Loss Function Adjustment Accounting for Synthetic Bias to Evaluate Coverage Measurement for Census 2000

Richard Griffin
U.S. Census Bureau, Washington, DC 20233

Introduction

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 Accuracy and Coverage Evaluation (A.C.E.) sample about the population as a whole, to individuals living in geographic areas smaller than post-strata. The synthetic assumption is necessary to permit correction for small geographic areas based on a sample. 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.

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. Note that the census also has net coverage that varies across areas.

The loss function results reported in Navarro and Asiala (2001) did not include a measure of error due to the synthetic assumption. Griffin and Malec (2001) presented the effect of this bias on the loss function results. They used one of eight sets of assumptions dealing with correlation bias and processing error and one of two methods to synthetically distribute total error model targets to states and congressional districts (Model 6 and Synthetic Method 1, see Overview of methodology). The bias estimates used were from the 1990 Post Enumeration Survey. The 2000 A.C.E. was found to have overstated the undercount due to missing erroneous enumerations. Revised 2000 A.C.E. estimates are due to be completed by the end of 2002.

This report is a sensitivity analysis of the effect of varying these eight assumptions and two methods on the assessment of the effect of synthetic error on the loss function analysis. Two additional artificial populations are studied in addition to the four artificial populations examined by Griffin and Malec.

Since implementation of the methodology of this paper used 2000 A.C.E. results which will be revised, this paper is presented for the methodology of analysis and the results given only for the purpose of illustration. Results do not indicate any comparison of Census 2000 and the final revised A.C.E. estimates

Overview of methodology

This section describes the essence of estimating the effect of synthetic error on loss function results. The Appendix provides the mathematical details of the methodology.

Creation of 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). We adjusted one surrogate variable to weighted omissions and another to weighted erroneous enumerations. This is done by distributing the post-stratum level weighted omissions (weighted erroneous enumerations) proportional to the weighted omissions surrogate variable (weighted erroneous enumeration surrogate variable) for the congressional districts. These are added and subtracted to census counts to form an artificial population count. A correction for the bias in the post-stratum level dual system estimate (for alternative correlation bias and processing error assumptions) is allocated to the artificial population count for each congressional district. Congressional Districts are added to get state counts. (see Appendix). Unlike other approaches, this strategy can provide both net over- and under- coverage between local areas within a post-stratum. It is possible that the surrogates that are best for weighted omissions are different than those that are best for weighted erroneous enumerations. All artificial population counts summed over congressional districts and post-strata are equal to the target counts used in the loss function analysis (for alternative correlation bias and processing error assumptions).

The surrogate variables considered are:

- Allocations - Households with more than a specified amount of item nonresponse (Items include race, Hispanic origin, relationship, sex, and age)

- Number of Non-Mail Returns
- Number of Substitutions - Whole-household imputes and/or partial household substitutions
- Number of duplicates added back (late adds)
- 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 weighted omissions and one for weighted erroneous enumerations. Because of the possibly large amount of geocoding error at the block cluster level, these correlations will likely be small. Large correlations may merely mean that our artificial populations are related to geocoding error. Whatever the case, the correlations may be 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.

Sensitivity of Loss Function Results

The loss function results reported in Navarro and Asiala (2001) do not include an error component for the failure of the synthetic assumption used to create the target counts. An expression for a bias correction to a squared error loss function difference, Loss(Census) - Loss(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 analysis. 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.

Variations in assumptions used in Sensitivity analysis

Loss function results for states and congressional districts are reported for eight different sets of correlation bias and processing error assumptions and for two methods of synthetically carrying down targets from the evaluation post-stratum level to the production post-stratum level.

The eight sets of correlation bias and processing error for states and congressional districts are shown in Table 1.

The two methods of synthetically carrying down targets for states and congressional districts are Method 1 - Proportional to the Gross Dual System Estimate (DSE) and Method 2 - Proportional to the Gross Undercount.

Results

Artificial population creation

Based on the block cluster level correlation analysis, four sets of artificial population surrogate variables were selected as described in Table 2 for Artificial Populations 1, 2, 3, and 4. For each of these four artificial populations the count was corrected for DSE bias proportional to the census counts. Note that for Artificial Populations 2 and 4 the same surrogate variable is used for weighted omissions and weighted erroneous enumerations. Thus if the post-stratum has an overall undercount (overcount) all local areas will have an undercount (overcount) correction for that post-stratum for these artificial populations. Artificial populations 5 and 6 use the same surrogate variables as Artificial Populations 2 and 4 respectively. For these two artificial populations the count was corrected for DSE bias proportional to the single surrogate variable. See the Appendix for details. Among all the combinations of weighted omissions and weighted erroneous enumerations surrogates considered, these were the four that had the highest correlations. Artificial population 4 had the highest correlation among potential artificial populations that excluded remainder surrogates (such as, excludes surrogates formed by subtracting the number of persons with a characteristics such as substituted from the total number of persons). Typical correlations obtained ranged from slightly negative to around 0.26.

Effect of synthetic error on the weighted squared error loss function analysis

There are 96 combinations of bias model (8 models), artificial population (6 populations) and synthetic method (2 methods). Table 3 summarizes the results for state shares. This is presented only to demonstrate the analysis since the A.C.E. data has uncorrected bias. A shaded box indicates the bias correction changes a loss function decision.

- 18 of the 96 combinations have a bias correction which changes the decision.
- 6 of the 8 bias models have 2 combinations with a change
- 14 of the 18 combinations with a change have total error model target distribution method proportional to the undercount
- 16 of these 18 change a decision in favor of adjustment to in favor of the census.
- the 2 of the 18 which change the other way are both for artificial population 3

Summary

- Loss function results do not include a measure of synthetic error
- This paper develops a bias correction using artificial populations that can be added to loss function results to correct for synthetic error
- The purpose of the paper is to present methodology
- Results do not indicate any comparison of Census counts and revised A.C.E. estimates

References

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.

Griffin, R. and Malec D. (2001). "Accuracy and Coverage Evaluation: Assessment of Synthetic Assumption" DSSD Census 2000 Procedures and Operations Memorandum Series B-13*, February 28, 2001.

Hengartner, N. and T.P. Speed (1993). "Assessing Between-Block Heterogeneity Within the Post-Strata of the 1990 Post Enumeration Survey." Journal of the American Statistical Association, 88, 1047-1057.

Kim, J.J., A. Zaslavsky, and R. Blodgett (1995). "Between-State Heterogeneity of Undercount Rates and Surrogate Variables in the 1990 U.S. Census." Survey Methodology, 21, 1, pp.55-62.

Navarro, A. and M. Asiala, (2001). "Accuracy and Coverage Evaluation: Comparing Accuracy." DSSD Census 2000 Procedures and Operations Memorandom Series B-13*, February 28, 2001.

Schindler, E, (2001). " Accuracy and Coverage Evaluation: "Alternative Assessment of Synthetic Assumption.", DSSD Census 2000 Procedures and Operations Memorandum Series Q-xx, September 2001.

APPENDIX

Forming artificial populations

Let X denote a surrogate for weighted non-matches and Y denote a surrogate for weighted erroneous enumerations.

DSE_j = the Dual System Estimate for Post-stratum j

E_j = the weighted E sample total in 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 weighted non-matches as follows:

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

Define the estimated weighted erroneous enumerations as follows:

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

Denote the estimated DSE bias (estimated from the total Error Model including correlation bias) as $\hat{D}_j$

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

$$N_{ij} = Cen_{ij} + X_{ij}\frac{NONMATCH_j}{X_{.j}} - Y_{ij}\frac{ERR_j}{Y_j} - Cen_{ij}\frac{\hat{D}_j}{Cen_{.j}} \quad (1)$$

$$N_{.j} = Cen_{.j} + NONMATCH_j - ERR_j - \hat{D}_j$$
$$= Cen_{.j} + DSE_j - Cen_{.j} - \hat{D}_j$$
$$= DSE_j - \hat{D}_j$$

Equation (1) was used for Artificial Populations 1, 2, 3, and 4. For Artificial Populations 2 and 4, X and Y represented the same variable. In order to consider alternatives that use a surrogate variable instead of the Census counts to allocate the DSE bias, $\hat{D}_j$, Artificial Populations 5 and 6 were created using the single surrogate variable for Artificial Populations 2, and 4 respectively. Denoting the single surrogate variable by X, equation (2) is the artificial population count used for Artificial Populations 5 and 6.

$$N_{ij} = Cen_{ij} + X_{ij}\frac{(DSE_j - Cen_{.j} - \hat{D}_j)}{X_{.j}} \quad (2)$$

The artificial populations were selected by computing the, within post-strata, correlation between the coverage gap, z=(Weighted P-sample Non-matches)- (Weighted E-sample erroneous enumerations), and $N_{ij} - Cen_{ij}$, 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 Δ_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.

Note: For this derivation, assume the same weight is used for the A.C.E. Loss and the Census Loss. For state counts and state shares, the input loss function difference used A.C.E. data for the A.C.E. weight and Census data for the Census weight. For the bias correction term, we assume that Census data was used for both the A.C.E. Loss and the Census Loss. This assumption has negligible effect on results. For CD and County Shares, the input loss function difference used Census data for both the A.C.E. Loss and the Census loss so no assumption is necessary.

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 = $\sum_j \frac{C_{ij}}{C_{.j}}(DSE_j - \hat{D}_j)$

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

$= \sum_j \frac{C_{ij}}{C_{.j}} DSE_j$

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:

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

$$D_t = \sum_i [w_i(Cen_i - N_i)^2 - w_i(\hat{N}_i - N_i)^2]$$

$$= D_g + 2\sum_i w_i(\tilde{N}_i - N_i)(Cen_i - \hat{N}_i)$$

The resulting expected difference is:

$$\Delta_t = \Delta_g + 2\sum_i w_i(\tilde{N}_i - N_i)(Cen_i - a_i)$$

$$= \Delta_g + 2\sum_i w_i(\tilde{N}_i - N_i)(Cen_i - \tilde{N}_i - b_i),$$

So B = 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.

Table 1: Sensitivity Analysis Bias Models for States and Congressional Districts

Model 1 - Corr. Bias Males 18+; 100% Proc. Error
Model 2 - Corr. Bias Males 18+, except Non-Black Males 18-29; 0% Proc. Error
Model 3 - Corr. Bias Males 18+, except Non-Black Males 18-29; 25% Proc. Error
Model 4 - Corr. Bias Males 18+, except Non-Black Males 18-29; 50% Proc. Error
Model 5 - Corr. Bias Males 18+, except Non-Black Males 18-29; 75% Proc. Error
Model 6 - Corr. Bias Males 18+, except Non-Black Males 18-29; 100% Proc. Error
Model 7 - Corr. Bias Black Males 18+; 100% Proc. Error
Model 8 - No Corr. Bias; 100% Proc. Error

For Models 1 through 7 the degree of correlation bias is 100 percent.

Table 2: Surrogate Variables used to Create Artificial Populations

	Correlations (weighted analysis)	**Undercount Surrogate**	**Overcount Surrogate**	**Correction for DSE bias proportional to:**
Artificial Population 1	0.26	# non-substituted persons in households	#persons for whom reported date of birth and reported age were consistent (allocation not required)	Census Counts
Artificial Population 2	0.27	# non-substituted persons in households	# non-substituted persons in households	Census Counts
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)	Census Counts
Artificial Population 4	0.25	# persons whose household did not mail back the questionnaire	# persons whose household did not mail back the questionnaire	Census Counts
Artificial Population 5	0.27	# non-substituted persons in households	# non-substituted persons in households	Surrogate Variable
Artificial Population 6	0.25	# persons whose household did not mail back the questionnaire	# persons whose household did not mail back the questionnaire	Surrogate Variable

Household Persons only (Group Quarters Persons are Excluded)

Table 3. A.C.E. or Census More Accurate for State Shares?

Shaded cell indicates a change in loss function decision due to synthetic bias

Model	DSE Bias[1]	Distr. Method	Synthetic Bias Model (Artificial Population)						
			None	1	2	3	4	5	6
1	Corr. Bias Males 18+; 100 % Proc. Error	DSE	ACE	ACE	ACE	ACE	ACE	ACE	ACE
		UC	ACE	ACE	ACE	ACE	CEN	ACE	CEN
2	Corr. Bias Males 18+, except NB Males 18-29; 0 % Proc. Error	DSE	ACE	ACE	ACE	ACE	CEN	ACE	CEN
		UC	CEN	CEN	CEN	ACE	CEN	CEN	CEN
3	Corr. Bias Males 18+, except NB Males 18-29; 25 % Proc. Error	DSE	ACE	ACE	ACE	ACE	CEN	ACE	CEN
		UC	CEN	CEN	CEN	ACE	CEN	CEN	CEN
4	Corr. Bias Males 18+, except NB Males 18-29; 50 % Proc. Error	DSE	ACE	ACE	ACE	ACE	ACE	ACE	ACE
		UC	ACE	ACE	ACE	ACE	CEN	ACE	CEN
5	Corr. Bias Males 18+, except NB Males 18-29; 75 % Proc. Error	DSE	ACE	ACE	ACE	ACE	ACE	ACE	ACE
		UC	ACE	ACE	ACE	ACE	CEN	ACE	CEN
6	Corr. Bias Males 18+, except NB Males 18-29; 100 % Proc. Error	DSE	ACE	ACE	ACE	ACE	ACE	ACE	ACE
		UC	ACE	ACE	ACE	ACE	CEN	ACE	CEN
7	Corr. Bias Black Males 18+; 100 % Proc. Error	DSE	ACE	ACE	ACE	ACE	ACE	ACE	ACE
		UC	ACE	ACE	ACE	ACE	CEN	ACE	CEN
8	No Corr. Bias; 100 % Proc. Error	DSE	ACE	ACE	ACE	ACE	ACE	ACE	ACE
		UC	ACE	ACE	ACE	ACE	CEN	ACE	CEN

[1]Except for correlation bias, the other bias components are based on 1990 PES evaluations.