BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-88/12

THE DIFFICULTY OF IMPROVING STATISTICAL SYNTHETIC ESTIMATION

by

*Michael Lee Cohen and **Xiao Di Zhang
University of Maryland and Bureau of the Census
ASA/Census Program
Washington, D.C.

Bureau of the Census
ASA/Census Program
Washington, D.C. 20233

*ASA/Census Fellow
**ASA/Census Associate

Recommended by: Kirk M. Wolter

Report completed: March, 1988

Report issued: March 8, 1988

The Difficulty of Improving Statistical Synthetic Estimation

by

Michael Lee Cohen
University of Maryland and Bureau of the Census

and

Xiao Di Zhang
Bureau of the Census

Abstract: Statistical synthetic estimation, a technique widely suggested as a method for carrying down estimates to local levels can be shown to be a member of more general classes of estimators. It would then seem to follow that, by using information in the data, there would exist estimators that could select members of these classes based on this information, and that these estimators would outperform statistical synthetic estimation. We argue that these estimators are sometimes unavailable, and if they were available, would provide only modest improvements over the performance of the statistical synthetic estimator.

1. Definition of Statistical Synthetic Estimator

Assume we have a parent region composed of n subregions. We are provided with the census counts x_i, i=1,...,n, for each of the subregions and the census count X and the true count U, for the parent region, where $\sum_{i=1}^{n} x_i = X$. Let μ_i represent the unobserved true counts for the n subregions.

The problem is to estimate the μ_i under the constraint of internal consistency, i.e., where the estimates for the subregions sum to U. This property is important in census applications where it is expected that population counts for subregions add to population counts for parent regions. In this first case where we have no disaggregation for demographic subgroups, the statistical synthetic estimate is defined by:

$$(1) \qquad \hat{\mu}_i = (U/X)\, x_i \quad .$$

Now assume that we are provided with the census counts disaggregated demographically. Let us denote x_{ij} as the census count in the i-th subregion for the j-th demographic subgroup and we have that $x_i = \sum_{j=1}^{K} x_{ij}$. Also let $X_j = \sum_{i=1}^{n} x_{ij}$ be the census count of the parent region for the j-th demographic subgroup, and let U_j be the true count of the parent region for the j-th demographic subgroup. Let U_j be assumed known. In this case, the procedure is to separately apply the computation given above for each demographic subgroup and then add the results to arrive at the estimate:

$$(2) \qquad \mu_i^* = \sum_{j=1}^{K} (U_j/X_j)\, x_{ij}$$

It is clear that these estimates satisfy the constraint of internal consistency, i.e., that the estimates for a subregion add to the true count U for the parent region (where by the true count we might in practice intend only an improved count).

This allocation problem was examined by Deming (1938) with the following situation as motivation. Three measurements are taken of the three interior angles of a traingle. The three measurements will undoubtedly not add to 180°, and therefore need to be adjusted to incorporate this knowledge into the estimates. The methods Deming used are extended here to examine more general situations.

There are other situations which require internal consistency. One example arises if one asks respondents to supply probabilities for mutually exclusive and exhaustive events. Often, due to mistakes, the probabilities supplied will not sum to 100%. Without recontacting the respondent how should the observed percentages be modified so that they have the required property?

2. The Model for Statistical Synthetic Estimation

A. The case aggregated by demographic group -- the 1-dimensional case.

There are two simple models which result in statistical synthetic estimation in the 1-dimensional case. First, if we restrict ourselves to estimates of μ_i of the form:

$$\overset{0}{\mu}_i = Kx_i \ ,$$

and if we require that $\sum_{i=1}^{n} \mu_i^0 = U$, then K must equal U/X , and

$$\mu_i^0 = (U/X)\, x_i \ .$$

Likewise, if we wish to solve the minimization:

$$\min_K \sum_{i=1}^{n} (Kx_i - \mu_i)^2/x_i$$

(where x_i is playing the role of a variance as well as an observed value) we find that K = U/X. In this second model we are fortunate that the dependence of K on the individual unknown μ_i is solely through the known sum U.

Both of these models seem overly simplistic and narrow. It would be comforting to users of statistical synthetic estimation if it could be shown to be optimal for a wider class of estimates. We now show this to be the case. Following Deming (1938), we assume:

(3) $$x_i \overset{\text{ind.}}{\sim} N(\mu_i \ , \sigma_i^2) \ , \ i=1,\ldots,n \ ,$$

and furthermore we assume that $\sum_{i=1}^{n} \mu_i = U$ is know. The objective is to estimate the n μ_i's from the n x_i's under the constraint that the sum of the estimates equals U.

(A difficulty with this model in the census context is that we have little reason to assume that the x_i are unbiased. The bias does,

however, appear indirectly through the difference between $\sum_{i=1}^{n} \mu_i$ and $\sum_{i=1}^{n} x_i$.)

Constrained maximum likelihood estimation of (3), using the method of Lagrange multipliers, results in:

$$\mu_i^{\sigma} = x_i + [U-X] \left(\sigma_i^2 \Big/ \sum_{j=1}^{n} \sigma_j^2\right) . \tag{4}$$

This formula has an interesting interpretation. We distribute the overage, U-X, in proportion to the variability of each subregion. (We will avoid the philosophical discussion of what a variance for census counts means, except to say that the notion does have a frequentist interpretation.)

The statistical synthetic estimate arises as a special case when we set $\sigma_i^2 = x_i$. Then (4) becomes:

$$\hat{\mu}_i = x_i + [U-X] \left(x_i/X\right) = (U/X)\, x_i .$$

This estimate has the interpretation of spreading the overage to subregions in proportion to their population size. It is important to point out that using a random variable as a variance is at least awkward, but $\hat{\mu}_i$ can be considered as an approximation to an estimate where σ_i^2 is unknown but close to x_i. One possibility might result from the model:

$$x_i \overset{\text{ind.}}{\sim} N(\mu_i , \mu_i) .$$

What this argument shows is that statistical synthetic estimation can result from a nonparametric model. Also we have now demonstrated statistical synthetic estimation as a member of a general class of estimates (4) which raises the possibility of using members for specific situations which outperform $\hat{\mu}_i$.

B. The case disaggregated by demographic group -- the K-dimensional case.

The estimate $\mu_i^* = \sum_{j=1}^{K} [U_j/X_j]\, x_{ij}$

results from constraining a parametric class of estimates (and not optimizing) as was the case for $\hat{\mu}_i$. Consider the class of estimates:

$$\mu_i^0 = \sum_{j=1}^{K} K_j\, x_{ij} \ .$$

We introduce the constraints that the estimates for demographic subgroups in subregions should add to the assumed known estimates U_j for demographic subgroups in the parent region. Thus:

$$\sum_{i=1}^{n} K_j\, x_{ij} = U_j \text{ which implies that } K_j = U_j/X_j$$

Nonparametrically, we can derive μ_i^* from the following generalization of the 1-dimensional case. Let:

$$x_{ij} \overset{\text{ind.}}{\sim} N(\mu_{ij}, \sigma_{ij}^2)$$

where the μ_{ij} are unknown means for the j-th demographic subgroup in the i-th subregion, and where:

$$\sum_{i=1}^{n} \mu_{ij} = U_j \text{ , for } j=1,\ldots,K.$$

Just as in the 1-dimensional case, a Lagrange multiplier argument can be used to solve this problem of constrained maximum likelihood, resulting in:

$$\mu_{ij}^{\sigma} = x_{ij} + [U_j - X_j] \, (\sigma_{ij}^2 \, / \, \sum_{m=1}^{n} \sigma_{mj}^2)$$

3. Generalization of the 1-Dimensional Case

We now examine the more general model:

$$\underset{\sim}{X} \sim N \, (\underset{\sim}{\mu} \, , \, \underset{\sim}{\Sigma}) \, ,$$

where $\underset{\sim}{x}' = (x_1,\ldots,x_n)$, $\underset{\sim}{\mu}' = (\mu_1,\ldots,\mu_n)$, and $\underset{\sim}{\Sigma}$ is the nxn covariance matrix of $\underset{\sim}{x}$. We wish to find the maximum likelihood estimates of $\mu_1,\ldots,\mu_n$ subject to the constraint $\sum_{i=1}^{n} \mu_i = U$.

To accomplish this we use Lagrange multipliers. The maximum likelihood criterion reduces to minimizing:

$$(\underset{\sim}{x} - \underset{\sim}{\mu})' \, \underset{\sim}{\Sigma}^{-1} \, (\underset{\sim}{x} - \underset{\sim}{\mu}) \, ,$$

subject to: $\underset{\sim}{\mu}' \, \underset{\sim}{1} - U = 0$.

We have:

$$-2(\underset{\sim}{x} - \underset{\sim}{\mu})' \underset{\sim}{\Sigma}^{-1} = \lambda \underset{\sim}{1}'$$

(5) $$\text{or} \quad \underset{\sim}{\mu}' = \underset{\sim}{x}' + (\lambda/2)\, \underset{\sim}{1}' \underset{\sim}{\Sigma}$$

Multiplying (5) on the right by $\underset{\sim}{1}$ gives us:

$$\underset{\sim}{\mu}' \underset{\sim}{1} = \underset{\sim}{x}' \underset{\sim}{1} + (\lambda/2)\, \underset{\sim}{1}' \underset{\sim}{\Sigma} \underset{\sim}{1}$$

which implies that:

$$\lambda = (2/\underset{\sim}{1}' \underset{\sim}{\Sigma} \underset{\sim}{1})\, (\underset{\sim}{\mu}' \underset{\sim}{1} - \underset{\sim}{x}' \underset{\sim}{1})$$

And so:

(6) $$\tilde{\underset{\sim}{\mu}}' = \underset{\sim}{x}' + (\underset{\sim}{1}' \underset{\sim}{\Sigma} / \underset{\sim}{1}' \underset{\sim}{\Sigma} \underset{\sim}{1})\, (\underset{\sim}{\mu}' \underset{\sim}{1} - \underset{\sim}{x}' \underset{\sim}{1})$$

Letting σ^0_{ij} represent the (i,j-th) element of $\underset{\sim}{\Sigma}$, we can rewrite (6) as:

$$\tilde{\mu}_i = x_i + \left(\sum_{i=1}^{n} \sigma^0_{ij} \Big/ \sum_{i,j=1}^{n} \sigma^0_{ij}\right) (U-X) \ .$$

As a check, when $\sigma_{ij} = 0$ for all $i \neq j$, we have:

$$\mu_i^{\sigma} = x_i + \left(\sigma^0_{ii} \Big/ \sum_{j=1}^{n} \sigma^0_{jj}\right) (U-X) \ ,$$

as demonstrated before.

It is important to mention that in the census application the covariances σ_{ij} are unlikely to be estimable, and it is unclear that even rough estimates of the σ_{ii} would be available. Thus it would be comforting to know that the cost of using the wrong weights is often not great.

Consider the case where $\sigma_{ij} = 0$ for all $i \neq j$. Then:

$$\text{Var}\{\mu_i^{\sigma}\} = E\left[x_i + \left(\sigma_{ii} \Big/ \sum_{j=1}^{n} \sigma_{jj}\right)(U-X) - E\left(\mu_i^{\sigma}\right)\right]^2$$

$$= \sigma_{ii}\left[1 - \sigma_{ii} \Big/ \sum_{j=1}^{n} \sigma_{jj}\right] \tag{7}$$

Equation (7) makes it clear that the principal objective of statistical synthetic estimation must be internal consistency, since there is no great gain in precision. For example, if the variances for the n subregions are roughly comparable, we have:

$$\text{Var}\{\mu_i^{\sigma}\} \approx \sigma_{ii}\{1 - \tfrac{1}{n}\} = [(n-1)/n]\,\sigma_{ii} \ .$$

Given that it is difficult to beat the census by a large amount, it is surprising that one can misguess the σ_{ii} by quite a bit and still outperform the census. Suppose instead of the $\sigma_{ii} \Big/ \sum_{i=1}^{n} \sigma_{jj}$, we use weights f_i. Then we have:

$$\text{Var}\{\mu_i^{f}\} = \sigma_{ii}\left[1 - 2f_i + f_i^2 \left(\sum_{j=1}^{n} \sigma_{jj} \Big/ \sigma_{ii}\right)\right]$$

It is easy to show that $\mathrm{Var}\{\mu_i^f\} \leq \mathrm{Var}\{x_i\}$ whenever $0 \leq f_i \leq 2\,\sigma_{ii}$. Therefore, if one is able to estimate the variances of the counts for the subregions within a factor of 2, one can outperform the census counts, but not by a great deal.

4. Generalization of the K-Dimensional Case

A similar generalization of the K-dimensional case to the generalization given above for the 1-dimensional case could be developed. Instead we follow a different course. A straightforward approach to take is to choose a loss function and a parametric form of an estimate, along with the constraints imposed by internal consistency, and solve the constrained optimization problem. An obvious class of estimates is:

$$\sum_{j=1}^{K} K_j \, x_{ij} \, .$$

The loss function that is probably proposed most often in this setting is (see Tukey, 1983):

$$(8) \qquad \sum_{i=1}^{n} (a_i - \mu_i)^2 \, / \, \mu_i$$

where a_i is some estimate of the population count in the i-th subregion. If we were to apply all K constraints at once, namely:

$$\sum_{i=1}^{n} K_j \, x_{ij} = U_j \text{ for all } j \, ,$$

we must have that $K_j = (U_j/X_j)$. It is of interest to see what gains are possible if we use (8) with only one constraint (which would likely not

satisfy census requirements):

$$\sum_{i=1}^{n} \sum_{j=1}^{K} K_j \, x_{ij} = U \, .$$

Let us examine the case where K=2. We have:

$$(9) \qquad \min_{K_1,K_2} \sum_{i=1}^{n} (K_1 x_{i1} + K_2 \, x_{i2} - \mu_i)^2 \, / \, \mu_i$$

$$\text{such that} \quad \sum_{i=1}^{n} (K_1 \, x_{i1} + K_2 \, x_{i2}) = U \, .$$

Solving using Lagrange multipliers again, we find:

$$(10) \qquad \hat{K}_1 = \Big(\sum_{i=1}^{n} F_i \, G_i / \mu_i \Big) \, / \, \Big(\sum_{i=1}^{n} G_i^2 / \mu_i \Big)$$

where: $F_i = (U/X_2) \, x_{i2}$ and $G_i = x_{i1} - (X_1/X_2) \, x_{i2}$,

Also $\hat{K}_2 = (U - \hat{K}_1 \, X_1) \, / \, X_2$.

So $\hat{K}_1$ (and $\hat{K}_2$) has the interpretation of being a weighted regression coefficient, regressing F_i on G_i . F_i can be interpreted as an estimate for μ_i given the data x_{i2}, and G_i can be interpreted as the component of x_{i1} unexplained by x_{i2}. (Note: similar calculations result from the case K=3.)

$\hat{K}_1$ cannot be used in practice because it is a function of the unknown μ_i . At first glance this does not seem crucial since the μ_i appear in (10) simply as weights and one could seemingly substitute x_i or μ_i^* in for μ_i with little loss in performance. However, this turns out not to be the case. If one

substitutes x_i for μ_i in (10), $\hat{K}_1$ and $\hat{K}_2$ turn out to be equal, obviously rarely optimal. Furthermore, if one substitutes μ_i^* in for μ_i in (10) $\hat{K}_1$ and $\hat{K}_2$ turn out to be (U_1/X_1) and (U_2/X_2) respectively.

An interesting question is if the μ_i were available for use as weights, how much would (10) outperform μ_i^* . To answer this question we used four artificial data sets the Bureau of the Census has developed which are believed to approximate many of the features of undercoverage and which provide a true count and a census count for subregions and demographic subgroups. These are fully described in Isaki, Schultz, et al. (1987).

We computed the estimate in (10) and μ_i^* for artificial populations II and III, at the levels of states and counties, for 2 and 3 demographic subgroups. When we used 2 demographic subgroups Blacks and Hispanics were combined into one group, otherwise the three demographic subgroups were defined as (i) Blacks, (ii) Hispanics, and (iii) White and Others. The results are presented in Table 1.

Table 1. Comparison of Optimal Estimate with Statistical Synthetic Estimate Using Artificial Populations - Comparison Made Using Loss Function $\sum (a_i - \mu_i)^2/\mu_i$ [a].

A. 2 Demographic Groups I = Black + Hispanic II = White + Other

Level	Data Set	Loss: Optimal (1)	Loss: Stat. Synth. Est. (2)	Efficiency of Stat. Synth. Estimator (1)/(2)
State	AP2	9754.7	10666.5	.91
	AP3	8235.4	9176.5	.90
County	AP2	38343.3	39825.0	.96
	AP3	39420.8	41384.3	.95

B. 3 Demographic Groups I = Black II = Hispanic III = White + Other

Level	Data Set	Loss: Optimal (1)	Loss: Stat. Synth. Est. (2)	Efficiency of Stat. Synth. Estimator (1)/(2)
State	AP2	7847.3	9245.8	.85
	AP3	8234.9	9249.7	.89
County	AP2	34871.8	36890.9	.95
	AP3	39402.6	41564.8	.95

[a] For K=2, optimal estimate is given in (9). For K=3, similar calculations yield a more complicated expression.

Table 1 indicates that the efficiency of statistical synthetic estimation will often be very acceptable, especially given that the extra constraints are likely necessary for the census application of these methods, as well as the fact that the μ_i are unknown. The efficiencies are usually above .90. The difference between AP2 and AP3 is that in AP3 the undercoverage for Hispanics is assumed to be identical to that for Blacks, while in AP2 the undercoverage for Hispanics is assumed to be identical to that for Whites and Others. Thus, it is expected that the AP3 results would be essentially identical for 2 and 3

demographic groups. Finally, there is a hint that the efficiencies will fall with a rise in the number of demographic groups used.

5. Conclusion

Statistical synthetic estimation will likely play an important role in any adjustment of the 1990 Census. It has been suggested as the method by which estimates derived from the Post-Enumeration Survey will be "carried down" to small geographic areas. Also, should a timely adjustment be decided upon (and the Post-Enumeration Survey-based adjustment turns out not to be timely) statistical synthetic estimation based on national estimates of undercoverage derived from demographic analysis could be calculated in time to meet the existing statutory deadlines.

We have shown that this simple but important estimator is surprisingly hardy. In the case for K=1, if one tries to better estimate the weights to use it is unlikely that the resulting estimator will outperform statistical synthetic estimaton by much. In the case for K>1, if one relaxes some of the constraints the defficiency of statistical synthetic estimation is unlikely to be more than 15%. If one tries to substitute known quantities in as weights one merely recreates the statistical synthetic estimate or worse. These results should provide some comfort to prospective users of this methodology.

References

Deming, W. Edwards (1938), "Statistical Adjustment of Data," Dover Publications, Inc., New York.

Isaki, C.T., Schultz, L.K., Smith, P.J., and Diffendal, G.J. (1987), "Small Area Estimation Research for Census Undercount-Progress Report," pp. 219-238 in Small Area Statistics: An International Symposium, John Wiley and Sons, New York, NY.

Tukey, John W. (1983), Affidavit Presented to District Court, Southern District of New York, Mario Cuomo, et al., Plaintiffs(s), Malcolm Baldrige, et al., Defendants, 80 CIV. 4550 (JES).