

ESTIMATION OF MATCHING ERROR IN CENSUS UNDERCOUNT ESTIMATES

Mary C. Davis and Mary Mulry, U.S. Bureau of the Census
Paul Biemer, Research Triangle Institute
Mary Davis, U.S. Bureau of the Census, Washington, D.C. 20233

KEY WORDS: Capture-Recapture, Postcensal Estimates

1. INTRODUCTION

As part of the continuing research on census coverage error, the U.S. Bureau of the Census conducted a Postcensal Review (PCR) program of the 1990 Post Enumeration Survey (PES) to determine whether the undercount estimates from the 1990 Decennial Census could be used to adjust the postcensal estimates. A potential source of error in undercount estimates is a matching operation which determines whether the respondents in the population sample are enumerated in the census and whether the enumerations in the enumeration sample are correct. This paper discusses the 1992 Matching Error Study (MES), an evaluation of the matching of the data from the PCR.

The PES was really two samples: a) the enumeration sample or E sample consisting of census enumerations which measured erroneous enumerations, and b) the population sample or the P sample selected independently of the census which measured census omissions (Hogan, 1991). The same blocks were selected for both the E sample and the P sample, resulting in overlapping samples. The two samples were used in dual system estimation to produce an estimate of the census coverage error.

Matching is a critical part of the PCR. There are two basic types of errors which may occur as a result of matching -- random or systematic. Systematic errors are a particular concern since these errors may be associated with particular geographic areas and/or demographic population groups. The effect of matching error is that false nonmatches result in an overstatement of the actual coverage error. False matches, on the other hand, tend to understate the level of coverage error.

Dual system estimation assumes that the P-sample respondents can be linked, or matched, correctly to their census enumerations. Also, there is the assumption that census enumerations in the

E sample can be properly identified as correct or erroneous.

The Census Bureau's dual system estimator for an area or domain is given by

$$\hat{N} = \frac{CN_P}{M}$$

where $C = CEN - II - UM - EE$, CEN is the size of the original enumeration for the area or domain, II is the number of imputed persons, UM is the estimate of the unmatchable census enumerations, and EE is the estimate of the number of erroneous enumerations in the original enumeration. N_P is the estimate of the total population from the P sample universe, and M is the estimate of the number of "matchable" persons in both the census and the P sample. In what follows, we refer to the combination of EE and UM (EE+UM) as EE+. The goal of the MES is to evaluate the matching error in EE+ and M.

Section 2 discusses the methodology for estimating the error in M and EE+. Section 3 describes both the PCR and MES designs, including sampling. Section 4 contain the results and section 5 gives a summary and conclusions.

2. METHODOLOGY

The major focus of the MES is the estimation of the bias in EE+ and M. That is, suppose we were able to develop a "perfect" matching system, i.e., one in which no errors were made in classifying a case as an "EE+" or "not an EE+" and as an "M" or "not an M." Let $EE+_T$ denote the value of EE+ and let M_T denote the value of M from the perfect system. Let $EE+_p$ and M_p denote the values of EE+ and M from the production system. For this analysis, the production system is the PCR. The bias in the production system for $EE+_p$ is $B(EE+) = (EE+_p - EE+_T)$ and for M_p is $B(M) = E(M_p - M_T)$ where $E(\bullet)$ denotes the expectation taken over all possible samples and all appropriate nonsampling error distributions.

Estimates of matching error biases are only as good as the so-called perfect matching system that

produces the quantities M_T and $EE+_T$, i.e. the rematching system. In reality, no rematching system, no matter how well designed, can be expected to produce $EE+$ and M having no misclassification error biases. However, if the errors in EE_T and M_T are small relative to the size of $B(EE+)$ and $B(M)$, respectively, then estimates of $B(EE+)$ and $B(M)$ based upon the rematching system should still be useful for evaluating the production matching components. This was the objective in designing the MES rematch operation.

The bias in the DSE is directly affected by the bias in $EE+$ and M . To see this, let RDR denote the "relative difference rate" defined by $RDR = (\text{production} - \text{rematch})/\text{rematch}$. Thus,

$$RDR(M) = \frac{M_P - M_T}{M_T} \quad (1)$$

and

$$RDR(EE+) = \frac{EE+_P - EE+_T}{EE+_T} \quad (2)$$

Since the numerator of the RDR is an estimator of the bias $B(M)$ (for (1)) and $B(EE+)$ (for (2)), the RDR is a measure of the relative bias. It can be shown that the relative bias in \hat{N} , denoted by $RB(\hat{N})$, is given by

$$RB(\hat{N}) \approx E \left[\frac{RDR(C) - RDR(M)}{1 + RDR(M)} \right]$$

where

$$RDR(C) = \frac{C_P - C_T}{C_T}$$

for $C_P = \text{CEN-II-EE}+_P$ and $C_T = \text{CEN-II-EE}+_T$.

It can be seen from the expression that a positive bias in $EE+_P$ (i.e., $RDR(C) < 0$) will cause a downward bias in \hat{N} (i.e. $RB(\hat{N}) < 0$), ignoring the effect of $RDR(M)$. Likewise, a positive bias in M_P (i.e. $RDR(M) > 0$) also will cause a downward bias in \hat{N} , ignoring the effect of the bias in $EE+_P$ on $RDR(C)$. If $RDR(C)$ and $RDR(M)$ have like signs, then their affects are somewhat offsetting and $RB(\hat{N}) = 0$ when $RDR(C) = RDR(M)$.

Ultimately, every case in the E sample is classified as a correct enumeration, CE, an erroneous enumeration, EE, or as unresolved, UR. Every case in the P sample is classified as a match, M, a nonmatch, NM, unresolved, or out of scope, OS. A match probability (for the P sample) and a probability of correct enumeration (for the

E sample) are imputed for the unresolved cases in the PES imputation process. However, the analysis described in this report focuses on the data which emerged from the PCR matching operation before unresolved cases have been resolved through the imputation process.

3. STUDY DESIGN

3.1 Postcensal Review Design

After the interviewing phase of the 1990 PES, a matching operation determined whether the P-sample respondents were enumerated in the census and whether the E-sample cases were correctly enumerated. When necessary, cases were sent to followup to collect additional information to resolve the match status (Davis, et al, 1991).

Following the July 15, 1991 decision not to adjust the 1990 Census, revisions were made to the 1990 PES estimates. The Selective Cluster Review (SCR) rematch was conducted in the 104 clusters with the largest leverage index, L, in an effort to reduce the coverage error as a result of matching error. By computing a leverage index it was determined that 104 clusters, representing less than 2 percent of the PES sample, accounted for 25 percent of the coverage error measured by the dual system estimates. The leverage index is given by $L = \text{abs}(NM - EE)$ where NM is the weighted omissions in the census as determined by the P sample and EE is the weighted erroneous enumerations.

The rematch was performed by a group of Matching Review Specialists (MRS), the most highly trained matcher, and PES technicians. Additional training was conducted for the SCR due to several matching rules that were expanded to handle unique cases that went beyond the existing production matching rules. Although both MRS and techs worked on the rematch, the MRS had responsibility for the final codes.

After the SCR matching was completed, the 1990 production match codes were replaced by the SCR codes and the PCR estimates were calculated. Modifications to the computer editing of the data after clerical matching are included in the PCR results as well as a new poststratification scheme.

3.2 Matching Error Study Design

The 1992 MES differs from the 1990 MES (Davis, et al, 1991) by the inclusion of the PCR results and the Hispanic Cluster Review (HCR), a supplemental sample of 100 clusters. A leverage value was calculated for only the Hispanics in the cluster, although all persons within a selected cluster were rematched regardless of race or Hispanic origin. The HCR was conducted in an effort to reduce the large standard errors on the estimates of matching error for Hispanics from the 1990 MES.

The HCR rematching operation was conducted at the conclusion of the SCR using the same matching personnel and identical guidelines. As in the SCR, the MRS had responsibility for the final codes. Therefore, all final MES match codes were reviewed by an MRS.

The 1990 MES was a dependent rematch of the 919 block cluster evaluation sample conducted following the termination of all PES operations. "Dependent" means that the matchers had access to match codes assigned at every stage of production matching. However, procedures were implemented to insure that the assignment of MES match codes was not influenced by the production matching operation. For a detailed description of the MES matching operation, see Davis, et al (1991).

Of the 104 clusters rematched for the SCR, 32 overlapped with the original 1990 MES sample. For these clusters, the original MES match codes were replaced by the SCR match codes since the SCR matching is considered the best matching for those clusters. Therefore, no matching error is being measured for these 32 clusters since both PCR and MES codes are identical. Note that other clusters in the evaluation sample have no matching error, although not by design, so these 32 overlapping clusters are not unique.

3.3 Sampling

The 1990 PES evaluation sample was a stratified systematic subsample of the 5300 PES sample block clusters. The PES sampling strata were first grouped into thirteen evaluation sampling groups. Within each sampling group, block clusters were furthered grouped and sorted using criteria aimed at minimizing the variances of the estimated error rates. An unequal probability sample of 919 block clusters were drawn from the 13 sampling groups. For the allocation of the sample clusters to the evaluation groups, an optimal allocation strategy was followed. For a more detailed description of the 1990 evaluation sampling operation see Davis and Biemer (1991).

The HCR supplemental sample was a certainty sample, as was the SCR sample. A leverage index was computed for all PES clusters not already in the evaluation sample or in the SCR sample. The leverage index was calculated for only the Hispanics in the cluster. The clusters were sorted by descending leverage value and the first 100 clusters were selected for the rematch. The resulting evaluation sample used in this analysis is a 1019 cluster sample.

In order for the results from the HCR and SCR samples to be generalizable to the total population, two additional strata were formed. Each cluster in the HCR stratum was given an evaluation subsampling weight of 1. Another stratum was formed for the SCR sample. Since 32 of the 104 clusters were selected in the original evaluation sample, a subsampling weight of 3.25 was used for those clusters. All other subsampling weights for the remaining 887 clusters in the original evaluation sample were also adjusted to account for the forming of the two additional strata.

4. DATA ANALYSIS

4.1 Objectives

The primary objective of the MES is to evaluate the quality of the E-sample and P-sample PCR matching operations and its impact on the total population estimates. Determinations as to whether or not matching error was controlled in the PCR are based primarily upon results at the evaluation poststratum (EPS) levels given in Figure 1. However, matching error is also examined for demographic subgroups.

Figure 1. The 10 Evaluation Poststrata

- 1 Non-Hispanic (NH) White and Other, Owner in Urban Areas (UA) 250k+
- 2 NH White and Other, Owner in Other UAs
- 3 NH White and Other, Owner in Non-UAs
- 4 NH White and Other, Non-Owner in UAs 250k+
- 5 NH White and Other, Non-Owner in Other UAs
- 6 NH White and Other, Non-Owner in Non-UAs
- 7 Black, Non-Black Hispanic, API, Owner in UAs 250k+
- 8 Black, Non-Black Hisp., API, Non-Owner in UAs 250k+
- 9 Black, Non-Black Hisp., API, Owner in Other Urban & Non-UAs
- 10 Black, Non-Black Hisp., API, Non-Owner in Other Urban & Non-UAs

4.2 E-Sample Analysis

The estimates for the RDRs for EE+ defined by Equation (2) in Section 2 were computed for the 10 EPS, the U.S., 7 race and Hispanic origin categories given in Figure 2, and tenure breakdowns. A large absolute value of the RDR indicates more inaccuracy in the matching operation. A negative RDR implies that the DSE of the total population size is biased upward.

Table 1 gives the RDR(EE+) for all EPS and the respective 95% confidence interval. Only EPS 3 (NH White and Other, Owner in Other UAs) with an RDR(EE+) of -0.102 is significant indicating an underestimation of the PCR EE+ rate. At the U.S. level, the EE+ rate is also significantly underestimated by the PCR.

When the estimates are calculated by race/Hispanic origin categories, we see the RDR(EE+) of -0.048 for Non-Hispanic Whites and Others is significant indicating that the PCR EE+ rate was underestimated as much as 9 percent and no less than 1 percent with 95% confidence. For Non-Black Hispanics, the PCR EE+ rate was significantly overestimated (RDR(EE+)=0.053) at the 5% level. However, the lower confidence limit of 0.1 percent is very close to zero.

When the error estimates are calculated for minority and nonminority subgroups, the nonminority estimate of -0.050 is significant at the 5% level which is consistent with the race/Hispanic origin findings since Non-Hispanic Whites and Others make up the majority of the weight for the nonminority subgroup. When we examine RDR(EE+) by tenure groups (owner and non-owner), the number of owners classified as EE+ in the PCR is significantly underestimated (RDR(EE+)=-0.053).

Figure 2. Race/Hispanic Origin Categories

- 1 White, Non-Hispanic, Others
- 2 Black
- 3 Non-Black Hispanic
- 4 American Indian, Non-Hispanic
- 5 American Indian, Hispanic
- 6 Asian, Non-Hispanic
- 7 Asian, Hispanic

4.3 P-Sample Analysis

The estimates for the RDRs for NM were computed for the same subgroups as for the E-sample analysis. The computation of RDR(NM) is similar to RDR(M) defined by Equation (2) in Section 2 using the weighted number of nonmatches instead of weighted matches. The RDR(NM)

formula is used so that the P-sample matching error estimate would be of the same magnitude as the E-sample error. A large absolute value of the RDR(NM) indicates more inaccuracy in the matching operation. A positive RDR for NM implies that the DSE of the total population size is biased upward.

As seen in Table 1, EPS 7 and the U.S. have significant RDR(NM)s which indicate the PCR overestimated the number of nonmatches. However, the lower confidence limits of the 95% confidence intervals are very close to zero.

The only significant RDR for the 7 race/hispanic origin categories is for the Black population: RDR(NM) = 4.4 percent. Note that the lower confidence limit is 0.2 percent. After regrouping the data by minority and nonminority groups, neither group was found to have a significant RDR. The same is true for the estimates by tenure.

4.4 Effect of Matching Error on Undercount

As described in section 2, the Census Bureau's dual system estimate is given by $\hat{N} = (CN_p)/M$. The DSE is used to estimate the net undercount rate, \hat{U} , in the original enumeration where

$$\hat{U} = 100 * \frac{(\hat{N} - CEN)}{\hat{N}}$$

Several components of PES error are measured by the Census Bureau's total error model which assesses the overall accuracy of the PES and the census estimates of population size (Mulry and Spencer, 1992). Two net component errors involve P-sample matching error. The net error m_m is the error in assigning P-sample cases the status of enumerated or not enumerated, the same error measured by RDR(NM). The error affects the denominator of the DSE, M, the weighted number of matches. The error n_{pm} which occurs during the P-sample matching operation when cases may be declared as out of scope affects N_p , the size of the P-sample population. Cases declared as out of scope by the PCR may be declared as in scope by the MES and vice versa. This type of error was not included in the discussion of P-sample matching error earlier in this paper. However, the error from m_m and n_{pm} are combined for this discussion of P-sample matching error. A third net component error in the total error model measures E-sample office processing error, the same error measured by RDR(EE+). The error c_o arises during the processing of the E sample when respondents are misclassified as to whether they are correctly or

erroneously enumerated in the original enumeration. This error affects C, the number of distinct people enumerated in the census.

The estimate of the individual effect of matching error on the DSE, the mean, variance and covariance of m_m , n_{pm} , and c_o were estimated by EPS using the total error methodology. Simulations were run using three different combinations of the error components, assuming no other errors were present in the model: 1) P-sample matching error only: m_m and n_{pm} , 2) E-sample matching error only: c_o , and 3) P- and E-sample matching error combined: m_m , n_{pm} , and c_o . The simulations were used to calculate an error-free DSE and net undercount rate. The bias in the net undercount rate $B(\hat{U})$ was estimated by the difference between \hat{U} and the mean of the simulated distribution of the undercount rate. Table 2 gives the net undercount rate for each evaluation poststratum, its bias, $B(\hat{U})$, and the standard error for the three simulations described above. The asterisk (*) denotes significance at the 5 percent level.

The bias in the DSE of the undercount due to E-sample matching error is positive, indicating a reduction in the net undercount, in all EPS except EPS 6 and 10, both of which are minority poststrata. However, neither of these estimates are significant at the 5% level. For the P-sample components, only EPS 9 has a negative bias. When the E- and P-sample errors are combined, EPS 9 is the only EPS to have a negative bias, indicating that the undercount rate would increase based on matching error. Note, however, that the estimate is only slightly significant at the 5% level.

The results for the E-sample given in Table 2 are consistent with those given in section 4.2. The RDR(EЕ+) for nonminorities is significant as is the bias in the undercount for 3 of the 6 nonminority poststrata (EPS 1-6). EPS 3 has a significant bias and RDR(EЕ+). Of the 5 owner poststrata (EPS 1,2,3,7,and 9), 3 have a significant positive bias which is consistent with the results in Table 1. At the national level, both measures of E-sample matching error are also significant.

For the P-sample matching error results in Table 2, all of the minority poststrata (EPS 7-10) have a significant bias at the 5% level. All renter poststrata (EPS 4,5,6,8and 10) are also significant. However, these results are different from those in section 4.3 using RDR(NM). As noted above, the results presented here for the P-sample include measuring the movement from out of scope to in scope and vice versa which is not included in the

estimates of RDR(NM). This can explain the difference in the results for the P-sample.

When the P- and E-sample matching error estimates are combined, the P-sample error appears to contribute more than the E-sample error since $B(\hat{U})$ is significant in all EPS that were significant for the P-sample separately. In addition, the bias for EPS 2 and the U.S. is significant.

5. SUMMARY AND CONCLUSIONS

5.1 E Sample

When the MES data are grouped by evaluation poststrata, only EPS 3, Non-Hispanic White and Other, Owner in Non-Urban Areas was found to have a significant RDR(EЕ+). The U.S. estimate was also significant at the 5% level. Estimates of RDR(EЕ+) by race/Hispanic Origin were significant for Non-Hispanic Whites and Others and Non-Black Hispanics. There is also evidence that EЕ+'s were underestimated for nonminorities and owners. The results from the total error analysis examining the bias in the undercount due to E-sample matching error, assuming no other contributing errors, support the RDR(EЕ+) findings.

5.2 P Sample

The following RDR(NM)s were only slightly significant: EPS 7 with an RDR(NM) of 4.3%, the U.S. with and RDR of 3.0% and Blacks (national level) with an RDR of 4.4%. No other effects examined were significant at the 5% level including minorities, nonminorities, owners and non-owners. The results from the total error analysis examining the bias in the undercount due to P-sample matching error, assuming no other contributing errors, produced more significant results than the RDR(NM) findings. The difference in the results may be attributable to the measurement in the total error model of n_{pm} , the error in determining whether a P-sample person is in-scope or out-of-scope.

* This paper reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

ACKNOWLEDGEMENTS

The authors wish to thank Phil Gbur for his work on variance estimation.

Table 1. 95% Confidence Intervals for RDR(EE+) and RDR(NM) by Evaluation Poststratum

Evaluation Poststratum	E Sample Results			P Sample Results		
	RDR(EE+)	Lower CL	Upper CL	RDR(NM)	Lower CL	Upper CL
1	-0.027	-0.113	0.059	0.027	-0.028	0.082
2	-0.037	-0.142	0.068	-0.027	-0.157	0.104
3	-0.102	-0.195	-0.009	0.025	-0.019	0.069
4	0.013	-0.023	0.048	0.044	-0.017	0.105
5	-0.126	-0.253	0.002	0.069	-0.081	0.220
6	-0.012	-0.114	0.090	0.047	-0.026	0.119
7	-0.051	-0.117	0.014	0.043	0.002	0.085
8	0.011	-0.035	0.057	0.009	-0.020	0.037
9	-0.002	-0.083	0.086	-0.006	-0.036	0.024
10	-0.009	-0.102	0.084	0.096	-0.050	0.242
U.S.	-0.037	-0.067	-0.008	0.030	0.003	0.057

Table 2. Effect of Matching Error on the Net Undercount Rate by Evaluation Poststratum

Evaluation Poststratum	\hat{U}	Bias for c_o	Standard Error	Bias for m_m & n_{pm}	Standard Error	Bias for m_m, n_{pm}, c_o	Standard Error
1	-0.50	0.08	0.06	0.07	0.06	0.15	0.08
2	0.11	0.18*	0.07	0.01	0.07	0.19	0.10
3	-0.22	0.60*	0.12	0.09	0.05	0.69*	0.13
4	2.33	0.01	0.05	0.29*	0.08	0.30*	0.09
5	2.92	0.82*	0.19	0.98*	0.29	1.81*	0.43
6	5.30	-0.11	0.15	0.52*	0.14	0.40*	0.20
7	1.33	0.34*	0.09	0.24*	0.06	0.58*	0.11
8	7.13	0.03	0.09	0.66*	0.12	0.69*	0.16
9	2.07	0.10	0.10	-0.37*	0.09	-0.27*	0.13
10	6.44	-0.01	0.14	0.66*	0.30	0.65*	0.32
U.S.	1.61	0.22*	0.03	0.24*	0.03	0.47*	0.05

REFERENCES

Davis, M., Biemer, P., Mulry, M., and Parmer, R. (1991) "The Matching Error Study for the 1990 Post Enumeration Survey", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 248-253.

Davis, M. C. and Biemer, P. (1991) Post Enumeration Survey Evaluation Project P7: Estimates of P-sample Clerical Matching Error from a Rematching Evaluation. Coverage Studies and Evaluation Memorandum Series #H-2, Statistical Support Division, Washington, D.C.

Hogan, H.R. (1991) "1990 Post Enumeration Survey Operations and Results" *Proceedings of the Section on Social Statistics*, American Statistical Association.

Mulry, M.H. and Spencer, B. (1992) "Accuracy of Undercount Estimates for the 1990 Census", *Proceedings of the 1992 Annual Research Conference*, Bureau of the Census, Washington, D.C. in press.