Paper 336-2011

Small Area Estimation for Survey Data Analysis Using SAS® Software

Pushpal K Mukhopadhyay and Allen McDowell, SAS Institute Inc., Cary, NC

ABSTRACT

Small area estimation is important in survey analysis when domain (subpopulation) sample sizes are too small to provide adequate precision for direct domain estimators. Popular techniques for small area estimation use implicit or explicit statistical models to indirectly estimate the small area parameters of interest. Indirect estimation requires you to go beyond the survey data analysis methods that are available in the SAS/STAT® survey procedures. This paper describes the use of the MIXED, IML, and MCMC procedures to fit unit-level and area-level models, and to obtain small area predictions and the mean squared error of predictions. Hierarchical Bayes models are also discussed as extensions to the basic models.

INTRODUCTION

Estimating quantities of interest for subpopulations (also known as domains) with survey data is a common practice. Domains can be defined by any characteristics that partition the population into a set of mutually exclusive subpopulations. Common characteristics by which domains are defined are geographic areas such as states, counties, or municipalities, and demographic groups such as age, race, or gender. Domain estimators that are computed using only the sample data from the domain are known as *direct estimators*. Although direct estimators have several desired design-based properties, direct estimates often lack precision when domain sample sizes are small. Domains for which direct estimates of adequate precision cannot be produced are known as small areas. Survey designs usually focus on achieving a particular degree of precision for estimates at a much higher level of aggregation than that of small areas; therefore, the sample sizes for small areas are typically small. Producing estimates for small areas with an adequate level of precision often requires *indirect estimators* that use auxiliary data or values of the variable of interest from related areas, or both.

The traditional indirect estimators, such as synthetic and composite estimators, rely on implicit linking models. *Synthetic estimators* for small areas are derived from direct estimators for a large area that covers several small areas under the assumption that the small areas have the same characteristics as the large area. *Composite estimators* are essentially weighted averages of direct estimators and synthetic estimators. Both synthetic and composite estimators can yield estimates that provide higher precision compared to direct estimators. However, both types of estimators share a common tendency to be design-biased, and the design bias does not necessarily decrease as the sample size increases.

More recently, explicit linking models provide significant improvements in techniques for indirect estimation. Based on mixed model methodology, these techniques incorporate random effects into the model. The random effects account for the between-area variation that cannot be explained by including auxiliary variables. Most small area models can be defined as an area-level model, a unit-level model, or a hybrid. Area-level models relate small-area direct estimators to area-specific auxiliary data. Unit-level models relate the unit values of a study variable to unit-specific auxiliary data. Hybrid models involve both unit-level and area-level auxiliary variables.

This paper describes a unit-level model, a basic area-level model, and an unmatched sampling and linking area-level model. The section "UNIT-LEVEL SMALL AREA MODELS" illustrates the unit-level model with an example that considers the prediction of crop areas for some counties in Iowa. The MIXED procedure provides estimates of the model parameters and the small area predictions. Small area predictions are based on the empirical best linear unbiased predictors (EBLUP), and the mean squared error of predictions (MSEP) measures the precision of the predictions. The section "AREA-LEVEL SMALL AREA MODELS" illustrates the basic area-level model with an example that considers the prediction of wind erosion for some counties in Iowa. The model parameters are estimated with the MIXED procedure, and then the the EBLUPs and the MSEPs are computed with the IML procedure. The section "UNMATCHED MODELS" illustrates an unmatched sampling and linking area-level model with an example that considers estimating the undercoverage count and the undercoverage rate for provinces in the Canadian census. Standard linear mixed model theory cannot be applied to unmatched sampling and linking models. Instead, a hierarchical Bayes (HB) approach to estimation is taken. This approach uses the MCMC procedure to estimate the means and variances of the posterior distributions of the small area parameters of interest.

UNIT-LEVEL SMALL AREA MODELS

Unit-level models relate the unit values of a study variable to unit-specific auxiliary data. For example, suppose you have a survey of firms that is designed to estimate total wages and salaries paid to workers. Perhaps the survey is designed

so that estimates of a specified degree of precision can be made at the state level. After the survey is conducted, you decide you want to estimate total wages and salaries by industry, but the sample sizes for some industries are so small that the variances of the estimates are unacceptably large. To improve the precision of the estimates, you can use auxiliary data such as firm-level values of gross business income to fit a linear mixed model with industry-specific random effects to improve the efficiency of your estimates (Rao and Choudry 1995).

More formally, suppose y_{ij} is the value of a study variable in area i and unit j, for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, N_i$, where m is the number of small areas and N_i is the number of population units in area i. Assume that unit-specific auxiliary information $x_{ij} = (x_{ij1}, ..., x_{ijq})^T$ is available for every unit in the population, where q is the number of the auxiliary variables. A basic unit-level model relates the y_{ij} to the x_{ij} through a nested error regression model of the form

$$y_{ij} = x_{ij}^T \boldsymbol{\beta} + u_i + \epsilon_{ij} \tag{1}$$

where $\boldsymbol{\beta}$ is a fixed set of regression parameters, u_i are area-specific random effects, and ϵ_{ij} are the sampling errors. Suppose the parameters of interest are the small area means which are defined as

$$\theta_i = \bar{\boldsymbol{x}}_{i(p)}^T \boldsymbol{\beta} + u_i \tag{2}$$

where $\bar{\boldsymbol{x}}_{i(p)} = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}$ is the population mean. Assume $u_i \overset{ind.}{\sim} (0, \sigma_u^2)$, $\epsilon_{ij} \overset{ind.}{\sim} (0, \sigma_e^2)$, and that u_i and ϵ_{ij} are independent. Also assume that a sample of size n_i is selected from the N_i units in area i and that the sample values also satisfy population model (1).

If the variance parameters σ_u^2 and σ_e^2 are known, the best linear unbiased predictor (BLUP) for the small area means θ_i are given by

$$\tilde{\theta}_i = \bar{\boldsymbol{x}}_{i(p)}^T \tilde{\boldsymbol{\beta}} + \gamma_i \left(\bar{y}_{i.} - \bar{\boldsymbol{x}}_{i.}^T \tilde{\boldsymbol{\beta}} \right) \tag{3}$$

where $\gamma_i = \left(\sigma_u^2 + \sigma_e^2 / n_i \right)^{-1} \sigma_u^2$, $\tilde{\boldsymbol{\beta}}$ is the BLUP of $\boldsymbol{\beta}$, and $\bar{\boldsymbol{x}}_{i.}$ and $\bar{y}_{i.}$ are the sample means for area i. For areas with large n_i, γ_i is close to 1 and the predictor (3) is close to the regression predictor $\bar{y}_{i.} + \left(\bar{\boldsymbol{x}}_{i(p)} - \bar{\boldsymbol{x}}_{i.} \right)^T \tilde{\boldsymbol{\beta}}$. If $(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ is an estimator of (σ_u^2, σ_e^2), then an empirical best linear unbaised predictor (EBLUP) for θ_i is given by

$$\hat{\theta}_i = \bar{\boldsymbol{x}}_{i(p)}^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_i \left(\bar{y}_{i.} - \bar{\boldsymbol{x}}_{i.}^T \hat{\boldsymbol{\beta}} \right) \tag{4}$$

where $\hat{\gamma}_i = \left(\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i \right)^{-1} \hat{\sigma}_u^2$ and $\hat{\boldsymbol{\beta}}$ is an EBLUP for $\boldsymbol{\beta}$. Assuming the normality of the error components, the MSEP for $\hat{\theta}_i$ can be obtained following Kackar and Harville (1984), Prasad and Rao (1990), and Kenward and Roger (1997). The MSEP for $\hat{\theta}_i$ has the approximate form

$$\text{MSEP}(\hat{\theta}_i) \approx g_{1i}(\sigma_u^2, \sigma_e^2) + g_{2i}(\sigma_u^2, \sigma_e^2) + g_{3i}(\sigma_u^2, \sigma_e^2) \tag{5}$$

(6)

where

$$g_{1i}(\sigma_u^2, \sigma_e^2) = \gamma_i \sigma_e^2 / n_i \tag{7}$$

$$g_{2i}(\sigma_u^2, \sigma_e^2) = \left(\bar{\boldsymbol{x}}_{i(p)} - \gamma_i \bar{\boldsymbol{x}}_{i.} \right)^T \left(\sum_{i=1}^{m} A_i \right)^{-1} \left(\bar{\boldsymbol{x}}_{i(p)} - \gamma_i \bar{\boldsymbol{x}}_{i.} \right) \tag{8}$$

$$g_{3i}(\sigma_u^2, \sigma_e^2) = n_i^{-2} \left(\sigma_u^2 + \sigma_e^2 / n_i \right)^{-3} h(\sigma_u^2, \sigma_e^2) \tag{9}$$

with

$$A_i = \sigma_e^{-2} \sum_{j=1}^{n_i} \left(\boldsymbol{x_{ij}} \boldsymbol{x_{ij}}^T - \gamma_i n_i \bar{\boldsymbol{x_{i.}}} \bar{\boldsymbol{x_{i.}}}^T \right) \tag{10}$$

and

$$h(\sigma_u^2, \sigma_e^2) = \sigma_e^4 \bar{V}_{uu}(\boldsymbol{\delta}) + \sigma_u^4 \bar{V}_{ee}(\boldsymbol{\delta}) - 2\sigma_e^2 \sigma_u^2 \bar{V}_{ue}(\boldsymbol{\delta}) \tag{11}$$

where $\boldsymbol{\delta} = (\sigma_u^2, \sigma_e^2)^T$, $\bar{V}_{uu}(\boldsymbol{\delta})$, and $\bar{V}_{ee}(\boldsymbol{\delta})$ are the asymptotic variances of the estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$, and $\bar{V}_{ue}(\boldsymbol{\delta})$ is the asymptotic covariance of $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$.

Assuming normality of the errors u_i and ϵ_{ij}, an estimator of the MSEP (Rao 2003) is given by

$$\text{msep}(\hat{\theta}_i) = \text{g}_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + \text{g}_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + 2\text{g}_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) \tag{12}$$

where $(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ is an unbiased estimator of (σ_u^2, σ_e^2).

PROC MIXED provides estimators of (σ_u^2, σ_e^2) that satisfy these assumptions when you specify the Type 1 or residual (restricted) maximum likelihood (REML) estimation methods together with the Kenward-Roger method of covariance estimation. Thus, PROC MIXED can compute the EBLUPs and their MSEPs.

Example: Prediction of County Crop Areas

This example from Battese, Harter, and Fuller (1988) considers the prediction of areas planted with corn and soybeans for 12 counties in north-central Iowa. The area of corn and soybeans in the 37 segments (primary sampling units) of the 12 counties was determined by interviewing farm operators. Each segment represents approximately 250 hectares. This information is augmented by auxiliary data derived from satellite imagery readings. Crop areas for each segment are estimated from satellite images by counting the number of individual pixels in the satellite photographs. Each pixel, which can be either a corn or soybean crop, represents approximately 0.45 hectares. The objective of the study is to generate a predictor of mean crop areas per segment in the sample. The model assumes that there is a linear relationship between the survey and satellite data with county-specific random effects.

The survey results and satellite data are contained in the following SAS data set Corndata. The data set includes the variables County, Segments, Cornhec, Cornpix, and Soypix. The variable Segments is the total number of segments within each county. In the first 36 observations, Cornhec records the number of hectares of corn reported in the survey, Cornpix is the number of pixels reported for corn, and Soypix is the number of pixels for soybeans. In the last 12 observations of the data set, the variable Cornhec is set to missing, Cornpix contains the population mean number of pixels per segment for corn, and Soypix contains the population mean number of pixels per segment for soybeans. After the model parameters have been estimated, the population mean numbers of pixels for corn and soybeans are used to compute the EBLUPS in equation (4).

```
data corndata;
   length county $ 12;
   input county $ segments n @@;
   do i = 1 to n;
      drop i;
      input cornhec cornpix soypix @@;
      output;
   end;
   label county   = 'County'
         segments = 'Total Segments'
         n        = 'Sampled Segments'
         cornhec  = "Reported Hectares for Corn"
         cornpix  = "Number of Pixels for Corn"
         soypix   = "Number of Pixels for Soybeans";
datalines;
CerroGordo    545 1  165.76 374  55
Hamilton      566 1   96.32 209 218
Worth         394 1   76.08 253 250
Humbolt       424 2  185.35 432  96  116.43 367 178
```

```
Franklin      564 3  162.08 361 137  152.04 288 206  161.75 369 165
Pocahontas    570 3   92.88 206 218  149.94 316 221   64.75 145 338
Winenbago     402 3  127.07 355 128  133.55 295 147   77.7  223 204
Wright        567 3  206.39 459  77  108.33 290 217  118.17 307 258
Webster       687 4   99.96 252 303  140.43 293 221   98.95 206 222
                     131.04 302 274
Hancock       569 5  114.12 313 190  100.6  246 270  127.88 353 172
                     116.9  271 228   87.41 237 297
Kossuth       965 5   93.48 221 167  121    369 191  109.91 343 249
                     122.66 342 182  104.21 294 179
Hardin        556 5   88.59 220 262  165.35 355 160  104    261 221
                      88.63 187 345  153.7  350 190
CerroGordo    545 1 . 295.29 189.7    Hamilton      566 1 . 300.4  196.65
Worth         394 1 . 289.6  205.28   Humbolt       424 1 . 290.74 220.22
Franklin      564 1 . 318.21 188.06   Pocahontas    570 1 . 257.17 247.13
Winenbago     402 1 . 291.77 185.37   Wright        567 1 . 301.26 221.36
Webster       687 1 . 262.17 247.09   Hancock       569 1 . 314.28 198.66
Kossuth       965 1 . 298.65 204.61   Hardin        556 1 . 325.99 177.05
;
```

The following SAS statements use the MIXED procedure to estimate the regression parameters and the variance parameters for a unit-level small area model. The METHOD= option in the PROC MIXED statement specifies that the Type 1 estimation method be used. The Type 1 method provides a method of moments estimator which produces an unbiased estimate of the residual variance. The ASYCOV option requests the asymptotic covariance matrix for the variance parameters.

The DDFM=KENWARDROGER option in the MODEL statement performs the MSEP and the degrees-of-freedom calculations detailed by Kenward and Roger (1997). This method is based on taking more of the true nonlinearity of the mixed model estimates into account to achieve a higher order of accuracy for the estimated covariance of effects. The RANDOM statement specifies that a county-level random effect be included in the model.

```
proc mixed data = corndata method = type1 asycov order=data;
   class county;
   model cornhec = cornpix soypix / solution covb outp=pred
                                    ddfm=kenwardroger;
   random county / cl;
run;
```

Output 1 displays the results.

Output 1 Parameter Estimates

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate
county	139.68
Residual	149.56

Asymptotic Covariance Matrix of Estimates

Row	Cov Parm	CovP1	CovP2
1	county	7664.58	-714.68
2	Residual	-714.68	1960.89

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	51.0466	25.2010	29.5	2.03	0.0519
cornpix	0.3287	0.05165	28.9	6.36	<.0001
soypix	-0.1344	0.05731	29.9	-2.34	0.0259

The estimate of σ_u^2 is 139.68, and the estimate of σ_e^2 is 149.56. Standard errors for the estimated covariance parameters are the square root of the diagonals of the estimated asymptotic covariance matrix. Thus, the standard error of the estimate for σ_u^2 is 87.54 (the square root of 7664.58), and the standard error of the estimate for σ_e^2 is 44.28 (the square root of 1960.89).

The OUTP= option in the MODEL statement produces an output data set that contains the predicted values and their standard errors. Output 2 displays a part of the input data set with the prediction statistics. The coefficient of variation (CV) for the small area predictor is defined as the ratio of the estimated standard error and the predicted value. The last two columns are the model CV for the EBLUP and the design CV for the direct estimates. The direct estimates and the design CV for the direct estimates are obtained by using the DOMAIN statement in PROC SURVEYMEANS (statements not shown). The use of the DOMAIN statement is common for survey data when you have adequate sample sizes within domains. From the table, you can see that the model CVs for the EBLUPs are always lower than the design CVs for the direct estimates except for the Franklin, Hancock, and Kossuth counties. Design CVs for the direct estimates are not available for Cerro Gordo, Hamilton, and Worth counties. All these counties have only one sampled segment.

Output 2 Predicted Small Area Means and Prediction Errors

County	Total Segments	Sampled Segments	Number of Pixels for Corn	Number of Pixels for Soybeans
CerroGordo	545	1	295.29	189.70
Franklin	566	3	300.40	196.65
Hamilton	394	1	289.60	205.28
Hancock	424	5	290.74	220.22
Hardin	564	5	318.21	188.06
Humbolt	570	2	257.17	247.13
Kossuth	402	5	291.77	185.37
Pocahontas	567	3	301.26	221.36
Webster	687	4	262.17	247.09
Winenbago	569	3	314.28	198.66
Worth	965	1	298.65	204.61
Wright	556	3	325.99	177.05

Direct Estimates for the Mean Hectares of Corn per Segment	EBLUP for the Mean Hectares of Corn per Segment	Standard Error of Prediction	CV for the EBLUP	CV for the Direct Estimates
165.76	122.22	10.13	0.08	.
158.62	126.20	10.04	0.08	0.02
96.32	106.80	9.85	0.09	.
109.38	108.51	8.45	0.08	0.06
120.05	144.22	6.73	0.05	0.12
150.89	112.10	6.78	0.06	0.16
110.25	112.85	6.78	0.06	0.04
102.52	122.00	6.88	0.06	0.20
117.60	115.29	5.91	0.05	0.08
112.77	124.43	5.48	0.04	0.13
76.08	106.95	5.37	0.05	.
144.30	142.98	5.79	0.04	0.18

AREA-LEVEL SMALL AREA MODELS

Area-level models relate area-specific direct survey estimates to area-specific auxiliary data. For example, suppose you have a survey designed to estimate per capita income. Estimates of per capita income at the state level might be measured with adequate precision; but if you want estimates for munincipalities with populations less than 1,000 people, the sample sizes can be very small and the estimates can have large variances. To improve the precision of the estimates, you can use auxiliary data such as county-level values of per capita income, tax return data, and housing data to fit a linear mixed model to improve the efficiency of your estimates (Fay and Herriot 1979).

Suppose the population is divided into M mutually exclusive and exhaustive areas and that there are survey estimates available for m, $m \leq M$, of the areas. $\bar{y}_i$ is the survey estimate of the mean for area i, and $\bar{\boldsymbol{x}}_i = (\bar{x}_{i1}, ..., \bar{x}_{ip})^T$ is a known population mean vector of auxiliary variables for area i. A basic area-level model relates the $\bar{y}_i$ to the $\bar{\boldsymbol{x}}_i$ through a linear mixed model of the form

$$\bar{y}_i = \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta} + u_i + \bar{e}_i \tag{13}$$

where $\boldsymbol{\beta}$ is a fixed set of regression parameters, u_i are area-specific random effects, and $\bar{e}_i$ are the sampling errors. Assume that $u_i \overset{ind.}{\sim} (0, \sigma_u^2)$, $\bar{e}_i \overset{ind.}{\sim} (0, D_i)$ and that u_i and $\bar{e}_j$ are independent for all i and j.

The unknown mean for area i is

$$\theta_i = \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta} + u_i \tag{14}$$

If $\boldsymbol{\beta}$, σ_u^2, and D_i are known, the BLUP of u_i is

$$\hat{u}_i = \gamma_i (u_i + \bar{e}_i) \tag{15}$$

where

$$\gamma_i = \left(\sigma_u^2 + D_i\right)^{-1} \sigma_u^2 \tag{16}$$

The BLUP for θ_i is

$$\tilde{\theta}_i = \begin{cases} \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta} + \gamma_i \left(\bar{y}_i - \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta}\right) & \text{if } i \in A \\ \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta} & \text{if } i \notin A \end{cases} \tag{17}$$

where A is the index set for small areas in which $\bar{y}_i$ is observed.

For area-level small area models, the sampling variances D_i are typically estimated from the survey data (usually by pooling information across several related areas) or from other sources and then assumed to be known. Therefore, the variability of estimating D_i is often ignored for MSEP computation. When $\boldsymbol{\beta}$, and σ_u^2 are unknown, the EBLUP is given in equation (17) with the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_u^2$ replacing the unknown parameters. The EBLUP for the observed small areas can also be written as

$$\hat{\theta}_i = \hat{\gamma}_i \bar{y}_i + (1 - \hat{\gamma}_i) \bar{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}$$

With this representation, the EBLUP for the observed small areas are convex combinations of the direct estimators ($\bar{y}_i$) and the synthetic estimators ($\bar{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}$). For large areas where $\hat{\gamma}_i$ are close to 1, the EBLUP is close to the direct estimator; for small areas where $\hat{\gamma}_i$ are close to zero, the EBLUP is close to the synthetic estimator.

If $\hat{\sigma}_u^2$ is an unbaised estimator of σ_u^2, then an estimator of the MSEP is

$$\text{msep}(\hat{\theta}_i) = \begin{cases} \hat{\gamma}_i D_i + (1 - \hat{\gamma}_i)^2 \bar{\boldsymbol{x}}_i \hat{V}\{\hat{\boldsymbol{\beta}}\} \bar{\boldsymbol{x}}_i^T + 2\left(\hat{\sigma}_u^2 + D_i\right) \hat{V}\{\hat{\gamma}_i\} & \text{if } i \in A \\ \hat{\sigma}_u^2 + \bar{\boldsymbol{x}}_i \hat{V}\{\hat{\boldsymbol{\beta}}\} \bar{\boldsymbol{x}}_i^T & \text{if } i \notin A \end{cases} \tag{18}$$

where

$$\hat{V}\{\hat{\gamma}_i\} = \left(\hat{\sigma}_u^2 + D_i\right)^{-4} D_i^2 \bar{\mathrm{V}}\left(\hat{\sigma}_u^2\right)$$

and $\bar{\mathrm{V}}(\hat{\sigma}_u^2)$ is the asymptotic variance of $\hat{\sigma}_u^2$ (Prasad and Rao 1990).

You can use the MIXED procedure along with the GDATA= option to estimate the parameters of the small area model. Then, you compute the predicted values and the MSEP with the IML procedure, as shown in the next example.

Example: Predicting Wind Erosion for Counties in Iowa

This example considers the prediction of wind erosion in Iowa for the year 2002 (Fuller 2009, section 5.5). The data are a small subset from the U.S. National Resources Inventory with a few modifications to facilitate the example. Forty-four counties in Iowa report measures of wind erosion. The survey provides observations for all 44 counties, but an additional 4 counties with no observations are included in the example for the purposes of illustration. Each county is divided into segments, and the segments are the primary sampling units of the survey. The sample of segments

in a county are treated as a simple random sample. The soils of Iowa have been mapped, so population values for a number of soil characteristics are available. The mean of the soil erodibility index for each county is the auxiliary data in the small area model.

The survey results and auxiliary data are contained in the following SAS data set Erosion:

```
data erosion;
input County TotalSegments SampleSegments Erodibility y @@;
datalines;
  3 1387 13 -1.2317 0.429    15 2462 18 -0.0431 0.665   21 2265 14  0.6593 1.083
 27 2479 19 -1.1273 0.788    33 2318 18 -0.6198 0.869   35 1748 12  1.3130 1.125
 41 2186 16  0.0079 0.683    47 3048 19 -0.9243 0.408   59 1261 12 -0.5306 0.839
 63 1822 15  0.4563 0.754    67 1597 11 -1.4053 0.690   71 1345 15 -0.2193 0.927
 73 1795 12 -0.4818 0.945    75 2369 13 -1.8049 0.619   77 2562 15 -1.0395 0.475
 79 1899 11  1.5981 0.790    83 2486 16 -0.1545 0.647   85 2241 19  0.7700 0.727
 91 2066 15 -0.2882 1.120    93 1385 10  0.2830 0.677  109 2752 18  0.5255 0.968
119 1753 29  0.2605 0.703   129 1270 12 -0.0261 0.616  131 1232 10 -1.0261 0.422
133 2943 24  1.4121 1.045   135 1190 15 -1.3583 0.363  141 1567 11  2.2911 1.424
143 1511 10 -0.2771 0.975   145 1772 16 -1.8138 0.451  147 2716 17  0.1811 0.945
149 3877 16  2.1541 1.065   151 1823 10  0.4190 0.918  153 1580 18 -1.0497 0.670
155 4405 21  0.3348 0.619   157 2121 13 -1.4538 0.578  161 2423 16  0.7551 0.719
165 2327 12 -1.1504 0.376   167 3180 44  1.3262 0.954  169 1862 16 -0.4206 0.583
187 3011 15 -0.0580 0.874   189 1644 10  1.7335 1.256  193 2319 17  1.6142 0.905
195 1290 16 -1.2512 0.599   197 1754 11  1.2380 0.577  201 1822 15  0.4563  .
202 1511 10 -0.2771  .      203 3877 16  2.1541  .     204 3011 15 -0.0580  .
;
```

The data set includes the variables County, TotalSegments, SampleSegments, Erodibility, and Y. The variable County records an identification number for each county, TotalSegments records the total number of segments in the county, and SampleSegments records the sampled number of segments in the county. The variable Erodibility records a standardized population mean of the soil erodibility index for each county, and Y records the survey sample mean (direct estimate) of a variable that is related to wind erosion for each county. The first 44 observations contain the observed data, and the last 4 observations contain the hypothetical counties for which there were no observations in the survey. Consequently, the variable Y is set to missing in those 4 observations.

Unlike the unit-level model in the previous example, the area-level model attempts to model the relationship between area-level means and auxiliary data. However, without repeated measures for each area with which to estimate the within-area variability, the parameters σ_u^2 and D_i are not identified. Therefore, exogenous information is necessary in order to successfully model the data.

A preliminary analysis suggests that the assumption of a common population variance for the counties is reasonable (Fuller 2009). Therefore, assume that the variance of the mean wind erosion for county i is $n_i^{-1}\sigma_e^2 (= D_i)$, where n_i is the sampled number of segments in county i and σ_e^2 is the common variance with a value of 0.0971. This exogenous information makes it possible to identify and estimate the remaining parameters of the model. However, the estimation procedure is not as straightforward as it was for the unit-level model in the previous example. To understand why, you need a little knowledge of how PROC MIXED works.

Recall that a mixed model is of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ represents univariate data, $\boldsymbol{\beta}$ is an unknown vector of fixed effects with known model matrix $\mathbf{X}$, $\boldsymbol{\gamma}$ is an unknown vector of random effects with known model matrix $\mathbf{Z}$, and $\boldsymbol{\epsilon}$ is an unknown random error vector.

A key assumption is that $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are normally distributed with

$$\mathrm{E}\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

$$\mathrm{Var}\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

The variance of $\mathbf{y}$ is therefore $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. You can model $\mathbf{V}$ by setting up the random-effects design matrix $\mathbf{Z}$ and by specifying covariance structures for $\mathbf{G}$ and $\mathbf{R}$.

PROC MIXED constructs a mixed model according to the specifications in the MODEL, RANDOM, and REPEATED statements. The MODEL statement names a single dependent variable and the fixed effects, which determine the $\mathbf{X}$ matrix of the mixed model. The RANDOM statement defines the $\mathbf{Z}$ matrix of the mixed model, the random effects in the $\boldsymbol{\gamma}$ vector, and the structure of $\mathbf{G}$. The REPEATED statement specifies the $\mathbf{R}$ matrix in the mixed model. If no REPEATED statement is specified, $\mathbf{R}$ is assumed to be equal to $\sigma^2\mathbf{I}$.

The GDATA= option in the RANDOM statement provides you with complete control over the specification of the **G** matrix. Specification of the **R** matrix in the REPEATED statement is restricted to a limited number of covariance structures. However, this does not mean that you cannot use PROC MIXED to fit the area-level model. One remedy is to switch the roles of the **G** and **R** matrices. That is, you use the GDATA= option to specify the covariance structure of the residuals (the $n_i^{-1}\sigma_e^2$) and you do not include a REPEATED statement so that **R** is assumed to be equal to $\sigma_u^2\mathbf{I}$.

As a result of this subterfuge, the output from PROC MIXED is reversed, so that the covariance parameter estimate and the standard error that is reported for the residual are in fact the estimate and standard error for the random effect variance σ_u^2. Also, the EBLUPs and their standard errors depend on **G**, and because the roles of **G** and **R** have been switched, PROC MIXED computes the EBLUPs and their standard errors incorrectly. Fortunately, the EBLUPs and their standard errors can be easily computed using the ODS output data sets from PROC MIXED and a little programming using the IML procedure.

The following DATA step creates a sampling variance data set named G2 that is later provided to PROC MIXED using the GDATA= option in the RANDOM statement:

```
data g2;
    set erosion;
    row=_n_;
    col=_n_;
    value=0.0971/SampleSegments;
    keep row col value;
run;
```

The following SAS statements estimate the regression parameters and the covariance parameter for the area-level model. The METHOD=REML option in the MODEL statement specifies that the residual (restricted) maximum likelihood method be used to estimate the covariance parameters. The CLASS statement declares the variable County to be a class variable. The MODEL statement specifies Y as the dependent variable and Erod_Ind as the only independent variable in the model. The SOLUTION option produces a solution for the fixed-effects parameters, and the COVB option produces the approximate variance-covariance matrix of the fixed-effects parameter estimates $\hat{\boldsymbol{\beta}}$. The RANDOM statement defines the random effects, and the GDATA= option specifies that the **G** matrix be read from the SAS data set G2. The ODS OUTPUT statement specifies that the covariance matrix of fixed-effects parameter estimates, the fixed-effects solution vector, the estimated covariance parameters, and the asymptotic covariance matrix of covariance parameters be saved in the SAS data sets Covbeta, Beta, Sigma2, and Acovsigma2, respectively. These data sets are used later to compute the EBLUPs and their standard errors.

```
proc mixed data = erosion asycov method = reml;
   class county;
   model y = Erodibility / solution covb;
   random county / gdata = g2;
   ods output covb = covbeta
              solutionF = beta
              covparms = sigma2
              asycov = aCovSigma2;
run;
```

The parameter estimates are reported in Output 3.

Output 3 Parameter Estimates

```
                    The Mixed Procedure

                 Covariance Parameter
                       Estimates

                 Cov Parm     Estimate

                 Residual      0.02405

                 Asymptotic Covariance
                  Matrix of Estimates

              Row    Cov Parm        CovP1

                1    Residual     0.000046
```

Output 3 *continued*

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	0.7700	0.02642	42	29.14	<.0001
Erodibility	0.1554	0.02461	0	6.31	.

The parameter estimates reported in Output 3 are in agreement with those reported in Fuller (2009), with some numerical differences due to the fact that Fuller (2009) reports maximum likelihood estimates while the estimates reported in Output 3 are residual maximum likelihood (REML) estimates. Keep in mind that the 0.2405 that is reported as the residual variance is in fact the estimate of the variance of the random effects, σ_u^2.

With the estimates produced by PROC MIXED, you can now compute the EBLUPs and their standard errors. The first step in this process is to generate a variable in the Erosion data set to represent the sample variance for each county.

```
data erosion;
   set erosion;
   dsgvar   = 0.0971/SampleSegments;
run;
```

Now, use PROC IML to read into matrices the ODS output data sets that PROC MIXED created:

```
proc iml;
   use beta;       read all var {estimate}          into bet;
   use covbeta;    read all var _num_               into covb;
   use sigma2;     read all var {estimate}          into sigma2;
   use aCovSigma2; read all var {CovP1}             into acSigma2;
   use erosion;    read all var {y Erodibility dsgvar} into dat;

   nobs = nrow(dat);
   np = nrow(bet);
   y = dat[,1];
   one = J(nobs,1,1);
   XI = J(nobs,1,1) || dat[,2:np];
   d = dat[,np + 1];
   sigma2Vec = sigma2*one;
   covb = covb[,2:np+1];
   gamma = sigma2Vec/(sigma2Vec + d);
```

Next, the following statement computes the predicted means for the observed counties using equation (17):

```
   EBLUP = gamma#y + (1-gamma)#(XI*bet);
```

The following SAS statements compute the predicted means for the unobserved counties:

```
   do i = 1 to nobs;
     if y[i] = . then
        EBLUP[i] = XI[i,]*bet;
   end;
```

Next, the following statements compute the MSE of the predicted means for the observed counties using equation (18):

```
   g1i = gamma#d;

   XCovBXT = XI * covb * XI`;
   g2i = (one - gamma)##2 # vecdiag(diag(XCovBXT));

   avSigma2 = 1/sum( (sigma2Vec + d)##(-2) );
   avSigma2 = 2*avSigma2;
   g3i =( (d##2) # ( (d+sigma2Vec) ##(-3) ) ) * acSigma2;

   mse = g1i + g2i + 2*g3i;
```

The following statements compute the MSE for the unobserved counties:

```
   do i = 1 to nobs;
     if y[i] = . then
        mse[i] = XI[i,]* covb * XI[i,]` + sigma2;
   end;
```

The following statements create a SAS data set Outdata, which contains the small area predictions and their variances:

```
    create outData var{EBLUP mse} ;
    append;
    close outData;
quit;
```

The Outdata data set is now merged with the original data set Erosion, and the standard area of the prediction is computed and stored as the variable SE_EBLUP. Labels are also generated for the variables in preparation for printing the results.

```
data outData;
   merge outData erosion;
   SE_EBLUP  = sqrt(mse);
   CV_EBLUP  = SE_EBLUP/EBLUP;
   CV_Direct = sqrt(dsgvar)/y;
   keep County TotalSegments SampleSegments Erodibility y EBLUP
        SE_EBLUP CV_EBLUP CV_Direct;
run;
```

Finally, the data set is printed; Output 4 displays the results.

```
proc print data = outdata noobs;
   var county TotalSegments SampleSegments Erodibility y EBLUP
   SE_EBLUP CV_EBLUP CV_Direct;
   format Erodibility  8.3
          y            6.3
          EBLUP        6.3
          SE_EBLUP     6.3
          CV_EBLUP     6.2
          CV_Direct    6.2;
run;
```

Output 4 displays a part of the input data set with the prediction statistics. The last two columns are the model CV for the EBLUP and the design CV for the direct estimate. You can see that the model CV for the EBLUP is smaller than the design CV for the direct estimates for all counties except for counties 91 and 133 where they are the same. For counties with a large number of sampled segments, the EBLUP is close to the direct estimates.

Output 4 Predicted Small Area Means and Prediction Errors

County	Total Segments	Sample Segments	Erodibility	y	EBLUP	SE_EBLUP	CV_EBLUP	CV_Direct
3	1387	13	−1.232	0.429	0.464	0.077	0.17	0.20
15	2462	18	−0.043	0.665	0.683	0.067	0.10	0.11
21	2265	14	0.659	1.083	1.036	0.075	0.07	0.08
27	2479	19	−1.127	0.788	0.754	0.066	0.09	0.09
33	2318	18	−0.620	0.869	0.833	0.067	0.08	0.08
35	1748	12	1.313	1.125	1.087	0.080	0.07	0.08
41	2186	16	0.008	0.683	0.701	0.071	0.10	0.11
47	3048	19	−0.924	0.408	0.446	0.066	0.15	0.18
59	1261	12	−0.531	0.839	0.801	0.079	0.10	0.11
63	1822	15	0.456	0.754	0.772	0.073	0.09	0.11
67	1597	11	−1.405	0.690	0.653	0.082	0.13	0.14
71	1345	15	−0.219	0.927	0.886	0.073	0.08	0.09
73	1795	12	−0.482	0.945	0.882	0.079	0.09	0.10
75	2369	13	−1.805	0.619	0.588	0.078	0.13	0.14
77	2562	15	−1.040	0.475	0.503	0.073	0.14	0.17
79	1899	11	1.598	0.790	0.851	0.083	0.10	0.12
83	2486	16	−0.155	0.647	0.667	0.071	0.11	0.12
85	2241	19	0.770	0.727	0.756	0.066	0.09	0.10
91	2066	15	−0.288	1.120	1.036	0.073	0.07	0.07
93	1385	10	0.283	0.677	0.716	0.085	0.12	0.15
109	2752	18	0.526	0.968	0.947	0.067	0.07	0.08
119	1753	29	0.261	0.703	0.716	0.055	0.08	0.08
129	1270	12	−0.026	0.616	0.654	0.079	0.12	0.15
131	1232	10	−1.026	0.422	0.476	0.085	0.18	0.23
133	2943	24	1.412	1.045	1.037	0.060	0.06	0.06
135	1190	15	−1.358	0.363	0.405	0.073	0.18	0.22
141	1567	11	2.291	1.424	1.344	0.083	0.06	0.07
143	1511	10	−0.277	0.975	0.904	0.085	0.09	0.10
145	1772	16	−1.814	0.451	0.458	0.071	0.16	0.17
147	2716	17	0.181	0.945	0.917	0.069	0.08	0.08
149	3877	16	2.154	1.065	1.073	0.072	0.07	0.07
151	1823	10	0.419	0.918	0.894	0.085	0.09	0.11
153	1580	18	−1.050	0.670	0.658	0.068	0.10	0.11
155	4405	21	0.335	0.619	0.652	0.063	0.10	0.11
157	2121	13	−1.454	0.578	0.570	0.077	0.14	0.15
161	2423	16	0.755	0.719	0.753	0.071	0.09	0.11
165	2327	12	−1.150	0.376	0.430	0.080	0.18	0.24
167	3180	44	1.326	0.954	0.956	0.045	0.05	0.05
169	1862	16	−0.421	0.583	0.608	0.071	0.12	0.13
187	3011	15	−0.058	0.874	0.850	0.073	0.09	0.09
189	1644	10	1.734	1.256	1.194	0.086	0.07	0.08
193	2319	17	1.614	0.905	0.927	0.069	0.07	0.08
195	1290	16	−1.251	0.599	0.594	0.071	0.12	0.13
197	1754	11	1.238	0.577	0.680	0.082	0.12	0.16
201	1822	15	0.456	.	0.841	0.158	0.19	.
202	1511	10	−0.277	.	0.727	0.157	0.22	.
203	3877	16	2.154	.	1.105	0.166	0.15	.
204	3011	15	−0.058	.	0.761	0.157	0.21	.

UNMATCHED MODELS

You can use the techniques described in the previous two examples to estimate the means or totals for small areas. However, sometimes the small area parameter of interest is a nonlinear function of the small area totals y_i. For example, you might want to estimate small area rates or proportions such as census undercoverage rates, the proportion of a population below a certain poverty level, or the illiteracy rate in the population at smaller subdivisions such as counties or school districts. In such situations, the sampling model is

$$\bar{y}_{i.} = y_i + \epsilon_i$$

along with the linking model

$$\theta_i = g(y_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i$$

where $\epsilon_i \overset{ind.}{\sim} (0, D_i)$, $u_i \overset{ind.}{\sim} (0, \sigma_u^2)$, and $i = 1, 2, ..., m$. Here, the sampling model does not match the linking model. That is, you cannot naively combine the sampling model with the linking model to produce a linear mixed-effects model for small area estimation. See Rao (2003) for more information about unmatched small area models.

To fit an unmatched small area model using the hierarchical Bayes (HB) approach, you first specify a prior distribution $f(\boldsymbol{\beta}, \sigma_u^2)$ on the model parameters. You then apply Bayes' rule to derive the posterior distributions of the model parameters and the small area parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$. The posterior means of the small area parameters are the Bayes estimators for the small areas. See Gelman et al. (2004) for an introduction to Bayesian analysis.

Assuming normality, the small area model can be written as

$$\bar{y}_i|y_i, D_i \sim N(y_i, D_i),$$
$$\theta_i = g(y_i)|\boldsymbol{\beta}, \sigma_u^2, \bar{\boldsymbol{x}}_i \sim N(\bar{\boldsymbol{x}}_i^{\boldsymbol{T}}\boldsymbol{\beta}, \sigma_u^2),$$
$$f(\boldsymbol{\beta}) \propto 1,$$
$$\sigma_u^2 \sim IG(a, b),$$

where $IG(a, b)$ denotes the inverse gamma distribution with the shape parameter a and the scale parameter b. Note that D_i are assumed to be known as in the section "AREA-LEVEL SMALL AREA MODELS" on page 5.

Evaluating the posterior distribution often involves multidimensional integration. When the solution is analytically intractable, as is often the case, you can use Markov chain Monte Carlo (MCMC) methods. The MCMC method is a general simulation method for sampling from posterior distributions and computing posterior quantities of interest. The MCMC procedure is designed specifically for this purpose. Chen (2009) describes how the MCMC procedure is used for Bayesian modeling.

Example: Estimating Census Undercoverage

This example from Rao (2003, section 10.4) applies the HB approach to an unmatched sampling and linking model to estimate the undercoverage count M_i and the undercoverage rate U_i for each province in the 1991 Canadian census. After the census is taken, a follow-up survey is conducted in order to provide a direct estimate $(\hat{M}_i)$ of M_i, the number of persons missed by the census. The undercoverage rate is then calculated as $U_i = M_i/(M_i + C_i)$, where C_i is the actual census count.

Thus, the sampling model is

$$\hat{M}_i = M_i + \epsilon_i \tag{19}$$

where $\epsilon_i \sim N(0, D_i)$ and $i = 1, 2, \ldots, 10$ denote the 10 provinces in Canada. The sampling variances D_i are estimated through a generalized variance function model of the form $V(\hat{M}_i) \propto C_i^{\gamma}$ and are treated as known in the sampling model (19).

The linking model is

$$\log(U_i) = \log\left(M_i/(M_i + C_i)\right) = \beta_0 + \beta_1 \log(C_i) + u_i \tag{20}$$

for $i = 1, \ldots, 10$. One specification of the model using the HB framework is

$$\hat{M}_i|M_i, D_i \overset{ind.}{\sim} N(M_i, D_i) \tag{21}$$

$$\log(U_i)|\beta_0, \beta_1, \sigma_u^2, C_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 \log(C_i), \sigma_u^2) \tag{22}$$

$$f(\beta_0, \beta_1, \sigma_u^2) = f(\boldsymbol{\beta}) f(\sigma_u^2) \propto \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma_u^2}\right)^{a+1} e^{(-b/\sigma_u^2)} \tag{23}$$

where $i = 1, 2, \ldots, 10$.

Equation (23) specifies the prior distributions of the model parameters. Specifically, β_0 and β_1 are specified as having "flat" priors such that $f(\boldsymbol{\beta}) \propto 1$ to reflect a lack of prior information regarding these parameters. The prior distribution for the parameter σ_u^2 is specified as an inverse-gamma with shape and scale parameters a and b, respectively. The shape and scale parameters are typically set to be very small.

The following SAS statements use the MCMC procedure to estimate the model parameters and the small area undercoverage counts and rates. The input data set is named Undercoverage, and it contains the variables Index, Province, CensusCount, Missing, and D. The variable Missing contains the direct estimates of the undercoverage count M_i, and the variable D contains the known variances. The data set is similar to Rao (2003, example 10.2.2).

```
data undercoverage;
input Index Province $ CensusCount Missing D;
datalines;
    1  Nfld   569640  11566   3424572.3136
    2  PEI    129963   1220    133956
    3  NS     899549  17329  12011769.64
    4  NB     722797  24280  11554560.64
    5  Que   6965643 184473 217793841.47
    6  Ont  10088786 381104 929537656.42
    7  Man   1091728  20691  18879980.912
    8  Sask   987783  18106  11834563.22
    9  Alta  2526533  51825  60431189.063
   10 BC    3286372  92236  85074796.96
;
```

The following SAS statements specify the model in PROC MCMC. The NMC= option in the PROC MCMC statement specifies the number of MCMC iterations, excluding the burn-in iterations. The NTHIN= option controls the thinning rate of the simulation, and the NBI= option specifies the number of burn-in iterations. The OUTPOST= option names the output data set for posterior samples of parameters. The MONITOR= option directs PROC MCMC to output analysis for the specified symbols of interest.

```
proc mcmc data=undercoverage nmc=45000 nthin=10 nbi=5000 seed=123456
       outpost=o1 monitor=(_parms_ m u)
       stats=(summary interval) diag=none;
   array m[10];
   array u[10];
   parm (beta0 beta1) 1;
   parm s2;
   prior beta: ~ general(0);
   prior s2 ~ igamma(shape=0.01, scale=0.01);
   random gamma ~ n(beta0 + beta1*log(censuscount), var=s2) subject=province;
   m[index] = censuscount*exp(-gamma)/(1-exp(-gamma));
   u[index] = exp(-gamma);
   model missing ~ n(m[index], var=d);
   ods output postsummaries=est;
run;
```

The two ARRAY statements specify that the arrays M and U be constructed. The arrays define the undercoverage count (M_i) and the undercoverage rate (U_i) for each province.

The next two PARMS statements specify the parameters of the model. The first PARMS statement specifies that the two regression coefficients be named Beta0 and Beta1 and that both have initial values equal to 1. The second PARMS statement specifies that the random effects variance parameter be named S2.

For each parameter, you must specify a prior distribution. The first PRIOR statement specifies a general(0) distribution for the regression coefficients, which implements the notion of a "flat" prior. The prior for the random effects variance parameter S2 is specified as an inverse-gamma with shape and scale parameters equal to 0.01.

The RANDOM statement defines a random effect and its prior distribution. The SUBJECT= option identifies the subjects in the random effects model. The random effects parameters associated with each subject are assumed to be conditionally independent of each other given other parameters in the model. In this case, the random effect is named Gamma, and it is defined to have a normal distribution with a mean equal to $\beta_0 + \beta_1 \log(C_i)$ and a variance of S2.

NOTE: The RANDOM statement in PROC MCMC is available only in SAS/STAT 9.3 and later. You can fit the model using earlier releases, but more programming is required. For an example of how to do this in SAS/STAT 9.2, see "Example 52.5 Random Effects Models" in the *SAS/STAT 9.2 User's Guide*.

The next two statements simply define the equations for M_i and U_i as derived from the linking equation (20). The results of these computations are stored in the previously declared arrays M and U.

Next, the MODEL statement specifies the complete small area model, which now encompasses both the sampling model for M_i and the linking model for $\log(U_i)$.

Finally, the ODS OUTPUT statement directs the procedure to create a data set named Est to store the basic statistics for each parameter; these statistics include the posterior summaries (namely, the sample size, mean, standard deviation, and percentiles).

Output 5 displays the MCMC results. The "Posterior Summaries" table displays the number of posterior samples, the posterior mean and standard deviation estimates, and the percentile estimates. The "Posterior Intervals" table displays

the equal-tail and highest posterior density (HPD) interval (Gelman et al. 2004) estimates for each parameter.

Output 5 PROC MCMC Results

The MCMC Procedure

Number of Observations Read 10
Number of Observations Used 10

Parameters

Block	Parameter	Sampling Method	Initial Value	Prior Distribution
1	beta0	N-Metropolis	1.0000	general(0)
	beta1		1.0000	general(0)
2	s2	Conjugate	0.00990	igamma(shape=0.01, scale=0.01)

Random Effects Parameters

Parameter	Subject	Levels	Prior Distribution
gamma	Province	10	normal(beta0 + beta1*log(censuscount), var=s2)

The MCMC Procedure

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
beta0	4500	7.0153	1.0397	6.3606	6.9717	7.6065
beta1	4500	-0.2227	0.0721	-0.2635	-0.2194	-0.1777
s2	4500	0.0531	0.0534	0.0207	0.0374	0.0681
m1	4500	10784.6	1535.7	9726.6	10745.3	11790.9
m2	4500	1467.0	289.2	1270.9	1467.3	1654.5
m3	4500	17241.5	2566.5	15505.5	17168.5	18889.2
m4	4500	18707.9	3557.0	15984.2	18514.7	21118.8
m5	4500	188535	14029.8	179288	188683	197803
m6	4500	370690	29578.6	350916	369700	390504
m7	4500	21257.0	3170.6	19144.3	21258.8	23371.6
m8	4500	18677.8	2621.6	16886.1	18638.7	20423.5
m9	4500	54963.0	6555.2	50846.9	55037.8	59323.5
m10	4500	89967.3	8286.4	84317.7	89620.7	95426.9
u1	4500	0.0186	0.00260	0.0168	0.0185	0.0203
u2	4500	0.0112	0.00218	0.00968	0.0112	0.0126
u3	4500	0.0188	0.00275	0.0169	0.0187	0.0206
u4	4500	0.0252	0.00467	0.0216	0.0250	0.0284
u5	4500	0.0263	0.00191	0.0251	0.0264	0.0276
u6	4500	0.0354	0.00273	0.0336	0.0353	0.0373
u7	4500	0.0191	0.00279	0.0172	0.0191	0.0210
u8	4500	0.0186	0.00256	0.0168	0.0185	0.0203
u9	4500	0.0213	0.00249	0.0197	0.0213	0.0229
u10	4500	0.0266	0.00239	0.0250	0.0265	0.0282

Posterior Intervals

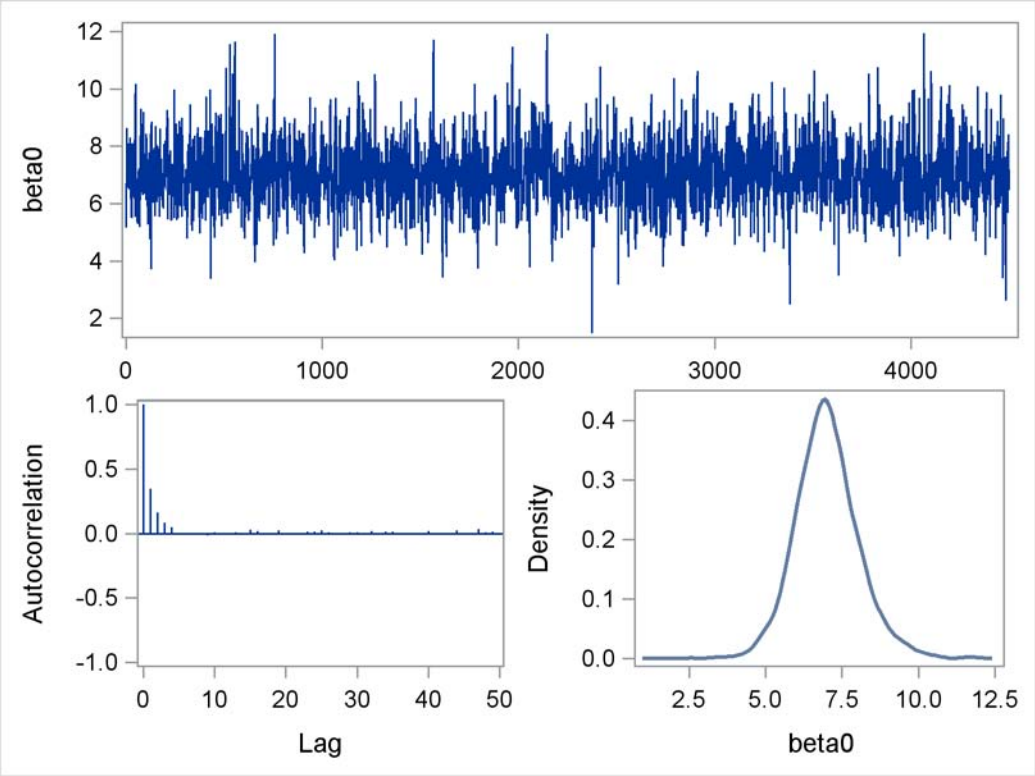
Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.050	5.0450	9.2706	4.9166	9.0390
beta1	0.050	-0.3799	-0.0870	-0.3605	-0.0745
s2	0.050	0.00627	0.1819	0.00306	0.1474
m1	0.050	7887.8	13938.5	7763.8	13769.9
m2	0.050	896.2	2043.6	920.1	2062.8
m3	0.050	12423.7	22436.8	12376.5	22309.4
m4	0.050	12676.5	26194.8	12355.2	25693.3
m5	0.050	161318	216311	161824	216623
m6	0.050	314273	431855	309297	425381
m7	0.050	15004.2	27582.3	14982.6	27512.6
m8	0.050	13665.6	23823.8	13741.7	23877.0
m9	0.050	41907.8	67791.9	41180.5	66942.1
m10	0.050	73968.4	106327	73516.8	105669
u1	0.050	0.0137	0.0239	0.0134	0.0236
u2	0.050	0.00685	0.0155	0.00703	0.0156
u3	0.050	0.0136	0.0243	0.0136	0.0242
u4	0.050	0.0172	0.0350	0.0168	0.0343
u5	0.050	0.0226	0.0301	0.0227	0.0302
u6	0.050	0.0302	0.0410	0.0297	0.0405
u7	0.050	0.0136	0.0246	0.0135	0.0246
u8	0.050	0.0136	0.0236	0.0137	0.0236
u9	0.050	0.0163	0.0261	0.0160	0.0258
u10	0.050	0.0220	0.0313	0.0219	0.0312

You might notice in the "Posterior Summaries" table that the algebraic signs of the regression coefficients β_0 and β_1 are opposite of what you might expect. This is due to the parameterizations of Gamma in the preceding statements. The rationale for reversing the sign on Gamma in the equations for M_i and U_i is that experimentation with this model indicates that doing so provides a better range for the estimated quantities of interest in the intermediate computations.

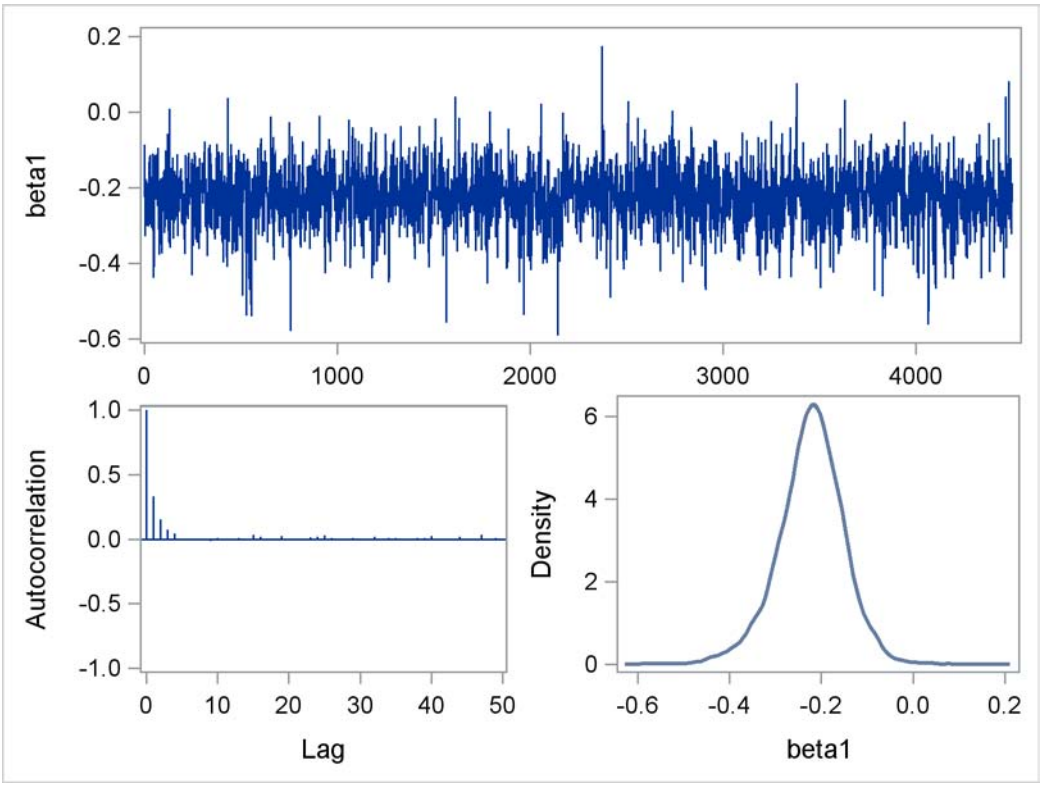
PROC MCMC automatically generates the trace, autocorrelation, and kernel density plots that are shown in Output 6. A trace plot provides you with evidence of whether the Markov chain has converged to its stationary distribution. A trace can also tell you whether the chain is mixing well. A chain might have reached stationarity if the distribution of points is not changing as the chain progresses. The aspects of stationarity that are most recognizable from a trace plot are a relatively constant mean and variance. The trace plots in Output 6 indicate that the Markov chain has stabilized and appears constant for all three variables. The trace plots also appear to indicate that the Markov chains have good mixing. A chain that mixes well traverses its posterior space rapidly, and it can jump from one remote region of the posterior to another in relatively few steps.

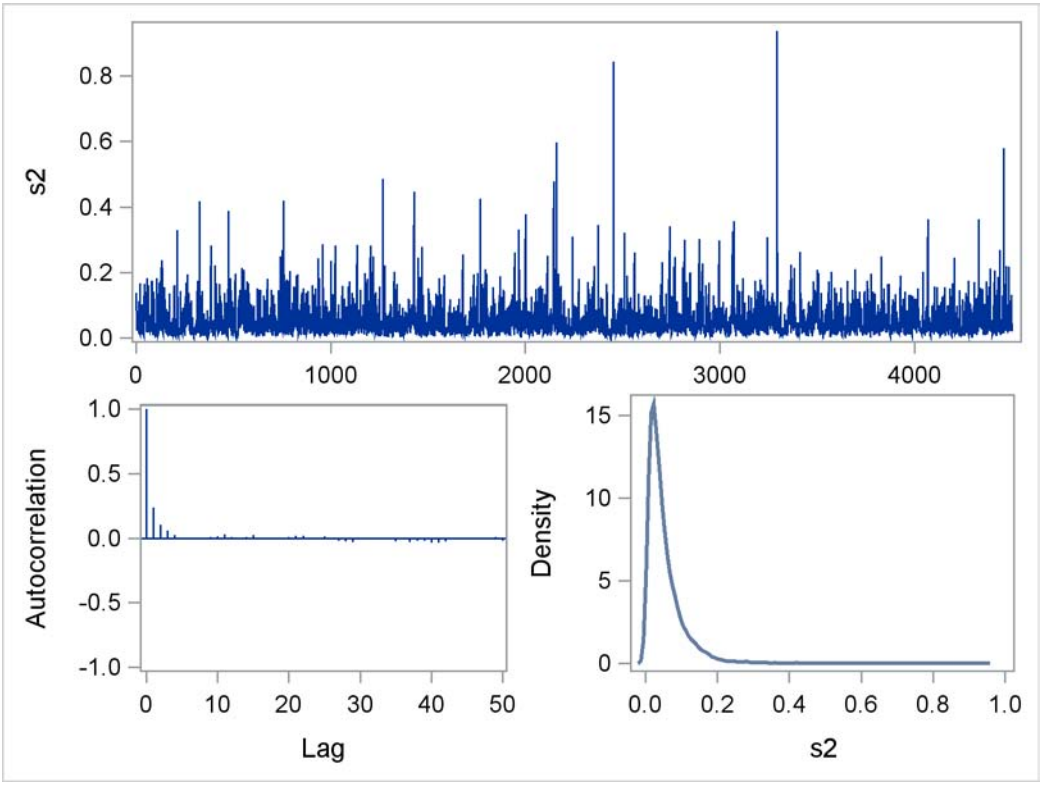
The autocorrelation plots do not indicate any significant autocorrelations for all three small area parameters. The NTHIN= option in the PROC MCMC statement controls the thinning, which might control the autocorrelations among the posterior samples. NTHIN = 10 is sufficient in this example.

Output 6 MCMC Diagnostic Plots



Output 6 *continued*





The following SAS statements use the output data set Est to generate two tables of HB estimates of the posterior means of the undercoverage count M and undercoverage rate U, and coefficients of variation for the estimates.

```
data estcount;
   merge undercoverage est (firstobs=4 obs=13);
   CVHB = StdDev/Mean;
```

```
   CVD  = sqrt(D)/Missing;
   label Missing = 'Direct Estimate for Undercount'
         Mean    = 'HB Estimate for Undercount'
         StdDev  = 'Standard Deviation for the HB Estimator'
         CVHB    = 'CV for the HB Estimator'
         CVD     = 'CV for the Direct Estimator';
run;
data estrate;
   merge undercoverage est (firstobs=14);
   CVHB = StdDev/Mean;
   UR   = Missing/(Missing+CensusCount);
   label UR = 'Direct Estimate for Undercoverage Rate'
         Mean    = 'HB Estimate for Undercoverage Rate'
         StdDev  = 'Standard Deviation for the HB Estimator'
         CVHB    = 'CV for the HB Estimator';
run;
proc print data=estcount label noobs;
   var Province CensusCount Missing Mean StdDev CVHB CVD;
   format Mean 10.1
          CVHB  4.2
          CVD   4.2;
run;
proc print data=estrate label noobs;
   var Province CensusCount UR Mean StdDev CVHB;
   format UR     6.3
          Mean   6.3
          StdDev 6.3
          CVHB   6.3;
run;
```

The results are displayed in Output 7 and Output 8. Output 7 displays the census counts, direct estimates for undercount, HB estimates for undercount, estimated standard deviations for the HB estimates, model CV for the HB estimates, and the design CV for the direct estimates.

Output 7 HB Estimates for the Undercounts

Province	Census Count	Direct Estimate for Undercount	HB Estimate for Undercount	Standard Deviation for the HB Estimator	CV for the HB Estimator	CV for the Direct Estimator
Nfld	569640	11566	10784.6	1535.7	0.14	0.16
PEI	129963	1220	1467.0	289.2	0.20	0.30
NS	899549	17329	17241.5	2566.5	0.15	0.20
NB	722797	24280	18707.9	3557.0	0.19	0.14
Que	6965643	184473	188534.8	14029.8	0.07	0.08
Ont	10088786	381104	370689.9	29578.6	0.08	0.08
Man	1091728	20691	21257.0	3170.6	0.15	0.21
Sask	987783	18106	18677.8	2621.6	0.14	0.19
Alta	2526533	51825	54963.0	6555.2	0.12	0.15
BC	3286372	92236	89967.3	8286.4	0.09	0.10

From the table in Output 7 you can see that the model CV for the HB estimate is lower than the design CV for the direct estimate for every province in Canada except for New Brunswick (NB). For provinces with large sample sizes such as Ontario (Ont) or Quebec (Que), the CV for the direct estimates and the HB estimates are similar.

Output 8 represents the prediction statistics for the undercoverage rate. The model CV for the HB estimates for the undercoverage rate range from 7.2% to 19.5%.

Output 8 HB Estimates for the Undercoverage Rates

Province	Census Count	Direct Estimate for Undercoverage Rate	HB Estimate for Undercoverage Rate	Standard Deviation for the HB Estimator	CV for the HB Estimator
Nfld	569640	0.020	0.019	0.003	0.140
PEI	129963	0.009	0.011	0.002	0.195
NS	899549	0.019	0.019	0.003	0.146
NB	722797	0.032	0.025	0.005	0.185
Que	6965643	0.026	0.026	0.002	0.072
Ont	10088786	0.036	0.035	0.003	0.077
Man	1091728	0.019	0.019	0.003	0.146
Sask	987783	0.018	0.019	0.003	0.138
Alta	2526533	0.020	0.021	0.002	0.117
BC	3286372	0.027	0.027	0.002	0.090

CONCLUSION

Small area estimation techniques are useful for subpopulation (domain) analysis when direct domain estimators do not have adequate precision due to small sample sizes. Indirect estimation for small areas uses statistical models and auxiliary variables to borrow strength from similar areas. This paper describes three approaches for indirect estimation of small area parameters using the three most commonly used small area models. The MIXED, IML, and MCMC procedures are used to predict the small area parameters and their prediction errors. SAS/STAT users can use the techniques described in this paper to compute indirect estimators for small area statistics.

ACKNOWLEDGMENTS

The authors are grateful to Fang Chen, Tony An, Donna Watts, Randy Tobias, Tianlin Wang, Min Zhu, Maura Stokes, Tim Arnold, and Anne Baxter of the Statistical R&D Division at SAS Institute Inc. for their contributions to the preparation of this manuscript.

REFERENCES

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28–36.

Chen, F. (2009), "Bayesian Modeling Using the MCMC Procedure," in *Proceedings of the SAS Global Forum 2009 Conference*, Cary, NC: SAS Institute Inc.

Fay, R. E. and Herriot, R. A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the Americal Statistical Association*, 74, 269–277.

Fuller, W. A. (2009), *Sampling Statistics*, Hoboken, New Jersey: John Wiley & Sons.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Second Edition, London: Chapman & Hall.

Kackar, R. N. and Harville, D. A. (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association*, 79, 853–862.

Kenward, M. G. and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*, 53, 983–997.

Prasad, N. G. N. and Rao, J. N. K. (1990), "The Estimation of Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85, 163–171.

Rao, J. N. K. (2003), *Small Area Estimation*, Hoboken, New Jersey: John Wiley & Sons.

Rao, J. N. K. and Choudry, G. H. (1995), "Small Area Estimation: Overview and Empirical Study," in B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, eds., *Business Survey Methods*, 527–542, New York: John Wiley & Sons.

CONTACT INFORMATION

Pushpal K Mukhopadhyay
SAS Institute Inc.
SAS Campus Drive
Cary, NC, 27513
919-531-2123
pushpal.mukhopadhyay@sas.com

Allen Mcdowell
SAS Institute Inc.
SAS Campus Drive
Cary, NC, 27513
919-531-6837
allen.mcdowell@sas.com