ANALYSIS OF GENERALIZED VARIANCE ESTIMATION FOR THE CENSUS 2000 DRESS REHEARSAL

Michael D. Starsinic and Machell K. Town, U.S. Bureau of the Census
Michael D. Starsinic, U.S. Bureau of the Census, Washington, DC 20233

KEY WORDS: generalized variance, census, small area estimation

1. INTRODUCTION[1]

It is the policy of the U.S. Census Bureau to provide measures of how reliable its published estimates are. Due to the very large number of published estimates for Census 2000, it is not feasible to report a standard error for each estimate. Instead, it was decided to compute generalized variance parameters for a set of general characteristics for data product users to compute an estimate of the variance for any desired estimate at any desired geographic level. Computing a generalized variance model also eases the problem of instability associated with estimating standard errors for very small populations, such as the census's redistricting (Public Law 94-171) data released at the block and tract level, crosstabulated by race, Hispanic origin, and age. A method of computing the generalized variances using a weighted least-squares regression (Wolter 1985) was implemented in the 1995 Census Test (Krenzke and Navarro 1996). Basing our efforts on that work, the model was used again to calculate the generalized variances for the Census 2000 Dress Rehearsal, and it is planned to be the method used in production for Census 2000. This paper analyzes the results of the modeling from the Census 2000 Dress Rehearsal. Sections two and three give brief overviews of the sampling, estimation, and direct variance estimation processes, and results of the variance generalization are found in section four.

2. SAMPLE DESIGN AND ESTIMATION

The Census 2000 Dress Rehearsal was conducted at three sites: the city of Sacramento, California (henceforth referred to as the Sacramento site, or simply Sacramento); Menominee County, Wisconsin, including the Menominee American Indian Reservation (henceforth referred to as the Menominee site, or simply Menominee); and eleven counties is South Carolina, including the city of Columbia and town of Irmo (henceforth referred to as the South Carolina site, or simply South Carolina). In accordance with the Congressional directions, a different combination of sampling procedures was conducted at each site. In Sacramento, sampling was done for the follow-up of non-respondents (NRFU) and U.S. Postal Service-identified undeliverable-as-addressed (UAA) vacants, while 100% follow-up was done in the other two sites. An Integrated Coverage Measurement (ICM) survey was conducted in all three sites in order to correct for undercoverage, although the ICM in South Carolina was done in the form of a post-enumeration survey (PES) and the official, published results do not include the ICM adjustment. The South Carolina PES is not considered in this paper. Additionally, data products were prepared and released for Sacramento and Menominee "without statistical methods", which meant that ICM adjustments were ignored. Under these conditions, Sacramento's estimates were still subject to the sampling variability due to NRFU and UAA vacant estimation, but Menominee had no sampling error.

In Sacramento, a systematic sample of nonrespondent addresses were selected in each tract to be followed-up by enumerators. As many housing units were selected for the sample to bring the response rate to 90%. For example, if a tract had a 60% mail response rate, a three-in-four sample would be selected, which would bring the total response rate up to 90%. Tracts with an 85% or higher mail response rate would be sampled at a fixed one-in-three rate. Housing units identified as UAA vacants were sampled at a fixed rate of three-in-ten. In blocks in the ICM sample, all non-respondents and UAA vacants were followed-up, which would likely raise the final response rate above 90% in tracts containing those blocks. The remaining nonsampled nonrespondents and nonsampled UAA vacants were then estimated using a nearest-neighbor hot deck imputation.

In both Sacramento and Menominee, block clusters were selected to become part of the ICM sample. The E-Sample essentially consisted of the initial phase results from the sampled blocks. The P-Sample was an independent sample conducted in the same selected blocks. The results were compared to identify each

[1] 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.

individual in the E-Sample as a correct or erroneous enumeration, and each individual in the P-Sample as a match or nonmatch to an E-Sample individual.

Coverage factors for each of 84 poststrata based on combinations of age, sex, tenure, race, and Hispanic origin were then computed using:

$$CF = \frac{IP - II}{IP} \times \frac{CE}{E} \times \frac{P}{M}$$

where IP is the weighted initial phase estimate for the poststratum, II is the weighted number of persons in the initial phase whose data were wholly imputed, CE is the weighted estimate of the number of persons in the initial phase who were correctly enumerated, E is the weighted estimate of the number of persons in the initial phase, P is the weighted estimate of the number of persons found by the independent ICM collection procedures, and M is the weighted estimate of the number of persons found by the independent ICM collection procedures who can be matched to persons enumerated in the initial phase. All the estimates were based on the results of the ICM sample blocks. The dual system estimates, obtained by multiplying the coverage factors by the initial phase estimates for each poststrata, were then raked & the results rounded to produce final population estimates. Coverage factors greater than one indicated an undercount and, those less than one indicated an overcount.

3. DIRECT ESTIMATION OF VARIANCE

The variance of estimates from the Census 2000 Dress Rehearsal in Sacramento was composed of two components, the variance due to NRFU & UAA vacant sampling and the variance due to ICM sampling. The variance in Menominee was just the variance due to ICM sampling, since NRFU and UAA vacant sampling did not occur at that site. These are the only two components of sampling variability that the variance was intended to capture.

To calculate the variance due to NRFU and UAA vacant sampling, each housing unit was assigned 300 0-1 replicate weights. All respondents and sampled nonrespondents have all their replicate weights equal to one. In each of the replicates, a prespecified subset of the NRFU-estimated households were given a zero weight and "replaced" with the household information for the "second nearest neighbor", the household that would have been used according to the nearest neighbor imputation rules if the actual nearest neighbor had not been in the sample. Each individual in a housing unit received that housing unit's set of replicate weights. The weights were multiplied by each individual's coverage factor, and the new weights were then summed to give 300 estimates of the site's population. The NRFU/UAA component of the variance was then computed from the 300 estimates. The variance formula can be applied to any subset of the site, so block- and tract-level estimates are easily computed. The theoretical basis for this variance estimator is laid out in Fay and Town (1998).

The variance due to ICM sampling at the site level is calculated using a more straightforward jackknife procedure, with each sampling stratum/substratum zero-weighted in turn. As the imputation for missing data in the ICM sample is included in the recalculation of the dual system estimates in each jackknife replicate, its variability is incorporated into the ICM sampling variance. ICM variances at levels lower than site were computed using a site-level covariance matrix for the poststrata coverage factors.

The total variance for any geographic level in Sacramento is simply the sum of the NRFU/UAA vacant and ICM variances for that level.

3.1 Comparison with 1995 Census Test Results

For Sacramento, we can look at the relative contributions of the ICM and NRFU/UAA vacant components to the total variance. Table 1 gives the relative contributions at different geographic levels for the redistricting category "All Persons". ICM blocks, which had 100% nonresponse followup, do not have variance due to NRFU/UAA vacants, and therefore reduce the overall contribution.

Table 1. Relative Variance Contributions in Sacramento for "All Persons" Redistricting Category

Geographic Level	**Median % Var. Due To NRFU/UAA Vacant**	**Median % Variance Due To ICM**
Block (Excluding ICM Blocks)	91.69%	8.31%
Block (All)	86.51%	13.49%
Tract	23.43%	76.57%
Site	0.44%	99.56%

As expected, variance due to NRFU and UAA vacants clearly dominates at the block level, ICM variance is usually the larger percentage at the tract level, and the ICM variance dominates at the site level. These percentages are generally smaller than a similar comparison for the Oakland, CA 1995 Census Test site in Krenzke & Navarro (1996).

4. GENERALIZED ESTIMATION OF VARIANCE

In the 1995 Census Test, the redistricting data consisted of 39 items for each site, 13 race and Hispanic origin categories by three age categories. For the Census 2000 Dress Rehearsal, the number of redistricting items had increased to 86 (43 race and Hispanic origin by two age), primarily due to allowing respondents to indicate multiple responses for the race question. Whereas the generalized variance function for the 1995 Census Test was used to create generalized variance parameters at the site level only, the plan for the Census 2000 Dress Rehearsal was to create separate generalized variance parameters for each of the 86 redistricting items for each site

The generalized variance function used was:

$$V_x^2 = V_y^2 + b\left(\frac{1}{x} - \frac{1}{y}\right)$$

where:

x = estimated redistricting item population
y = estimated site population
V_x^2 = relative variance of x
V_y^2 = relative variance of y
b = estimated regression parameter for the model

and where the relative variance is defined as

$$V_x^2 = \frac{Var(x)}{x^2}$$

This was the model chosen for production out of the seven analyzed in Krenzke and Navarro (1996), and there are some theoretical results that suggest this is a "good" model to use (Wolter, 1986; Valliant, 1987). Specifically, this is the only generalized variance model of this type which has the desirable property that the variance of a proportion is equal to the variance of the complementary proportion (Tomlin, 1974).

It was advantageous to fix the intercept term in the regression to be V_y^2, as this ensured positive variances and forced $V_x^2 = V_y^2$ when x=y. Thus, the equation becomes

$$(V_x^2 - V_y^2) = b\left(\frac{1}{x} - \frac{1}{y}\right)$$

For each redistricting data item, the weighted regression (with the weights proportional to the inverse of the square of the predicted values of V_x^2) was run nine times at the specified geographic level(s). After each run, outliers are identified as those observations with a standardized residual greater than 3, or if

$$ARD(V^2) = \frac{|V^2_{X,\,Direct} - V^2_{X,\,Predicted}|}{V^2_{X,\,Direct}} \geq .50$$

The outliers were then removed, the data reweighted, and the regression re-run.

After the ninth iteration, the **a** and **b** parameters were output, one pair for each redistricting data item. The **b** parameter is the single regression parameter output by the program. The **a** parameter is calculated as

$$a = V_y^2 - \frac{b}{y}$$

so

$$V_x^2 = a + \frac{b}{x}$$

For publication purposes, due to the small size of the parameters, both parameters were multiplied by 1000, and the formulas which the users use to estimate the standard error of a estimated population or proportion were modified accordingly:

$$SE(\hat{x}) = \sqrt{\frac{a\hat{x}^2 + b\hat{x}}{1000}}$$

$$SE(\hat{p}) = \sqrt{\frac{1}{1000}\left(\frac{b}{\hat{y}}\right)(\hat{p}(1 - \hat{p}))}$$

4.1 Menominee

It was planned to use block, block cluster, and tract data, or some combination thereof, to estimate the generalized variance parameters. However, the Menominee site encompassed only one tract, so only block and block cluster data were available.

In testing the generalized variance modeling, two statistics relating to the fit were key in determining which set of geographic levels to use for each set of published

parameters: the median absolute relative deviation (ARD) of the standard errors

$$ARD(SE) = \frac{|SE_{Direct}(x) - SE_{Predicted}(x)|}{SE_{Direct}(x)}$$

and the proportion of the observations for which the relative deviation of the standard errors was greater than zero. This second variable was an indicator of whether the parameters were over- or underestimating the variance.

The initial fit on block data did not work well. Most of the estimated standard errors for larger blocks were much less than the calculated standard errors, and the median ARD(SE) was higher than desired for most of the redistricting categories. For many of the 86 redistricting items, there were a large number of blocks which had only a very small (positive) number of persons in the particular race by Hispanic origin by age category. The model was re-fit using blocks with census counts greater than one, and the median ARD(SE) for the "All Persons" redistricting item was reduced to .2241, a more acceptable value (see Table 2). The percent of RD(SE) greater than zero (indicating overestimates) for "All Persons" was 60.07%, but there were still some problems underestimating the standard errors for the larger blocks.

The model was also fit using block cluster data, also omitting blocks by redistricting items with fewer than two persons. However, the parameters from the block cluster model, when applied to the block level data, gave an obviously unacceptable fit. For "All Persons", the estimated standard error was an overestimate for every block, and the estimated standard error was an overestimate 90% of the time for three-fourths of the redistricting items. It was decided the parameters to be used for publication would be from the block-level data alone.

For 26 of the 86 redistricting categories, there were not enough individuals to be able to estimate the parameters, and these were omitted from the production tables. For many other categories, the population in them was large enough to allow parameters to be estimated, but it was unclear how stable the parameter estimates were.

4.2 Sacramento

For Sacramento, the regression model was fit on tracts, blocks, and tracts & blocks together. Generally, the combined parameters fit the data similarly to the block parameters. While the tract parameters showed better median ARD(SE)'s for the tract data, they fit the block data rather poorly. The combined parameters did slightly better than the block parameters on the tract data, so the combined parameters were the ones selected.

The comparison of the directly estimated and predicted standard errors for the "All Persons" redistricting item for tracts in the Sacramento site and using the combined parameters is shown in Figure 1. The fit, in general, is adequate, but the predicted values begin to underestimate the direct values as the population increases. Correcting this is a key point of the ongoing research. Plots for other redistricting items in Sacramento, as well as plots for Menominee and Sacramento without ICM sampling, produce similar results.

4.3 Sacramento, Excluding ICM Sampling

The Sacramento data with the ICM excluded was also fit on blocks, tracts, and blocks and tracts combined. Results were quite similar to Sacramento including ICM, and the parameters from the combined data were selected to be the ones published.

Table 2. Median ARD(SE) and % RD(SE) > 0, "All Persons" Redistricting Item

Site	Geog. Levels Used	Median ARD(SE)		% RD(SE) > 0	
		Blocks	Tracts	Blocks	Tracts
Menominee	blocks only	.2241	-	60.07%	-
Sacramento, w/ ICM	blocks & tracts	.3444	.1068	62.45%	26.26%
Sacramento, w/o ICM	blocks & tracts	.3419	.1952	61.63%	50.49%

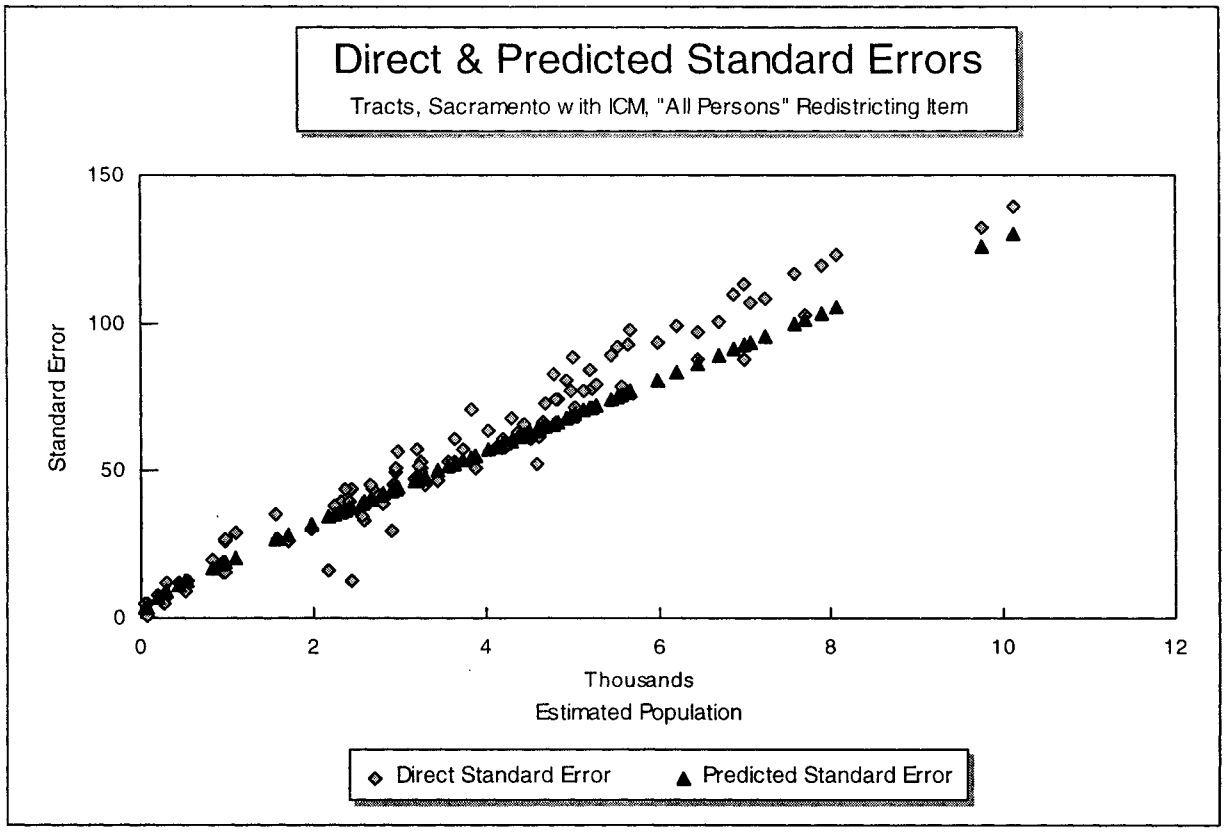

Figure 1. Comparison of Directly Estimated and Predicted Standard Errors
(Tracts, Sacramento with ICM, "All Persons" Redistricting Item)

4.4 Comparison with 1995 Census Test Results

The median ARD(SE)'s for Menominee and Sacramento compare favorably to those from Krenzke and Navarro (1996) for the 1995 Census Test sites of Paterson, NJ (.526 for blocks and .466 for tracts), Oakland, CA (.660 for blocks and .537 for tracts), and six parishes in Northwest Louisiana (.711 for block and .585 for tracts). One factor that likely contributed to the decrease in median ARD(SE)'s was a change in outlier detection during the regression to obtain the parameters. For the 1995 Census Test, the standardized residual cutoff value was 3.5, and ARD(V^2) was not used to detect outliers. For the Census 2000 Dress Rehearsal, a standardized residual cutoff of 3.0 was used, and ARD(V^2) was used to detect outliers.

Also, predicted standard errors (for the "All Persons" redistricting category, which is equivalent to the site-level parameter estimation done in 1995), especially at lower population estimates, were much smaller for the Census 2000 Dress Rehearsal than for the 1995 Census Test. (See Table 3 below.)

Table 3. Predicted Standard Errors for Given Populations, Dress Rehearsal and Census Test

Site	SE's for Given Population				
	10	**100**	**1000**	**10000**	**100000**
Menominee	0.38	2.11	18.48	N/A	N/A
Sacramento, with ICM	1.50	4.88	19.12	128.10	1199.74
Sacramento, without ICM	1.51	4.79	15.14	47.97	154.62
Oakland	4.54	14.48	49.23	238.36	1956.05
Paterson, NJ	7.70	24.40	78.98	301.20	1934.48
NW Louisiana	4.79	15.21	49.86	205.49	1469.44

This is due in part to the decrease in (or absence of) the NRFU/UAA vacant variance, which is the largest contributor to the variance at small geographic (and thus population) levels. The NRFU sampling implemented in the 1995 Census Test had a fixed sampling rate of one-in-three (Treat, 1996), the lowest rate possible under the design for the Census 2000 Dress Rehearsal. Most tracts had a rate much higher than that, leading to a lower NRFU/UAA vacant variance.

5. FUTURE RESEARCH

Several key issues need to be investigated prior to production use of this method for Census2000. Many of these are results of policy and methodology changes from the Census 2000 Dress Rehearsal to Census 2000 itself which have been made or are under consideration. First, and most importantly, sampling for NRFU and UAA Vacant housing units has been eliminated from the plan for Census 2000. Nonrespondents and UAA Vacants will be subject to 100% follow-up, thus eliminating one component of the variance. The generalized variance function needs to be re-fit for Sacramento using just the ICM component of the variance to see how well the model holds, considering the less-than-ideal fit of the model to the Menominee data.

In Census 2000, parameters derived from the model will be calculated at the state level (or possibly higher) and must perform adequately for counties, congressional districts, metropolitan areas and other groupings intermediate in size between tracts and the state itself. The present implementations of the generalized variance model in the 1995 Census Test and the Census 2000 Dress Rehearsal were on relatively small areas. The sites generally did not contain more than one county or more than one complete congressional district. It is unclear how well a model fit with only smaller geographic areas such as blocks and tracts would work at estimating the variances of the larger areas. Data from the Post-Enumeration Survey of the 1990 Census will be used to provide insight into this potential problem.

These and other issues will be investigated over the coming months to allow improvements in the production of the generalized variances for Census 2000.

6. REFERENCES

Fay, R. E., and Town, M. K. (1998), "Variance Estimation for the 1998 Census Dress Rehearsal", Proceedings of the Section on Survey Research Methods, American Statistical Association.

Krenzke, T., and Navarro, A.(1996), "Sampling Error Estimation in the 1995 Census Test for Small Areas", Proceedings of the Section on Survey Research Methods, American Statistical Association.

Tomlin, P. (1974), "Justification of the Functional Form of the GATT Curve and Uniqueness of Parameters for the Numerator and Denominator of Proportions", unpublished memorandum, U.S. Bureau of the Census.

Treat, J. (1996), "Analysis Comparing the NRFU Block Sample and NRFU Unit Sample Nonresponse Follow-Up Evaluation", 1995 Census Test Results Memorandum #31, U.S. Bureau of the Census

Valliant, R. (1987), "Generalized Variance Functions in Stratified Two-Stage Sampling", Journal of the American Statistical Association, 82, 398, 499-508.

Wolter, K. (1985), Introduction to Variance Estimation, New York, Springer-Verlag.