Survey Methodology, December 1994
Vol. 20, No. 2, pp. 149–158
Statistics Canada

Dual System Estimation of Census Undercount in the Presence of Matching Error

YE DING and STEPHEN E. FIENBERG[1]

ABSTRACT

Dual system estimation (DSE) has been used since 1950 by the U.S. Bureau of Census for coverage evaluation of the decennial census. In the DSE approach, data from a sample is combined with data from the census to estimate census undercount and overcount. DSE relies upon the assumption that individuals in both the census and the sample can be matched perfectly. The unavoidable mismatches and erroneous nonmatches reduce the accuracy of the DSE. This paper reconsiders the DSE approach by relaxing the perfect matching assumption and proposes models to describe two types of matching errors, false matches of nonmatching cases and false nonmatches of matching cases. Methods for estimating population total and census undercount are presented and illustrated using data from 1986 Los Angeles test census and 1990 Decennial Census.

KEY WORDS: Capture-recapture; Matching bias; Modelling matching error; Multinomial likelihood.

1. INTRODUCTION

The problem of undercount in the U.S. census has been of special concern since the first census of 1790 (Jefferson 1986). The DSE (or capture-recapture) approach has been used in conjunction with the census to evaluate population coverage as part of what is called the post-enumeration survey (PES) program. Ericksen and Kadane (1985) and Wolter (1986) describe the use of the DSE approach in the context of the 1980 decennial census. A new design for the PES was planned for the 1990 decennial census and refinements in methodology were examined in connection with a 1986 test census in central Los Angeles County, referred to as the Test of Adjustment Related Operations (TARO). Diffendal (1988) discusses methodology, operations, and the results of TARO, and Hogan and Wolter (1988) and Schenker (1988) provide evaluation of the operations and assumptions underlying the DSE approach.

The PES approach to dual-system estimation uses two samples, called the P-sample and the E-sample. The P-sample which is drawn separately from the census, helps to measure census omissions; the E-sample drawn from the census enumerations, helps to measure census erroneous enumerations. For the 1986 TARO, the dual-system estimator for the population size, N, which combines the information from the P-sample and the E-sample takes the form:

$$\hat{N} = (\text{CEN} - \text{EE} - \text{SUB}) \cdot N_p/M,$$

where CEN is the unadjusted census count; EE is the estimated number of erroneous enumerations and unmatchable persons included in the census; SUB is the number of whole-person substitutions in the census; N_p is the number of people in the P-sample; M is the estimate of the number of people in both census and the P-sample. For details see Diffendal (1988) or Wolter (1986). For the variation on this formula as used in conjunction with the 1990 census, see Hogan (1992, 1993).

DSE and the matching problem gained considerable attention in the 1970's due to its use in estimating births and deaths in developing countries, and it is thought by some that perhaps the greatest problem with the dual-system estimation approach used in 1980 census was the rate of matching error (Fienberg 1989). Jaro (1989) describes the technological innovations for matching introduced by the Bureau of the Census for 1990 and the test of the related matching methodology in a 1985 pretest. Biemer (1988) considers models for evaluating the impact of matching error on estimates of census coverage error without attempting to correct for the matching bias in the usual dual-system estimate. The actual procedure used in the 1990 census included not only a computer matching algorithm and various clerical follow-ups but also logistic regression models for unresolved cases in both the P-sample and E-sample (see Belin *et al.* 1993).

Matching is used to determine the census enumeration status of the people enumerated in the P-sample. Specifically, those people in the P-sample who are matched to the census are considered to have been enumerated. People in the P-sample who do not match are, for the most part, considered to have been missed by the census. Matching errors can occur for two general reasons:

[1] Ye Ding is Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg is Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

1. The information reported by the respondents/interviewers was incorrect.
2. Correct information was reported, but it was not correctly used.

Moreover, two types of errors can occur: false matches of nonmatching cases and false nonmatches of matching cases. False matches of nonmatching cases may be divided into

(a) instances in which a P-sample case was erroneously matched to the enumeration of another person, but a match to that actual E-sample case should have been made, and
(b) instances in which no match should have been made.

The former case is not "serious" for the purposes of estimating N, since such false matches would have been, in fact, correctly classified as a match to the census. In the second case, however, the number of nonmatches becomes understated. False nonmatches to the census, on the other hand, have the effect of overestimating the nonmatch rate. Fay, Passel, Robinson and Cowan (1988) note that false nonmatches probably represent a greater concern than false matches. False matches are less common than false nonmatches because matches can be reviewed easily.

In Section 2, we propose models for matching errors and then, in Section 3 and 4, we present a systematic procedure for the estimation of the population total and thus the census undercount. In Section 5, we analyze the data from 1986 Los Angeles test census and 1990 Decennial Census to show how our method accounts for matching errors in the undercount estimates.

2. MODELING MATCHING ERRORS

For simplicity, we assume that the matching mechanism is constrained, in the sense that no individual in one sample can be matched with more than one individual in another sample. Moreover, we implicitly assume a version of simple random sampling, within strata, and this yields a standard multinomial sampling model for dual system estimation. This simplification allows us to focus on the impact of matching and its mechanisms. In what follows, we provide a way to view the recapture data, for the purpose of setting up models for matching.

Let $Z_{N\times 1}$ be the characteristic vector for the whole population, such that the i-th component of $Z_{N\times 1}$ contains the characteristics for the i-th individual, where $1 \le i \le N$. Not all the components in $Z_{N\times 1}$ can be observed in any one sample. The object is to estimate N, the size of the population, from information from two samples. One could view drawing a sample from the population as drawing some components in $Z_{N\times 1}$ at random to form a new vector Y. Then, missing or misreporting of certain characteristics in those components drawn may cause matching errors. Henceforth we will refer to the first sample as Y_1 and the second sample as Y_2, and in the following discussion they will be the two capture-recapture samples for dual system estimation.

Two types of matching errors can occur: false nonmatches of matching cases, and false matches of nonmatching cases. We will refer to the former as a type 1 error and the latter as a type 2 error. We can focus on modeling one or both types of error. Under perfect matching, each component in Y_1 or Y_2 contains the same information as in $Z_{N\times 1}$, and the number of matches will be the number of elements common to Y_1 and Y_2. When faced with uncertain matching, we consider the following simple model:

Model (A):

(i) Assume that those matched pairs of components under perfect matching will still be matched, each with common probability α, $0 < \alpha \le 1$.
(ii) All those unmatched will remain unmatched, *i.e.*, no false matches.

Model (A) characterizes a mechanism for type 1 matching error with error probability $1 - \alpha$, assuming that type 2 matching error is negligible.

To develop a model for both types of matching error, we need to consider carefully all the possibilities that lead to false matches. When there is no matching error, one can write $Y_1 = (M_1, N_1)$ and $Y_2 = (M_2, N_2)$, so that sets M_1 and M_2 have the same size and every individual in M_1 is correctly matched with one individual in M_2 and vice versa, N_1 is the set of those in sample Y_1 who are not matched with any one in sample Y_2, and N_2 is the set of those in sample Y_2 who are not matched with any one in sample Y_1. When matching errors are present, false matches can occur in the following ways:

(a) A person in M_1 is matched incorrectly with a person in M_2.
(b) A false match occurs between M_1 and N_2.
(c) A false match occurs between M_2 and N_1.
(d) A false match occurs between N_1 and N_2.

We note that each of (a), (b), (c) happens only when at least 2 errors are made, that is, the correct match is not made and an incorrect match is made. Since such errors occur with small probability, we assume for simplicity that cases (a), (b), (c) have negligible probability of occurrence in the next model.

Model (B):

(i) Assume, as in model (A), that matching pairs between M_1 and M_2 will still be matched, but with probability α, $0 < \alpha \le 1$.
(ii) Assume that false matches of types (a), (b), (c) are negligible.
(iii) Assume that each person in N_1 will be matched with someone in N_2 with a common probability β, $0 \le \beta < 1$.

Even though, in theory, both α and β can vary from 0 to 1, in the census context we expect that $\alpha \approx 1$, and $\beta \approx 0$.

We can also consider instances in which the matching error probabilities and capture probabilities potentially vary over identifiable population subgroups. In other words, the population can be divided into strata, by demographic (*e.g.*, age, race, sex) and geographic variables, within which the matching error probabilities and capture probabilities could be assumed to be more homogeneous than in the whole population. Suppose the whole population consists of l strata. Let $Z^i_{N_i \times 1}$ be the characteristic vector for the population of the i-th stratum with unknown size N_i, and let Y_{i1}, Y_{i2} be two samples taken from the i-th stratum which are used to get an estimate $\hat{N}_i$. Then we can form an estimate of the overall population size by setting $\hat{N} = \sum_{i=1}^{l} \hat{N}_i$. We can refine models (A) and (B) as follows:

Model (A′):

Assume model (A) holds within each stratum, and let α_i be the probability of a match for matching components in stratum i, $0 < \alpha_i \leq 1$, $1 \leq i \leq l$.

Model (B′):

Assume model (B) holds within each stratum, and let the two probability parameters for i-th stratum be α_i, β_i, $1 \leq i \leq l$.

For 1990 PES, the P-sample matching was conducted using the sample blocks plus a ring of surrounding blocks (Hogan 1993). Geocoding errors may lead to false matches across geographically defined post-strata, and false matches are possible for demographically defined post-strata. Models (B′) implicitly assumes that there are no false matches across post-strata. Further, all of the models represent a simplification of the underlying sample design of the PES.

3. ESTIMATE THE POPULATION TOTAL

In this section, we consider estimation of the population total under the various matching models, (A), (A′), (B), and (B′), assuming the validity of usual assumptions of independence of the two samples and homogeneous probabilities of inclusion in the samples. For models involving heterogeneous catchability and/or dependence, see the three-sample approach in Darroch *et al.* (1993) and the approach in Alho *et al.* (1993).

Let N be the number of individuals in the population under consideration, x_{1+} the number of individuals in Y_1, x_{+1} the number of individuals in Y_2, and x_{11} the number of individuals in both samples. The number of individuals observed in Y_2 but not Y_1 is $x_{21} = x_{+1} - x_{11}$ and the number observed in Y_1 but not Y_2 is $x_{12} = x_{1+} - x_{11}$.

One can arrange the capture-recapture data in a 2×2 contingency table with one missing cell:

		Sample Y_2 present	Sample Y_2 absent
Sample Y_1	present	x_{11}	x_{12}
	absent	x_{21}	$-$

where we use symbol "$-$" to indicate the missing cell, and standard notation for marginal totals: $x_{1+} = x_{11} + x_{12}$, $x_{+1} = x_{11} + x_{21}$. There is a corresponding 2×2 table of probabilities, $p_{ij} = \Pr[\text{any individual falls into } (i,j) \text{ cell}]$,

		Sample Y_2 present	Sample Y_2 absent
Sample Y_1	present	p_{11}	p_{12}
	absent	p_{21}	p_{22}

with the usual linear constraint

$$\sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1.$$

Let n be the number of observed different individuals in the two samples, *i.e.*, $n = x_{11} + x_{12} + x_{21}$. If we assume that the samples are randomly selected with homogeneous selection probabilities, then the numbers of individuals in the four cells have a multinomial distribution

$$(x_{11}, x_{12}, x_{21}, N - n) \sim \text{Mult}(N, p_{11}, p_{12}, p_{21}, p_{22}).$$

We use the conditional likelihood approach developed by Sanathanan (1972). For fixed n, (x_{11}, x_{12}, x_{21}) has a multinomial distribution with likelihood function

$$L_1(p_{11}, p_{12}, p_{21}) = \frac{n!}{x_{11}!x_{12}!x_{21}!} \cdot \frac{p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}}}{(p_{11} + p_{12} + p_{21})^n}. \quad (1)$$

Then n is viewed as being binomially distributed with sample size N and probability $p_{11} + p_{12} + p_{21}$, and the corresponding likelihood is

$$L_2(N) = \frac{N!}{n!(N-n)!} (p_{11} + p_{12} + p_{21})^n [1 - (p_{11} + p_{12} + p_{21})]^{N-n}. \quad (2)$$

In the conditional approach we derive maximum likelihood estimates for the cell probabilities based on the likelihood (1), then find the value of N which maximizes (2), given

the values of the cell probabilities. Sanathanan (1972) has shown that under suitable regularity conditions both conditional and unconditional likelihood estimates of N are consistent and have the same asymptotic multivariate normal distribution. The conditional approach is particularly suitable for a large sample problem like ours.

Under the equal catchability assumption, we let p_1 be the probability that any individual in the population is included in Y_1, and similarly we let p_2 be the probability of inclusion in Y_2. The probabilities p_1 and p_2 are usually referred to as capture probabilities and they do not depend on how the matching mechanism operates. Then the probability that an individual is in both samples is p_1p_2, and the probability of being in set N_1 is $p_1(1 - p_2)$. Since model (A) is a special case of model (B) with $\beta = 0$, we focus on formulating the problem under model (B). To do this, we first need to work out the parametric specification of the cell probabilities. An individual will fall into the (1,1) cell in the 2×2 table only in two cases, *i.e.*, the individual is actually in both samples and a match is made, or, using the notation in the last section, an individual who is actually in N_1 is incorrectly matched with some one in N_2. Here the matching direction from N_1 to N_2 is implicitly assumed in (iii) of model (B). The probability that the former case occurs is αp_1p_2, and the probability that the latter case occurs is $\beta p_1(1 - p_2)$. Furthermore, the two cases are mutually exclusive. Thus, we have $p_{11} = \alpha p_1p_2 + \beta p_1(1 - p_2)$, and, $p_{12} = p_1 - p_{11} = p_1 - \alpha p_1p_2 - \beta p_1(1 - p_2)$, $p_{21} = p_2 - p_{11} = p_2 - \alpha p_1p_2 - \beta p_1(1 - p_2)$. Rao (1957) studied regularity conditions under which there exist unique maximum likelihood estimates of parameters in a multinomial distribution. His conditions are satisfied by the parameterization of $\{p_{ij}\}$ here.

For $\alpha = 1, \beta = 0$, this setup reduces to the usual two sample problem and there exist well known solutions in closed form for resulting likelihood equations for the conditional likelihood (1) (*cf.* Bishop *et al.* 1975, chap. 6, p. 232), leading to the usual dual-system estimator, $\hat{N}_{\text{DSE}} = x_{1+}x_{+1}/x_{11}$. Otherwise, the maximum likelihood estimates cannot be written in closed form. Once we have $\hat{p}_1$ and $\hat{p}_2$, however, the conditional maximum likelihood estimates for p_1 and p_2, the conditional maximum likelihood estimate for N can be written as

$$\hat{N} = \frac{n}{\hat{p}_1 + \hat{p}_2 - (\alpha - \beta)\hat{p}_1\hat{p}_2 - \beta\hat{p}_1}, \tag{3}$$

(*cf.* Chapman 1951). Under model (A′) or (B′), for the i-th stratum, one can use the estimates of the parameters computed under model (A) or (B) for the data of that stratum, and then sum over strata for an estimate of the population total.

4. ESTIMATE MATCHING ERROR RATES BY REMATCH STUDY DATA

In what follows, we give estimates of the matching error rate parameters α and β using the data from the Matching Error Study (rematch study), one of the operations conducted by the Census Bureau in the 1986 Los Angeles test census to evaluate the PES. Briefly, the rematch typically operates for a sample of cases, using more extensive procedures, highly qualified personnel and reinterviews to obtain estimates of the bias associated with the previous matching process. For further details, see Childers, Diffendal, Hogan and Mulry (1989). In their discussion of the Matching Error Study in Los Angeles TARO, Hogan and Wolter (1988) state that "The rematch was done independently of the original match, and the discrepancies between the match and the rematch results are adjudicated. Because of this intensive approach to the rematch, we believe the rematch results represent true match status, while differences between the match and rematch results represent the bias in the original match results."

The data collected in a rematch study can be displayed as in the following table

Rematch Study Data

		Rematch Classification: Matched	Rematch Classification: Not Matched
Original Classification	Matched	y_{11}	y_{12}
	Not Matched	y_{21}	y_{22}

To estimate α and β, we assume that in the original matching process, errors are made according to model (B) and that errors in the rematch process can be disregarded, *i.e.*, the rematch is assumed to be perfect. It then follows that $y_{11} + y_{21}$ is the true number of matches, and thus is fixed, while y_{11} is a random variable having a binomial distribution, *i.e.*, $y_{11} \sim \mathcal{B}(y_{11} + y_{21}, \alpha)$. Thus the maximum likelihood estimate of α is $\hat{\alpha} = y_{11}/(y_{11} + y_{21})$, and the maximum likelihood estimate of the false nonmatch rate γ is $\hat{\gamma} = 1 - \hat{\alpha} = y_{21}/(y_{11} + y_{21})$. By the same argument, $y_{12} \sim \mathcal{B}(y_{12} + y_{22}, \beta)$, and the maximum likelihood estimate of the false match rate is $\hat{\beta} = y_{12}/(y_{12} + y_{22})$.

We can use the estimates of the matching error rates derived here to analyze the data from the rematch study from the Los Angeles test census. Very often, in addition to estimating the size of a population, it is of interest to estimate the size of a subpopulation such as black, white, or a subpopulation at a certain geographical location. In such case, it is more appropriate to allow for heterogeneity

of matching error rates across various population strata by using estimates of matching error rates for each stratum of interest. Such estimates can be obtained by conducting a rematch study within each stratum and then using the derived estimates. Data for applying model (B′) are available from 1990 Census and are analyzed here.

5. APPLICATIONS

5.1 Application of One Stratum Model to 1986 TARO

Hogan and Wolter (1988) present the rematch data from the 1986 Los Angeles TARO. The rematch results for the P-sample are given in Table 1 in the form of a cross-tabulation of match statuses as assigned from the original TARO match and the rematch. Table 2 presents the two way table of data for the 1986 TARO, with no post-stratification. The estimate of the number missed by both systems, 5,870 is approximately the same order of magnitude as census substitutions 5,259 and erroneous enumerations 6,426 (Hogan and Wolter 1988). Rematch results for the E-sample are presented in Table 3. Let CP, EP be the total correct enumeration and erroneous enumeration by production classification, and let CR, ER be the total correct enumeration and erroneous enumeration by rematch classification, then based on the data in Table 3, Hogan and Wolter (1988) conclude that the original rate of erroneous enumerations (EE), EP/(CP + EP) = 325/(325 + 19,269) = .016 should be increased to about ER/(CR + ER) = 411/(411 + 19,334) = .021.

Table 1

Results of 1986 Los Angeles Test Census Rematch Study: P-Sample. Source: Hogan and Wolter (1988)

Original Match Classification	Rematch Classification			
	Matched	Not Matched	Un-resolved	Total
Matched	16,623	18	55	16,696
Not matched	88	2,164	56	2,308
Unresolved	17	0	132	149
Total	16,728	2,182	243	19,153

Table 2

Data and Dual-System Estimate for 1986 Los Angeles Test Census. Source: Hogan and Wolter (1988)

		PES		
		Counted	Missed	Total
Correct Census Enumerations*	Counted	298,204	45,463	343,667
	Missed	38,503	5,870	44,373
	Total	336,707	51,333	388,040

* Correct Enumerations = Total Census Enumerations − Substitutions − Erroneous Enumerations.

Table 3

Results of 1986 Los Angeles Test Census Rematch Study: E-Sample. Source: Hogan and Wolter (1988)

Original Match Classification	Rematch Classification			
	Correct Enumer-ation	Erroneous Enumer-ation	Unresolved	Total
Correct enumeration	19,153	28	88	19,269
Erroneous enumeration	41	283	1	325
Unresolved	140	100	223	463
Total	19,334	411	312	20,057

We now reanalyze the data in Table 2 using model (B), but ignoring the unresolved cases in Table 1 because their classification status are unavailable to us. From the data in Table 1 we estimate $\hat{\gamma} = 1 - \hat{\alpha} = 88/(16{,}623 + 88) = .53\%$, and $\hat{\beta} = 18/(18 + 2{,}164) = .82\%$. In Table 4, we present the estimates and associated standard deviations under model (B) and those from the traditional DSE. The standard deviations are computed using asymptotic normality, for details, see Ding (1990, 1993a, 1993b). The estimated undercount is then defined to be undercount = $(\hat{N} - \text{CEN})/\hat{N} \times 100\%$, and CEN is the total census enumerations, *i.e.*, CEN = Correct Census Enumeration + Substitutions + EE = 343,667 + 5,259 + 6,426 = 355,352. The estimates on the last row of Table 4 indicates that the undercount estimate provided by the DSE should be reduced by 8.42% − 8.05% = .37%. We recall that Hogan and Wolter (1988) argue that the original rate of EE should be increased by 2.1% − 1.6% = .5% as a result of information in the rematch study. This then gives an additional adjustment to the estimated undercount of about .5%. Overall, we estimate that the undercount estimate was biased upward by about .9% (assuming the overlapping is negligible, even though two components are not strictly additive).

Table 4

Comparison of Estimates for 1986 Los Angeles Test Census

Parameter	DSE (SD)	MLE from Model (B) (SD)
p_1	.8856 (5.48×10^{-4})	.8892 (5.51×10^{-4})
p_2	.8677 (5.78×10^{-4})	.8712 (5.86×10^{-4})
N	388,040 (87)	386,470 (79)
Undercount (%)	8.42%	8.05%

Table 5

13 Evaluation Post-strata (EPS) for 1990 PES

1	Northeast, Central City, Minority
2	Northeast, Central City, Nonminority
3	U.S., Noncentral City, Minority
4	Northeast, Noncentral City, Nonminority
5	South, Central City, Minority
6	South, Central City, Nonminority
7	South, Noncentral City, Nonminority
8	Midwest, Central City, Minority
9	Midwest, Central City, Nonminority
10	Midwest, Noncentral City, Nonminority
11	West, Central City, Minority
12	West, Central City, Nonminority
13	West, Noncentral City, Nonminority + Indian

Table 6

Dual System Data for 13 EPS of 1990 PES

EPS	x_{1+} (Census)	x_{+1} (P-sample)	x_{11}
1*	5,966,529	4,656,305.09	4,284,132.78
2	9,235,705	8,685,235.79	8,626,362.34
3*	24,255,611	22,628,349.88	21,068,045.55
4	31,173,378	30,150,266.34	29,966,142.62
5*	9,985,055	8,809,620.02	8,249,407.92
6	13,977,529	13,582,482.34	13,278,614.01
7	47,548,548	44,059,397.93	42,987,517.59
8*	4,060,286	3,714,168.27	3,520,314.04
9	11,826,352	10,058,288.52	9,854,052.95
10	39,343,787	38,358,735.32	38,031,852.01
11*	7,283,885	5,743,998.39	5,365,961.67
12	11,073,872	10,512,339.59	10,222,147.69
13	26,415,232	26,721,116.28	26,025,370.25

*Corresponds to minority post-stratum.

Table 7

Results of Rematch Study for 13 EPS of 1990 PES: P-Sample

EPS	y_{11}	y_{21}	y_{12}	y_{22}
1*	14,301	124	31	2,773
2	15,051	36	16	1,136
3*	28,784	293	49	4,166
4	32,753	703	27	2,058
5*	28,674	189	18	3,738
6	21,757	69	36	1,156
7	48,061	47	20	3,278
8*	14,800	58	21	2,527
9	16,527	39	20	874
10	43,721	120	107	1,664
11*	12,522	133	11	2,097
12	15,122	59	8	1,078
13	43,356	232	108	4,583

Table 8

Results of Rematch Study for 13 EPS of 1990 PES: E-Sample

EPS	CP	EP	CR	ER
1*	17,027	1,415	17,106	1,645
2	15,821	879	15,631	932
3*	32,420	2,430	32,322	2,446
4	33,369	1,242	32,922	1,665
5*	32,412	1,880	33,030	2,044
6	24,392	1,225	24,336	1,284
7	51,107	2,908	50,929	3,047
8*	17,174	1,518	17,133	1,526
9	18,279	648	18,228	656
10	44,450	1,604	44,584	1,631
11*	13,644	985	13,693	909
12	15,647	522	15,590	583
13	49,647	2,062	49,545	2,334

5.2 Application of Multiple Strata Model to 1990 Census

We now analyze stratified data from the evaluation of the PES carried out as part of 1990 decennial census. Hogan (1993) describes operations and results for the 1990 PES, Mulry and Spencer (1991, 1993) present total error analysis, and Davis *et al.* (1991) report on the PES Matching Error Study (MES). The MES was conducted for each of 13 Evaluation Post-strata (EPS) by geographic region and ethnic group. Of the 13 EPS listed in Table 5, five correspond to substantial minority populations (Blacks and Hispanics), *i.e.*, EPS 1, 3, 5, 8 and 11. In Table 6, we present the dual system data for each of the 13 EPS, and we give, in Table 7 and Table 8, relevant rematch data for the P-sample and E-sample. These data are drawn from the final reports on PES evaluation projects P7 and P10 by the Census Bureau (Davis and Biemer 1991a, 1991b). The P-sample for the 1990 PES consisted of about 172,000 housing units (Hogan 1992). The P-sample data are weighted to get estimates of x_{+1} (P-sample total) and x_{11} (total matches) in the usual analysis of the dual system data and the analysis presented here. Nevertheless, the actual unweighted P-sample data can be used to make inference, see Appendix for comparison between estimates from actual P-sample data and estimates from weighted P-sample data.

In Table 9, we give the usual dual system estimates and standard deviations of the capture probabilities (*i.e.*, coverage rate by Census or P-sample) for each of the 13 EPS. Estimates in Table 10 indicate that there is significant variation in matching error rates across the EPS. Among three EPS with $\hat{\gamma}$ larger than .01%, EPS 3 and EPS 11 are minority post-strata. This suggests that the nonmatch rate may be higher for minority post-strata than for the remainder. On the other hand, there is no clear evidence from the estimates of $\hat{\beta}$ that the false match rate is higher

Table 9

Usual Dual System Estimates and Standard Deviations for 13 EPS of 1990 PES

EPS	$\hat{p}_1$ (SD)	$\hat{p}_2$ (SD)	$\hat{N}$ (SD)
1*	0.92007 (12.57×10^{-5})	0.71803 (18.42×10^{-5})	6,484,855 (470)
2	0.99322 (2.78×10^{-5})	0.93402 (8.17×10^{-5})	9,298,737 (67)
3*	0.93105 (5.33×10^{-5})	0.86858 (6.86×10^{-5})	26,051,987 (540)
4	0.99389 (1.42×10^{-5})	0.96127 (3.46×10^{-5})	31,364,919 (88)
5*	0.93641 (8.22×10^{-5})	0.82618 (11.99×10^{-5})	10,663,134 (390)
6	0.97763 (4.01×10^{-5})	0.95000 (5.83×10^{-5})	14,297,391 (131)
7	0.97567 (2.32×10^{-5})	0.90408 (4.27×10^{-5})	48,734,156 (359)
8*	0.94781 (11.54×10^{-5})	0.86701 (16.85×10^{-5})	4,283,875 (190)
9	0.97969 (4.45×10^{-5})	0.83322 (10.84×10^{-5})	12,071,466 (224)
10	0.99148 (1.48×10^{-5})	0.96665 (2.86×10^{-5})	39,681,946 (108)
11*	0.93419 (10.35×10^{-5})	0.73669 (16.32×10^{-5})	7,797,041 (443)
12	0.97240 (5.05×10^{-5})	0.92309 (8.01×10^{-5})	11,388,243 (164)
13	0.97396 (3.08×10^{-5})	0.98524 (2.35×10^{-5})	27,121,400 (104)

Table 10

Estimates of Matching Error Rates for 13 EPS of 1990 PES

EPS	$\hat{\gamma}$(%)	$\hat{\beta}$(%)
1*	0.009	0.011
2	0.002	0.014
3*	0.010	0.012
4	0.021	0.013
5*	0.007	0.005
6	0.003	0.030
7	0.001	0.006
8*	0.004	0.008
9	0.002	0.022
10	0.003	0.060
11*	0.011	0.005
12	0.004	0.007
13	0.005	0.023

Table 11

MLEs from Model (B′) and Standard Deviations for 13 EPS of 1990 PES

EPS	$\hat{p}_1$ (SD)	$\hat{p}_2$ (SD)	$\hat{N}$ (SD)
1*	0.92406 (12.68×10^{-5})	0.72114 (18.79×10^{-5})	6,456,833 (446)
2	0.99464 (2.79×10^{-5})	0.93536 (8.30×10^{-5})	9,285,474 (92)
3*	0.93896 (5.38×10^{-5})	0.87597 (7.01×10^{-5})	25,832,352 (279)
4	0.99999 (2.65×10^{-5})	0.98070 (3.64×10^{-5})	30,731,889 (781)
5*	0.94166 (8.28×10^{-5})	0.83080 (12.13×10^{-5})	10,603,717 (306)
6	0.97922 (4.03×10^{-5})	0.95154 (6.03×10^{-5})	14,274,182 (64)
7	0.97600 (2.32×10^{-5})	0.90438 (4.30×10^{-5})	48,717,792 (338)
8*	0.95034 (11.59×10^{-5})	0.86933 (17.06×10^{-5})	4,272,459 (159)
9	0.97756 (4.47×10^{-5})	0.83141 (11.12×10^{-5})	12,097,806 (285)
10	0.99217 (1.50×10^{-5})	0.96733 (3.06×10^{-5})	39,654,306 (90)
11*	0.94239 (10.46×10^{-5})	0.74316 (16.58×10^{-5})	7,729,158 (359)
12	0.97561 (5.07×10^{-5})	0.92614 (8.10×10^{-5})	11,350,674 (101)
13	0.97895 (3.10×10^{-5})	0.99029 (2.42×10^{-5})	26,983,168 (355)

Table 12

Undercount Percentage and Bias Estimates for 13 EPS of 1990 PES

EPS	UC(DSE)	UC(P)	UC(E)	UC(T)	Bias(P)	Bias(E)	Bias(T)
1*	6.40	5.99	5.30	4.89	0.41	1.10	1.51
2	−0.69	−0.83	−1.05	−1.20	0.14	0.36	0.51
3*	5.59	4.79	5.53	4.72	0.80	0.06	0.87
4	−0.11	−2.17	−1.33	−3.39	2.06	1.23	3.29
5*	5.03	4.49	4.68	4.15	0.53	0.35	0.88
6	1.22	1.06	0.99	0.83	0.16	0.23	0.39
7	1.77	1.73	1.50	1.47	0.03	0.26	0.29
8*	3.52	3.26	3.46	3.20	0.26	0.06	0.32
9	1.05	1.26	1.00	1.21	−0.22	0.05	−0.17
10	0.41	0.34	0.36	0.29	0.07	0.05	0.12
11*	5.26	4.43	5.77	4.94	0.83	−0.51	0.32
12	1.89	1.56	1.51	1.19	0.32	0.38	0.70
13	1.79	1.29	1.28	0.78	0.50	0.51	1.01

for minority post-strata, or the other way around. In Table 11, we give maximum likelihood estimates and standard deviations under model (B′). Heterogeneity in the capture probabilities is significant. This heterogeneity together with the variation in the matching error rates suggests that model (B′) is more appropriate than model (B). The asymptotic standard deviations in Table 9 and 11 appear unusually small comparing to the sample size of N. Ding (1993b) shows that this is a typical feature of the dual system problem when the capture probabilities are very high, as it is the case in census application. Despite very narrow confidence intervals, simulation studies in Ding (1993b) show that the asymptotic normal approximation being used is highly accurate in terms of coverage probability.

Table 12 provides estimates of matching bias of various sources in the undercount estimate by the usual DSE. UC(DSE) is the undercount estimate from the DSE defined in the same way as for the 1986 TARO estimate; UC(P) is the undercount estimate computed by MLE from matching error model to adjust for matching bias in P-sample, and Bias(P) = UC(DSE) − UC(P). Again, following Hogan and Wolter (1988), we define the bias in E-sample operation by Bias(E) = ER/(CR + ER) − EP/(CP + EP), and the undercount estimate correcting for E-sample error by UC(E) = UC(DSE) − Bias(E). Finally the total matching bias by both P-sample and E-sample is Bias(T) = Bias(P) + Bias(E), and the undercount estimate correcting for both sources of error is UC(T) = UC(DSE) − Bias(T). Note that it is possible, as observed for EPS 2 and 4 in Table 12, that undercount estimate is negative, thus indicating an overcount instead. This happens when the DSE (or MLE) is less than CEN, the total census enumeration. The dual system data represents "corrected" census counts with erroneous and other incorrect enumerations excluded from CEN.

For each of Bias(P), Bias(E) and Bias(T), a positive estimate indicates a upward bias in the undercount estimate from the DSE by ignoring the corresponding source of error, that is, UC(DSE) should be reduced by the estimated bias to account for that source of error. For each of UC(DSE), UC(P), UC(E) and UC(T), we get significantly higher undercount figures for each of the five minority post-strata, *i.e.*, EPS 1, 3, 5, 8 and 11. For both Bias(P) and Bias(E), all the bias estimates are positive except for Bias(P) for post-stratum 9 and Bias(E) for post-stratum 11. This supports the common belief that there is usually an upward bias attributable to matching errors in the undercount estimate by the DSE, except for some non-minority geographical areas where in fact there is disproportionately large share of erroneous enumerations.

The effects of the two types of matching errors are well understood. False nonmatches results in upward bias and false matches produce downward bias. The nature of the overall matching bias is then dependent upon which type of matching error dominates. By computing undercount estimates for 1980 Census data with selective pair of γ and β, Ding (1990) concludes that due to high capture probabilities in the census application of the capture-recapture technique, the matching bias is dominated by the false nonmatch rate when the false nonmatch rate (γ) and the false match rate (β) are about the same magnitude. This point can be easily confirmed here. EPS 4 has the largest estimate of γ, $\hat{\gamma}$ = .021% and results in the largest Bias(P) = 2.06%. EPS 3 and EPS 4 have about the same estimate of β, $\hat{\beta}$, .012% and .013%, respectively, but EPS 3 has much smaller Bias(P) = .80%, due to smaller estimate of γ, $\hat{\gamma}$ = .010%. About a .01% difference in $\hat{\gamma}$ gives dramatic difference in Bias(P). For matches and nonmatches with complete data, Fay *et al.* (1988, p. 53) state "Because of sometimes difficult nature of the matching work, false nonmatches probably represent a greater concern than false matches". The data analyzed by our methods include both complete data and data produced as a result of the Bureau's imputation procedure. The sensitivity of our estimates to γ lends some support to the statement by Fay *et al.* when both matching for complete data and matching for imputed data are considered together. On the other hand, a downward bias can be observed when $\hat{\beta}$ is much larger than $\hat{\gamma}$. For EPS 9, $\hat{\beta}$ = .022%, about 10 times as large as $\hat{\gamma}$ = .002%. Thus false matches dominate false nonmatches for this stratum, and we see the only negative (downward) bias, Bias(P) = −.22%.

For a specific matching procedure there is an inevitable trade-off between matching errors and unresolved cases. Depending on the extent of unresolved cases and the imputation algorithm used, the resolution process might yield a significant number of false matches. The empirical evidence accumulated by the Bureau of the Census, as we note above, lends some support for the "unbiasedness" of the missing data mechanism used in the imputation process in our example, but further evidence on the issue is desirable.

6. SUMMARY

In this article, we have presented models and methods for the estimation of population total and census undercount that corrects for matching bias of the usual dual-system estimate in the presence of matching errors. Two sources of information are combined in the estimation procedure, the dual-system or capture-recapture census data, and the data from a matching error study (rematch study). The accuracy of our estimates relies on the assumption that the rematch is error free. Matching error rates are likely not to be homogeneous over different population strata. Model (B′) allows for heterogeneity of matching error rates across various population strata but requires stratified rematch data to estimate the error parameters within strata. The methods presented here generalize the standard theoretical framework for the use of maximum likelihood estimation to accommodate matching errors.

We can adjust for erroneous enumerations in the estimate of EE by the use of rematch data for the E-sample. We obtain an overall matching bias in the DSE by adding two bias components from the P-sample and the E-sample. Our analysis of the 1986 Los Angeles test census data indicates that the upward bias of the DSE in the estimate of the census undercount is just under 1%, thereby lending support to the 1% value used by Hogan and Wolter (1988) in their evaluation study. For the analysis on 1990 Census data, the computational results not only agree with understood aspects of matching bias, but also offer findings that were not previously known.

For simplicity, we have assumed that the PES is (allowing for stratification) based on simple random sampling. The models still need to be adapted to account for the complex sampling design actually used (see Hogan 1992, 1993).

It has been known that the perfect matching assumption does not hold in the application of dual system estimation in the U.S. census. The matching problem in the use of the DSE has two components. The first component involves the missing P-sample enumeration status. The second involves errors in classifying P-sample people as enumerated or not. The present paper provides a method to address both components using dual system data adjusted for imputed enumeration probabilities, and can be of possible value in future censuses provided that the models are adapted to handle the complex survey design of the PES. Ding (1993c) develops estimates to directly address the first component by modifying the usual DSE method and describes the relationship between the proposed estimates and those that result from the application of the Census Bureau's imputation scheme for missing P-sample enumeration status (Schenker 1988, Belin *et al.* 1993).

ACKNOWLEDGMENTS

Fienberg's work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada to York University, Toronto, Canada. The authors are grateful to Mary Mulry for furnishing data on 1990 Decennial Census, to Joe Sedransk for suggestions, and to Jay Kadane, Larry Wasserman and Mike Meyer for commenting on an earlier version of this work. An Associate Editor and two referees provided comments that have led to a sharpening of the discussion. The basic models in this manuscript were first developed as part of the first author's Ph.D. thesis at Carnegie Mellon University.

APPENDIX

Comparison of Estimates from Weighted and Unweighted P-Sample Data

For simplicity, we assume a weight $k > 1$ for the P-sample and consider the usual dual system estimation problem. Let $\{x_{ij}\}$ be the cell counts in the 2×2 table for weighted P-sample data and census enumerations, $i, j = 1, 2$ and $ij \neq 22$. One could make inference with unweighted P-sample data and census enumerations deflated by a factor of k to get cell counts $\{y_{ij}\}$, $i, j = 1, 2$ and $ij \neq 22$. Then $x_{ij} = ky_{ij}$, $ij \neq 22$, and $x_{1+} = ky_{1+}$, $x_{+1} = ky_{+1}$. Let the usual dual system estimates derived from $\{x_{ij}\}$ be $\hat{p}_1, \hat{p}_2$ and $\hat{N}_w$, and estimates from $\{y_{ij}\}$ be $\hat{q}_1$, $\hat{q}_2$ and $\hat{N}_u$. The estimates are (Bishop *et al.* 1975, chap. 6) $\hat{p}_1 = x_{11}/x_{+1} = y_{11}/y_{+1} = \hat{q}_1, \hat{p}_2 = x_{11}/x_{1+} = y_{11}/y_{1+} = \hat{q}_2, \hat{N}_w = x_{1+}x_{+1}/x_{11} = ky_{1+}y_{+1}/y_{11} = k\hat{N}_u$. Thus if one considers the unweighted P-sample data and uses $\hat{N}_* = k\hat{N}_u$ to estimate the population total, then $\hat{q}_1$, $\hat{q}_2$ and $\hat{N}_*$ give the same point estimates as $\hat{p}_1$, $\hat{p}_2$ and $\hat{N}_w$ from weighted P-sample data. From the asymptotic normal distribution of the estimates (Ding 1993b), we have $\mathrm{Var}(\hat{N}_w) = k\mathrm{Var}(\hat{N}_u)$, $\mathrm{Var}(\hat{q}_1) = k\mathrm{Var}(\hat{p}_1)$, $\mathrm{Var}(\hat{q}_2) = k\mathrm{Var}(\hat{p}_2)$. Then $\mathrm{Var}(\hat{N}_*) = k\mathrm{Var}(\hat{N}_w)$, and $\hat{q}_1$, $\hat{q}_2$ and $\hat{N}_*$ have larger variance than $\hat{p}_1$, $\hat{p}_2$ and $\hat{N}_w$, respectively. To compute estimates with unweighted P-sample data, one needs to know k and $\{y_{ij}\}$. We emphasize that the trivial case of a constant sampling weight for all cases in the same post-stratum is assumed here for simplicity of discussion. However, the real situation can be complex. For example, Blacks may be sampled at a low probability in a White stratum and are then combined with other Blacks sampled with much higher probabilities.

REFERENCES

ALHO, J.M., MULRY, M.H., WURDEMAN, K., and KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.

BELIN, T.R., DIFFENDAL, G.J., MACK, S., RUBIN, D.B., SCHAFER, J.L., and ZASLAVSKY, A.M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*, 88, 1149-1166.

BIEMER, P.P. (1988). Modeling matching error and its effect on estimates of census coverage error. *Survey Methodology*, 14, 117-134.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: M.I.T. Press.

CHAPMAN, D.C. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, 1, 131-160.

CHILDERS, D., DIFFENDAL, G., HOGAN, H., and MULRY, M. (1989). Coverage Evaluation Research: the 1988 Dress Rehearsal. Paper presented to the Census Advisory Committee of the American Statistical Association and the Census Advisory Committee on Population Statistics at the Joint Advisory Committee Meeting, Alexandria, VA.

DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of American Statistical Association*, 88, 1137-1148.

DAVIS, M.C., MULRY, M., PARMER, R., and BIEMER, P. (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. (1991a). Estimates of P-Sample Clerical Matching Error from a Rematching Evaluation. Report on Post-Enumeration Survey Evaluation Project P7, U.S. Department of Commerce, Bureau of the Census.

DAVIS, M.C., and BIEMER, P. (1991b). Measurement of the Census Erroneous Enumerations: Clerical Error Made in the Assignment of Enumeration Status. Report on Post-Enumeration Survey Evaluation Project P10, U.S. Department of Commerce, Bureau of the Census.

DING, Y. (1990). Capture-recapture Census with Uncertain Matching. Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.

DING, Y. (1993a). On the asymptotic normality of multinomial population size estimates with application to the backcalculation estimates of AIDS epidemic. To appear in *Biometrika*.

DING, Y. (1993b). On the asymptotic normality of dual system estimates. Unpublished manuscript.

DING, Y. (1993c). Capture-recapture census with probabilistic matching. Submitted for publication.

DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in central Los Angeles county. *Survey Methodology*, 14, 71-86.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *Journal of American Statistical Association*, 80, 98-131.

FAY, R.E., PASSEL, J.S., ROBINSON, J.G., and COWAN, C.D. (1988). The Coverage of Population in the 1980 Census. U.S. Department of Commerce, Bureau of the Census.

FIENBERG, S.E. (1989). Undercount in the U.S. decennial census. *Encyclopedia of Statistical Sciences, Supplement Volume*, 181-185.

JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 test census of Tampa, Florida. *Journal of American Statistical Association*, 84, 414-420.

JEFFERSON, T. (1986). Letter to David Humphreys. *The Papers of Thomas Jefferson*, 22, 62.

HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.

HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.

HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology*, 14, 99-116.

MULRY, M.H., and SPENCER, B.D. (1991). Total error in PES estimates of population: the dress rehearsal census of 1988 (with discussion). *Journal of American Statistical Association*, 86, 839-854.

MULRY, M.H., and SPENCER, B.D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of American Statistical Association*, 88, 1080-1091.

RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.

SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.

SCHENKER, N. (1988). Handling missing data in coverage estimation with application to the 1986 test of adjustment related operations. *Survey Methodology*, 14, 87-98.

WOLTER, K. (1986). Some coverage error models for census data. *Journal of American Statistical Association*, 81, 338-346.