CURRENT POPULATION SURVEY SMALL AREA ESTIMATION FOR CONGRESSIONAL DISTRICTS

Richard Griffiths
U.S. Bureau of the Census, Suitland, MD 20233

Key Words: Small area, SPREE, Regression, Composite

I. What is the Problem?

This paper reports preliminary results of the Census Bureau's investigation of small area estimation methodology for Congressional Districts (CDs) of the 103rd Congress based on Current Population Survey (CPS) data.

The CPS is a monthly national sample survey of the population of the United States. Its sample is a two-stage stratified probability sample drawn independently within each of the 50 states and the District of Columbia. It is designed to provide accurate estimation of labor force characteristics at the national and state levels. The survey also gathers data on the employment and earnings status of the population.

The CPS sample was not designed, however, to provide for estimation at the CD level. The achieved sample sizes for CDs tend to be too small to allow precise sample-based estimation.

In this paper we discuss research into determining a methodology for producing CD-level estimators which are more precise than the direct sample-based estimators. This research falls into the category of small area estimation. There are a number of articles which give overviews of different types of small area estimation, for instance, Kalton (1990), Ghosh and Rao (1994), and Purcell and Kish (1979).

The approaches to CD-level estimation investigated in this paper combine model-based small area estimation with CPS sample-based estimation. The resulting estimators are composite estimators.

II. The Composite Estimator

Given in univariate form, the composite small area estimator of characteristic Y for CD h is

$$\hat{Y}_h = w_h \hat{M}_h + (1 - w_h) y_h \qquad \text{(II-1)}$$

where

w_h is the weight for the model-based small area estimator; $0 \leq w_h \leq 1$;

$\hat{M}_h$ is the model-based estimator of characteristic Y for CD h; and

y_h is the sample-based estimator of characteristic Y for CD h.

In (II-1), we have a composite estimator which includes both a model-based and a sample-based estimator. We plan to compare two model-based methods for CD-level estimation: the Structure Preserving Estimation (SPREE) of Purcell (1979) and a regression-based method.

Why Consider Two (as Opposed to One) Methods?

Regression methods, in particular empirical Bayesian methods, appear to be the currently-preferred method of small area estimation in many settings, assuming sufficient auxiliary information is available. See Ghosh and Rao (1994) or Fay (1986,1987,1988). So, we want to consider such a model.

But also appealing is the SPREE estimation of Purcell (1979). SPREE is based on preserving data relationships which existed at the small area level in a previous time period. Assuming these relationships hold over time, they should serve as a good model for the current time period. Since we have 1990 census data available at the CD level for some estimands, and thus a fairly recent structure at the small area level, we'd like to see how this method performs.

This paper, then, examines the performance of two composite estimators: one based on the SPREE method and one based on a regression method. We discuss the sample-based and model-based portions of the composite estimators in the following sections. The paper also contains sections detailing preliminary results obtained using the two composite estimators for basic March 1994 CPS data for the state of Iowa.

III. The Sample-based Estimator

The sample-based part of estimator (II-1) is a direct estimator of the CD-level characteristics. It uses data collected for the March 1994 CPS.

The sample-based estimator of characteristic Y has the form

$$y_h = \sum_{i=1}^{n_h} SW_h y_{hi} \qquad \text{(III-1)}$$

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

where

n_h is the number of sample units in CD h;

SW_h is the sample weight applied to each unit in order to weight the sample values to the CD h level; and y_{hi} is the value of characteristic Y for unit i in CD h.

For the calculation of the sample-based estimator used to produce the results given in this paper, we assumed the units were from a simple random sample within each CD and that the sample was proportionately distributed to each CD. Under this assumption, SW_h in (III-1) equals the inverse of the probability of selection for each unit multiplied by an adjustment for nonresponse.

This estimator does not reflect the stratified, cluster design of the CPS. In future research we hope to provide a sample-based estimator which better reflects the sample design.

IV. The SPREE-based Composite Estimator

The SPREE portion of the composite estimator follows the work of Purcell (1979). A description of the SPREE method can be found in Purcell (1979) and Purcell and Kish (1980).

What is SPREE?

The SPREE method is a categorical data analysis approach to the problem of small area estimation, being applicable to the estimation of frequencies. It makes two assumptions concerning the availability of data. The first is that there exist current estimates for the variables of interest by subgroups for the large area; the second is that estimates of variables of interest are available by the same subgroups at the small area level from some previous time period.

The SPREE method of estimation uses data from a previous time period to allocate current data at the large area level to the small areas. The data from the previous time period are known at the small area level and known for cross-classifications (subgroups) of some auxiliary variables. For our purposes, this previous time period is the 1990 Decennial Census from which estimates are available at the CD level.

The data from the previous time period are known as the association structure. The data from the current time period at the large area level are known as the allocation structure. The SPREE method allocates the current data (in the allocation structure) to the small area level by retaining the relationship of the data given in the association structure.

The SPREE estimator of characteristic Y total for CD h is

$$\hat{M}_{hy.} = \sum_{g=1}^{G} \hat{M}_{hyg}\ , \text{ with } \hat{M}_{hyg} = \frac{N_{hyg}}{N_{.yg}} m_{.yg} \qquad \text{(IV-1)}$$

where

N_{hyg} is the total number of persons counted in the 1990 Decennial Census in CD h with the y^{th} level of characteristic Y (for example, unemployed and employed could be the two levels of an employment characteristic) in subgroup g;

$N_{.yg}$ is the sum of N_{hyg} over the CDs; and

$m_{.yg}$ is the estimated number of persons with the y^{th} level of characteristic Y in subgroup g at the state level from the 1994 March CPS data.

In (II-1) the weights w_h determine how much the composite estimator depends on the sample-based estimator and how much it depends on the model-based estimator.

Weights

For the SPREE-based composite estimator the weights would be calculated as follows:

$$w_h = \frac{mse(y_h) - E(y_h - Y_h)(\hat{M}_h - Y_h)}{mse(y_h) + mse(\hat{M}_h) - 2E(y_h - Y_h)(\hat{M}_h - Y_h)}$$

as given in Drew, Singh, Choudhry (1982), where Y_h is the true value of characteristic Y in CD h. If $\hat{M}_h$ is unbiased for Y_h, then $E(y_h - Y_h)(\hat{M}_h - Y_h)$ is equal to the covariance between $\hat{M}_h$ and y_h.

If $E(y_h - Y_h)(\hat{M}_h - Y_h)$ is negligible relative to $mse(y_h)$ and $mse(\hat{M}_h)$, we may use as the weight

$$w_h = \frac{mse(y_h)}{mse(y_h) + mse(\hat{M}_h)} \qquad \text{(IV-2)}$$

In order to determine the appropriate weights for (II-1), we need to calculate the mean square error (MSE) of the SPREE estimator.

Calculation of the Mean Square Error

The MSE of the SPREE estimator includes both a bias and a variance term.

From the allocation structure, there is sampling variability, since the allocation structure is based on the March 1994 CPS sample. If the assumption that the association structure relationships hold for the 1994 data fails, there is a bias. Estimates of the sampling variability and bias taken together provide the estimates of the MSE.

Estimation of the Variance

Assuming the sampling error of the N_{hyg} from the census data is negligible, the only random variable in (IV-1) is $m_{.yg}$. The variance of $\hat{M}_{hyg}$ is then

$$V(\hat{M}_{hyg})=\left(\frac{N_{hyg}}{N_{.yg}}\right)^2 V(m_{.yg})$$

where $V(m_{.yg})=\sum_{h=1}^{H} V(m_{hyg})$, under the assumption of independent sampling within CDs.

Estimation of the Bias

The bias of $\hat{M}_{hy.}$ is given by

$$B_{hy.}=E(\hat{M}_{hy.})-M_{hy.}$$

Assuming the sample-based CD estimators are unbiased (though they may have large sampling variances), let $m_{hy.}$ be the unbiased sample-based estimator for the y^th level of characteristic Y for CD. Then, following work given in Purcell (1979, pp. 171-173), a squared estimate of this bias can be shown to be

$$b^2_{hy.}=(\hat{M}_{hy.}-m_{hy.})^2-\hat{V}(\hat{M}_{hy.})-\hat{V}(m_{hy.})+2\hat{Cov}(\hat{M}_{hy.},m_{hy.})$$

The Mean Square Error

The mean square error of the SPREE estimator is then calculated as follows:

$$M\hat{S}E(\hat{M}_{hy.})=\hat{b}^2_{hy.}+\hat{V}(\hat{M}_{hy.})$$

$$=(\hat{M}_{hy.}-m_{hy.})^2-\hat{V}(\hat{M}_{hy.})-\hat{V}(m_{hy.})+2\hat{Cov}(\hat{M}_{hy.},m_{hy.})+\hat{V}(\hat{M}_{hy.})$$

$$=(\hat{M}_{hy.}-m_{hy.})^2-\hat{V}(m_{hy.})+2\hat{Cov}(\hat{M}_{hy.},m_{hy.})$$

Similarly, the mean square error of the SPREE-based composite estimator is found to be

$$M\hat{S}E(\hat{Y}_h)=(\hat{Y}_h-m_{hy})^2-\hat{V}(m_{hy})+2\hat{Cov}(\hat{Y}_h,m_{hy})$$

V. The Regression-based Composite Estimator

The regression-based composite estimator is based on a components-of-variance model of CD-level estimates. In this sense, it follows work described in Fay (1986, 1987, 1988).

Basis

The components-of-variance model on which the regression-based composite estimator is founded may be expressed as

$$\mathbf{y}=\mathbf{Xb}+\mathbf{a}+\mathbf{d}$$

where

y is the vector of sample-based estimates;
a is a vector of random CD effects;
d is the vector of random sampling errors;
X is a matrix of auxiliary information; and
b is the regression coefficient vector.

In Fay (1986, 1988), **a** has a multivariate normal distribution with mean **0** and variance-covariance matrix **A**; **d** is also distributed multivariate normally with mean **0** and variance-covariance matrix **D**. **a** and **d** are assumed independent. We make these same assumptions in this paper.

Form

The regression-based composite estimator may be expressed as

$$\hat{\mathbf{y}}=(\,\mathbf{I}-\mathbf{A}(\mathbf{D}+\mathbf{A})^{-1})\mathbf{X}\hat{\mathbf{b}}+\mathbf{A}(\mathbf{D}+\mathbf{A})^{-1}\mathbf{y} \qquad \text{(V-1)}$$

where

$\hat{\mathbf{y}}$ is the vector of estimated CD-level characteristics;
I is the identity matrix; and

$\hat{\mathbf{b}}$ is an estimator of the regression coefficient vector.

See also Morris (1983) for a description of (V-1).

The best linear unbiased estimator of **b** in (V-1) is, under the assumption that **A** and **D** are known,

$$\hat{\mathbf{b}}=(\mathbf{X}'(\mathbf{A}+\mathbf{D})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{A}+\mathbf{D})^{-1}\mathbf{Y} \qquad \text{(V-2)}$$

A and **D** are not known, however, and we must estimate them in order to evaluate (V-2).

We estimate **D** with design-based estimates of the sampling errors. For the results given in this paper, we estimate **A** using a quadratic moment estimator as given in Prasad and Rao (1990).

MSE

For the results given in this paper, we also calculate estimated MSEs of the regression-based estimators. From Ghosh and Rao (1994) and Prasad and Rao (1990), we get a formula for these estimated MSEs:

$$M\hat{S}E(\hat{y}_{hi})=M(A^*_{hi})+2D^2_{hi}(A^*_{hi}+D_{hi})^{-3}v^{a}(A^*_{hi})$$

where

$\hat{y}_{hi}$ is the regression-based composite estimator of the i^th estimand in CD h;

A^*_{hi} is the estimated variance of the random CD effect for estimand i in CD h;

D_{hi} is the estimated sampling variance of the sample-based estimator of estimand i in CD h;

$$M(A^*_{hi})=\left(\frac{A^*_{hi}}{A^*_{hi}+D_{hi}}\right)D_{hi}+\left(1-\frac{A^*_{hi}}{A^*_{hi}+D_{hi}}\right)^2\mathbf{X}_{hi}(\mathbf{X}'(\mathbf{A}+\mathbf{D})^{-1}\mathbf{X})\ ;$$

$\mathbf{X}_{hi}$ is the row of **X** corresponding to estimand i in CD h; and

$v^{a}(A^{*}_{hi})=\frac{2}{H^{2}}\sum_{h=1}^{H}(D_{hi}+A^{*}_{hi})^{2}$ and H is the number of CDs.

VI. An Investigation of the SPREE estimator

Using the basic March 1994 CPS data, we calculated SPREE-based composite estimates for two characteristics for the state of Iowa.

The levels of the characteristics for which estimates were obtained are categories of employment and household income. Employment categories are

- employed,
- unemployed, and
- other, which includes persons under 16 years of age and persons in the armed forces.

Household income categories are

- households with total income of less than $10,000;
- households with total income between $10,000 and $25,000;
- households with total income between $25,000 and $50,000;
- households with total income between $50,000 and $75,000; and
- households with total income of more than $75,000.

The state of Iowa was chosen since it represents a simple yet nontrivial example of the problem posed by CD-level estimation: each of its counties is wholly located in one of its five CDs. This represents a much simpler case than that of many states in which county boundaries cross CD boundaries.

The SPREE-based composite estimator was used to estimate the number of persons in each CD in Iowa falling into each of these categories.

Results

Using the techniques for the SPREE-based composite estimation described above in the paper, we obtained the results given in Table 1.

This table gives the ratios of the MSE of the SPREE estimator to the sample-based estimator and of the MSE of the SPREE-based composite estimator to the sample-based estimator and the reduction in MSE derived by using the SPREE-based composite estimator over the sample-based estimator. The MSEs given in this table have been averaged over the five CDs.

From this table we can see that the composite estimator offers an improvement over the sample-based estimator. The reduction in MSE, when MSEs are averaged over all CDs, ranges from 17.1% to 78.5%.

VII. An Investigation of the Regression-based Composite Estimator

As we did for the SPREE-based composite estimator, we tested the regression-based composite estimator for the eight estimands using March 1994 CPS data for the state of Iowa.

For each of the eight estimands the model included two auxiliary variables. The auxiliary variables were total number of persons in the CD and the corresponding estimate of the characteristic from the 1990 census for the employment characteristics. For the household income characteristics, the auxiliary variables were the total number of housing units in the CD and the corresponding estimate of the characteristic from the 1990 census.

Results

Some results for the regression-based estimator are included in Table 2. In this table we have the results for one CD: the sample-based, regression, and regression-based estimates for each estimand and the MSEs of the regression-based composite estimators.

In this table we make no comparison of the MSEs of the regression-based composite estimator to the sample-based estimator and the SPREE-based composite estimator. The reason for this is that the estimated MSEs for the sample-based estimators and the SPREE-based composite estimators are calculated using a jackknife methodology and are thus sample-based MSE estimators. The estimated MSEs for the regression-based composite estimators are calculated under the assumptions of the components-of-variance model and are thus model-based MSE estimators. So, the sets of MSEs are not comparable.

Without this comparison, then, we can talk only about what needs to be done in the near future.

VIII. What Else is There to do?

At this point in our research, we have no conclusions to offer. We have not yet made a comparison between the SPREE-based and regression-based composite estimators. This must be done before reaching any conclusions. Instead, we are left with a list of things to do:

- Compare the precision of the SPREE-based composite estimators and the regression-based composite estimators.
- Investigate the possibility of controlling the estimates for current estimates of population counts.
- Examine the similarity of definitions between the CPS and the census. If the census and CPS estimates are based on different definitions, it makes very little sense, in terms of the SPREE estimation, to construct current CPS-based estimates which have the same structure as previous census estimates.

- Produce estimates for states which have counties that cross CD boundaries.
- Extend the sample-based estimator to account for the actual sample design, rather than assuming simple random sampling.

There is also the possibility of including the previous year's estimate in the composite estimator. This needs to be examined as does the possible use of a time series approach to CD-level estimation.

Time Series

Estimation of CD characteristics will be based on March CPS data for each year, since the March CPS provides a wealth of information. Estimates from March of one year to March of the next will be correlated since four of the eight rotation groups in the CPS sample are the same from March of one year to March of the next. This autocorrelation suggests that a time series application to CD-level estimation might be appropriate. In fact, the Bureau of Labor Statistics has used time series estimation to produce monthly small area estimates of employment from CPS data (see Tiller et. al 1993).

There are several reasons why we did not explore the use of a time series in this paper:

- A time series approach would require many years of CPS March data. At the present we have access to only March 1994 data;
- The data needed would have CD codes. This would be required in order to produce sample-based CD-level estimates. At present, the only CPS data file having CD codes is from the March 1994 CPS.
- Congressional District boundaries may change dramatically every 10 years. It may be problematic to extend a time series for more than 10 years. Also, less severe boundary changes occur during intercensal years, though they would be easier to account for by adjusting previous data to correspond to current CD boundaries.

If these problems can be overcome, we feel consideration should be given to using a time series approach to CD-level estimation .

IX. Acknowledgments

No paper, no matter how inconsequential, is complete without a note of thanks to those who helped in its development. Thanks go to Donna Kostanich and Debbie Fenstermaker who helped most in developing the ideas presented. Thanks also go to four Census Bureau reviewers — Jenny Thompson, Aref Dajani, Vicki Huggins, and Machell Town — who helped immensely with their comments on drafts of this paper.

X. REFERENCES

Drew, J.D., M.P. Singh, and G.H. Choudhry (1982), "Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey," Survey Methodology, Vol. 8, pp. 17-47.

Fay, R.E. (1986), "Multivariate Components of Variance Models as Empirical Bayes Procedures for Small Domain Estimation," Proceedings of the Survey Research Methods Section, American Statistical Association.

Fay, R.E. (1987), "Application of Multivariate Regression to Small Domain Estimation," in Small Area Statistics, An International Symposium, R. Platek et. al, editors, New York, Wiley.

Fay, R.E. (1988), "Empirical Bayes Estimation for Multiple Characteristics," Proceedings of the Section on Survey Research Methods, American Statistical Association.

Fay, R.E., C. T. Nelson, and L. Litow (1993), "Estimation of Median Income for 4-Person Families by State," in Statistical Policy Working Paper 21, Indirect Estimators in Federal Programs.

Ghosh, M. and J.N.K. Rao (1994), "Small Area Estimation: An Appraisal," Statistical Science, Volume 9, pp. 55-93.

Kalton, G. (1990), "Methods of Small Area Estimation: A Review," Proceedings of the Consensus Conference on Small Area Analysis, pp. 89-95, Washington, DC: Health Resources and Services Administration.

Morris, C.N. (1983), "Parametric Bayes Inference: Theory and Application," Journal of the American Statistical Association, Volume 78, pp. 47-55.

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

Purcell, N.J. (1979), "Efficient Estimation for Small Domains: A Categorical Data Analysis Approach," Unpublished doctoral dissertation, University of Michigan.

Purcell, N.J. and L. Kish (1979), "Estimation for Small Domains," Biometrics, 35, pp. 365-384.

Purcell, N.J. and L. Kish (1980), "Postcensal Estimates for Local Areas (or Domains)," International Statistical Review, 48, pp. 3-18.

Tiller, R., S. Brown, and A. Tupek (1993), "Bureau of Labor Statistics' State and Local Area Estimates of Employment and Unemployment," Chapter 5 of Statistical Working Paper 21, Federal Committee on Small Area Estimation, Federal Committee on Statistical Methodology.

Table 1 SPREE-based Composite Estimator Results

	Variance sample-based estimator	MSE SPREE estimator	MSE SPREE-based composite estimator			
	V(m)	MSE(Mhat)	MSE(Yhat)	MSE(Mhat) /V(m)	MSE(Yhat) /V(m)	Reduction in MSE
EMPLOYMENT						
employed	316022754.4	693417499	261839617.4	2.194	0.829	17.1%
unemployed	16687991.38	10205174	10853947.5	0.612	0.650	35.0%
other	427657348.6	830499032	344155636.9	1.942	0.805	19.5%
HOUSEHOLD INCOME						
<$10,000	42822226.15	26222439	29494197.8	0.612	0.689	31.1%
$10-25,000	67277331.11	115689623	52827499.5	1.720	0.785	21.5%
$25-50,000	73019212.55	10066188	37899295.9	0.138	0.519	48.1%
$50-75,000	39778729.2	28517480	27734087.7	0.717	0.697	30.3%
>$75,000	20466117.39	0	4402854.1	0.000	0.215	78.5%

Table 2 Regression-based Composite Estimator Results

	Sample-based estimate	Regression estimate	Regression-based composite estimate	
	m	Xb	yhat	MSE(yhat)
CD 1				
EMPLOYMENT				
employed	308901	275267.15	295795.14	355501275
unemployed	17994	14569.69	15408.93	27202987
other	292406	248296.11	276228.04	550760306
HOUSEHOLD INCOME				
<$10,000	35988	36897.55	36897.55	71350596
$10-25,000	65979	59130.69	59982.38	101287121
$25-50,000	77975	70611.77	72989.26	108060713
$50-75,000	43486	35707.83	39381.17	63073178
>$75,000	22493	22850.09	22850.09	32367567