SMALL AREA ESTIMATION IN THE NATIONAL RESOURCES INVENTORY

Junyuan Wang, Wayne A. Fuller, and Jean Opsomer
Iowa State University
Junyuan Wang, Department of Statistics, Iowa State University, Ames, IA 50011 USA
(jywang@iastate.edu)

Key Words: natural resources; components-of-variance model

Abstract:

The National Resources Inventory is a large nationwide survey of the U.S. land area. Measurements are made on many characteristics, including land use. Estimates of land uses that represent a relatively small fraction of total area, such as roads, are desired for counties and portions of counties. A small area estimation scheme using auxiliary information and a components-of-variance model is described and estimation results presented.

1. Introduction

Few national surveys are designed to provide direct sample estimates for the smallest geographic areas of interest. For example, the U.S. Current Population Survey is of a size to provide direct estimates only for the larger states. In the same way, the crop reports issued by the U.S. Department of Agriculture are based on direct sample estimates at the state level. Thus, it is reasonable to consider model-based procedures called *small area estimation* when estimates are desired for smaller geographic units. An example of such units are counties in the United States. We describe the application of small area estimation procedures for a large U.S. survey called the *National Resources Inventory.*

2. The U.S. National Resources Inventory

The Iowa State Statistical Laboratory cooperates with the U.S. Natural Resources Conservation Service on a large survey of land use in the United States. The survey is a panel survey and was conducted in 1982, 1987, 1992, and 1997.

The survey collects data on soil characteristics, land use, land cover, wind erosion, water erosion, and conservation practices. The data are collected by employees of the Natural Resources Conservation Service. Iowa State University has responsibility for sample design and for estimation. See Nusser and Goebel (1997) for a complete description of the survey.

The sample is a stratified sample of all states and Puerto Rico. The sampling units are areas of land called *segments.* The segments vary in size, from 40 acres to 640 acres. Data are collected for the entire segment on items such as urban land and water area. Detailed data on soil properties and land use are collected at a random sample of points within the segment. Generally, there are three points per segment, but 40-acre segments contain two points and the samples in two states contain one point per segment. Some data, such as total land area, federally owned land and area in large water bodies, are collected on a census basis external to the sample survey. The current sample contains about 300,000 segments and about 800,000 points.

The sample size is such that direct estimates have acceptable variances for subdivisions of the surface area called *hydrologic units.* Hydrologic units are, essentially, drainage areas for major streams. There are about 200 hydrologic units in the United States. The estimation procedure is designed to reproduce the correct acreage for counties where conuties are important political subdivisions. There are about 3,100 counties in the United States. Because the sample must provide consistent acreage estimates for both counties and hydrologic units, the basic tabulation unit is the portion of a hydrologic unit within a county. This unit is called a *HUCCO.* There are about 5,000 HUCCOs. Some HUCCOs are relatively small and may contain only one segment. Some are relatively large and contain more than 100 segments.

This research was funded in part by cooperative agreement 68-3A75-43 between the USDA Natural Resources Conservation Services and Iowa State University. Computing for the research reported in this paper was done with eqipment purchased with funds provided by an NSF SCREMS grant award DMS 9707740.

3. Small Area Estimation

In the National Resources Inventory, small area estimation is used in the estimation of area in roads and in the estimation of acres in urban and built-up areas. Urban land is divided into two categories on the basis of the size of the tract. We present the analysis for the change in the sum of the two categories of urban land acreage from 1992 to 1997.

3.1 Small area estimation model for urban change

The urban area was determined for each year of the NRI survey. Let $U92_{kl}$ and $U97_{kl}$ denote urban area in HUCCO l of county k in 1992 and in 1997 respectively. The data used in small area estimation for urban change are $D_{kl} = U97_{kl} - U92_{kl}$, the direct estimated change from 1992 to 1997. The auxiliary information is the 1992 population and 1997 population. Population data are only available for counties, not for HUCCOs. Therefore, we defined two variables

$$\begin{aligned} \tilde{z}_{kl,1} &= (U92_k^{-1} U92_{kl}) P92_k \\ \tilde{z}_{kl,2} &= (U92_k^{-1} U92_{kl})(P97_k - P92_k), \end{aligned}$$

where $U92_k$ is the urban acres in county k in 1992, $P92_k$ and $P97_k$ are the populations of county k in 1992 and in 1997 respectively. We expect both variables to be positively related to the change in urban acres. Because a reduction in urban area is extremly rare, we set the population change variable to a small positive number if the population change is negative.

Some heavily urbanized areas have a large population and very little area in non-urban uses that can be converted to urban use. Therefore the actual regression variables used in the analysis were constructed to recognize the availability of potentially convertible land. The variables are

$$\begin{aligned} z_{kl,1} &= min(R_1 \tilde{z}_{kl,1}, 0.5 C_{kl}) \\ z_{kl,2} &= min(R_2 \tilde{z}_{kl,2}, 0.5 C_{kl}), \end{aligned}$$

where R_j is the ratio of the total change in urban for the state to the sum of $\tilde{z}_{kl,i}$ for the state, and C_{kl} is the total area available for conversion to urban use in HUCCO kl.

In the sequel, we use the sigle subscript i in place of the double subscript kl as the index for the HUCCO. Our goal is to predict d_i, the unobservable true value of change in urban area. The distribution of D_i is highly skewed. Therefore, a power transformation $Y_i = D_i^{0.375}$ is used in the estimation. A model for small area estimation is

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + n_i^{0.375} b_i \qquad (3.1) \\ Y_i &= y_i + e_i, \qquad (3.2) \end{aligned}$$

where $y_i = d_i^{0.375}, x_{1i} = z_{1i}^{0.375}, x_{2i} = z_{2i}^{0.375}$, n_i is the number of sample points in the i-th HUCCO, b_i is the area effect, e_i is the sampling error, and $(b_l,\ e_i), i = 1,\ 2,\ \ldots,\ m$, are independent vectors with normal distribution

$$\begin{pmatrix} b_i \\ e_i \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_b^2 & 0 \\ 0 & \sigma_{ei}^2 \end{pmatrix} \right). \qquad (3.3)$$

We designed the small area estimation procedure to be fully automated because estimates were to be constructed for about 5,000 small areas using more than fifty analysis units, where the typical analysis unit is a state. Therefore, we used relatively simple procedures in place of complicated procedures that might produce marginal gains in efficiency. One could design an iterative estimation procedure for β_1 and β_2 in which the estimated between area component of variance is used to estimate the covariance matrix of $u_i = n_i^{0.375} b_i$. We used a simple weighted least squares procedure where the weights were a function of n_i. The estimators of $(\beta_1, \beta_2)'$ is

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \mathbf{H}^{-1} \begin{pmatrix} \sum_i n_i^{-0.25} x_{1i} Y_i \\ \sum_i n_i^{-0.25} x_{2i} Y_i \end{pmatrix}, \qquad (3.4)$$

where

$$\mathbf{H} = \begin{pmatrix} \sum_i n_i^{-0.25} x_{1i}^2 & \sum_i n_i^{-0.25} x_{1i} x_{2i} \\ \sum_i n_i^{-0.25} x_{1i} x_{2i} & \sum_i n_i^{-0.25} x_{2i}^2 \end{pmatrix}.$$

Because the variance of Y_i is closely related to $n_i^{-0.25}$, the procedure is close to estimated generalized least squares. Our model estimator of y_i is

$$\hat{y}_i = x_{1i}\hat{\beta}_1 + x_{2i}\hat{\beta}_2 \qquad (3.5)$$

Let $V(D_i|d_i) = V(e_i^\eta), \eta = 0.375^{-1}$, be the within HUCCO sample error. The variance of e_i^η can be estimated directly from the sample data. However, the sample size is relatively small in some HUCCOs and, hence, the variance of the estimated variance is large. Therefore, a model was developed for $V(D_i|d_i)$ to provide an improved variance estimator for small HUCCOs.

If we had a simple random sample of points with a zero-one indicator for urban change, the sample variance of D_i would be

$$\begin{aligned} V\left(e_i^\eta\right) &= V\left(N\hat{p}\right) = N^2 n_i^{-1} PQ \\ &= N n_i^{-1}\left(NPQ\right) \doteq N n_i^{-1}\left(NP\right) \ ,(3.6) \end{aligned}$$

where N is the population number of sample units, P is the proportion of the units that change into urban, and $\hat{p}$ is the corresponding sample proportion. In our notation, $D_i = N\hat{p}$. Now $n_i^{-1}N$ is constant under proportional sampling. Thus, $V\left(e_i^\eta\right)$ should be proportional to $E\left(D_i\right) = NP$. In our data the variance may increase at a slightly faster rate. If P were nearly constant, then $n_i^{-1}E\left(D_i\right) = n_i^{-1}NP$ would be nearly constant too. This means that $E\left(D_i\right)$ would be almost proportional to n_i. It seems reasonable to approximate the variance of D_i with a function of both n_i and $E\left(D_i\right)$. Our approximation is

$$V\left(e_i^\eta\right) \doteq C_1 \ n_i^{-0.25} E\left(D_i\right)^{1.25}, \qquad (3.7)$$

where C_1 is a constant to be determined. By Taylor approximation, the variance of e_i is

$$V\left(e_i\right) = \left[E\left(D_i\right)\right]^{2(\eta-1)} \eta^2 V\left(e_i^\eta\right). \qquad (3.8)$$

For $\eta = 0.375$, we have

$$V\left(e_i\right) = \sigma_w^2 n_i^{-0.25}. \qquad (3.9)$$

where σ_w^2 is a constant to be determined.

Let $\hat{V}\left(e_i^\eta\right)$ be the direct estimated variance of D_i from the sample data. The estimated variance of e_i is

$$\hat{V}\left(e_i\right) = 0.96\left(D_i^*\right)^{2(\eta-1)} \eta^2 \hat{V}\left(e_i^\eta\right) \qquad (3.10)$$

where 0.96 is an empirical adjustment and D_i^* is an estimator of $E\left\{D_i\right\}$. An estimator of the within component of variance is

$$\hat{\sigma}_w^2 = \left(\sum_i \delta_i\right)^{-1} \sum_i n_i^{0.25} \hat{V}\left\{e_i\right\}\delta_i \ , \qquad (3.11)$$

where

$$\delta_i = \begin{cases} 1 & \text{if } n_i > 2 \\ 0 & \text{otherwise.} \end{cases}$$

An estimator of the between-component of variance is

$$\hat{\sigma}_b^2 = \left(\sum_i n_i\delta_i\right)^{-1} \sum_i \ddot{\sigma}_{b,i}^2 n_i \delta_i, \qquad (3.12)$$

where $\ddot{\sigma}_{b,i}^2 = \left(n_i^{-0.375}\right)^2\left[\left(\hat{Y}_i - Y_i\right)^2 - \hat{V}\left(e_i\right)\right]$.

The predictor of y_i for the i-th HUCCO is

$$\begin{aligned} \tilde{y}_i &= \hat{y}_i + \hat{\alpha}_i\left(Y_i - \hat{y}_i\right) \\ &= \hat{\alpha}_i Y_i + \left(1 - \hat{\alpha}_i\right)\hat{y}_i \ , \qquad (3.13) \end{aligned}$$

where

$$\hat{\alpha}_i = \left(\hat{\sigma}_b^2 + n_i^{-1}\hat{\sigma}_w^2\right)^{-1}\hat{\sigma}_b^2 \ . \qquad (3.14)$$

Under the model, the error in $\tilde{y}_i$ as an estimator of y_i is

$$\begin{aligned} &\tilde{y}_i - y_i = \hat{\alpha}\left(u_i + e_i\right) - u_i \\ &\quad + \left(1 - \hat{\alpha}_i\right)\left[x_{1i}(\hat{\beta}_1 - \beta_1) + x_{2i}(\hat{\beta}_2 - \beta_2)\right], \ (3.15) \end{aligned}$$

where $u_i = n_i^{0.375} b_i$. If $\hat{\alpha}_i$ were treated as a fixed quantity and the possible covariance between $(e_i, \ u_i)$ and $(\hat{\beta}_1, \hat{\beta}_2)$ ignored,

$$\begin{aligned} &\hat{V}\left\{\tilde{y}_i - y_i\right\} = n_i^{0.75}\left[(1 - \hat{\alpha}_i)^2\hat{\sigma}_b^2 + \hat{\alpha}_i^2 n_i^{-1}\widehat{\sigma_w}^2\right] \\ &\quad + (1 - \hat{\alpha}_i)^2(x_{1i}, x_{2i})\hat{V}\{(\hat{\beta}_1, \hat{\beta}_2)\}(x_{1i}, x_{2i})'. \ (3.16) \end{aligned}$$

The estimator (3.16) does not contain a contribution to the variance from estimating the variance components.

The predictor of change in total urban from 1992 to 1997 for HUCCO i is

$$\tilde{d}_i = \tilde{y}_i^\eta. \qquad (3.17)$$

The corresponding variance of $\tilde{d}_i$ is

$$\hat{V}\left\{\tilde{d}_i - d_i\right\} = (0.96)^{-1}(D_i^*)^{1.25}(0.375)^{-2}\hat{V}\left\{\tilde{y}_i - y_i\right\}. \qquad (3.18)$$

For confidence limits of $\tilde{d}_i$, it is preferable to set limits for $\tilde{y}_i$ and then exponentiate those limits. The $E\left(\tilde{y}_i^\eta\right)$ is not equal to $E\left(\tilde{d}_i\right)$. As a partial adjustment for this bias, and to maintain the design unbiasedness of the state estimated change, the small area estimates are ratio adjusted so that the sum of the estimates is equal to the sum of the direct estimates for the state in the production program.

3.2 Results and Model Checks

We present some plots and statistics for a set of HUCCOs in a state in the Midwest. The analyses are preliminary as the data have not been released. There are 163 HUCCOs in the state. The HUCCOs are divided into ten groups on the basis of the model predicted change in acres. The first group is the group of HUCCOs with $(x_{1i}, x_{2i}) = (0, 0)$. This is the group of HUCCOs with zero estimated change.

There are either 15 or 16 HUCCOs in the remaining nine groups.

Table 1 contains summary statistics for the grouped data. While the groups were ordered on the predicted change, the number of segments is generally larger for HUCCOs with larger changes. The fact that the mean model (3.1) fits quite well is demonstrated by agreement between the two columns Y_i and $\hat{y}_i$ defined in (3.5), which is shown in Figure 1.

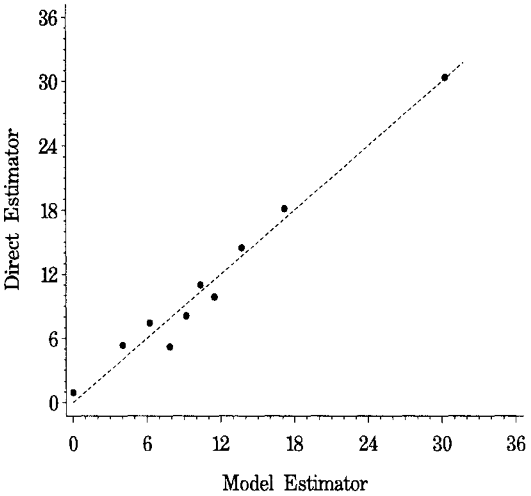

Figure 1: Mean Model

The t-statistics in the column "t-stat" in Table 1 were calculated as the difference between the group means divided by the standard error of the difference. The standard error was calculated using the mean square of the individual differences. The sum of squares of the ten t-statistics is 12.27. This value can be compared to the 5% tabular value of 18.31 for the chi-square distribution with 10 degrees of freedom. The model is easily accepted on the basis of this test. The R^2 computed for the regression of Y_i on (x_{1i}, x_{2i}) is 0.79.

Table 1: Summary statistics for mean model

	HUCCOs in group	Sample size mean	Group mean		t-stat
			Y_i	$\hat{y}_i$	
1	21	10.2	0.9	0.0	1.36
2	15	20.3	5.4	4.0	1.25
3	16	41.2	7.6	6.1	1.37
4	16	65.8	5.3	7.8	-2.14
5	16	74.1	8.2	9.1	-0.65
6	16	73.3	10.1	10.3	-0.14
7	16	78.2	10.2	11.5	-1.22
8	16	64.4	14.5	13.7	0.53
9	16	75.9	18.1	17.2	0.43
10	15	109.9	30.1	30.2	0.10

Figure 2 is the plot of grouped means of $[\hat{V}(e_i^\eta)]^{0.5}$ against the grouped means of $[(D_i^*)^{1.25} n_i^{-0.25}]^{0.5}$. The plot indicates that (3.7) is a resonable approxi-amtion to the sample variance $V(D_i|d_i) = \hat{V}(e_i^\eta)$.

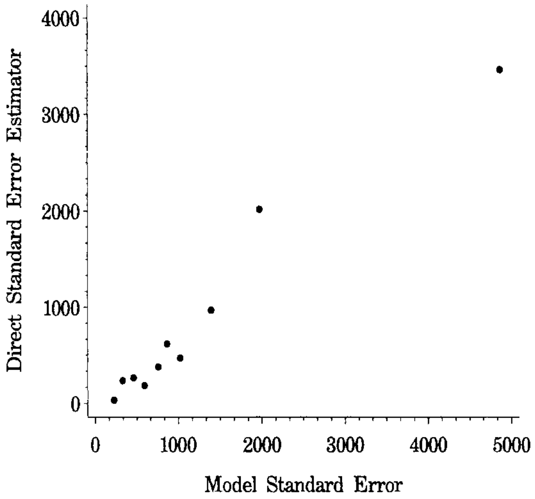

Figure 2: Variance Model

Table 2 contains statistics for sample standard errors using the groups of Table 1. The column headed $[\hat{V}(e_i)]^{0.5}$ contains the group averages of the square root of $\hat{V}(e_i)$, where $\hat{V}(e_i)$ is defined in (3.9). The averages of $(n_i^{-0.25}\hat{\sigma}_w^2)^{0.5}$ in the adjacent column decrease as one moves down the column because the average sample size increases. The entries in the column for $(n_i^{-0.25}\hat{\sigma}_w^2)^{0.5}$ were ratio adjusted so that the sum is equal to the sum of $[\hat{V}(e_i)]^{0.5}$. There are many direct estimates of zero in the first group, so that a comparison of $[\hat{V}(e_i)]^{0.5}$ and $(n_i^{-0.25}\hat{\sigma}_w^2)^{0.5}$ for that group says little about model adequacy. The sum of squares for the other nine t-statistics associated with groups two through ten is 13.44, considerably less than the tabular value of 16.92 for chi-square with nine degrees of freedom. Thus, there is no reason to reject the variance model.

The root transformation has a strong variance stabilizing effect. The standard errors of the original estimates of change for the largest group are about 3386 and the corresponding standard errors for the second group are about 110.

In this particular state, the estimator of the between-HUCCO component of variance was neg-

Table 2: Summary statistics for variance model

Group	HUCCOs in group	Sample size mean	Group mean		t-stat	Mean of $[\hat{V}\{\tilde{y}_i - y_i\}]^{0.5}$	Efficiency
			$[\hat{V}(e_i)]^{0.5}$	$(n_i^{-0.25}\hat{\sigma}_w^2)^{0.5}$			
1	21	10.2	1.16	4.02	-	0.50	-
2	15	20.3	3.79	3.76	0.02	0.84	4.48
3	16	41.2	3.66	3.35	0.37	1.38	2.42
4	16	65.8	1.76	3.16	-2.04	1.79	1.77
5	16	74.1	2.98	3.10	-0.14	1.94	1.60
6	16	73.3	3.63	3.14	0.33	1.93	1.62
7	16	78.2	2.59	3.04	-0.86	2.06	1.48
8	16	64.4	3.86	3.14	0.90	1.83	1.72
9	16	75.9	5.79	3.07	1.34	2.06	1.49
10	15	109.9	3.89	2.92	2.38	2.45	1.19

ative. In the production version of the program we impose a lower bound of 0.008 on the ratio of $\hat{\sigma}_b^2$ to $\hat{\sigma}_w^2$ in the computation of $\hat{\alpha}$. The lower bound means that the direct estimator in a HUCCO with 125 segments receives a weight of 0.5.

Because the estimated value for $\hat{\sigma}_b^2$ is zero, the estimated variance of the prediction error for $\tilde{y}_i$ was computed as

$$\begin{aligned}\hat{V}\{\tilde{y}_i - y_i\} &= \hat{\alpha}_i^2 n_i^{-0.25}\hat{\sigma}_w^2 \\ &+ (1-\hat{\alpha}_i)^2(x_{1i}, x_{2i})\hat{V}\{(\hat{\beta}_1, \hat{\beta}_2)\}(x_{1i}, x_{2i})'. \quad (3.19)\end{aligned}$$

The seventh column of Table 2 contains the mean of the prediction standard errors, where the standard error is the square root of the model variance of the predictor computed with equation (3.19). The last column of Table 2 is the ratio of the fifth column to the seventh column. There are large estimated gains in efficiency from using the small area model, particularly for HUCCOs with a small number of segments.

4. References

Fuller, Wayne A. (1997). "Small area inference for binary variables in the National Health Interview Survey", *Journal of American Statistical Association* **92**, 815-826.

Ghosh, M. and Rao, J.N.K. (1994). "Small area estimation: an appraisal", *Statistical Science*, **9**, 55-93.

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