County level variance estimates for small area estimation for the National Resources Inventory

Pushpal Mukhopadhyay*, Tapabrata Maiti, and Wayne A. Fuller
Iowa State University

Abstract

The National Resources Inventory (NRI) is a longitudinal national survey of U.S. non federal lands. The objectives of the survey are to produce estimates such as land use, land cover and soil erosion at the national and sub-national level.

The cover and crop management factor (C-factor) is one of the inputs used to determine soil erosion in the NRI survey. We develop an estimation model for the C-factor that can be used in small-area estimation for counties. The NRI data set contains a significant proportion of imputed values, where the unobserved values are determined by the sampling design. An existing small area procedure is extended to adjust for the effect of imputed data and is applied to the NRI.

Keywords: *Small Area Estimates, Imputation, Area Level Model, Soil Erosion, Universal Soil Loss Equation.*

1 Introduction

Survey statisticians frequently encounter small area estimation (SAE) problems. In small area estimation problems, estimates are sought for a domain with a "small" or "moderate"' sample size. Because of this small sample size, direct domain estimates have low precision. Estimation approaches that "borrow strength" from similar areas using explicit or implicit models are described in (Rao, J.N.K. (2003)). In almost all large surveys, some form of imputation is used. Several approaches have been taken to produce a valid estimate and its variance when imputed data are present, for example see (Rao, J.N.K. and Shao, J. (1992)) and (Sarndal, C.E. (1992))). Not much work has been done to consider the effects of imputation on SAE.

The National Resources Inventory (NRI) collects annual data on US non-federal land. Among other variables, the C factor (a variable highly related with soil erosion) is recorded for each selected sample point for which erosion is to be calculated. In this work we will estimate the C factor for each county in Iowa for the year 2002. In practice a number of variables including erosion would be estimated, but we study only the C factor. The NRI collects data through a supplemented panel design, where a fixed panel (core) is observed every year. Although the entire core panel is usually observed in each year, only a random sample of the core was observed in the year 2002. The unobserved part of the core is imputed using a hot-deck type single imputation procedure.

If we consider a small area model for the C factor, then the design variances of county means are required at the first stage of the model and these variances depend on the imputation procedure. The sampling errors from two different counties are not independent because of imputation. We will (i) estimate the sampling error covariance matrix adjusted for imputation, (ii) fit a multivariate area level model using the sampling error covariance, and (iii) contrast the predicted values from complete data analysis with those from available data analysis. To estimate the sampling error covariance matrix we fit a regression model within each imputation cell which closely matches the imputation procedure used in the NRI.

In section 2 we describe the design of the NRI survey and the current imputation procedure. In section 3 we describe small area models for county level means and propose a method for estimating the sampling error covariance matrix. Results and findings are given in section 4 and conclusions are in section 5.

2 The NRI Survey

The NRI is a longitudinal survey conducted by the US Department of Agriculture's (USDA) Natural Resources Conservation Service (NRCS) in cooperation with the Iowa State University Center for Survey Statistics and Methodology (CSSM). The survey was designed to assess conditions and trends for land cover, soil, water, and related natural resources on non-federal lands in the United States. The NRI survey is a stratified two-stage area survey. The primary sampling units (PSU) are the divisions of the US land defined by the Public Land Survey System (PLS). Three sample points are then selected within

*204 Snedecor Hall, Iowa State University, Ames, IA, 50011; E-mail: pushpal@iastate.edu

most PSU's according to a restricted randomization procedure (Nusser, S.M. and Goebel, J.J. (1997)). Since 2000, the full panel structure of the NRI has been replaced by a two-phase supplemented panel sampling design in which the 1997 NRI segments serve as a first phase, and each year a partially overlapping panel is selected through a stratified sampling design as a second phase. The annual sample includes approximately 42,000 "core" segments that are to be visited every year. An additional 30,000 segments are selected from the remaining 268,000 PSUs each year to from a supplemental sample.

2.1 Variables of Interest

Data collection to estimate soil loss is a major focus for the NRI. Soil loss is estimated using the universal soil loss equation (USLE). The USLE is not collected directly, rather it is calculated using several factors related to soil properties and farming practices. The cover and crop management factor (C factor) is one important factor in the USLE. Other factors are soil support factor (P factor), rainfall factor (R factor), soil erodibility factor (K factor), slope length (SL), and slope percent (SP). R factor, P factor, and K factor can be obtained for each point from administrative records (NRCS, soil science data base). SP and SL are directly observed in the field. In this article we will focus only on the C factor, as it is observed for each selected sample point and unobserved values are imputed for all points that require USLE. The C factor in the USLE measures the combined effect of all the interrelated cover and crop management variables. It is defined as the ratio of soil loss from land maintained under specified conditions to the corresponding loss from continuous tilled bare fallow. The value of C is usually expressed as an annual value for a particular cover and crop management system but is calculated from the soil loss ratios for short periods of time within which cover and management effects are relatively uniform. The soil loss ratios are combined in proportion to the applicable percentages of erosion index (EI) to derive annual C values. Broad use (BU) of the land and land cover use (LU) are related to the C factor. Also slope percent, irrigation practice, rotation of crop and cowardin classification of wetland systems are used to determine the value for the C factor ((Rosewell, C.J. (1993))).

2.2 Current Imputation Procedure

The C factor was only observed for half of the core (P00.1) in 2002 and the missing half of the core (P00.2) was imputed. Imputation cells were created using matching BU, LU, SP, irrigation type (IT), and cowardin wetland classification (COW) for every recipient point. Then a donor is chosen from the same imputation cell as the recipient point. Finally, the missing value is imputed using a ratio adjusted donor value based on values for the years 1997 and 2003.

The imputation cells and small areas (counties) are not the same. For a missing observation in county i the donor can come from county j (both the donor and the missing observation must be in the same imputation cell). Hence estimated county means are not independent. If we assume that observed values in two distinct counties are independent then the correlation between two county estimates is due to the imputed values.

3 County Level Estimates

Let the finite population $\mathcal{F}$ with index set $U = \{1, 2, ..., N\}$ be divided into m subdivisions (counties) $\{U_i\}_{i=1}^m$. Let A_1 be a set of indexes of a sample of size n from the above population, A_r be a set of indexes of r observed values and A_m be a set of indexes of $n - r$ unobserved values. Assume that A_1 can be divided into G poststrata (imputation cells) such that $A_1 = U_{g=1}^G A_{1g}$, $A_r = U_{g=1}^G A_{rg}$, and $A_m = U_{g=1}^G A_{mg}$. Further, let y_{igk} be the k^{th} C factor in county i and imputation cell g, and π_{igk} be the selection probability of y_{igk} in A_1. A design based estimate of the mean C factor in county i, denoted by $\bar{y}_{i..}$, is

$$\bar{y}_{i..} = N_{i+}^{-1}\{\sum_{g=1}^{G}\sum_{k\in A_{rig}} w_{igk}y_{igk} + \sum_{g=1}^{G}\sum_{k\in A_{mig}} w_{igk}z_{igk}\}, \quad (1)$$

where N_{i+} is the population size of county i, $w_{igk} = \pi_{igk}^{-1}$ and z_{igk} are imputed values.

3.1 Small Area Model

If we have reasonable county level covariates then the estimated sample means, $\bar{y}_{i..}$, can be used to fit an area level small area model. Let θ_i be the true unobserved mean C factor for county i. Then with $\mathbf{y} = (\bar{y}_{1..}, ..., \bar{y}_{m..})^T$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)^T$, $\mathbf{u} = (u_1, ..., u_m)^T$, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ we write,

$$\begin{aligned} \mathbf{y} &= \theta + \mathbf{e}, \\ \boldsymbol{\theta} &= X\boldsymbol{\beta} + \mathbf{u} \end{aligned} \quad (2)$$

where X is a $m \times p$ matrix of covariates, $\mathbf{e}$ is the sampling error, and $\mathbf{u}$ is an area specific random quantity. We further assume that

$$(\mathbf{u}^T, \mathbf{e}^T) \sim N(\mathbf{0}, blockdiag.\{\begin{bmatrix} \sigma^2 I & \mathbf{0} \\ \mathbf{0} & \Sigma_{ee} \end{bmatrix}\}). \quad (3)$$

Then with $\Sigma_{zz} = \sigma^2 I + \Sigma_{ee}$ and assuming σ^2 and Σ_{zz} are known, the best linear unbiased predictor (BLUP) of $\boldsymbol{\theta}$ is,

$$\tilde{\boldsymbol{\theta}} = X\tilde{\boldsymbol{\beta}} + \sigma^2\Sigma_{zz}^{-1}(\mathbf{y} - X\tilde{\boldsymbol{\beta}}), \tag{4}$$

where

$$\tilde{\boldsymbol{\beta}} = (X^T\Sigma_{zz}^{-1}X)^{-1}X^T\Sigma_{zz}^{-1}\mathbf{y}. \tag{5}$$

Since σ^2 is unknown, the empirical BLUP (EBLUP) estimator of $\boldsymbol{\theta}$ can be obtained by substituting the estimated σ^2 in (4) and (5). The explicit form is

$$\hat{\boldsymbol{\theta}} = X\hat{\boldsymbol{\beta}} + \hat{\sigma}^2\tilde{\Sigma}_{zz}^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}}), \tag{6}$$

where

$$\hat{\boldsymbol{\beta}} = (X^T\tilde{\Sigma}_{zz}^{-1}X)^{-1}X^T\tilde{\Sigma}_{zz}^{-1}\mathbf{y}, \tag{7}$$

$\tilde{\Sigma}_{zz} = \hat{\sigma}^2 I + \Sigma_{ee}$ and Σ_{ee} assumed is known.

The hot deck imputation procedure used in the NRI will affect the sampling part of the small area model (2) but the relation between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ will remain unchanged. In section (3.2) we propose a methodology to estimate the design covariance matrix when imputed values are used.

(Datta, et al. (1992)) obtained a second-order approximation for the covariance matrix $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ as,

$$MSE(\hat{\boldsymbol{\theta}}) \approx \mathbf{G}_1(\sigma^2) + \mathbf{G}_2(\sigma^2) + \mathbf{G}_3(\sigma^2) \tag{8}$$

where $\mathbf{G}_1(\sigma^2) = \Sigma_{ee} + \Sigma_{ee}\Sigma_{zz}^{-1}\Sigma_{ee}$, $\mathbf{G}_2(\sigma^2) = \Sigma_{ee}\Sigma_{zz}^{-1}X(X^T\Sigma_{zz}^{-1}X)^{-1}X^T\Sigma_{zz}^{-1}\Sigma_{ee}$, and $\mathbf{G}_3(\sigma^2) = \Sigma_{ee}K^3\Sigma_{ee}V(\hat{\sigma}^2)$ with $K = \Sigma_{zz}^{-1} - \Sigma_{zz}^{-1}X(X^T\Sigma_{zz}^{-1}X)^{-1}X^T\Sigma_{zz}^{-1}$. The first term in expression (8) is the prediction covariance matrix if all parameters are known, the second term is due to the uncertainty of estimating β and the third term is due to the uncertainty of estimating σ^2.

A second-order approximation to the estimator of $MSE(\hat{\theta})$ can be obtained by replacing σ^2 by its estimator $\hat{\sigma}^2$ and by accounting for the bias associated with $\mathbf{G}_1(\hat{\sigma}^2)$. So if $\hat{\sigma}^2$ is an unbiased estimator of σ^2 then $MSE(\hat{\theta})$ can be estimated by,

$$mse(\hat{\theta}) \approx \mathbf{G}_1(\hat{\sigma}^2) + \mathbf{G}_2(\hat{\sigma}^2) + 2\mathbf{G}_3(\hat{\sigma}^2). \tag{9}$$

3.2 Estimation of the Covariance and the Imputation Model

A direct survey estimate of $V(\bar{y}_{i..})$, ignoring the fact that some values were imputed, may seriously under estimate the true variance (Sarndal, C.E. (1992)). To estimate the variance for an imputed data set, we must consider the response mechanism, the survey design, and the imputation model. The response mechanism defines the distribution governing the response given the sample (which is defined by the supplemented panel design in the NRI), the survey design defines how the sample is chosen, and the imputation model (implicit or explicit) defines how the missing values are imputed.

To estimate the covariance matrix Σ_{ee} we will define an explicit imputation model (ξ) which closely matches the model of the current imputation method.

Result 1. Let E_ξ be the expectation with respect to the imputation model (ξ) and assume

(A1) Uniform nonresponse mechanism within imputation group g.

(A2) Assume the z_{igk} are obtained using a hot deck imputation such that

$$E_*[z_{igk}] = \{\sum_{k\in A_{rig}} w_{igk}\}^{-1} \sum_{k\in A_{rig}} w_{igk}y_{igk} \tag{10}$$

and

$$V_*[\sum_{k\in A_{mig}} w_{igk}z_{igk}] =$$
$$(1-p_g)^2 N_{ig}^2(1 - n_{r+g}^{-1}n_{m+g})(n_{m+g})^{-1}\sigma_g^2$$
$$(= (1-p_g)^2 N_{ig}^2(1 - n_{r+g}^{-1}n_{m+g})(n_{m+g})^{-1}S_{A_{r+g}}^2)$$

where $*$ denotes the expectation and variance with respect to the hot deck imputation and $p_g = n_{1+g}^{-1}n_{r+g}$ denotes response rate within imputation cell g.

(A3) If there are no missing values, the county sampling errors are independent.

(A4) Missing values are imputed through a hot-deck type imputation using imputation cells and a cell-model of the form $\xi : y_{igk} = \mathbf{q}_{igk}^T\boldsymbol{\gamma} + \epsilon_{igk}$, where $\epsilon_{igk} \overset{ind.}{\sim} (0, \sigma_g^2)$ within each imputation cell and $\mathbf{q}_{igk}$ are unit level covariates.

(A5) The same donor is not used twice.

(A6) The number of donors from county i used to impute missing values in county i' is known and denoted by $\tau_{ii'}$.

Then we have the following results:

(R1) $E[\bar{y}_{i..}|\mathcal{F}] = E[E\{E_*(\bar{y}_{i..})|A_1, \mathcal{F}\}|\mathcal{F}]$ is conditionally asymptotically unbiased for θ_i.

(R2) The covariance of the direct county means $\bar{y}_{i..}$ and $\bar{y}_{i'..}$ is given by

$$Cov(\bar{y}_{i..}, \bar{y}_{i'..}) = (n_i n_{i'})^{-1}\sum_{j=1}^{J}\{(n_{+j})^{-1} V(\bar{y}_{.j.})(\tau_{ii'j} + \tau_{i'ij})\} \tag{11}$$

and an estimate of the covariance in (11) is

$$cov(\bar{y}_{i..}, \bar{y}_{i'..}) = (n_i n_{i'})^{-1}(\tau_{ii'} + \tau_{i'i}) J^{-1}\sum_{j=1}^{J}\{(n_{+j})^{-1}\hat{\sigma}_j^2\}. \tag{12}$$

(R3) The variance of the county mean $\bar{y}_{i..}$ is

$$V([\bar{y}_{i..}|\mathcal{F}]) = V[\tilde{y}_{1i..}|\mathcal{F}] + N_{i+}^{-2}\sum_g E[p_g(1-p_g) \sum_{k\in A_{1ig}}\{w_{igk}y_{igk} - R_{A_{1ig}}w_{igk}\}^2|\mathcal{F}] + N_{i+}^{-2}\sum_g(1-p_g)^2 N_{ig}^2(1-n_{r+g}^{-1}n_{m+g})(n_{m+g})^{-1} \sigma_g^2 + O(n_{1i+}^{-3/2}), \quad (13)$$

and assuming simple random nonreplacement sample within county (SRSWOR) an estimate of the variance in (13) is

$$\hat{V}[\bar{y}_{i..}|\mathcal{F}] = (1-f_{1i})n_{1i+}^{-1}\hat{S}_i^2 + n_{1i+}^{-2}\sum_g(1-p_g) \{(f_{1i}N_{ig}-1)\hat{S}_{rig}^2 + (f_{1i}N_{ig}-n_{1ig})\bar{y}_{rig}^2\} + N_{i+}^{-2}\sum_g(1-p_g)^2N_{ig}^2 (1-n_{r+g}^{-1}n_{m+g})(n_{m+g})^{-1}\hat{\sigma}_g^2, \quad (14)$$

where $n_{1i+} = \sum_{g=1}^{G}\sum_{k\in A_{1ig}} w_{igk}$ is the weighted sum of sample points in county i, $n_{1+g} = \sum_{i=1}^{m}\sum_{k\in A_{1ig}} w_{ijk}$ is the weighted sum of sample points in imputation cell g, and $\hat{\sigma}_g^2$ denotes the estimate of imputation model variance σ_g^2,

$$\hat{S}_i^2 = N_{i+}^{-1}\sum_g p_g^{-1}\sum_{k\in A_{rig}}(y_{igk} - n_{ri+}^{-1}\sum_g\sum_{k\in A_{rig}} y_{igk})^2 \quad (15)$$

and

$$\hat{S}_{rig}^2 = (n_{rig}-1)^{-1}\sum_{k\in A_{rig}}(y_{igk} - n_{rig}^{-1}\sum_{k\in A_{rig}} y_{igk})^2, \quad (16)$$

f_{1i} is the original sampling fraction in county i, p_g is the nonresponse rate in imputation cell g, and $\tilde{y}_{i..}$ is the mean of y_{igk} in county i if there were no missing values.

Using equations (12) and (14) the complete form of Σ_{ee} is known and we now can apply methods discussed before to estimate the county level mean for C factor. Under mild assumptions it can be shown that $V^{-1}\hat{V} \xrightarrow{p} 1$. Assumptions (A1) - (A3) are standard and are justified for our setup. We will justify the assumption (A4) in the next section through numerical results. Assumptions (A5) and (A6) are not very usual and need more explanations. For the NRI surveys, we know exactly how many points from county i are used as donors for recipient points in county i'. Given this information, (A6) is known and also there are very few points that are used twice as a donor. Although (A5) is not exactly true, the number of times the same donors are used twice is small. Proof of the result is omitted.

Cell	Num. Obs.	Num. Miss.	Mean	S.E.
1	2321	918	0.210	0.0584
5	25	13	0.130	0.0672
9	87	32	0.092	0.0580
21	113	50	0.016	0.0258
33	439	156	0.013	0.0095
46	270	133	0.004	0.0001

Table 1: *Final Imputation cells with mean and standard error of C factor. Num. Obs. and Num. Miss. denote number of observed sample points and number of missing points in each cell respectively.*

4 Results

We have considered the estimation of C factor at county level for Iowa for the year 2002. There are a total of 99 counties for which small area estimations are required. We have used point level data for the survey years 1997, 2001, and 2002 and from panels P00.1, P00.2, and the 2002 supplement (P02). The panel P02 is replaced by the 2001 supplement panel (P01) for the year 2001. There are a total of 8340 sample points but only 4557 sample points required USLE and hence C factor (note that although there are a total of 12 BU categories C factor/USLE is required only for four categories; viz., cultivated cropland, non-cultivated cropland, pasture land, and conservation reserve practices). Among 4557 sample points, C factor is observed for 3255 sample points and 1302 sample points have imputed values. In our complete data analysis we will include all of these 4557 units and we will compare it with observed data analysis where we will use only 3255 observed sample points.

4.1 Imputation Model

In the NRI survey imputation cells were created based on the BU, LU, SP, IT, and COW. BU has 4 categories, COW has 3 categories, IT has 2 categories, SP is divided into 4 categories ([0.1,2.0), [2.0,4.0), [4.0,8.0), and [8.0,36]), and LU is divide into 4 categories to form 384 possible cells. Most of these cells have no observation or very few observations. We merged some of these small imputation cells to obtain cells with a reasonable number of sample points. Table (1) shows the final six imputation cells with the number of observed points, number of missing points, estimated design mean and standard deviation for the C factor.

	Q_1	Median	Mean	Q_3
Fitted	0.014	0.193	0.155	0.233
Imputed	0.020	0.190	0.156	0.240

Table 2: *Summary of fitted values from the imputation model. Q_1 and Q_3 are the first and third quartiles respectively.*

We considered the C factor for 1997 (C97) and C factor for 2001 (C01) as covariates for imputation models. Although slope percent is a continuous variable, we have treated it as a factor with four levels to create imputation cells. We have also considered slope percent for the year 2002 (SP02) as a possible continuous covariate in the imputation model. Design weights vary across sample points and hence are also considered as a possible covariate for the imputation model. Given these possible covariates, we searched for the best parsimonious imputation model within each imputation cell. We fit an overall mean in cell 21, 33, and 46; and a model of the form $y_{ijk} = \gamma_0 + \gamma_1 C01_{ijk} + \gamma_2 C97_{ijk} + \gamma_3 SP02_{ijk} + \gamma_4 Weight_{ijk} + \epsilon_{ijk}$ in cells 1, 5, and 9. A summary of the predicted values from the fitted imputation model against the original imputed values for the panel P00.2 is shown in the table (2). This table suggests that the predictions from our explicit imputation model closely match the original imputed values.

4.2 Proposed small area model and county level estimates

Once a reasonable estimator of Σ_{ee} is obtained from the unit level information, several soil properties, such as, soil erodibility index, soil support factor, soil texture, erosion index, and slope percent are the possible covariates for the small area model. These soil information can be obtained from the USDA soil science databases and can be treated as known. Other soil information can be obtained from the 1997 NRI survey.

5 Conclusion

We considered the effect of imputation on a small area model, the missing values were imputed through hot deck imputation using regression. An unit level imputation model was built which closely matches the imputation procedure used in the NRI. The residual sum of squares from this imputation model is used to estimate the extra variability due to the imputation. Since the imputation cells cross county boundaries, the county estimates are correlated. A method of estimating these correlations using the fitted imputation model is proposed. A multivariate small area model can then be fitted assuming the estimated design variance as known and the EBLUP estimator and its estimated MSE can be used to produce county estimates. It is shown that the inflation of design variance due to imputed values has an effect on small area estimates and the proposed methodology will adjust for this effect.

References

Bell A., Drignei D., et al. (2003), "Estimation procedure for 2002 NRI." *Center for Survey Statistics and Methodology, National Resource Conservation Service, Unpublished technical report* 2003.

Datta, G.S., Ghosh, M., Huang, E.T., Isaki, C.T., Shultz, L.K., and Tsay, J.H., (1992), "Hierarchical and empirical Bayes methods for adjustment of census undercount: The 1998 Missouri dress rehearsal data" *Survey Methodology*, **18**, 95-100.

Fay, R.E., and Herriot, R.A., (1979), "Estimation of income from small places: An application of James-Stein procedures to census data." *Journal of the American Statistical Association*, **74**, 269-277.

Fuller, W.A., (2003), "Sample selection for the 2000 NRI-2004 NRI surveys." Unpublished manuscript.

Nusser, S.M., and Goebel, J.J., (1997), "The national resource inventory: a long-term multiresource monitoring program." *Environmental and Ecological Statistics*, **4**, 181-204.

Rao, J.N.K., (2003), "Small Area Estimation," *Wiley Series in Survey Methodology.*

Rao, J.N.K., Shao, J., (1992), "Jackknife variance estimation with survey data under hot deck imputation." *Biometrika* 1992; **79**, 811-822.

Rosewell C.J., (1993), "A program to assist in the selection of management practices to reduce erosion." *Technical Handbook No. 11* 1993; Soil Conservation Service, ISBN 0 7305 5931 9.

Sarndal, C.E., (1992), "Methods for estimating the precision of survey estimates when imputation has been used." *Survey Methodology* 1992; **18**, No.2, 241-252.