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

Nanak Chand, Charles H. Alexander, U.S. Bureau of the Census

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 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 \( A_i \), \( i = 1, \ldots, m \). 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 \( x_i = (x_{i1}, \ldots, x_{is})' \) 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 \( g_i = g(P_i), i = 1, \ldots, m \).

We consider the small area model,

\[ g = X\beta + \varepsilon + \omega, \]

where \( g, \beta, \varepsilon, \) and \( \omega \) are mx1 vectors, \( \beta \) is a vector of random area effects, \( \varepsilon \) represents random sampling errors, and \( g \) has a multivariate normal distribution. \( X \) is a mxs design matrix and \( \beta \) is a
sx1 vector of unknown parameters. \( \mathbf{L} \) and \( \mathbf{G} \) are statistically independent. Let \( \sum \mathbf{L} \) and \( \mathbf{V} \) be mxm diagonal matrices with the \((i, i)\)th elements respectively equal to \( \tau^2 \) and \( \delta_i^2 \). We also assume that

\[
E(\mathbf{G} | \mathbf{G}) = \mathbf{Q}, \quad \text{Var}(\mathbf{G} | \mathbf{G}) = \mathbf{V} ,
\]

and \( \mathbf{L} \sim \mathcal{N}(\mathbf{Q}, \sum) \).

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

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

(Cox and Snell (1989)). The variance components \( \delta_i^2 \) 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 \( \tau^2 \) 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 \( \beta \) and \( \tau^2 \) minimize the expression

\[
\ln(|\mathbf{V}|) + \ln(|\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|) + (\mathbf{G} - \mathbf{X} \hat{\beta})^T \mathbf{V}^{-1} (\mathbf{G} - \mathbf{X} \hat{\beta})
\]

where \( \mathbf{V} \) 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

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

The RML estimators of \( \beta \) and \( \tau^2 \) minimize

\[
\ln(|\mathbf{V}|) + \ln(|\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|) + (\mathbf{G} - \mathbf{X} \hat{\beta})^T \mathbf{V}^{-1} (\mathbf{G} - \mathbf{X} \hat{\beta})
\]

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

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

with

\[
\pi(\tau^2) = \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}.
\]

The FH estimator of \( \tau^2 \) is obtained by simultaneously solving

\[
(\mathbf{G} - \mathbf{X} \hat{\beta})^T \mathbf{V}^{-1} (\mathbf{G} - \mathbf{X} \hat{\beta}) = m - s, \quad \text{and}
\]

\[
\hat{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{G}.
\]

The QM estimator of \( \tau^2 \) is given by

\[
(m - s)^{-1} \left[ (\mathbf{G} - \mathbf{X} \hat{\beta})^T (\mathbf{G} - \mathbf{X} \hat{\beta}) - \sum_{i=1}^{m} \delta_i^2 \right.
\]

\[
+ \sum_{i=1}^{m} \delta_i^2 \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \right]
\]

where \( \hat{\beta} \) is the ordinary least square estimator of \( \beta \) given by

\[
\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{G},
\]

and \( \mathbf{x}_i^T \) is the \( i \)th row of the design matrix \( \mathbf{X} \).

Under normality, the variances of FH and QM estimators of \( \tau^2 \) are given by

\[
\text{V}(\text{FH}) = \text{V}(\text{QM}) = 2m^{-2} \sum_{i=1}^{m} (\delta_i^2 + \tau^2)^2.
\]
IV. EMPIRICAL BEST LINEAR UNBIASED PREDICTORS (EBLUP) AND THEIR MEAN SQUARE ERRORS (MSE)

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

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

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

$$\gamma_i = \frac{\tau^2}{(\tau^2 + \delta_i^2)},$$

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

$$\hat{g}_i = \gamma_i g_i + (1-\gamma_i)X^T\hat{\theta},$$

where $\gamma_i$ is the value of $\gamma_i$ when $\tau^2$ 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 $j$th person in small area $A_i$ in a household survey and let $\ell_{ij}$ 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

$$b_{ij} = 1, \text{ if } (i, j) \text{ belongs to the base population},$$

$$0 \text{ otherwise},$$

$$c_{ij} = 1, \text{ if } (i, j) \text{ has characteristic } C,$$

$$0 \text{ otherwise}.$$

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 \[ w_i = \frac{b_i}{\sum_{i=1}^{m} b_i} \].

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

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 \( i = 1, \ldots, m \),

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

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

\[ \sum_{i=1}^{m} w_i W_i = 1 \]

If we thus define

\[ \hat{P}_i^{\text{mod}} = \hat{P}_i + W_i \left( P - \sum_{i=1}^{m} w_i \hat{P}_i \right), \]

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

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

\[ x_{i2} = 2 \sin^{-1} \left( \sqrt{p_i \text{cm}} \right), i = 1, \ldots, m. \]

where \( p_i \text{cm} \) 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 \( \{ \delta_j^2 \} \) of \( \Sigma \) are calculated from the sampling variance formulas for CPS. This gives

\[ \delta_j^2 = a^2 b^G / N_i, \]

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

\[ r_i = \frac{g_i - x_i^T \hat{\beta}}{\sqrt{\hat{\tau}^2 + \delta_i^2}}, \]

\( i = 1, \ldots, m \) 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 \( \hat{\beta} \) and \( \hat{\tau}^2 \).

<table>
<thead>
<tr>
<th></th>
<th>RML</th>
<th>ML</th>
<th>FH</th>
<th>QM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>1.6193</td>
<td>1.6305</td>
<td>1.6504</td>
<td>1.7131</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>.5424</td>
<td>.5428</td>
<td>.5436</td>
<td>.5464</td>
</tr>
<tr>
<td>( \hat{\tau}^2 )</td>
<td>.0414</td>
<td>.0386</td>
<td>.0338</td>
<td>.0209</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Tract-Group</th>
<th>CPS</th>
<th>RML</th>
<th>ML</th>
<th>FH</th>
<th>QM</th>
</tr>
</thead>
<tbody>
<tr>
<td>05</td>
<td>6.4</td>
<td>08.9</td>
<td>09.0</td>
<td>09.2</td>
<td>09.8</td>
</tr>
<tr>
<td>10</td>
<td>13.5</td>
<td>08.4</td>
<td>08.3</td>
<td>08.0</td>
<td>06.9</td>
</tr>
<tr>
<td>20</td>
<td>06.6</td>
<td>06.9</td>
<td>06.9</td>
<td>06.9</td>
<td>07.0</td>
</tr>
<tr>
<td>30</td>
<td>09.3</td>
<td>16.1</td>
<td>16.3</td>
<td>16.6</td>
<td>17.6</td>
</tr>
<tr>
<td>40</td>
<td>07.2</td>
<td>07.3</td>
<td>07.3</td>
<td>07.4</td>
<td>07.4</td>
</tr>
<tr>
<td>RML for the County: 8.44504</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE B

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

<table>
<thead>
<tr>
<th>Tract-Group (MODIFIED)</th>
<th>CPS</th>
<th>RML</th>
<th>ML</th>
<th>FH</th>
<th>QM</th>
</tr>
</thead>
<tbody>
<tr>
<td>05</td>
<td>6.4</td>
<td>09.6</td>
<td>09.7</td>
<td>09.9</td>
<td>10.5</td>
</tr>
<tr>
<td>10</td>
<td>13.5</td>
<td>09.3</td>
<td>09.2</td>
<td>08.9</td>
<td>07.7</td>
</tr>
<tr>
<td>20</td>
<td>06.6</td>
<td>07.6</td>
<td>07.6</td>
<td>07.7</td>
<td>07.9</td>
</tr>
<tr>
<td>30</td>
<td>09.3</td>
<td>16.7</td>
<td>16.9</td>
<td>17.2</td>
<td>18.1</td>
</tr>
<tr>
<td>40</td>
<td>07.2</td>
<td>08.4</td>
<td>08.5</td>
<td>08.6</td>
<td>09.1</td>
</tr>
</tbody>
</table>

TABLE C

1994 UNEMPLOYMENT RATES
Alameda County, CA
100xMSE

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

553
REFERENCES


