INDIRECT ESTIMATION OF RATES AND PROPORTIONS FOR SMALL AREAS WITH CONTINUOUS MEASUREMENT

Nanak Chand, Charles H. Alexander, U.S. Bureau of the Census
Nanak Chand, U.S. Bureau of the Census, Washington, D.C. 20233

I. INTRODUCTION

This paper develops methods to estimate rates and proportions for small areas using information from the national household surveys in combination with simulated continuous measurement (CM) data. The problem is complicated because the existing models are mainly designed for continuous variables and because microdata from the surveys and the CM are not matchable.

Typically, household surveys are designed to provide unbiased estimates of characteristics of interest at the national or state levels, but their sample size is not large enough for small area estimation. The census long form or the CM survey has a larger sample size but may result in estimates with much larger bias than the survey estimates.

Two types of small area models, which take into account random area-specific effects, have been developed in the literature. In the first type, auxiliary data are available for each of the population elements. Such models are considered by Battese, Harter and Fuller (1988), Datta and Ghosh (1991), Dempster and Raghunathan (1987), Fuller and Harter (1987), Kleffe and Rao (1992), and MacGibbon and Tomberlin (1989).

In the second type of models, only area-specific auxiliary data are available. These models are considered by Cressie (1989, 1990, 1992), Datta el al (1992) Ericksen and Kadane (1985, 1987, 1992), Fay (1987), Fay and Herriot (1979), Ghosh, Datta, and Fay (1991), and Prasad and Rao (1990). Ghosh and Rao (1994) give a comprehensive review of both types of small area models.

In this paper, we adapt the above methods to use national household surveys and CM data to estimate rates and proportions at the census tract level. Some census tracts may be collapsed to assure a nonzero number of observations in the resulting groups.

We apply the adapted methods to develop indirect estimates of unemployment rates taking into account the 1994 Current Population Survey (CPS) data along with the simulated CM data for Alameda County, California. This application and others will be investigated in joint research between the Census Bureau and the Bureau of Labor Statistics on how best to integrate the CM and the CPS data. The methods may also be applicable to other such surveys.

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 national household surveys.

The auxiliary data $\underline{x}_i = (x_{i1}, \ldots, x_{is})^T$ are available from these surveys and from CM 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 continuous first derivative. Let

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

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, $\underline{t}$ is a vector of 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 $\sum$ and ∇ be mxm diagonal matrices with the (i, i)th elements respectively equal to τ^2 and δ_i^2. We also assume that

$$E(\underline{e} \mid \underline{g}) = \underline{0}, \; Var(\underline{e} \mid \underline{g}) = \nabla ,$$

and $\underline{t} \sim N(\underline{0}, \sum)$.

For our applications, we choose g as the variance stabilization function given by

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

(Cox and Snell (1989)). The variance components δ_i^2 are then given by the sampling variance formulas appropriate for the respective household survey. The suitability of the above assumptions under this transformation is tested in Section VII.

III. VARIANCE COMPONENT ESTIMATION

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)).

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 asymptotic variance of $\hat{\tau}^2$ (ML) is given by

$$\text{V(ML)} = [\frac{1}{2}\sum_{i=1}^{m} (\delta_i^2 + \tau^2)^{-2}]^{-1} .$$

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

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

The asymptotic variance of $\hat{\tau}^2$ (RML) is given by

$$\text{V(RML)} = [\frac{1}{2} trace(\pi(\tau^2)\pi(\tau^2))]^{-1} ,$$

with

$$\pi(\tau^2) = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1} .$$

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

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

$$\underline{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}\underline{g}$$

The QM estimator of τ^2 is given by

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

where

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

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

and $\underline{x}_i^T$ is the ith row of the design matrix X. Under normality, the variances of FH and QM estimators of τ^2 are given by

$$\text{V(FH)} = \text{V(QM)} = 2m^{-2}\sum_{i=1}^{m} (\delta_i^2 + \tau^2)^2 .$$

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

With τ^2 estimated by one of the four methods in Section III., let $\hat{\beta}$ be the best linear unbiased estimator of β given by

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

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

$$\gamma_i = \tau^2 / (\tau^2 + \delta_i^2) ,$$

be the measure of uncertainty in the model relative to the total variance. Then the regression synthetic estimator of $\hat{g}$ is $X^T\hat{\beta}$ and the EBLUP of $g(P_i)$ is given by

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

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

The MSE of $\hat{g}_i$ (Cressie (1992), Kacker and Harville (1984), and Prasad and Rao (1990)) consists of three parts.

The first part is due to the measure of uncertainty in the model relative to the total variance. The second part is due to estimation of unknown parameters in the model. The third part is due to estimation of variance components of the random area effects.

V. ADJUSTMENT OF EBLUP ESTIMATORS

Since national household surveys are designed to provide unbiased estimates for large areas, we will make an adjustment to the EBLUP estimators for each A_i such that an appropriately weighted sum of these adjusted estimators equals the household survey estimate for the large area.

Let (i, j) denote the jth person in small area A_i in a household survey and let f_{ij} be the final survey weight assigned to (i, j), i = 1, ..., m, j = 1 ,..., n_i, n_i being the number of persons in the sample in the base population with respect to which the characteristic C of interest is measured.

We define the variables b_{ij} and c_{ij} as

b_{ij} = 1, if (i, j) belongs to the base population,
= 0 otherwise

c_{ij} = 1, if (i, j) has characteristic C,
= 0 otherwise

The household survey estimate p_i of proportion P_i of persons with characteristic C in A_i is defined as

$$p_i = \sum_{j=1}^{n_i} f_{ij} c_{ij} / \sum_{j=1}^{n_i} f_{ij} b_{ij} = c_i / b_i ,$$

where c_i and b_i are respectively the weighted number of persons with characteristic C, and weighted base population with respect to which C is measured, in A_i , i = 1, ..., m.

The corresponding household survey estimate for the large area is

$$p = \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_{ij} c_{ij} / \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_{ij} b_{ij}$$

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

where $$w_i = b_i / \sum_{i=1}^{m} b_i .$$

Thus the household survey estimate for the large area is a weighted sum of the household survey estimates of small areas with weights w_i, i = 1, ..., m.

We define the modified EBLUP $\hat{p}_i^{\text{mod}}$ 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,

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

with $\hat{M}_i = MSE(\hat{p}_i)$ defined in Section IV,

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

If we thus define

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

it follows that

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

Thus the weighted average of the modified EBLUP estimators equals the household survey estimate for the large area. We note that the above calculation of $\hat{p}_i^{\text{mod}}$ does not require person-specific data.

VI. ESTIMATION OF UNEMPLOYMENT RATES

We illustrate the above estimation procedures by taking { A_i, $i = 1, \ldots, m$ } as the census tracts in Alameda County, California. Some tracts are combined to result in nonzero number of observations in each A_i in the CPS samples during 1994.

The direct estimate p_i of the unemployment rate in A_i is calculated as the ratio of weighted number of unemployed to the total weighted labor force sixteen years or older, in the twelve monthly samples in 1994. The function g is taken as described in Section II.

The design matrix X is defined with s = 2 as

$$x_{i1} = \frac{1}{2} mk \sin^{-1}(1 / N_i \sqrt{N_i}) , \text{ and}$$

$$x_{i2} = 2 \sin^{-1}(\sqrt{p_i^{cm}}) \quad , i = 1, \ldots, m.$$

where p_i^{cm} is the unemployment rate observed in A_i in the simulated CM sample, k = 1,000 is a normalizing constant, and N_i is the total labor force represented by the sample in A_i.

The diagonal elements { δ_i^2 } of ∇ are calculated from the sampling variance formulas for CPS. This gives

$$\delta_i^2 = a^2 b^G / N_i ,$$

where a is the adjustment of the standard error estimate from monthly to annual data, and b^G is the variance generalization parameter for the total unemployed (U.S. Bureau of Labor Statistics (October, 1993)).

There are a total of seventy two tracts in the 1994 CPS sample for Alameda County. We collapsed thirty two tracts, giving m = 40.

VII. CHECKING THE SUITABILITY OF THE ASSUMED MODEL

When the model is correct, the standardized residuals given by

$$r_i = (g_i - \underline{x}_i^T\beta) / \sqrt{(\hat{\tau}^2 + \delta_i^2)},$$

i = 1 ,..., m are approximately distributed as N(0, 1) variables.

We first verified that the skewness and kurtosis of the standardized residuals for each of the four methods of estimation lie within the 95 percent confidence intervals for these statistics.

We also applied the Shapiro-Wilk test for testing the hypothesis that the standardized residuals are a random sample from the N(0, 1) distribution. This test accepted the null hypotheses for each of the estimation methods.

VIII. A COMPARISON OF THE VARIANCE COMPONENT ESTIMATION METHODS

The four estimation methods, when applied to the Alameda County data, gave the following estimates of $\underline{\beta}$ and τ^2 .

	RML	ML	FH	QM
$\hat{\beta}_1$	1.6193	1.6305	1.6504	1.7131
$\hat{\beta}_2$	.5424	.5428	.5436	.5464
$\hat{\tau}^2$	.0414	.0386	.0338	.0209

Table A shows the four sets of EBLUP estimators of unemployment rates along with the weighted CPS estimates, for five of the tract groups.

Table B shows the modified EBLUP estimators of unemployment rates. An appropriately weighted sum of these estimates, for the forty tract groups, equals the CPS estimate of unemployment rate for the whole county. This latter rate is equal to 9.37757 percent.

Table C gives MSE estimates associated with the four EBLUP estimators. While RML has other advantages, RML consistently results in higher MSE than ML.

TABLE A

1994 UNEMPLOYMENT RATES
Alameda County, CA (%)
Weighted CPS: 9.37757

Tract-Group	CPS	RML	ML	FH	QM
05	06.4	08.9	09.0	09.2	09.8
10	13.5	08.4	08.3	08.0	06.9
20	06.6	06.9	06.9	06.9	07.0
30	09.3	16.1	16.3	16.6	17.6
40	07.2	07.3	07.3	07.3	07.4

RML for the County: 8.44504

TABLE B

1994 UNEMPLOYMENT RATES
Alameda County, CA (%)
Weighted CPS: 9.37757

Tract-Group	CPS	(MODIFIED) RML	ML	FH	QM
05	06.4	09.6	09.7	09.9	10.5
10	13.5	09.3	09.2	08.9	07.7
20	06.6	07.6	07.6	07.7	07.9
30	09.3	16.7	16.9	17.2	18.1
40	07.2	08.4	08.5	08.6	09.1

TABLE C

1994 UNEMPLOYMENT RATES
Alameda County, CA
100xMSE

Tract-Group	Sample-Size	RML	ML	FH	QM
05	26	.217	.210	.211	.176
10	54	.154	.147	.144	.110
20	51	.131	.127	.130	.115
30	11	.465	.443	.424	.309
40	253	.042	.041	.043	.045

REFERENCES

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

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

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

[4] 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.

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

[6] Datta, G.S. and Ghosh, M. (1991) Bayesian prediction in linear models: Applications to small area estimation. Ann. Statist. 19 1748-1770.

[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] Dempster, A.P. and Raghunathan, T.E., Using a covariate for small area estimation: a common sense Bayesian approach. In Small Area Statistics (R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh, eds.) 77-90. Wiley, New York.

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

[10] 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.

[11] 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.

[12] 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.

[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] Fuller, W.A. and Harter, R.M. (1987). The multivariate components of variance model for small area estimation. In Small Area Statistics (R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh, eds.) 103-123. Wiley, New York.

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

[16] Ghosh, M., Datta, G.S. and Fay, R.E. (1991) Hierarchical and empirical multivariate Bayes analysis in small area estimation. In Proceedings of the Bureau of the Census Annual Research Conference 63-79. Bureau of the Census, Washington, D.C.

[17] 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.

[18] Kleffe, J. and Rao, J.N.K. (1992) Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. J. Multivariate Ana. 43 1-15.

[19] MacGibbon, B. and Tomberlin, T.J. (1989). Small area estimation of proportions via empirical Bayes techniques. Survey Methodology 15 237-252.

[20] 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.

[21] U.S. Department of Labor (October 1993). Employment and Earnings, Bureau of Labor Statistics, Washington, D.C. 20212.