WEIGHT ESTIMATION FOR LARGE SCALE RECORD LINKAGE APPLICATIONS

John Armstrong, Maher Saleh, Elections Canada
Elections Canada, 257 Slater Street, Ottawa, Canada K1A 0M6

KEY WORDS: Fellegi-Sunter model, mixture model, EM algorithm

1. INTRODUCTION

Record linkage refers to the problem of bringing together records that refer to the same entity, often the same individual, when the identifying information on the records is not unique. Modern record linkage began with the work of Howard Newcombe (Newcombe et al., 1959) who introduced the concept of a decision rule that classified record pairs based on an odds ratio of frequencies. Large values of this odds ratio, or weight, provided evidence that both records in a pair represented the same individual. Fellegi and Sunter (1969) provided a mathematical model for record linkage that formalized the ideas of Newcombe. In their model, Newcombe's odds ratio appeared as a ratio of probabilities.

Estimation of weights is a critical component of applied record linkage work. In this paper, some of the conventional weight estimation methods that are available in widely used software packages are compared to statistical methods based on estimation of mixture models. The methods are compared in the context of a large scale application of record linkage at Elections Canada. Previous empirical comparisons of conventional weight estimation methods with statistical methods based on mixture models were performed by Winkler (1994). Although the statistical methods led to only marginal improvements in matching for the data considered in those comparisons, Winkler (1999, p. 78) suggests that the question "For what types of files and matching situations can general dependence-based probabilities and decision rules improve matching?" is still open. The work reported in this paper is relevant to the issue.

The paper is organized as follows. Section 2 contains discussion of record linkage concepts and statistical weight estimation methods. The Elections Canada application, the baseline methodology that relied on conventional weight estimation methods and the corresponding results are described in section 3. Results from the use of statistical weight estimation methods involving maximum likelihood estimation of mixture models are compared with the results from the baseline methodology in section 4. Conclusions can be found in section 5.

2. RECORD LINKAGE

Background information about record linkage is provided in this section. Record linkage concepts are described in section 2.1 and section 2.2 includes discussion of statistical methods for weight estimation. Two important issues for most record linkage applications, blocking and the constraint of one-to-one matching, are examined in section 2.3.

2.1 Concepts

The problem of record linkage can be formulated using data file A containing N_A records and data file B containing N_B records. The objective of record linkage is to partition the set of record pairs $C = \{(a,b) | a \in A, b \in B\}$ into two disjoint sets – the set M of true matches and the set U of true non-matches.

A record linkage process classifies record pairs based on the results of comparisons of data items. In an application involving name and address data, the data items compared could include first name, family name, day of birth, postal code, etc. If K data items are compared for record pair (a,b) the results may be represented by the outcome vector $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ where $x_i = 0$ if data item i disagrees and $x_i = 1$ if data item i agrees. Components of $\mathbf{x}$ may assume other values that reflect the degree of similarity between data items that disagree. The results of comparisons of family names, for example, may reflect the fact that while both the pair "MacLean" and "Maclaine" and the pair "MacLean" and "Brown" disagree, the former pair are more similar than the latter.

In addition, components of $\mathbf{x}$ may assume other values to reflect the relative frequency of the specific value agreed on for data items that agree. Agreement on the value "Brzozowski" for family name provides more evidence that both records refer to the same individual than agreement on the value "Smith" since, provided that data for an English speaking population are used, the former agreement is much less likely to occur by chance. The probability density function for outcome vector $\mathbf{x}$ is a mixture of two probability density functions given by

$$P(\mathbf{x}) = P(\mathbf{x}|M)P(M) + P(\mathbf{x}|U)P(U), \quad (1)$$

where $P(\mathbf{x}|M) = P(\mathbf{x}|(a,b) \in M)$ and $P(\mathbf{x}|U) = P(\mathbf{x}|(a,b) \in U)$.

Fellegi and Sunter (1969) proposed classification of record pairs based on the ratio of probabilities $R(\mathbf{x}) = P(\mathbf{x}|M)/P(\mathbf{x}|U)$. Large values for $R(\mathbf{x})$ provide evidence that record pair (a,b) is a true match. The decision rule provided by Fellegi and Sunter classifies each record pair as a member of one of three sets – the set of links ($\tilde{M}$), the set of non-links ($\tilde{U}$) and the set of possible links (Q) – using two thresholds. The rule is given by

$$\begin{aligned} (a,b) \in \tilde{M} \quad &\text{if} \quad R(\mathbf{x}) > \tau_1 \\ (a,b) \in Q \quad &\text{if} \quad \tau_2 \le R(\mathbf{x}) \le \tau_1 \\ (a,b) \in \tilde{U} \quad &\text{if} \quad R(\mathbf{x}) < \tau_2. \end{aligned} \quad (2)$$

Fellegi and Sunter defined $R(\mathbf{x})$ as the weight associated with outcome vector $\mathbf{x}$. Weights have been defined using other monotonic transformations of $R(\mathbf{x})$ by other authors. The choice of monotonic transformation has no implications for the estimation of $R(\mathbf{x})$.

The thresholds τ_1 and τ_2 are calculated to satisfy two classification error bounds – a bound on the proportion of true matches classified as non-links and a bound on the proportion of true non-matches classified as links – that are specified before linkage is conducted. Possible links are reclassified as links or non-links after manual review. It is assumed that this review process does not produce any classification errors. Fellegi and Sunter showed that (2) is optimal in the sense that, for any admissible pair of classification error bounds, τ_1 and τ_2 can be calculated so that the size of Q is minimized while the bounds are satisfied.

2.2 Statistical Weight Estimation Methods

Statistical weight estimation methods that rely on observed data are based on (1), the probability density function of the outcome vector $\mathbf{x}$. If each component of $\mathbf{x}$ represents either agreement or disagreement then, under the assumption that the components are mutually statistically independent, the probability density function is subject to the constraints

$$\begin{aligned} P(\mathbf{x}|M) &= \prod_1^K P(x_k|M) \\ P(\mathbf{x}|U) &= \prod_1^K P(x_k|U). \end{aligned} \quad (3)$$

Fellegi and Sunter (1969) provided the seven equations that can be solved under this assumption in the case $K=3$ to estimate the seven parameters in (3) using the method of moments, as well as the closed form expressions for the solutions, as part of their Method II. For $K>3$, the equation system corresponding to the method of moments estimator must be solved using numerical methods.

The expectation-maximization (EM) algorithm (Dempster, Laird and Rubin 1977) can be used to estimate the parameters in (3). Jaro (1989) described estimation using EM under the independence assumption and presented empirical results. Two important extensions of the independent EM algorithm – estimation incorporating convex constraints on estimated probabilities and estimation incorporating dependencies between outcomes of comparisons for different data items – were introduced by Winkler (1989). The convex constraints were designed to constrain estimates towards values that were expected to produce good decision rules in practice. Dependencies were introduced using loglinear structures that included all three-way interactions. Armstrong and Mayda (1993) noted that loglinear structures could be used to introduce more general dependencies. They described the EM algorithm for parameterizations of $\log(P(\mathbf{x}|M))$ and $\log(P(\mathbf{x}|U))$ using hierarchical loglinear models defined by selected interactions and provided some empirical results. Armstrong and Mayda used the same loglinear structure for true matches and true non-matches. Thibaudeau (1993) worked with a model that included specific sets of interactions for true non-matches and allowed for independence among true matches.

Another important generalization of the independence model can be motivated by the observation that the set of all record pairs for a linkage of records on individuals can be partitioned into a set representing different individuals at the same address and a set representing different individuals at different addresses as well as a set representing the same individuals. Winkler (1992) applied the EM algorithm to estimate the probability density function for $\mathbf{x}$ corresponding to three classes, given by

$$P(\mathbf{x}) = P(\mathbf{x}|M)P(M) + \sum_{i=1}^{2} P(\mathbf{x}|U_i)P(U_i) \quad (4)$$

where U_1 and U_2 both include true non-matches. Empirical results from EM estimation of (4) with all three-way interactions, selected interactions common to all three classes and convex constraints were given by Winkler (1995). Later authors, including Larsen

and Rubin (2000) have estimated versions of (4) and the original two-class model, (1), that include different selected interactions in each class.

These methods often cannot be directly applied in situations in which more than two outcomes are allowed for each data item compared – the large number of parameters cannot be estimated. For example, they cannot usually be directly applied when components of $\mathbf{x}$ can assume values that reflect the relative frequency of the value agreed on or when components of $\mathbf{x}$ can assume values that reflect the similarity of data items that disagree. Weighting schemes that incorporate the notions of relative frequency and similarity, as well as other types of partial agreement, can nevertheless be constructed by adjusting estimated weights for agreement and disagreement. While the adjustments required by these schemes are constrained by properties of the data and the estimated agreement and disagreement weights, they are ad hoc in the sense that they are not derived from the probability density function of $\mathbf{x}$. Winkler (1994) discusses adjustments for the similarity of strings that disagree. The same sort of adjustments were included in the empirical work reported in section 4.

2.3 Blocking and One-to-one Matching

Fellegi and Sunter (1969) recognized the importance of "blocking" data files to reduce the computations required for record linkage applications. When a file is blocked all record pairs that do not agree on a subset of variables called blocking variables are classified as non-links. Computations are reduced since the data item comparisons required to construct outcome vectors for these record pairs are no longer necessary. Use of a single set of blocking variables in a record linkage application does not present any complications for weight estimation since the data used in weight estimation can be restricted to record pairs that agree on the blocking variables. Of course, use of a single set of blocking variables has the disadvantage that any true matches that do not agree on the blocking variables are implicitly classified, incorrectly, as non-links.

In order to minimize the number of true matches that are missed due to blocking, applications often involve matching data more than once using different sets of blocking variables. Two approaches to weight estimation can be considered if more than one set of blocking variables are used. One approach is to estimate weights for outcome vectors that include a component for each data item comparison made under at least one combination of blocking variables, as well as components for the blocking variables themselves. Under this approach the weights assigned to record pairs are all directly comparable regardless of the blocking variables that were used when particular pairs were matched. The computational complexity of this approach is a disadvantage in practice.

Another approach is to estimate weights separately for each combination of blocking variables using an outcome vector that includes only the data item comparisons made for that blocking variable combination. Under this alternative, weights are estimated conditionally given agreement on blocking variables. While weight estimation is simplified for a particular combination of blocking variables, practitioners using this approach must deal with the issue of reconciling weights for record pairs classified as links or possible links under different combinations of blocking variables. The conditional approach was used in the Elections Canada record linkage application described in section 3.

When matching two files that do not contain duplicates records, it is logical to impose the constraint that each record should be included in at most one link. This constraint is imposed for matching work at Elections Canada. The requirement for one-to-one matching leads to two methodological questions. First, there is the issue of how to modify weight estimation methods to incorporate the one-to-one matching requirement. The second issue is how to solve the assignment problem that arises in one-to-one matching when a record is included in more than one record pair that is classified as a link or possible link by decision rules.

Winkler (1999) reports that weights estimated using the EM algorithm without taking the requirement for one-to-one matching into account are known to work well in one-to-one matching situations. None of the weight estimation methods described in sections 3 and 4 involve any explicit adjustment for one-to-one matching.

Relatively early software implementations of record linkage, like the CANLINK program (Hill and Pring-Mill, 1985) developed at Statistics Canada, used a "edgewise" greedy algorithm to solve the assignment problem related to one-to-one matching. This approach performed well in many record linkage applications at Statistics Canada and has been incorporated in more recent computer software (Statistics Canada, 1996). The edgewise greedy algorithm looks at all record pairs that do not satisfy the one-to-one requirement. The record pair with the highest weight is retained. All other record pairs that conflict with this best pair, meaning that they include records that are also in the best pair, are discarded.

The procedure is repeated until the one-to-one requirement is satisfied.

Jaro (1989) introduced a linear sum assignment algorithm to force one-to-one matching. The algorithm maximizes the sum of weights over all record pairs that are retained after one-to-one matching is forced. Practical experience with this procedure indicates that it can produce incorrect assignments. Suppose for example, that the record pair "John Smith" and "Jon Smith" have a weight of 20 while "John Smith" and "John Smolinski", as well as "Jon Smith" and "Jon Smolko", have weights of 11. The linear sum assignment algorithm may incorrectly retain the latter two record pairs. Winkler (1994) describes an algorithm that minimizes this type of incorrect assignment by discarding record pairs with weights below a threshold value before any assignment alternatives are considered.

3. APPLICATION – BASELINE METHODOLOGY

Large scale applications of record linkage methods at Elections Canada, conducted to maintain the National Register of Electors, are described in this section. Section 3.1 contains information about the administrative and operational context of the record linkage applications. The methods used in the production version of an important application, an update of the match of Register data with information from Canada Customs and Revenue Agency, are described in section 3.2. Conventional weight estimation methods were used. The methods described in section 3.2 constitute the baseline methodology that will be compared with methods involving statistical estimation of weights in section 4. The results from production use of the baseline methodology are reported in section 3.3.

3.1 Context

Elections Canada is the agency of the Canadian federal government that is responsible for the conduct of federal elections and referendums. The National Register of Electors is a permanent list of electors that contains name, residential address, mailing address, date of birth and gender data for Canadian citizens who are 18 years of age and older. The establishment of the National Register of Electors has permitted Elections Canada to replace the door-to-door enumerations that were once conducted during election campaigns with new procedures that give eligible voters greater opportunities to revise information on voters lists during campaigns and are more cost-effective.

The Register is maintained between electoral events using information from administrative data sources. Updates to Register data are required to reflect the demographic changes that have a continuous impact on the electoral population. The four largest categories of demographic change are elector moves (14% per year), elector deaths (1%), Canadian citizens who reach 18 years of age and qualify as electors for the first time (2%) and adults who become Canadian citizens and qualify as electors for the first time (0.7%). The Register maintenance program includes update activities that are directed against each of these change categories and the update activities for each change category rely, to some extent, on record linkage. Discussion of the application of record linkage at Elections Canada in this paper will focus on matching carried out to facilitate the processing of elector moves.

Elector address data that can be used to update the Register to reflect elector moves are obtained from Canada Customs and Revenue Agency (CCRA, formerly Revenue Canada) as well as provincial and territorial drivers license files for all jurisdictions except Alberta, Manitoba and Quebec. Data from the provincial electoral registers maintained by Quebec and British Columbia – the only provinces with registers – are also used for updating. Shortly after the establishment of the Register each of these major data suppliers was asked to establish a unique number or data transfer identifier for each of their clients that would be sent to Elections Canada each time that data were provided. Complete files of data from all of the major suppliers of elector address data were matched with Register information using record linkage methods. Those matching activities established over 34 million links between elector numbers and data transfer identifiers. The work is described in detail in Armstrong, Block and Saleh (1999).

Only data supplier records that included data transfer identifiers that had been linked to elector numbers in those initial matches were used to change the addresses of electors during maintenance of the Register between June 1997 and December 1999. Over this period, data for new individuals were added to the Register and appeared in the files received from major suppliers of address data. As a result, the proportion of elector records whose addresses could be kept up to date declined, as did the proportion of address changes received from data suppliers that could be used for updating. In early 2000 the matching, using record linkage methods, of Register records with elector numbers that had not been linked in the initial matches to data supplier records with

unlinked data transfer identifiers was initiated using December 1999 data. The discussion in this paper will focus on the match update for CCRA information.

Elections Canada receives name, address and date of birth data from CCRA for tax filers who give consent for the transfer of this information on their tax return. Consent given on a tax return for year Y that is completed early in year Y+1 is valid only until the end of year Y+1. CCRA creates two files annually for transfer to Elections Canada, in July and in December. The file sent to Elections Canada in December 1999 included records for 17.5 million tax filers who had given consent for the transfer of personal information to Elections Canada on their 1998 tax returns. It included over 83% of the 20.9 million individuals who filed 1998 tax returns in calendar year 1999. The number of individuals who file tax returns for a particular year in the following calendar year is smaller than the number of individuals represented on the CCRA database since not all individuals on the database file a return for a particular year and not all of those who file do so in the following calendar year. Almost 3.1 million CCRA records received by Elections Canada in December 1999 (17% of the file received) included data transfer identifiers that were not linked to elector numbers. At the same time, the Register contained information for about 1.5 million electors that were not linked to CCRA data transfer identifiers.

3.2 Methods

The methods described in this section were used in production to match 1.5 million records from the December 1999 version of the National Register of Electors that did not have a link to a CCRA data transfer identifier against 3.1 million CCRA records received by Elections Canada in December 1999 that did not have a link to an elector number. They are similar to those used in the initial match of Register and CCRA information described in Armstrong, Block and Saleh (1999). A review of commercially available software that was conducted before the initial Register-CCRA match had identified Automatch as the best record linkage software available for Elections Canada applications. The results from the initial match did not suggest that there was any need to revisit this decision so Automatch was also used for the match update.

Standardization of name and address information is an essential prerequisite for any record linkage application. Given names were converted to standardized "roots" using a table built by Elections Canada that included about 3,700 names. The roots were designed to remove the effects of nicknames common spelling variations and differences related to language. For example, "Liz" and "Beth" were converted to "Elizabeth", "Allen" and "Allan" were converted to "Al" and "Pierre" was converted to "Peter". Both roots and original names were later used in record linkage. Commercial software (Group 1 Software, Inc., 1997) was used to identify the components of addresses (street name, street type, municipality, etc.) when they were not already separated on data files, to correct postal codes and to convert component values to standard values used by Canada Post by correcting typographical errors and eliminating spelling variations.

Records were linked in a number of "passes" using different blocking variables and data item comparisons at each pass. Records classified as links in any pass were removed from the files being linked before they were used in subsequent passes. One-to-one matching was forced for each pass using the linear sum assignment algorithm available in the Automatch software. Postal code, as well as a data item related to name or date of birth, were typically used as blocking variables in the first two passes. These passes were followed by two passes in which the first three characters of postal code was used as a blocking variable. The final two passes did not use any address information for blocking. The use of a number of different sets of blocking variables militated against the misclassification of true matches because they did not agree on a particular set of blocking variables. Weights were estimated separately for each pass, conditional on blocking variables. Weights for record pairs classified as links were recalculated after all matching and manual review processes were completed to provide an approximate reconciliation of weights assigned during different passes.

The Automatch record linkage software does not provide the capability to estimate weights using any of the statistical methods described in section 2.2 Weights for agreement and disagreement results were calculated using conventional methods. $P(x_k)$ was used as an estimate of $P(x_k|U)$ for each data item k. Initial values for $P(x_k|M)$, established using data quality information, were refined after manual review of the classification decisions based on the corresponding weights. The refinement process relied mainly on ad hoc adjustment and further review.

The weights for agreement and disagreement results were adjusted to allow for some more complex data comparisons. The adjustment for a complex comparison for data item k, for example, involved interpolating between the weight for outcome vector $\mathbf{x}_1$ that includes agreement on data item k, and the

weight for outcome vector $\mathbf{x}_2$ that is identical to $\mathbf{x}_1$ except that it includes disagreement on item k. Interpolation adjustments for character string comparisons used the results from string proximity measures that quantify the similarities between two character strings by assigning a value between zero and one. When data item k, say family name, disagreed for a record pair but the string proximity measure of the similarities between the two family names, α, was greater than a cutoff, θ, the adjusted weight for the record pair was calculated as

$$\log(R^*) = \frac{\alpha - \theta}{1 - \theta} \log R(x_1) + \frac{1 - \alpha}{1 - \theta} \log R(x_2). \quad (5)$$

Interpolation adjustments for comparisons of numeric data items, like year of birth, were calculated analogously using absolute differences between values and a maximum difference or cutoff.

Weights for record pairs that agreed on character data items were also adjusted, during the first four passes of matching, to take the relative frequency of the value agreed into account. The Automatch implementation of frequency adjustments relies on the independence assumption, (3), as well as the use of $P(x_k)$ to estimate $P(x_k|U)$. If x_k represents the result of the comparison of family name, for example, the proportion of records with value "Smith" was used as an estimate of $P(x_k = \text{"}Smith\text{"}|U)$. More information about weight adjustments can be found in Matchware Technologies, Inc. (1998). Weight adjustments for relative frequency were not used in the final two matching passes. Most record pairs classified as links in the final two passes agreed on relatively few identifiers. If frequency weights were used, the links could not be easily separated from other record pairs that appeared to be non-links but agreed on rare values for a very small number of identifiers.

Decision rules for the first five matching passes used a single threshold value. An initial value for the threshold was chosen based on experience with similar data. The initial value was refined using ad hoc adjustments based on manual review of classification decisions. Record pairs above this threshold were classified as links and the records involved were removed from subsequent passes. The thresholds used in the first four passes were set relatively high to militate against the linking of records that could have found better matches in later passes as well as the incorrect assignments that can be generated by the linear sum assignment algorithm. In an additional effort to militate against matching errors, links from the first four passes that had relatively poor agreement on date of birth were discarded and the records included again for possible matching in passes five and six.

Two threshold values were used for pass six. A semi-automated review process was used to reclassify the possible links produced by this pass. They were divided into ten groups based on the results of data item comparisons done using a program developed to process the possible links. A sample of record pairs from each group was reviewed manually and a classification decision for the group was made based on the results of the review.

If a group was classified as links, for example, the estimated number of true non-matches in the group, obtained from the manual review, was used in error rate calculations. No attempt was made to remove all of the misclassified records through more comprehensive review. Estimates of numbers of true non-matches misclassified as links during the linkage passes were also obtained from manual review of samples. These estimates were combined with estimates of classification errors made during the semi-automated review of possible links to obtain estimates of error rates for the application.

3.3 Results

The results of the Register-CCRA match update are given in Table 1. The numbers of records and links shown for each processing step refer to the number of records that were included in the step and the number of record pairs classified as links during the step. Overall results are given on the last line of the table. Percentages are calculated with respect to the total number of elector records included in the match. Links found during the semi-automated review of possible matches are included with the links found in passes five and six.

Some comments on the relatively low linkage rates shown in Table 1, as well as the number of CCRA records that were not linked, are appropriate. There are conceptual differences between the population represented on the National Register of Electors and the population represented on the CCRA file. First, not all individuals eligible to vote in Canada file tax returns. Second, although CCRA filters the records of tax filers who give consent by age before providing information to Elections Canada to ensure that all records provided represent individuals who are 18 years of age or older, no filtering on citizenship is done. Citizenship data are not collected by CCRA. Although the guide that is distributed with tax returns indicates that Canadian citizenship is required to vote, this requirement is not mentioned on the tax return. It is reasonable to expect that some non-citizens give consent because they are not aware

that their data cannot be used by Elections Canada or because they expect that they will become citizens in the future. According to Statistics Canada's 1996 Census, over 1.6 million adult residents of Canada were not Canadian citizens in 1996. Third, over 300,000 CCRA records represented individuals who were not included in the December 1999 Register data because they had recently reached 18 years of age. Finally, comparison of counts of Register records against demographic estimates of population and citizens based on 1996 Census data suggest that over 1 million electors older than 18 years of age are not represented on the Register.

It is important to note that, after integration of the match update results with the initial match results, 95% of the elector records in the National Register of Electors have a link to a CCRA data transfer identifier as of April 2000 (excluding records that were linked in the initial match and refer to electors who are now deceased). The quality of the information in the National Register depends on this linkage rate, as well as linkage rates with other data sources, consent rates and a variety of other proportions. The 95% Register-CCRA linkage rate exceeds the target of 85% for this linkage rate that was determined before the National Register of Electors program was implemented.

4. APPLICATION – STATISTICAL WEIGHT ESTIMATION METHODS

Empirical work involving the application of statistical weight estimation methods to data from the Register-CCRA match update is reported in this section. The data used for the empirical work, common elements of the estimation methodology and the measures used to evaluate the effectiveness of the weights are described in section 4.1. The results obtained from estimating a number of models are reported in section 4.2, with discussion.

4.1 Data and Methodology

The benefits of statistical weight estimation methods for the Register-CCRA production match were investigated using the data remaining from the first four passes of production matching, after records included in links founds in the first four passes had been removed. The same blocking variables used for pass five production matching – first initial of last name and date of birth – were employed. Weights were estimated conditional on the blocking variables. Only outcome vectors for record pairs that agreed on the blocking variables were used for the modeling of $P(\mathbf{x})$. The data items compared were root name (called variable a in the presentation of models in section 4.2), middle initial (variable b), family name (c) , civic number (d), the first eight characters of street name (e) and postal code (f). While most of the links involving records with the same address had been found during the first four passes, the pass five data included some record pairs with partial name agreement and partial address agreement as well as a few pairs with strong agreement on both name and address that had not been classified as links during the first four passes due to processing errors. A sample of the data including all record pairs in about 20% of blocks was used for statistical weight estimation.

Mixture models were estimated using the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1997). For models with interactions, the general computational algorithm of Winkler (2000) was used. This algorithm is an example of the multi-cycle expectation-conditional maximization (MCECM) algorithm (Meng and Rubin, 1993). The models allowed two outcomes – agreement and disagreement – for root name, family name and postal code. An additional outcome, corresponding to a missing value for one or both records in a pair, was included for middle initial, civic number and street name so that the outcomes used for statistical weight estimation were the same as those used in the calculation of production weights. Both production weights and statistical weights for record pairs that disagreed on family name, street name or postal code (or on any combination of these variables) were adjusted using string proximity measures. For productions weights and for statistical weights based on a model without interactions the adjustments were calculated using (5) and if adjustments were needed for more than one data item in a record pair then they were calculated independently. If adjustments to a statistical weight based on a model with interactions were required for more than one data item, the adjustments calculated using (5) were multiplied by a scale factor. The factor was chosen to ensure that the weight for the record pair, after incorporating all adjustments, was always bounded above by the weight for the outcome vector in which all adjusted data items agreed (and results for other data items did not change) and below by the weight for the outcome vector in which all adjusted data items disagreed.

In order to compare weights estimated using statistical methods with production weights, the effectiveness with which the two sets of weights separated true matches and true non-matches was examined. In production, pass five matching involved

a single threshold analogous to the upper threshold of the Fellegi-Sunter decision rule. All record pairs with weight below the threshold were included in pass six. For the comparison of weights two pass five thresholds, corresponding to the upper and lower thresholds of the decision rule, τ_2 and τ_1 respectively, were set up using the production methodology for threshold determination. A third threshold τ_1^*, corresponding to an extended lower threshold, was also established. An edgewise greedy algorithm was used to force one-to-one matching. If n record pairs had production weights above a particular threshold after forcing one-to-one matching, the corresponding threshold for each set of statistical weights was set so that n record pairs had weights above the threshold after one-to-one matching was forced. This approach ensured that comparisons of methods focused on the effectiveness of the estimated weights rather than issues associated with the setting of thresholds. Samples of record pairs from each set of n pairs were reviewed manually and the proportion of each set that were true matches was estimated.

4.2 Models and Results

Results using three statistical models for weight estimation are reported in this section. Each model involved a different parameterization of the probability density for the two-class mixture, (1). The two-class model was used rather than the three-class alternative, (4), because the data files included few instances in which more than one record referred to the same household. The interactions that were included in models were chosen based on their plausibility, given the dependencies that were expected to exist in the data. No attempt was made to search over a large number of models to determine the best possible fit or to fit models that did not seem plausible a priori. This approach was chosen in an attempt to ensure that the time and effort required to estimate weights using statistical methods would not be greater than time and effort required for conventional estimation methods.

The weight comparisons were based on a sub-sample of the record pairs used for statistical weight estimation. In Table 2, the results of the comparisons have been extrapolated from the sample to all the data involved in production pass five. Results for statistical weight estimation methods are given in the final three rows of the table. The "independent - independent" weights were calculated through EM estimation of (1) with independence in each class. The weights for "independent – (a, b, c, def)" were based on EM estimation (via the MCECM algorithm) of a model with interactions for true non-matches only. For true non-matches the model included the three-way interaction between the address variables. It was hierarchical in the sense that the two way interactions de, ef, and df were also included, as well as all main effects. Since agreement on civic number and street name generally implies agreement on postal code, interactions between address variables might be expected. The weights for "(ab,c,d,e,f) – independent" were based on EM estimation of a hierarchical model with a two-way interaction for true matches. This interaction was included because transposition of first and middle names had been frequently observed among true matches. The probabilities corresponding to production weights were used as starting values for all EM estimation.

Recall that thresholds for the statistical weights were set so that the number of record pairs with weights above each threshold was equal to the number of pairs with weights above the corresponding production threshold. The estimated proportions of true matches among record pairs with weights above each threshold given in Table 2 are indicators of the effectiveness of each weight set. According to this measure, the weights based on EM estimation of the mixture model with independence for true matches and true non-matches outperform the production weights for all three thresholds. Weights from the model "independent – (a, b, c, def)" are roughly as effective as the production weights in concentrating true matches above various thresholds. Weights from the model "(ab, c , d, e, f) – independent" are also roughly equivalent to production weights for the upper thresholds and lower threshold but they are substantially more effective than production weights in concentrating true-matches above the extended lower threshold. Although the extended lower threshold produces a large number of possible links, it could be used in practice in combination with a comprehensive semi-automated review process.

5. CONCLUSIONS

Although statisticians have developed weight estimation methods based on maximum likelihood estimation of mixture models, conventional weight estimation methods are still used almost exclusively by record linkage practitioners. Conventional methods are widely available in commercial software packages while statistical methods are much less accessible. Advances in computer technology are bringing large scale record linkage applications within the reach of relatively small organizations, like Elections Canada,

that do not have extensive resources for software development.

The results reported in this paper indicate that the use of statistical weight estimation methods can provide improvements over conventional methods typically used by practitioners. The results do not demonstrate that statistical methods always yield improvements, nor can they used to provide guidelines to situations in which improvements should be expected. Nevertheless, they suggest that statistical weight estimation methods should be more widely available.

ACKNOWLEDGEMENTS

The authors would like to thank William Winkler for providing computer code for the maximum likelihood estimation of mixture models that was the basis of the mixture model estimation code used for this paper. Thanks are also due to both William Winkler and Michael Larsen for advice on mixture model estimation and to Dean Judson for comments during his discussion of the paper at the Joint Statistical Meetings that led to improvements.

REFERENCES

Armstrong, J., Block, C. and Saleh, M. (1999). Record linkage for electoral administration. Statistical Society of Canada, Proceedings of the Survey Methods Section, 57-64.

Armstrong, J.B. and Mayda, J.E. (1993). Model-based estimation of record linkage error rates. Survey Methodology, 19, 137-147.

Belin, T.R. and Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. Journal of the American Statistical Association, 90, 694-707.

Dempster, A.P., Laird, N.M. and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

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

Group 1 Software, Inc., (1997). Canadian Code-1 Plus Users Guide, Release 2.0. Lanham, Maryland.

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, 89, 414-420.

Larsen, M.D., and Rubin, D.B. (2000). Iterative automatic record linkage using mixture models. In preparation.

Hill, T., and Pring-Mill, F. (1985). Generalized Iterative Record Linkage System. In Record Linkage Techniques – 1985 (Proceedings of the Workshop in Exact Matching Methodologies Arlington, Virginia, May 9-10, 1985), eds. B. Kilss and W. Alvey. Washington, D.C.: Department of the Treasury, Internal Revenue Service, 327-333.

Matchware Technologies, Inc. (1998). Automatch: Individual matching, geocoding and file unduplicating. Kennebunk, Maine.

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., Fair, M.E. and Lalonde, P. (1992). The use of names for linking personal records. Journal of the American Statistical Association, 87, 1193-1204.

Statistics Canada (1996). GRLS v3 User Guide. Systems Development Division. Ottawa, Ontario.

Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. Survey Methodology, 19, 31-38.

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

Winkler, W.E. (1992). Comparative analysis of record linkage decision rules. Proceedings of the Survey Research Methods Section, American Statistical Association, 829-834.

Winkler, W.E. (1994). Advanced methods for record linkage. Proceedings of the Survey Research Methods Section, American Statistical Association, 467-472.

Winkler, W.E. (1995). Matching and record linkage. In Business Survey Methods, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), New York: John Wiley and Sons, Inc., 355-384.

Winkler, W.E. (1999). The state of record linkage and current research problems. Statistical Society of Canada, Proceedings of the Survey Methods Section, 73-79.

Winkler, W.E. (2000). Personal communication.

Table 1. Results of Register-CCRA Match Update

Processing Step	Elector Records	CCRA Records	Links	Estimated False Links	Estimated False Non-Links
Passes 1-4	1,585,349	3,065,197	429,221 (27.1%)	3,119 (0.2%)	Not Applicable
Passes 5-6	1,156,128	2,635,976	115,520 (7.3%)	2,268 (0.1%)	Not Applicable
Overall	1,585,349	3,065,197	544,741 (34.4%)	5,387 (0.3%)	18,338 (1.2%)

Table 2. Results of Weight Comparisons

Weights	Estimated No. of Record Pairs with Blocking Var. Agreement	Record Pairs with Weights Above Threshold (Estimated Proportion of True Matches)		
		$\tau_1 (= 20)$	$\tau_2 (= 16)$	$\tau_2^* (= 14)$
Production	9,540,000	95,000 (95.1%)	120,000 (91.6%)	182,000 (64.7%)
Independent – Independent	9,540,000	95,000 (96.5%)	120,000 (92.3%)	182,000 (67.8%)
Independent – (a, b, c, def)	9,540,000	95,000 (95.3%)	120,000 (91.4%)	182,000 (64.4%)
(ab, c, d, e, f) – Independent	9,540,000	95,000 (94.1%)	120,000 (92.8%)	182,000 (69.8%)