Small Area Estimation with Administrative Records and Continuous Measurement

Nanak Chand, Charles H. Alexander

U.S. Census Bureau
Washington, DC 20233

Presented to the Annual Meeting of the American Statistical Association (ASA), Chicago, Illinois, August 1996.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

I. INTRODUCTION

The American Community Survey (ACS) component of the Continuous Measurement program is designed to provide reliable direct estimates of the various population characteristics for substate areas. For small areas, such as census tracts, it is desirable to improve the ACS estimates by borrowing strength from other areas and other sources of data. In this project, we will develop procedures to derive indirect estimates of characteristics of interest by integrating ACS data with administrative records and the previous census data.

Synthetic estimators which borrow strength from similar areas may be sensitive to the similarity assumption. Regression synthetic estimators based on auxiliary data taken from other sources for the same and similar areas will be less sensitive to this assumption. The composite estimation (Singh, Gambino and Mantel (1994)) combines direct and synthetic estimators, and thus balances the potential bias of synthetic estimators against the instability of the direct estimators. In addition, the procedure may provide estimators with between area variation much smaller than the prior known variance (Spjotvoll and Thomsen (1987)).

However, composite estimators under fixed effect models provide best linear unbiased estimators which reduce to synthetic estimators for areas with small sampling fractions, irrespective of the size of between area variance relative to the within area variance. This limitation is avoided by using models which take into account random area effects (Chand and Alexander (1995), Cressie (1989, 1990, 1992), Datta et al (1992), Ericksen and Kadane (1985, 1987, 1992), Fay (1987), Fay and Herriot (1979), Ghosh and Rao (1994), and Prasad and Rao (1990)).

The paper adapts the small area methods for application to the ACS variables of interest such as proportion of population below poverty level. The applications pertain to developing estimates and their mean squared errors of such proportions for census tracts.

II. ASSUMPTIONS

A large area A is composed of m small areas A_i , i = 1, ..., m. The parameter of interest for A_i is the true population proportion P_i .

A direct estimator p_i of P_i is available from the ACS. The auxiliary data $\underline{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are available from administrative records and from previous censuses for each A_i . These data are related to P_i .

The transformation g is a function of a single variable and has a nonzero and continuous first derivative. Let

$$g_i = g(p_i), i = 1, \ldots, m.$$

We consider the small area model,

$$\underline{g} = X\underline{\beta} + \underline{t} + \underline{e},$$

where $\underline{g}, \underline{t}$ and $\underline{e}$ are mx1 vectors, $\underline{t}$ represents random area effects, $\underline{e}$ represents random sampling errors, and $\underline{g}$ has a multivariate normal distribution. X is a mxs design matrix and $\underline{\beta}$ is a sx1 vector of unknown parameters. $\underline{t}$ and $\underline{e}$ are statistically independent. Let Σ and Δ be mxm diagonal matrices with the (i, i)th elements respectively equal to τ^2 and δ_i^2 . We also assume that

$$E(\underline{e} \mid \underline{g}) = \underline{0}, Var(\underline{e} \mid \underline{g}) = \Delta, \text{ and } \underline{t} \sim N(\underline{0}, \Sigma).$$

In this paper, we consider two transformations. The first is the variance stabilization function given by

$$g_i = 2\sin^{-1}(\sqrt{p_i}),$$

and the second is the logistic function given by

$$g_i = \ln[p_i/(1 - p_i)],$$

i = 1,..., m. (Cox and Snell (1989)).

III. EMPIRICAL BEST LINEAR UNBIASED PREDICTORS (EBLUP) AND THEIR MEAN SQUARE ERRORS (MSE)

We consider four estimators of the variance component τ^2 under the model of the previous section. These are the maximum likelihood (ML) estimator, the restricted maximum likelihood (RML) estimator (Cressie (1989, 1992)), the Fay and Herriot (FH) estimator (Fay and Herriot (1979)), and a quadratic moment (QM) estimator (Prasad and Rao (1990) and Ghosh and Rao (1994)).

While the calculations of the first three estimators require iterative solutions, the last one has an explicit solution.

The ML estimators of $\underline{\beta}$ and τ^2 minimize the expression

$$\ln(|V|)+(\underline{g}-X\underline{\beta})^T V^{-1}(\underline{g}-X\underline{\beta})$$

where V is a mxm diagonal matrix with the (i, i)th element equal to $\tau^2+\delta_i^2$.

The RML estimators of $\underline{\beta}$ and τ^2 minimize

$$\ln(|V|)+\ln(|X^T V^{-1} X|)$$
$$+(\underline{g}-X\underline{\beta})^T V^{-1}(\underline{g}-X\underline{\beta})$$

The FH estimator of τ^2 is obtained by simultaneously solving

$$(\underline{g}-X\underline{\beta})^T V^{-1}(\underline{g}-X\underline{\beta})=m-s \text{ , and}$$

$$\underline{\beta}=(X^T V^{-1} X)^{-1} X^T V^{-1}\underline{g}.$$

The QM estimator of τ^2 is given by

$$(m-s)^{-1}[(\underline{g}-X\underline{\hat{b}})^T(\underline{g}-X\underline{\hat{b}})-\sum_{i=1}^{m}\delta_i^2$$
$$+\sum_{i=1}^{m}\delta_i^2\underline{x_i}^T(X^T X)^{-1}\underline{x_i}]$$

where $\underline{\hat{b}}$ is the ordinary least square estimator of $\underline{\beta}$ given by

$$\underline{\hat{b}}=(X^T X)^{-1} X^T\underline{g},$$

and $\underline{x_i}^T$ is the ith row of the design matrix X. With τ^2 estimated by one of the above four methods, let $\underline{\hat{\beta}}$ be the best linear unbiased estimator of $\underline{\beta}$ given by

$$\underline{\hat{\beta}}=(X^T U^{-1} X)^{-1} X^T U^{-1}\underline{g},$$

where U is the mxm matrix obtained from V by replacing τ^2 by its estimator $\hat{\tau}^2$.

The measure of uncertainty in the model relative to the total variance is defined as the ratio of the variance component of the random area effects to the total variance, and is given by

$$\gamma_i=\tau^2/(\tau^2+\delta_i^2), \text{ i = 1, ..., m.}$$

The regression synthetic estimator of the vector of outcome variables is the product of transpose of the design matrix and the best linear unbiased estimator of the vector of unknown parameters. Thus the regression synthetic estimator of $g(P_i)$ is $X^T\underline{\hat{\beta}}$.

The EBLUP of the outcome variable is the weighted average of the transformed direct ACS estimate and the regression synthetic estimator, the weight being the estimated measure of uncertainty in the model.

Thus the EBLUP of $g(P_i)$ is given by

$$\hat{g}_i = \hat{\gamma}_i g_i + (1-\hat{\gamma}_i)\,\underline{x}_i^T \underline{\hat{\beta}},$$

where $\hat{\gamma}_i$ is the value of γ_i when τ^2 is replaced by its estimator $\hat{\tau}^2$.

The corresponding estimator $\hat{p}_i$ of p_i is taken as $\sin^2\left(\frac{\hat{g}_i}{2}\right)$ for the variance stabilization model and as $\frac{e^{\hat{g}_i}}{1+e^{\hat{g}_i}}$ for the logistic model.

The MSE of the EBLUP, defined as the expected value of its squared deviation from the true value, consists of three parts. Part one is the sampling error variance times the measure of uncertainty in the model relative to the total variance. The second part is due to estimating the unknown parameters in the model. The third part is due to estimation of the variance component of the random area effects.

The MSE of $\hat{g}_i$ (Cressie (1992), Kackar and Harville (1984), and Ghosh and Rao (1994)) is given by

$$M_i^g = M_{ai}(\tau^2) + \delta_i^4 (\tau^2 + \delta_i^2)^{-3} v^a(\hat{\tau}^2),$$

where $v^a(\hat{\tau}^2)$ is the asymptotic variance of $\hat{\tau}^2$ and

$$M_{ai}(\tau^2) = \gamma_i \delta_i^2 + (1-\gamma_i)^2\, \underline{x}_i^T (X^T V^{-1} X)^{-1} \underline{x}_i.$$

An approximately unbiased estimator of M_i^g (Prasad and Rao (1990)) is given by

$$\hat{M}_i^g = M_{ai}(\hat{\tau}^2) + 2\,\delta_i^4 (\hat{\tau}^2 + \delta_i^2)^{-3} v^a(\hat{\tau}^2).$$

This estimator of MSE, using the moment estimators of τ^2, is valid under moderate nonnormality of the random effects $\underline{t}$.

IV. ADJUSTMENT OF EBLUP ESTIMATORS

Since ACS is designed to provide unbiased estimates for large areas, we make an adjustment to the EBLUP estimators for each A_i such that an appropriately weighted sum of these adjusted estimators equals the ACS estimate for the large area.

Let $w_i = B_i / \sum_{i=1}^{m} B_i$ be the ratio of the base population in A_i with respect to p_i, to the total base population in A.

Then the ACS estimate for A is the weighted sum of the ACS estimates for A_i with weights w_i, i = 1, ..., m.

We define the modified EBLUP $\hat{p}_i^{\,\mathrm{mod}}$ of p_i in the following steps:

This modification is similar to the one suggested by Battese, Harter, and Fuller (1988). Their model assumes that element- specific auxiliary data are available for each A_i.
Defining for i=1, ..., m,

$$W_i = w_i \hat{M}_i / \sum_{i=1}^{m} w_i^2 \hat{M}_i,$$

$\hat{M}_i$ being MSE of $\hat{p}_i$, we have,

$$\sum_{i=1}^{m} w_i W_i = 1$$

If we thus define

$$\hat{p}_i^{\,\mathrm{mod}} = \hat{p}_i + W_i (p - \sum_{i=1}^{m} w_i \hat{p}_i),$$

it follows that

$$\sum_{i=1}^{m} w_i \hat{p}_i^{\,\mathrm{mod}} = p.$$

This derivation of $\hat{p}_i^{mod}$ does not require element-specific data.

V. ESTIMATION OF PROPORTION OF PERSONS BELOW POVERTY LEVEL

We illustrate the above estimation procedures by taking $\{A_i, i = 1,...,m\}$ as the census tracts in Alameda County, California.

The direct estimate p_i of the proportion below poverty level in A_i is calculated as the ratio of weighted number of persons below poverty level to the total weighted ACS population, simulated from the 1990 census long form data. The function g is chosen as described in Section II. The sources of auxiliary data are the simulated administrative records data such as income of tax filers in the tract, and the census data such as number of persons with hispanic origin.

For the logistic model, the design matrix X is defined with s = 4 as

$$X_i1 = 1, X_i2 = \ln[\frac{C_i + .5}{B_i - C_i + .5}],$$

$$X_{i3} = \ln(T_i),$$

$$X_{i4} = \ln[\frac{H_i + .5}{B_i - H_i + .5}],$$

where, for area A_i, B_i is the base population, C_i is the number of persons with a college degree, H_i is number of persons with hispanic origin, and T_i is the simulated median income of tax filers, i = 1, ..., m.

For the variance stabilization model, the design matrix is defined with s = 4 as

$$X_i1 = 1, X_i2 = 2\sin^{-1}\sqrt{C_i / B_i},$$

$$X_i3 = \ln(T_i), X_{i4} = 2\sin^{-1}\sqrt{H_i / B_i}$$

The variance components δ_i^2 are estimated by the Jackknife method using the VPLX program (Fay (1990)).

There are a total of 291 tracts in the above ACS sample for Alameda County, giving m = 291. The suitability of the assumed models is verified by demonstrating that the standardized residuals are approximately normally distributed with mean zero and variance one.

VI. A COMPARISON OF THE VARIANCE COMPONENT ESTIMATION METHODS

The four estimation methods, when applied to the Alameda County data, gave the following estimates of τ^2.

Variance Stabilization Model (VSTM)				
	RML	ML	FH	QM
$\hat{\tau}^2$	.0688	.0678	.0696	.0710
Logistic Model (LGM)				
	RML	ML	FH	QM
$\hat{\tau}^2$	.7416	.7300	.7636	.7960

Tables A1-A2 show the four sets of EBLUP estimators of percent of persons below poverty level along with the weighted ACS estimates, for five of the 291 tracts. The four methods of variance component estimation provide similar results for each of the two models.

Tables B1-B2 show the modified EBLUP estimators of percent below poverty level. An appropriately weighted sum of these estimators equals the ACS estimate of the percent below poverty level for the whole county. This latter percent is equal to 11.01. For comparison, the weighted average of the unadjusted RML for the county is 10.73 under VSTM and is 10.94 under LGM.

Tables C1-C2 give MSE estimates associated with the four EBLUP estimators. The tables show the small levels of MSE of the EBLUP estimators for each of the estimation methods.

SIMULATED ACS SAMPLE

TABLE A1

Percent Below Poverty, Alameda County (VSTM)

Tract	ACS	RML	ML	FH	QM
4004	18.5	17.8	17.8	17.8	17.8
4052	08.1	07.9	07.9	07.9	07.9
4087	19.3	19.1	19.1	19.1	19.1
4101	06.7	07.3	07.3	07.2	07.2
4229	30.7	26.6	26.6	26.5	26.5

TABLE A2

Percent Below Poverty, Alameda County (LGM))

Tract	ACS	RML	ML	FH	QM
4004	18.5	17.9	17.9	17.9	17.9
4052	08.1	07.8	07.7	07.8	07.8
4087	19.3	19.2	19.2	19.2	19.2
4101	06.7	07.3	07.3	07.3	07.3
4229	30.7	27.9	28.0	28.0	28.1

TABLE B1

Percent Below Poverty, Alameda County (VSTM)
(MODIFIED)

Tract	ACS	RML	ML	FH	QM
4004	18.1	18.1	18.1	18.1	18.1
4052	08.1	08.0	08.0	08.0	08.0
4087	19.3	19.7	19.7	19.7	19.7
4101	06.7	07.3	07.3	07.3	07.3
4229	30.7	27.0	26.9	27.0	27.0

TABLE B2

Percent Below Poverty, Alameda County (LGM)
(MODIFIED)

Tract	ACS	RML	ML	FH	QM
4004	18.1	18.0	18.0	18.0	18.0
4052	08.1	07.8	07.8	07.8	07.8
4087	19.3	19.3	19.3	19.3	19.3

4101	06.7	07.3	07.3	07.3	07.3
4229	30.7	28.1	28.1	28.2	28.3

TABLE C1

Proportion Below Poverty, Alameda County
(MSEx10000 - VSTM)

Tract	RML	ML	FH	QM
4004	07.0	07.0	07.0	07.0
4052	02.6	02.6	02.6	02.6
4087	06.4	06.4	06.4	06.4
4101	02.2	02.2	02.2	02.2
4229	16.5	16.4	16.5	16.5

TABLE C2

Proportion Below Poverty, Alameda County
(MSEx10000 - LGM)

Tract	RML	ML	FH	QM
4004	07.1	07.1	07.1	07.1
4052	02.4	02.4	02.4	02.4
4087	06.5	06.5	06.5	06.5
4101	02.4	02.4	02.4	02.4
4229	17.3	17.3	17.4	17.5

REFERENCES

[1] Battese, G.E., Harter, R.M., and Fuller, W.A. (1988) An error-components model for prediction of county crops using survey and satellite data. J. Amer. Statist. Assoc. 83 28-36.

[2] Chand, N., and Alexander, C.H. (1995) Indirect estimation of rates and proportions for small areas with continuous measurement, 1995 Proceedings of the Section on Survey Research Methods, American Statistical Association, pp 5549-554.

[3] Cox, D.R., and Snell, E.J. (1989) The Analysis of Binary Data (2nd Edition). Methuen, London.

[4] Cressie, N. (1989) Empirical Bayes estimation of undercount in the decennial census. J.Amer. Statist. Assoc. 84 1033-1044.

[5] Cressie, N. (1990) Small area prediction of undercount using the general linear model. In Symposium 90-Measurement and Improvement of Data Quality-Proceedings 93-105. Statistics Canada, Ottawa.

[6] Cressie, N. (1992) REML estimation in empirical Bayes smoothing of census undercount. Survey Methodology 18 75-94.

[7] Datta, G.S., Ghosh, M., Huang, E.T., Isaki, C.T., Schultz, L.K., and Tsay, J.H. (1992) Hierarchical and empirical Bayes methods for adjustment of census undercount. Survey Methodology 18 95-108.

[8] Ericksen, E.P. and Kadane, J.B. (1985) Estimating the population in census year (with discussion). J. Amer. Statist. Assoc. 80 98-131.

[9] Ericksen, E.P. and Kadane, J.B. (1987) Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census. In Small Area Statistics (R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh, eds.) 23-45. Wiley, New York.

[10] Ericksen, E.P. and Kadane, J.B. (1992) Comment on "Should we have adjusted the U.S. Census of 1980," by D.A. Freedman and W.C. Navidi. Survey Methodology 18 52-58.

[11] Fay, R.E. (1987). Application of multivariate regression to small domain estimation. In Small Area Statistics (R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh, eds.) 91-102. Wiley, New York.

[12] Fay, R.E. (1990). VPLX: Variance Estimation for Complex Surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp 266-271.

[13] Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. J. Amer. Statist. Assoc. 74 269-277.

[14] Ghosh, M. and Rao, J.N.K. (1994) Small area estimation: An appraisal. Statistical Science. 9 55-93.

[15] Kackar, R.N., and Harville, D.A. (1984) Approximations for standard errors of estimators for fixed and random effects in mixed models. J. Amer. Statist. Assoc. 79 853-862.

[16] Prasad, N.G.N., and Rao, J.N.K. (1990) The estimation of mean squared errors of small area estimators. J. Amer. Statist. Assoc. 85 163-171.

[17] Singh, M.P., Gambino J., and Mantel, H.J. (1994) Issues and strategies for small area data. Survey Methodology 20 3-14.

[18] Spjotvoll, E. and Thomsen, I. (1987). Application of some empirical Bayes methods to small area estimation. Bulletin of the International Statistical Institute 2 435-439