

Census 2000 Accuracy and Coverage Evaluation: Assessment of Synthetic Assumption

Richard A. Griffin and Donald J. Malec

U.S. Census Bureau

The synthetic assumption states that census net coverage does not vary within post-strata. For example, the synthetic assumption implies that census counts in St. Louis, Missouri in a given post-stratum have the same net coverage as the census counts in the same post-stratum but in Milwaukee, Wisconsin. The synthetic assumption within post-strata will permit the Census Bureau to draw conclusions from the A.C.E. sample about the population as a whole, to individuals living in geographic areas smaller than post-strata. The synthetic assumption must hold to permit correction for small geographic areas based on national level, post-strata sample estimates. This adjustment is only correcting for systematic biases and not local census errors. The error that is introduced when the synthetic assumption does not hold is called synthetic error.

Assessments of the 1990 PES were concerned with the possibility that synthetic error introduced error in the PES, especially for low levels of geography such as blocks. Synthetic error is of greater concern for small areas than for larger geographic aggregations. It is acknowledged that synthetic error will likely result in the population of some blocks being overestimated and the population of other blocks being underestimated; statistical correction is not expected to produce unqualified improvement in the smallest geographic areas, like blocks.

While the accuracy of the A.C.E.'s synthetic estimates depends on the degree in which net coverage varies within post-strata, it is important to understand that perfectly equal net coverage cannot exist within all post-strata. The Census Bureau's evaluation of synthetic error should focus on whether the variability of net coverage is so great as to prevent an improvement from using the A.C.E. Additionally, the A.C.E. was designed to reduce the variability of net coverage as compared with the 1990 PES. The A.C.E. design has enhanced post-strata, including variables for mail return rate and type of enumeration areas. In

---

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.

addition, the census has net coverage that varies across areas.

This paper presents alternative methods to document and measure synthetic error in the A.C.E. and the effects, if any, these violations had on the overall accuracy of the A.C.E., both numeric and distributive.

The two components of error in synthetic estimates are: (1) **Synthetic population bias** due to applying the same coverage correction factor to areas with different net census coverage and (2) **Bias in the post-stratum level Dual System Estimate (DSE)** including correlation bias. Synthetic bias is measured at the Congressional district and state levels and is compared to error in the census.

### Overview of methodology

This section describes the essence of estimating bias in synthetic estimates. The Appendix provides the mathematical details of the methodology.

### Creation of artificial populations

The basic methodology used to estimate the synthetic population bias component of synthetic error is artificial populations.

We use census variables thought to be related to coverage to produce artificial populations. Call these variables surrogates. We use methodology similar to one method suggested by Freedman and Wachter (1994). Adjust one surrogate variable to gross undercount and another to gross overcount. This is done by distributing the post-stratum level gross undercount (gross overcount) proportional to the gross undercount surrogate variable (gross overcount surrogate variable) for the congressional districts (see Appendix). These are added and subtracted to census counts to form an artificial population. Unlike other approaches, this strategy can provide both net over- and under- coverage between local areas within a poststrata.

The surrogate variables considered are:

- Allocations -households with more than a specified amount of item nonresponse
- Number of Non-Mail Returns
- Number of Substitutions -whole-household imputes and/or partial household substitutions
- Units at basic street address

Allocations, substitutions, multi-unit, and non-mail back were surrogates used by Freedman and Wachter (1994). They also used mobility and poverty which are Census 2000 long form data items not available at this time.

At the block cluster level, a correlation between a "coverage gap" and each artificial population's estimated true net coverage error (see Appendix for details) can be made. Note that each artificial population uses two surrogate variables, one for gross undercount and one for gross overcount. The correlations are used to help rank the artificial populations in order of importance.

From this analysis, multiple sets of artificial populations are selected for calculation of the error of synthetic estimates.

### *Bias due to synthetic estimation*

The first component, synthetic population bias is estimated from an artificial population; it is the synthetic estimate minus the population count estimated from the artificial population.

The second component is estimated using post-stratum biases, estimated as part of the Total Error Model and Loss function work. It is the post-stratum level estimated biases,  $\hat{D}_{ij}$ , in the DSE allocated to the state and congressional district levels.

The estimated bias for shares accounts for the same two components of error as for levels.

The estimate of bias for area i takes the following form:

$$\begin{aligned}\hat{B}_i &= \text{Syn}\hat{B}_i + D\hat{S}E\hat{B}_i \\ &= (\tilde{N}_{i\cdot} - N_{i\cdot}) + \sum_j \frac{\text{Cen}_{ij}}{\text{Cen}_{\cdot j}} \hat{D}_{ij}\end{aligned}$$

where  $\tilde{N}_{i\cdot}$  is a synthetic target (unbiased at the post-stratum level) estimate and  $N_{i\cdot}$  is the true population total, for area level i.

The bias for the synthetic estimator of a population share for area i takes the following form:

$$\hat{B}_{share, i} = \frac{N_i + \text{Syn}\hat{B}_i + D\hat{S}E\hat{B}_i}{\sum_i (N_i + \text{Syn}\hat{B}_i + D\hat{S}E\hat{B}_i)} - \frac{N_i}{\sum_i N_i}$$

*Bias in census counts (shares)*

The bias in the census count (share) for an area is the census count (share) minus the population count (share) estimated from the artificial population.

## **Results**

### **Artificial population creation**

Based on the block cluster level correlation analysis, four artificial populations were created as described in Table 1. Among all the combinations of overcount and undercount surrogates considered, these were the four that had the highest correlations.

Note that for Artificial Populations 2 and 4 the same surrogate variable is used for undercount and overcount. Thus if the post-strata has an overall undercount (overcount) all local areas will have an undercount (overcount) for that post-strata for these artificial populations. See the Appendix for details.

### **Components of bias in synthetic estimates for states**

Table 2 provides the mean and standard deviation of the relative synthetic bias (both components as defined above) for states for counts and shares for each of the four artificial populations.

The results for Artificial Populations 1 and 2 are similar. Artificial Population 4 is similar to these on average but has more variation. Artificial Population 3 is different than the others.

### **Effect of synthetic Error on the Weighted Squared Error Loss Function Analysis**

The loss function analysis, employed by the Census Bureau, does not, traditionally, include an error component for the failure of the synthetic assumption (Fay and Thompson (1993)). An expression for a bias correction to a weighted squared error loss function difference,  $\text{Loss}(\text{Census}) - \text{Loss}(\text{A.C.E.})$ , is shown in the Appendix. This bias correction term can be added to loss function results to correct for the bias of excluding synthetic error in the loss function target estimates. The interpretation of the bias correction term is most relevant in terms of the sign of the squared error loss function difference. If the loss function difference is positive, indicating adjustment is favorable, only a negative bias correction can change this making adjustment unfavorable. Similarly, if the difference is negative, indicating adjustment is not favorable, this can be reversed only if the bias correction is positive. The

amount of bias being added or subtracted must be larger than the absolute difference to reverse the outcome.

Tables 3 and 4 show the bias correction term for congressional districts for estimated counts and estimated shares. In each table results are shown for each of the four artificial populations. Column (1) is the census weighted squared error loss minus the adjusted weighted squared error loss. This has a bias due to excluding synthetic error. Column (2) is the synthetic bias correction term. Column (3) is the relative bias (column (2) / column (1)). Column (4) is the bias corrected loss function difference (column (1) + column (2)).

For congressional district count estimates (Table 3), three of the four artificial populations show a negative bias correction is necessary. However, in all three of these cases this negative bias correction is less than 8 percent of the difference in census and A.C.E. Thus, correcting for the bias would not reverse the loss function results. For the other artificial population, the loss function analysis is conservative.

For congressional district share estimates (Table 4), the bias correction is positive for two of the four artificial populations and the loss function analysis is conservative. For the other two artificial populations, the bias correction is negative (12.04 percent and 3.7 percent) but much smaller in absolute value than the loss function difference. Thus correcting for the bias would not reverse the loss function results.

Similar tables for states are given in Griffin and Malec (2001).

#### **Bias estimates from the total error model and loss function analysis used for examining the effect of synthetic bias on loss function analysis**

All loss function results cited in this report use the model which includes correlation bias except for Non-Blacks ages 18-29 and uses the Gross DSE to distribute target estimates. Using Gross Undercount to distribute target estimates keeping the correlation bias assumption fixed would produce results of similar magnitude and sign. We did not run alternative correlation bias assumptions; we think these results are reasonable under these alternatives but we are not completely confident of this. Work on a sensitivity analysis is in progress. See Navarro and Asiala (2001) for information on how results differ with different DSE bias assumptions.

#### **References**

- Griffin, R. and Malec, D. (2001), "Accuracy and coverage Evaluation: Assessment of Synthetic Assumption", DSSD Census 2000 Procedures and Operations Memorandum Series B-14\*
- Fay, R.E. and J. Thompson (1993). "The 1990 Post Enumeration Survey Statistical Lessons in Hindsight." *Proceedings of the 1993 Annual Research Conference*. U.S. Bureau of the Census, 71-91.
- Freedman, D. and K. Wachter (1994). "Heterogeneity and Census Adjustment for the Intercensal Base." *Statistical Science*, 476-485.
- Navarro, A. and M. Asiala, (2001). "Accuracy and Coverage Evaluation: Comparing Accuracy." DSSD Census 2000 Procedures and Operations Memorandum, Series B-13\*, February 28, 2001.

**Table 1: Surrogate Variables used to Create Artificial Populations**

	Correlations (weighted analysis)	Undercount Surrogate	Overcount Surrogate
Art. Pop. 1	0.26	# non-substituted persons in households	#persons for whom reported date of birth and reported age were consistent (allocation not required)
Art. Pop. 2	0.27	# non-substituted persons in households	# non-substituted persons in households
Artificial Population 3	0.26	# persons with 2 or more items allocated	#persons for whom reported date of birth and reported age were consistent (allocation not required)
Artificial Population 4	0.25	# persons whose household did not mail back the questionnaire	# persons whose household did not mail back the questionnaire

Household Persons only (Group Quarters Persons are Excluded)

**Table 2: Average and Standard Deviation of State Relative Synthetic Bias**

Statistic	Artificial Pop. 1		Artificial Pop. 2		Artificial Pop. 3		Artificial Pop. 4	
	Count	Share	Count	Share	Count	Share	Count	Share
Mean	0.0079	0.0007	0.0079	0.0007	0.0086	0.0014	0.0079	0.0007
Standard Deviation	0.0017	0.0017	0.0017	0.0017	0.0065	0.0065	0.0030	0.0030

**Table 3: Weighted Loss Function Synthetic Bias Correction for Congressional District Counts**

Weighted				
Artificial Population	Census Loss minus A.C.E. Loss (1)	Synthetic Bias Correction (2)	Relative Bias (3)	Corrected Loss (4)
1	2.07E+04	-4.99E+02	-2.41%	2.02E+04
2	2.07E+04	-8.69E+01	-0.42%	2.06E+04
3	2.07E+04	5.64E+03	27.22%	2.64E+04
4	2.07E+04	-1.61E+03	-7.79%	1.91E+04

**Table 4: Weighted Loss Function Synthetic Bias Correction for Congressional District Shares**

Weighted				
Artificial Population	Census Loss minus A.C.E. Loss (1)	Synthetic Bias Correction (2)	Relative Bias (3)	Corrected Loss (4)
1	2.09E-04	-2.51E-05	-12.04%	1.84E-04
2	2.09E-04	-7.73E-06	-3.70%	2.01E-04
3	2.09E-04	4.99E-04	238.79%	7.07E-04
4	2.09E-04	3.83E-05	18.36%	2.47E-04

## APPENDIX

### Forming artificial populations

Let X denote a surrogate for gross undercount and Y denote a surrogate for gross overcount.

$DSE_j$  = the Dual System Estimate for Post-stratum j

$CE_j$  = the weighted E sample number of correct enumerations in post-stratum j

$EE_j$  = the weighted E sample number of erroneous enumerations in post-stratum j

$Cen_{.j}$  = the census count in post-stratum j

Note that for any variable V,  $V_{.j}$  is the sum of  $V_{ij}$  over areas i.

Define the estimated number of omissions as follows:

$$OMISS_j = DSE_j - Cen_{.j} \left( \frac{CE_j}{E_j} \right)$$

Define the estimated erroneous enumerations as follows:

$$ERR_j = Cen_{.j} \left( \frac{EE_j}{E_j} \right)$$

$N_{ij}$  is the artificial population value and  $Cen_{ij}$  is the census count for area i, post-stratum j.

$$N_{ij} = Cen_{ij} + X_{ij} \frac{OMISS_j}{X_{.j}} - Y_{ij} \frac{ERR_j}{Y_j}$$

$$N_{.j} = Cen_{.j} + OMISS_j - ERR_j = DSE_j$$

The artificial population surrogates were selected by computing the, within post-strata, correlation between the coverage gap

$z = (\text{Weighted P-sample Non-matches}) - (\text{Weighted E-sample erroneous enumerations})$ .

and  $N_{ij} - Cen_{ij}$  (estimated true net coverage error), at the A.C.E. block cluster level.

### Correction for Synthetic Bias in Loss Function Analysis

Notation:

$D_g$  = the census squared error loss minus the A.C.E. squared error loss using synthetic target estimates.

$D_t$  = the census squared error loss minus the A.C.E. squared error loss using "true" target estimates

The loss function analysis output is in terms of expected losses using the synthetic target estimates, i.e.,  $\Delta_g = E(D_g)$ . However, we would like to know

$\Delta_t = E(D_t)$ . Therefore, we develop

an expression for a bias correction term, B, to be added to  $\Delta_g$  to correct loss function results for synthetic bias so that

$$\Delta_t = \Delta_g + B.$$

Define:

$w_i$  = the squared error loss function weight for area i (set equal to 1 for an unweighted squared error loss)

$Cen_i$  = the census count for area i

$N_i$  = the "true" target estimate for area i

$\tilde{N}_i$  = the synthetic target estimate for area i

$\hat{N}_i$  = the A.C.E. synthetic estimate for area i (including DSE post-stratum biases)

$b_i$  = bias in the post-stratum level DSE including correlation bias allocated to area i

By definition,

$$a_i = E(\hat{N}_i) = \tilde{N}_i + b_i$$

Using this notation:

$$D_g = \sum_i [w_i(Cen_i - \tilde{N}_i)^2 - w_i(\hat{N}_i - \tilde{N}_i)^2], \text{ and}$$

$$\begin{aligned} D_t &= \sum_i [w_i(Cen_i - N_i)^2 - w_i(\hat{N}_i - N_i)^2] \\ &= D_g + 2 \sum_i w_i(\tilde{N}_i - N_i)(Cen_i - \hat{N}_i) \end{aligned}$$

The resulting expected difference is:

$$\begin{aligned} \Delta_t &= \Delta_g + 2 \sum_i w_i(\tilde{N}_i - N_i)(Cen_i - a_i) \\ &= \Delta_g + 2 \sum_i w_i(\tilde{N}_i - N_i)(Cen_i - \tilde{N}_i - b_i), \end{aligned}$$

so  $\sum_i B_i =$  bias correction term  $= 2 \sum_i w_i(\tilde{N}_i - N_i)(Cen_i - \tilde{N}_i - b_i)$ .

Estimates for this bias term are made by using artificial population values for the terms  $N_i$  and  $\tilde{N}_i$  and by estimating  $b_i$  with  $\sum_j \frac{Cen_{ij}}{Cen_{.j}} \hat{D}_j$ . An analogous approach is used for shares.