ESTIMATION OF RECORD LINKAGE MODELS USING DEPENDENT DATA

J.B. Armstrong and J.E. Mayda, Statistics Canada
J.E. Mayda, Statistics Canada, 11-H R.H. Coats Bldg., Tunney's Pasture, Ottawa, Canada, K1A 0T6

KEY WORDS: mixture model, iterative scaling

1. INTRODUCTION

Microdata files containing information about individuals, businesses or dwellings are used in many statistical applications. The linking of microdata 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 microdata records involved in an application contain a unique identifier, exact matching is trivial. Record linkage methods deal with the problem of exact matching of microdata records when a unique identifier is not available. In that case, each microdata record typically includes a number of data fields containing identifying information that could be used for matching. Some of the 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.

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. The starting point for record linkage is the set of record pairs formed as the cross-product of A and B, denoted by $C = \{ (a,b); a \in A, b \in B \}$. 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.

Record linkage methods that can be applied in practice when files A and B are large involve classification of most record pairs in C using an automated procedure based on a statistical model. Since an automated matching procedure inevitably involves some misclassification, determination of classification error rates is an important issue. In most applications, classification decisions are based on the model introduced by Fellegi and Sunter (1969). When this model is used for classification, two methods for error rate estimation can be employed. First, estimates can be obtained by determining the true match status of a sample of record pairs. Alternatively the Fellegi-Sunter model provides a method for calculation of error rates as a direct by-product of linkage.

Classification error rate estimation methods involving sampling may be both costly and cumbersome to implement. Although model-based error rate estimates can be easily calculated, they often have poor properties in practice due to incorrect parameterization of the Fellegi-Sunter model and problems with estimation of model parameters.

In this paper three alternative combinations of model parameterization and estimation method are evaluated. The approaches described are appropriate when no auxiliary information about the true match status of record pairs is available. Model-based error rate estimates obtained using each alternative are compared with true error rates using various data sets.

The plan of the paper is as follows. In Section 2 some details of the Fellegi-Sunter model, including the calculation of model-based error rate estimates, are provided. Three alternative approaches to parameterization and estimation of the model are described in Sections 3, 4 and 5. The results of comparisons of the three approaches using synthetic data are reported in Section 6. The results of evaluation work with information from a real application are described in the Section 7. Section 8 contains some concluding remarks.

2. FELLEGI - SUNTER MODEL

This section contains a summary of relevant aspects of the theory for record linkage developed by Fellegi and Sunter (1969). Section 2.1 contains details of model-based classification error rate estimation. Classification error rate estimates are calculated using estimates of certain probabilities. The Fellegi-Sunter theory does not completely determine the relationship between these probabilities and underlying unknown parameters. A general parameterization is given in Section 2.2. The parameterizations used in Sections 3, 4 and 5 are particular cases of the general parameterization.

2.1 Classification Decisions and Error Rate Estimation

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. For the purposes of the present paper, the outcome of each comparison is either agreement or disagreement. For the case of K matching fields, we introduce the outcome vector for record pair j, $\underline{x}^j = (x_1^j, x_2^j, ..., x_K^j)$. 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.

The main outputs of a record linkage procedure are a set of record pairs classified as matches, $\tilde{M}$, and a set of record pairs classified as non-matches, $\tilde{U}$. Newcombe et al. (1959) introduced the idea that record linkage classification decisions should be based on the ratio

$$R(\underline{x}) = P(\underline{x}|M) / P(\underline{x}|U),$$

where $\underline{x} = (x_1, x_2, ..., x_K)$ is the generic outcome vector, $P(\underline{x}|M)$ is the probability that comparisons for a record pair that is a true match will produce outcome vector $\underline{x}$, and $P(\underline{x}|U)$ 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, the best record linkage rule involves classifying records pairs according to

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

$$j \in \tilde{U} \text{ if } \omega^j < \tau_2,$$

where $\tau_1 \geq \tau_2$ and the "weight" ω^j is defined as

$$\omega^j = 10 \log_2 (R(\underline{x}^j)).$$

Denote the size of set M by $|M|$. The classification error rate for true matches is given by $\mu = |M \cap \tilde{U}|/|M|$ and the error rate for true non-matches is $\lambda = |U \cap \tilde{M}|/|U|$. The model-based estimate of the classification error rate for true matches is given by

$$\hat{\mu} = \sum_{\underline{x} \in L(\tau_2)} P(\underline{x}|M)$$

where $L(\tau_2) = \{\underline{x}; 10 \log_2(R(\underline{x})) < \tau_2\}$. The model-based estimate is the sum of the probability of outcome vector $\underline{x}$ among true matches, calculated over all outcome vectors with weights less than τ_2. The model-based estimate for true non-matches is analogous.

2.2 A Model For Outcome Probabilities

Calculation of model-based classification error rate estimates requires estimation of $P(\underline{x}|M)$ and $P(\underline{x}|U)$ for each of the 2^K possible values of $\underline{x}$. It is evident that these outcome probabilities will depend on the frequency distributions of true values for various data fields, as well as the manner in which errors are introduced on files A and B. In the present paper we consider situations in which the outcome vectors x^j represent the only information available for estimation of outcome probabilities.

The probability density function for $\underline{x}$ is given by

$$f(\underline{x}) = p \cdot P(\underline{x}|M) + (1-p) \cdot P(\underline{x}|U),$$

where p is the probability that a record pair chosen at random is a true match. A log-linear structure for the outcome probabilities is the most general parameterization. Use of log-linear models for record linkage has been considered by Winkler (1989) and Thibaudeau (1989). The saturated log-linear model for outcome probabilities for true matches is

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

with the usual restrictions. The saturated model for $P(\underline{x}|U)$ is analogous.

If saturated log-linear models for $P(\underline{x}|M)$ and $P(\underline{x}|U)$ are employed, the density function includes $1+2^{K+1}$ unknown parameters. Since no auxiliary information is available, it is not possible to identify all of these parameters. 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

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

and the corresponding K-vectors by $\underline{m}$, $\underline{u}$. Outcome probabilities can be written as

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

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

3. ESTIMATION UNDER INDEPENDENCE ASSUMPTION METHOD OF MOMENTS

A method of moments estimator of $P(\underline{x}|M)$ and $P(\underline{x}|U)$ can be employed in the case of independence. The probability density function for record pairs under independence is a mixture of distributions involving $2 \cdot K+1$ unknown parameters. The method of moments estimator is based on a system of $2K+1$ equations that provide expressions for functionally independent moments of $\underline{x}$ in terms of the parameters. 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 (1969), 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.

4. ESTIMATION UNDER INDEPENDENCE ASSUMPTION - ITERATIVE METHOD

The iterative method was developed by record linkage practitioners. Although the method is not based on the probability distribution of record pairs, 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 agreement probabilities for true matches and true 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 assumed that these probabilities are equal to the probabilities of agreement among record pairs chosen at random. 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 $q_{Ak1}, q_{Ak2}, \ldots, q_{AkJ(k)}$ and denote the file B frequencies by $q_{Bk1}, q_{Bk2}, \ldots, q_{BkJ(k)}$. The initial estimate of u_k is

$$\hat{u}_k = \sum_{j=1}^{J(k)} (q_{Akj} \cdot q_{Bkj})/(N_A \cdot N_B).$$

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

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

Next, frequency counts among record pairs in the sets M^0 and U^0 are used as new estimates of 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.

5. RELAXING THE INDEPENDENCE ASSUMPTION - ESTIMATION USING ITERATIVE SCALING

The iterative scaling estimation methods developed by Haberman (1976) were first applied to record linkage by Winkler (1989) and Thibaudeau (1989). Consider information about the frequency of all possible outcome vectors among record pairs that are true matches and true non-matches in a record linkage application involving K matching variables. If the true match status of each record pair was known, this data could be summarized in a contingency table of dimension 2^{K+1}. In practice, true link status is unknown and the available table of dimension 2^K can be modelled using a latent variable model with one latent variable at two levels.

The components of the categorical latent variable models described by Everitt (1984) are typically independence models. That is, the contingency table corresponding to each level of the latent variable is independent. Use of latent variable models nevertheless facilitates the incorporation of certain types of dependence. Consider the case K=4 and suppose one would like to allow for dependence between the outcomes of comparisons for matching variables one and two with other variables independent. A latent variable model with one unknown classification variable at two levels and independent sub-models could be fit to a 4x2x2 table, where the classification variable at four levels corresponds to the four possible outcomes for the joint comparison of matching variables one and two.

Subsequent discussion will be facilitated by additional notation. We will denote the model with independent outcomes for all K matching variables by G(1), G(2),..., G(K) . A model incorporating dependence between matching variables one and two, with other variables independent, will be denoted by G(1)G(2), G(3),..., G(K). It is apparent that a model in which the symbol for each matching variable appears in only one term can be estimated using a latent variable model with independent sub-models.

The Haberman estimation method operates by raking tables of counts. Denote by $c(\underline{x}^h)$ the number of record pairs with outcome vector $\underline{x}^h$, h = 1,2,...,2^K . Let $c(\underline{x}^h \mid M)$ and $c(\underline{x}^h \mid U)$ denote the unknown true counts for outcome vector $\underline{x}^h$ among true matches and true non-matches, respectively. Denote the estimated counts after i iterations of the Haberman algorithm by $\hat{c}(\underline{x}^h \mid M)^i$ and $\hat{c}(\underline{x}^h \mid U)^i$. Starting values $\hat{c}(\underline{x}^h \mid M)^0$ and $\hat{c}(\underline{x}^h \mid U)^0$ can be constructed using estimates of agreement probabilities and proportions 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 introduced above, a set of rakes is performed for each term in the model. For a particular term, a rake is done for every level of the corresponding classification variable. Let S_{gl} denote the set of outcome vectors at level l of term g . The iteration i rake of the table of true matches for level l of term g involves computation of

$$c^*(x^h|M)^i = \hat{c}(x^h)\cdot\hat{c}(x|M)^{i-1}/(\hat{c}(x^h|M)^{i-1} + \hat{c}(x^h|U)^{i-1}),$$

$$\hat{c}(x^h|M)^i = \hat{c}(x^h|M)^{i-1}\cdot\sum_{h\in S_{gl}} c^*(x^h|M)^i / \sum_{h\in S_{gl}} \hat{c}(x^h|M)^{i-1},$$

$$\forall\ x^h \in S_{gl}.$$

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 relative maximum of the likelihood function rather than the maximum likelihood estimate. Experiments with different starting values using data sets employed in the evaluation reported in Section 6 did not yield any examples of this problem.

6. COMPARISON STUDY - SYNTHETIC DATA

In this section, the results of comparisons of the combinations of model parameterizations and estimation methods described in Sections 3, 4 and 5 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. Marginal distributions of identifiers were taken from the Canadian Mortality Data Base for 1988. Each synthetic data set was generated so that the independence assumption was violated among true matches. Frequency data for various outcome vectors for true matches obtained from record linkage projects conducted by the Canadian Centre for Health Information (CCHI) at Statistics Canada was used to introduce violations of the independence assumption.

The value of the likelihood ratio test statistic for the independence hypothesis, computed using the CCHI frequency data is 6240. This value is extreme relative to the reference distribution -- the X^2 distribution with 11 degrees of freedom. 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 subsequently is based on 50 Monte Carlo trials. Each trial involved generation of files A and B of size 500, estimation of m_1, m_2, m_3, m_4 and u_1, u_2, u_3, u_4, calculation of thresholds corresponding to various model-based classification error rate estimates and calculation of actual error rates corresponding to the thresholds.

The method of moments equation system was solved using a variation of Newton's method that is described in detail in Moré (1980). Computer code from IMSL (1987) was employed.

The properties of the iterative method depend on the

definitions of the initial sets of matches and non-matches, M^0 and U^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

$$P(j \in U | \omega^j > \tau_1^0) + \gamma \cdot P(j \in U | \omega^j = \tau_1^0) = \lambda^0,$$

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 defined such that the corresponding estimated false match rate was λ^0. Starting values for m_k, $k = 1,2,3,4$, were set to 0.9.

Information about classification error rates for true matches is given in Table 1. An estimated rate of 0.02 based on the method of moments, for example, corresponds to an actual rate of 0.058. Results for the iterative method are given for $\lambda^0 = 0.00025$ and 0.001. Biases in estimated classification error rates for true matches using the iterative method are smaller than for the method of moments, and are particularly small for $\lambda^0 = 0.001$. Among latent variable models that could be identified for all synthetic data sets, the model that gives the best fit to the CCHI frequency data is G(1)G(2), G(3), G(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 X^2 reference distribution with 10 degrees of freedom.

Information about model-based classification error rate estimates obtained using the latent variable model with dependence between given name and middle initial and a series of synthetic data sets is given in the fifth column of Table 1. Comparing column five with the iterative method and the method of moments results, one notes that use of the latent variable model with dependence removes most of the bias in model-based classification error rate estimates based on the independence assumption. The biases using the latent variable model are nevertheless large relative to their Monte Carlo standard errors. The evidence that model-based classification error rates estimates obtained using G(1)G(2), G(3), G(4) are biased is consistent with the lack of fit of this model to the CCHI frequency data.

The rightmost column of Table 1 is based on a series of synthetic data sets generated using a modified version of the CCHI frequency data. In particular, expected values of the frequency counts under the model G(1)G(2), G(3), G(4) were employed. The biases in model-based classification error rates based on the independence assumption are apparently eliminated when the correct latent variable model is used.

7. COMPARISON STUDY - REAL DATA

7.1 Objectives

The main objective of the study was to provide error rate estimates to personnel in the Canadian Centre for Health Information (CCHI) at Statistics Canada, based on weights derived using various models. The intention of this study was that if the error rate estimates from a particular model were satisfactory, then the model could be used in future applications to generate weights and estimate error rates.

The data that was used for the comparison study was obtained from the CCHI. The same data sets were used as were involved in evaluations reported in Fair and Lalonde (1987). The main file contained data on Ontario Miners for whom social insurance identifiers were available from the Workers' Compensation Board. This file was linked to a subset (1964-1977) of the Canadian Mortality Data Base. These files included the true link status (alive, dead, lost to follow-up) that was determined by 100% manual resolution in the original Fair and Lalonde study. The status "dead" was used in this study as the indicator of a true match. This allowed us to calculate the actual error rates and compare them to the estimated error rates.

7.2 Methodology Used

Four identifiers -- surname, first given name, birth month and birth day -- were chosen as matching fields for this study. Weights were estimated using the alternative methods. The CANLINK system was used to link the data files, based on the estimated weights from each method.

7.2.1 Iterative Method

Weights were estimated for this method using a function available in the CANLINK system. The unlinked set components for each matching field were derived based on the probability of agreement among record pairs chosen at random from the population. In order to generate the linked set components for each matching field, initial guesses for the agreement probabilities for record pairs that are true matches are required. Based on previous experience, 0.10 (or 10%) was arbitrarily chosen as an initial value. For this study missing outcomes were assumed equally likely for true matches and true non-matches.

At the stage where the matched pairs were identified, the records were first blocked into smaller groups based on the NYSIIS code for the surname. This is done in the CANLINK system in order to reduce the number of possible pairs to be compared. This is very practical as the system only looks for links within pockets, which reduces the processing time and complexity.

Once the unlinked set components of weights were generated by the CANLINK system, and the linked set components were assigned to the matching fields, thresholds to identify an initial set of matches and an initial set of non-matches in the data files had to be determined. The initial thresholds were set up in this study in order to identify the initial set of matches and non-matches to be used for the recalculation of the weights in subsequent steps. After examining the weights and some preliminary runs through the CANLINK system with various limits, thresholds were determined such that the number of record pairs identified as links was not too large.

It is at this point that the iteration process is usually carried out, and new weights can be generated using the outcomes of comparisons for record pairs classified as links and those classified as non-links. For this study, we did not iterate because initial attempts at iteration did not appear to affect the weights significantly. Since there was no iteration, the next step in CANLINK was the group resolution phase where a one-to-one mapping of the links was carried out in order to have groups formed with only one record from each file in a matched pair. Conflicts that

had to be resolved due to duplicates on the data files were dealt with automatically by the system by examining all links and then breaking a link in conflict if there was another link with a higher weight. The "best" links were therefore determined by the linkage weight alone.

7.2.2 Method of Moments

The computer software mentioned in Section 6 was used for estimation. Only record pairs that agreed on the NYSIIS code for surname were included in the set used for estimation. Note that this type of blocking was not used during calculation of probability estimates for the iterative method.

From the agreement probabilities for each matching field calculated by this program, the agreement and disagreement weights for both the linked and unlinked sets were calculated using the formula mentioned in Section 2.1. Once the weights were determined, they were assigned to the matching fields and the CANLINK system was again used to determine the links. The processing in CANLINK followed the procedures described in Section 7.2.1.

7.3 Evaluation and Results

Once the set of linked records was identified using estimated weights from each method, the first step in the evaluation was to identify the true links from the CCHI study. This was done using the status available on the data files which was generated when the original evaluations were done by the CCHI. Next, the links identified by the iterative method and method of moments estimators were each merged with the true links. This allowed us to examine the number of true matches identified by each weight estimation method, as well as the number of false matches made or true matches missed for various thresholds.

Biases in estimated classification error rates based on the iterative method are shown in Table 2 for various thresholds. The threshold corresponding to an estimated rate of .05, determined using the methods described in Section 2.1, produced an actual rate of .04866. Table 3 gives similar information for the method of moments. The threshold corresponding to an estimated rate of 0.219 produced an actual rate of 0.383. It was assumed that missing outcomes were equally likely for true matches and true non-matches.

The biases in the estimated error rates for true matches generated by the iterative method are important, but are much smaller than the biases in method of moments estimates. The estimated error rates for true matches based on the method of moments method differ by 58-75% from the actual error rates for various thresholds. These relative biases also increase as the global weight decreases, indicating that the estimated error rates may be less stable when the estimates are smaller.

8. CONCLUSIONS

In this paper, the issue of classification error rate estimation for record linkage has been discussed. Fellegi-Sunter theory allows for the calculation of model-based classification error rate estimates. Model-based estimates are rarely used by practitioners because they often have poor properties. The poor properties of these estimates are related to incorrect parameterization of the Fellegi-Sunter model. In practice, error rates are estimated by selecting a sample of record pairs and determining the true link status of selected records. This sampling procedure is expensive and often cumbersome to implement. Consequently, it would be desirable if the quality of Fellegi-Sunter classification error rate estimates was sufficient to allow their use in practice.

Three combinations of parameterization of the Fellegi-Sunter model and estimation method have been evaluated using synthetic data as well as information from a real application. Model-based classification error rate estimates were calculated for each combination and compared to actual error rates.

Classification error rate estimates obtained using the method of moments, which relies heavily on the assumption of independence, included substantial bias. This was particularly evident using the real data. For the simulated data, the magnitude of the bias in classification error rate estimates obtained using the iterative method depended on the definition of an initial set of matches. Some definitions of the initial set of matches led to relatively small biases -- others produced estimates with biases larger than those obtained using the method of moments. Using the real data, the iterative method error rate estimates differed from the actual rates, but were closer than the method of moments estimates.

The third alternative, specification of a latent variable model and estimation using iterative scaling, does not require the assumption of independence. For the synthetic data sets with lack of independence, model-based classification error rates estimates obtained using iterative scaling included much smaller biases than estimates obtained using the independence assumption. The biases remaining in the error estimates based on iterative scaling reflect the fact that while the latent variable model employed fit the data much better than a model based on the independence assumption, it still exhibited significant lack of fit. When the synthetic data was modified so that a latent variable model that did not exhibit significant lack of fit could be identified, there was no evidence of bias in model-based classification error rate estimates.

The results reported here suggest that practical use of model-based classification error rates estimates may be possible if record linkage models are estimated using an appropriate parameterization. The use of latent variable models and iterative scaling provide a method of incorporating dependencies between outcomes of comparisons for different data fields in record linkage models. It should be noted that both the method of moments and the iterative method can be modified to allow for dependencies. Compared to modified versions of these methods, the latent variable - iterative scaling approach has the advantage of simplicity. The evaluation results reported here suggest that it should be used more frequently in practice.

The next phase of the research presented in this paper will be to identify an appropriate latent variable model for the real data application, and apply it to the data in order to estimate classification error rates based on that model.

ACKNOWLEDGEMENTS

The authors would like to thank William Winkler for introducing them to the iterative scaling method and providing the computer code that was the basis of the iterative scaling estimation program used to obtain results reported here. Thanks are also due to Martha Fair and Pierre Lalonde for making available the Ontario miners' data and the CCHI frequency data for true matches.

REFERENCES

Everitt, B.S. (1984). An introduction to latent variable models. London: Chapman and Hall.

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

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

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

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

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. American Statistical Association, Proceedings on the Statistical Computing Section, 283-288.

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

TABLE 1
Classification Error Rates for True Matches -- Synthetic Data

Estimated Rate	Bias -- Method of Moments	Bias -- Iterative Method $\lambda^0 = 0.00025$	Bias -- Iterative Method $\lambda^0 = 0.001$	Bias-- Iterative Scaling	Bias -- Iterative Scaling Modified data
0.02	-0.0380	-0.0307	0.0051	-0.0050	-0.0008
0.04	-0.0373	-0.0335	0.0041	0.0055	-0.0015
0.06	-0.0366	-0.0354	-0.0060	0.0046	-0.0008
0.08	-0.0359	-0.0365	-0.0066	0.0041	-0.0005
0.10	-0.0348	-0.0319	-0.0025	0.0043	-0.0007

TABLE 2
Classification Error Rates for True Matches
Real Data -- Iterative Method

Estimated Rate	Bias
0.05	0.0134
0.20	-0.0259
0.34	-0.0554

TABLE 3
Classification Error Rates for True Matches
Real Data -- Method of Moments

Estimated Rate	Bias
0.2190	-0.1640
0.2368	-0.1773
0.4061	-0.2405
0.6125	-0.3634
0.6903	-0.4029