modeling, is simply μ_{jk}.

The Gibbs sampling algorithm for model 4 can be found in the appendix.

4 Results

Results are presented for the four models of section 3. Comparison results are computed using Stata survey software (StataCorp 2003a, 2003b). In model 1, the points are treated like observations on units in the state. In model 2, the points are observations within strata defined by counties. In model 3, the points are grouped into clusters defined by PSUs. Model 4 uses the counties as strata and the PSUs as clusters. Survey sample weights for each point are calculated in the NRI survey to reflect state and other control totals.

Table 1: Estimates, standard errors, and intervals for the average soil loss due to wind erosion per acre on Iowa cultivated cropland in 2003: estimates based on model 1 and on survey design ignoring clustering and stratification. *Artificial results; not official NRI results.*

Source	Estimate	SE	95% interval	Interval type
Stata	17.46	2.19	(13.08, 21.84)	Confidence
Model 1	17.92	2.00	(13.96, 22.01)	Posterior

4.1 Results for Model 1: points in the state

The Gibbs sampling algorithm for model 1 from the appendix was run from three independent starting values for one-thousand ten iterations each. After examining plots of parameter values versus iteration number, the first ten iterations were discarded as an initialization period. Means of counties and the overall mean (μ_0) were examined on their original scale. Variances and scale values ($\kappa's$) were examined on a log scale. Based on plots of the series, it is apparent that the Gibbs sampling algorithms quickly converges to sampling from the posterior distribution of the parameters. Estimates after a few hundered and after a few thousand iterations are almost the same, which suggests that the algorithms have converged. See the appendix for additional discussion of convergence diagnosis.

Table 1 presents (modified) results for model 1 and the corresponding survey design estimate using survey weights. The statements of results throughout the paper present artificial results that do not correspond to official NRI results. The statements do nonetheless illustrate phenomena encountered when fitting models to real NRI data. Model 1 and design-based estimates estimates of the amount of soil loss per acre due to wind erosion are about the same. The standard deviation of estimates from the simulation (recorded in the SE column) is a little smaller than the standard error produced by Stata. No substantial difference was noted when modeling at the state level.

4.2 Model 2 results

The Gibbs sampling algorithm for model 2 from the appendix was run from three independent starting values for nine-thousand iterations each. Convergence was monitored as it was for model 1 and as described in the appendix. Plots of parameter values, including county-level means, and estimates versus the last four-thousand iterations shows only random scatter about a constant line. This number of iterations was much more than needed for convergence, because, as for model 1, the Gibbs sampling algorithm quickly converged.

Table 2 presents results for model 2 and the corresponding survey design estimate with stratification using survey weights. These are not actual results, but illustrate what was observed with the real NRI data. The estimates using model 2 are higher than the survey design estimate. One reason this could be happening is that the log normal distribution is highly skewed to the right and the transformed means are especially sensitive to large values of $\mu_j + \sigma^2/2$. Gelman *et al.* (2004, chapter 9) also noted the sensitivity of inferences to the log normal model assumption. As noted by these authors, the mean is affected greatly by the extreme tail of the log normal distribution. Although a natural choice for these highly skewed data, the impact of the distributional assumption has a deleterious impact on estimation of the mean. Future work will have to examine modeling alternatives to address this problem. Karlberg (2000) proposed an alternative for situations such as the one encountered here and will be studied in future work. The phenomenon is explored further for model 3 in the next section.

Table 2: Estimates, standard errors, and intervals for the average soil loss due to wind erosion per acre on Iowa cultivated cropland in 2003: estimates based on model 2 and on survey design using counties as strata. *Artificial results; not official NRI results.*

Source	Estimate	SE	95% interval	Interval type
Stata	17.46	2.00	(13.46, 21.46)	Confidence
Model 2 with equal variance	21.26	1.80	(17.62, 24.60)	Posterior
Model 2 with unequal variance	21.80	1.90	(17.92, 25.33)	Posterior

The estimate and the standard error under model 2 are higher with unequal variance than with equal variances. Since the unequal variance model introduces several additional parameters (about two-thirds of the counties had positive wind erosion values), the standard error result is expected. The higher mean would be consistent with comments above concerning the log normal modeling assumption. The hyperprior values for the prior distribution on variances made more of an impact on the unequal variance model. This is a believable result, because only the observations in a county contribute to estimation of the county-level variance. The standard error of the survey estimator is smaller than the corresponding value in table 1, which is consitent with the usual performance of a stratified estimator.

4.3 Model 3 results

The Gibbs sampling algorithm for model 3 from the appendix was run from three independent starting values for nine-thousand iterations each. Series plots of parameter values and esimates versus iteration look very similar to those from previous models: very shortly after initialization there is no apparent pattern other than random scatter about a horizontal line. Nine-thousand iterations was many more iterations than necessary.

Table 3 presents artificial results for model 3 and the corresponding survey design estimate with clustering by segments (PSUs) using survey weights. The model is applied to the data on both the original and logarithmic scales. The estimates using model 3 and the logarithmic data are higher than the survey design estimate and higher than the estimates for model 2. It appears that the impact of the log normal distributional assumption is greater in model 3, which has many more clusters (segments, or PSUs) than model 2 had strata (counties). Few values in a segment directly influence the mean estimate within a segment.

The estimate for model 3 using data on the original scale is close to the estimate using the traditional survey estimator. The standard error under model 3 (the standard deviation of the simulated estimates), however, is lower than using the survey variance estimator. This lower standard deviation results because the missing data (the points not included in the sample) and their means were not simulated from an appropriate model. Instead, the observed points were simply weighted to relfect the population total. The imputation of missing values at the PSU means are reflected in the weighting scheme. This practice, as is apparent in the estimate, greatly understates variability within the cluster sample.

Two alternative strategies were studied, but details are not reported here. The first generated values for each missing point. If an observed point was zero, then a number of zero points equal to one less than the size of the weight for that point were added. If an observed point was positive, then a number of new points equal to one less than the weight for that point were drawn from a normal distribution with mean μ_j and variance σ^2. This method, as would be expected, basically reproduced the results of table 3 using data on the original scale.

The second alternative method, on an individual simulation run, generated new means for new PSUs to contain the additional positive wind erosion values. These means were drawn from a normal distribution with mean μ_0 and variance σ^2/κ_0. That is, the value of the new PSU means were effectively imputed for each simulation run. New observations were then drawn from distributions with these means and variance σ^2. This procedure, simulating means for unobserved PSUs and values for unobserved points, generates more variability. Since accurate descriptions of the population of segments in Iowa were not available to the author, two points per new segment were generated; this ratio approximates the average number of positive points per segment containing at least one positive point in the observed data.

Some additional work was done, but is not reported here due to restrictions on releasing information on unpublished NRI results for 2003. The inability of these simple simulation procedures to replicate the observed data suggest that the model, and in particular the equal variance and normality assumptions, are inadequate. The author currently does not have adequate data on the population in order to conduct further work on this simulation. A simulation considering characteristics from previous waves of data collection on unobserved segments and points will be considered in future work.

Table 3: Estimates, standard errors, and intervals for the average soil loss due to wind erosion per acre on Iowa cultivated cropland in 2003: estimates based on model 3 and on survey design using segments as clusters (PSUs). *Artificial results; not official NRI results.*

Source	Estimate	SE	95% interval	Interval type
Stata	17.46	2.45	(12.56, 22.36)	Confidence
Model 3 with data on logarithmic scale	21.90	1.82	(18.26, 25.53)	Posterior
Model 3 with data on original scale	17.16	1.44	(14.25, 20.00)	Posterior

4.4 Model 4 results

The Gibbs sampling algorithm for model 4 from the appendix was run from three independent starting values for nine-thousand iterations each. Series plots of parameter values and estimates versus iteration look very similar to those of the other models. It is clear thta the algorithms converge quickly to sampling from the desired posterior distributions. Table 4 presents (modified) results for model 4 and the corresponding survey design estimate with stratification by counties and clustering by segments using survey weights. As usual in this paper, these are not official NRI results. The model is applied to the data on the original scale only. The estimate using model 4 is similar to the survey design-based estimate, but as with model 3 the variance is less. Again, the weighting mean adjustment, equivalent to imputing the posterior means within PSUs, leads to an understatement of uncertainty. Future work will simulate missing values to incorporate the proper amount of uncertainty into estimates.

The second line of results in table 4 presents the replication variance estimate (Fuller 1998, Kim and Sitter 2003). The replicaton variance estimator uses replicate survey weights (multiple sets of replicate weights in the case of the NRI) to reflect the complex survey design. As is apparent, the replicate variance estimation method gives an even higher estimate of standard error and wider confidence interval.

5 Discussion and Future Work

The four models of section 3 were fit using the algorithms of the appendix to soil wind erosion data from the 2003 NRI sample of cultivated crop land in Iowa. Modified results are reported that illustrate patterns of results under the various models and survey estimators, but no official NRI results are presented in this paper. It is relatively easy to fit the models, but two issues arise when doing so. First, the log normal distributional assumption tends

Table 4: Estimates, standard errors, and intervals for the average soil loss due to wind erosion per acre on Iowa cultivated cropland in 2003: estimates based on model 4 and on survey design using segments as clusters (PSUs) and counties as strata. *Artificial results; not official NRI results.*

Source	Estimate	SE	95% interval	Interval type
Stata	17.46	2.46	(12.54, 22.38)	Confidence
Replication variance		2.63	(12.20, 22.72)	Confidence
Model 4 with data on original scale	17.51	1.32	(14.87, 20.15)	Posterior

to inflate estimates, especially when small areas being modeled had small sample sizes. Alternative modeling strategies should be investigated. These strategies should be based in subject matter understanding. One option is to truncate distributions beyond the upper limit of the soil wind erosion distribution; NRCS subject matter experts will be consulted about this option. A second possibility is to explore different specifications of the prior distributions on the variance components, as in Gelman (2004). A third route is to incorporate methods of Karlberg (2000) and of Maiti and Slud (2002), Slud and Maiti (2004), Maiti (2004), and Ghosh and Maiti (2004).

The second issue that arises when fitting the models of this paper is that conditioning on weights produces understatement of varibility. One strategy for future consideration is simulation of unobserved means and values. In the case of most large-scale sample surveys, this could be quite a task, especially with small portions of the population included in the sample and numerous iterations of the simulated output from sampling the posterior distribution of parameters. Although the goal is not to mimic exactly the survey design estimates, it is necessary to convince consumers of the results that the hierarchical modeling produces reasonable results, especially in cases for which the consumer is likely to have some insight. Finally, future work needs to consider assessment of model adequacy as well as the issues discussed below.

A number of challenges need to be addressed if multilevel modeling is going to be used with NRI data. It will be necessary to include all states in the United States and all twelve broad uses of land (groups of land cover and land use categories) in analyses. Many outcomes, such as amount of soil erosion due to water and wind, amount of prime farmland, condition of wetland areas, and extent of urban land will need to be modeled simultaneously. The challenge of building and evaluating models for small areas, as illustrated with the soil wind erosion data from Iowa in 2003, will be a significant challenge. Data over mulitple years are available, and should be analyzed together. Indeed, longitudinal analyses and the study of trends in natural resource availability and use are key uses of NRI survey data. New analyses will have to conform to past work and maintain historical consistency.

There are many reasons to believe that multilevel models and modeling of NRI data will be beneficial. Efforts to model data can help one learn about patterns in the data, which could be useful in forming hypotheses. Data analysis related to model 1 revealed a mixture structure in soil wind erosion in Iowa. Further work could attempt to relate the mixture struture to the factors described in section 2 related to soil wind erosion. Many of the biological and ecological relationships between factors and erosion will be similar over time. Efforts to understand them will be useful in future analyses and in other modeling applications. There are a lot of data, many variables on numerous segments and points, so there is an opportunity to fit complex models. The models could be used for imputation of missing values. Currently, missing values are filled-in or imputed with donor values. The model-based imputations could serve as a check on the reasonableness of the donor-based methods. Even though there are a lot of data nationwide, there are small sample sizes of land with certain broad uses in PSUs and in some counties. Hierarchical models tend to shrink estimates toward common values. This shrinkage effect could reduce mean squared errors of estimates at local levels and in small areas.

Future work will involve expanding efforts to model NRI survey data. This will include using several states for modeling of multiple outcome variables over a few years, using point-level covariate information to predict outcomes, modeling the spatial relationships among sampling units, and relating segment- and county-level effects to outcomes. Efforts will be made to model variables for the purposes of imputing for missing observations and for intentionally and unintentionally unobserved units. Subject matter knowledge as expressed by the NRCS technical staff will be critial to most of these efforts. Future work also will develop checks (Rubin 1984; Gelman, Meng, and Stern 1996) for model adequacy.

6 References

Brejda, J.J., Karlen, D.L., Smith, J.L., and Allan, D.L. (2000b). "Identification of regional soil quality factors and indicators: II. Northern Mississippi Loess Hills and Palouse Prairie." *Soil Science Society of America Journal*, 64(6): 2125-2135.

Brejda, J.J., Mausbach, M.J., Goebel, J.J., Allan, D.L., Dao, T.H., Karlen, D.L., Moorman, T.B., and Smith, J.L. (2001). "Estimating surface soil organic carbon content at a regional scale using the National Resource Inventory." *Soil Science Society Of America Journal*. 65(3): 842-849.

Brejda, J.J., Moorman, T.B., Karlen, D.L., and Dao, T.H. (2000a). "Identification of regional soil quality factors and indicators: I. Central and southern high plains." *Soil Science Society of America Journal*, 64(6): 2115-2124.

Brooks, S.P. , and Gelman, A. (1998). "General methods for monitoring convergence of iterative simulations", *Journal of Computational and Graphical Statistics*, 7 , 434-455

Fuller, W.A. (1998). "Replication variance estimation for two-phase samples," *Statistica Sinica*, 8(4): 1153-1164.

Gelman, A. (2004). "Prior distributions for variance parameters in hierarchical models." Unpublished. Available on October 22, 2004, at `http://www.stat.columbia.edu/ ∼ gelman/research/unpublished/tau7.pdf`.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd edition. Chapman & Hall: London

Gelman, A., Meng, X.L., Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies," *Statistica Sinica*, 6(4): 733-760.

Gelman, A., and Rubin, D.B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science* 7, 457-472.

Ghosh, M., and Maiti, T. (2004). "Small-area estimation based on natural exponential family quadratic variance function models and survey weights." *Biometrika*, 91(1): 95-112.

Karlberg, F. (2000). "Survey estimation for highly skewed populations in the presence of zeroes." *Journal of Official Statistics*, 16(3): 229-241.

Kim, J.K., and Sitter, R.R. (2003). "Efficient replication variance estimation for two-phase sampling," *Statistica Sinica*, 13(3): 641-653.

Kurkalova, L., Kling, C.L., and Zhao, J.H. (2004). "Multiple benefits of carbon-friendly agricultural practices: Empirical assessment of conservation tillage." *Environmental Management*, 33(4): 519-527.

Maiti, T. (2004). "Applying jackknife method of mean squared prediction error estimation in SAIPE." *Statistics in Transition*, 6(5): 685-695.

Maiti, T. and Slud, E.V. (2002). "Comparison of Small Area Models in SAIPE." Technical Report, U.S. Bureau of the Census.

Mummey, D.L., Smith, J.L., and Bluhm, G. (2000). "Estimation of nitrous oxide emissions from US grasslands." *Environmental Management*, 25(2): 169-175.

Nusser, S.M., Breidt, E.J., and Fuller, W.A. (1998). "Design and estimation for investigating the dynamics of natural resources." *Ecological Applications*, 8(2): 234-245.

Nusser, S.M., and Goebel, J.J. (1997). "The National Resources Inventory: A long-term multi-resource monitoring programme." *Environmental and Ecological Statistics*, 4(3): 181-204.

Rubin, D.B. (1984). "Bayesianly justifiable and relevant frequency calculations for the applied statistician." *Annals of Statistics*, 12(4): 1151-1172.

Slud, E. V., and Maiti, T. (2004). "MSE estimation in transformed Fay-Herriot models, with applications to SAIPE." *Technical report*.

StataCorp. (2003a). *Stata Statistical Software: Release 8.0*. College Station, TX: Stata Corporation.

StataCorp. (2003b). *Survey Data Reference Manual*. College Station, TX: Stata Press.

Appendix: Gibbs sampling details

Models were described in section 3. The sections below describe the Gibbs sampling algorithms used to sample from the posterior distributions. Experiments were conducted with multiple independent series in models 1 and 2 to assess convergence of the algorithms (Gelmand and Rubin 1992, Brooks and Gelman 1998). Simulations did not show multimodal behavior and converged after a few hundered iterations. Estimates for the four models were computed based on a few thousand iterations after a "burn-in" period and again after a few thousand more iterations. Estimates under all the models did not change appreciably when additional iterations were added.

Model 1: Points within states

Gibbs sampling was used to simulate mixture models parameters for the positive wind erosion values on the log scale. In order to sample from the mixture distribution, latent variables $z_i, i = 1, \ldots, n$ were created. The value of z_i is one if point i is from the first mixture class and 0 if it belongs to the second mixture class.

The Gibbs sampling algorithm proceeds as follows.

1. Choose initial values of unknown parameters: $p = p^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \sigma_1^{2(0)}, \sigma_2^{2(0)}$.

2. Repeat the following steps numerous times T until the series of values converges to the joint posterior distribution of pa-

rameters. At iteration $t+1, t = 0, \ldots, T-1$, the steps are given below.

(a) For $i = 1, \ldots, n$, draw a value of z_i from a Bernoulli distribution with probability p_i, where $p_i = \frac{p^{(t)}\phi(y_i|\mu_1^{(t)},\sigma_1^{2(t)})}{p^{(t)}\phi(y_i|\mu_1^{(t)},\sigma_1^{2(t)})+(1-p^{(t)})\phi(y_i|\mu_2^{(t)},\sigma_1^{2(t)})}$, where $\phi(y|\mu,\sigma^2)$ is the normal density function for a normal distribution with mean μ and variance σ^2 evaluated at y. Let $z_i^{(t+1)}$ be the drawn value of z_i. Compute $n_1^{(t+1)} = \sum_{i=1}^n z_i^{(t+1)}$ and $n_2^{(t+1)} = n - n_1^{(t+1)} = \sum_{i=1}^n (1 - z_i^{(t+1)})$.

(b) Draw $p^{(t+1)}$ from a Beta($\alpha_p + n_1^{(t+1)}$, $\beta_p + n_2^{(t+1)}$) distribution.

(c) Draw $\sigma_1^{2(t+1)}$ from an Inverse Gamma distribution with parameters $\alpha_1 + n_1^{(t+1)}/2$ and $\beta_1 + \sum_{i=1}^n z_i^{(t+1)}(y_i - \overline{y_1})^2/2 + \frac{1}{2}(\mu_{01} - \overline{y_1})^2 \frac{\kappa_1 n_1^{(t+1)}}{\kappa_1 + n_1^{(t+1)}}$, where $\overline{y_1} = \sum_{i=1}^n z_i y_i / \sum_{i=1}^n z_i$.

Draw $\sigma_2^{2(t+1)}$ from an Inverse Gamma distribution with parameters $\alpha_2 + n_2^{(t+1)}/2$ and $\beta_2 + \sum_{i=1}^n z_i^{(t+1)}(y_i - \overline{y_2})^2/2 + \frac{1}{2}(\mu_{02} - \overline{y_2})^2 \frac{\kappa_2 n_2^{(t+1)}}{\kappa_2 + n_2^{(t+1)}}$, where $\overline{y_2} = \sum_{i=1}^n (1 - z_i) y_i / \sum_{i=1}^n (1 - z_i)$.

(d) Draw $\mu_1^{(t+1)}$ from a normal distribution with mean $(\sum_{i=1}^n z_i^{(t+1)}\overline{y_1} + \kappa_1\mu_{01})/(n_1^{(t+1)} + \kappa_1)$ = $(n_1^{(t_1)}\overline{y_1} + \kappa_1\mu_{01})/(n_1^{(t+1)} + \kappa_1)$ and variance $\sigma_1^{2(t+1)}/(n_1^{(t+1)} + \kappa_1)$.

Draw $\mu_2^{(t+1)}$ from a normal distribution with mean $(\sum_{i=1}^n (1 - z_i^{(t+1)})\overline{y_2} + \kappa_2\mu_{02})/(n_2^{(t+1)} + \kappa_2)$ = $(n_2^{(t+1)}\overline{y_2} + \kappa_2\mu_{02})/(n_2^{(t+1)} + \kappa_2)$ and variance $\sigma_2^{2(t+1)}/(n_2^{(t+1)} + \kappa_2)$.

The mixture classes were well-separated as noted in section 2. The mean μ_1 on the left was initialized below zero, whereas the mean μ_2 on the right was above zero. In four thousand iterations, the value of μ_1 was always less than the value of μ_2. Technically, the labels on the classes (μ_1 corresponds to the small mean, μ_2 to the large mean) should switch if the series of simulated values is to travel throughout the parameter space. It was judged in this case that the likelihood of a switch is extremely low given the separation of the two modes and the large sample sizes around each mode.

Model 2: Points within counties

The steps for estimation of the model 2 with equal variances are given below. In the ideal case, all unknown parameters would be drawn at once conditional only on the observed data and independently of previous draws of paramters. At the other extreme, one parameter would be drawn at a time conditional on all the other current parameter values and the data. In the version of the algorithm given here, sets of parameters are drawn together. First, σ^2 and the county-level means $\{\mu_j, j = 1, \ldots, J\}$ are drawn together. Second, κ_0 and μ_0 are drawn together. This formulation of the algorithm should produce draws that are less dependent across iterations than an algorithm drawing one parameter value at a time.

1. Choose initial values for unknown parameters: $\mu_j^{(0)}, j = 1, \ldots, J$, $\sigma^{2(0)}$, $\mu_0^{(0)}$, and $\kappa_0^{(0)}$.

2. Repeat the following steps numerous times T until the series of values converges to the joint posterior distribution of parameters. At iteration $t+1, t = 0, \ldots, T-1$, the steps are given below.

   (a) Draw $\sigma^{2(t+1)}$ from an Inverse Gamma distribution with parameters $\alpha + n/2$ and $\beta + \sum_{j=1}^J \sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2/2 + \sum_{j=1}^J \frac{n_j\kappa_0^{(t)}}{n_j + \kappa_0^{(t)}}(\mu_0^{(t)} - \overline{y}_j)^2/2$.

   (b) For $j = 1, \ldots, J$, independently draw $\mu_j^{(t+1)}$ from a normal distribution with mean $(n_j\overline{y}_j + \mu_0^{(t)}\kappa_0^{(t)})/(n_j + \kappa_0^{(t)})$ and variance $\sigma^{2(t+1)}/(n_j + \kappa_0^{(t)})$.

   (c) Draw $\kappa_0^{(t+1)}$ from a Gamma($\alpha_0 + \frac{J-1}{2}$, $\beta_0 + \sum_{j=1}^J (\mu_j^{(t+1)} - \overline{\mu}^{(t+1)})^2/(2\sigma^{2(t+1)})$) distribution, where $\overline{\mu}^{(t+1)} = \sum_{j=1}^J \mu_j^{(t+1)}/J$

   (d) Draw $\mu_0^{(t+1)}$ as Normal($\overline{\mu}^{(t+1)}$, $\sigma^{2(t+1)}/(J\kappa_0^{(t+1)})$).

When variances within counties are unequal, step 2 is modified as follows.

2. (a) For $j = 1, \ldots, J$, independently draw $\sigma_j^{2(t+1)}$ from an Inverse Gamma distribution with parameters $\alpha + n_j/2$ and $\beta + \sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2/2 + \frac{n_j\kappa_0^{(t)}}{n_j + \kappa_0^{(t)}}(\mu_0^{(t)} - \overline{y}_j)^2/2$.

   (b) For $j = 1, \ldots, J$, independently draw $\mu_j^{(t+1)}$ from a normal distribution with mean $(n_j\overline{y}_j + \mu_0^{(t)}\kappa_0^{(t)})/(n_j + \kappa_0^{(t)})$ and variance $\sigma_j^{2(t+1)}/(n_j + \kappa_0^{(t)})$.

   (c) Draw $\kappa_0^{(t+1)}$ from a Gamma($\alpha_0 + \frac{J-1}{2}$, $\beta_0 + \sum_{j=1}^J (\mu_j^{(t+1)} - \overline{\mu}^{(t+1)})^2/(2\sigma_j^{2(t+1)})$) distribution, where $\overline{\mu}^{(t+1)} = \left(\sum_{j=1}^J \mu_j^{(t+1)}/\sigma_j^{2(t+1)}\right) / \left(\sum_{j=1}^J 1/\sigma_j^{2(t+1)}\right)$.

   (d) Draw $\mu_0^{(t+1)}$ as Normal($\overline{\mu}^{(t+1)}$, $(\sum_{j=1}^J \kappa_0^{(t+1)}/\sigma_j^{2(t+1)})^{-1}$).

Model 3: Points within PSUs

The Gibbs sampling algorithm for model 3 is the same as for model 2. The index j ($j = 1, \ldots, J$) in model 3 represents PSUs rather than counties.

Model 4: Points within PSUs within counties

As in the Gibbs sampling algorithm for models 2 and 3, some of the parameters are drawn in sets. The sets are $\{\sigma^2, \mu_{jk}, k = 1, \ldots, K, j = 1, \ldots, J_k\}$, $\{\kappa_1\}$, $\{\mu_k, k = 1, \ldots, K\}$, and $\{\kappa_0, \mu_0\}$.

1. Choose initial values for unknown parameters: $\mu_{jk}^{(0)}, k = 1, \ldots, K, j = 1, \ldots, J_k$, $\mu_k^{(0)}, k = 1, \ldots, K$, $\sigma^{2(0)}$, $\mu_0^{(0)}$, $\kappa_0^{(0)}$, and $\kappa_1^{(0)}$.
2. Repeat the following steps a large number of times T until the series of values converges to the joint posterior distribution of parameters. At iteration $t + 1, t = 0, \ldots, T - 1$, the steps are given below.
   (a) Draw $\sigma^{2(t+1)}$ from an Inverse Gamma distribution with parameters $\alpha + n/2 + K/2$ and $\beta + \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{i=1}^{n_j}(x_{ijk} - \overline{x}_{jk})^2/2 + \sum_{k=1}^{K}\sum_{j=1}^{J_k}\frac{n_{jk}\kappa_1^{(t)}}{n_{jk}+\kappa_1^{(t)}}(\mu_k^{(t)} - \overline{x}_{jk})^2/2 + \kappa_0^{(t)}\sum_{k=1}^{K}(\mu_k^{(t)} - \mu_0^{(t)})^2/2$.
   (b) For $k = 1, \ldots, K$, $j = 1, \ldots, J_k$, independently draw $\mu_{jk}^{(t+1)}$ from a normal distribution with mean $(n_{jk}\overline{x}_{jk} + \mu_k^{(t)}\kappa_1^{(t)})/(n_{jk} + \kappa_1^{(t)})$ and variance $\sigma^{2(t+1)}/(n_{jk} + \kappa_1^{(t)})$.
   (c) Draw $\kappa_1^{(t+1)}$ from a Gamma($\alpha_1 + \frac{n_{\text{psu}}}{2}$, $\beta_1 + \sum_{k=1}^{K}\sum_{j=1}^{J_k}(\mu_{jk}^{(t+1)} - \mu_k^{(t)})^2/(2\sigma^{2(t+1)})$) distribution, where $n_{\text{psu}} = \sum_{k=1}^{K} J_k$ is the number of PSUs.
   (d) For $k = 1, \ldots, K$, independently draw $\mu_k^{(t+1)}$ from a normal distribution with mean $(\kappa_1^{(t+1)}J_k\overline{\mu_k}^{(t+1)} + \kappa_0^{(t)}\mu_0^{(t)})/(\kappa_1^{(t+1)}J_k + \kappa_0^{(t)})$ and variance $\sigma^{2(t+1)}/(\kappa_1^{(t+1)}J_k + \kappa_0^{(t)})$, where $\overline{\mu_k}^{(t+1)} = \sum_{j=1}^{J_k}\mu_{jk}^{(t+1)}/J_k$.
   (e) Draw $\kappa_0^{(t+1)}$ from a Gamma($\alpha_0 + \frac{K-1}{2}$, $\beta_0 + \sum_{k=1}^{K}(\mu_k^{(t+1)} - \overline{\mu}^{(t+1)})^2/(2\sigma^{2(t+1)})$) distribution, where $\overline{\mu}^{(t+1)} = \sum_{k=1}^{K}\mu_k^{(t+1)}/K$.
   (f) Draw $\mu_0^{(t+1)}$ as Normal($\overline{\mu}^{(t+1)}$, $\sigma^{2(t+1)}/(K\kappa_0^{(t+1)})$).

If models with unequal variances in the counties or the PSUs were used, then the above algorithm for model 4 would be modified in a manner analogous to the variation in the algorithm presented for model 2 with unequal variances.

Figure 2: Histograms of positive soil wind erosion values on cropland in six Iowa counties in 2003, log scale, modified scales. The histograms do not represent sets of actual data values, but some counties do have histograms with similar patterns.