UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

December 31, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-39

MEMORANDUM FOR Donna Kostanich
Chair, A.C.E. Revision II A.C.E. Planning Group

From: Mary Mulry DK for MM
Chair, A.C.E. Revision II Assessment Subgroup

Prepared by: Richard A. Griffin
Chief, Estimation Staff
Decennial Statistical Studies Division

Subject: A.C.E. Revision II - Synthetic Error Evaluation Study Plan

1. BACKGROUND

The goal of the A.C.E. Revision II Assessment of the Synthetic Assumption is to (1) estimate bias in synthetic estimates for states and counties of population 100,000 or greater and (2) correct loss function study output for synthetic estimation bias. Point (2) is necessary because the loss function analysis does not include an error component for failure of the synthetic assumption.

Similar assessments of the Synthetic Assumption were done for the original Census 2000 A.C.E. estimates produced in March 2001.

2. QUESTIONS TO BE ANSWERED

This study answers the following questions:

- How much bias is there in A.C.E. Revision II synthetic estimates for states and counties of population 100,000 or more?

  For each state and county with a population of 100,000 or more, the synthetic population bias, post-stratum level DSE bias, total bias, and relative bias will be calculated. Synthetic population bias will be compared with post-stratum level DSE bias.

- What is the effect of a synthetic bias correction on loss function results for A.C.E. Revision II?

The loss function analysis will compute the census loss minus the A.C.E. Revision II loss for eleven groups of geographic areas. For each of the 11 groups the synthetic bias correction will be computed for each of six artificial populations. Thus, for each of 66 combinations of geographic group and artificial population, we will determine if a synthetic bias correction changes the loss function decision.

3. METHODOLOGY

3.1 Bias in Synthetic Estimates

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). Synthetic bias will be measured for states and counties with greater than 100,000 population. For this study plan post-stratum refers to an "estimation cell" which is a cross of an E-sample and P-sample post-stratum. The decomposition of the error in a synthetic estimate into these two components is shown in the Appendix.

The basic methodology used to estimate the synthetic population bias component of synthetic error is artificial populations. We will use the same surrogate variables used for the production A.C.E. synthetic error assessment. These are shown in Table 1. Note that the correlations used to select these variables were computed using production A.C.E. weighted P-sample non-matches and E-sample erroneous enumerations. We will not do a new analysis using A.C.E. Revision II results to potentially choose different surrogate variables. The methodology for creating the artificial population counts is shown in the Appendix.

For each of the six artificial populations for each state level (i.e. estimated count) and county level (only counties greater than 100,000 estimated count), both components of error in synthetic estimates will be computed as well as the ratio of the synthetic population bias to the post-stratum level DSE bias.

3.2 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. An expression for a bias correction to a weighted squared error loss function difference, Loss(Census) - Loss(A.C.E. Revision II), 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 A.C.E. Revision II is favorable, only a negative bias correction can change this making A.C.E. Revision II unfavorable. Similarly, if the difference is negative, indicating A.C.E. Revision II 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.

For A.C.E. Revision II there will be only one set of loss functions. For each geographic grouping detailed in Table 2, the sum of the weighted squared error loss function difference over geographic entities in the group , Loss(Census) - Loss(A.C.E. Revision II), will be provided as input to this evaluation.

Notation:
D = Loss(Census) - Loss(A.C.E. Revision II)
B = synthetic bias correction term
B/D = relative bias in D due to loss functions not including a synthetic error component
D + B = Bias corrected loss function difference.

For each of the six artificial population, for each grouping in Table 2, D, B, B/D, and D + B will be produced. Thus for a given grouping and artificial population we will know if the synthetic bias correction would change a loss function decision.

Table 1: Surrogate Variables used to Create Artificial Populations

	Correlations (weighted analysis from A.C.E. production)	**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 2 : Groupings for the loss functions*

Levels
All Counties with population of 100,000 or less
All Counties with population greater than 100,000
All places with population at least 25,000 but less than 50,000
All places with population at least 50,000 but less than 100,000
All places with population greater than 100,000
Shares within state
All Counties
All places
Shares within U.S.
All places with population at least 25,000 but less than 50,000
All places with population at least 50,000 but less than 100,000
All places with population greater than 100,000
All states

*Loss functions are weighted. The weight is the reciprocal of the census count in the area.

4. DATA REQUIREMENTS

- Surrogate variable file from the Decennial Statistical Studies Division (DSSD) providing detailed post-stratum and geographic codes
- File from the Planning, Research, and Evaluation Division (PRED) with total error model person DSE bias estimate for each detailed post-stratum
- File from DSSD with Loss Function output for each geographic group studied
- A.C.E. Revision II DSE census, E-sample, and P-sample output files from DSSD

5. DIVISION RESPONSIBILITIES

- DSSD will be responsible for managing the study, preparing the study plan, and writing the report.
- The Statistical Research Division (SRD) will do the programming for state and county level estimates of synthetic bias and for computing the artificial population bias correction estimates.
- DSSD and SRD will collaborate on methodology and evaluation.

6. MILESTONE SCHEDULE

Activity	**Person(s) Responsible**	**Planned Start**	**Planned Finish**
Study Plan	Rick	11/21/02	12/17/02
DSE Bias Input Files	Katie	11/01/02	12/17/02
Surrogate Variable input file	Randy	11/22/02	12/16/02
Loss Function Input File	Randy	11/01/02	12/20/02
DSE Input files	Dawn	10/01/02	12/16/02
SRD Programming (includes testing)	Don	11/22/02	12/17/02
SRD Output	Don	12/17/02	12/20/02
Analysis	Rick and Bob		
Draft Report		12/23/02	12/24/02
Final Report		12/30/02	12/31/02

7. LIMITATIONS

Artificial populations were created using surrogate variables, available for small areas, correlated with gross undercount and/or gross overcount. The surrogate variables are not the variable of interest and the correlations of the selected surrogates were smaller than we would have preferred. No artificial population provides the true population count for any geographic area.

8. RELATED STUDIES

- SRD will do a loss function analysis that does not consider synthetic bias. Output from the Loss Function Analysis is input to this evaluation.

- SRD will do a A.C.E. Revision II Error Model Study. Output from the A.C.E. Revision II Error Model Study is input to this evaluation.

9. REFERENCES

Griffin, R. And Malec, D. (2001). "Accuracy and coverage Evaluation: Assessment of the Synthetic Assumption." DSSD Census 2000 Census Procedures and Operations Memorandum Series B-14*

Griffin, R. And Malec, D. (2001). "Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report 23: Sensitivity Analysis for the Assessment of the Synthetic Assumption." DSSD Census 2000 Census Procedures and Operations Memorandum Series Q-72*

APPENDIX

1. Forming artificial populations

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

j indicates a non-zero post-stratum formed by the crossing of the E and P sample post-stratifications. Each j will be associated with a E-sample component based on the E-sample post-stratification and a P-sample component based on the P-sample post-stratification.

DSE_j = the Dual System Estimate for Post-stratum j

E_j = the weighted E sample total associated with post-stratum j

CE_j = the weighted E sample number of correct enumerations associated with post-stratum j

EE_j = the weighted E sample number of erroneous enumerations associated with 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 associated with post-stratum j as follows:

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

Define the estimated weighted erroneous enumerations associated with post-stratum j as follows:

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

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

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

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

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

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.

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

2. Decomposition of the Error in a Synthetic Estimate into Two Additive Components.

Notation

$N_{i.}$ = the true population for area i

cen_{ij} = census count for area i, post-stratum j

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

$CF_{.j} = \frac{N_{i.}}{cen_{.j}}$ = true coverage correction factor for post-stratum j

$\hat{CF}_{.j} = \frac{DSE_j}{cen_{.j}}$ = estimated coverage factor for post-stratum j

$\hat{N}_{i.} = \sum_j \hat{CF}_{.j}\ cen_{ij}$ = the A.C.E. Revision II synthetic estimate for area i

$\tilde{N}_{i.} = \sum_j CF_{.j} cen_{ij}$ = the known population synthetic estimate for area i

Then $\hat{N}_{i.} - N_{i.} = (\tilde{N}_{i.} - N_{i.}) + (\hat{N}_{i.} - \tilde{N}_{i.})$

Define:

$B_i = E(\hat{N}_{i.} - N_{i.})$, the bias in the synthetic estimate

$SynB_i = \tilde{N}_{i.} - N_{i.}$, the error due to carrying down the true post-stratum coverage correction factors to area i. Since the true coverage correction factors are used, bias in the DSE at the post-stratum level is excluded from this error.

$DSEB_i = E(\hat{N}_{i.} - \tilde{N}_{i.})$, the error due to using the estimated coverage correction factors instead of the true coverage correction factors for each post-stratum. This error is due to bias in the

DSE including correlation bias.

3. Specifying Bias due to Synthetic Estimation

The first component of the synthetic bias is estimated using artificial populations, the second component is estimated using post-stratum biases, estimated as part of the A.C.E. Revision II Error Model and Loss Function work. The estimate of bias for area i takes the following form:

$$\hat{B}_i = Sy\hat{n}B_i + DS\hat{E}B_i = (\tilde{N}_{i\,.} - N_{i\,.}) + \sum_j \frac{Cen_{ij}}{Cen_{.\,j}} \hat{D}_j .$$

Here, the first part is estimated from an artificial population; it is the known artificial population synthetic count minus the actual population count from the artificial population.

The second part contains the post-stratum bias, $\hat{D}_j$, (estimated elsewhere) which is an estimate of: (E(DSE_j)-the true population of post-stratum j). The true population of post-stratum j is estimated using results from the A.C.E. Revision II Error Model Analysis. In this second term, we weight the post-stratum bias by the proportion of post-stratum census counts in area i .

4. Correction for Synthetic Bias in Loss Function Analysis

Notation:

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

D_t = the census squared error loss minus the A.C.E. Revision II 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. Revision II Loss and the Census 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 = $\sum_j \frac{C_{ij}}{C_{.\,j}}(DSE_j - \hat{D}_j)$

$\hat{N}_i$ = the A.C.E. Revision II 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 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.