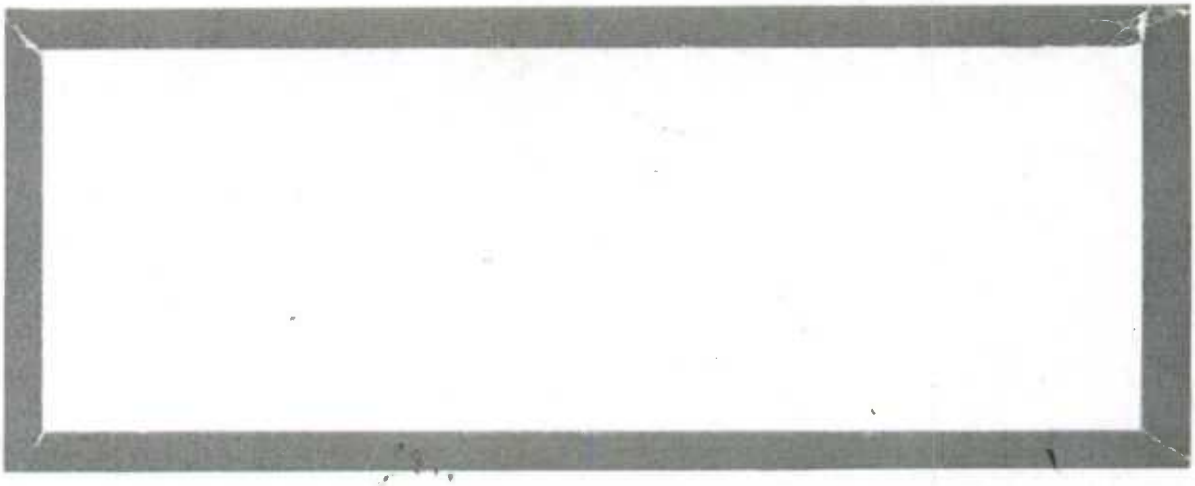
Statistique
Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes
entreprises

Canada

WORKING PAPER NO. BSMD-90-015E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-90-015E

DIRECTION DE LA MÉTHODOLOGIE

WEIGHT ESTIMATION FOR RECORD LINKAGE

by

J.B. Armstrong

December 1990

WEIGHT ESTIMATION FOR RECORD LINKAGE

John B. Armstrong[1]

RESUME

Le jumelage d'enregistrements est une méthode d'appariement exacte. On peut appliquer la méthode lorsqu'il n'est pas possible d'indentifier uniquement l'entité associée à chaque enregistrement dans un ou plusieurs fichiers. L'ensemble de paires d'enregistrements est formé de concordances (paires dont les éléments représentent la même entité) et de non-concordances (paires dont les éléments représentent des entités différentes). Le modèle de Fellegi et Sunter (1969) est souvent utilisé dans les applications. Ce modèle permet de classifier chaque paire d'enregistrements comme étant lien (concordance designée), non-lien ou cas indéterminé (paire pour laquelle on reporte une décision). Pour effectuer la classification, on utilise les poids. Le poid de concordance associé a un champ est une fonction de la probabilité de concordance du champ pour les paires concordantes et de la probabilité de concordance du champ pour les paires non-concordantes. En appliquant le modèle de Fellegi et Sunter, on pose habituellement l'hypothèse d'indépendance des champs. Etant donnée cette hypothèse, on peut considérer plusieurs estimateurs des probabilités de concordance. On peut calculer les estimations des taux d'erreur de classification à partir des estimations des probabilités. On examine les propriétés des estimations des probabilités et des estimations correspondantes des taux d'erreur de classification obtenues en utilisant diverses méthodes. On considère l'importance de l'hypothèse d'indépendance.

1. INTRODUCTION

Microdata files containing information about individuals, businesses or dwellings are used in many statistical applications. Exact matching of microdata records that refer to the same entity is often required. In some cases the existence of a unique identifier renders such a matching operation trivial. Record linkage is a technique for exact matching of microdata records when a unique identifier is not available. Typically, each microdata record includes a number of data fields containing identifying information. Each of these fields may contain errors. Positive identification of the entity associated with a particular microdata record is generally not possible without considering all identifying data fields.

Applications of record linkage include the unduplication of lists of dwellings or businesses obtained from various sources to create survey frames. The unduplication of dwelling address lists to obtain a single list intended for use as a census coverage improvement tool is described in Drew, Armstrong and Dibbs

---

1 John B. Armstrong, Senior Methodologist, Statistics Canada, 11-N RH Coats Bldg, Tunney's Pasture, Ottawa, Ontario, K1A 0T6

(1987). 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). Fair and Lalonde (1988) examine the extent to which the presence or absence of various identifiers effects the accuracy of record linkage done as part of follow-up studies.

Without loss of generality, the record linkage problem can be formulated using two data files. The file A contains N_A records and the file B contains N_B records. These two files are samples taken from the same population or different populations that contain some common entities.

The starting point for record linkage is the set of record pairs formed as the cross-product of A and B, denoted by $C = \{(a,b) \mid a \in A, b \in B\}$. 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 plan of the paper is as follows. In the second section, some details are provided concerning the mathematical model that is the basis of most applications. In section three, some methods of estimating the parameters of the model are described. The methods considered include an approach that is often used in practice, as well as alternatives that have a stronger theoretical basis. The results of some simulations, intended to provide some evidence concerning the possibility of improving current practice, are discussed in the fourth section. Section five contains some concluding remarks.

2. FELLEGI - SUNTER MODEL

This section contains a summary of aspects of the theory for record linkage developed by Fellegi and Sunter (1969) that are relevant for subsequent discussion of weight estimators and their evaluation. Subsection 2.1 includes some information related to the definition of outcomes of comparisons of data fields. Subsection 2.2 contains a description of the optimality criterion that is the basis of the theory as well as a method of estimating classification error rates. An independence assumption that is a key component of most applications is noted in subsection 2.3.

2.1 Comparisons and Outcomes

In order to obtain information related to the classification of a record pair in the set C as a member of the set of true matches, M, or the set of true non-matches, U, data fields containing identifying information are compared. For example, in an application involving personal identifiers separate comparisons

of family names, given names, and dates of birth might be performed. Three types of outcomes from such comparisons can be distinguished. These are: (i) agreement and disagreement; (ii) partial agreement; and (iii) value-specific agreement. Agreement and disagreement outcomes must always be used. These outcomes are sometimes supplemented by partial agreement and value-specific agreement in applications.

The use of partial agreement outcomes makes it possible to distinguish between cases in which the values of a data field for a record pair are similar, although not identical, and cases in which they are completely different. It is plausible to suppose that two records with family names "JOHNSTON" and "JOHNSTONE" are more likely to refer to the same individual than two records with family names "JOHNSTON" and "SMITH". If partial agreement outcomes are used in addition to agreement and disagreement, this idea can be incorporated in a record linkage application. For example, comparison of family names in an application might involve the following outcomes. The names are declared to agree if they match exactly. If they do not match exactly but they have the same NYSIIS representation, there are declared to partially agree. (NYSIIS is a phonetic encoding scheme designed to militate against the effects of common spelling errors.) If the NYSIIS representations of the names differ, they are declared to disagree. Multiple levels of partial agreement can be employed. For pairs of family names that are both longer than five characters, for example, the first level of partial agreement could be defined as agreement on the first five characters and the second level could be defined as agreement on the first three characters.

When value-specific agreement outcomes are employed, agreement of a data field for two records on a relatively rare value can be considered to provide more evidence that the record pair is a true match than agreement on a more common value. For example, agreement of two records on family name "XHIGNESSE" can be considered to provide more evidence that the records refer to the same entity than agreement on family name "SMITH". Value-specific partial agreement outcomes can also be employed for some definitions of partial agreement.

During the discussion of the Fellegi-Sunter model for record linkage in this section, it will be assumed that only agreement and disagreement outcomes are employed. For a situation involving K matching fields, we introduce the generic outcome vector $\underline{x} = (x_1, x_2, \ldots x_K)$. and the outcome vector for record pair j, $j = 1, 2, \ldots, N$, denoted by $\underline{x}' = (x_1', x_2', \ldots x_K')$. We have $x_k' = 1$ if record pair j agrees on data field k and $x_k' = 0$ if record pair j disagrees on data field k.

2.2 Optimality and Classification Error Estimation

Two types of classification error are associated with record linkage. In particular, practitioners might be concerned about the number of true non-matches incorrectly classified as matches and the number of true matches incorrectly classified as non-matches. Fellegi and Sunter pointed out that, in order to

control both these types of classification error at predetermined levels, it is necessary to employ a record linkage rule involving partition of the set C into three components, denoted here by $\tilde{M}$, $\tilde{U}$ and Q.

Record pairs in the set $\tilde{M}$ are classified as matches and those in the set $\tilde{U}$ are classified as non-matches. The link status of record pairs in the set Q cannot be determined using the record linkage rule. In order to ensure that classification errors are controlled at pre-determined levels, a manual procedure must be used to determine the true match status of record pairs in Q without error.

Fellegi and Sunter suggested that an optimal record linkage rule is a rule that minimizes the size of Q and consequently the cost of a record linkage operation. The main result of their paper states that the best record linkage rule involves use of the ratio

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

where $P(\underline{x}^j \mid M)$ is the probability that comparisons for a record pair will produce outcome vector $\underline{x}^j$, given that the record pair is a true match, and $P(\underline{x}^j \mid U)$ is the probability of $\underline{x}^j$, given that the record pair is a true non-match. The weight associated with record pair j is

$$\omega^j = 10 \cdot \log_2(R^j). \tag{2}$$

The problem of estimating ω^j is equivalent to that of estimating $P(\underline{x}^j \mid M)$ and $P(\underline{x}^j \mid U)$.

The best record linkage rule involves classifying records pairs in $\tilde{M}$, $\tilde{U}$ and Q according to

$$\begin{aligned} j \in \tilde{M} \quad &if \quad \omega^j > \tau_1, \\ j \in Q \quad &if \quad \tau_2 < \omega^j < \tau_1, \\ j \in \tilde{U} \quad &if \quad \omega^j < \tau_2, \end{aligned} \tag{3}$$

using appropriate values for the thresholds τ_1 and τ_2. For record pairs with weights equal to one of these thresholds, a random decision is necessary to ensure that classification errors are controlled exactly at the specified levels.

Estimates of τ_1 and τ_2 can be calculated using estimates of $P(\underline{x}^j \mid M)$, $P(\underline{x}^j \mid U)$, $j = 1, 2, \ldots N$, as follows. There are 2^K possible values of the outcome vector $\underline{x}$. To estimate τ_1 and τ_2, these vectors can be arranged in descending order according to the weights associated with them. Let $x^{(l)}$ denote the l^{th} outcome vector in this ordering and let $\omega^{(l)}$ denote the corresponding weight. Suppose that classification errors for true non-matches and true matches must be controlled at levels μ and λ, respectively, and that the pair of error levels

(μ, λ) is admissible, using the terminology of Fellegi and Sunter. (A pair of error levels is admissible provided that they are not both too large. More details are given at the end of subsection 2.2.) Let L_1 and L_2 denote, respectively, the smallest value of L for which $\sum_{i=1}^{L} P(\underline{x}^{(i)} \mid U) \geq \mu$ and the smallest value of L for which $\sum_{i=L}^{2^K} P(\underline{x}^{(i)} \mid M) \leq \lambda$. Estimates of τ_1 and τ_2 are given, respectively, by $\omega^{(L_1)}$ and $\omega^{(L_2)}$.

The random decision needed for record pairs with weights τ_1 or τ_2 can also be specified. Record pairs with weight τ_1 are classified in $\tilde{M}$ with probability P_μ and those with weight τ_2 are classified in $\tilde{U}$ with probability P_λ, where

$$P_\mu = \frac{\mu - \sum_{1}^{L_1 - 1} P(\underline{x}^{(i)} \mid U)}{P(\underline{x}^{(L_1)} \mid U)}, \tag{4}$$

$$P_\lambda = \frac{\lambda - \sum_{L_2+1}^{2^K} P(\underline{x}^{(i)} \mid M)}{P(\underline{x}^{(L_2)} \mid M)}.$$

Otherwise, record pairs with weights equal to one of the thresholds are classified in Q.

Fellegi and Sunter note that a record linkage rule that controls rates of classification errors for true matches and true non-matches at levels μ and λ can be constructed if $L_1 < L_2$ or $L_1 = L_2$ and $P_\mu + P_\lambda \leq 1$. In this case the pair of error levels (μ, λ) is admissible. If (μ, λ) is inadmissible, a record linkage rule with lower error rates that does not require classification of record pairs in Q can be constructed.

2.3 Independence

Weight estimation requires estimation of $P(\underline{x} \mid M)$ and $P(\underline{x} \mid U)$ for each of the 2^K possible values of $\underline{x}$. In order to reduce the number of parameters that must be estimated in applied work, it is typically assumed that the outcomes of comparisons for different data fields are independent. That is, is it assumed that

$$P(\underline{x} \mid M) = \prod_{k=1}^{K} P(x_k \mid M). \tag{5}$$

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

If independence is assumed, the number of parameters that must be estimated is reduced to $2 \cdot K$. For reasonably large numbers of matching fields, this reduction can be substantial.

3. ESTIMATION METHODS

In this section, five methods of estimating weights for the Fellegi-Sunter model that have been described in the literature are reviewed. The methods examined include two methods suggested by Fellegi and Sunter (1969), a method of moments related to method II of Fellegi and Sunter, maximum likelihood using the EM algorithm (Jaro 1989) and an iterative approach advocated by Newcombe (1988). The review of Fellegi-Sunter method I in subsection 3.1 involves discussion of estimation of weights for value-specific agreement outcomes, as well as agreement and disagreement outcomes. Discussion of the other methods, in subsections 3.2 through 3.4, is restricted to estimation of agreement and disagreement weights.

Comments concerning estimation of value-specific agreement weights using all methods except Fellegi-Sunter method I are found in subsection 3.5. This subsection also includes discussion of estimation of weights for partial agreement outcomes. It is assumed that data fields for which weight estimates are required are never missing. Missing values can be incorporated without fundamental changes in any of the weight estimation methods.

It is important to note that all the estimation methods considered here involve use of the independence assumption, (5). Given independence, new notation can be introduced. In particular, we denote the probabilities of agreement among record pairs that are true matches and true non-matches, respectively, by

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

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

and the corresponding K-vectors by $\underline{m}$, $\underline{u}$.

3.1 Fellegi-Sunter Method I

The weight estimation method described in this subsection is due to Fellegi and Sunter (1969). Suppose that there are J_k different true values for data field k for entities represented on files A and B. Denote the frequencies of these true values for entities represented on file A by $h_{Ak1}, h_{Ak2}, \ldots, h_{AkJ_k}$. Similarly, denote the frequencies for entities represented on file B by $h_{Bk1}, h_{Bk2}, \ldots, h_{BkJ_k}$.

Suppose that the set of true matches, M, contains N_M record pairs and denote the frequencies of true values for the corresponding entities by $h_{k1}, h_{k2}, \ldots, h_{kJ_k}$. It is important to note that these are frequencies for true values. In most applications the data on files A and B contain errors and these frequencies are unknown.

Further, denote the rates of errors in information for data field k on files A and B by e_{Ak} and e_{Bk}, respectively, and the probability of a change in the true value of the data field k by e_{Tk}. (In practice, a change in the true value of data field k is possible if files A and B refer to different time points.) The probability that a record pair will agree for data field k on the j_k^{th} value in the list of the J_k possible true values is given by

$$P(x_k = 1, result = j_k \mid M) = (h_{kj_k}/N_M)\cdot(1-e_{Ak})\cdot(1-e_{Bk})\cdot(1-e_{Tk})$$
$$\approx (h_{kj_k}/N_M)\cdot(1-e_{Ak}-e_{Bk}-e_{Tk}) \qquad (6)$$

for record pairs that are true matches and by

$$P(x_k = 1, result = j_k \mid U) = ((h_{Akj_k}\cdot h_{Bkj_k} - h_{kj_k})/(N_A\cdot N_B))\cdot(1-e_{Ak})\cdot(1-e_{Bk})\cdot(1-e_{Tk})$$
$$\approx (h_{Akj_k}/N_A)\cdot(h_{Bkj_k}/N_B)\cdot(1-e_{Ak}-e_{Bk}-e_{Tk}) \qquad (7)$$

for record pairs that are true non-matches. The approximations are reasonable if e_{Ak}, e_{Bk} and e_{Tk} are small and, in the case of (7), $h_{Akj_k}\cdot h_{Bkj_k}$ is large relative to h_{kj_k}.

In order to apply these formulas to estimate agreement and disagreement weights, some assumptions about the way in which errors are introduced in files A and B are necessary. In particular, if it is assumed that the probability is zero that both records in a pair that is a member of M will have erroneous but identical information for data field k, we have

$$m_k \approx (1-e_{Ak}-e_{Bk}-e_{Tk}). \qquad (8)$$

If one assumes that an erroneous value for data field k on file A, for example, will not appear on file B one can obtain

$$u_k \approx (1-e_{Ak}-e_{Bk}-e_{Tk})\cdot\sum_{j=1}^{J_k}(h_{Akj}/N_A)\cdot(h_{Bkj}/N_B). \qquad (9)$$

A similar assumption concerning the uniqueness of each value of data field k on file A that differ from the file B value for the same entity due to a true change is also required.

In applied work, it is often the case that all of the quantities involved in these formulas are unknown. If no information other than the data in files *A* and *B* is available, direct estimation of $\underline{m}$ and $\underline{u}$ is simpler than separate estimation of the quantities involved in the formulas above. The four methods of weight estimation considered in sections 3.2 through 3.4 involve direct estimation of $\underline{m}$ and $\underline{u}$. The Fellegi-Sunter method I is not examined in the simulation experiments reported in section 4. In applications in which, for example, estimates of e_{Ak} and e_{Bk} based on information not found on files *A* and *B* are available, use of (8) and (9) might be preferred to other weight estimation methods.

3.2 Fellegi-Sunter Method II and Method of Moments

If we denote by p the probability that a record pair chosen at random is a true match, the probability density function for the outcome vector $\underline{x}$ is

$$f(\underline{x}) = p \cdot P(\underline{x} \mid M) + (1-p) \cdot P(\underline{x} \mid U) \qquad (10)$$

where

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

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

This density function is a mixture of distributions involving $2 \cdot K + 1$ unknown parameters, namely $(\underline{m}, \underline{u}, p)$. The first term consists of the probability that a record pair is a true match multiplied by the probability of a particular vector of outcomes among record pairs that are true matches. The second term, referring to true non-matches, is analogous.

Fellegi-Sunter method II is based on solution of a method of moments equation system. In general, the method requires definition of an outcome vector of length three, say $\underline{z}$. Each component of $\underline{x}$ is associated with a component of this modified outcome vector. "Agreement" and "disagreement" for each component of $\underline{z}$ is defined as a combination of one or more outcome configurations for the associated components of $\underline{x}$. For example, in the case $K = 4$, the first two components of $\underline{x}$ might be associated with the first component of $\underline{z}$. "Agreement" for the first component of $\underline{z}$ might be defined as agreement on at least one of the first two components of $\underline{x}$. Let (π, μ, ρ) denote the unknown parameters of the distribution of $\underline{z}$ which is a mixture analogous to (10). Estimates of these parameters can be obtained by equating estimates of seven functionally independent moments of $\underline{z}$, obtained using data on files *A* and *B*, to expressions for the moments in terms of the parameters. The equation system given by Fellegi and Sunter is

$$E\left(\prod_{k=1}^{3} z_k\right) = \rho \cdot N \cdot \prod_{k=1}^{3} \pi_k + (1-\rho) \cdot N \cdot \prod_{k=1}^{3} \mu_k, \quad i = 1,2,3.$$

$$E(z_i) = \rho \cdot N \cdot \pi_i + (1-\rho) \cdot N \cdot \mu_i, \quad i = 1,2,3. \tag{11}$$

$$E\left(\prod_{k=1}^{3} z_k\right) = \rho \cdot N \cdot \prod_{k=1}^{3} \pi_k + (1-\rho) \cdot N \cdot \prod_{k=1}^{3} \mu_k$$

If $\pi_k \neq \mu_k$, $k = 1,2,3$. and some mild conditions on the relative sizes of sample moments (details in Fellegi and Sunter) are satisfied, the solution of this equation system exists and can be obtained analytically.

The probability of each configuration of $\underline{x}$ involved in "agreement" or "disagreement" for each of the components of $\underline{z}$ can also be estimated for true matches and true non-matches. Suppose that "agreement" for the first component of $\underline{z}$ includes L_1 configurations of agreement and disagreement involving two or more components of $\underline{x}$. Further, suppose that z_1^l is one if the l^{th} configuration occurs and is otherwise zero. Consistent estimates of $P(z_1^l = 1 \mid M)$ and $P(z_1^l = 1 \mid U)$ can be obtained using the equations

$$E(z_1^l \cdot z_2 \cdot z_3) = \rho \cdot N \cdot \pi_2 \cdot \pi_3 \cdot P(z_1^l = 1 \mid M) + (1-\rho) \cdot N \cdot \mu_2 \cdot \mu_3 \cdot P(z_1^l = 1 \mid U),$$

$$E(z_1^l \cdot z_2) = \rho \cdot N \cdot \pi_2 \cdot P(z_1^l = 1 \mid M) + (1-\rho) \cdot N \cdot \mu_2 \cdot P(z_1^l = 1 \mid U). \tag{12}$$

A similar approach can be used for configurations associated with other components of $\underline{z}$. Note that Fellegi-Sunter method II involves the assumption that the components of $\underline{z}$ are independent but does not require independence among the components of $\underline{x}$. If K is large, many ways of associating components of $\underline{x}$ and $\underline{z}$ and defining "agreement" and "disagreement" for components of $\underline{z}$ are available.

For $K > 3$ it is also possible, of course, to obtain estimates of $(\underline{m}, \underline{u}, p)$ by solving a general version of (11) based on independence of the components of $\underline{x}$ and incorporating $2 \cdot K + 1$ equations. A closed form solution is not available but standard numerical methods can be employed. Subsequently, this approach will be called the method of moments. The method of moments and Fellegi-Sunter method II are equivalent when $K = 3$.

Some general comments concerning the use of the method of moments for small samples, based on the general discussion of the use of method of moments estimators for parameters of mixtures in Titterington, Smith and Makov (1985) are appropriate. These comments apply in situations in which the independence assumption is true. First, for $K > 3$ it is possible that the solution to the equation system may not be a feasible solution (all probabilities may not lie in the interval $[0,1]$). Second, there may be more than one solution to the equation system for $K > 3$. Consequently, the solution obtained using an iterative numerical method may be sensitive to starting values. (It should be noted that sensitivity to starting values can occur even when the solution is unique if, from a numerical point of

view, it is not clearly defined.) Finally, weight estimates obtained using the method of moments approach may have higher sampling variances than estimates obtained using alternative methods when files *A* and *B* are small. The final comment also applies to method II, as Fellegi and Sunter point out. Information related to the importance of these comments for the method of moments is provided in section 4.

3.3 Maximum Likelihood

Maximum likelihood using the EM algorithm is a method for estimating the parameters of mixtures that has been applied in many contexts. (Refer to Titterington, Smith and Makov (1985).) Work on use of maximum likelihood to estimate weights for record linkage has been done by Winkler (1988) and Jaro (1989). In order to use maximum likelihood it is necessary to introduce vectors of dummy variables $\underline{z}^j$, $j = 1,2,...N$, where

$$\underline{z}^j = (1,0) \quad if \quad j \in M,$$

$$\underline{z}^j = (0,1) \quad if \quad j \in U.$$

These dummy variables are unknown and are considered as missing data during estimation of the parameters $(\underline{m}, \underline{u}, p)$. Denote by X the N x K matrix with row j given by $\underline{x}^j$ and by Z the N x 2 matrix with row j given by $\underline{z}^j$. The complete data log-likelihood is

$$l(X, Z \mid \underline{m}, \underline{u}, p) = \sum_{j=1}^{N} \underline{z}^j \cdot (\ln(P(\underline{x}^j \mid M), P(\underline{x}^j \mid U))' \qquad (13)$$

$$+ \sum_{j=1}^{N} \underline{z}^j \cdot (\ln p, \ln(1-p))'$$

Application of the EM algorithm involves starting with initial parameter estimates $(\underline{\hat{m}}, \underline{\hat{u}}, \hat{p})$. At the E or expectation step of the algorithm, $\underline{z}^j$ is replaced by $\underline{\hat{z}}^j = (\hat{z}_1^j, \hat{z}_2^j)$, where

$$\hat{z}_1^j = \hat{P}(\underline{x}^j \mid M)/(\hat{P}(\underline{x}^j \mid M) + \hat{P}(\underline{x}^j \mid U)), \qquad (14)$$

$$\hat{z}_2^j = \hat{P}(\underline{x}^j \mid U)/(\hat{P}(\underline{x}^j \mid M) + \hat{P}(\underline{x}^j \mid U)),$$

and $\hat{P}(\underline{x}^j \mid M)$ and $\hat{P}(\underline{x}^j \mid U)$ are calculated using $\underline{\hat{m}}$ and $\underline{\hat{u}}$, respectively.

At the M or maximization step, parameter estimates that maximize the log-likelihood, conditional on the $\underline{\hat{z}}^j$, are calculated. This maximization problem has a closed form solution. In particular, denote the 2^K possible outcome vectors

by $\underline{x}^l$, $l = 1,2,\ldots 2^K$, and the corresponding values of $\hat{z}_1^l$ and $\hat{z}_2^l$ by $\hat{z}_1^l$ and $\hat{z}_2^l$, $l = 1,2\ldots 2^K$. In addition, if one denotes the frequency of $\underline{x}^l$ for record pairs in the set C (the set of all record pairs) by $g(\underline{x}^l)$, we have

$$\hat{m}_k = \sum_{l=1}^{2^K} \hat{z}_1^l \cdot x_k^l \cdot g(\underline{x}^l) / \sum_{l=1}^{2^K} \hat{z}_1^l \cdot g(\underline{x}^l). \tag{15}$$

The expression for $\hat{u}_k$ is similar. The probability that a record pair chosen at random is a true match is estimated by

$$\hat{p} = \sum_{l=1}^{2^K} \hat{z}_1^l \cdot g(\underline{x}^l) / \sum_{l=1}^{2^K} g(\underline{x}^l). \tag{16}$$

More details are given in Jaro (1989).

If the independence assumption, (5), is satisfied then maximum likelihood estimates obtained using the EM algorithm are statistically consistent. In large samples, maximum likelihood estimates are equivalent to those obtained by Fellegi-Sunter method II and the method of moments. For small samples, use of the EM algorithm could have certain advantages. In particular, the form of the equations used during iteration ensures that the algorithm will not converge to an infeasible solution. In addition, Titterington, Smith and Makov note that maximum likelihood estimates of the parameters of mixture distributions often have smaller variances than estimates obtained using a method of moments approach. On the other hand, the EM algorithm may be sensitive to starting values. Information concerning these issues will be provided in section 4.

3.4 Iterative Method

The iterative method was developed by record linkage practitioners. Its use 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. Initial guesses for the agreement probabilities for record pairs that are true matches, m_k, $k = 1,2,\ldots K$, are required. These starting values are typically obtained from previous related linkage studies. In order to estimate probabilities of agreement for various data items among record pairs that are true non-matches, it is assumed that these probabilities are equal to the probabilities of agreement among record pairs chosen at random, namely that,

$$u_k = P(x_k = 1), \quad k = 1,2,\ldots K. \tag{17}$$

Suppose that J_k values appear on file A and/or file B for data field k. Denote the frequencies of these values on file A by $q_{Ak1}, q_{Ak2}, \ldots q_{AkJ_k}$ and denote the file B frequencies by $q_{Bk1}, q_{Bk2}, \ldots q_{BkJ_k}$.

The estimate of u_k corresponding to (17) is

$$\hat{u}_k = \sum_{j=1}^{J_k} (q_{Akj} \cdot q_{Bkj})/N. \qquad (18)$$

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

$$j \in M^0 \text{ if } \omega^j > \tau_1^0,$$
$$j \in U^0 \text{ if } \omega^j < \tau_2^0.$$

Next, frequency counts among record pairs in the sets M^0 and U^0 are used as new estimates of $\underline{m}_k, \underline{u}_k$, $k = 1,2,\ldots K$. 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 the estimates of u_k, $k = 1,2,\ldots K$. However, the first iteration often produces large changes in the estimates of m_k, $k = 1,2,\ldots K$. Typically, there are no substantial changes at the second iteration.

The properties of the iterative method depend on the choice of the initial thresholds τ_1^0, τ_2^0. In practice, these thresholds are determined subjectively, incorporating information from similar linkage projects. The simulations reported in the next section provide information about the effects of various initial thresholds.

3.5 Estimation of Value-Specific and Partial Agreement Weights

Use of value-specific and partial agreement outcomes is an important part of some applications. Consequently, some comments concerning weight estimation for these outcomes are appropriate. Given auxiliary information about the probabilities of errors on files used in record linkage, Fellegi-Sunter method I can be used directly for the estimation of frequency weights. Formulas for estimation of weights for partial agreement outcomes using the approach of Fellegi-Sunter method I can be developed. For example, Eagen and Hill (1987) provide a formula for the case in which character strings are compared and partial agreement is defined as agreement on the first k characters. However, the fact that such formulas include a number of parameters that are usually unknown in practice militates against use of Fellegi-Sunter method I.

The iterative method allows for estimation of weights for partial and value-specific agreement using an initial definition for the set of true matches. The iterative approach is employed by users of record linkage methodology in the Canadian Center for Health Information at Statistics Canada to obtain weight estimates for partial agreement. Starting values for partial agreement weights are determined based on experiences in similar record linkage projects. When initial sets of matches and non-matches have been defined, frequencies of record pairs that do not agree exactly but agree partially are used to produce new estimates of probabilities of partial agreement among true matches and true non-matches.

The value-specific weight for agreement of a record pair on the j_k^{th} value in the list of the J_k possible true values for data field k is

$$\omega(x_k = 1, result = j_k) = 10 \cdot \log_2(P(x_k = 1, result = j_k \mid M)/P(x_k = 1, result = j_k \mid U)). \quad (19)$$

To estimate this weight in practice, the iterative approach is not used. Instead, recalling the notation of subsection 3.1 and assuming that file A is the larger of the two data files, $P(x_k = 1, result = j_k \mid M)$ is approximated by $m_k \cdot q_{Ak j_k}/N_A$, and $P(x_k = 1, result = j_k \mid U)$ by $P(x_k = 1, result = j_k)$. In addition, it is assumed that

$$P(x_k = 1, result = j_k) = P(x_k = 1) \cdot F(j_k), \quad (20)$$

where the adjustment factor $F(j_k)$ can be defined using frequencies of true values for data field k as

$$F(j_k) = h_{Ak j_k} \cdot h_{Bk j_k} / \sum_{j=1}^{J_k} (h_{Akj} \cdot h_{Bkj}). \quad (21)$$

In practice these true values are unknown and the adjustment factor is estimated by

$$\hat{F}(j_k) = q^2_{Ak j_k} / \sum_{j=1}^{J_k} q^2_{Akj}. \quad (22)$$

It should be noted that this approach to estimation of value-specific agreement weights for data field k can be employed using any method to obtain an estimate of m_k. Winkler (1989b) describes another approach that requires estimates of $\underline{m}$ and $\underline{u}$ but does not depend on the method used to obtain these estimates.

Generalization of weight estimation by maximum likelihood, Fellegi-Sunter method II or the method of moments to accommodate partial agreement outcomes is straightforward. In the case of one level of partial agreement for each matching field, the probability density function for the outcome data is

$$f(\underline{x},\underline{y}) = p \cdot P(\underline{x},\underline{y} \mid M) + (1-p) \cdot P(\underline{x},\underline{y} \mid U) \qquad (23)$$

where

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

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

y_k is one if there is partial agreement and is otherwise zero, $\underline{\eta}$ and $\underline{\upsilon}$ are probabilities of partial agreement among record pairs that are true matches and true non-matches respectively, and $\underline{x}$, $\underline{m}$, $\underline{u}$ and p are defined as in (10). Using Fellegi-Sunter method II, estimates of probabilities of outcome configurations involving partial agreement can be obtained by appropriate definition of "agreement" and "disagreement" outcomes for the components of $\underline{z}$. Definition of "disagreement" to include partial agreement and use of equations analogous to (12) to obtain estimates of partial agreement probabilities given method of moments estimates of agreement and "disagreement" probabilities is possible for $K > 3$. The parameters of (23) can also be estimated using the EM algorithm. As Wu (1990) indicates, the expectation step of the algorithm has the same form as (14). (Probabilities for outcome vector $\underline{x}^i$ are replaced by probabilities for the pair of outcome vectors $\underline{x}^i$, $\underline{y}^i$.) The solution to the maximization step has a closed form.

In the simulations described in the next section, attention is restricted to estimation of weights for agreement and disagreement. Although the properties of weight estimates for partial and value-specific agreement obtained using various methods are of interest, estimation of these weights does not, in principle, pose a problem for any method.

4. SIMULATIONS

In this section, the results of some simulation experiments involving three weight estimation methods - maximum likelihood using the EM algorithm, the method of moments and the iterative method - are presented. The objectives of the simulations and the simulation strategy is described in subsection 4.1. Results obtained when the independence asssumption, (5), is true are described in subsection 4.2. In subsection 4.3, simulations conducted using data sets that do not satisfy the independence assumption are discussed. The extent to which the independence assumption is violated can be varied by adjusting a parameter involved in data generation. In the fourth subsection, results obtained using

data sets with departures from independence based on those found in data used in record linkage applications in the Canadian Center for Health Information at Statistics Canada are presented.

4.1 Objectives and Strategy

The simulations reported here were conducted in order to obtain information about the feasibility of improving current record linkage practice at Statistics Canada. Currently, the iterative method is used to estimate weights. In order to estimate thresholds required to control classification error rates at given levels, the true link status of a sample of record pairs is usually determined. Recall that, in the context of the Fellegi-Sunter model, threshold estimates can be calculated based on estimates of the probabilities $\underline{m}$, $\underline{u}$, as described in section 2. Compared to use of Fellegi-Sunter threshold estimates, estimation of thresholds based on the true match status of a sample of record pairs is costly and difficult to implement properly for large applications. Consequently, the main objective of the simulations was to examine the properties of Fellegi-Sunter threshold estimates. Estimated rates of classification error corresponding to Fellegi-Sunter thresholds were compared to actual classification error rates for each weight estimation method.

Given thresholds required to control classification error rates for true non-matches and true matches at specified levels, the match status of record pairs classified in the set Q (refer to (3)) must be determined manually. Subsequently, $Q(\mu,\lambda)$ will be used to denote the manual resolution set corresponding to classification error rates for true non-matches and true matches of μ and λ, respectively. The actual size of $Q(\mu,\lambda)$ based on (unknown) true thresholds, as well as the estimated size based on estimated thresholds, were examined for each weight estimation method. The actual size of $Q(\mu,\lambda)$ provides some indication of how well true matches and true non-matches are separated. Comparison of actual and estimated sizes of $Q(\mu,\lambda)$ provides evidence about the accuracy of classification error rate estimates.

A number of issues related to the performance of various estimation methods were mentioned in section 2. In particular, the following questions can be posed:

(i) how do the estimation methods perform when the independence assumption, (5), is false?

(ii) are the method of moments and/or maximum likelihood sensitive to starting values?

(iii) does the method of moments often produce infeasible estimates?

(iv) how do small sample properties of the method of moments and maximum likelihood estimates of $\underline{m}$, $\underline{u}$ compare?

The discussion in subsections 4.3 and 4.4 is focused on issue (i). The other three questions are of less practical interest and the relevant information from the simulations will be summarized now. No evidence was found that the method of moments or maximum likelihood using EM were sensitive to starting values for the data sets involved in these simulations. In addition, there were no cases in which the method of moments produced infeasible estimates. The simulations reported here each involved 50,000 record pairs. No evidence to prefer the method of moments or the EM algorithm for data sets of this size was found. Given that both methods produce statistically consistent estimates and most applications involve considerably more than 50,000 record pairs, the methods appear to be equivalent with respect to (ii), (iii) and (iv) from a practical point of view.

Synthetic data records containing four personal identifiers (family name, initial, given name, date of birth) were used in all simulations. The generation of synthetic records has the advantage that it is possible to control the extent to which the independence assumption, (5), is satisfied. Marginal distributions of identifiers were taken from the Canadian Mortality Database for 1988. Each record on this database documents an individual death and it is frequently used in Canadian record linkage applications related to health. Generation of each synthetic record involved selection of family name, initial, given name and date of birth from the appropriate marginal distribution. Only the 100 most common non-francophone family names and the 100 most common francophone family names were used. In addition, only the 50 most common francophone given names and the 50 most common non-francophone given names were employed.

Initially, the files *A* and *B* involved in each Monte Carlo trial for a particular set of simulation conditions were identical. Each record on file *A* was a true match with exactly one file *B* record. In order to create a situation requiring use of record linkage, changes were introduced in file *B* records. The changes were introduced independently for each identifier in data sets for which the independence assumption, (5), was true. For data sets used in subsections 4.3 and 4.4, dependent changes were introduced. This data generation procedure creates violations of the independence assumption for record pairs that are true matches, but not for record pairs that are true non-matches. Dependence among true matches is a more important practical problem than dependence among true non-matches simply because sets of record pairs selected at random are dominated by non-matches. Consequently, testing for independence using record pairs chosen at random is effectively equivalent, in practice, to testing for independence among true non-matches.

Each set of simulation results reported in subsequent subsections is based on 50 Monte Carlo trials. Each trial involved generation of files *A* and *B* of size 500 using the appropriate scheme to introduce errors, estimation of $\underline{m}$ and $\underline{u}$, estimation of thresholds corresponding to various classification error rates and calculation of actual error rates corresponding to these thresholds. Note that

the set C containing all record pairs includes 249,500 true non-matches for each Monte Carlo trial. In order to reduce computing costs, only 49,500 true non-matches were used during estimation.

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

$$j \in M^0 \text{ if } \omega^j > \tau_1^0,$$
$$j \in U^0 \text{ if } \omega^j < \tau_2^0.$$

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 $\hat{P}(\cdot)$ is based on the initial iterative estimates of $\underline{u}$ obtained using (17). Record pairs with weight τ_1^0 were classified in M^0 with probability γ. The initial set of matches used by the iterative method was consequently defined such that the corresponding estimated false match rate was μ^0. Starting values for m_k, $k = 1, 2, \ldots K$ were set to 0.9. The importance of these starting values for the properties of the iterative method was not examined.

Note that classification error rates for true non-matches are defined relative to the total number of non-matches in section 2. From a practitioner's point of view, it is more natural to think about these rates relative to the number of record pairs classified as matches by a record linkage procedure. For the simulation conditions, a classification error rate for true non-matches of 0.0002 corresponds to an error rate of approximately 0.02 relative to the number of record pairs classified as matches.

4.2 Independence Assumption True

To generate data sets used in the simulations reported in this subsection, file B identifiers were changed independently at a rate of 8%. When an identifier was changed, it was re-selected from the corresponding marginal distribution. It was assumed that the probability that a re-selected identifier would correspond to the original identifier was zero.

Results for classification error rates for the method of moments and maximum likelihood using EM are given in Tables 1 and 2. The results for these two methods are similar and, since the independence assumption is satisfied, it is not surprising that the biases in estimated classification error rates are small relative to their Monte Carlo standard errors.

Results for classification error rates for the iterative method for three different values of μ^0 are given in Tables 3, 4 and 5. The value of 0.0025 for μ^0, for example, indicates that according to the initial iterative estimates of $\underline{u}$ obtained using (18), the initial set of matches contained 125 records that were true non-matches (since there are approximately 50,000 records in U). The results in the tables indicate that estimated classification error rates produced using the iterative method are biased. The biases for true non-matches are small relative to estimated error rates for all values of μ^0. Although some of the biases for true non-matches are greater than two times their Monte Carlo standard errors, biases and standard errors are of the same order of magnitude. Biases in classification error rates for true matches are larger, relative to both estimated error rates and Monte Carlo standard errors. The sizes of these biases depend on the definition of the initial set of matches. The smallest biases are obtained for $\mu^0 = 0.0025$.

The three different values of μ^0 for which iterative method results are reported here were chosen arbitrarily. It is plausible to suggest that, for very small values of μ^0, the set M^0 would contain few record pairs that do not agree on all identifiers. Consequently, the estimates of probabilities of agreement among true matches used at the second iteration of the iterative method would overestimate the true probabilities and there would be a tendency to underestimate the number the number of true matches classified as non-matches. Conversely, one would expect a tendency to overestimate the number of true matches classified as non-matches for large values of μ^0 The biases in classification error rates for true matches in Tables 3, 4 and 5 are consistent with these suggestions.

Information concerning actual and estimated sizes of $Q(0.0001, 0.01)$ is given in Table 6. The pair of error rates (0.0001, 0.01) was feasible in all Monte Carlo trials. The results in Table 6 provide no evidence that weight estimation method has any importance for the actual size of the manual resolution set. Actual and estimated sizes are close for the method of moments, maximum likelihood and the iterative method with $\mu^0 = 0.0025$. There are large differences between actual and estimated sizes for the iterative method with $\mu^0 = 0.0003125$ and the iterative method with $\mu^0 = 0.005$. These results are consistent with the observed biases in classification error rate estimates.

4.3 Parametric Dependence

To generate data for the simulations reported in this subsection, independent errors in identifiers were introduced for a fraction, δ, of records in file B at a rate of $(.08)/\delta$. Using this data generation procedure, we have $P(x_k \neq 1 \mid M) = 0.08$ for $k = 1, 2, \ldots K$. In addition, outcomes of comparisons of different data fields for record pairs that are true matches are not independent. In particular, we have

$$P(x_k \neq 1 \mid x_{k'} \neq 1) = P(x_k \neq 1)/\delta, \quad k \neq k'.$$

When $\delta = 1$ the independence assumption (15) is satisfied; violation of the assumption

becomes more extreme as δ decreases. The smallest admissible value is $\delta = 0.08$.

As noted for the simulations reported in subsection 4.1, the results obtained using the method of moments and maximum likelihood were similar. Only method of moments results are reported here. Information about estimated classification error rates for the method of moments is found in Table 7 for $\delta = 0.667$ and Table 8 for $\delta = 0.5$. There are important biases in classification error rates for true matches for both values of δ, and the biases are larger for $\delta = 0.5$. Biases in classification error rates for true non-matches are small.

Results obtained with the iterative method for data sets generated using $\delta = 0.5$ are given in Tables 9, 10 and 11. These results illustrate two points. First, the biases in classification error rate estimates obtained using the iterative method depend on the definition of the initial set of matches. Second, these biases can be smaller or larger than those for the method of moments. Biases in classification error rates for true matches using the iterative method with $\mu^0 = 0.0003125$ are larger than those obtained using the method of moments. Using other values, particularly $\mu^0 = 0.005$, the iterative method produces smaller biases than the method of moments.

Note that $\mu^0 = 0.0025$, the value that produced the best estimates of classification error rates for data sets in which the independence assumption is true (section 4.2), is not the best choice here. Biases for $\mu^0 = 0.005$ are close to zero. For larger values of μ^0, one would expect biases to be larger and positive. (See Tables 3, 4 and 5 and related discussion in section 4.2.)

Actual sizes of $Q(0.0001, 0.01)$ did not provide statistically significant evidence that any of the three estimation methods separated true matches and true non-matches more effectively than the others. Differences between actual and estimated sizes of $Q(0.0001, 0.01)$ were consistent with results for biases in estimated classification error rates. Detailed results concerning $Q(0.0001, 0.01)$ are not reported here.

4.4 Dependence Based on Real Data

In the data used for the simulations described in this subsection, dependence in the set of true matches followed the pattern found in 27,794 true matches obtained from various record linkage projects conducted by the Canadian Center for Health Information at Statistics Canada. Most of these projects involved matching a cohort file to the Canadian Mortality Database. The match status of each record pair had been manually verified during the linkage operation. The frequency of each outcome vector among these true matches is shown in Table 12. Given the large amount of data, one would expect to detect any lack of independence. Nevertheless, the X^2 value corresponding to a test of the independence hypothesis is 6240 on only 3 degrees of freedom, indicating major departures from independence.

To generate data, identical *A* and *B* files were created using the marginal distributions of four identifiers from the Canadian Mortality Database. For each file *B* record, an outcome vector was drawn from the distribution given in Table 12. Identifiers corresponding to zeros in this outcome vector were re-selected using the appropriate marginal distribution. It was assumed that the probability that a re-selected identifier would correspond to the original identifier was zero.

Information about classification error rates obtained using the method of moments is given in Table 13. (Maximum likelihood results were very similar.) Although the biases in estimated classification error rates for true non-matches are small, there are important biases in estimated rates for true matches. Results for the iterative method are given in Tables 14, 15 and 16 for $\mu^0 = 0.0003125$, ,0.0025, and 0.005, respectively. These results are qualitatively similar to those for parametric dependence in the set of true matches. Biases in estimated classification error rates for true non-matches using the iterative method with $\mu^0 = 0.0003125$ are larger than those obtained using the method of moments. However, biases for $\mu^0 = 0.0025$ and $\mu^0 = 0.005$ are smaller than those for the method of moments. Biases in classification error rates for both true matches and true non-matches are particularly small for $\mu^0 = 0.005$.

5. CONCLUSIONS

In this paper, methods of estimating weights required for record linkage using the Fellegi-Sunter model were reviewed. The Fellegi-Sunter model allows for the estimation of classification error rates based on weight estimates. These classification error rate estimates are not currently used by record linkage practitioners at Statistics Canada. Since their use would reduce the costs of record linkage applications, the properties of estimates obtained using different weight estimators are of interest. The results of some simulations conducted to examine these properties are reported.

The simulation results indicate that two weight estimation methods, the method of moments and maximum likelihood, produce results that are equivalent for practical purposes. (The method of moments is equivalent to Fellegi-Sunter method II when there are three matching fields and only agreement and disagreement outcomes are employed.) When the independence assumption involved in the Fellegi-Sunter model is true, these methods yield better estimates of classification error rates than the iterative method, the approach currently used in practice. Biases in classification error rate estimates obtained using the iterative method depend on the definition of an initial set of matches. Some definitions of the initial set of matches lead to small biases -- others produce estimates with large biases. There is no evidence that weight estimation method has an effect on the size of the manual resolution region corresponding to given classification error rates.

The independence assumption is often violated in applied work. The simulation results indicate that, when the independence assumption is false, classification error rate estimates obtained using the method of moments or maximum likelihood are biased. Biases in estimates obtained using the iterative method can be smaller than those from the method of moments for certain definitions of the inital set of matches. However, inappropriate definitions of the initial set of matches produce estimates with large biases. As one would expect, the definition of the initial set of matches that yields the best estimates of classification error rates depends on the data. The definition that produced the best estimates for data sets in which the independence assumption was false differed from the definition that produced the best estimates when the independence assumption was true.

Superficially, the simulation results presented here provide some evidence to support the use of the iterative method to estimate weights in practice. When the independence assumption is false, classification error rate estimates obtained using the iterative method to estimate weights often include smaller biases than error rate estimates obtained using other weight estimators. However, the biases are large enough to strongly militate against use of the estimates in practice. In addition, the performance of the iterative method is irregular, since it depends on the definition of an initial set of true matches. The simulation results indicate that the best definition of the initial set of matches for a particular application depends on the data. A mathematical solution to the problem of determining the best definition is not available. Consequently, use of classification error rate estimates obtained from the iterative method would require calibration using a sample of record pairs for which true match status is known.

Rather than attempting to calibrate the iterative method, other methods of dealing with dependence can be considered. For example, Winkler (1989b) describes the use of a sample of record pairs for which true match status is known to improve weight estimates when the independence assumption used during estimation is false. Alternatively, a calibration sample could be used to identify a mixture model incorporating departures from independence that could be estimated using the full data set. Fellegi-Sunter method II does not require independence between outcomes of comparisons for all matching fields. Instead, it is assumed that it is possible to define "agreement" and "disagreement" such that outcomes for three composite matching fields are independent. The approach of combining matching fields and assuming independence only between outcomes for composite fields can also be used when weights are estimated by maximium likelihood or the method of moments.

ACKNOWLEDGEMENTS

The author would like to thank Ivan Fellegi and Jeff Wu for comments on a draft version of this paper that lead to large improvements. Jeff Smith and Patricia Whitridge also provided useful suggestions.

REFERENCES

Drew, J.D., Armstrong, J.B. and Dibbs, R. (1987). Research into a register of residential addresses for urban areas of Canada. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 300-305

Eagen, M. and Hill, T. (1988). Record linkage - methodology and its application. *Statistics Canada, Proceedings of the Symposium on the Statistical Uses of Adminstrative Data*, 139-150.

Fair, M.E. and Lalonde, P. (1988). Missing identifiers and the accuracy of individual follow-up. *Statistics Canada, Proceedings of the Symposium on the Statistical Uses of Administrative Data*, 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.

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

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

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley.

Winkler, W.E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 667-671.

Winkler, W.E. (1989a). Methods for adjusting for lack of independence in an application of the Fellegi-Sunter model of record linkage. *Survey Methodology*, 15, 101-117.

Winkler, W.E. (1989b). Frequency-based matching in the Fellegi-Sunter model of record linkage. Paper presented at the annual meeting of the American Statistical Association.

Wu, J. (1990). Minutes of the meeting of the Advisory Committee on Statistical Methods, October 15-16, 1990. Internal document, Statistics Canada.

Table 1: Classification Error Rates, Method of Moments

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000004	0.000012	0.02	-0.0013	0.0010
0.0004	-0.000022	0.000015	0.04	0.0004	0.0013
0.0006	-0.000024	0.000019	0.06	0.0009	0.0013
0.0008	-0.000019	0.000022	0.08	0.0010	0.0013
0.001	-0.000015	0.000024	0.10	0.0011	0.0013

Table 2: Classification Error Rates, Maximum Likelihood

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000006	0.000012	0.02	-0.0012	0.0009
0.0004	-0.000026	0.000016	0.04	0.0005	0.0012
0.0006	-0.000031	0.000019	0.06	0.0010	0.0013
0.0008	-0.000025	0.000023	0.08	0.0011	0.0013
0.001	-0.000016	0.000025	0.10	0.0012	0.0013

Table 3: Classification Error Rates, Iterative Method
$\mu^0 = 0.0003125$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000009	0.000012	0.02	-0.0087	0.0012
0.0004	-0.000034	0.000016	0.04	-0.0108	0.0011
0.0006	-0.000054	0.000021	0.06	-0.0133	0.0011
0.0008	-0.000065	0.000025	0.08	-0.0161	0.0011
0.001	-0.000075	0.000026	0.10	-0.0196	0.0011

Table 4: Classification Error Rates, Iterative Method
$\mu^0 = 0.0025$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000024	0.000013	0.02	0.0006	0.0008
0.0004	-0.000031	0.000016	0.04	0.0081	0.0010
0.0006	-0.000044	0.000020	0.06	0.0076	0.0012
0.0008	-0.000053	0.000023	0.08	0.0051	0.0013
0.001	-0.000067	0.000025	0.10	0.0045	0.0013

Table 5: Classification Error Rates, Iterative Method
$\mu^0 = 0.005$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000021	0.000013	0.02	0.0079	0.0007
0.0004	-0.000036	0.000016	0.04	0.0081	0.0008
0.0006	-0.000071	0.000019	0.06	0.0174	0.0010
0.0008	-0.000087	0.000022	0.08	0.0283	0.0013
0.001	-0.000078	0.000025	0.10	0.0330	0.0014

Table 6: Size of $Q(0.0001, 0.01)$

Method	Actual		Estimated		Difference	
	Mean	Std. Error	Mean	Std. Error	Mean	Std. Error
Method of Moments	97.07	6.87	91.46	4.07	5.61	7.04
Maximum Likelihood	95.95	6.84	90.74	4.28	5.21	7.03
Iterative Method $\mu^0 = 0.0003125$	93.15	6.98	53.44	2.60	39.72	6.91
Iterative Method $\mu^0 = 0.0025$	96.11	7.02	99.69	3.81	-3.58	6.84
Iterative Method $\mu^0 = 0.005$	94.28	6.84	159.9	3.73	-65.62	6.32

Table 7: Classification Error Rates, Method of Moments, Parametric Dependence, $\delta = 0.667$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	0.000006	0.000010	0.02	-0.0170	0.0015
0.0004	-0.000005	0.000014	0.04	-0.0188	0.0015
0.0006	-0.000006	0.000018	0.06	-0.0182	0.0015
0.0008	-0.000000	0.000020	0.08	-0.0176	0.0015
0.001	0.000004	0.000022	0.10	-0.0172	0.0015

Table 8: Classification Error Rates, Method of Moments, Parametric Dependence, $\delta = 0.50$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	0.000001	0.000011	0.02	-0.0387	0.0018
0.0004	-0.000006	0.000015	0.04	-0.0390	0.0016
0.0006	-0.000012	0.000018	0.06	-0.0380	0.0017
0.0008	-0.000009	0.000021	0.08	-0.0370	0.0017
0.001	-0.000000	0.000024	0.10	-0.0363	0.0016

Table 9: Classification Error Rates, Iterative Method, Parametric Dependence, $\mu^0 = 0.000312S$, $\delta = 0.S$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000009	0.000012	0.02	-0.0438	0.0016
0.0004	-0.000035	0.000015	0.04	-0.0446	0.0016
0.0006	-0.000060	0.000019	0.06	-0.0453	0.0016
0.0008	-0.000075	0.000023	0.08	-0.0471	0.0016
0.001	-0.000091	0.000026	0.10	-0.0489	0.0015

Table 10: Classification Error Rates, Iterative Method, Parametric Dependence, $\mu^0 = 0.002S$, $\delta = 0.S$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000019	0.000012	0.02	-0.0213	0.0013
0.0004	-0.000031	0.000014	0.04	-0.0241	0.0015
0.0006	-0.000051	0.000018	0.06	-0.0244	0.0016
0.0008	-0.000068	0.000022	0.08	-0.0224	0.0016
0.001	-0.000086	0.000026	0.10	-0.0178	0.0014

Table 11: Classification Error Rates, Iterative Method, Parametric Dependence, $\mu^0 = 0.00S$, $\delta = 0.S$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000017	0.000011	0.02	-0.0069	0.0011
0.0004	-0.000032	0.000015	0.04	-0.0053	0.0010
0.0006	-0.000064	0.000019	0.06	-0.0010	0.0015
0.0008	-0.000081	0.000023	0.08	0.0032	0.0014
0.001	-0.000085	0.000026	0.10	0.0074	0.0014

Table 12: Outcome Frequencies, Set of 27794 True Matches

Outcome by Identifier Disagreement=0, Agreement=1				Frequency	
Given Name	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	1848	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	1751	6.30
1	1	1	1	21860	78.65

Table 13: Classification Error Rates, Method of Moments, Dependence Based on Real Data

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	0.000009	0.000008	0.02	-0.0378	0.0013
0.0004	0.000016	0.000010	0.04	-0.0372	0.0013
0.0006	0.000026	0.000013	0.06	-0.0365	0.0013
0.0008	0.000035	0.000016	0.08	-0.0359	0.0014
0.001	0.000045	0.000020	0.10	-0.0349	0.0014

Table 14: Classification Error Rates, Iterative Method, Dependence Based on Real Data, $\mu^0 = 0.0003125$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000008	0.000008	0.02	-0.0962	0.0042
0.0004	-0.000008	0.000013	0.04	-0.0973	0.0040
0.0006	-0.000026	0.000015	0.06	-0.0952	0.0039
0.0008	-0.000055	0.000017	0.08	-0.0922	0.0036
0.001	-0.000086	0.000021	0.10	-0.0904	0.0035

Table 15: Classification Error Rates, Iterative Method, Dependence Based on Real Data, $\mu^0 = 0.0025$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000009	0.000010	0.02	-0.0042	0.0014
0.0004	-0.000010	0.000016	0.04	-0.0221	0.0011
0.0006	-0.000013	0.000019	0.06	-0.0258	0.0011
0.0008	-0.000023	0.000021	0.08	-0.0291	0.0012
0.001	-0.000036	0.000023	0.10	-0.0285	0.0013

Table 16: Classification Error Rates, Iterative Method, Dependence Based on Real Data, $\mu^0 = 0.005$

True Non-matches			True Matches		
Estimated Rate	Bias	Std. Error of Bias	Estimated Rate	Bias	Std. Error of Bias
0.0002	-0.000007	0.000010	0.02	0.0051	0.0008
0.0004	-0.000000	0.000016	0.04	0.0041	0.0018
0.0006	-0.000000	0.000019	0.06	-0.0060	0.0014
0.0008	-0.000000	0.000022	0.08	-0.0066	0.0017
0.001	-0.000001	0.000026	0.10	-0.0025	0.0020