BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-88/17

MODEL BIAS AND SAMPLING ERROR CONSIDERATIONS IN SMALL AREA COVERAGE ERROR ESTIMATION

by

Cary T. Isaki, Elizabeth T. Huang and Linda K. Schultz
Statistical Research Division
Bureau of the Census
Room 3132, F.O.B. #4
Washington, D.C. 20233 U.S.A.

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended by: Nash J. Monsour

Report completed: July 1988

Report issued: July 1988

Model Bias and Sampling Error Considerations in Small Area Coverage Error Estimation

by

Cary T. Isaki, Elizabeth T. Huang and Linda K. Schultz

Abstract: The effect of model bias on the performance of small area estimators of population coverage in a census situation is examined. The effects of sample based estimates of variance in regression smoothing are also investigated. The results are compared with previously obtained results under ideal situations.

I. Introduction

As part of an evaluation program of the 1990 Decennial Census, small area estimates of the population will be produced, at least at the county level. Given the large number of counties, 3137 and the budgeted sample size of 150,000 housing units for a post enumeration survey (PES), direct estimates for each county are not possible.

Instead, synthetic and regression estimation procedures are planned for use in estimation. A PES will provide direct estimates of the population for approximately 100 geographic areas crossed in some instances by race and tenure (in the largest metropolitan areas). The PES will be an independently designed block sample, unlike the 1980 effort that piggy-backed onto a monthly labor force survey sample. This sampling strategy has some advantages to be discussed shortly.

By direct estimates we mean a dual system estimate or a capture-recapture method. We use the term direct estimate to contrast with synthetic estimate which is used to obtain small area estimates. A dual system estimation procedure consists of two independent samples of the population of interest and a matching of units to identify common elements in both samples. In the coverage evaluation context the census is viewed as one sample and the PES is viewed as the other sample. Matching of persons in the PES to that in the census causes some difficulty when information used to match is incomplete. The block sample of persons makes matching less error prone by concentrating the geographic search area.

In addition to matching error, the dual system estimator is subject to misclassification error, correlation bias and other sources of error. We focus on the above three sources of error in the bias of the dual system estimator and its effect on the synthetic estimator. Previous work on coverage error estimation had assumed that the dual system estimator was unbiased.

The synthetic estimator of total population for an area α is composed of two principal parts. One part consists of census count categories for area α while the other consists of coverage evaluation factors that are ratios of the dual system estimates and the corresponding census counts. For example, a county in the northeastern rural part of the country may be composed of two dozen age-race-sex categories each with coverage evaluation factors based on the entire northeastern rural portion of the U.S. The synthetic estimator of total for the county would be the sum of its component census counts multiplied by the associated coverage evaluation factors.

It has been proposed by Tukey (1981) and others that the coverage evaluation factors be smoothed before application in synthetic estimation. The smoothing procedure involves modelling the coverage evaluation factors via regression and averaging the predicted factors with the original factors inversely proportional to their variances. Previous work using simulation has tended to support this procedure. However, in the simulation, sample survey variances were assumed known.

We extend the above work by using estimated variances in regression smoothing and assess its effects on small area coverage estimates. The assessment is done with respect to two artificial populations labelled AP2 and AP3. We resorted to constructing artificial populations because the census is the only source providing detailed small area counts and it is also being assessed as to its performance.

We chose the variable census substitutions as a proxy for the number of missed persons. Census substitutions are the result of imputing people into housing units. For example, people were substituted into the census 1) when no form was completed but people may have lived in the housing unit, 2) when we know only the number of people living in the unit, 3) for machine failure or 4) when field counts for an area (enumeration district (ED) or block) were larger than the processed counts. Preliminary analysis using 1980 Post Enumeration Program (PEP) state data indicated that the census substitution rate was the most important explanatory variable of several types of nonmatch rates in the PEP. The nonmatch rate in the PEP refers to the ratio of estimated total number of persons in the PEP not matched to the census to the PEP estimated total number of persons. Since the nonmatch rate estimates the miss rate of the census (under ideal conditions) and census substitutions were available by age-race-sex at the ED level we focused on census substitutions as a proxy for undercount.

The artificial populations are constructed by age-race-sex at the ED level.

$$\text{AP2} = \text{census count} + F_{DA1} \times \text{substitution count}$$

$$\text{AP3} = \text{census count} + F_{DA2} \times \text{substitution count.}$$

The F_{DA} factors are defined so that the U.S. total equals that from Demographic Analysis with an assumed illegal count of 3.5 million. Basically, AP2 and AP3 differ in the treatment of Hispanics. Hispanics in AP2 are assumed to have an undercount similar to that for non-Blacks while in AP3, Hispanics are assumed to have an undercount similar to Blacks. We use the artificial populations to assess the performance of the synthetic estimator.

II. Dual System Estimation Model

In this section we examine the effects of bias in the estimator of coverage evaluation factors due to response correlation, matching error and classification error. We then utilize the biassed coverage evaluation factors in a synthetic estimator for estimating total population of such small areas as EDs and counties using the artificial populations. These results are then compared to the results when unbiased coverage evaluation factors are assumed.

The model underlying the bias of the dual system estimator is based on a multinomial distribution with 3 sets of parameters. The first set of parameters deals with probabilities of response to the census and the PES and response correlation by race (Black, White). The second set of parameters deals with classification error. Here, probabilities of correctly classifying the race of a unit are introduced for both the census and the PES. The proportion of race in the population and the true size of the population is introduced in the second set as well. The third set of parameters deals with matching error. We consider here, the probability of a match given the same classification states (should possibly be matched) while the other parameter is the probability of a match given different classification states (should possibly not be matched).

Set 1.

a. Let (R,R) denote that a unit has been captured by the census and the PES, respectively. Define (R,NR), (NR,R) and (NR,NR) in the same manner where NR stands for non-capture.

b. Let P_W (R,R) be the probability of capture of a unit that is truly White and

$P_W(R,R) = P_{WC}P_{WP} + \rho_W g_W$ where

P_{WC} = Probability of a White person responding to the Census,

P_{WP} = Probability of a White person responding to the PES,

ρ_W = Correlation between responses to the Census and the PES,

$$g_W = (P_{WC}\ \overline{P}_{WC}\ P_{WP}\ \overline{P}_{WP})^{\frac{1}{2}} \quad , \quad \overline{P}_{WC} = 1-P_{WC}$$

P_B (R,NR), etc. are defined similarly to that of Whites.

Set 2.

a. Let (W,W) denote the classification of a respondent to both the census and PES as White.

b. Let f_W denote the probability of a White respondent being classified White in the census and s_W denote the probability of a White respondent being classified White in the PES.

c. Let P_W (W,W) denote the probability of a truly White person being classified White in both the census and the PES and that the person has responded to both the census and the PES-

$P_W[(W,W)] = P_W(R,R)\ f_W s_W$. Likewise, the eight other probabilities, $P_W[(W,NR)]$, ..., $P_W[(NR,NR)]$ and $P_B[(B,B)]$, ..., etc. for Black can be obtained.

d. Let N_W and N_B be the true number of Whites and Blacks, respectively, $N_W = N x \alpha$ where $N = N_W + N_B$.

Set 3.

a. Let X_i, i = 1, 2, ..., 9 denote the number of Whites classified as (W,W), (W,B), (B,W), (B,B), (W,NR), (B,NR), (NR,W), (NR,B) and (NR,NR). Let Y_i, i = 1, ..., 9 hold for Blacks.

b. Assume $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ are each multinomially distributed N_W and N_B, respectively with probabilities $P_W[(W,W)]$, ..., $P_W[(NR,NR)]$ and $P_B[(W,W)]$, ..., $P_B[(NR,NR)]$.

c. Let $\underset{\sim}{T} = \underset{\sim}{X} + \underset{\sim}{Y}$. Then $\underset{\sim}{T}$ represents the distribution of couplets prior to matching. We do not observe the individual components of $\underset{\sim}{T}$.

d. We do observe $T_1 + T_2 + T_5$ for census classified White and $T_1 + T_3 + T_7$ for PES classified White. For Blacks we observe $T_3 + T_4 + T_6$ in the census and $T_2 + T_4 + T_8$ in the PES.

e. Let ϕ_M denote the probability of a match in T_1 and T_4 cases where the individual responds to both the census and the PES and is classified identically. Let ϕ_0 denote the probability of a match between other respondents to the census and respondents to the PES classified with the same race (e.g., White).

f. Assume that the expected number of matches for estimating total White conditional on capture and classification is

$$E[m_W] = T_1\ \phi_M + \phi_0 \min\left(T_2 + T_5,\ T_3 + T_7\right)$$

while that for Blacks is

$$E[m_B] = T_4 \phi_M + \phi_0 \min (T_3 + T_6, T_2 + T_8).$$

If we let the number of census classified White be denoted N_{WC}, then

$N_{WC} = T_1 + T_2 + T_5$

$N_{WP} = T_1 + T_3 + T_7$ in the PES

$N_{BC} = T_3 + T_4 + T_6$ classified Black in the census

$N_{BP} = T_2 + T_4 + T_8$ in the PES.

The usual dual system estimator for total White is then $\text{d.s.e.}_W = m_W^{-1} N_{WC}N_{WP}$ and for Black, $\text{d.s.e.}_B = m_B^{-1} N_{BC}N_{BP}$. To obtain the bias of the d.s.e., we assume that $E[\text{d.s.e.}_W] \doteq E[m_W]^{-1} E[N_{WC}] E[N_{WP}]$ where for example,

$$E[T_1] = E[X_1 + Y_1] = N_W P_W[(W,W)] + N_B P_B[(W,W)]$$

and

$$E[m_W] \doteq \phi_M E[T_1] + \phi_0 \min (E[T_2 + T_5], E[T_3 + T_7]).$$

A fuller description of the model can be found in Isaki (1988).

III. Model Bias - Tweaked Factors

Specifying parameter values the above expressions were used to produce biased coverage evaluation factors and these biased factors were used in synthetic estimation using the artificial populations. In this way, small area synthetic estimation results using biased factors were produced for small areas in the state of New Jersey. The true factors and the biased factors (termed tweaked factors) are presented in Table 1 for AP2 and AP3. Another set of factors was also examined and produced similar results. In Table 2, set of factors was also examined and produced similar results. In Table 2, the measures of performance of Syn 2 coverage estimates using the tweaked factors and the true factors are presented along with the census. The measures of performance are presented in the Appendix.

Table 1. **Tweaked Factors for New Jersey AP2/AP3**

I. AP2 ($F_W = F_B = .999$; $\phi_M = .995$; $\phi_0 = .03$, $P_{WC} = P_{WP}$; $P_{BC} = P_{BP}$)

15 Factors	True Factor	P_C	α_W	ρ	Tweaked Factor
A. N.Y.C.					
1. Black	1.074	.931	.70	.25	1.061
2. Hispanic	1.014	.986	.70	.20	1.018
3. Rest	1.011	.989	.60	.10	1.014
B. Central City 250,000+					
4. Black	1.106	.904		.25	1.082
5. Rest	1.008	.992	.60	.10	1.011
C. Central City 50,000-250,000					
6. Black	1.072	.932		.25	1.063
7. Rest	1.008	.992	.80	.10	1.010
D. Central City 50,000+					
8. Hispanic	1.016	.984	.90	.20	1.014
E. N.Y.C. Balance of SMSA					
9. Rest	1.005	.995	.90	0	1.008
F. Balance of SMSA 250,000+ excluding N.Y.C.					
10. Rest	1.004	.996	.90	0	1.007
G. Balance of SMSAs 250,000+					
11. Black and Hispanic	1.033	.968	.90	.30	1.025
H. Balance of SMSA with Central City 50,000-250,000					
12. Rest	1.006	.994	.95	.10	1.009
I. Cities 10,000-50,000					
13. Rest	1.005	.995	.95	.10	1.009
J. Rural					
14. Rest	1.008	.992	.95	.10	1.011
K. Remainder					
15. Black and Hispanic	1.021	.979		.20	1.019

II. AP3 ($F_W = F_B = .999$; $\phi_M = .977$; $\phi_0 = .03$, $P_{WC} = P_{WP}$, $P_{BC} = P_{BP}$)

15 Factors	True Factor	P_C	α_W	ρ	Tweaked Factor
A. N.Y.C.					
1. Black	1.074	.931	.70	.25	1.086
2. Hispanic	1.061	.943	.70	.20	1.077
3. Rest	1.008	.992	.60	.10	1.032
B. Central City 250,000+					
4. Black	1.106	.904		.25	1.104
5. Rest	1.006	.994	.60	.10	1.030
C. Central City 50,000-250,000					
6. Black	1.072	.933		.25	1.085
7. Rest	1.005	.995	.80	.10	1.028
D. Central City 50,000+					
8. Hispanic	1.071	.934	.90	.20	1.099
E. N.Y.C. Balance of SMSA					
9. Rest	1.003	.997	.90	0	1.027
F. Balance of SMSA 250,000+ excluding N.Y.C.					
10. Rest	1.003	.997	.90	0	1.027
G. Balance of SMSAs 250,000+					
11. Black and Hispanic	1.039	.962	.90	.30	1.050
H. Balance of SMSA with Central City 50,000-250,000					
12. Rest	1.003	.997	.95	.10	1.026
I. Cities 10,000-50,000					
13. Rest	1.003	.997	.95	.10	1.026
J. Rural					
14. Rest	1.005	.995	.95	.10	1.028
K. Remainder					
15. Black and Hispanic	1.030	.971		.20	1.088

Table 2. Measures of Performance for New Jersey Counties and EDs for AP2 and AP3 Using Known and Tweaked Factors

I. AP2 - Measures

Counties (21)	Syn 2	Syn 2 - tweaked	Census
1. Number of counties where ARE(c_i) < ARE(e_i)	4	6	
2. Number of counties where ADP(c_i) < ADP(e_i)	4	3	
3. MARE	.0070	.0080	.0131
4. Max (ARE)	.0399	.0460	.0716
5. α	1866	2356	5752
6. SADP	.0114	.0126	.0161
7. PI	.779	.847	
8. ϕ	1770	2254	3598
9. IMP1 x 10^{-3}	.2401	.3058	.5015

EDs (7657)			
1. Number of EDs where ARE(c_i) < ARE(e_i)	4165	4541	
2. Number of EDs where ADP(c_i) < ADP(e_i)	1549	1438	
3. MARE	.0148	.0158	.0153
4. Max (ARE)	.6293	.6347	.6587
5. α	10774	11866	17809
6. SADP	.0173	.0188	.0232
7. PI	.804	.822	
8. ϕ	10678	11764	15656
9. IMP1 x 10^{-3}	1.4485	1.5961	2.1822

II. AP3 - Measures

Counties (21)	Syn 2	Syn 2 - tweaked	Census
1. Number of counties where ARE(c_i) < ARE(e_i)	4	18	
2. Number of counties where ADP(c_i) < ADP(e_i)	3	3	
3. MARE	.0073	.0249	.0129
4. Max (ARE)	.0444	.0359	.0793
5. α	2240	5215	6835
6. SADP	.0122	.0136	.0178
7. PI	.847	.847	
8. ϕ	2137	2843	4452
9. IMP1 x 10^{-3}	.2896	.3693	.6210
EDs (7657)			
1. Number of EDs where ARE(c_i) < ARE(e_i)	4192	6280	
2. Number of EDs where ADP(c_i) < ADP(e_i)	1458	1545	
3. MARE	.0150	.0326	.0157
4. Max (ARE)	.6643	.6619	.6930
5. α	12200	16264	20336
6. SADP	.0181	.0194	.0255
7. PI	.818	.819	
8. ϕ	12097	13891	17952
9. IMP1 x 10^{-3}	1.6399	1.8041	2.5045

Except for the ARE related measures in AP3, the tweaked factors Syn 2 estimator remains superior to the census. While the measures have been degraded, they still indicate the superiority of Syn 2. Similar results were obtained for places.

IV. Smoothing - Regression Estimator

We consider using regression estimation in two separate ways. In the first way, we model the estimated coverage evaluation factors via regression. In the second way, we model the estimated percent net coverage error at the state level and use it to estimate the net coverage error for counties (this latter invoking a synthetic assumption concerning state versus county relationships).

We briefly outline the regression approach using terminology relevant to the second situation above. Let S_i denote the true population count and C_i denote the census count for state i. Let $\underset{\sim}{Y} = (Y_1, \ldots, Y_{51})^T$ denote the vector of true net coverage error $(Y_i = (S_i - C_i)/S_i)$ and assume that the regression model

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \quad , \quad \underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}\ ,\ \sigma^2 \underset{\sim}{I}) \tag{1}$$

holds, where $\underset{\sim}{X}$ is a 51 x p matrix and $\underset{\sim}{\beta}$ is a p x 1 vector. Assume also, that a PES is conducted to measure $\underset{\sim}{Y}$ and let $\hat{\underset{\sim}{Y}} \sim N(\underset{\sim}{Y}, \underset{\sim}{D})$ where $\underset{\sim}{D}$ is a diagonal variance - covariance matrix whose diagonal elements d_i are the sampling variances of $\hat{\underset{\sim}{Y}}$. In our context, $\hat{Y}_i = (\hat{S}_i - C_i)/\hat{S}_i$ is the estimated net undercount and the $\hat{S}_i$ are synthetic estimates of state total population using estimated coverage evaluation factors based on dual system estimation. Combining, we have

$$\hat{\underset{\sim}{Y}} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{u} \text{ where } \underset{\sim}{u} \sim N(\underset{\sim}{0},\ \sigma^2\ \underset{\sim}{I} + \underset{\sim}{D}). \tag{2}$$

In previous work we have assumed that $\underset{\sim}{D}$ is known and estimated σ^2 and $\underset{\sim}{\beta}$ using maximum likelihood.

Ericksen and Kadane (1985) using a hierarchical Bayesian framework developed by Lindley and Smith (1972) proposed the estimation of net undercount of states, cities and balance of states using $\hat{\underset{\sim}{Y}}_{EK}$ where

$$\hat{\underset{\sim}{Y}}_{EK} = [\underset{\sim}{D}^{-1} + \sigma^{-2}\ \underset{\sim}{I}]^{-1}\ [\underset{\sim}{D}^{-1}\ \hat{\underset{\sim}{Y}} + \sigma^{-2}\ \underset{\sim}{X}\ \hat{\underset{\sim}{\beta}}] \tag{3}$$

and the use of $\underset{\sim}{X}\ \hat{\underset{\sim}{\beta}}$ to estimate for smaller areas (those areas where PES estimates are not available or where sampling error is too large). Fuller and Harter (1987), using a components of variance model approach, obtained similar results. Our present interests are in the effect of using estimated d_i in (3).

In the next section, we present the results using estimated d_i as well as known d_i using a sample replicate. We begin with the results of modelling coverage evaluation factors termed <u>Smoothed Factors</u> (or Smoothed F). The other results are termed <u>Smoothed State</u> (or Smoothed S).

A. Smoothed Factors

Using sample replicate data and the estimated sampling variances, $\hat{d}_i$, the re-estimated regressions for AP2 and AP3 smoothed coverage evaluation factors are

$$\hat{y}_i^{AP2} = 1.013 + .017\, X_{Bi} - .005\, X_{Ri} - .003\, X_{G1i}$$

$$\hat{\sigma}^2_{AP2} = .592 \times 10^{-5} \qquad \text{and} \tag{4}$$

$$\hat{y}_i^{AP3} = 1.005 + .023\, X_{NWi} - .002\, X_{G2i}$$

$$\hat{\sigma}^2_{AP3} = .307 \times 10^{-5}$$

where X_{Bi}, X_{Ri} and X_{NWi} are the proportions of Black, non-Black or non-Hispanic, and Black or Hispanic persons in the associated geographic category, respectively. The X_{G1i} and X_{G2i} are collections of divisional indicator variables. The previous regressions, using known d_i, were

$$\hat{y}_i^{AP2} = 1.019 + .040\ X_{Bi} -.007\ X_{Ri} -.006\ X_{G1i}$$

$$\hat{\sigma}^2_{AP2} = .939 \times 10^{-5} \quad \text{and} \qquad (5)$$

$$\hat{y}_i^{AP3} = 1.008 + .052\ X_{NWi} -.004\ X_{G2i}$$

$$\hat{\sigma}^2_{AP3} = .429 \times 10^{-5}\ .$$

There have been changes in the estimated regression coefficients and in the model variances. The average of the estimated d_i for both AP2 and AP3 are both lower than the average of the d_i themselves. This alone tends to place less weight on the regression in the construction of the smoothed factors. However, the model variances are also smaller and this tends to counter-act the effect of the smaller estimated d_i.

In Table 3 below, we illustrate the performance of the smoothed factor method in providing coverage evaluation estimates of total population for states and for counties. The results for both states and county indicate some loss in performance using Smoothed F Syn 2 ($\hat{d}_i$) (smoothed coverage evaluation factors with estimated d_i) over the known d_i method. If proportions are of interest, Smoothed F Syn 2 ($\hat{d}_i$) is generally superior to Syn 2 but the α measure indicates that Syn 2 is somewhat superior in the case of large units for ARE.

B. Smoothed State

In a previous report, Isaki, et al. (1988) a smoothed state model was presented assuming known sampling variances, $\underset{\sim}{D}$. The resulting regression equations were

$$\hat{Y}_i^{AP2} = -.709 + .224\, Z_{Ai} + .096\, Z_{Ri}$$

$$\hat{\sigma}^2_{AP2} = .083 \qquad (6)$$

$$\hat{Y}_i^{AP3} = -.257 + .069\, Z_{Mi} + .094\, Z_{Ai}$$

$$\hat{\sigma}^2_{AP3} = .003$$

where

Y_i is the percent net undercount at the state level

Z_{Ai} is the percent allocation

Z_{Ri} is the percent minority renter and

Z_{Mi} is the percent minority.

Using the estimated sampling variances, $\hat{\underset{\sim}{D}}^o$, the comparable regressions are

$$\hat{Y}_i\, {}^{AP2}_{\hat{d}_i} = -.406 + .171\, Z_{Ai} + .077\, Z_{Ri}$$

$$\hat{\sigma}^2_{AP2} = .178$$

and (7)

$$\hat{Y}_i\, {}^{AP3}_{\hat{d}_i} = -.143 + .0387\, Z_{Mi} + .099\, Z_{Ai}$$

$$\hat{\sigma}^2_{AP3} = .082$$

Using the estimated sampling variances instead of the known variances, the $\hat{\sigma}^2$ values for both AP2 and AP3 in (7) increased over that in (6). Furthermore, the estimated sampling variances are typically smaller than the known variances. This translates to giving less weight to the regression model and more weight to the sample estimates when using the smoothed EK Bayes method shown in (3).

A comparison of the measures of improvement of the synthetic estimators compared to the census using both known and estimated variances can be made using Tables 3 and 4.

Table 3. Measures of Performance of Some Coverage Evaluation Estimators for States for AP2 and AP3

(51 states)

A. AP2

	Measure		Smoothed F	Smoothed F		Smoothed S	Smoothed S	Smoothed	
		Syn 2	Syn 2 (d_i)	Syn 2 ($\hat{d}_i$)	Syn DA	Syn 2 (d_i)	Syn 2 ($\hat{d}^o_i$)	EK ($\hat{d}^o_i$)	Census
1.	No. of states where ARE(c_i) < ARE(e_i)	6	5	1	8	5	4	3	-
2.	No. of states where ADP(c_i) < ADP(e_i)	20	14	15	13	11	9	13	-
3.	MARE	.0060	.0054	.0070	.0053	.0039	.0043	.0053	.0147
4.	Max (ARE)	.0218	.0350	.0550	.0288	.0131	.0258	.0255	.0771
5.	α	12189	9333	21896	9282	8160	8828	8686	77310
6.	SADP	.0056	.0046	.0051	.0048	.0045	.0045	.0051	.0067
7.	PI	.481	.687	.607	.757	.646	.730	.671	-
8.	ϕ	11985	8823	10942	9282	8007	8265	8432	17368
9.	MP1 x 10^{-3}	.0525	.0388	.0487	.0408	.0351	.0364	.0371	.0788

B. AP3

	Measure		Smoothed F	Smoothed F		Smoothed S	Smoothed S	Smoothed	
		Syn 2	Syn 2 (d_i)	Syn 2 ($\hat{d}_i$)	Syn DA	Syn 2 (d_i)	Syn 2 ($\hat{d}^o_i$)	EK ($\hat{d}^o_i$)	Census
1.	No. of states where ARE(c_i) < ARE(e_i)	8	7	1	8	7	5	4	-
2.	No. of states where ADP(c_i) < ADP(e_i)	17	11	9	9	12	8	11	-
3.	MARE	.0060	.0048	.0072	.0049	.0041	.0048	.0051	.0136
4.	Max (ARE)	.0362	.0355	.0569	.0290	.0186	.0393	.0391	.0773
5.	α	19227	10608	29028	9180	7650	12895	12071	82339
6.	SADP	.0068	.0046	.0059	.0046	.0042	.0048	.0052	.0078
7.	PI	.635	.696	.852	.703	.639	.770	.667	-
8.	ϕ	18968	9843	13878	9129	7650	9716	10085	22048
9.	MP1 x 10^{-3}	.0822	.0433	.0619	.0401	.0335	.0430	.0445	.1000

Table 4. Measures of Performance of Some Coverage Evaluation Estimates for Counties for AP2 and AP3

(3137 counties)

A. AP2

		Smoothed F	Smoothed F		Smoothed S	Smoothed S	
Measures	Syn 2	Syn 2 (d_i)	Syn 2 ($\hat{d}_i$)	Syn DA	Syn 2 (d_i)	Syn 2 ($\hat{d}_i^o$)	Census
1. No. of counties where ARE(c_i) < ARE(e_i)	1104	1194	815	1254	1048	939	-
2. No. of counties where ADP(c_i) < ADP(e_i)	999	947	1067	862	864	804	-
3. MARE	.0092	.0087	.0081	.0087	.0081	.0078	.0128
4. Max (ARE)	.2200	.2179	.2196	.2192	.2072	.2102	.2236
5. α	44859	37958	54725	36703	33566	33961	115606
6. SADP	.0093	.0086	.0097	.0085	.0078	.0080	.0115
7. PI	.625	.710	.712	.702	.698	.722	-
8. ϕ	44515	37330	43771	36703	33566	33376	55663
9. MP1 x 10^{-3}	.1953	.1650	.1950	.1617	.1471	.1470	.2525

B. AP3

		Smoothed F	Smoothed F		Smoothed S	Smoothed S	
Measures	Syn 2	Syn 2 (d_i)	Syn 2 ($\hat{d}_i$)	Syn DA	Syn 2 (d_i)	Syn 2 ($\hat{d}_i^o$)	Census
1. No. of counties where ARE(c_i) < ARE(e_i)	1122	1221	806	1358	1154	1141	-
2. No. of counties where ADP(c_i) < ADP(e_i)	821	721	668	702	753	640	-
3. MARE	.0081	.0075	.0072	.0077	.0075	.0069	.0111
4. Max (ARE)	.3007	.2998	.3024	.2720	.2672	.2830	.3065
5. α	61485	44545	70929	41095	37330	45993	134476
6. SADP	.0098	.0087	.0105	.0084	.0078	.0088	.0131
7. PI	.680	.762	.812	.743	.736	.787	-
8. ϕ	61172	43918	55779	41045	37017	42891	74186
9. MP1 10^{-3}	.2676	.1934	.2490	.1798	.1630	.1897	.3366

It can be seen from Tables 3 and 4 that Syn 2 is inferior to Syn DA in all cases when the sample replicate is used for both states and counties. This is a sharp contrast to the case of known factors (not shown; See Isaki, et al. (1987)). Syn DA is competitive with some of the smoothing methods using both known and estimated variances. In general, MARE, α, SADP, ϕ and IMP1 perform in the same manner. Smoothed state and smoothed EK Bayes (not shown) perform the best followed by smoothed state $\hat{d}_i^o$ and smoothed EK Bayes $\hat{d}_i$. Smoothed factor, Syn DA, smoothed factor $\hat{d}_i$ round out the group. In general, smoothed factor $\hat{d}_i$ performs poorly. The PI measure behaves differently than the other measures with smoothed factor $\hat{d}_i$ as one of the better methods. Given the definition of the PI measure, one can see how it is possible to get misleading results. Since a state or county population count is either all included or all not included, there is no in between.

V. Summary

We have attempted to illustrate two effects of model considerations in coverage estimation. Using a model for expectation in dual system estimation and using various parameter values the effect of the resulting bias in the coverage evaluation factors was not so severe as to seriously affect the measures of performance. The one exception is the MARE measure for counties and EDs using AP3 where the factors were biased upward.

In the case of using estimated sampling variances in regression modelling, the net effect on the measures of performance was a poorer performance than when known variances were used. The results indicated however, that the smoothed state method remained superior to the other methods. At any rate, some type of smoothing is likely to improve over the performance of Syn 2.

Our evaluation results are dependent on the AP2 and AP3 populations that were used as a standard. Lacking the correct counts and having the need to assess the performance of the various potential coverage evaluation methods, we proceeded in the manner described. The reader will note that Syn DA performed rather well for states and counties. It did not do as well when places or EDs were examined. In addition, it is unlikely to do well in central city places where minority undercount is expected to be quite a bit different from the national undercount rates.

Finally, the PES estimates are not planned to be the final coverage estimates. Rather, auxiliary information to be provided by demographic analysis estimates are to be used as well. The nature of this combination is currently being researched.

References

1. Chandrasekaran, C. and Deming, W. (1949). "On a Method of Estimating Birth and Death Rates and the Extent of Registration", Journal of the American Statistical Association, 44, pg. 101-115.

2. Citro, C.F. and Cohen, M.L. (1985). "The Bicentennial Census - New Directions for Methodology in 1990", National Academy Press.

3. Citro, C.F. and Pratt, J.W. (1986). "The USA's Bicentennial Census: New Directions for Methodology in 1990", Journal of Official Statistics, 2 (4), pg. 359-380.

4. Cowan, C.D. (1984). "The Effects of Misclassification on Estimates From Capture-Recapture Studies", Ph.D. dissertation, George Washington University, Graduate School of Arts and Sciences, Washington, D.C.

5. Cramer, H. (1963). Mathematical Methods of Statistics. Princeton University Press: Princeton.

6. Ericksen, E. and Kadane, J. (1985). "Estimating the Population in a Census Year: 1980 and Beyond", Journal of American Statistical Association, 80, pg. 98-109.

7. Freedman, D.A. and Navidi, W.C. (1986). "Model for Adjusting the Census", Statistical Science, 1, pg. 3-11.

8. Fuller, W.A. and Harter, R. (1987). "The Multivariate Components of Variance Model for Small Area Estimation", in Small Area Statistics - An International Symposium, John Wiley and Sons: New York

9. Isaki, C.T. (1986). "Bias of the Dual System Estimator and Some Alternatives", Communications in Statistics - Theory and Methods, 15 (5), pg. 1435-1450.

10. Isaki, C.T., Diffendal, G.J. and Schultz, L.K. (1987). "Statistical Synthetic Estimates of Undercount for Small Areas", Proceedings of the Bureau of the Census' Second Annual Research Conference, March 23-26, 1986, pg. 557-569.

11. Isaki, C.T. and Schultz, L.K. (1987). "The Effects of Correlation and Matching Error in Dual System Estimation", Communications in Statistics - Theory and Methods, 16 (8), pg. 2405-2427.

12. Isaki, C.T., Schultz, L.K., Diffendal, G.J. and Huang, E.T. (1988). "On Estimating Census Undercount in Small Areas", paper submitted to Journal of Official Statistics, 39 pages.

13. Isaki, C.T. (1988). "Bias of the Dual System Estimator Under a Simple Model", submitted to Journal of Official Statistics, 28 pages.

14. Krotki, K.J. (1978). Developments in Dual System Estimation of Population Size and Growth. The University of Alberta Press: Edmonton.

15. Lindley, D.V. and Smith, A.F.M. (1972). "Bayes Estimates for the Linear Model", Journal of the Royal Statistical Society, Ser. B, 34, pg. 1-19.

16. Seber, G.A.F. (1982). The Estimation of Animal Abundance and Related Parameters. MacMillan: New York.

17. Seltzer, W. and Adlakha, A. (1974). "On the Effects of Errors in the Application of the Chandrasekaran-Deming Technique," Laboratories for Population Statistics, Reprint Series No. 14, University of North Carolina at Chapel Hill.

18. Tukey, J.W. (1981). Discussion of "Issues in Adjusting the 1980 Census Undercount", by Barbara Bailar and Nathan Keyfitz, paper presented at the Annual Meeting of the American Statistical Association, Detroit, MI.

19. Wilks, S.S. (1962). Mathematical Statistics. John Wiley and Sons: New York.

Appendix

Measures of Performance

All census defined geographic areas larger than EDs are collections of EDs. Hence, having defined two artificial populations at the ED level we have two sets of standards with which to compare the performance of the census (enumeration) and the coverage estimation methods. A number of measures of performance were developed and additional ones suggested by others (Citro and Cohen 1985). In defining the measures, c represents the enumeration (census), e represents an estimate of the population, s represents the artificial population used as the standard and N denotes the number of areas. The measures consider both estimates of level (total population) as well as proportion of the population.

Measures of Performance

1. Number of areas where $\mathrm{ARE}(c_i) < \mathrm{ARE}(e_i)$

   where

   $\mathrm{ARE}(e_i) = |(e_i - s_i)/s_i|$

   (ARE = absolute relative error)

2. Number of areas where $\mathrm{ADP}(c_i) < \mathrm{ADP}(e_i)$

   where

   $\mathrm{ADP}(e_i) = |P_i^e - P_i^s|$

   and

   $P_i^e = e_i / \sum_i^N e_i$ for the i-th area

   (ADP = absolute difference in proportions)

3. $\text{MARE} = \frac{1}{N} \sum_{i}^{N} \left|\frac{e_i - s_i}{s_i}\right|$ (MARE = mean ARE)

4. Maximum ARE(e)

5. Weighted squared relative error

$$\alpha(e) = \sum_{i}^{N} s_i \, [(e_i - s_i) / s_i]^2$$

6. Sum of absolute difference of proportions

$$\text{SADP}(e) = \sum_{i}^{N} |P_i^e - P_i^s|$$

7. Proportion of population improved

$$\text{PI} = \sum_{i}^{N} \text{IMPV}_i / M$$

$$M = \sum_{i}^{N} s_i, \quad \text{IMPV}_i = \begin{cases} s_i & \text{if ADP}(e_i) < \text{ADP}(c_i) \\ 0 & \text{otherwise} \end{cases}$$

8. Weighted squared relative error differences

$$\phi(e) = \sum_{i}^{N} s_i \, [\{(e_i - s_i)/s_i\} - \{(\sum_{i}^{N} e_i - \sum_{i}^{N} s_i)/\sum_{i}^{N} s_i\}]^2$$

9. Sum of weighted squared ADP

$$\text{IMP1}(e) = \sum_{i}^{N} [\text{ADP}(e_i)]^2 / P_i^s$$