

GENERALIZED VARIANCE MODELING FOR THE CENSUS 2000 A.C.E.

Michael D. Starsinic, Charles D. Sissel, U.S. Census Bureau[†]
Michael D. Starsinic, U.S. Bureau of the Census, Washington, DC 20233

Key Words: Generalized Variance, Census, Small Area Estimation

1. INTRODUCTION

It is Census Bureau policy to release measures of error for estimates in all its data products. For Census 2000, that included being prepared to release estimates of sampling error for the 100-percent census data as adjusted by the results of the Accuracy and Coverage Evaluation (A.C.E.). The vast number of small area population estimates produced for Census 2000 make it impractical to provide a direct standard error for each estimate. Instead, generalized variance parameters are made available for certain demographic and geographic characteristics so users can approximate the standard error of any desired A.C.E.-adjusted estimate. Computing a generalized variance model also eases the problem of instability associated with estimating standard errors for very small populations. Initially, we planned to use the 1998 Census 2000 Dress Rehearsal generalized variance methodology (Starsinic & Town, 1999) to produce Census 2000 estimates. However, questions arose about these initial weighted least squares regression models, necessitating a change in approach to a generalized coefficient of variation methodology. This paper analyzes the results of the generalized variance modeling for Census 2000. Section 2 provides a brief overview of the sampling, estimation, and variance estimation for the A.C.E. Section 3 gives the original and final generalized variance methodologies considered, and Sections 4 and 5 analyze results from several alternative methodologies.

2. A.C.E. SAMPLING, ESTIMATION, AND VARIANCE ESTIMATION

The A.C.E. was intended to estimate the net coverage of Census 2000, taking into account persons enumerated erroneously or more than once and persons missed by the census enumeration. The estimates were produced using

a post-stratified, dual-system estimator, based on a sample of approximately 310,000 housing units.

An initial sample of 29,136 clusters (collections of contiguous census blocks) was selected, but this was reduced through further sampling to a final size of 11,303 clusters to meet a target for the total number of sampled housing units. An enumeration of these clusters based on an address list independent of the census was called the P Sample, and the E Sample was formed essentially from the census results for the same clusters. These two samples were used to identify E-Sample individuals as correct or erroneous enumerations and P-Sample individuals as matches or nonmatches to census persons. The matching and identification results were used to produce dual-system estimates at the post-stratum level, where 448 person-level post-strata were defined using a combination of demographic and geographic characteristics.

A multi-phase jackknife variance estimator was developed and implemented, which directly computed variances at the post-stratum level. The sample, although large, cannot support direct variance estimates at small geographic levels. For consistency, all other variances - including those for geographic areas used in the computation of the generalized variance parameters - were computed synthetically using post-stratum variances and covariances.

For more information on these operations, see ZuWallack et al. (2000), Haines (2001), and Navarro & Sands (2001).

3. CENSUS 2000 GENERALIZED VARIANCE METHODOLOGY

3.1 Why We Need Generalized Variances

The first Census 2000 release for detailed geographic areas was the Public Law (PL) 94-171 data. These are census block-level counts which are used by states to redefine legislative district boundaries ("redistricting"). The most detailed sets of data are 286 overlapping combinations of race, Hispanic origin, and age, which we will call redistricting (or PL) categories. For each state, we prepared generalized variance parameters for 62 of the

[†]This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

286 largest groups, with four additional “catch-all” categories for the redistricting categories not modeled separately.

3.2 Initial Methodology - Generalized Variance Functions

As mentioned above, the Census 2000 Dress Rehearsal used a weighted least squares regression generalized variance function methodology for its published generalized variance parameters. The parameters were published for 86 groups used in the Dress Rehearsal as a preliminary prototype for the Census 2000 redistricting categories.

The form of this initial generalized variance function (GVF) was:

$$V_x^2 = V_y^2 + b \left(\frac{1}{x} - \frac{1}{y} \right) \quad (3.1)$$

where:

- x = estimated population for redistricting group in small geographic area
- y = estimated population for redistricting group in entire geographic area
- V_x^2 = relative variance of x, defined as $\frac{\text{Var}(x)}{x^2}$
- V_y^2 = relative variance of y
- b = estimated regression parameter for the model

For our regression models, we fixed the intercept at V_y^2 to force $V_x^2 = V_y^2$ when $x=y$ (i.e., when the redistricting group = the entire geographic area). We then ran nine iterations of our weighted regression model (with the weights equal to $(1/V_x^2)^2$) to obtain the “b” regression parameter and set

$$a = V_y^2 - \frac{b}{y} \quad (3.2)$$

With an appropriate pair of parameters, users could approximate the standard error of an estimate, \hat{x} , as

$$SE(\hat{x}) = \sqrt{a\hat{x}^2 + b\hat{x}} \quad (3.3)$$

Ideally, \hat{x} would exclude the out-of-scope A.C.E. population (specifically, persons living in group quarters and those enumerated in the Remote Alaska operation), as they do not contribute sampling error. However, the majority of published estimates, including the first-released PL 94-171 counts, are not broken down by in- and out-of-scope. Since a user only sees the estimates including the out-of-scope population, the A.C.E. out-of-scope population is used in the modeling.

This GVF methodology has some theoretical justifications and advantages (Wolter 1985, Valliant 1987, and Tomlin 1974), and appears to have worked well in the Census 2000 Dress Rehearsal.

We were unable to generalize our results to larger geographies such as places, counties, and congressional districts using the Dress Rehearsal data alone because of the small size of the Dress Rehearsal sites. Instead, we used variances from a simulation designed to approximate the 2000 variances (Asiala 2001) based on known A.C.E. sample sizes and weights combined with data from the 1990 Post-Enumeration Survey (PES), a coverage survey similar in many ways to the A.C.E. Including variance estimates from places, counties, and congressional districts with variance estimates for census tracts seemed to improve the fit. Omitting the tract-level variances and using only the variance estimates from the larger geographic areas produced parameters which gave poor approximations to areas with small populations.

However, a serious problem arose with the results from the simulated data. With the simulated data, we produced parameters for 26 redistricting categories (the others dealt with persons identifying with more than one race group, which was not possible in the 1990 census) for 51 states, including the District of Columbia but excluding Puerto Rico, for a total of 1,326 sets of parameters. Of these, 69 parameter pairs (about five percent) had a negative value of “b”, the regression parameter. It can be seen from Equation 3.3 that

$$\text{Var}(\hat{x}) < 0 \text{ if } \hat{x} < \frac{-b}{a} \quad (3.4)$$

As long as both parameters are positive - and in our results the “a” parameter was always positive - the variance estimate can never be negative. However, with negative values for “b”, these 69 specific problem sets of parameters yielded negative variance estimates for population estimates less than $-b/a$. This is obviously not a desirable result.

It was unknown whether these occurrences were aberrations caused by something specific to the simulated data, or whether they could occur with the actual Census 2000 data. Considering the time pressures we would be under in producing the parameters, there was no way we could take the risk of producing bad parameter estimates, and so this model was abandoned. Alternative models, one of which is discussed in Section 5 below, eliminated the negative “b” cases and even gave improved empirical fits, but their theoretical justifications were on much less stable ground. Clearly, we needed to find an alternative generalized variance methodology.

3.3 Final Methodology - Generalized Coefficients of Variation

When the detailed 1990 PES redistricting estimates were released in 1998, they were accompanied by generalized variance tables using a generalized coefficient of variation (GCV) methodology. This was the approach we used in computing the Census 2000 GCV parameters.

The GCV parameter estimation process worked identically in each of the state by redistricting category cells. The coefficient of variation (CV) was calculated for all tracts in a state with a nonzero population in the particular redistricting category. Tracts composed entirely of persons out-of-scope for the A.C.E. sample had no sampling variance (and therefore a CV of 0) and were removed from the processing. Also removed were tracts with a very small population in the redistricting category, as these were shown in the Dress Rehearsal analysis to have a disproportionate downward effect on the parameters. A two-tiered cutoff was employed to prevent removing an overly-large fraction of “small” tracts from a given redistricting category.

The CVs of the remaining tracts were averaged within redistricting category to produce an initial GCV. Outliers were identified using the relative absolute deviation (RAD)

$$RAD = \frac{|tract\ population \times GCV - SE_{tract}|}{SE_{tract}} \quad (3.5)$$

Tracts with an RAD above the cutoff were removed, and a new GCV was computed using the CVs of the remaining tracts. This cycle of identifying and removing outliers and recalculating the GCV went through four iterations. The value of the GCV after the fourth iteration is the production value for that state by category combination. Using one (appropriate) GCV parameter, users could calculate an approximate standard error of any estimate as

$$SE(\hat{x}) = \hat{x} \times GCV \quad (3.6)$$

The GCV methodology was quick and easy to implement. Moreover, it was less burdensome on the user than the earlier proposed GVF methodology.

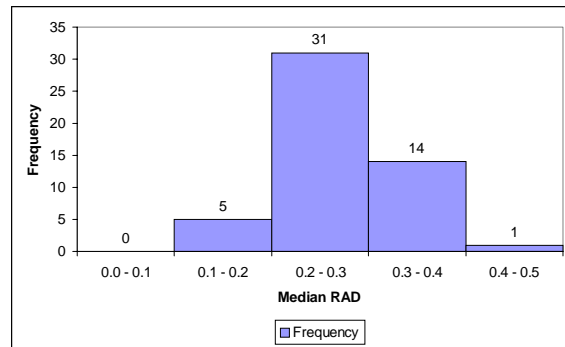
4. GENERALIZED VARIANCE RESULTS

Overall, the fit of the GCV model to the synthetic standard errors is good, and where the fit wasn't as good as we would have liked, the GCV generally produced an

overestimate of the variance, as opposed to the underestimation seen at times in the Dress Rehearsal GVF model (Starsinic & Town, 1999).

Figure 1 shows a typical distribution of the state-level values of the median RAD, specifically for the “Black Alone” redistricting category for all tracts. The state-level values were computed by applying the state “Black Alone” GCVs to the “Black Alone” population in each tract to produce approximate standard errors, and these were used to produce tract-level RADs. These (and following) RAD estimates include all tracts, including those removed as outliers during the GCV calculation.

Figure 1: Median RAD for States - All Tracts, “Black Alone” Redistricting Category



The GCVs also performed well at larger geographic levels. Figure 2 shows the distribution of median RADs for counties, which were not included in the modeling process. Again, as with tracts, when the GCV is off, it is usually too high.

Figure 2: Median RAD for States - All Counties, “All Persons” Redistricting Category

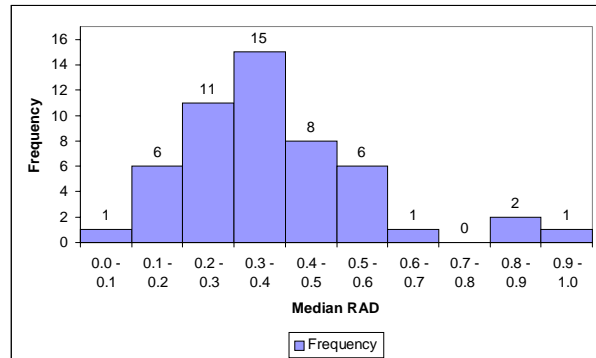
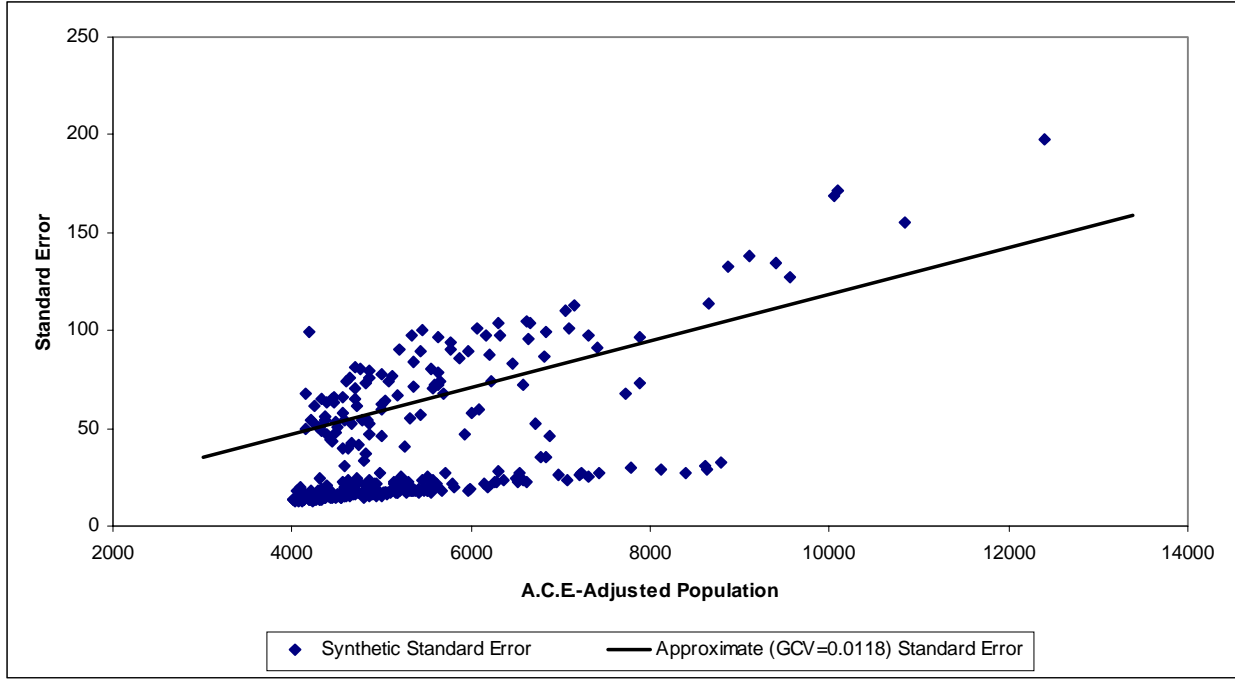


Figure 3: State “X” - Tracts Over 4,000 Population, “White Alone” Redistricting Category



One phenomenon that was first observed during the simulation work also occurred in the Census 2000 data. For some states and some redistricting categories, the tract-level synthetic standard errors fall into two or more bands, so that no single line (GCV or GVF) could describe the distribution well. Figure 3 shows the distribution of tracts with a population of at least 4,000 in the “white alone” redistricting category in a specific state. (All tracts were used to estimate the GCV parameter. The cutoff of 4,000 is used to emphasize the two distinct groups of tracts.) Two separate bands of observations are fairly well-defined, with the GCV value producing approximate standard errors (the line on the graph) which more closely follow the upper band.

What is causing this? It could be the interplay between the post-stratum definitions and the redistricting category definitions. The 448 post-strata were based on combinations of race, Hispanic origin, age, sex, tenure, size of metropolitan area, type of census enumeration, tract-level return rate, and census region. Persons of many different post-strata can be included in any redistricting category. When a redistricting category in a state contains groups of people from post-strata with markedly different variance characteristics, this “banding” can result. In Figure 3, the tracts in the upper grouping contain people that come from post-strata with higher variances than the post-strata making up the population in the tracts in the lower grouping. Fitting a simple variance function to a split distribution like this

would be difficult for any generalized variance methodology.

Ideally, the redistricting category definitions and the post-strata should more closely parallel one another. That is unlikely to happen, though, as post-stratification is a statistical process and defining redistricting categories for publication is largely a political one.

5. RESULTS OF OTHER METHODOLOGIES

After production was finished, we tested several other GVF regression models on real census data. After examining several, the GCV method seems to have performed better than any of the GVF models.

The most promising new GVF model was

$$\ln(V_x^2) = \ln(V_y^2) + b(x - y) \quad (5.1)$$

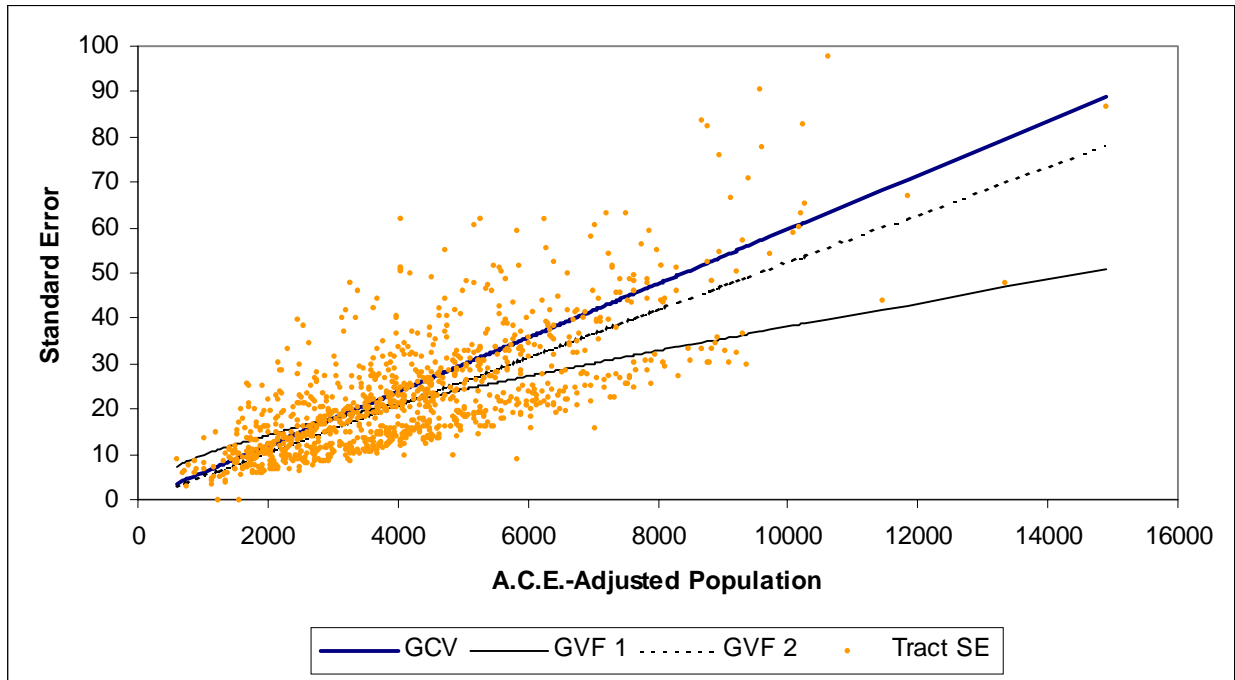
with $\ln(V_y^2)$ fixed as the intercept. After obtaining the “b” regression parameter using iterated weighted regression, we set the second parameter to

$$a = \ln(V_y^2) - by \quad (5.2)$$

The approximate standard error for an estimate, \hat{x} , is:

$$SE(\hat{x}) = \hat{x} e^{\frac{1}{2}(a + b\hat{x})} \quad (5.3)$$

Figure 4: Comparison of GCV, GVF 1, and GVF 2 Fits - State “Y”, “All Persons” Redistricting Category



Note that the variance (and hence standard error) of an estimate cannot be negative in this model. Let the model just described be GVF 2, and the original model described in Section 3.2 (Equation 3.1) be GVF 1.

In the majority of fits to the census data, GVF 2 outperformed GVF 1, based on median RADs and graphical comparisons of standard errors. Both were, however, inferior to the production GCV method.

Figure 4 shows the distribution of the standard errors for redistricting category 1 (“all persons”) for tracts in a certain state. This pattern is typical across most states and redistricting categories. For the smallest tracts, the GCV and GVF 2 models give almost identical approximate standard errors, while GVF 1 gives much higher standard errors. At about a population of 3,000, GVF 1 falls below GCV, and at about 4,000, GVF 1 falls below GVF 2. Its plot continues to flatten out as the population increases, and diverges from GCV. For the largest tract in Figure 4, GVF 1 is about 43 percent lower than GCV. GVF 2 gives nearly a straight line, although the slope of the curve is very slowly decreasing; both the “a” and “b” parameters here are naturally negative, and the function will eventually tend towards a slope of zero. The graph of GVF 2 is generally closer to GCV than to GVF 1, but it offers no real improvement on the fit of the GCV. GCV offers users a much simpler formula for computing the approximate standard error than GVF 1 or GVF 2.

6. FUTURE RESEARCH

One avenue of research is reflected in Section 5 above, involving a search for a theoretically sound regression GVF methodology that would give consistently good approximations to users and would not be vulnerable to the negative variance problem.

A second approach would be to further divide each state \times redistricting category into groups by size of area, and compute separate GCVs for each group. This was the methodology used with the published PES GCVs which used four divisions: less than 5,000, 5,000-25,000, 25,000-100,000, and greater than 100,000. The fit may be improved, but it is not clear what the optimal number of population-size groups should be (or even how to determine the optimal number), or where to make the divisions between each size class. This method also could introduce large discontinuities at the division between two adjacent size groups, where the difference of just one person would mean a large difference in the approximated standard errors. Still, if further subdividing each category results in greatly improved fits, this approach would have to be considered.

A third area for further research is to examine why the original GVF methodology - Equation 3.1 - yielded the negative “b” parameters. Using the original methodology on the Census 2000 data, there were 12 instances of

negative values for “b”. This may be related to the relationship between the redistricting categories and the post-stratification definitions described in Section 4.

As of this writing, A.C.E.-adjusted estimates have not been publicly released, and it is not known when or if they will be. A set of generalized variance parameters must be ready if the decision to publish is made. For now, it is this set of GCV parameters we have just discussed. But it does not necessarily need to remain this set, if a superior parameter set can be created.

7. REFERENCES

Asiala, M. (2001), “Comparison of Estimates of Small-Area Variances”, 2001 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Haines, D. (2001), “Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Dual System Estimation (U.S.)”, DSSD Census 2000 Procedures and Operations Memorandum Series Q-48, March 12, 2001.

Navarro, A., and Sands, R. (2001), “2000 Census A.C.E. Variance Estimates”, 2001 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Starsinic, M., and Town, M. K. (1999), “Analysis of Generalized Variance Estimation for the Census 2000 Dress Rehearsal”, 1999 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.

Valliant, R. (1987), “Generalized Variance Functions in Stratified Two-Stage Sampling”, *Journal of the American Statistical Association*, Vol. 82, No. 398, p. 499-508.

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

ZuWallack, R., Salganik, M., Cromar, R., and Mule, V. (2000), “Final Sample Design for the Census 2000 Accuracy and Coverage Evaluation”, 2000 Proceedings of the Section on Survey Research Methods, American Statistical Association.