USING ADMINISTRATIVE RECORDS FOR SMALL AREA ESTIMATION IN THE AMERICAN COMMUNITY SURVEY

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

1. Introduction

This paper describes methods to estimate rates and proportions for small areas by integrating data from administrative records with those of the American Community Survey (ACS). ACS is designed to provide reliable estimates of characteristics of interest for substate areas, but its sample size may not be large enough for smaller areas such as census tracts. We consider a class of small area procedures which borrow strength from neighboring areas and outside sources of data, the outside source for this paper being the administrative records.

Two types of small area models, which take into account random area 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), 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 Chand and Alexander (1995), Cressie (1989, 1990, 1992), Datta el al (1992) Ericksen and Kadane (1985, 1987, 1992), Fay (1987), Fay and Herriot (1979), Ghosh and Rao (1994), Ghosh, Datta, and Fay (1991), Kackar and Harville (1984), Prasad and Rao (1990), Singh, Gambino and Mantel (1994), and Spjotvoll and Thomsen (1987). The background and motivation of these methods is described in detail in Ghosh and Rao (1994).

Subsequent sections describe the underlying model and assumptions as they pertain to our situation, summarize four different methods for estimating the variance components, give formulas for deriving the empirical Bayes (EB) estimators and their mean square errors, and provide an adjustment to the EB estimators of proportions such that a suitably weighted sum of the modified estimators for small areas equals the corresponding ACS estimate for the large area. The paper also illustrates the methods by developing estimators of poverty rates at the census tract level, respectively for the simulated ACS data for Alameda county, California, and for three of the 1996 ACS sites. In addition, the paper compares the estimates of parameters of the model under the proposed methods, and provides additional statistics.

2. Model and Assumptions

A large area is composed of m small areas. The parameter of interest for a particular small area is the true population proportion . A direct estimator of is available from the ACS. The auxiliary data are available from administrative records and other sources for each of the small areas. In this paper, we are addressing the problem of using auxiliary data to reduce the variance of the ACS estimates. We are not considering any measurement errors in the ACS, which of course need to be addressed in a full treatment. The transformation g is a function of a single variable and has a nonzero continuous first derivative. Let

.

We consider the small area model,

,

where and are mx1 vectors, represents random area effects, represents random sampling errors, X is a mxs

design matrix and is a sx1 vector of unknown parameters.

We assume that the random area effects and the random sampling errors are statistically independent, are uncorrelated within themselves, have zero mean, and a normal distribution.

The paper studies two transformations, the variance stabilization transformation and the logistic transformation. In the first case is given by

and for the logistic case, we have,

(Cox and Snell (1989).) The process uses error variance components given by the sampling variance formulas appropriate for ACS. We also test the suitability of the underlying assumptions under each of these transformations.

3. Variance Component Estimation

We consider four methods of estimating the variance components for the random area effects.

The resulting estimators are the maximum likelihood (ML) estimator, the restricted maximum likelihood (RML) estimator, the Fay-Herriot (FH) estimator, and a quadratic moment (QM) estimator. The maximum likelihood and the restricted maximum likelihood estimators require iterative solutions to the likelihood equations. These are described in Chand and Alexander (1995) and Cressie (1989, 1992). The RML estimators of and minimize

.

The Fay-Herriot estimator also requires an iterative solution, and is obtained by equating to one the ratio of error component of variance to the error mean square for the weighted least square analysis. Calculation of the quadratic moment estimator and its variance does not require iterative solution, it is described in Prasad and Rao (1990), and is given by

4. Empirical Bayes (EB) Estimators, their Mean Square Errors, and Modified Estimators

The regression synthetic estimator of the outcome vector is the product of transpose of the design matrix and the best linear unbiased estimator of the vector of unknown parameters.

Defining the measure of uncertainty in the model as the ratio of the variance component of the random area effects to the total variance, the EB estimator of the outcome variable is the weighted average of the transformed direct estimate and the regression synthetic estimator, the weight being the estimated measure of uncertainty given by

,

being the error variance component.

The mean square error of an estimator is the expected value of its squared deviation from the true value. The mean square error of the EB estimator of the outcome variable 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 components of the random area effects. Denoting the first part by , MSE of is given by,

where is the asymptotic variance of .

We modify the EB estimators for each of the small areas such that an appropriately weighted sum of the resulting estimators equals the direct survey estimate for the large area. The modified estimator is similar to the one given in Battese, Harter, and Fuller (1988), and is the sum of the EB estimator for the particular area and a predetermined weight times the difference between the direct survey estimate and the weighted average of the EB estimators for each of the small areas:

where the weights satisfy ,

being the ratio of base population of the ith tract to that of the respective ACS site.

5. Estimation of Proportion Below Poverty Level

5a. Simulated ACS Data: Alameda County, California

We illustrate the above estimation procedures first by taking { } as the census tracts in Alameda County, California. This example provides comparisons between the logistic and the variance stabilization transformations as well as among the four methods used to estimate variance components of the random area effects.

The direct estimate of the proportion below poverty level in 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 before. 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 (LGM), the design matrix X is defined with s = 4 as

where, for area ,

is the base population, is the number of persons with a college degree, is number of persons with hispanic origin, and is the simulated median income of tax filers, i = 1, ..., m.

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

.

There are a total of 291 tracts in the above ACS sample for Alameda County, giving m = 291.

We tested the appropriateness of the assumed models by verifying that the standardized residuals

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

Tables A1-A2 show the four sets of EB estimators of proportions below poverty level along with the weighted ACS estimates, for randomly selected tracts. The four sets of estimators provide values which are close to one another under the two models.

Tables B1-B2 show the modified EB 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 estimates of the MSE associated with the four EB estimators. The tables show the small levels of the MSE of EB estimators for each of the estimation methods.

5b. 1996 ACS Sites

The second illustration consists of taking {[illegible]} as the census tracts respectively in Brevard County Florida, Multnomah County/Portland Oregon, and Rockland County New York. The direct estimate [illegible] of the proportion below poverty level in [illegible] is calculated as the ratio of weighted number of persons below poverty level to the total weighted ACS population in the respective tract. The function g is taken as described before.

The design matrix X is defined with s = 6 based on the Internal Revenue Service variables as

[illegible] [illegible]

[illegible]

and

[illegible]

We tested the suitability of the assumed model, obtaining results similar as in the case of simulated data. Table A shows the four sets of EB estimators of proportions below poverty level along with the weighted ACS estimates for randomly selected tracts, with the four methods providing comparable values. Table B shows the modified EB estimators of proportions below poverty level. The modifications meet the large area matching requirements.

Tables C gives MSE estimates associated with the four EB estimators. The table shows the small levels of MSE of the EB estimators for each of the estimation methods. The following table shows the reduction in variance achieved by the estimation process.

Reduction in Variance County Averages

Site	m	ACS Variance x 1000	MSEx1000	Percent Reduction
Brevard	86	0.4727	0.3065	35.16%
Multnomah	164	1.0728	0.3775	64.81%
Rockland	39	0.1401	0.1129	19.41%
Composite		0.5619	0.2656	52.73%

6. Analysis Applicable to Ultimate ACS Size Levels and other Future Research Issues

Since the ultimate ACS sample will be about twenty percent of the 1996 sample, we perform the following analysis appropriate for the ultimate size levels. For area i, let denote the direct estimate of proportion of persons in poverty in the kth systematic sample of one-fifth size taken from the full ACS sample for a specified site, and let denote the corresponding estimate from the remaining four-fifth sample, i = 1, ..., m; k = 1, ..., 5. Also, let and be the corresponding transformed values. We repeat the analysis of sections 2 - 4 replacing by , i = 1, ..., m ; k= 1, ..., 5.

Let and be the kth sample estimators derived similar to the full sample case, and let

and , be the corresponding estimates of the their mean squared errors. Also let

and the variance estimates of and respectively. Then we study the following 2m test statistics:

, i = 1, ..., m, and

, i = 1, ..., m.

These statistics provide a measure to test the difference between the model estimators given by the one-fifth sample as compared with the larger complementary four-fifth sample, for each of the m areas. Table D gives values of and , for the first, third, and fifth samples for randomly selected areas for Multnomah County/ Portland.

Other future research issues pertain to comparisons among the various alternative estimation procedures measured by criteria such as simplicity and reduction in the mean squared errors. Use of multi-year averages may involve questions pertaining to optimum number of years and appropriate weights and methods applicable to direct, model based, and various composite estimates.

There may be additional issues regarding the use of traditional time series methods when a number of years= data are available. The application of analysis of previous sections for estimating year to year differences may raise questions as to change in tax laws and other similar factors.

The following are Tables A1-A2, B1-B2, and C1-C2 for the simulated ACS data, and Tables A-D for Multnomah County/Portland. Reference list is available from the authors.

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 (VSTM) (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: Percent 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: Percent 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

Table A: ESTIMATES OF 1996 POVERTY RATES
Multnomah County/Portland Oregon

Tract	Weighted Full ACS Estimate	Restricted Maximum Likelihood EBLUP	Maximum Likelihood EBLUP	Fay-Herriot EBLUP	Quadratic Moment EBLUP
00301	0.16921	0.16095	0.16059	0.16126	0.16196
02301	0.31061	0.30436	0.30419	0.30451	0.30486
03301	0.2961	0.32044	0.32132	0.31969	0.31795
06601	0.05882	0.05698	0.05691	0.05705	0.05719
10406	0.14781	0.14505	0.14491	0.14516	0.14541

Table B: MODIFIED ESTIMATES OF 1996 POVERTY RATES
Multnomah County/Portland Oregon

Tract	MODIFIED RML EBLUP	MODIFIED ML EBLUP	MODIFIED FH EBLUP	MODIFIED QM EBLUP
00301	0.16279	0.16249	0.16305	0.16362
02301	0.30719	0.30710	0.30728	0.30746
03301	0.32422	0.32521	0.32339	0.32143
06601	0.05750	0.05744	0.05755	0.05766
10406	0.14716	0.14711	0.14720	0.14730

Table C: MEAN SQUARE ERRORS OF ESTIMATES OF 1996 POVERTY RATES
Multnomah County/Portland Oregon

Tract	MSE RML EBLUP	MSE ML EBLUP	MSE FH EBLUP	MSE QM EBLUP
00301	0.00035793	0.00035317	0.00036540	0.00037370
02301	0.00083616	0.00082158	0.00085910	0.00088520
03301	0.00097850	0.00096083	0.00100660	0.00103840
06601	0.00015764	0.00015544	0.00016110	0.00016500
10406	0.00022753	0.00022554	0.00023080	0.00023410

Table D: TEST STATISTICS FOR SAMPLE j FOR THE 1996 POVERTY RATES
Multnomah County/Portland Oregon
j=1

Tract	RML Statistic for g	RML Statistic for p	ML Statistic for g	ML Statistic for p
00301	-0.65136	-0.66296	-0.65825	-0.66986
02301	0.60909	0.60607	0.59314	0.59037
03301	-0.38188	-0.38351	-0.36814	-0.36959
06601	0.47352	0.46549	0.46122	0.45378
10406	-1.20907	-1.27181	-1.21365	-1.27585

Table D: TEST STATISTICS FOR SAMPLE j FOR THE 1996 POVERTY RATES
(Continued)
Multnomah County/Portland Oregon
j=1

Tract	FH Statistic for g	FH Statistic for p	QM Statistic for g	QM Statistic for p
00301	-0.6354	-0.64741	-0.63918	-0.65126
02301	0.58981	0.58653	0.57516	0.57208
03301	-0.36889	-0.37063	-0.35668	-0.35827
06601	0.46065	0.45219	0.44932	0.44132
10406	-1.17437	-1.23743	-1.17424	-1.23693

Table D: TEST STATISTICS FOR SAMPLE j FOR THE 1996 POVERTY RATES
(Continued)
Multnomah County/Portland Oregon
j=3

Tract	RML Statistic for g	RML Statistic for p	ML Statistic for g	ML Statistic for p
00301	0.363140	0.359220	0.333450	0.330190
02301	-0.414930	-0.418120	-0.412980	-0.416030
03301	-1.366590	-1.391890	-1.318830	-1.341090
06601	-0.581270	-0.598090	-0.590270	-0.607180
10406	0.201090	0.200500	0.192770	0.192240

Table D: TEST STATISTICS FOR SAMPLE j FOR THE 1996 POVERTY RATES (Continued)
j=3

Tract	FH Statistic for g	FH Statistic for p	QM Statistic for g	QM Statistic for p
00301	0.342200	0.338730	0.313870	0.310990
02301	-0.412690	-0.415790	-0.410620	-0.413580
03301	-1.331420	-1.354760	-1.285820	-1.306470
06601	-0.586300	-0.603250	-0.594790	-0.611840
10406	0.195230	0.194680	0.187190	0.186690

Table D : TEST STATISTICS FOR SAMPLE j FOR THE 1996 POVERTY RATES (Continued)
Multnomah County/Portland Oregon
j=5

Tract	RML Statistic for g	RML Statistic for p	ML Statistic for g	ML Statistic for p
00301	0.21970	0.21857	0.19808	0.19718
02301	-0.10651	-0.10662	-0.11352	-0.11364
03301	1.10300	1.08657	1.11543	1.09924
06601	0.69552	0.66291	0.67377	0.64357
10406	-0.22915	-0.23067	-0.23923	-0.24087

Table D: TEST STATISTICS FOR SAMPLE j FOR THE 1996 POVERTY RATES (Continued)
Multnomah County/Portland Oregon
j=5

Tract	FH Statistic for g	FH Statistic for p	QM Statistic for g	QM Statistic for p
00301	0.21651	0.21541	0.19741	0.19651
02301	-0.10715	-0.10726	-0.11328	-0.11340
03301	1.10150	1.08503	1.11198	1.09571
06601	0.69097	0.65863	0.67155	0.64132
10406	-0.22995	-0.23149	-0.23873	-0.24038