Survey Methodology, June 1988
Vol. 14, No. 1, pp. 117-134
Statistics Canada

Modeling Matching Error and its Effect on Estimates of Census Coverage Error

PAUL P. BIEMER[1]

ABSTRACT

Dual system estimators of census undercount rely heavily on the assumption that persons in the evaluation survey can be accurately linked to the same persons in the census. Mismatches and erroneous nonmatches, which are unavoidable, reduce the accuracy of the estimators. Studies have shown that the extent of the error can be so large relative to the size of census coverage error as to render the estimate unusable. In this paper, we propose a model for investigating the effect of matching error on the estimators of census undercount and illustrate its use for the 1990 census undercount evaluation program. The mean square error of the dual system estimator is derived under the proposed model and the components of MSE arising from matching error are defined and explained. Under the assumed model, the effect of matching error on the MSE of the estimator of census undercount is investigated. Finally, a methodology for employing the model for the optimal design of matching error evaluation studies will be illustrated and the form of the estimators will be given.

KEY WORDS: Undercount; Dual system estimation; Capture-recapture; Nonsampling error; Processing error.

1. INTRODUCTION

The use of capture-recapture methods for census evaluation and the evaluation of birth-death registration was first suggested by Sekar and Deming (1949). For estimating census coverage error, the method involves matching persons from a sample survey of the population to the census in order to determine the number of individuals which were enumerated in both the sample survey and the census. There are a number of difficulties which may occur in the capture-recapture method to cause substantial biases in an estimate of the total population size, N (see for example Burnham *et al.* 1987 and Wolter 1986). A problem which occurs quite often in applications of the procedure is the failure to accurately match persons from the sample survey to the census. Seltzer and Adlakta (1974) demonstrated that matching error can result in relative biases as large as 33% and may be positive or negative depending upon whether false nonmatches or false matches predominate (see also Scheuren and Oh 1985). Wolter (1983) notes that suspected matching errors in the 1980 Post Enumeration Program were a part of the reason not to adjust the 1980 U.S. Census.

This paper provides a basic framework for evaluating the matching error in capture-recapture studies (particularly for applications to human populations) and for assessing the impact of the errors on the accuracy of the estimate of N. To provide a simple and familiar basis for the discussion of matching error, we shall adopt the original Sekar-Deming capture-recapture model. Extensions of the Sekar-Deming technique are given in Marks, Seltzer and Krotki (1974), and Wolter (1986).

[1] Paul P. Biemer,Head, Department of Experimental Statistics, Director, University Statistics Center, New Mexico State University, Las Cruces, New Mexico, United States.

Consider a population U and let N denote the size of U. A census is conducted and N_c persons are counted. We wish to estimate N-N_c (referred to as the coverage error of the census) which is equivalent to estimating N. A post enumeration survey (PES) is conducted which employs the same reference period as the census. We assume that: (a) both the census and the PES contain no spurious events (i.e., duplications, fabrications, out-of-scope persons or unidentifiable persons) or that the number of such events can be accurately estimated and subtracted from N_c; and (b) the event of being counted in the census is independent of the event of being counted in the PES.

The PES persons are matched to the census in order to determine the number of PES persons who were also counted in the census. Let x_{11} denote the design unbiased estimator of the total number of persons in both the PES and the census populations and let N_p denote the design unbiased estimator of the PES population size. The Sekar-Deming estimator (more recently referred to as the dual system estimator or DSE) of N is

$$\hat{N} = \frac{N_p N_c}{x_{11}} \tag{1}$$

As we shall see, $\hat{N}$ is subject to two sources of error: sampling error and nonsampling error. Although there may be several sources of nonsampling error, the source of the error of concern here is matching error; i.e., the misclassification of PES persons as enumerated in the census (false positive errors) or not enumerated in the census (false negative errors).

Using Taylor series expansions, general forms for the moments of $\hat{N}$ can be derived. It can be shown that, to terms of order $1/n$, where n is the PES sample size,

$$\text{Bias } (\hat{N}) \doteq -N[\text{Relbias } (\hat{p}_{11}) - \text{Relvar } (\hat{p}_{11})] \times [1 + \text{Relbias } (\hat{p}_{11})]^{-1} \tag{2}$$

and

$$\text{Var } (\hat{N}) \doteq N^2 \text{ Relvar } (\hat{p}_{11}) [\, 1 + \text{Relbias } (\hat{p}_{11}) \,]^{-2} \tag{3}$$

where $\hat{p}_{11} = x_{11}/N_p$ is an estimator of p_{11}, the true proportion of the PES population falling in the census population; Relbias $(\hat{p}_{11})$ = Bias $(\hat{p}_{11})/p_{11}$; and Relvar $(\hat{p}_{11})$ = Var$(\hat{p}_{11}) \times E^{-2}(\hat{p}_{11})$. Here we have assumed that N_c, the census counts, has a variance of zero. This is a simplification since, as we mentioned, an estimate of the census spurious events may have been subtracted from the census count to obtain N_c and this correction may be subject to sampling and other errors. Nevertheless, the assumption is consistent with our emphasis in this paper on matching error and its effect on $\hat{N}$. The last section discusses an extension of the methodology which allows error in the estimator N_c.

From (2) and (3) we note that the total mean square error (MSE) of $\hat{N}$ depends upon the total MSE of $\hat{p}_{11}$. In the following section, we consider some models for evaluating the effects of matching error on $\hat{p}_{11}$. Letting j $(j=1,\ldots,n)$ be the index for the j^{th} individual in the PES sample, we define α_j as the probability that individual j is misclassified in the matching process and consider alternative assumptions regarding the probabilities α_j.

2. MATCHING ERROR MODELS

2.1 Uncorrelated Matching Error

Assume:

1. The event {unit j is misclassified} is independent of the event {unit j' is misclassified} for all $j \neq j'$.
2. $\alpha_j = \theta$ if unit j is truly in the census, referred to as the probability of a false negative error, and $\alpha_j = \phi$ if unit j is truly not in the census, referred to as the probability of a false positive error.

To fix the ideas, we assume simple random sampling for the PES and that n is small relative to N, then

$$E(\hat{p}_{11}) = p_{11}(1-\theta) + (1-p_{11})\phi, \tag{4}$$

$$\text{Bias}(\hat{p}_{11}) = -p_{11}\theta + (1-p_{11})\phi \tag{5}$$

$$\text{Var}(\hat{p}_{11}) = n^{-1} E(\hat{p}_{11})\,(1-E(\hat{p}_{11})) \tag{6}$$

$$= n^{-1}\,(SV + SMV),$$

where SV, denoting *sampling variance*, is given by

$$SV = p_{11}(1-p_{11})\,(1-\theta-\phi)^2 \tag{7}$$

and where SMV, denoting *simple matching variance*, is given by

$$SMV = p_{11}\theta(1-\theta) + (1-p_{11})\,\phi(1-\phi) \tag{8}$$

(proof in the appendix).

Readers familiar with the Hansen, Hurwitz, and Pritzker (1964) response error model will recognize the correspondence of their simple response variance and SMV in this model. Hansen, *et al.* define a measure I, referred to as the "index of [response] inconsistency," to be the ratio of the simple response variance to the total variance of a single response, i.e., the proportion of variance which is response variance. For survey responses, I is an indicator of the response reliability of the survey information. An analogous measure can be obtained for matching error to indicate the effect on the variance of $\hat{p}_{11}$ of matching unreliability. This measure, denoted by I_M, is given by

$$I_M = \frac{SMV}{SV + SMV}. \tag{9}$$

For some applications, assumptions (1) and (2) may be too restrictive. The independence assumption (1) is violated, for example, when unit B in the PES is erroneously matched to unit A in the census causing the correct match, unit A in the PES, to be erroneously classified as a nonmatch. Since this implies that the errors for units A and B are negatively correlated, the consequence is that $\text{Var}(\hat{p}_{11})$ will be smaller than given by (6). However, $E(\hat{p}_{11})$ is not affected by correlated errors. Another form of correlated matching error arises when matching is performed by clerks who may vary in their tendencies to commit false positive and false negative errors. The next section provides a model that describes these errors.

Assumption 2 specifies that the misclassification probabilities α_j are homogeneous across the PES population. This too may be a simplification since some individuals, perhaps the majority, may be classified with relatively little risk of error while other individuals are more difficult to match. Basically, matching problems arise from inaccurate or incomplete information about the characteristics of each individual in either or both systems. Therefore, if the PES sample can be post-stratified on the basis of the completeness of the information to be used for matching, the assumption may hold (at least approximately) within each stratum. The overall matching error rate is thus an aggregation of the individual stratum error rates. The last subsection explores this model.

Finally, the assumption of simple random sampling greatly reduces the complexity of the formula for $\text{Var}(\hat{p}_{11})$. Since PES samples are complex samples, the assumption is a simplification, yet it still provides useful formulas for: (a) identifying which components of matching error are likely to have the greatest impact on the total MSE of $\hat{N}$; and (b) allocating resources for and designing matching error evaluation studies. In many situations, an adjustment of SV by a "design effect" constant will account for most of the effect of complex sampling on $\text{Var}(\hat{p}_{11})$. Further, $E(\hat{p}_{11})$ is essentially unaffected by more complex forms of sampling than simple random sampling as long as $\hat{p}_{11}$ is appropriately weighted. Thus, the form of $B(\hat{p}_{11})$ does not depend upon this assumption.

2.2 Modeling Clerical Error

Suppose the PES is matched clerically to the census using k clerks. Let m_i denote the number of PES individuals classified by clerk i, $i=1,\ldots,k$. Let the double index (i,j) denote the j^{th} individual in the i^{th} clerk's assignment.

Assume:

1. The event {unit (i,j) is misclassified} and the event {unit (i',j') is misclassified} are independent when $i \neq i'$ and conditionally independent given clerk i for $i=i'; j\neq j'; i=1, \ldots, k; j,j' = 1,\ldots,m_i$.
2. $\alpha_{ij} = \theta_i$ if individual (i,j) is truly in the census, and $= \phi_i$ if individual (i,j) is truly not in the census.
3. $E(\theta_i) = \theta$; $E(\phi_i) = \phi$; $\text{Var}(\phi_i) = \sigma_\phi^2$; $V(\phi_i) = \sigma_\phi^2$; and $\text{Cov}(\theta_i, \phi_i) = \sigma_{\phi\theta}$.

For the subset of individuals in the i^{th} clerk's assignment, 1 and 2 are analogous to assumptions 1 and 2 for the model of the last section. Assumption 3 specifies that clerk matching error probabilities are independent and identically distributed random variables. This assumption is analogous to the assumptions made for interviewer errors in interviewer effect models (see for example Kish 1962, Hartley and Rao 1978 and Biemer and Stokes 1985). The assumption is appropriate if our interest lies in estimating the parameters of a much larger pool of clerks of which the k PES clerks are a representative sample.

It is shown in the appendix that, assuming simple random sampling, $E(\hat{p}_{11})$ is still given by (4). The general formula for $\text{Var}(\hat{p}_{11})$ is given by (A.3) in the appendix; however, a useful simplification results if we can assume that the assignment sizes m_i are approximately equal to m, the average size, and that each clerk's assignment has the same expected number of matches (i.e., clerk assignments are interpenetrated). Then

$$\text{Var}(\hat{p}_{11}) = \frac{1}{n}(SV + SMV) + \frac{m-1}{m}\frac{1}{k}CC \tag{11}$$

where CC, denoting the *correlated component of matching variance*, is

$$CC = p_{11}^2 \sigma_\theta^2 + (1-p_{11})^2\sigma_\phi^2 - 2p_{11}(1-p_{11})\,\sigma_{\phi\theta} \tag{12}$$

and SV, SMV are given by (7) and (8), respectively.

Note that CC is a consequence of the between clerk variability of the misclassification probabilities θ_i and ϕ_i. Further, by noting that CC is the variance of $-p_{11}\,\theta_i + (1-p_{11})\phi_i$ and the similarity of these terms with (5), we see that CC is the variance of the *net* biases among clerks. This latter fact proves that CC must be positive. Therefore, the effect of clerk variance is to increase the variance of $\hat{p}_{11}$.

Borrowing again from the response variance literature, we can define a parameter ρ_M which is analogous to the intra-interviewer correlation coefficient, ρ, defined by Kish (1962). We shall refer to ρ_M as the intra-clerk correlation since it is the correlation between the match classifications of any two units in the same clerk assignment. Under the model,

$$\rho_M = \frac{CC}{SV + SMV}$$

is the ratio of the correlated component of variance to the total variance associated with a single classification. It may be interpreted as the degree to which clerks "influence" the match rates within their assignments. Now, an alternative formula for Var $(\hat{p}_{11})$ which is equivalent to (11) is

$$\text{Var}(\hat{p}_{11}) = \frac{SV + SMV}{n}\,[1 + (m-1)\rho_M] \tag{13}$$

2.3 Post-stratification

Both the model for uncorrelated error and the model for clerical error assume (essentially) that individuals in the PES sample do not differ in the degree of difficulty of determining their true match classification (assumption 2 for both models). For example, for the clerical error model, the misclassification probability vector (θ_i, ϕ_i) is the same for all units in the i^{th} clerk's assignment. In reality, however, some individuals are much more difficult to classify than others depending upon such factors as the completeness of the matching information, whether a mover or non-mover, whether in single family home or apartment, etc.

A simple approach for modeling this situation is to stratify PES sample according to some variable, say Z, which is correlated with the misclassification probabilities α_j. The variable Z may be an indicator of the completeness of the information, the type of unit, etc.

Suppose there are L such strata indexed by h. Let (i,h,j) denote the j^{th} unit in the h^{th} stratum in the i^{th} clerks assignment where $i=1,\ldots,k$; $h=1,\ldots,L$, $j=0,\ldots,m_{ih}$; and m_{ih} is the number of units in stratum h for the i^{th} clerk. We shall again assume (1) as for the clerical error model; however, in addition assume:

2. $\alpha_{ihj} = \theta_{ih}$ if individual (i,h,j) is truly in the census.
   $= \phi_{ih}$ if individual (i,h,j) is truly not in the census.

3. $E(\theta_{ih}) = \theta_h$; $E(\phi_{ih}) = \phi_h$
   $\text{Var}(\theta_{ih}) = \sigma_{\theta h}^2$; $Var(\phi_{ih}) = \sigma_{\phi h}^2$;
   $Cov(\theta_{ih'}, \theta_{ih}) = \sigma_{\phi\theta h}$ if $h=h'$
   $= 0$ if $h \neq h'$

Under these assumptions, we have Bias $(\hat{p}_{11}) = \Sigma\pi_h$ Bias $(\hat{p}_{11})$ and Var $(\hat{p}_{11}) = \Sigma\pi_h^2$ Var $(\hat{p}_{11h}) + \Sigma\pi_h [E(\hat{p}_{11h}) - E(\hat{p}_{11})]^2$ where Bias $(\hat{p}_{11h})$, $E(\hat{p}_{11h})$, and Var $(\hat{p}_{11h})$ are given by (5), (4), and (6), respectively, indexing the clerk error parameters and p_{11} by h and where $\pi_h = E(n_h/n)$, the proportion of the population in the h^{th} stratum.

3. DEMONSTRATION OF THE EFFECT ON TOTAL ERROR

The models of the previous section can be useful for demonstrating the effect of matching error on the total mean square error of $\hat{N}$ and $\hat{p}_{11}$. In the illustrations that follow, we shall assume values of the model parameters which are typical given our experience and which are consistent with current 1990 PES design parameters.

In the PES, estimates of N will be made for a number of census strata. We assume that the desired coefficient of variation of the estimates is 1%. Matching will be conducted in a number of processing sites by teams of clerks. (More details on the matching operation are given in the next section). To illustrate the effect of matching error on the DSE, we consider a "typical" PES stratum. For this stratum, let $p_{11} = .85$ and k, the number of matching clerks in one processing site, be 10. In our analysis, we considered values of θ which varied from 0 to .10 and a number of typical values for the ratio $\gamma = \theta/\phi$, i.e., the ratio of the probability of false negatives to the probability of false positives. Little information exists which would indicate the typical range of ρ_M since no study has ever measured ρ_M for matching error. However, if we assume that the clerk error probabilities θ_i and ϕ_i follow a unimodal beta-distribution and are uncorrelated, we can obtain a maximum value for ρ_M corresponding to given values of the expected error probabilities θ and ϕ. Algebraically, the maximum value of ρ_M is given by

$$\rho_M^* = CC^* / (SMV + SV) \tag{14}$$

where $CC^* = p_{11}^2\,\theta^2\,(1-\theta)/(1+\theta) + (1-p_{11})^2\,\phi^2(1-\phi)/(1+\phi)$ (see Johnson and Kotz 1970, for the underlying theory). If θ_i and ϕ_i are positively correlated, then the assumption of zero correlation further exaggerates the effect of CC. Thus, the illustrations which follow indicate the maximum impact of matching variance on the estimates.

To illustrate the maximum effect of correlated variance on the precision of $\hat{p}_{11}$, the coefficient of variation of $\hat{p}_{11}$, denoted by $CV(\hat{p}_{11})$, was graphed as a function of θ for various values of γ. For these calculations, ρ_M^* was substituted for ρ_M in (13). The range of θ was $0 \le \theta \le .10$ and γ was $.5 \le \gamma \le 5$; i.e., $\phi = .2\theta$ to $\phi = 2\theta$. This range of values of γ seems reasonable since, typically, ϕ is smaller than θ. Figure 1 shows the function for $\gamma = 1$. There was no discernible difference for other values of γ in the range of interest. Thus, it appears that the size of ϕ has negligible effect on $CV(\hat{p}_{11})$. In fact, we see from the expression for CC^* that when $p_{11} = .85$, no more than 3% of the correlated variance is contributed by the variance of ϕ_i even when ϕ is the same size as θ. Figure 1 also suggest that $CV(\hat{p}_{11})$ may be increased two-fold to 2% for values of θ as small as 5%.

In Figure 2, the relative bias of $\hat{p}_{11}$, denoted by $RB(\hat{p}_{11})$ is illustrated for the same range of θ; i.e., $0 \le \theta \le .1$, and γ; i.e., $.5 \le \gamma \le 5$. The graph clearly indicates that bias is smaller for smaller values of γ. In fact, the bias is zero when $\gamma = (1-p_{11})/p_{11}$ or .18 assuming $p_{11} = .85$ as in this example. For θ as small as 5%, the relative bias is between -2% and -4%, depending upon the size of γ. Comparing this with the maximum increase in $CV(\hat{p}_{11})$ of one percentage point, we see that bias has the potential to be much more serious than correlated variance.

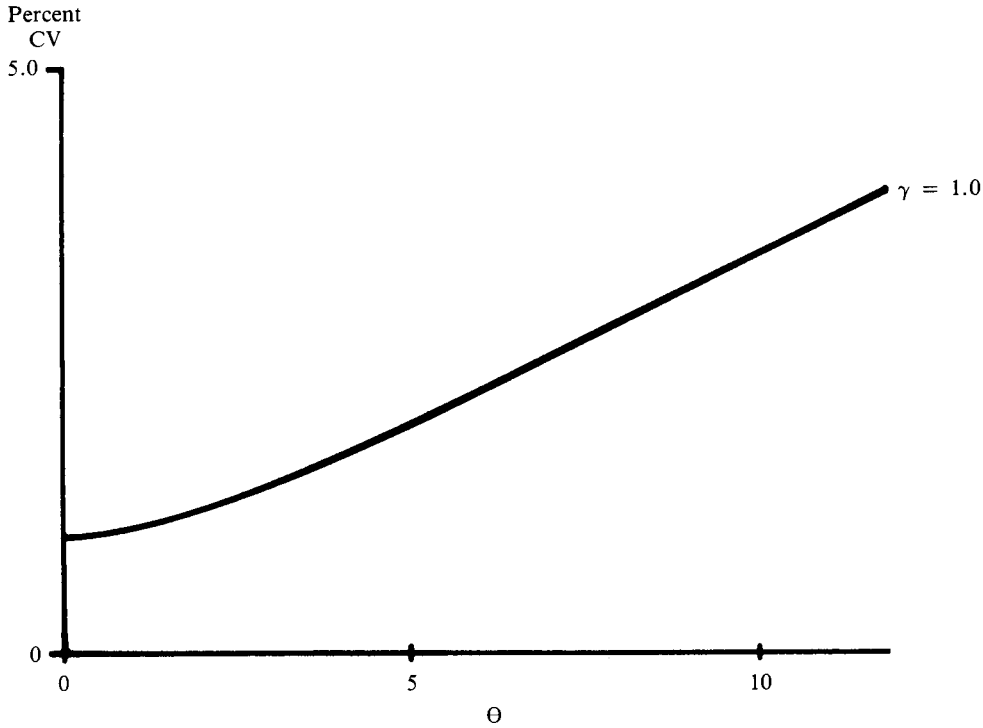

Figure 1. Coefficient of Variation of $\hat{p}_{11}$ as a Function of Θ for $\gamma = 1$

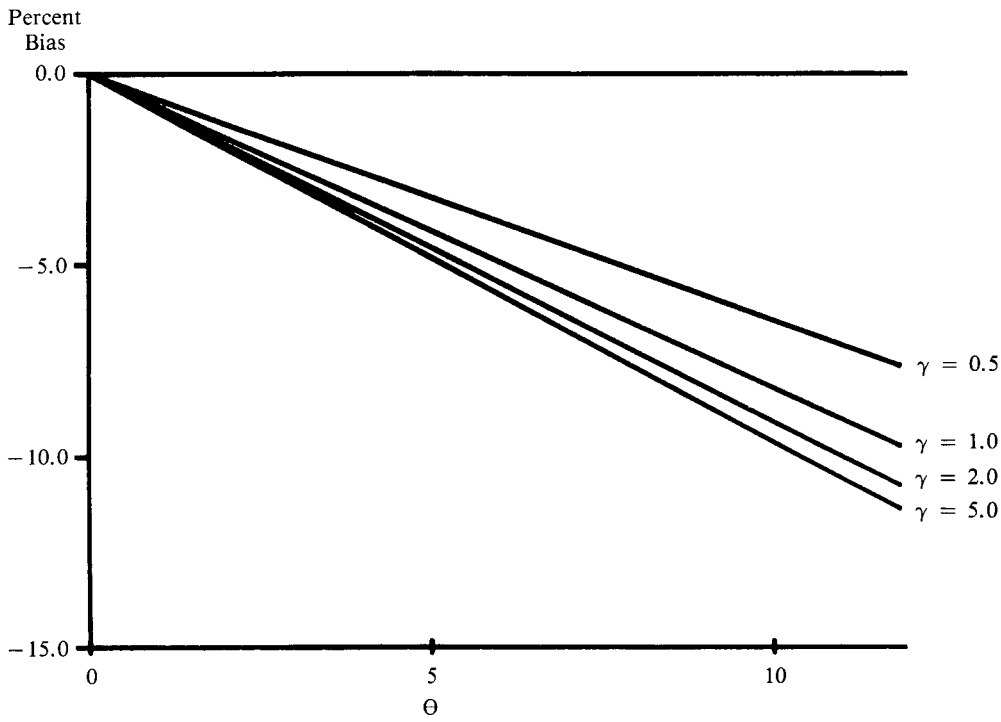

Figure 2. Relative Bias of $\hat{p}_{11}$ as a Function of Θ for Selected Values of γ

To indicate the potential effects of matching error on $\hat{N}$, the increase in total error as a function of θ and for selected values of γ was computed. Let $M(\theta;\gamma)$, $V(\theta;\gamma)$, and $B(\theta;\gamma)$ denote the mean square error, variance, and bias, respectively of $\hat{N}$ for given values of θ and γ. $M(0;\gamma)$ is the mean square error of $\hat{N}$ without matching error (i.e. $\theta = \phi = 0$) and thus $M(0;\gamma)^{1/2}$ is approximately the standard error of $\hat{N}$. Define $RM(\theta;\gamma) = (M(\theta;\gamma)/M(0;\gamma) - 1)^{1/2}$; $RV(\theta;\gamma) = (V(\theta;\gamma)/M(0;\gamma) - 1)^{1/2}$; and $RB(\theta;\gamma) = (B^2(\theta;\gamma)/M(0;\gamma))^{1/2}$.

Thus, $RM(\theta;\gamma)$ is the square root of the increase in the total mean square error of $\hat{N}$ for given θ and γ relative to the root MSE of $\hat{N}$ with no matching error. $RV(\theta;\gamma)$ is the contribution of this increase due to matching variance while $RB(\theta;\gamma)$ is the contribution due to matching bias. Hence, we have $RM(\theta;\gamma)^2 = RV(\theta;\gamma)^2 + RB(\theta;\gamma)^2$. Figures 3 and 4 show these functions for two extreme values of γ, $\gamma = .5$ and 5, respectively, and for $0 \leq \theta \leq .1$. Again, the maximum value of the correlated variance, $CC*$, was used for the variance computations. Thus, the contribution of matching variance to total error is probably substantially exaggerated.

These figures indicate that for these values of θ and γ, most of the error is contributed by bias, although the contribution to variance can be non-trivial. Further, as suggested earlier for Figures 1 and 2, the matching bias dominates the total matching error whenever false negative error dominates over false positive error.

4. ESTIMATION FROM REMATCH STUDIES

Methods for estimating the components of response error in sample surveys have been well documented in the literature (see for example Hansen, Hurwitz and Pritzker 1964, Hansen, Hurwitz and Bershad 1961). The techniques for estimating the components of matching error are essentially the same. For example, to estimate the correlated component of matching variance, CC, the assignments of the clerks must be "interpenetrated." This procedure, which is described in detail in Kish (1962), randomizes the assignment of PES cases to clerks so that each clerk's assignment has the same expected number of matched persons. Then, an estimator of CC is formed by the difference between the between clerks and within clerks mean squares from the analysis of variance of clerks. For more details of this procedure, refer to Bureau of the Census (1985).

In this section, the focus is on the analysis of data from rematch studies, the most commonly used method for evaluating matching error. There are two types of rematch studies. One attempts to replicate the original match operation for a sample of cases using the same procedures, training, match rules, etc. This type of rematch has the objective of estimating SMV, the simple matching variance or, equivalently, I_M, the index of match inconsistency. The second type of rematch aims at obtaining the most correct match possible and, therefore, uses more extensive procedures, highly qualified and expert clerks, and adjudication, i.e., resolving disagreements among the original and rematch classifications by a third, expert matcher. This type of rematch as the objective of estimating the matching bias. Further, as we will see, an estimate of SMV is also possible from these data.

The (unweighted) data collected in a rematch study can be displayed as in Table 1. Assume that the rematch sample is a simple random sample of r persons from the PES. Further we may assume either the uncorrelated error model or the clerical error model of the last section for both the match and rematch. Let μ_t ($t = a, b, c, d$) denote the mean observed proportion of the cell corresponding to t in Table 1.

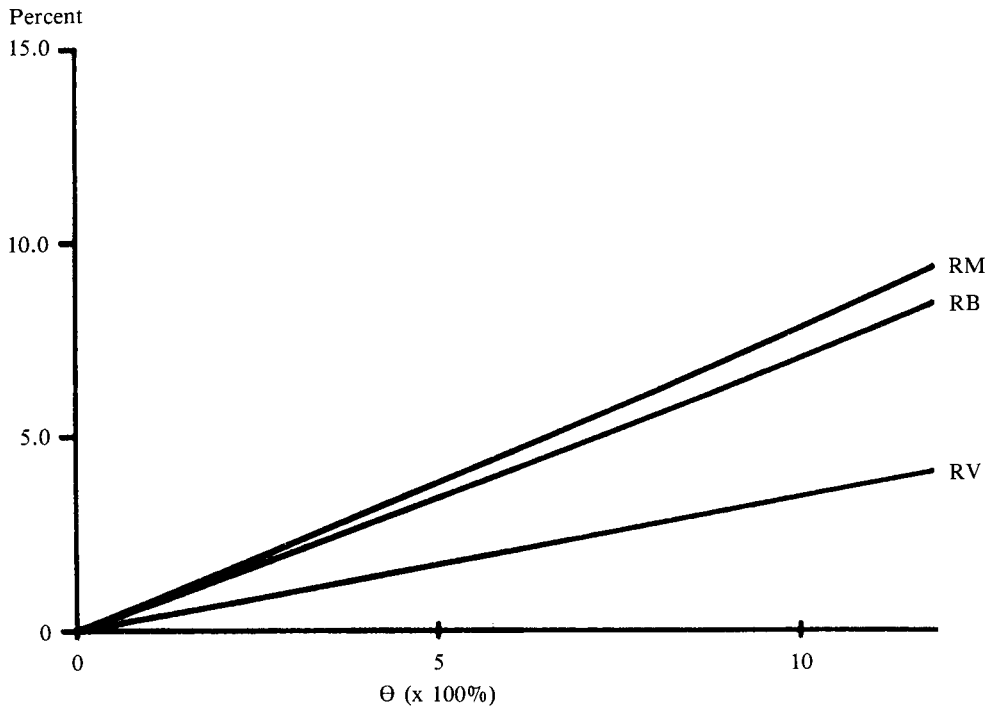

Figure 3. RM(Θ ; γ), RV(Θ ; γ), and RB(Θ ; γ), as a Function of Θ for γ = .5

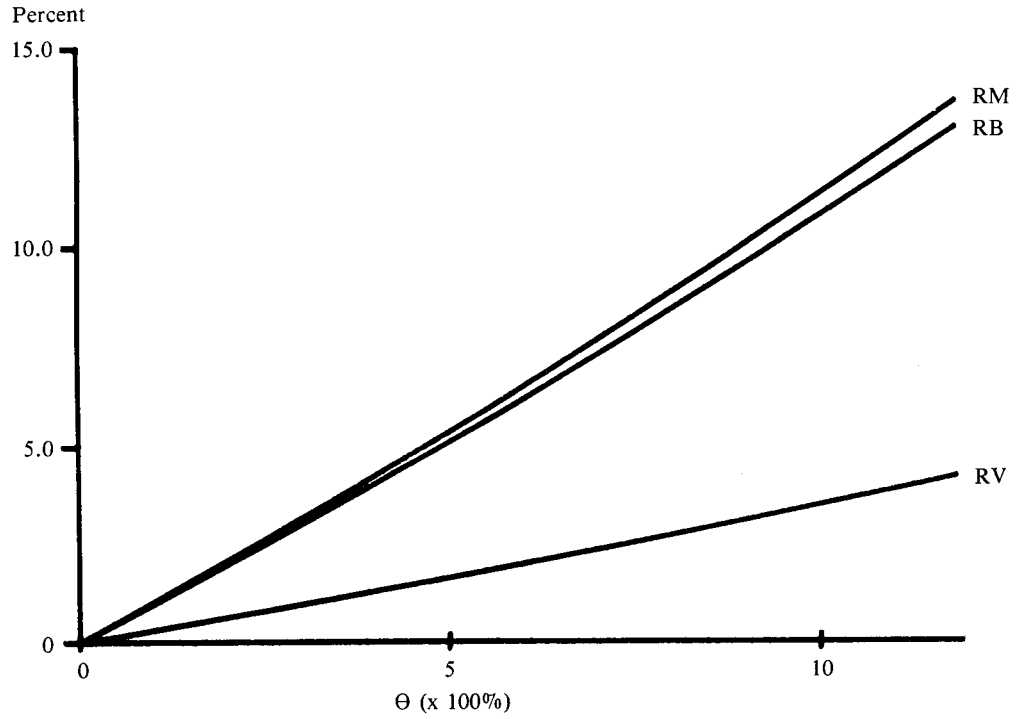

Figure 4. RM(Θ ; γ), RV(Θ ; γ), and RB(Θ ; γ), as a Function of Θ for γ = 5

Table 1

Rematch Study Data

Original Classification	Rematch Classification	
	Matched	Not Matched
Matched	a	b
Not Matched	c	d

Then

$$\mu_a = p_{11}\,(1-\theta_A)\,(1-\theta_B) + (1-p_{11})\phi_A\phi_B \tag{15}$$

$$\mu_b = p_{11}\,(1-\theta_A)\,\theta_B + (1-p_{11})\phi_A\,(1-\phi_B) \tag{16}$$

$$\mu_c = p_{11}\,\theta_A\,(1-\theta_B) + (1-p_{11})\,(1-\phi_A)\phi_B \tag{17}$$

$$\mu_d = p_{11}\,\theta_A\,\theta_B + (1-p_{11})\,(1-\phi_A)(1-\phi_B) \tag{18}$$

where the index A denotes original match and B denotes the rematch.

Define

$$\mu_A = E\left(\frac{a+b}{r}\right) = p_{11}(1-\theta_A) + (1-p_{11})\,\phi_A \tag{19}$$

and

$$\mu_B = E\left(\frac{a+c}{r}\right) = p_{11}(1-\theta_B) + (1-p_{11})\,\phi_B. \tag{20}$$

Note that μ_A and μ_B are expected values of the estimates of p_{11} based upon the original and the rematch classifications, respectively. The difference of these two estimates of p_{11}, i.e., $(b-c)/r$ is referred to as the *net difference rate* (NDR). Its expected value is

$$E(NDR) = \mu_A - \mu_B = -p_{11}(\theta_A - \theta_B) + (1-p_{11})\,(\phi_A - \phi_B). \tag{21}$$

Finally, the proportion of the r sample individuals having rematch classifications which disagree with the original match classification is $(b+c)/r$, referred to as the gross difference rate (GDR). Its expected value is

$$E(GDR) = \mu_b + \mu_c$$

$$= p_{11}\,[\theta_A\,(1-\theta_B) + (1-\theta_A)\theta_B] + (1-p_{11})[(1-\phi_A)\phi_B + \phi_A\,(1-\phi_B)]. \tag{22}$$

We shall now consider the estimation of the components of Var $(\hat{p}_{11})$ and Bias $(\hat{p}_{11})$ under three sets of assumptions for the rematch study. In the first case, we assume that the rematch study is conducted under the same general conditions as the original match so that the error parameters associated with both classifications are very nearly the same. For example, the clerks for both operations received the same training, have the same skill level, and use the same

procedures. The second case assumes that the rematch is perfect, i.e, the rematch classification may be considered the true classification. The third case falls somewhere between case 1 and 2. More extensive and improved matching procedures are used in the rematch; however, we are not willing to assume that the rematch classifications are without error. Instead we assume that fewer errors are made in the rematch than in the original match.

Case 1. Same General Conditions for the Match and Rematch

Assume that $\theta_A = \theta_B = \theta$ and $\phi_A = \phi_B = \phi$, i.e., the expected rates of misclassification are the same for both trials. Then, from (21), $E(NDR) = 0$ and no estimate of Bias $(\hat{p}_{11})$ can be computed from the data. However, from (22) and (8)

$$\tfrac{1}{2} E(GDR) = SMV \tag{23}$$

Further, an estimator of I_M in (9) is

$$\hat{I}_M = GDR / [2\hat{p}_{11}(1-\hat{p}_{11})] \tag{24}$$

where $\hat{p}_{11}$ is the PES estimator of p_{11} as defined for (2). Alternatively, an estimator of $E(\hat{p}_{11})$ can be obtained from Table 1; for example, see the estimators in (19) and (20).

Case 2. Perfect Rematch

Assume that $\theta_B = \phi_B = 0$, i.e., the rematch is conducted without misclassification error. Then, from (21),

$$\begin{aligned} E(NDR) &= -p_{11}\theta_A + (1-p_{11})\phi_A \\ &= \text{Bias } (\hat{p}_{11}). \end{aligned} \tag{25}$$

Further, the probability of false negative error, θ_A, is estimated by

$$\hat{\theta} = c/(a+c). \tag{26}$$

and, the probability of false positive error, ϕ_A, is estimated by

$$\hat{\phi} = b/(b+d). \tag{27}$$

An estimator of SMV is

$$\widehat{SMV} = \frac{1}{r}\left(\frac{ac}{a+c} + \frac{bd}{b+d}\right) \tag{28}$$

and, thus, an estimator of I_M is

$$\hat{I}_M = \widehat{SMV}/\hat{p}_{11}(1-\hat{p}_{11}) \tag{29}$$

where $\hat{p}_{11}$ is an estimator of $E(\hat{p}_{11})$ obtained either from the PES or from Table 1.

Case 3. Rematch Has Smaller Error But is Not Perfect

Assume that $0 < \theta_B < \theta_A$ and $0 < \phi_B < \phi_A$; i.e., the misclassification probabilities for the rematch are smaller than for the original match but are not zero. Then no unbiased estimator of Bias $(\hat{p}_{11})$ exists. However, $|E(NDR)|$ will be smaller than $|\text{Bias } (\hat{p}_{11})|$ if $\mu_A - p_{11}$ and $\mu_B - p_{11}$ both have the same sign; i.e., the estimator of p_{11} based on the match and the rematch data are biased in the same direction. Thus, under these conditions, $|NDR|$ is a lower bound estimator of $|\text{Bias } (\hat{p}_{11})|$.

Further, there is no unbiased estimator of SMV. However, it can be seen from (22) that

$$E(GDR) - 2SMV = p_{11}(\theta_B - \theta_A)(1 - 2\theta_A) + (1 - p_{11})(\phi_B - \phi_A)(1 - 2\phi_A).$$

Thus, whenever θ_A and ϕ_A are both less than .5, which is true in most practical applications, we have

$$E(GDR) < 2SMV$$

and $\hat{I}_M$ defined in (24) will underestimate I_M.

5. APPLICATION TO THE 1990 CENSUS

In the 1990 Census, the PES sample will consist of about 5000 "blocks" or groups of about 30 contiguous housing units and attempts will be made to match each person in every block to the census. The variables used for matching will include Name, Address, Relation to Head of Household, Sex, Birthdate, Marital Status, Race, and Hispanic Origin. The matching process will involve four separate stages as follows:

Stage 1. A computer match operation using the Fellegi and Sunter (1969) technique. Each PES person will be classified as either matched to the census, not matched, or possibly matched (i.e., requiring clerical review) by computer.

Stage 2. A first clerical review to correct any mismatches or erroneous non-matches made by the computer. In addition, a standardized set of matching rules will be applied to each possible match. Thus, each PES person will be classified as either a match, a non-match, a possible match or an unresolved case.

Stage 3. A second clerical review to reconsider, by applying greater human judgment, the classification made at the two earlier stages. The clerks for this stage, referred to as the special matching group (SMG), may also decide that for some households further field follow-up is required.

Stage 4. An "after field follow-up" review. Cases are reconsidered on the basis of any additional information obtained in the follow-up. The final classification codes are matched (enumerated), not matched (not enumerated) or unresolved (match status to be imputed in the final processing stage).

The procedures for imputing "matched" or "not matched" for unresolved cases are described in Schenker (1987). These cases which account for about 1% of the PES sample are not included in the tables which follow since the imputed match statuses of the unresolved cases were not available for this test. Nevertheless, imputation error can be an important source of matching error — one which poses special problems for the evaluation. For example, it is likely that some of the PES unresolved cases will also be unresolved in the rematch and no direct estimate of misclassification error can be computed for these cases. In the test described below, 83% of the unresolved PES cases remained unresolved in the rematch. Conversely, 41% of the cases which were unresolved in the rematch, were resolved in the PES match. If one assumes that imputations for those cases which were unresolved in the rematch are erroneous, an upper bound on the imputation error can be obtained. Likewise, a lower bound can be obtained by assuming all these imputations are correct. However, unless the proportion of imputations is very small, this "worst-case, best-case" analysis may yield bounds which are too wide to be useful.

In 1986, a pretest of these PES matching procedures was conducted in Los Angeles. A sample of about 4000 persons were matched to the Los Angeles test census and then rematched by census professional staff to evaluate matching bias. Special procedures were used in the rematch to ensure a very accurate match classification. Table 2 displays the rates of disagreement among the four stages of matching and the rematch. Note the improvement of the classifications at each higher stage indicated by the decreasing disagreement rate in the rematch column. The data also indicate that few classifications are affected in the "after follow-up" stage (.68% disagreement with stage 3). Further, the GDR for the final stage (relative to the rematch) is very low, less than 1%.

Under the assumption that the rematch process yields the true match status, Table 3 gives the estimates of θ, the probability of false negative error, and ϕ, the probability of false positive error, for each stage of matching. It appears, that for the computer match and the first level clerical match, the false nonmatch rate predominates. However, the opposite is true for the final two stages of matching.

Table 2

Comparison of Disagreement Rates for Stages of Matching (%)

	Stage 2	Stage 3	Stage 4	Rematch
Stage 1	2.9	4.4	4.7	5.5
Stage 2	0	3.3	4.0	4.8
Stage 3	3.3	0	.68	1.6
Stage 4	4.0	.68	0	.87

Table 3

Estimates of θ and ϕ for Stages of Matching

Stage of matching	Estimate of θ (x100%) (false nonmatch rate)	Estimate of ϕ (x100%) (false match rate)
1	6.2	2.3
2	5.1	3.3
3	1.5	2.1
4	.1	.3

Table 4

Results of the Rematch Study (weighted)

Original Match Classification	Rematch Classification	
	Matched	Not Matched
Matched	16690	9
Not Matched	85	2178

Table 5a

Rematch Results For Cases With Agreement On All Four Stages.

Original Match Classification	Rematch Classification	
	Matched	Not Matched
Matched	14458	0
Not Matched	64	1775

Table 5b

Rematch Results For Cases With Disagreement On at Least One Stage.

Original Match Classification	Rematch Classification	
	Matched	Not Matched
Matched	2223	9
Not Matched	21	403

Using the methodology of the previous section, we can estimate Relbias $(\hat{p}_{11})$, Relbias $(\hat{N})$, and I_M, the index of match inconsistency. Table 4 gives the results of the rematch study, weighted for the rematch sample probabilities of selection. For this table, the estimate of Relbias $(\hat{p}_{11})$ is $-.4\%$ and therefore, the estimate of Relbias $(\hat{N})$ is .4%, computed from (2) assuming a 1% coefficient of variation for $\hat{p}_{11}$ and replacing Relbias $(\hat{p}_{11})$ by its estimate. I_M is estimated to be .49% which is in the very low range. The false positive rate is $\hat{\phi} = .004$ and the false negative rate is $\hat{\theta} = .005$.

As mentioned in the second section, the probability of matching error may depend upon the completeness of the PES or census information, among other things. To indicate the extent to which match error rates vary, the rematch sample was partitioned into two subsamples. The first subsample was composed of cases which were classified as "matched" or "not matched" consistently across all stages of matching, i.e., for which all four stages agreed. The remainder of the sample made up the second subsample, i.e., cases for which at least one of the stages disagreed. This division approximates a division based upon completeness of the matching information since most of the cases having no disagreement between stages are those where information is the most complete. The weighted results are shown in tables 5a (complete cases) and 5b (incomplete cases).

For "complete" cases, the false negative rate is .44% while the false positive rate is 0. Thus, none of the cases were erroneously matched although a modest number were erroneously called nonmatches. These data may provide evidence of the greater skill of the rematch staff at finding matches for PES cases. The estimate of I_M is .39%, very low. For "incomplete" cases, the false negative rate is .93% while the false positive rate is 2.18%. The estimate of I_M is 1.1%, still quite low. However, these data indicate a much higher risk of false matches for the "incomplete" cases.

The data from this study indicates that matching error causes a small negative bias ($-.4\%$) in $\hat{N}$ which amounts to an underestimate of approximately one million persons (assuming $N = 250$ million persons). Even for the more difficult cases the bias is only $-.7\%$. It would be interesting to look at certain demographic subgroups of the population — movers, proxy respondents, and apartment dwellers — to see the extent of matching error for these domains. Unfortunately, the information that would allow this analysis is not currently available.

6. SUMMARY

The models and MSE formulas developed in this paper can be useful for evaluating the impact of matching error on estimates of census coverage error. In the context of the 1990 U.S. census matching error bias appears to be the largest and most important component of the $MSE(\hat{N})$. Preliminary studies of the magnitude of matching error bias for the 1990 Census indicate that this component is small, less than one half of one percent. This estimate does not reflect imputation error which affects about 1% of the *PES* cases. Moreover, estimates of bias depends heavily on the assumption that the rematch process yields the true match classification. More work is needed to check the validity of this assumption.

In the development of the formulas for the total mean square error of $\hat{N}$, we assumed that N_c was not prone to error. However, in actual practice, an estimate of the numbers of census spurious events (or erroneous enumerations), denote by EE, may be subtracted from N_c. Since this estimator is obtained from a match of a sample of the census units to the PES, EE is also subject to sampling error and matching error. For example, a person may be classified as an erroneous enumeration when they were correctly enumerated (false positive error), or they may be classified as correctly enumerated when they are erroneously enumerated (false negative error). The model and methodology formulated for evaluating the effect of false positive and false negative errors for x_{11} can be easily extended for the estimator of erroneous enumerations. Note that the Taylor approximation formulas for the bias and variance of $\hat{N}$, (2) and (3), will now contain terms for the bias and variance of EE.

For future research, studies of matching error correlated variance are needed to inform us of the extent to which the clerk variance contributes to the total error of $\hat{N}$. We suspect that $CC*$, the maximum effect of correlated error, substantially over estimates the impact of clerks. Research is also needed from rematch studies to identify the characteristics of persons or households prone to matching error. Perhaps then special efforts could be directed toward these cases. For this objective, the use of logistic models should be explored for predicting the probability a case is misclassified from the various characteristics of the case.

ACKNOWLEDGMENTS

This work was supported though a Joint Statistical Agreement with the U.S. Bureau of the Census. I wish to thank Aref Dejani of the Census Bureau for providing some computer support for the preparation of this paper. Thanks are also due to Bernice Garrett for typing and proof reading of the paper.

APPENDIX

Derivation of the MSE Formulas

Let U denote the population of size N to be enumerated. Let U_c denote the subset of U which is enumerated in the census. Let S denote the PES sample and S_c denote $S \cap U_c$, the set of PES persons enumerated in the census. Denote the n units in S as $u_1, \ldots, u_n$. Define the variables

$$\eta_i = 1 \text{ if } u_i \varepsilon S_c$$

$$= 0 \text{ if } u_i \notin S_c$$

and

$$y_i = 1 \text{ if } u_i \text{ classified (by the matching process) in } S_c.$$

$$= 0 \text{ if } u_i \text{ not classified in } S_c.$$

Model for Correlated Error

Assume: (1) y_i is a random variable with $P(y_i = 1 \mid \eta_i = 0) = \phi$ and $P(y_i = 0 \mid \eta_i = 1) = \theta$, and (2) y_i and y_j are independent given η_i and η_j for $i \neq j$. Let $E(\cdot \mid S)$ and $V(\cdot \mid S)$ denote conditional expectation and variance, respectively, given S. Then, $\hat{p}_{11} = \Sigma y_i / n$ and $E(\hat{p}_{11} \mid S) = (1-\theta)\tilde{p}_{11} + \phi(1-\tilde{p}_{11})$ where $\tilde{p}_{11} = \Sigma \eta_i / n$. Taking expectation with respect to S yields the result in (5).

Further, $V(y_i \mid \eta_i = 0) = \phi(1-\phi)$ and $V(y_i \mid \eta_i = 1) = \theta(1-\theta)$. Therefore, $V(\hat{p}_{11} \mid S) = \phi(1-\phi)(1-\tilde{p}_{11})/n + \theta(1-\theta)\tilde{p}_{11}/n$.

Taking expectation with respect to S yields SMV in (8).

Finally, combining $VE(\hat{p}_{11} \mid S)$ and $EV(\hat{p}_{11} \mid S)$ yields the result in (6).

Model for Clerical Error

Let (i,j) denote the j^{th} person in the i^{th} clerk's assignment. Let y_{ij} *and* η_{ij} be defined in analogy to y_i and η_i. Assume (1) — (3) for the clerical error model. Let E_2, V_2, and C_2 denote conditional expectation, variance, and covariance with respect to the clerk error distributions holding the sample of clerks fixed. Let E_1, V_1, and C_1, denote the corresponding expectation, variance and covariance with respect to the random selection of the k clerk parameter vectors, as per assumption (3), holding the sample S fixed. Then

$$E_1 E_2 (\hat{p}_{11}) = E_1 \left\{ \sum_i \left[(1-\theta_i) \frac{n_{1i}}{n} + \phi_i \frac{n_{0i}}{n} \right] \right\}$$

$$= (1 - \theta)\tilde{p}_{11} + \phi(1 - \tilde{p}_{11})$$

where $n_{1i} = \sum_j \eta_{ij}$ and $n_{0i} = \sum_j (1 - \eta_{ij})$. Hence, (4) follows upon taking expectation of (A.1) with respect to S.

Consider the variance of $\hat{p}_{11}$. We have $\mathrm{Var}(\hat{p}_{11}) = \mathrm{VE}(\hat{p}_{11} \mid S) + \mathrm{EV}(\hat{p}_{11} \mid S)$ where $\mathrm{E}(\hat{p}_{11} \mid S)$ is given by (A.1). Further $n^2 V(\hat{p}_{11} \mid S) = \sum_i \sum_i V(y_{ij} \mid S) + \sum_i \sum_{j \neq j'} \mathrm{Cov}(y_{ij}, y_{ij'} \mid S)$

where $V(y_{ij}|\, S) = V_2(y_{ij}) + V_1E_2(y_{ij})$ and $\text{Cov}(y_{ij},y_{ij'}|\, S) = C_1\,[E_2(y_{ij}), E_2(y_{ij'})]$, the term $E_1C_2(y_{ij},y_{ij'})$ being zero. Since $E_2(y_{ij}) = \phi_i$, for $\eta_{ij} = 0$, and $E(y_{ij}) = 1-\theta_i$ for $\eta_{ij} = 1$, we have $V_1E_2(y_{ij}) = \sigma_\phi^2$, if $\eta_{ij} = 0$, and $= \sigma_\theta^2$ if $\eta_{ij} = 1$. Further $V_2(y_{ij}) = \phi_i(1-\phi_i)$ for $\eta_{ij} = 0$ and $V_2(y_{ij}) = \theta_i(1-\theta_i)$ for $\eta_{ij} = 1$. Thus

$$E_1V_2(y_{ij}) = \phi(1-\phi) - \sigma_\phi^2 \text{ if } \eta_{ij} = 0$$

$$= \theta(1-\theta) - \sigma_\theta^2 \text{ if } \eta_{ij} = 1$$

Similarly, it can be shown that, for $j \neq j'$,

$$C_1\,\{E_2(y_{ij}), E_2(y_{ij'})\} = \sigma_\theta^2 \text{ if } (\eta_{ij}, \eta_{ij}') = (1,1)$$

$$= -\sigma_{\phi\theta} \text{ if } (\eta_{ij}, \eta_{ij}') = (1,0)$$

$$= \sigma_\phi^2 \text{ if } (\eta_{ij}, \eta_{ij}') = (0,0).$$

Therefore,

$$V(\hat{p}_{11}|\, S) = (\Sigma m_i^2 - n)\,/\,n^2\, CC + SMV/n. \tag{A.2}$$

Finally, combining (A.1) and (A.2) in the identity

$$V(\hat{p}_{11}) = VE(\hat{p}_{11}|\, S) + EV(\hat{p}_{11}|\, S), \text{ we have}$$

$$V(\hat{p}_{11}) = 1\,/\,n(SV+SMV) + (\Sigma m_i^2 - n)\,/\,n^2\, CC. \tag{A.3}$$

If we further assume that $m_i = m$ for all i we obtain the form in (11).

REFERENCES

BIEMER, P. P., and STOKES, S. L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of American Statistical Association* 80, 158-166.

Bureau of the Census (1985). *Evaluating censuses of population and housing*, Statistical Training Document. United States Bureau of the Census, Washington, D.C.

BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C., and POLLOCK, K. H. (1987). *Design and Analysis Methods for Fish Survival Experiments Based on Release — Recapture.* American Fisheries Society Monograph 5.

FELLEGI, I. P., and SUNTER, A. B. (1969). A Theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HANSEN, M. H., HURWITZ, W. N., and BERSHAD, M. A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

HANSEN, M. H., HURWITZ, W. N., and PRITZKER, L. (1964). The Estimation and interpretation of cross differences and the simple response variance. In *Contributions to Statistics* (Ed. C. R. Rao), Oxford: Pergamon Press, 111-136.

HARTLEY, H. O., and RAO, J. N. K. Estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement*, (Ed. N.K. Namboodiri), New York: Academic Press.

JOHNSON, N. L., and KOTZ S. (1969). *Continuous Univariate Distributions II.* Boston: Houghton Mifflen.

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

MARKS, E. S., SELTZER, W., and KROTKI, K. J. (1974). *Population Growth Estimation*. New York: Population Council.

RICHER, W. E. (1958). *Handbook of Computations for Biological Statistics of Fish Populations*. Fisheries Research Board of Canada. Ottawa: Queen's Printer and Controller of Stationery.

SEBER, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (3rd. ed.). New York: MacMillian.

SEKAR, C. C., and DEMING, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

SELTZER, W., and ADLAKTA A. (1974). On the effect of errors in the application of the Chandrasekaran — Deming techniques, Reprint 14. Laboratory for Population Statistics, University of North Carolina.

SCHENKER, N. (1987). Report on missing data in the 1986 test of adjustment related operations. Survey Research Division Report Series, Bureau of the Census, RR-87/09.

SCHEUREN, F., and OH, H. L. (1985). Fiddling around with nonmatches and mismatches. *Proceedings of the Workshop on Exact Matching Methodologies*. Arlington, Virginia.

WOLTER, K. M. (1983). Affidavit, Mario Cuomo *et al. vs.* Malcolm Baldrige *et al.* U. S. District Court, Southern District of New York, 80 Civ. 4550.