Survey Methodology, December 1993
Vol. 19, No. 2, pp. 137–147
Statistics Canada

Model-Based Estimation of Record Linkage Error Rates

J.B. ARMSTRONG and J.E. MAYDA[1]

ABSTRACT

Record linkage is the matching of records containing data on individuals, businesses or dwellings when a unique identifier is not available. Methods used in practice involve classification of record pairs as links and non-links using an automated procedure based on the theoretical framework introduced by Fellegi and Sunter (1969). The estimation of classification error rates is an important issue. Fellegi and Sunter provide a method for calculation of classification error rate estimates as a direct by-product of linkage. These model-based estimates are easier to produce than the estimates based on manual matching of samples that are typically used in practice. Properties of model-based classification error rate estimates obtained using three estimators of model parameters are compared.

KEY WORDS: Mixture model; Latent variable model; Iterative scaling.

1. INTRODUCTION

Computer files containing information about individuals, businesses or dwellings are used in many statistical applications. The linking of records that refer to the same entity is often required. The process of linking records referring to the same entity is called exact matching. If all records involved in an application have been accurately assigned a unique identifier, exact matching is trivial. Record linkage methods deal with the problem of exact matching when a unique identifier is not available. In that case, each record typically includes a number of data fields containing identifying information that could be used for matching. Problems in matching are due to errors in these data or due to the same value for a particular field being valid for more than one entity.

Applications of record linkage include the unduplication of lists of dwellings or businesses obtained from various sources to create survey frames. In addition, record linkage is widely used in applications related to health and epidemiology. Work in this area typically involves matching records containing information on individuals in industrial or occupational cohorts to records documenting the illness or death of individuals. For example, record linkage methodology for follow-up studies of persons exposed to radiation is discussed in Fair, Newcombe and Lalonde (1988).

The record linkage problem can be formulated using two data files that correspond to two populations. Each file may contain information for all entities in the corresponding population or information for a random sample of entities. The file A contains N_A records and the file B contains N_B records. The set of record pairs formed as the cross-product of A and B is denoted by $C = \{(a,b); a \in A, b \in B\}$. C contains $N = N_A \cdot N_B$ record pairs. The objective of record linkage is to partition the set C into two disjoint sets – the set of true matches, denoted by M, and the set of true non-matches, U.

The theoretical framework introduced by Fellegi and Sunter (1969) is the basis of a great deal of applied work. For each record pair, a decision is taken concerning whether or not the records refer to the same entity after examining data recorded on files A and B. The possible decisions are link (A_1), non-link (A_3) and possible link (A_2). There are two types of errors. First, decision A_1 may be taken for a record pair that is a member of U, the set of true non-matches. Second, decision A_3 may be taken for a record pair that is a member of set M, the set of true matches. Acceptable levels of classification error are specified before the files are linked. A record pair is classified as a possible link if the data do not provide sufficient evidence to justify classification of the pair as a link or non-link at error levels less than or equal to those specified. Accurate estimation of classification error rates associated with various decision rules is necessary to determine an appropriate rule. The classification error rate for true non-matches is $P(A_1 \mid U)$. The error rate for true matches is $P(A_3 \mid M)$.

Estimates of classification error rates can be obtained by selecting a sample of record pairs from the set C and manually determining the true match status of sampled pairs. Applications of this approach are described in Bartlett *et al.* (1993). Sampling may be both costly and cumbersome to implement, particularly when the same linkage must be done for a number of pairs of files, each with slightly different characteristics. Belin and Rubin (1991) describe another method of error rate estimation

[1] J.B. Armstrong and J.E. Mayda, Statistics Canada, Business Survey Methods Division, 11-RH Coats Bldg, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

that requires true match status for record pairs in a pilot study. In contrast to the straightforward sampling approach, the Belin-Rubin method provides a framework for the application of information obtained from the pilot study to larger linkages involving similar data.

The Fellegi-Sunter framework provides a method for calculation of error rate estimates using estimates of probabilities that record pairs will agree on various combinations of data fields. Calculation of these model-based error rate estimates is straightforward and manual determination of the true match status of record pairs is not required. However, they often have poor properties in applied work. See, for example, Belin (1990). In this paper, the potential for improvement of the properties of model-based error rate estimates through careful estimation of agreement probabilities is examined.

Three alternative estimation methods are evaluated. The approaches described use only the information on files A and B. They do not rely on auxiliary information. Model-based error rate estimates obtained using each alternative are compared with actual error rates using both synthetic data that incorporate important characteristics of data from health applications of record linkage, and information from an actual record linkage application.

The plan of the paper is as follows. Section 2 includes details of the model-based classification error rate estimation method introduced by Fellegi and Sunter. The model for agreement probabilities that forms the basis of subsequent discussion of estimation methods is also specified. Two estimation methods that rely on an important independence assumption are described in Section 3. A third alternative that does not require independence is discussed in Section 4. The results of comparisons of the three approaches using synthetic data are reported in Section 5. The results of evaluation work with information from a real application are described in Section 6. Section 7 contains some concluding remarks.

2. THEORETICAL CONCEPTS

Relevant aspects of the theory for record linkage developed by Fellegi and Sunter (1969) are summarized in this section. In the Fellegi-Sunter framework, estimates of classification error rates are calculated using estimates of probabilities of agreement on various combinations of data fields. Applications of the theory of Fellegi and Sunter usually involve the assumption that the probability that a record pair will agree on a particular data field is independent of the results of comparisons for other fields. The theory is nevertheless very flexible, allowing for any pattern of dependence between results of comparisons for different data fields. A parameterization of dependence in terms of loglinear effects is given.

2.1 Model-Based Classification Error Rate Estimation

To obtain information related to the classification of a record pair as a link (A_1), non-link (A_3) or possible link (A_2), data fields containing identifying information are compared. In an application involving records referring to persons, separate comparisons of family names, given names, and dates of birth might be performed. The outcome of a comparison is a numerical code representing a statement like "names agree", "names disagree", "name missing on one or both files", "names agree and both are George" or "names disagree but their first two characters agree". The outcome codes used in applied work differ between applications and between comparisons in the same application. The smallest number of outcome codes that can be used for any comparison is two – corresponding to agreement and disagreement. An outcome code corresponding to "missing on one or both files" is usually needed in applied work. The agreement outcome may be replaced by a number of value-specific outcomes (such as "names agree and both are George"). Certain disagreements may be coded as partial agreements (such as "names disagree but their first two characters agree").

For present purposes, we consider agreement and disagreement outcomes only. In the case of K matching fields, we introduce the outcome vector $\underline{x}^j = (x_1^j, x_2^j, \ldots, x_K^j)$ for record pair j. We have $x_k^j = 1$ if record pair j agrees on data field k and $x_k^j = 0$ if record pair j disagrees on data field k.

Newcombe *et al.* (1959) introduced the idea that decisions concerning whether or not a pair of records represent the same entity should be based on the ratio

$$R(\underline{x}) = P(\underline{x} \mid M)/P(\underline{x} \mid U), \tag{1}$$

where $\underline{x} = (x_1, x_2, \ldots, x_K)$ is the generic outcome vector, $P(\underline{x} \mid M)$ is the probability that comparisons for a record pair that is a true match will produce outcome vector $\underline{x}$, and $P(\underline{x} \mid U)$ is the probability of $\underline{x}$ for a record pair that is a true non-match. The optimality of record linkage methods involving this ratio was demonstrated by Fellegi and Sunter.

In the Fellegi-Sunter framework, a linkage rule assigns a probability of each classification decision (A_1, A_2 and A_3) to each outcome vector. The decision function corresponding to outcome vector $\underline{x}$ is $d(\underline{x}) = (P(A_1 \mid \underline{x}), P(A_2 \mid \underline{x}), P(A_3 \mid \underline{x}))$. Acceptable rates of classification error for true non-matches and true matches are specified before linkage is conducted. We denote these pre-specified error rates by μ and λ respectively. Among the class of record linkage rules satisfying the relations $P(A_1 \mid U) \leq \mu$ and $P(A_3 \mid M) \leq \lambda$ for fixed values of μ and λ, Fellegi and Sunter define the optimal linkage rule as the rule that minimizes $P(A_2)$, the probability that a record pair will be classified as a possible link. The optimal rule has the form

$$
\begin{aligned}
d(\underline{x}^j) &= (1,0,0) \quad \text{if} \quad \omega^j > \tau_1 \\
d(\underline{x}^j) &= (P_\mu, 1 - P_\mu, 0) \quad \text{if} \quad \omega^j = \tau_1 \\
d(\underline{x}^j) &= (0,1,0) \quad \text{if} \quad \tau_2 < \omega^j < \tau_1 \\
d(\underline{x}^j) &= (0, 1 - P_\lambda, P_\lambda) \quad \text{if} \quad \omega^j = \tau_2 \\
d(\underline{x}^j) &= (0,0,1) \quad \text{if} \quad \omega^j < \tau_2
\end{aligned} \tag{2}
$$

where $\tau_1 \geq \tau_2$, the "weight" ω^j is defined as $\omega^j = \log(R(\underline{x}^j))$ and P_μ and P_λ are positive constants in the interval [0,1). (Refer to Fellegi and Sunter (1969) for full details.) Determination of τ_1 and τ_2 requires the estimation of classification error rates corresponding to various choices for these threshold values, underscoring the importance of accurate estimation of classification error rates in the Fellegi-Sunter framework.

Model-based estimates of classification error rates can be calculated using estimates of outcome probabilities for true matches and true non-matches. Let $\hat{P}(\underline{x} \mid M)$ and $\hat{P}(\underline{x} \mid U)$ denote estimates of the probabilities of outcome vector $\underline{x}$ for true matches and true non-matches and denote the ratio of these estimates by $\hat{R}(\underline{x})$. The model-based estimate of the classification error rate for true matches based on decision rule (2) is

$$
\hat{\lambda} = \sum_{\underline{x} \in L(\tau_2)} \hat{P}(\underline{x} \mid M) + P_\lambda \sum_{\underline{x} \in Q(\tau_2)} \hat{P}(\underline{x} \mid M) \tag{3}
$$

where $L(\tau_2) = \{\underline{x};\ \log(\hat{R}(\underline{x})) < \tau_2\}$ and $Q(\tau_2) = \{\underline{x};\ \log(\hat{R}(\underline{x})) = \tau_2\}$.

The model-based estimate of the classification error rate for true non-matches is

$$
\hat{\mu} = \sum_{\underline{x} \in G(\tau_1)} \hat{P}(\underline{x} \mid U) + P_\mu \sum_{\underline{x} \in Q(\tau_1)} \hat{P}(\underline{x} \mid U) \tag{4}
$$

where $G(\tau_1) = \{\underline{x};\ \log(\hat{R}(\underline{x})) > \tau_1\}$ and $Q(\tau_1) = \{\underline{x};\ \log(\hat{R}(\underline{x})) = \tau_1\}$.

2.2 A Model For Outcome Probabilities

Calculation of model-based classification error rate estimates requires estimation of $P(\underline{x} \mid M)$ and $P(\underline{x} \mid U)$ for each of the 2^K possible values of $\underline{x}$. The probability density function for $\underline{x}$ is a mixture of two probability densities given by

$$
f(\underline{x}) = pP(\underline{x} \mid M) + (1 - p)\, P(\underline{x} \mid U), \tag{5}
$$

where p is the probability that a record pair chosen at random is a true match. The outcome probabilities depend on the frequency distributions of identifiers for entities represented on files A and B, as well as the probabilities that errors are introduced when identifiers are recorded on the files. Fellegi and Sunter (1969, pp. 1192-1194) describe a method of estimating agreement probabilities involving their definition in terms of frequency distributions and error probabilities. They recommend use of the method when prior information is available.

In the present paper, we consider situations in which the data on files A and B and the outcome vectors $\underline{x}^j$, $j = 1, 2, \ldots, N$, represent the only information available for estimation of outcome probabilities. A loglinear structure for the outcome probabilities is the most general parameterization. The saturated loglinear model for outcome probabilities for true matches is

$$
\begin{aligned}
\log(P(\underline{x} \mid M)) &= M(0) + M(1)_{x_1} + M(2)_{x_2} + \ldots \\
&\quad + M(K)_{x_K} + M(1)\, M(2)_{x_1, x_2} + \ldots \\
&\quad + M(K-1)\, M(K)_{x_{K-1}, x_K} + \ldots \\
&\quad + M(1)\, M(2) \ldots M(K)_{x_1, x_2, \ldots, x_K},
\end{aligned} \tag{6}
$$

with the usual restrictions

$$
\sum_{x_J} M(J)_{x_J} = 0, \quad J = 1, 2, \ldots, K,
$$

$$
\sum_{x_{J_1}} M(J_1) M(J_2)_{x_{J_1}, x_{J_2}} = \sum_{x_{J_2}} M(J_1) M(J_2)_{x_{J_1}, x_{J_2}} = 0,
$$

$$
\forall\ J_1, J_2, \quad \textit{etc.},
$$

as well as the restriction

$$
\sum_{\underline{x}} P(\underline{x} \mid M) = 1.
$$

The saturated model for $P(\underline{x} \mid U)$ is analogous.

If saturated loglinear models for $P(\underline{x} \mid M)$ and $P(\underline{x} \mid U)$ are employed, the density function includes $2^{K+1} - 1$ unknown parameters. It is not possible to identify all these parameters when no auxiliary information is available. In order to obtain a model that can be identified and to simplify the estimation problem, the assumption that the outcomes of comparisons for different data fields are independent is often employed. Under the assumption of independence, we denote the probabilities of agreement among record pairs that are true matches and true non-matches, respectively, by

$$
m_k = P(x_k = 1 \mid M), \quad k = 1, 2, \ldots, K,
$$

$$
u_k = P(x_k = 1 \mid U), \quad k = 1, 2, \ldots, K.
$$

Outcome probabilities can be written as

$$P(\underline{x} \mid M) = \prod_{k=1}^{K} m_k^{x_k} (1 - m_k)^{(1-x_k)},$$

$$P(\underline{x} \mid U) = \prod_{k=1}^{K} u_k^{x_k} (1 - u_k)^{1-x_k}.$$

This model involves $2 \cdot K + 1$ unknown parameters, namely $(\underline{m}, \underline{u}, p)$, where $\underline{m} = (m_1, m_2, \ldots, m_k)$, $\underline{u} = (u_1, u_2, \ldots, u_k)$. There are, of course, a number of intermediate models between the saturated model and the independence model. Methods that can be used to estimate the independence model are described in Section 3. Estimation of intermediate models is discussed in Section 4.

3. ESTIMATION UNDER INDEPENDENCE ASSUMPTION

3.1 Method of Moments

A methods of moments estimator of $P(\underline{x} \mid M)$ and $P(\underline{x} \mid U)$ can be employed in the case of independence. The estimator is based on a system of $2 \cdot K + 1$ equations that provide expressions for functionally independent moments of $\underline{x}$ in terms of the parameters. The equations are

$$E\left(\prod_{k \neq i}^{K} x_k\right) = pN \prod_{k \neq i}^{K} m_k + (1 - p) N \prod_{k \neq i}^{K} u_k, \quad i = 1, 2, \ldots, K$$

$$E(x_i) = pNm_i + (1 - p) Nu_i, \quad i = 1, 2, \ldots, K, \tag{7}$$

$$E\left(\prod_{k=1}^{K} x_k\right) = pN \prod_{k=1}^{K} m_k + (1 - p) N \prod_{k=1}^{K} u_k.$$

To obtain estimates of the parameters using the method of moments, it is necessary to solve the equations after expectations have been replaced by averages calculated using record pairs in C. The equation system for $K = 3$ was given by Fellegi and Sunter, who also derived a closed form solution that exists if some mild conditions are satisfied. Their paper included a word of caution concerning use of the method in the case of departures from independence. For $K > 3$, a closed form solution is not available but standard numerical methods can be used. Parameter estimates obtained using the method of moments are statistically consistent if the independence assumption is true.

3.2 Iterative Method

The iterative method was developed by record linkage practitioners. Although the method is not based on the probability distribution of the outcome vector, it does make use of the independence assumption. Application of the iterative method is described by several authors, including Newcombe (1988). Statistics Canada's record linkage software, CANLINK, is set up to facilitate use of the iterative method.

The method requires initial estimates of the agreement probabilities for true matches and non-matches. For true matches, guesses based on previous experience must be employed. To obtain initial estimates of agreement probabilities among record pairs that are true non-matches it is typically assumed that these probabilities are equal to the probabilities of agreement among record pairs chosen at random, namely that,

$$u_k = P(x_k = 1), \quad k = 1, 2, \ldots, K.$$

Suppose that $J(k)$ different values for data field k appear on file A and/or file B. Denote the frequencies of these values on file A by $f_{k1}, f_{k2}, \ldots, f_{kJ(k)}$ and denote the file B frequencies by $g_{k1}, g_{k2}, \ldots, g_{kJ(k)}$. For a particular value one, but not both, of the counts may be zero. The initial estimate of u_k is

$$\hat{u}_k^0 = \sum_{j=1}^{J(k)} (f_{kj}\, g_{kj})/N. \tag{8}$$

Given these probability estimates, initial sets of matches and non-matches, denoted by M^0 and U^0 respectively, are obtained using a decision rule

$$j \in M^0 \quad \text{if} \quad \omega^j > \tau_1^0,$$

$$j \in U^0 \quad \text{if} \quad \omega^j < \tau_2^0.$$

Next, frequency counts among record pairs in the sets M^0 and U^0 are used as new estimates of agreement probabilities. These estimates are used to obtain new sets of matches and non-matches and the iterative process is continued until consecutive estimates of agreement probabilities are sufficiently close.

In most applications, the assumption that the probability of agreement among record pairs that are true non-matches is equal to the probability of agreement among all record pairs is a good one and iteration does not lead to any important changes in estimates of non-match agreement probabilities. However, the first iteration often produces large changes in agreement probability estimates for true matches. Typically, there are no substantial changes at the second iteration.

It should be noted that the statistical properties of the iterative method are unclear. In practice, performance of the method will depend on the choice of the initial thresholds τ_1^0, τ_2^0. These thresholds are typically chosen subjectively. The simulations reported in Section 5 provide information about the effects of various initial thresholds.

4. RELAXING THE INDEPENDENCE ASSUMPTION - ESTIMATION USING ITERATIVE SCALING

Methods of estimation for latent variable models can be used to estimate agreement probabilities when the dependence between outcomes of comparisons for different matching fields is parameterized in terms of loglinear effects. Winkler (1989) and Thibaudeau (1989) have estimated agreement probabilities using loglinear models including all interaction terms up to third or fourth order to parameterize dependencies. The formulation presented here facilitates use of loglinear models including selected interactions. Match status can be considered a latent variable with two levels (true match and true non-match). Let $c_{0,\underline{x}}$ and $c_{1,\underline{x}}$ denote the numbers of true non-matches and true matches, respectively, with outcome vector $\underline{x}$ in a record linkage application involving K matching variables. These counts are, of course, unobservable since the value of the latent variable for each record pair is unknown. Instead, $c_{\underline{x}} = c_{0,\underline{x}} + c_{1,\underline{x}}$ is observed.

Using the parameterization of dependence in terms of loglinear effects and a saturated model for true matches, we can write

$$\log(c_{1,\underline{x}}/(Np)) = M(0) + M(1)_{x_1} + M(2)_{x_2} + \ldots + M(K)_{x_K} + M(1)M(2)_{x_1,x_2} + \ldots + M(K-1)M(K)_{x_{K-1},x_K} + \ldots + M(1)M(2)\ldots M(K)_{x_1,x_2,\ldots,x_K},$$

with the usual restrictions. A similar expression for true non-matches is available. The latent variable model corresponding to these saturated loglinear models is

$$\log(c_{s,\underline{x}}/w_s) = G(0) + Z_s + G(1)_{x_1} + \ldots + G(K)_{x_K} + ZG(1)_{s,x_1} + \ldots + ZG(K)_{s,x_K} + \ldots + G(1)G(2)\ldots G(K)_{x_1,x_2,\ldots,x_K} + ZG(1)G(2)\ldots G(K)_{s,x_1,x_2,\ldots x_K},$$

where the index s has value zero for true non-matches and one for true matches, $w_0 = (1-p)N$ and $w_1 = pN$. The parameters are analogous to the parameters of a saturated loglinear model for a contingency table of dimension 2^{K+1}. The usual restrictions apply. For example, the term $ZG(1)_{s,x_1}$ represents the interaction of the latent variable and the first matching variable and

$$\sum_s ZG(1)_{s,x_1} = \sum_{x_1} ZG(1)_{s,x_1} = 0.$$

This model conforms to the general latent variable model of Haberman (1979, p. 561). Additional restrictions must be imposed to identify and estimate the parameters. For simplicity, we will consider only hierarchical models. In addition, we restrict attention to models that allow all non-zero effects to interact with the latent variable.

In subsequent discussion we will denote latent variable models using symbols $G(1), G(2), \ldots$, loglinear models for true matches using $M(1), M(2), \ldots$ and loglinear models for true non-matches using $U(1), U(2), \ldots$. In the case of four matching variables, for example, the model $G(1)G(2), G(3), G(4)$ is a latent variable model including a general level term, main effects for all four matching variables and a term for the interaction of matching variables one and two, as well as a main effects term for the latent variable (the interaction of the general level term and the latent variable), terms for the interaction of each matching variable and the latent variable and a term for the interaction of matching variables one and two and the latent variable. The model includes 12 parameters that must be estimated. The number of parameters that must be estimated in one of the latent variable models considered here is twice the number of parameters in the corresponding loglinear model.

The iterative scaling method of Haberman (1976) can be used to estimate latent variable models. The Haberman estimation method operates by raking tables that contain estimated counts for each outcome among true matches and true non-matches. Denote the estimated counts for outcome vector $\underline{x}$ after i iterations of the Haberman algorithm by $\hat{C}^i_{1,\underline{x}}$ and $\hat{C}^i_{0,\underline{x}}$ for true matches and true non-matches, respectively. Starting values $\hat{C}^0_{1,\underline{x}}$ and $\hat{C}^0_{0,\underline{x}}$ can be constructed using estimates of agreement probabilities and the proportion of true matches obtained under the independence assumption. Each iteration of the algorithm involves a series of raking operations on the current table for true matches and the analogous rakes on the current table for true non-matches. Using the notation for hierarchical models introduced above, a set a raking operations is performed for each of the interaction terms that define the model. For four matching variables and the model $G(1)G(2), G(3)G(4)$, two sets of raking operations are performed - one for the $G(1)G(2)$ interaction and a second for the $G(3)G(4)$ interaction. For each iteraction, one raking operation is performed for every level of the corresponding classification variable. Let S_{gl} denote the set of outcome vectors at level l of term g. The raking operation on the table of true matches at iteration i for level l of term g involves computation of

$$\gamma_{1,\underline{x}} = c_{\underline{x}}\hat{c}^{i-1}_{1,\underline{x}}/(\hat{c}^{i-1}_{1,\underline{x}} + \hat{c}^{i-1}_{0,\underline{x}}),$$

$$\hat{c}^i_{1,\underline{x}} = \hat{c}^{i-1}_{1,\underline{x}} \sum_{\underline{x}\in s_{gl}} \gamma_{1,\underline{x}} \Big/ \sum_{\underline{x}\in s_{gl}} \hat{c}^{i-1}_{1,\underline{x}}, \quad \forall\, \underline{x}\in S_{gl}.$$

The algorithm is terminated when changes between estimated counts for consecutive iterations are smaller than a given tolerance.

Haberman (1976) notes that the iterative scaling algorithm may converge to a local maximum of the likelihood function rather than to the maximum likelihood estimate. Experiments with different starting values using data sets employed in the evaluation reported in Section 5 did not yield any examples of this problem.

5. COMPARISON OF ESTIMATION METHODS - SYNTHETIC DATA

In this section, the results of comparisons of the estimation methods described in Section 3 and Section 4 are presented. The comparisons involved application of each approach to a series of synthetic data sets generated using Monte Carlo methods.

Synthetic data records containing four personal identifiers (family name, middle initial, given name, date of birth) were employed. Information on possible values of each identifier, as well as their relative frequencies, was taken from the Canadian Mortality Data Base for 1988. This database, which is frequently used in health applications of record linkage, contains a separate record for each individual death.

The independence assumption was violated among true matches in each synthetic data set. Information on the frequency of outcome vectors for true matches obtained from various record linkage projects conducted by the Canadian Center for Health Information at Statistics Canada was used during data generation. Most of the projects involved matching a cohort file to the Canadian Mortality Data Base. The frequency of each outcome vector among the true matches is shown in Table 1. The dependence in these data is clear. Although approximately 88.3% of the true matches agree on given name, the probability of agreement on given name given disagreement on middle initial and agreement on family name and birth year is only 381/1366 - about 27.9%. The value of the likelihood ratio test statistic for the independence hypothesis is 3604. This value is very extreme relative to the chi-square reference distribution with 10 degrees of freedom. (Note that one degree of freedom is lost due to the zero count for the cell (1,0,0,0).)

For each synthetic data set, file A records were generated by selecting identifiers according to relative frequencies in the 1988 Canadian Mortality Data Base. In order to simplify the data generation process, the choice of family names was restricted to the 100 most common non-francophone family names and the 100 most common francophone family names found on the 1988 file. The choice of given name was restricted to the 50 most common francophone given names and the 50 most common non-francophone given names. All name choices excluded typographical variations. All middle initials and birth years found on the 1988 file were considered. Records with anglophone given names were more likely to receive an anglophone family name than records with francophone given names (reflecting the distribution of names in the Canadian population). Otherwise, identifiers were selected independently.

Table 1

Outcome Frequencies, Set of True Matches, Synthetic Data

Outcome by Identifier: 0 = Disagreement, 1 = Agreement				Frequency	
Given Name	Middle Initial	Family Name	Birth Year	Count	Percentage
0	0	0	0	7	0.03
0	0	0	1	33	0.12
0	0	1	0	125	0.45
0	0	1	1	985	3.54
0	1	0	0	5	0.02
0	1	0	1	39	0.14
0	1	1	0	202	0.73
0	1	1	1	1,848	6.65
1	0	0	0	0	0.0
1	0	0	1	13	0.05
1	0	1	0	50	0.18
1	0	1	1	381	1.37
1	1	0	0	44	0.16
1	1	0	1	451	1.62
1	1	1	0	1,751	6.30
1	1	1	1	21,860	78.65
			Total	27,794	100

The starting point for file B was an exact copy of file A. Each file B record was a true match with exactly one file A record. To introduce dependence among true matches, an outcome vector was drawn from the frequency distribution in Table 1 for each file B record. Identifiers corresponding to zeroes in the outcome vector were re-selected. Consequently, the set of outcome vectors for true matches was a sample from the Table 1 distribution. The synthetic data sets also included mild departures from the independence assumption for true non-matches since the selection of given and family names was not completely independent.

Each set of simulation results reported subsequently is based on 50 Monte Carlo trials. Each trial involved generation of files A and B of size 500, estimation of $\underline{m}$ and $\underline{u}$, determination of thresholds corresponding to various

model-based classification error rate estimates and calculation of actual error rates corresponding to the thresholds. The same series of 50 synthetic data sets was used for each set of simulations. Note that the set C contains 250,000 record pairs including 249,500 true non-matches for each Monte Carlo trial. In order to reduce computing time required by the simulations, only 49,500 true non-matches were used for each trial. (A small scale test was conducted to verify that reducing the number of true non-matches had a negligible effect on the estimated agreement probabilities.) True non-matches were removed from C by dividing files A and B into five corresponding blocks of size 100 and excluding record pairs involving records from blocks that did not correspond.

The method of moments equation system was solved using a variation of Newton's method that is described in detail in Moré *et al.* (1980). Computer code from IMSL (1987) was employed. Agreement probabilities of 0.9 for true matches and 0.1 for true non-matches for all matching fields were used as starting values for the solution of the equation system. The method did not appear sensitive to starting values.

The properties of the iterative method depend on the definitions of the initial sets of matches and non-matches, M^0 and U^0. Recall that, given initial probabilities, record pairs are classified according to

$$j \in M^0 \quad \text{if} \quad \omega^j > \tau_1^0,$$

$$j \in U^0 \quad \text{if} \quad \omega^j < \tau_2^0.$$

When the iterative method was implemented for the simulations reported here, τ_2^0 was set equal to τ_1^0. For each Monte Carlo trial, τ_1^0 was determined such that

$$\hat{P}(j \in U \mid \omega^j > \tau_1^0) + \gamma \cdot \hat{P}(j \in U \mid \omega^j = \tau_1^0) = \mu^0,$$

for some $\gamma \in [0,1)$, where the estimated probabilities are based on the initial iterative estimates of $\underline{u}$. Record pairs with weight τ_1^0 were classified in M^0 with probability γ. That is, the initial set of matches used by the iterative method was chosen to correspond to an estimated classification error rate of μ^0 for true non-matches. Starting values for m_k, $k = 1, 2, \ldots, 4$, were set to 0.9.

The zero count in Table 1 (agreement on given name, disagreement on all other identifiers) was treated as a structural zero during data generation. Among loglinear models involving no more than six parameters the model that gives the best fit to the Table 1 data is $M(1)M(2)$, $M(3), M(4)$. This model, involving dependence for outcomes of comparisons for given name and middle initial, does not fit particularly well. The likelihood ratio test statistic for lack of fit is 57.95 – an extreme value relative to the chi-square reference distribution with 9 degrees of freedom. The latent variable model $G(1)G(2)$, $G(3)$, $G(4)$ was estimated for each synthetic data set using iterative scaling. This model fit the synthetic data sets somewhat better than the model $M(1)M(2), M(3), M(4)$ fit the true match data. The largest lack of fit test statistic among the fifty synthetic data sets was 25.03 and the model was rejected only ten times at the 5% level of significance.

Averages of classification error rate estimates obtained using the synthetic data sets and the corresponding Monte Carlo standard errors are reported in Table 2 for true non-matches and Table 3 for true matches. After multiplication by 99, the error rates for true non-matches represent numbers of misclassified true non-matches divided by numbers of true matches. Results are given for the method of moments and iterative scaling, as well as the iterative method with $\mu^0 = 0.0000625$, 0.00025 and 0.001. The biases in estimated error rates for true non-matches are generally small. The iterative method with $\mu^0 = 0.001$

Table 2

Classification Error Rates, True Non-matches, Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate (× 99)	Actual Rate (× 99)				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0188 (0.0008)	0.0208 (0.0008)	0.0208 (0.001)	0.0207 (0.001)	0.0195 (0.001)
0.04	0.0381 (0.001)	0.0408 (0.0013)	0.0407 (0.0016)	0.0405 (0.0016)	0.0397 (0.0016)
0.06	0.057 (0.0012)	0.0626 (0.0015)	0.0615 (0.0018)	0.0602 (0.0019)	0.059 (0.0018)
0.08	0.076 (0.0015)	0.0855 (0.0017)	0.0838 (0.0019)	0.0804 (0.0022)	0.0785 (0.0019)
0.10	0.095 (0.0019)	0.1086 (0.0021)	0.1061 (0.0022)	0.1007 (0.0026)	0.0978 (0.0021)

Table 3

Classification Error Rates, True Matches, Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate	Actual Rate				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0580 (0.0013)	0.1179 (0.0041)	0.0507 (0.0014)	0.0149 (0.0008)	0.025 (0.0012)
0.04	0.0773 (0.0014)	0.1362 (0.004)	0.0735 (0.0012)	0.0359 (0.0018)	0.0455 (0.0016)
0.06	0.0966 (0.0014)	0.1542 (0.0038)	0.0954 (0.0012)	0.0660 (0.0014)	0.0646 (0.0018)
0.08	0.1159 (0.0014)	0.1722 (0.0036)	0.1165 (0.0012)	0.0866 (0.0017)	0.0841 (0.0019)
0.10	0.1348 (0.0014)	0.1904 (0.0035)	0.1319 (0.0014)	0.1025 (0.002)	0.1043 (0.002)

provides the best estimates, followed by iterative scaling. For true matches the performance of the iterative method is very sensitive to the choice of μ^0. Although the iterative method performs well for $\mu^0 = 0.001$, the biases for $\mu^0 = 0.0000625$ and $\mu^0 = 0.00025$ are substantial. Estimates of classification error rates for true matches obtained using the method of moments also include large biases. Biases in estimates based on iterative scaling are relatively small.

Table 4

Classification Error Rates, True Non-matches,
Modified Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate ($\times$ 99)	Actual Rate ($\times$ 99)	
	Method of Moments	Iter. Scaling
0.02	0.0189 (0.0008)	0.0194 (0.001)
0.04	0.0385 (0.0011)	0.0396 (0.0016)
0.06	0.0577 (0.0013)	0.0589 (0.0019)
0.08	0.0767 (0.0016)	0.0785 (0.002)
0.10	0.0957 (0.002)	0.0978 (0.0021)

The information in Tables 4 and 5 is based on a series of synthetic data sets generated using a modified version of Table 1. Expected values of Table 1 cell counts under the model $M(1)M(2), M(3), M(4)$ were used for data generation. The biases in model-based classification error rate estimates obtained using the method of moments are greatly reduced using the latent variable model $G(1)G(2), G(3), G(4)$ estimated using iterative scaling, particularly for true matches.

Table 5

Classification Error Rates, True Matches,
Modified Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate	Actual Rate	
	Method of Moments	Iter. Scaling
0.02	0.0553 (0.0014)	0.0208 (0.0011)
0.04	0.0747 (0.0014)	0.0415 (0.0016)
0.06	0.094 (0.0014)	0.0608 (0.0018)
0.08	0.1134 (0.0014)	0.0805 (0.002)
0.10	0.1325 (0.0015)	0.1007 (0.002)

6. COMPARISON OF ESTIMATION METHODS - REAL DATA

Results of comparisons of the three estimation methods using data from a record linkage application are presented in this section. Two data files used in empirical work reported by Fair and Lalonde (1987) were employed. The first file contained information on Ontario miners obtained from the Workmen's Compensation Board. The second file included information from the Canadian

Mortality Data Base (CMDB) for individual deaths during the period 1964 to 1977 inclusive. The miners' file included only those records with a valid social insurance number. The second file contained records that had survived an initial comparison exercise designed to eliminate records with no similarity to any of the records on the miners' file. The vital status of each miner at the end of 1977 had been classified as "confirmed dead", "confirmed alive" or "lost to follow-up" based on a previous linkage, combined with thorough follow-up procedures, including manual review. Records on the miners' file for individuals "confirmed dead" included the CMDB death registration number. More information on the construction of the files and the procedures used to determine true link status can be found in Fair and Lalonde.

Four identifiers - given name, NYSIIS code of mother's maiden name, day of birth and birth month - were chosen as matching fields for the comparison. Records on the miners' file with vital status "lost to follow-up" were eliminated. After records with missing values for at least one matching field or for birth year were also removed, file A (based on the miners' file) contained 45,638 records and file B (based on the CMDB) included 24,597 records. Restricting comparisons of the two files to pairs of records with the same NYSIIS representation of family name and birth years differing by at most one, there were 26,500 true non-matches and 2063 true matches.

Frequencies of outcomes among true matches and true non-matches are shown in Table 6. All loglinear models corresponding to a non-saturated latent variable model (that is, all models with fewer than eight parameters) are rejected by the frequency data for true non-matches at a very low level of significance. Among models with fewer than eight parameters the model $U(1)$, $U(2)U(4)$, $U(3)U(4)$ corresponds to the lowest likelihood ratio test statistic for lack of fit - 35.29. The model $M(1)$, $M(2)M(4)$, $M(3)M(4)$ provides an adequate fit to the true match data (likelihood ratio test statistic of 10.29).

Agreement probability estimates were computed using the method of moments, the iterative method and iterative scaling using the latent variable model $G(1)$, $G(2)G(4)$, $G(3)G(4)$. The likelihood ratio test statistic for the independence model corresponding to the method of moments estimator is 108 (six degrees of freedom). The independence model is rejected by the data at a very low significance level. In contrast, the likelihood ratio test statistic for the latent variable model $G(1)$, $G(2)G(4)$, $G(3)G(4)$ is 1.44 (two degrees of freedom), suggesting an adequate fit. Model-based estimates of classification error rates corresponding to each set of probability estimates were calculated for various thresholds. Actual classification error rates are compared to model-based estimates for true non-matches in Table 7 and true matches in Table 8. The error rates for true non-matches have been rescaled so that the number of true matches is in the denominator.

Table 6

Outcome Frequencies, Real Data

Outcome by Identifier: 0 = Disagreement, 1 = Agreement				Count	
Given Name	NYSIIS of Mother's Maiden Name	Day of Birth	Birth Month	True Matches	True Non-Matches
0	0	0	0	4	22,100
0	0	0	1	3	888
0	0	1	0	11	2,322
0	0	1	1	128	211
0	1	0	0	3	199
0	1	0	1	7	19
0	1	1	0	27	27
0	1	1	1	242	13
1	0	0	0	9	576
1	0	0	1	10	32
1	0	1	0	52	94
1	0	1	1	392	4
1	1	0	0	27	13
1	1	0	1	32	1
1	1	1	0	115	0
1	1	1	1	1,001	1
			Total	2,063	26,500

Model-based classification error rate estimates obtained using the iterative method are very inaccurate, particularly for true non-matches, regardless of the value of μ^0. Error rate estimates obtained using iterative scaling are slightly less accurate than estimates based on the method of moments for true matches. However, they are considerably more accurate than method of moments estimates for true non-matches.

Some words of caution are necessary. Even though the model $U(1)$, $U(2)U(4)$, $U(3)U(4)$ does not adequately describe the dependencies among true non-matches, the iterative scaling algorithm obtained a good fit using an estimate of the proportion of matched records (0.0747) that differs somewhat from the true value (0.0722). A similar fit can also be obtained using the model $G(1)G(2)$, $G(1)G(3)$, $G(4)$ and an estimate of 0.077 for the proportion of matches. Error rate estimates based on the model $G(1)G(2)$, $G(1)G(3)$, $G(4)$ are no better than estimates obtained using the method of moments.

Table 7

Classification Error Rates, True Non-matches, Real Data

Estimated Rate (× 12.84)	Actual Rate (× 12.84)				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0368	1.311	0.1859	0.186	0.0339
0.04	0.0796	1.314	0.1888	0.193	0.0649
0.06	0.1224	1.317	0.1917	0.1967	0.0684
0.08	0.1573	1.323	0.1990	0.1994	0.1106
0.10	0.1863	1.333	0.60	0.4066	0.1282

Table 8

Classification Error Rates, True Matches, Real Data

Estimated Rate	Actual Rate				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0166	0.0141	0.0193	0.0225	0.0105
0.04	0.0318	0.0264	0.029	0.0278	0.0263
0.06	0.0598	0.0383	0.0472	0.0326	0.0529
0.08	0.0782	0.0416	0.1372	0.0488	0.0784
0.10	0.0966	0.045	0.1393	0.1371	0.0958

7. CONCLUSIONS

In this paper, the issue of classification error rate estimation for record linkage has been discussed. The Fellegi-Sunter framework provides for the calculation of classification error rate estimates using estimates of agreement probabilities. These model-based estimates typically have poor properties in practice. It has been demonstrated that their properties can be improved through careful estimation of agreement probabilities. Three estimation methods have been evaluated using synthetic data as well as information from a real application.

For two of the three methods, the assumption that outcomes of comparisons for different data fields are independent was employed. This assumption was not valid for either the synthetic data or the real data. The synthetic data included strong dependencies for true matches and minor dependencies for true non-matches. Dependencies in the real data were particularly strong for true non-matches. Classification error rate estimates obtained using the method of moments, which relies on the assumption of independence, included substantial bias for synthetic data and were relatively inaccurate for real data. The magnitude of the bias in classification error rate estimates for synthetic data obtained using the iterative method depended on the definition of an initial set of matches. Although some definitions of the initial set of matches led to relatively small biases, others produced estimates with biases much larger than those obtained using the alternative methods. For the real data, all the definitions of the initial set of matches considered led to very inaccurate error rate estimates. There are no mathematical rules available for the choice of an initial set of matches for the iterative method. The results in this paper provide no evidence to recommend its use.

The third method relies on a parameterization of dependencies between outcomes of comparisons for different data fields using loglinear effects. Under this parameterization, estimates of agreement probabilities that do not rely on the independence assumption can be obtained through use of the iterative scaling method to estimate the parameters of a latent variable model. For the synthetic data sets with lack of independence, model-based classification error rate estimates obtained using iterative scaling included much smaller biases than estimates based on the independence assumption. Although the latent variable model fit most synthetic data sets better than a model based on the independence assumption, it sometimes exhibited significant lack of fit. When the synthetic data was modified to improve the fit of the latent variable

model, there was no evidence of bias in model-based classification error rate estimates. The real data included important departures from independence for both true matches and true non-matches. Model-based error rate estimates obtained using iterative scaling were slightly less accurate than estimates based on the method of moments for true matches and considerably more accurate for true non-matches.

The results reported here indicate that properties of model-based classification error rates estimates can be improved using an appropriate estimator of agreement probabilities. Latent variable models and iterative scaling provide a method of incorporating dependencies between outcomes of comparisons for different data fields during estimation of agreement probabilities.

ACKNOWLEDGEMENTS

The authors would like to thank William Winkler for providing the computer code that was the basis of the iterative scaling estimation program used to obtain our results, as well as Fritz Scheuren and three anonymous referees for comments on a earlier version of this paper that led to a significant improvement in both the content and the presentation. Thanks are also due to Martha Fair and Pierre Lalonde for making available the Ontario miners' data and the outcome frequency data for true matches.

REFERENCES

BARTLETT, S., KREWSKI, D., WANG, Y., and ZIELINSKI, J.M. (1993). Evaluation of error rates in large scale computerized record linkage studies. *Survey Methodology*, 19, 3-12.

BELIN, T.R. (1990). A proposed improvement in computer matching techniques. In *Statistics of Income and Related Administrative Record Research*: 1988-1989, U.S. Internal Revenue Service, 167-172.

BELIN, T.R., and RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 657-668.

FAIR, M.E., and LALONDE, P. (1987). Missing identifiers and the accuracy of individual follow-up. *Proceedings: Symposium on Statistical Uses of Administrative Data, Statistics Canada*, 95-107.

FAIR, M.E., NEWCOMBE, H.B., and LALONDE, P. (1988). Improved mortality searches for Ontario miners using social insurance index identifiers. Research report, Atomic Energy Control Board.

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

HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.

HABERMAN, S.J. (1979). *Analysis of Qualitative Data*. London: Academic Press.

IMSL (1987). Math/Library FORTRAN subroutines for mathematical applications. Houston: IMSL Inc.

MORÉ, J., GARBOW, B., and HILLSTROM, K. (1980). User guide for MINPACK-1. Argonne National Labs Report ANL-80-74.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.

THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.

WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 145-155.