BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: Census/SRD/RR-87/15

REPORT 4
CENSUS RATIO ADJUSTMENT FOR SYNTHETIC ESTIMATION

by

Beverley Causey
Statistical Research Division
Bureau of the Census
Room 3134, F.O.B. #4
Washington, D.C. 20233 U.S.A.

Recommended by: Kirk W. Wolter

Report Completed: June 3, 1987

Report issued: June 4, 1987

DRAFT 5/18/87
B. Causey:jo:bev4

Report 4
Census Ratio Adjustment for Synthetic Estimation

Report 1 considered across-the-board ratio adjustment of area census counts using a known population total T; Report 2 considered the same adjustment using an estimator, $\hat{T}$, of T. In this report suppose that, once again, we have areas (typically, states) indexed i=1,...,I. Suppose, also, that we categorize each person in a second manner, e.g., according to J age-race-sex cells as in our empirical results below. We number the cells from j=1 to J. Let T_j be the true total number of persons, summed over the I areas, that fall into cell j. In the manner of Report 1 suppose that we know T_j, from a PES or similar source, for each of our cells j.

With this knowledge of T_j, our "synthetic" estimation of the population total for area i proceeds as follows. Let y_{ij} be the census count of number of persons in area i and cell j; let Y_j be the census count of all persons in cell j; let p_{ij} be the proportion y_{ij}/Y_j. Our synthetic estimator for area i is

$$a_i = \sum_j y_{ij} T_j / Y_j = \sum_j p_{ij} T_j. \qquad (1)$$

Let y_i be the census count for area i, and $\underline{y}$ be, as in earlier reports, the vector of area census counts. As in (2) of Report 1 we once again want to compare the loss-function values $f_k(\underline{a})$ and $f_k(\underline{y})$ to decide whether adjustment is worthwhile. Thus as in Report 1 we compute the ratios $f_k(\underline{a})/f_k(\underline{y})$. A ratio far less than 1 suggests that synthetic estimation is worthwhile.

Next, in the manner of Report 2, we confront the fact that we do not know T_j exactly, but only have an estimator of it, which we call $\hat{T}_j$. Suppose that we view $\hat{T}_j$ as an unbiased estimator of T_j with a variance W_j. Also, the estimators $\hat{T}_j$ are mutually uncorrelated random variables: there is no correlation

among the J-departures of observed $\hat{T}_j$ from T_j. (Later in this paper we will consider positive correlation among these estimators.) Thus we envision a procedure for constructing $\hat{T}_j$ such that any large, uniform systematic upward or downward bias, of the $\hat{T}_j$'s as estimators of T_j's, has been removed. Perhaps one can achieve this by correcting cell totals to a known overall total. This premise may need to be scrutinized.

In the manner of Report 2 we want to consider a "breakeven" variance, which tells us how precisely we need to know $\hat{T}_j$ in order to make adjustment worthwhile. But whereas in Report 2 we had only a single variance W, we now have J variances W_j. Thus in this report we proceed as follows, in an effort to develop a meaningful statement as to breakeven variance.

Suppose that W_j is proportional to T_j^c : that is, T_j raised to the "c" power with the value of c independent of j. A value c=0 says that W_j itself is a constant, independent of j. A constant value for W_j, and thus a value c=0, do not seem plausible - because the uncertainty associated with $\hat{T}_j$ naturally tends to be largest, in <u>absolute</u> terms, when T_j is largest. That is, for example, for a cell of size 10, known to be very small, we will estimate T_j with an error of very small magnitude, say 10 or 15 persons; but for a cell of size 1 million the magnitude of our absolute error in estimation will be much more than 10 or 15 persons. A value c=2 says that W_j/T_j^2 is constant: that is, our c.v. is the same for all j. The value c=2 thus does not seem plausible, as the uncertainty associated with $\hat{T}_j$ naturally tends to be largest, in <u>relative</u> terms, when T_j is smallest. As an intermediary between c=0 and 2, we use c=1: variance is proportional to T_j. (This last pattern corresponds, for example, to the fact that the sum of n independent and identically distributed random variables has, along with an expected value proportional to n, a variance which, also, is proportional to n rather than to 1 or n^2.)

Thus we consider a general model of form

$$\mathrm{Var}(\hat{T}_j) = VT_j^c \qquad (2)$$

with values of both the power c and the coefficient V independent of j. We prefer c=1 and show calculations for it, but have also done some calculations for c=0 and 2. For each value of c we compute a breakeven value for V. Consider the synthetic estimator

$$\hat{a}_i = \sum_j y_{ij}\hat{T}_j/Y_j = \sum_j p_{ij}\hat{T}_j \qquad (3)$$

and, in analogy with $f_k(\underline{a})$ for $\underline{a}$ as in (1), the loss-function value $f_k(\hat{\underline{a}})$. In the manner of Report 2 we now compute the value of V such that

$$E(f_k(\hat{\underline{a}})) = f_k(\underline{y}) \ . \qquad (4)$$

Explicitly, we get this value of V as follows. The general form of our loss functions $f_k(\underline{x})$ for k=1,2,3 is, as in (2) of Report 1,

$$\sum(x_i - t_i)^2/d_i \qquad (5)$$

with t_i equal to the true total for area i, and the divisor d_i equal to 1 for k=1, y_i for k=2, and t_i for k=3. We have

$$E(f_k(\hat{\underline{a}})) = f_k(\underline{a}) + V \sum_i \sum_j p_{ij}^2 T_j^c /d_i \qquad (6)$$

with $p_{ij} = y_{ij}/Y_i$ as in (1). Accordingly, the breakeven value for V is

$$V = (f_k(\underline{y}) - f_k(\underline{a}))/(\sum_i \sum_j p_{ij}^2 T_j^c /d_i). \qquad (7)$$

(Incidentally, Report 2 and 3 discussed at length the partition of c.v. C_3, for f_3, into C_2 and B. Here, with J totals T_j instead of a single T, we do not get such a neat partition and interpretation.)

For c=0, V is just a constant variance, and for c=2 it is just a constant c.v., as noted above. For c=1, of special interest to us, the interpretation becomes more elaborate. We now present some empirical results.

In our empirical study we considered the 3 artificial populations (true counts and accompanying census counts) AP1, AP2 and AP3 of Report 3, as constructed by Isaki, Diffendal and Schultz. Here, area i is state; cell j corresponds to age (5 categories), race (3 categories), and sex (2 categories), with 5x3x2, or 30, cells in all. (Here we do not consider the sampling strata that were of major interest in Report 3.) We included the following nine states in the study:

ME, NH, VT, MASS, CT, RI, NY, NJ, PA.

In other words we consider the northeastern end of the U.S.: PA and everything to the north and east of it.

In summary, our results make a substantial case for considering adjustment based on synthetic estimation. We have already discussed our preference for c=1 and will show results only for that value of c; but the conclusion is the same for c=0 and 2. Likewise in Report 2 we explained a preference for our 3rd loss function (k=3), based on division in (5) by the true t_i. Thus we give results only for k=3, but the conclusion is the same for loss functions 1 and 2.

For AP3 the ratio $f_3(\underline{a})/f_3(\underline{y})$, of loss functions based on synthetic state estimates ($\underline{a}$) and census figures ($\underline{y}$), is only .00933. That is, synthetic estimation reduces error, as measured by the loss function f_3, by a factor of almost 107 - if we know the true cell counts, T_j, exactly. We of course do not know T_j exactly; we use the estimates $\hat{T}_j$. For c=1 we consider the model $\mathrm{Var}(\hat{T}_j) = VT_j$ as in (2). The breakeven value for V is given by (7), with c=1, k=3, and $d_i = t_i$.

This breakeven V is 21,038.7. If V is in fact less that

this amount, we prefer the synthetic estimates $\hat{a}_i$ to the census counts y_i. We would interpret this breakeven V as follows. For cell j we have a variance VT_j, and thus a c.v. $(V/T_j)^{1/2}$. Let $\bar{T}$ be the arithmetic mean of T_j over our J cells; here we have $\bar{T}$ = 1,653,206.9, and an average c.v. of 11.281%. To justify synthetic estimation we require that our average cell c.v. be less than 11.281%; such a requirement does not seem restrictive. For example, if a cell is of size 1,653,207 (i.e., $\bar{T}$) the required c.v. is 11.281% or less. Likewise a cell of size 3,306,414 (i.e., $2\bar{T}$) requires a c.v. of 7.977%, of size 826,603 (i.e., $\bar{T}/2$) requires 15.954%.

Keep in mind that we are making a statement about the average behavior of all c.v.'s viewed simultaneously. We are not saying that every c.v. has to be less than its respective bound $(V/T_j)^{1/2}$ - only that, on average, this bound is not exceeded. An alternative interpretation is as follows. If the variance W_j were in fact proportional to T_j - that is, if the ratio W_j/T_j were equal to a constant V for all j - then our breakeven value of V would be given by (7). Accordingly, if most or all of the actual ratios W_j/T_j are less (greater) than V, then we probably should (should not) adjust. As an overall rule of thumb, which gives proper relative weighting to large and small areas, we might compare the ratio R = W/T against V, with $T = \sum T_j$ and W equal to the value, to the best of our knowledge, of $\sum W_j$, the sum of variances. For R<V we would adjust, for R>V we would not.

The above loss-function ratio, $f_3(\underline{a})/f_3(\underline{y})$, and average c.v., for a cell of size $\bar{T}$, correspond to AP3, which as in Report 3 is preferred among our 3 artificial populations. In all we have Table 1. On this basis, although c.v.'s for AP1 and AP2

Table 1: States

	Loss-Function Ratio	Average c.v.
AP1	.00829	5.943%
AP2	.00868	7.270%
AP3	.00933	11.281%

are less conclusive than for AP3, we encourage the exploration of ratio adjustment of state totals based on synthetic estimation.

Thus for AP1-2-3 we have considered the 9 states as "areas". Next we consider 217 counties (and independent urban jurisdictions), into which these 9 states are divided, as areas. Our age-race-sex cells, indexed by "j", are unchanged; but our areas "i" now are counties rather than states. Thus the totals T_j (true) and Y_j (census) are unchanged; but y_{ij} and a_i in (1), and the totals t_i and y_i, are now computed county-by-county. For counties, Table 2 gives the results for loss function (k=) 3 and exponent (c=) 1, that appear in Table 1 for states.

Table 2: Counties

	Loss-Function Ratio	Average c.v.
AP1	.22780	5.801%
AP2	.20600	7.778%
AP3	.21464	11.466%

The loss-function ratios in Table 1 are much smaller than in Table 2: for AP3 we have .00933 for states and .21464 for counties. Thus suppose that in fact we knew the true cell totals T_j exactly; then, the use of the synthetic a_i in preference to census y_i would not enhance our accuracy in estimating the true county totals t_i, as dramatically as it enhances it in estimating the true state totals. Meanwhile the average c.v.'s are about the same as in Table 1: for AP3 we have 11.466% for counties, vs. 11.281% for states. Thus to justify synthetic estimation for counties we require that our average cell c.v., in estimating T_j, be less than 11.466%, whereas for states we required 11.281%. Our knowledge of the true cell totals T_j has to be about as precise in order to make synthetic estimation for counties worthwhile, as it has to be in order to make synthetic estimation for states worthwhile. For AP1 and AP2, also, the county-state

difference is small. On this basis we encourage the exploration of county adjustment based on synthetic estimation.

Next we consider the 53,727 ED's, into which the 9 states are divided, as areas. We obtain Table 3, once again based on k=3 and c=1. The loss-function ratios are now close to 1; use of the synthetic a_i does not greatly enhance our accuracy in

Table 3: ED's

	Loss-Function Ratio	Average c.v.
AP1	.81692	4.610%
AP2	.71684	6.764%
AP3	.73189	9.365%

estimating the true ED totals (even if we, hypothetically, know the true T_j's exactly). The average c.v.'s are somewhat smaller than for states and counties. Thus to make synthetic estimation for ED's worthwhile at all we must know the T_j's with somewhat more precision than we need to make it worthwhile for states and counties.

Generally, therefore, we observe the following pattern. As average area size decreases and number of areas increases (from 9 states to 217 counties to 53,727 ED's), with our 30 adjustment cells held constant, the loss-function ratio steadily increases toward 1 (for AP3, from .00933 to .21464 to .73189). That is, synthetic estimation can remove a fraction of error which is much larger when we have a few, very large areas than when we have many small areas. However, the changes in breakeven c.v.'s are not so striking (for AP3, from 11.281 to 11.464 to 9.365). We must make a decision as to whether we know the T_j's precisely enough to make adjustment worthwhile (however modest the gains from it); this decision depends somewhat, but not heavily, on the number and size of our areas.

Thus one might envision a decision as to whether to adjust all areas at all levels. In this way we would have consistency

in that the adjusted population for a whole (e.g. state) is the sum of the adjusted populations for its separate parts (e.g., counties which comprise the state).

Two further directions of inquiry, as follows, are: (1) loss functions based on proportions rather than counts, and (2) positive correlation among the errors in the cell-total estimators $\hat{T}_j$.

(1) Up to now our loss functions have been based on area counts: t_i true, y_i census, a_i and $\hat{a}_i$ based on adjustment. We now base them on proportions of total population. Let:

$g_i = t_i/T$, the true proportion of population that area i represents out of the total.

$h_i = y_i/Y$, the proportion of population, as measured by the census, that area i represents.

$r_i = a_i/T$ with a_i, as in (1), the synthetic total for area i. That is, r_i is the proportion of population based on exact synthetic totals.

$\hat{r}_i = \hat{a}_i/\hat{T}$ with $\hat{a}_i$ as in (3) the synthetic estimator of total for area i, and $\hat{T}$ the estimated grand total, that we use in practice. That is, $\hat{r}_i$ is the proportion for area i based on the synthetic figures that we use in practice.

In our 4 loss functions we replace t_i, y_i, and $\hat{a}_i$ by the corresponding g_i, h_i, and $\hat{r}_i$. Thus our criterion, as measured in 4 different ways, is closeness of estimated area proportions to true area proportions rather than of estimated area counts to true counts. Such proportions are vitally related to proper allocation of a pie among the areas, as for revenue sharing or apportionment of the House of Representatives. For Report 1-3 we adjusted area census counts across the board; such an adjustment did not change the area proportions. Yet the synthetic adjustment of this report does change them.

Once again we limit the discussion to loss function (k=)3, with exponent c=1. Letting f_3' (rather than f_3) denote our new

loss function, we now compare the census

$$f_{\$}(\underline{h})=\sum(h_i-g_i)^2/g_i \qquad (10)$$

against the synthetic expected value

$$E(f_{\$}(\hat{\underline{r}}))=E(\sum(\hat{r}_i-g_i)^2/g_i). \qquad (11)$$

A convenient approximation to the value of (11) might be based on a standard 1st-order Taylor linearization of the ratio estimator $\hat{r}_i$. For (11) we obtain

$$\sum(r_i-g_i)^2/g_i+V\sum_i[\sum_j(p_{ij}-r_i)^2T_j]/g_iT^2 \qquad (12)$$

whereupon we may compute a breakeven value for V, with accompanying interpretation, as before.

Use of this linearization is limited, however. This paragraph digresses to explain why we use it when we have many areas, such as (perhaps) 217 counties, but not few areas, such as 9 states. On a first reading one might skip the details of this paragraph. We linearize the ratio $\hat{r}_i=\hat{a}_i/\hat{T}$, with $\hat{T}$ always appearing in the denominator. We approximate that $1/\hat{T}$ behaves linearly, i.e. like

$$1/T - (\hat{T}-T)/T^2, \text{ or } 2/T - \hat{T}/T^2 .$$

The inaccuracy that results from this approximation has its relatively largest effect on the value of (12) (specifically, the 2nd large term in (12)) when the relvariance of $\hat{a}_i$, in the numerator, is the smallest. The latter happens when we have the fewest, largest areas: That is, $\hat{a}_i$ has large expected value and small relative variance. Accordingly, for 9 states we in general anticipate less accuracy in the use of (12) , than we do for 217 counties; and even for our 217 counties we are only using a convenient approximation. (Future work might be pursued here, such as Monte-Carlo simulation. One might also consider a 2nd-order Taylor approximation to enhance accuracy, but we then encounter 3rd and 4th moments which prohibit our results from being distribution-free.)

Accordingly, Table 4 gives approximate average c.v.'s for counties based on (12) for proportions. We may compare these to the average c.v.'s in Table 2 based on counts for states and counties. Table 4 somewhat exceeds Table 2 for AP2 and AP3

Table 4: Average c.v.'s Based on Proportions

	Counties
AP1	4.880%
AP2	9.531%
AP3	12.039%

with the reverse holding for AP1; differences between breakeven levels do not seem major.

(2) Our second direction of inquiry is positive correlation. Early in this report we presumed that the random errors in the quantities $\hat{T}_j$ were uncorrelated. We now presume that there is positive correlation among them: an erroneously high total for cell 1 suggests that other cell totals are likely to be erroneously high. To capture the effect of this correlation, we model that the linear correlation between $\hat{T}_j$ and $\hat{T}_{j'}$ for $j \neq j'$ is equal to a nonnegative value R. From this point it is straightforward to develop formulas (dependent on the value of R) for the breakeven V. These formulas are extensions of those for counts in (7) and proportions in (12).

Average c.v.'s for counts and proportions, for AP3 only, are in Tables 5a and 5b. The figures for R=0 (no correlation) coincide with those for AP3 in Tables 1,2 and 4.

Table 5a

Average c.v.'s for counts, positive correlation and AP3

R	0	.1	.3	.6
States	11.281	6.610	4.346	3.194
Counties	11.466	7.033	4.692	3.466

Table 5b

Average c.v.'s for proportions, positive correlation and AP3

R	0	.1	.3	.6
Counties	12.039	9.821	7.589	5.995

As one might surmise (and can analytically demonstrate), the breakeven c.v. drops sharply for counts as the correlation increases. Thus correlation among our errors implies that we must know the true cell totals all the more precisely, in order

to make synthetic estimation worthwhile for counts. For proportions the same pattern is there empirically although much weaker; the gap between county breakeven levels in 5a and 5b widens as the correlation increases.