Survey Methodology, June 1996
Vol. 22, No. 1, pp. 55–64
Statistics Canada

Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Errors

YE DING and STEPHEN E. FIENBERG[1]

ABSTRACT

The multiple capture-recapture census is reconsidered by relaxing the traditional perfect matching assumption. We propose matching error models to characterize error-prone matching mechanisms. The observed data take the form of an incomplete 2^k contingency table with one missing cell and follow a multinomial distribution. We develop a procedure for the estimation of the population size. Our approach applies to both standard log-linear models for contingency tables and log-linear models for heterogeneity of catchability. We illustrate the method and estimation using a 1988 dress rehearsal study for the 1990 census conducted by the U.S. Bureau of the Census.

KEY WORDS: Capture-recapture census; Estimates for total population size; Log-linear models; Matching errors; Multiple recapture census.

1. INTRODUCTION

The multiple recapture census technique has been used in many fields to estimate the size of a closed population. Cormack (1968) and Seber (1982) give excellent reviews of many techniques used. Here we consider a sequence of samples, $s_1, \ldots, s_k$, where the members of i-th sample are uniquely labeled, for example, by tagging or marking, and then returned to the population (Darroch 1958). Usual multiple recapture census methods make the following assumptions.

(1) **Perfect matching**. Individuals in one list (information source, sample) can be matched with those in another list without error. In other words, there are no misclassification errors with respect to determining whether a particular individual has been recorded by both information sources or only one of them.

(2) **Independence**. The lists are independent of one another, that is, the probability of an individual being included in one list does not depend on whether the individual was included in previous lists.

(3) **Homogeneity (Equal Catchability)**. All individuals in the population under study have equal probabilities of being observed (captured) in any list (sample).

(4) **Closure**. The population in question is "closed", so that there are no changes due to birth, death, emigration, or immigration during the period when the sampling takes place.

Darroch (1958) examined the multiple recapture census under these four assumptions. Fienberg (1972) adopted a log-linear model approach to allow for statistical dependence of specific types among samples, thereby dropping the independence assumption. Darroch, Fienberg, Glonek and Junker (1993) developed an extended log-linear model approach that allows for individual-level heterogeneity as well as dependence, but it requires at least three samples, *i.e.*, $k = 3$. In the context of the two-sample census approach used by U.S. Bureau of Census for census coverage evaluation, matching problems due to unavoidable mismatches and erroneous nonmatches have been explored by several authors. For example, Ding and Fienberg (1994) considered modeling matching errors in the two-sample census and developed systematic procedure for the estimation of population totals. The inclusion of a third sample, *e.g.*, drawn from the administrative records, in modeling and estimation of census coverage has been considered by the U.S. Bureau of Census in the past and remains an option to augment and evaluate the dual system approach. In this paper, we consider matching error models for the multiple sample census problem, allowing for both dependence and heterogeneity.

Here we view the observations from a multiple recapture census data as falling into a 2^k cross-classification, with absence or presence on the i-th sample defining the category for the i-th dimension. In this cross-classification, the cell corresponding to absence for all k samples is missing. The objective is to estimate the number of individuals in the population who are not observed, which corresponds to the missing cell in the 2^k incomplete contingency table. In Section 2, we investigate the effects of matching errors on the observed 2^k incomplete table. In Section 3, some models for matching errors are proposed to characterize an error-prone matching process. Based on these models and assumptions (3) and (4), we develop a procedure using log-linear model formulation for the estimation of the population size. In Section 5, we use the proposed methods to analyze data from 1988 Dress Rehearsal Census conducted by the U.S. Bureau of Census.

[1] Ye Ding, 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, Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

2. MATCHING ERRORS IN MULTIPLE SAMPLE CENSUS

We begin by classifying matching errors into two broad categories, mismatches and erroneous nonmatches. To understand the nature of matching errors in multiple-sample census, we review the case of a three-sample census. Suppose that there are no missing data or errors in recording the information for any individual in the population and one takes three samples from the population, s_1, s_2, and s_3. For instance, suppose that, in sample s_1, individuals 1, 3, 4 and 7 are seen, individuals 3, 4, and 8 are seen in s_2, and individuals 4, 9, and 10 in s_3. In vector notation, we can represent this as $s_1 = (1, 3, 4, 7)$, $s_2 = (3, 4, 8)$ and $s_3 = (4, 9, 10)$. Matching errors are not present provided that there is complete and correct information available. We thus have the following incomplete 2^3 table corresponding to these three samples:

Table 1

Original Table without Matching Errors

	s_1			
	Present		Absent	
	s_2		s_2	
s_3	Present	Absent	Present	Absent
Present	1	0	0	2
Absent	1	2	1	-

Suppose further that, because of missing data or incorrect information, we actually observe

$$s_1 = (1, 3, 4, 7), \quad s_2 = (3^*, 4^*, 8), \quad s_3 = (4, 9, 10),$$

where 3* and 4* are individuals 3 and 4 but with incorrect information leading to two erroneous nonmatches when the samples are matched. Assuming no erroneous matches, we then observe the incomplete 2^3 table:

Table 2

Observed Table with Matching Errors

	s_1			
	Present		Absent	
	s_2		s_2	
s_3	Present	Absent	Present	Absent
Present	0	1	0	2
Absent	0	3	3	-

The effects of matching errors are obvious from a comparison of Table 1 and 2:

(i) The number of observations may increase for some cells while decreasing for the others, and as a consequence, the marginal totals and especially the total number of different individuals observed in the three samples may change, subject to the constraint that the total number of observations in each sample, x_{1++}, x_{+1+}, and x_{++1}, remain the same. Changes in the total number of different individuals in all samples make our problem distinct from the usual misclassification problem in the analysis of categorical data, in which the possibility of making mistakes in classifying individuals into respective categories is considered. (*e.g.*, see Chen 1979).

(ii) In parallel, there may be changes in some cell probabilities subject to the constraint that the probability of being captured in a sample, p_{1++}, p_{+1+}, and p_{1++}, is unchanged.

Because of the complexity of matching errors in the three-sample case, we need some special terminology for descriptive convenience. We say that an individual is at state 1 with respect to sample s_1 if the individual is observed in s_1 and at state 0 if not. We use a triple (i,j,k), $0 \leq i, j, k \leq 1$, to denote an individual at state i, j, and k with respect to s_1, s_2 and s_3, respectively. For instance, (1,0,0) is an individual observed only in s_1, and (1,1,1) is an individual captured in three samples. We define the level of an individual (i,j,k) as $i + j + k$, *i.e.*, the number of samples in which the individual is included. There are four different levels, 0, 1, 2 and 3. An individual has level 0 if and only if he/she is not captured by any sample, and has level 3 if he/she is in three samples. For a (1,1,0) individual, if the correct match is not made according to the matching rule, this individual decomposes into "two different" individuals, a (1,0,0) and a (0,1,0), assuming no erroneous matches. On the other hand, a (1,0,0) individual matched incorrectly with a (0,1,0) will produce a single observed (1,1,0) individual. For convenience, we call such a decomposition or combination a *transition*. Then transitions can only go from level 3 or 2 to the same (if there is no matching error) or lower levels in the absence of erroneous matches. More specifically, a (1,1,1) person may make a transition into one of 5 possible sets of individuals

$$\{(1,1,1)\}, \quad \{(1,0,0), (0,1,1)\}, \quad \{(0,1,0), (1,0,1)\}$$

$$\{(0,0,1), (1,1,0)\}, \quad \{(1,0,0), (0,1,0), (0,0,1)\}.$$

For level 2 individuals, (1,1,0) can decompose into $\{(1,0,0),(0,1,0)\}$ or stay at $\{(1,1,0)\}$, and similarly for $\{(0,1,1)\}$ and $\{(1,0,1)\}$. From above discussions, we summarize the effect of matching errors by the following diagram:

$$\boxed{\text{Table 1}} \longrightarrow \{\text{Matching Process}\} \longrightarrow \boxed{\text{Table 2}}$$

where Table 1 is the original 2^k incomplete table with no matching errors and Table 2 is the observed 2^k incomplete table in the presence of matching errors. Henceforth, we denote the cell probabilities and expected cell counts associated with Table 1 by $\{r_{ijk}\}$ and $\{l_{ijk}\}$ and those of Table 2 by $\{p_{ijk}\}$, $\{m_{ijk}\}$, for $1 \le i,j,k \le 2$.

3. SOME MODELS FOR MATCHING ERRORS

We now propose models to describe the matching errors, each of which allows us to formulate the reallocation of cell probabilities and expected cell counts associated with Table 1.

Model (1). In addition to the homogeneity and closure assumptions in §1, we assume that: (i) There are no erroneous matches in the matching process; (ii) Any individual will stay at his original state with probability θ, and transition to any of a possible set of individuals with probability $(1-\theta)/(m-1)$, where m is the number of all possible sets of individuals to which the individual may transition. For example, for a (1,1,1) person discussed late in last section, $m = 5$.

Under this model, for the three-sample census, we can express the probabilities for the table with matching errors, $\{p_{ijk}\}$, in terms of probabilities of the table with no matching errors, $\{r_{ijk}\}$:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1-\theta}{4} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1-\theta}{4} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1-\theta}{4} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{211} + (1-\theta) r_{121} + r_{221}.$$

Let

$$\vec{p} = (p_{111}, p_{112}, p_{121}, p_{211}, p_{122}, p_{212}, p_{221})^T,$$

and

$$\vec{r} = (r_{111}, r_{112}, r_{121}, r_{211}, r_{122}, r_{212}, r_{221})^T,$$

then

$$\vec{p} = M_1 \times \vec{r}. \qquad (1)$$

Here M_1 is a 7 by 7 matrix determined by the above seven equations derived under Model (1).It is straightforward to verify that the probability of catching any individual in each sample is fixed, *i.e.*, $p_{1++} = r_{1++} = p_1$, $p_{+1+} = r_{+1+} = p_2, p_{++1} = r_{++1} = p_3$. This must be the case because the sample capture probabilities do not depend on how the matching mechanism operates.

We can easily generalize this formulation to handle the k-sample case; however, the algebra involved is quite messy for large k. We can simplify this model by requiring that the transitions can go downwards by at most one level, thus yielding Model (2):

Model (2). In addition to the homogeneity and closure assumptions in §1, we assume that: (i) there are no erroneous matches in the matching process; (ii) a transitions can only go downwards by at most one level; (iii) any individual will stay at his original state with probability θ, and transition to any of a possible set of individuals with probability $(1-\theta)/(m'-1)$, where m' is the number of sets of individuals to which transitions are possible and allowed.

We first consider the three-sample case. A (1,1,1) individual can decompose into three individuals, *i.e.*, $(1,1,1) \mapsto \{(1,0,0), (0,1,0), (0,0,1)\}$ (we use "$\mapsto$" to denote for decomposition), if three presumed matches are not made. Assumption (ii) of Model (2) assumes that this triple error has negligible probability when compared with the transition in which only one of the matches is not made so that $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$, or $(1,1,1) \mapsto \{(1,0,1),(0,1,0)\}$, or $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$.

For three sample case, the parametric model for expressing $\{p_{ijk}\}$ in terms of $\{r_{ijk}\}$ is:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1-\theta}{3} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1-\theta}{3} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1-\theta}{3} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{211} + (1-\theta) r_{121} + r_{221}.$$

Then

$$\vec{p} = M_2 \times \vec{r}, \qquad (2)$$

where M_2 is a 7 by 7 matrix determined by the above seven equations derived under Model (2). Again, the capture probabilities are unchanged, *i.e.*, $p_{1++} = r_{1++} = p_1$, $p_{+1+} = r_{+1+} = p_2$, $p_{++1} = r_{++1} = p_3$.

For the k-sample problem, let $p_{\bar{1}}$ be the probability of being captured in all samples, *i.e.*, $p_{\bar{1}} = p_{111\ldots1}$, and let $p_{\bar{1},\bar{2}(h_1,h_2)}$ be the cell probability corresponding to absence in the h_1-th, and h_2-th sample and presence in the others, *etc*. Under Model (2), we have $p_{\bar{1}} = \theta r_{\bar{1}}$. For $i \le k-2$, the probability of being missed by the h_1-th, h_2-th, ..., and h_i-th sample and captured by the others is

$$p_{\bar{1},\bar{2}(h_1,h_2,\ldots,h_i)} = \theta r_{\bar{1},\bar{2}(h_1,h_2,\ldots,h_i)} + \frac{1-\theta}{k-i+1} \sum_{j=1}^{i} r_{\bar{1},\bar{2}(\{h_1,h_2,\ldots,h_i\}\backslash h_j)}.$$

For $i = k-1$, the individual is included in only one sample. For example, the probability of being captured only by the first sample is

$$p_{1,\bar{2}} = r_{1,\bar{2}} + (1-\theta) \sum_{h \neq 1} r_{1,1(h),\bar{2}} + \frac{(1-\theta)}{3} \sum_{h_1,h_2 \ge 2} r_{1,1(h_1,h_2),\bar{2}} + \sum_{j=3}^{k-1} \sum_{h_1,h_2,\ldots,h_j \ge 2} \frac{(1-\theta)}{(j+1)} r_{1,1(h_1,h_2,\ldots,h_j),\bar{2}},$$

where $r_{1,1(h_1,h_2,\ldots,h_j),\bar{2}}$ is the cell probability in the original table which corresponds to presence in the first, h_1-th, h_2-th, ..., h_j-th sample and absence in the others. By symmetry, we can write down the expression for $p_{1(h),\bar{2}}$, the probability of being observed in the h-th sample only and missed in all others.

We can refine Model (2) by assuming unequal matching rates. For example, we consider two decompositions: $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$ and $(1,1,0) \mapsto \{(0,1,0),(1,0,0)\}$. It is common for both cases that one presumed match is not made. They differ in that one has two sources of information for that match while the other has only one. It is reasonable to assume different matching error probabilities for the two cases instead of a common one as proposed in Model (2). This leads to:

Model (3). In addition to (i) and (iii) in Model (2), we assume

$$(1,1,1) \mapsto \begin{cases} (1,1,1) & \text{with probability } \alpha_1 \\ \{(1,1,0),(0,0,1)\} & \text{with probability } (1-\alpha_1)/3 \\ \{(0,1,1),(1,0,0)\} & \text{with probability } (1-\alpha_1)/3 \\ \{(1,0,1),(0,1,0)\} & \text{with probability } (1-\alpha_1)/3 \end{cases}$$

$$(1,1,0) \mapsto \begin{cases} (1,1,0) & \text{with probability } \alpha_2 \\ \{(0,1,0),(1,0,0)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

$$(1,0,1) \mapsto \begin{cases} (1,0,1) & \text{with probability } \alpha_2 \\ \{(1,0,0),(0,0,1)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

$$(0,1,1) \mapsto \begin{cases} (0,1,1) & \text{with probability } \alpha_2 \\ \{(0,1,0),(0,0,1)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

and (1,0,0), (0,1,0), (0,0,1) stay the same with probability one.

Under this model, we can express the cell probability $\{p_{ijk}\}$ in Table 2 in terms of α_1, α_2 and the cell probabilities of Table 1, $\{r_{ijk}\}$. To do this, we need to consider all possible transitions that produce an individual that falls into the (i,j,k) cell in Table 2. For example, we consider an observed (1,0,0) individual. This person falls into cell (1,2,2) of Table 2. Let F be the event that an observed individual has a (1,0,0) status. Let E_{ijk} be the event that an individual falls into (i,j,k) cell in Table 1. Then

$$F = \bigcup_{\{i,j,k\}} (E_{ijk} \cap F).$$

According to Model (3), there are only four possible transitions as follows that can make F happen:

$$(1,1,1) \mapsto \{(1,0,0),(0,1,1)\},$$

$$(1,1,0) \mapsto \{(1,0,0),(0,1,0)\},$$

$$(1,0,1) \mapsto \{(1,0,0),(0,0,1)\},$$

$$(1,0,0) \mapsto \{(1,0,0)\}.$$

Therefore

$$F =$$

$$(E_{111} \cap F) \cup (E_{112} \cap F) \cup (E_{121} \cap F) \cup (E_{122} \cap F).$$

By the definitions of cell probabilities of the two tables, $p(F) = p_{122}$, and $p(E_{ijk}) = r_{ijk}$. By the assumptions in Model (3), $p(F \mid E_{111}) = (1 - \alpha_1)/3, p(F \mid E_{112}) = p(F \mid E_{121}) = \alpha_2$, and $p(F \mid E_{122}) = 1$.

Since $E_{111} \cap F$, $E_{112} \cap F$, $E_{121} \cap F$ and $E_{122} \cap F$ are four mutually exclusive possibilities that F can happen, thus

$$p_{122} = p(E_{111} \cap F) + p(E_{112} \cap F) + p(E_{121} \cap F) + p(E_{122} \cap F)$$

$$= p(F \mid E_{111}) \cdot p(E_{111}) + p(F \mid E_{112}) \cdot p(E_{112}) + p(F \mid E_{121}) \cdot p(E_{121}) + p(F \mid E_{122}) \cdot p(E_{122})$$

$$= \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122}.$$

In the same manner, we can derive the expressions of other cell probabilities of Table 2 to get

$$p_{111} = \alpha_1 r_{111},$$

$$p_{112} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{112},$$

$$p_{121} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{121},$$

$$p_{211} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{211},$$

$$p_{122} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122},$$

$$p_{212} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{211} + r_{212},$$

$$p_{221} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{211} + (1 - \alpha_2) r_{121} + r_{221}.$$

Then

$$\vec{p} = M_3 \times \vec{r}, \qquad (3)$$

where M_3 is a 7 by 7 matrix determined by the above seven equations derived under Model (3).

For $\alpha_1 = \alpha_2 = \theta$, we get the same formulation as under Model (2). For the special case with $\alpha_1 = \alpha_2 = 1$, $p_{ijk} = r_{ijk}$, reducing to the traditional problem. Again, the capture probabilities remain the same, *i.e.*, $p_{1++} = r_{1++}, p_{+1+} = r_{+1+}, p_{++1} = r_{++1}$.

4. ESTIMATING THE SIZE OF THE POPULATION

4.1 Log-linear Model Formulation

For purposes of exposition, we confine our attention to the three-sample census case, although extensions to the k-sample census for $k > 3$ are straightforward. As before, let l_{ijk} and m_{ijk} be expected cell counts for Table 1 and Table 2 respectively. The relationship between the cell probabilities and the expected cell counts is $l_{ijk} = r_{ijk}N$, and $m_{ijk} = p_{ijk}N$. Let

$$\vec{m} = (m_{111}, m_{112}, m_{121}, m_{211}, m_{122}, m_{212}, m_{221})^T,$$

and

$$\vec{l} = (l_{111}, l_{112}, l_{121}, l_{211}, l_{122}, l_{212}, l_{221})^T.$$

Since for each of the models we have proposed in the last section, there is a matrix M with entries depending on the matching probability parameters in the chosen model such that $\vec{p} = M \times \vec{r}$, multiplying through by N gives

$$\vec{m} = M \times \vec{l}. \qquad (4)$$

For any log-linear model specified for Table 1, it is straightforward to obtain the parameterization for m_{ijk}. For example, for any of the models suggested in Fienberg (1972), we can write the expected counts in terms of functions of u-term parameters:

$$l_{ijk} =$$

$$g_{ijk}(u, u_1(i), u_2(j), u_3(k), u_{12}(ij), u_{13}(ik), u_{23}(jk)), \qquad (5)$$

and then obtain the parameterization of $\{m_{ijk}, (ijk) \neq (222)\}$ from (4).

4.2 Estimating the Size of the Population

We now consider the matching rates in our various models as known. To obtain the estimate of the population size, we proceed as follows. First, following Sanathanan (1972), we compute the maximum likelihood estimates of u-term parameters from l_c, the conditional likelihood associated with Table 2 given n,

$$l_c = n! \prod_{\{(ijk) \neq (222)\}} \frac{(q_{ijk})^{x_{ijk}}}{x_{ijk}!},$$

where $n = \sum_{\{(ijk)\neq(222)\}} x_{ijk}$, and $q_{ijk} = m_{ijk}/n$. Sanathanan (1972) shows that, under suitable regularity conditions, the conditional maximum likelihood estimates and the unconditional ones are both consistent and have the same asymptotic normal distribution. If we remove redundant u-term parameters using the constraints associated with the specified log-linear model for Table 1, then the problem is to find the maximum of l_c subject to the following single constraint:

$$\sum_{\{(ijk)\neq(222)\}} m_{ijk} = n.$$

Numerically, this is a nonlinearly constrained optimization problem. Rao (1957) studied regularity conditions under which there exist unique maximum likelihood estimates of the parameters in a multinomial distribution. His conditions are satisfied by the parameterization of $\{q_{ijk}\}$. Once the conditional maximum likelihood estimates of the u-term parameters are obtained, we use the loglinear model specified for Table 1 to compute the conditional maximum likelihood estimates of $\{l_{ijk}\}$, the expected cell counts of Table 1 including the expected count of the missing cell. Then our estimate of N is

$$\hat{N} = \sum_{\{ijk\}} \hat{l}_{ijk}.$$

In the case of no matching errors, with $\alpha_1 = \alpha_2 = 1$ in Model (3), $m_{ijk} = l_{ijk}$. Thus

$$\hat{N} = n + \hat{m}_{222},$$

i.e., we get back to the estimation method for the traditional multiple recapture census problem developed by Fienberg (1972) when the log-linear models in Fienberg (1972) are considered.

As we have discussed earlier, a log-linear model is specified for Table 1 and the observations are viewed as falling into Table 2, whose parametric model of the expected cell counts is specified by the log-linear model and a chosen model for matching errors. To assess the appropriateness of a log-linear model specified for Table 1, we can apply the usual Pearson and likelihood ratio goodness-of-fit tests, X^2 and G^2, discussed in Fienberg (1972), to Table 2. Each statistic has an asymptotic χ^2 distribution under the null hypothesis that the model fits, with degrees of freedom equal to $2^k - 1 -$ (number of independent parameters in the model).

5. ANALYSIS OF 1988 ST. LOUIS DRESS REHEARSAL CENSUS DATA

Dual System Estimation (DSE), based on the standard two-sample census, has been employed by U.S. Bureau of Census for census coverage evaluation since 1950. In 1988, the Census Bureau conducted a Dress Rehearsal Census for the 1990 decennial census at three sites: St. Louis, Missouri; Columbia, Missouri; and western Washington State. Zaslavsky and Wolfgang (1993) present data for a population subgroup from the Post Enumeration Survey (PES) in the dress rehearsal census in St. Louis which focuses on urban Black male adults who are believed to be underestimated by dual system methods. The resulting data consists of three sources: the C-sample is the census itself; the P-sample was compiled from the PES; a third source of information was the Administrative List Supplement (ALS), compiled from pre-census administrative records of state and federal government agencies, encompassing Employment Security, driver's license, Internal Revenue Service, Selective Service, and Veteran's Administrative records. The C-sample and P-sample provide data for the implementation of the usual DSE or capture recapture approach. The ALS data can be combined with the Census and the P-sample for analysis from a three-sample perspective, though it was originally intended to improve the coverage of the P-sample. In Table 3, we present three-sample data for PES sampling stratum 11 in St. Louis obtained by collapsing the original data in Table 1 of Zaslavsky and Wolfgang (1993) over four poststrata defined by owners/renters × age 20-29, 30-44.

Table 3
Three-Sample Data for Stratum 11, St. Louis

	Census			
ALS	Present *P*-sample		Absent *P*-sample	
	Present	Absent	Present	Absent
Present	300	51	53	180
Absent	187	166	76	–

Such triple-system data can be analyzed with the matching error Model (2) and data from a separate Matching Error Study (MES, or rematch study) associated with the same sampling poststratum. The MES is one of the operations conducted by the Census Bureau to evaluate the PES, and 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. In the discussion of the Matching Error Study done in a 1986 test census in Los Angeles, 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."

Table 4

St. Louis Rematch Study: *P*-sample
Source: Mulry, Dajani and Biemer (1989)

Original Match Classification	Rematch Classification			
	Matched	Not Matched	Un-resolved	Total
Matched	2,667	7	8	2,682
Not matched	9	427	30	466
Unresolved	0	7	20	27
Total	2,676	441	58	3,175

The data from the MES thus provides a basis for estimating error rates in the original matching process. Mulry, Dajani and Biemer (1989) report the MES operation for the 1988 Dress Rehearsal and rematch data for all three test sites, and in Table 4, we reproduce those data relevant for our purposes.

Let α be the matching rate between the *C*-sample and the *P*-sample, and $\gamma = 1 - \alpha$ be the nonmatch error rate. We assume no errors in the rematch. Then from the data in Table 4, we can estimate α by $\hat{\alpha} = 2667/(2667 + 9) = 99.6637\%$, and γ by $\hat{\gamma} = 1 - \hat{\alpha} = .3363\%$. The parameter θ is a three-sample matching rate for the *C*-sample, *P*-sample and the ALS. It takes two matches, say, one between the *C*-sample and the *P*-sample, and the other one between the *P*-sample and the ALS, in order to reach a correct (1,1,1) three-sample classification. In the absence of evaluation of the match between the census and the ALS, we assume that these two matches are independent of each other and that the matching rate for the *P*-sample and ALS is the same for the *C*-sample and the *P*-sample. Thus we can use $\theta = \alpha^2$, and $\hat{\theta} = \hat{\alpha}^2 = 99.3285\%$. Based on other qualitative information, this seems to be unreasonably high match rate, and the match error rate for the census and the ALS is probably higher than the match error rate between the census and the *P*-sample. In the absence of better quantitative information, however, we proceed to use it in the calculations that follow.

Table 5

Estimates Under Various Models

Log-linear Model	Usual MLE		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
[C] [P] [A]	1091.48 (11.24)	248.31 (3)	1083.58 (10.93)	244.56 (3)
[CP] [A]	1204.14 (23.31)	90.60 (2)	1194.73 (22.86)	87.30 (2)
[PA] [C]	1108.34 (13.77)	247.93 (2)	1100.03 (13.40)	244.53 (2)
[CA] [P]	1068.87 (10.47)	230.66 (2)	1061.09 (10.10)	226.42 (2)
[CP] [CA]	1271.11 (52.55)	87.16 (1)	1256.77 (50.97)	84.37 (1)
[CP] [PA]	1598.88 (106.26)	17.55 (1)	1585.03 (104.93)	15.88 (1)
[CA] [PA]	1080.47 (13.38)	230.43 (1)	1072.19 (12.88)	226.44 (1)
[CP] [CA] [PA]	2360.82 (363.25)	– (0)	2309.55 (352.36)	– (0)

Table 5 gives the estimates of the population size for various log-linear models with estimates of standard errors and goodness-of-fit statistics. Standard errors are computed with the delta method as discussed in Fienberg (1972). The assumption of independence between the census and the *P*-sample has been questioned for the use of the DSE. The dual system method has limited capacity to test this assumption and to adjust for potential dependency, while both can be handled through log-linear models for three or more samples. There are four models listed in Table 5 that assume independence between the census and the *P*-sample: the independence model [C] [P] [A], [PA] [C], [CA] [P], and [CA] [PA]. All of them fit the data poorly. The three models with the interaction term for the census and the *P*-sample, [CP] [A], [CP] [CA], and [CP] [PA] fit the data much better. With the addition of an interaction term linking the census and the ALS, model [CP] [CA] fits only slightly better than [CP] [A], indicating that the census and the *P*-sample are together nearly independent from the ALS. The model [CP] [PA] fits the data the best, suggesting that the usual independence assumption for the DSE is invalid and that there is dependence between the *P*-sample and the ALS. For all seven non-saturated log-linear models, we obtain better fits under matching error Model (2), though only slightly so, due to the high match rate for the data from the 1988 U.S. Census Dress Rehearsal. For the [CP] [PA] model, there is a .8738% difference in the estimate of *N* associated with the nonmatch rate of .3363%. If the nonmatch rate had been 10%, *i.e.*, a 90% match rate, and assuming that the difference in the estimate of *N* is approximately linear in the nonmatch rate, there would have been a 26% difference between the usual maximum likelihood estimate of *N* and our estimate.

Table 6

Dual-System Data for Stratum 11, St. Louis

P-sample	Census		
	Present	Absent	Total
Present	487	129	616
Absent	217	–	
Total	704		

Table 6 presents the usual dual system data for stratum 11, St. Louis. The number of people in both the census and the *P*-sample is $y_{11} = 300$, the number of those in the census only is $y_{12} = 217$, and number in the *P*-sample only is $y_{21} = 129$. The total census count is $y_{1+} = y_{11} + y_{12} = 704$, the total *P*-sample count is $y_{+1} = y_{11} + y_{21} = 616$, the dual system estimate is $\widehat{\text{DSE}} = y_{1+}y_{+1}/y_{11} = 893$ (p. 232, Bishop, Fienberg and Holland 1975), and the estimated variance of $\widehat{\text{DSE}}$ is $\text{Var}(\widehat{\text{DSE}}) = y_{1+}y_{+1}y_{12}y_{21}/y_{11}^3 = 105.4$ (p. 233, Bishop *et al.* 1975). The standard error is $\text{SE}(\widehat{\text{DSE}}) = 10.27$.

The census undercount for the population estimate $\widehat{\text{DSE}}$ is $(\widehat{\text{DSE}} - y_{1+})/\widehat{\text{DSE}} \times 100\% = 21.17\%$. For our best fitting model, the census undercount is $(\hat{N} - y_{1+})/\hat{N} = 55.97\%$ for the estimate $\hat{N} = 1599$ assuming no matching error and 55.58% for $\hat{N} = 1585$ from matching error Model (2). Thus there is a $55.97\% - 55.58\% = 0.39\%$ upward bias by ignoring matching errors. This is quite close to the figure of 0.37% computed in Ding and Fienberg (1994) for the 1986 Los Angeles test census data using a two-sample match rate of 99.4734%, as compared to 99.6637% here for the St. Louis data. Our estimates show that the urban Black male adults targeted in the St. Louis Dress Rehearsal were heavily undercounted by the census, and that the undercount is severely underestimated by the usual dual-system or capture-recapture estimator of the population size. A third and qualitatively different sample might work well for this demographic group.

The homogeneity of the capture probabilities is one of the assumptions in the standard approach to the estimation of the size of a closed population. Darroch *et al.* (1993) developed a quasi-symmetry model and a partial quasi-symmetry model to allow for varying catchability of individuals. The quasi-symmetry model assumes that the pattern of heterogeneity is the same for all three samples, the partial quasi-symmetry model assumes that the pattern of heterogeneity is the same for two samples but different for the third sample. This is a sensible model given that the third sample is qualitatively quite different from the census and the PES and this model is equivalent to a combination of dependence and heterogeneity. For the multinomial cell probabilities including the missing cell, $R = (r_{111}, r_{112}, \ldots, r_{222})$, both are log-linear models of the form $\log R = A\beta$ with an appropriately chosen design matrix A and a vector of parameters β. The design matrices for both models are given in Darroch *et al.* (1993).

Table 7

Heterogeneous Catchability Models

Log-Linear Model	MLE from Darroch *et al.* (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
Full quasi-symmetry	1923.63 (216.84)	133.54 (2)	1906.61 (213.47)	133.50 (2)
Partial quasi-symmetry	2576.54 (413.28)	11.70 (1)	2557.08 (409.39)	11.72 (1)

Our proposed method can readily incorporate heterogeneous catchability to estimate the population size by assuming a heterogeneity model for Table 1 and then adopting the conditional likelihood estimation (Sanathanan 1972). Table 7 presents estimates from fitting the quasi-symmetry model and the partial quasi-symmetry model for the data from stratum 11. Again, the effect of the matching errors in this analysis is not substantial due to the high matching rate. The partial quasi-symmetry model fits much better than the quasi-symmetry model, indicating there seems to be plausible heterogeneity and the pattern of heterogeneity seems different in the ALS. The lack of fit of the independence model might also be explained in part by the dependence among the samples (in particular between the census and the *P*-sample) and in part by heterogeneous catchability.

The partial quasi-symmetry model incorporates the [CP] dependence and thus is an alternative to the model [CP] [PA] in Table 5. The two models yield similar fits to the data, but they give dramatically different estimates of *N*, with the model incorporating heterogeneity having a much larger estimate accompanied by a much larger estimated standard error. This suggests that there is a considerable instability associated with heterogeneity parameters and, although the two models are not nested and thus not directly comparable, it seems reasonable to opt for the smaller and more stable estimate which does not incorporate heterogeneity.

Darroch *et al.* (1993) considered four substrata for stratum 11 in their analysis. The two cross-classification variables for the four substrata O2, R2, O3 and R3 are whether residents owned or rented homes and whether they were age 20-29 or 30-44. The data for the four substrata are given in Table 8 where 1 corresponds to presence in a sample and 0 is for absence. We have reanalyzed them for comparison. Table 9 and Table 10 give estimates for both heterogeneity models. As pointed out earlier, the high match rate yields similar estimates and fits for models incorporating matching errors. The partial quasi-symmetry model shows significant improvement in fits over the full quasi-symmetry model with the best fits obtained for R2 and R3. If we add the estimates of *N* across the four substrata, the total for the matching error version of partial quasi-symmetry is $\hat{N} = 2980.8$, more than 16% larger than the estimate from the collapsed model in Table 7. Of course, the standard error of the estimate has increased by a similar magnitude.

Table 8

Three-Sample Data for Four Substrata of Stratum 11
Source: Table 2, Darroch *et al.* (1993)

Sample			Substratum			
C	*P*	*A*	O2	R2	O3	R3
0	0	1	59	43	35	43
0	1	0	8	34	10	24
0	1	1	19	11	10	13
1	0	0	31	41	62	32
1	0	1	19	12	13	7
1	1	0	13	69	36	69
1	1	1	79	58	91	72

Table 9

Estimates for Full Quasi-Symmetry

Sub-stratum	MLE from Darroch *et al.* (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
O2	780.83 (294.81)	11.70 (2)	777.98 (293.99)	11.69 (2)
R2	394.34 (56.45)	41.09 (2)	391.14 (55.29)	41.02 (2)
O3	765.45 (254.57)	25.99 (2)	759.97 (252.44)	25.98 (2)
R3	361.83 (47.33)	59.31 (2)	358.71 (46.20)	59.22 (2)

Table 10

Estimates for Partial Quasi-Symmetry

Sub-stratum	MLE from Darroch *et al.* (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
O2	605.66 (212.63)	7.51 (1)	601.44 (210.93)	7.52 (1)
R2	652.34 (205.12)	0.04 (1)	646.59 (202.58)	0.04 (1)
O3	1124.00 (473.26)	8.27 (1)	1126.90 (476.54)	8.22 (1)
R3	611.78 (200.82)	2.92 (1)	605.91 (198.26)	2.92 (1)

6. SUMMARY

In this paper, we have presented models for matching errors and models for the estimation of the population total and census undercount in a multiple sample census. We have illustrated our methods by reanalyzing census coverage data from the 1988 St. Louis Dress Rehearsal census. Two sources of information are considered in our analysis, the data from a Matching Error Study (MES), and triple-system data with every individual cross-classified according to presence or absence in each of three samples: the census, a post enumeration survey (P-sample) and an administrative list supplement. We imbed the standard log-linear model formulation of Fienberg (1972) into our estimation procedure to account for statistical dependency together with matching errors and to allow for formal goodness-of-fit test of various models. Our method applies to any model of a log-linear form and we have illustrated how heterogeneity models can be incorporated into our approach to allow for both matching errors and heterogeneous catchability.

Our matching error models assume that false matches are negligible. Sensitivity analysis in Ding (1990) shows that when both the false nonmatch rate and the false match rate are the same order of magnitude, the matching bias is dominated by the false nonmatch rate (see also Fay, Passel, Robinson and Cowan 1988, p. 53). This is because the capture probabilities in the census and the post enumeration survey are high, and thus a comparable change in both the false nonmatch and false match rates has substantially more impact on false nonmatches than false matches. For the 1986 Los Angeles test census data, the estimates of false nonmatch rate and false match rate computed in Ding and Fienberg (1994) are about 0.5% and 0.8%, respectively. Based on these empirical findings, we have some reason to believe that, at least in the census application described here, our models for false nonmatch errors are reasonable approximations to reality.

We have analyzed the St. Louis triple-system data with an estimate of the matching rate taken from the MES. Matching rates may not be homogeneous over different population strata, and we suggest that the MES data associated with the same sampling stratum be used. We have developed formulation in §3 for the k-sample census, and our approach can be readily applied to a k-sample census with $k \geq 4$.

REFERENCES

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

CHEN, T.T. (1979). Log-linear models for categorical data with misclassification and double sampling. *Journal of American Statistical Association*, 74, 481-488.

CORMACK, R.M. (1968). The statistics of capture-recapture methods. *Oceanography and Marine Biology, Annals Review*, 6, 455-506.

DARROCH, J.N. (1958). The multiple-recapture census, I: estimation of a closed population. *Biometrika*, 45, 343-359.

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.

DING, Y. (1990). Capture-recapture census with uncertain matching. Ph.D. dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.

DING, Y., and FIENBERG, S.E. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology*, 20, 149-158.

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

FIENBERG, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59, 591-603.

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

MULRY, M.H., DAJANI, A., and BIEMER, P. (1989). The Matching Error Study for the 1988 Dress Rehearsal. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 704-709.

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.

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

ZASLAVSKY, A.M., and WOLFGANG, G.S. (1993). Triple System Modeling of Census, Post-Enumeration Survey and Administrative List Data. *Journal of Business and Economic Statistics*, 11, 279-288.