GENERALIZED VARIANCE ESTIMATES DUE TO ADJUSTMENT OF THE 1990 CENSUS

Richard A. Griffin, Alfredo Navarro, and Lawrence Bates
Bureau of the Census

I. INTRODUCTION

The 1990 Census population counts may be adjusted based upon the results of the 1990 Post Enumeration Survey (PES). If the decision to adjust the census counts is made, the Census Bureau must have a system in place to produce reliability measures for the adjusted figures. The results of the 1990 Census Enumeration and the 1990 PES are used to measure census coverage and to produce adjustment factors (AF). The adjustment factor is a function of the census count and the dual system estimate (DSE) [1,2,3]. The DSEs are subject to nonsampling and sampling error as well. The adjusted census counts, the census counts inflated by the AFs, will be published in standard data products. These data are published for many geographic areas and demographics. It will not be feasible to publish a reliability measure for each adjusted count (estimate) and tabulation area.

The purpose of this research is to formulate a generalized variance strategy for the adjustment option.

This investigation is based on results for the 1988 Dress Rehearsal PES in Missouri.

II. DUAL SYSTEM ESTIMATION AND POSTSTRATIFICATION

A. Dual System Estimate

To get estimates of the total population, a dual system estimator is used. A typical DSE used by the census [2] is written

$$\text{DSE} = \frac{N_p\,(\text{CEN}-\text{SUB}-\text{EE})}{M} \qquad (1)$$

where

N_p = PES population estimate
CEN = unadjusted Census count
SUB = number of census whole-person substitutions
EE = estimate of the number of erroneous enumerations

and

M = estimator of the number of persons in both the census and the PES populations.

B. Poststratification

The 1988 Dress Rehearsal population and 1988 PES population were poststratified according to the following scheme. Each poststratum is thought to be homogenous with respect to the Census undercount mechanism.

Stratum	Description
1	White Non Hispanic nonowners in St. Louis
2	All other nonowners in St. Louis
3	White Non Hispanic owner in St. Louis
4	All other owners in St. Louis
5	White Non Hispanic person in Tape Address Register (TAR) areas in East Central MO
6	White Non Hispanic persons in non TAR in East Central MO
7	All other persons in East Central MO

These seven basic poststrata were further stratified by twelve age by sex categories. This poststratification is designed to reduce the bias of the DSE, which could be significant due to differential undercount.

Based on results from the PES, a DSE is calculated for each of the poststrata.

III. COUNT ADJUSTMENT

A. Adjustment Factors

The initial adjustment factor is defined as the ratio of the DSE, as described in Section II. A, to the unadjusted census count

$Y_i = DSE_i/CEN_i$, where
i denotes the poststratum. (2)

To reduce variability, the adjustment factors are then "smoothed" through a regression model before adjusting the block level counts. Adjustment factors are used as the dependent variable in the regression model.

The regression model is written as where

$$Y_i = B_o + B_1X_{i1} + \ldots. + B_p X_{ip} + S_i + E_i$$

Y_i = adjustment factor for the ith poststratum,

X_{ij} = independent variable $(j=1,\ldots..,p)$,

B_j = regression coefficient to be estimated,

S_i = sampling error of the adjustment factor, and

E_i = model error.

The possible independent variables were [4]

X_1 = indicator variable for St. Louis

X_2 = indicator variable for White Non Hispanic

X_3 = indicator variable for proportion Tape Address Register (TAR)

X_4 = indicator variable for proportion nonowner

X_5 = indicator variable for sex

X_6-X_{10} = indicator variables for age groups

X_{11} = proportion of cases substituted

A number of regression models using a subset of these independent variables were examined and the best for our purposes was selected.

A weighted average of the sample-based and the model-based adjustment factors is defined as

$$AF_i = (Y_i/\sigma^2{}_i + \sum_j X_{ij} \hat{B}_j / \epsilon^2)(\sigma_i{}^{-2} + \epsilon^{-2})^{-1}, \quad (3)$$

where $\sigma^2{}_i$ is the sampling variance of the adjustment factor in stratum i and ϵ^2 is the model error variance (constant across stratum).

AF_i is ultimately used to adjust the census block counts.

B. Estimation

Synthetic estimation is used to produce poststrata estimates down to the block level. The estimator is written as

$$\hat{N}_{(ij)} = AF_i * CEN_{ij} \quad (4)$$

where

$\hat{N}_{(ij)}$ = Adjusted item count for the i-th poststratum in the j-th block.

AF_i = Adjustment factor of the i-th poststratum [See (3)]

CEN_{ij} = Unadjusted census item count for the i-th poststratum in the j-th block.

$\hat{N}$ is usually a noninteger number. An intergerization mechanism is used to transform the noninteger values to integers (whole persons are enumerated) that has no detrimental effect in the use and quality of the adjusted figures.

The estimate of the total of an item at the block level is written

$$\hat{N}_{(j)} = \sum_{i=1}^{M} AF_i \; CEN_{ij} \quad (5)$$

where M is the number of poststrata.

To get estimates of the total of an item for higher geographic areas the adjusted block counts are added as necessary.

C. Variance Estimation

The variance of (4), the poststratum estimate of the total of an item is written

$$\text{Var}\ (\hat{N}_{(ij)}) = \text{Var}(AF_i{*}CEN_{ij}) \quad (6)$$

CEN_{ij} is not subject to sampling variability, therefore

$$\text{Var}(\hat{N}_{(ij)}) = CEN_{ij}{}^2\text{Var}(AF_i) \quad (7)$$

The variance of AF_i is obtained from the results in Freedman and Navidi [5]. The Undercount Research Staff of the Statistical Research Division has produced the variance-covariance matrix for the AF_i's.

In general, the variance for an estimate of the total of an item for a tabulation area, K, is written

$$\text{Var}(\hat{N}_{(K)}) = \sum_{i=1}^{M} CEN_{i(k)}^{2}\text{Var}(AF_i) + \sum_{i=1}^{M}\sum_{i\neq j}\text{COV}(AF_i,AF_j)CEN_{i(k)}CEN_{j(k)} \quad (8)$$

IV. GENERALIZATION SCHEMES

For each tract/Block Numbering Area (BNA) and selected data item, the coeffiecient of variation (CV) of the adjusted item count will be calculated. In general, the coefficient of variation is written:

$$CV(\hat{N}_{(lK)}) = \sqrt{\text{Var}\ [\hat{N}_{(lk)}]}/\hat{N}_{(lk)} \quad (9)$$

for item l.

For each data item the CV's will be (simple) averaged across tracts and BNA's within the state (for this empirical study, just tracts/BNA's in St. Louis and East Central Missouri).

$$CV(\hat{N}_{(l)}) = \sum_{k=1}^{n_k} CV[\hat{N}_{(lk)}]/n_k \quad (10)$$

(n_k = number of tracts/BNA's within the state)

Finally, a (weighted) average state CV will be calculated. The generalized CV is written

$$CV(\hat{N}) \ = \sum_{l=1}^{G} CV\ [\hat{N}_{(l)}]\ \frac{T_l}{T_{ls}}$$

where

$$T_l = \sum_{k=1}^{n_k} CEN_{lk} \text{ and } T_{ls} = \sum_{l=1}^{G}\ \sum_{k=1}^{n_k} CEN_{lk}$$

G is the number of items grouped together to be represented by one published CV.

Five methods of grouping items will be examined. (See Section VI).

V. ITEMS

In order to simulate each of the five grouping methods (Section VI) and in order to determine how efficient each method is at predicting actual standard errors, formula (8) was used to calculate variances at the tract and BNA levels for 457 data items. Twenty of these items are the Public Law (PL) 94-171 items defined by the five major races (White, Black, American Indian, Eskimo or Aleut, Asian and Pacific Islander, other) Hispanic or Non-Hispanic and age (less than 18 or 18+). There are 372 items contained in only one PES poststratum. Of these, 84 items are the actual poststrata, see II.B. The remaining 288 items of these 372 items are defined by race/origin group (5 major races plus Hispanic) by tenure crossed by sex and age in St. Louis and by race/origin group by type of enumeration area crossed by sex and age in East Central Missouri. The other 65 items are included in a combination of PES poststrata. These are as follows.

A. Each of the seven basic poststrata by sex summed across age categories (14 items).

B. Each of the seven basic poststrata summed across sex and age categories (7 items).

C. Each race or origin crossed by sex and tenure (24 items).

D. Each race or origin crossed by sex (12 items).

E. Each race or origin (6 items).

F. Total population.

G. Total count adjustment population.

In addition, the variances of the 20 PL items were calculated at the block group and block levels. Also the variances of a subset of the other items were calculated at the block group level.

VI. GROUPING METHODS[1]

A. Method 1

For each of the 12 race or origin by sex categories, group appropriate items from the 372 items contained in only one PES poststratum. For example, Black/male would consist of the 12 tenure by age items for black males in St. Louis and the 12 type of enumeration area by age items for black males in East Central Missouri. These 12 generalized CV's (weighted averages across items in a group) are published. In the case of a data item logically represented by more than one CV, the higher one is used.

B. Method 2

For each of the 12 race or origin by tenure categories, group the appropriate four items described in Section V.A. For example, Black/owner would consist of all other owner males in St. Louis, all other owner females in St. Louis, all other males in East Central Missouri and all other females in East Central Missouri. These 12 generalized CV's are published and the highest logical one is used.

C. Method 2A

For each of the 12 race or origin by tenure categories, group the appropriate items from the items contained in only one PES poststratum. For example, Black/owner would consist of the 12 sex by age items for black owners in St. Louis and the 12 sex by age items for black owners in East Central Missouri. These 12 generalized CV's are published and the highest logical one is used.

D. Method 3

Each of the 7 basic poststrata defines a group by itself. The simple average of the CV's for each of the 7 items across tracts is published. The highest logical CV is used for any estimated data item. For example, the estimated number of blacks in St. Louis would use the higher of the all other nonowner in St. Louis or all other owner in St. Louis published CV's.

E. Method 4

Each of the 24 race or origin by tenure by sex items defines a group by itself. The simple average of each of the CV's for each of the 24 items across tracts is published. The highest logical CV is used for any estimated data item. For example, the estimated number of black males would use the higher of the black male owner and black male non-owner published CV's.

VII. EMPIRICAL STUDY

The generalized CV's using each of the five grouping methods were calculated. For each method, for a given data item, this generalized CV ($CV[\hat{N}_{(1)}]$) was multiplied by the geographic area adjusted item count ($\hat{N}_{(1)}$) to obtain the generalized standard error. The actual standard errors were compared to the generalized standard errors for each of the five methods based on the mean, median, and maximum of the absolute relative difference across geographic areas. The PL data items were done for all tracts, block groups and blocks in St. Louis and East Central Missouri. (For median and maximum the data given here is for a 10 percent sample of blocks). The 65 items that are included in a combination of PES poststrata (defined in V.A-G) were done for all tracts and block groups in St. Louis and East Central Missouri. The weighted average mean, simple average median, and simple average maximum for a given geographic area type (tract, block group or block) were calculated across items. Only non-zero areas for a given item were included. The absolute relative difference was defined as

$$\frac{|SE[\hat{N}_{(1)}]-SE_G[\hat{N}_{(1)}]|}{SE_G[\hat{N}_{(1)}]}$$

where
$SE_G[\hat{N}_{(1)}] = CV[\hat{N}_{(1)}]*\hat{N}_{(1)} =$ Generalized Standard Error.

$SE[\hat{N}_{(1)}]$ = Actual Standard Error. Results are shown in Table 1 for the weighted average mean.

VIII. CONCLUSIONS AND LIMITATIONS

Generalization method 4 is clearly superior to the other methods in terms of the statistics calculated in this empirical study. The average across data items of the mean relative absolute difference was lowest at all geographic levels (tract, block group and block) for method 4. For the average across data items median and maximum relative absolute difference, method 4 was either the lowest or very close to the lowest for each geographic level.

The relative errors of method 4 compare favorably with those that we observed from the census sample data generalized variance methodology. Results from a 1980 Census empirical study on various methods to generalize 1980 Census variance estimates provide data that can be compared with the results of this empirical study for method 4[6]. The following table shows this comparison.

Average Mean, Median and Maximum Absolute Relative Difference between Predicted and Actual Standard Error.

	Selected 1980 Method- Weighting Area level (71 Items)	Method 4- Tract level (85 items)
Mean	.316	.223
Median	.253	.209
Max.	1.883	.647

There are limitations in using the results of this empirical study to make decisions for the 1990 Census. The poststratification scheme for the 1990 Census is different than for the 1988 Dress Rehearsal[1]. The major difference is in the race/origin categories. For 1990, they are black, non-black and Hispanic, and all other (non-black and non-Hispanic). For the Dress Rehearsal, they were White Non-Hispanic and all other. The sex/age poststratification schemes are the same and the geographic and tenure categories are similar. Also, the averaging across tract in the empirical study is for only one state and only part of that state (St. Louis and East Central Missouri sites). There are also very few Hispanics in the dress rehearsal sites. With these limitations in mind, we need to make a decision for the 1990 Census based on the best available information. The empirical study shows that averaging and publishing the coefficient of variation (CV) of the estimate for each of the 24 race or origin by tenure by sex items across tracts produces reasonable generalized standard errors for other items. At this point in time, we are recommending that each of these 24 average tract/BNA level CV's be published for each state for the 1990 Census if we have count adjustment. A data user will be instructed to multiply the highest logical generalized CV times the published adjusted count to produce a generalized standard error.

FOOTNOTES

[1] For all grouping methods, total population and total count adjustment population are each a group by themselves. The simple average of the CV's of both of these items are published. Thus, there is no difference between the methods for these two items.

References

1. Woltman, H., Alberti, N., and Moriarity, C., Sample Design for the 1990 Census Post Enumeration Survey, ASA, 1988.
2. Diffendal, Gregg, The 1986 Test of Adjustment Related Operations in Central Los Angeles County, Survey Method., June 1988, Vol. 14, No. 1, pp. 71-86.
3. Mulry, Mary H. and Spencer, Bruce D., Total Error in PES Estimates of Population: The Dress Rehearsal Census of 1988, ARC, 1990.
4. Isaki, C., Small Area Estimation Summary - 1988 Dress Rehearsal, Unpublished document, Bureau of the Census, 1989.
5. Freedman, D. A. and Navidi, W. C., Regression Models for Adjusting the 1980 Census, Statistical Science, 1986, Vol. 1, No. 1,. pp. 3-39.
6. Fan, Milton C., Preliminary Summary of Results from a Comparison of Methods to Present 1980 Census Variance Estimates, 1980 Census Preliminary Evaluation Results Memorandum No. 54, 1983.

Table 1 Weighted Average Mean Relative Absolute Difference Public Law Data (20 Items)

Area Method	Tract	Block Group	Block
1	0.310	0.290	0.327
2	0.252	0.293	0.329
2A	0.280	0.324	0.340
3	0.412	0.454	0.384
4	0.189	0.187	0.244

Other Data (65 Items)

Area Method	Tract	Block Group
1	0.394	0.375
2	0.235	0.286
2A	0.241	0.297
3	0.247	0.317
4	0.234	0.250