MATCHING AND RECORD LINKAGE

William E. Winkler

U.S. Bureau of the Census

Record linkage is used in creating a frame, removing duplicates from files, or combining files so that relationships on two or more data elements from separate files can be studied. Much of the record linkage work in the past has been done manually or via elementary but ad hoc rules. This chapter focuses on computer matching techniques that are based on formal mathematical models subject to testing via statistical and other accepted methods.

1. INTRODUCTION

Matching has a long history of uses in statistical surveys and administrative data development. A business register consisting of names, addresses, and other identifying information such as total financial receipts might be constructed from tax and employment data bases (see chapters by Colledge, Nijhowne, and Archer). A survey of retail establishments or agricultural establishments might combine results from an area frame and a list frame. To produce a combined estimator, units from the area frame would need to be identified in the list frame (see Vogel-Kott chapter).

To estimate the size of a (sub)population via capture-recapture techniques, one needs to accurately determine units common to two or more independent listings (Sekar and Deming 1949; Scheuren 1983; Winkler 1989b). Samples must be drawn appropriately to estimate overlap (Deming and Gleser 1959).

Rather than develop a special survey to collect data for policy decisions, it might be more appropriate to match data from administrative data sources. There are potential advantages. First, the administrative data sources might contain greater amounts of data and their data might be more accurate due to improvements over a period of years. Second, virtually all of the cost of the data collection would be borne by the administrative program. Third, there would be no increase in respondent burden due to a special survey. In a general context, Brackstone (1987) discusses the advantages of administrative sources as a substitute for surveys. As a possible application of matching two administrative sources, an economist might wish to link a list of companies and the energy resources they consume with a comparable list of companies and the types, quantities, and dollar amounts of the goods they produce. Methods of adjusting analyses for matching error in merged data bases are available (Neter, Maynes, and Ramanthan 1965; Scheuren and Winkler 1993).

This chapter addresses exact matching in contrast to statistical matching (Federal Committee on Statistical Methodology 1980). An *exact match* is a linkage of data for the same unit (e.g., establishment) from different files; linkages for units that are not the same occur only because of error. Exact matching uses identifiers such as name, address, or tax unit number. *Statistical matching*, on the other hand, attempts to link files that may have few units in common. Linkages are based on similar characteristics rather than unique identifying information because strong assumptions about joint relationships are made. Linked records need not correspond to the same unit.

The primary reasons computers are used for exact matching are to reduce or eliminate manual review and to make results more easily reproducible. Computer matching has the advantages of allowing central supervision of processing, better quality control, speed, consistency, and better reproducibility of results. When two records have sufficiently comparable information for making decisions about whether the records represent the same unit, humans can exhibit considerable ingenuity by accounting for unusual typographical errors, abbreviations, and missing data. For all but the most difficult situations, computerized record linkage can currently achieve results at least as good as a highly trained clerk. When two records have missing or contradictory name or address information, then the records can only be correctly matched if additional information is obtained. For those cases when additional information cannot be adjoined to files automatically, humans are often superior to computer matching algorithms because they can better deal with a variety of inconsistent situations.

The goal of this chapter is to explain how aspects of name, address, and other information in files can affect development of automated procedures. Algorithms are based on the optimal decision rules developed by Fellegi and Sunter (1969) to describe methods introduced by Newcombe (Newcombe *et al.* 1959). Record linkage involves (1) string comparator metrics, search strategies, and name and address parsing/standardization from computer science; (2) discriminatory decision rules, error rate estimation, and iterative fitting procedures from statistics; and (3) linear programming methods from operations research.

This chapter contains many examples because its main purpose is to provide background for practitioners. While proper theoretical ideas play an important role in modern record linkage, the intent is to highlight and summarize some theoretical ideas rather than present a rigorous development. Readers who are not as interested in the theory can skip all but the first three subsections of section 3. The seminal paper by Fellegi and Sunter (1969) is still the best reference on the theory and related computational methods.

1.1. **Terminology and Definition of Errors**

As much work and associated software development has been done by different groups working in relative isolation, this section gives terminology consistent with Newcombe (Newcombe et al. 1959; Newcombe 1988) and Fellegi and Sunter (1969). In the product space **A × B** of files A and B, a *match* is a pair that represents the same business entity and a *nonmatch* is a pair that represents two different entities. With a single list, a *duplicate* is a record that represents the same business entity as another record in the same list. Rather than regard all pairs in **A × B**, it may be necessary to consider only those pairs that agree on certain identifiers or *blocking criteria*.

Blocking criteria are sometimes also called pockets or sort keys. For instance, instead of making detailed comparisons of all 90 billion pairs from two lists of 300,000 records representing all businesses in a State of the U.S., it may be sufficient to consider the set of 30 million pairs that agree on U.S. Postal ZIP code. *Missed matches* are those false nonmatches that do not agree on a set of blocking criteria.

A record linkage decision rule is a rule that designates a pair either as a link, a possible link, or a nonlink. *Possible links* are those pairs for which identifying information is not sufficient to determine whether a pair is a match or a nonmatch. Typically, clerks review possible links and decide their match status. In a list of agricultural entities, name information alone is not sufficient for deciding whether "John K Smith, Jr, Rural Route 1" and "John Smith, Rural Route 1" represent the same unit. The second "John Smith" may be the same as "John K Smith, Jr" or may represent a father or grandfather. *False matches* are those nonmatches that are erroneously designated as links by a decision rule. *False nonmatches* are either (1) matches designated as nonlinks by the decision rule as it is applied to a set of pairs or (2) matches that are not in the set of pairs to which the decision rule is applied. Generally, link/nonlink refers to designations under decision rules and match/nonmatch refers to true status.

Matching variables such as common identifiers like names, addresses, annual receipts, or tax code numbers are used to identify matches. Where possible, an establishment name such as "John K Smith Company" is often parsed into components such as first name "John," initial "K", surname "Smith", and business keyword "Company". The parse allows better comparison of name information that can improve matching accuracy. Similarly, an address such as "1423 East Main Road" might be parsed into location number "1423", direction "East", street name "Main", and street type "Road." Matching variables will not necessarily uniquely identify matches. For instance, in constructing a frame of retail establishments in a city, name information such as "Hamburger Heaven" may not allow proper linkage if "Hamburger Heaven" has several locations.

The addition of address information may not help if many establishments have different addresses on different lists. In such a situation there is insufficient information to separate new units from existing units that have different mailing addresses associated with them.

Matching weight or score is a number assigned to a pair that simplifies assignment of link and nonlink status via decision rules. A procedure, or matching variable, has more *distinguishing* power if it is better able to delineate matches and nonmatches than another. *Establishment name* refers to the name associated with a business, institution, or agricultural entity.

1.2. **Improved Computer-assisted Matching Methods**

Historically, most record linkage consisted entirely of clerical procedures in which clerks reviewed lists, obtained additional information when matching information was missing or contradictory, and made linkage decisions for cases for which rules had been developed. To bring together pairs for detailed review, clerks typically reviewed listings sorted alphabetically by name or address characteristics. If a name contained an unusual typographical variation, the clerks might not find its matches. If files were large so that some matches were separated by several pages of printouts, then those matches might not be reviewed. Even after extensive training, the clerks' matching decisions were sometimes inconsistent. All work required extensive review. Each major update required training new sets of clerks.

The disadvantages of computer matching software are that its development may require person years by proficient computer scientists, and existing software may not work optimally on files having characteristics significantly different from those on which it was developed. The advantages of the automated methods far outweigh their disadvantages. First, in situations for which good identifiers are available, computer algorithms are fast, accurate, and yield reproducible results. Second, search strategies can be far faster and more effective than those applied by clerks. As an example, the best computer algorithms allow searches using spelling variations of key identifiers. Third, computer algorithms can better account for the relative distinguishing power of combinations of matching fields as input files vary. In particular, the algorithms can deal with the relative frequency that combinations of identifiers occur.

The following example describes creation of mailing lists for the U.S. Census of Agriculture in 1987 and 1992. It dramatically illustrates how enhanced computer matching techniques can reduce costs and improve quality. To produce the address list, duplicates are identified in six million records taken from 12 different sources. Absolute numbers are comparable because 1987 proportions are multiplied times the 1992 base of six million. Before 1982, listings were reviewed manually and an unknown proportion of duplicates remained in files. In 1987, the development of effective name parsing and adequate address parsing software allowed creation of an ad hoc computer algorithm for automatically designating links and creating subsets for efficient clerical review. Within pairs of records agreeing on U.S. Postal ZIP code, the ad hoc computer algorithm used a combination of surname-based information, the first character of the first name, and numeric address information to designate 6.6 percent (396,000) of the records as duplicates and 28.9 percent as possible duplicates that had to be clerically reviewed. 14,000 person hours (as many as 75 clerks for three months) were used in identifying an additional 450,000 duplicates (7.5 percent). Because many duplicates were not located, subsequent estimates based on the list may have been compromised.

In 1992, algorithms were developed that were based on the Fellegi-Sunter model and that used effective computer algorithms for dealing with typographical errors. The computer software designated 12.8 percent of the file as duplicates and another 19.7 percent as needing clerical review. 6500 person hours were needed to identify an additional 486,000 duplicates (8.1 percent). Even without further clerical review, the 1992 computer procedures identified almost as many duplicates as the 1987 combination of computer and clerical procedures. The cost of the development of the software was $110,000 in 1992. The rates of duplicates identified by computer plus clerical procedures were 14.1 percent in 1987 and 20.9 percent in 1992. The 1992 computer procedures lasted 22 days; in contrast, the 1987 computer plus clerical procedure needed three months.

As an adjunct to computer operations, clerical review is still needed for dealing with pairs having significant amounts of missing information, typographical error, or contradictory information. Even then, using the computer to bring pairs together and having computer-assisted methods of review at terminals is more efficient than review of printouts.

2. STANDARDIZATION AND PARSING OF LISTS

Appropriate parsing of name and address components is the most crucial part of computerized record linkage. Without it, many true matches would erroneously be designated as nonlinks because common identifying information could not be compared. For specific types of establishment lists, the drastic effect of parsing failure has been quantified (Winkler 1985b, 1986). DeGuire (1988) presents an overview of the ideas needed for parsing and standardizing addresses. Parsing of names requires similar ideas.

2.1. Standardization of Name and Address Components

The basic ideas of standardization are (1) to replace the many spelling variations of commonly occurring words with standard spellings such as a fixed set of abbreviations or spellings and (2) to use certain key words that are found during standardization as hints for parsing subroutines.

In standardizing names, words of little distinguishing power such as "Corporation" or "Limited" are replaced with consistent abbreviations such as "CORP" and "LTD," respectively. First name spelling variations such as "Rob" and "Bobbie" might be replaced with a consistent assumed original spelling such as "Robert" or an identifying root word such as "Robt" because "Bobbie" might refer to a woman with "Roberta" as her legal first name. The purpose of the standardization is to allow name-parsing software to work better, by presenting names consistently and by separating out name components that have little value in matching. If establishment-associated words such as "Company" or "Incorporated" are encountered, then flags are set that force entrance into different name-parsing routines than would be encountered if such names were not encountered.

Standardization of addresses operates like standardization of names. Words such as "Road" or "Rural Route" are typically replaced by appropriate abbreviations. For instance, when a variant of "Rural Route" is encountered, a flag is set that forces parsing into a set of routines different from the set of routines associated with a house-number/street-name type of address. If reference lists containing city, state or province, and postal code combinations are available from national postal services or other sources, then, say, city names in address lists can be placed in a form that is consistent with the reference list.

2.2. **Parsing of Name and Address Components**

Parsing divides a free-form name field into a common set of components that can be compared. Parsing algorithms often use hints based on words that are standardized. For instance, words such as "CORP" or "CO" might cause parsing algorithms to enter different subroutines than words such as "MRS" or "DR".

[Table 1 about here]

In the examples of Table 1, the word "Smith" is the name component with the most identifying information. PRE refers to a prefix, POST1 and POST2 refer to postfixes, and BUS1 and BUS2 refer to commonly occurring words associated with businesses. While exact, character-by-character comparison of the standardized but unparsed names would yield no matches, use of the subcomponent last name "Smith" might help designate some pairs as links. Parsing algorithms are available that can deal with either last-name-first types of names such as "John Smith" or last-name-last types such as "Smith, John." None are available that can accurately parse both types of names in a single file.

Humans can easily compare many types of addresses because they can associate corresponding subcomponents in free-form addresses. To be most effective, matching software requires corresponding address subcomponents in identified locations. As the examples in Table 2 show, parsing software divides a free-form address field into a set of corresponding subcomponents that are in identified locations.

[Table 2 about here]

2.3. **Examples of Names**

The main difficulty with business names is that even when they are properly parsed, the identifying information may be indeterminate. In each example of Table 3, the pairs refer to the same business entities that might be in a survey frame. Alternatively, in Table 4, each pair refers to different business entities that have name subcomponents that are similar.

[Tables 3 & 4 about here]

Because the name information in Tables 3 and 4 may not be sufficient for accurately determining match status, address information or other identifying characteristics may have to be obtained via clerical review. If the additional address information is indeterminate, then at least one of the establishments in each pair may need to be contacted.

3. **MATCHING DECISION RULES**

For many projects, automated matching decision rules have often been developed using ad hoc, intuitive approaches. For instance, the decision rule might be:

```
If the pair agrees on a specific three characteristics or agrees on four or more within a set of
   five characteristics, designate the pair as a link;
else if the pair agrees on a specific two characteristics, designate the pair as a possible link;
else designate the pair as a nonlink.
```

Ad hoc rules are easily developed and may yield good results. The disadvantage is that ad hoc rules may not be applicable to pairs that are different from those used in defining the rule. Users seldom evaluate ad hoc rules with respect to false match and false nonmatch rates.

In the 1950s, Newcombe (1959) introduced concepts of record linkage that were formalized in the mathematical model of Fellegi and Sunter (1969). Computer scientists independently rediscovered the model (Cooper and Maron 1979; Van Rijsbergen *et al.* 1981; Yu *et al.* 1982) and showed that the decision rules based on the model work best among a variety of rules based on competing mathematical models. The ideas of Fellegi and Sunter are a landmark of record linkage theory because they introduced many ways of computing key parameters needed for the matching process. Their paper (1) provides methods of estimating outcome probabilities that do not rely on intuition or past experience, (2) gives estimates of error rates that do not require manual intervention, and (3) yields automatic threshold choice based on estimated error rates.

In my view, the best way to build record linkage strategies is to start with the formal mathematical techniques based on the Fellegi-Sunter model and to make (ad hoc) adjustments only as necessary. The adjustments may be likened to the manner in which early regression procedures were informally modified to deal with outliers and colinearity.

3.1. **Crucial Likelihood Ratio**

The record linkage process attempts to classify pairs in a product space **A × B** from two files A and B into M, the set of true matches, and U, the set of true nonmatches. Fellegi and Sunter (1969), making rigorous concepts introduced by Newcombe (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U) \qquad (1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur.

3.2. **Theoretical Decision Rule**

The decision rule is given by:

If $R > UPPER$, then designate pair as a link.

If $LOWER \leq R \leq UPPER$, then designate pair as a possible link and hold for clerical review. (2)

If $R < LOWER$, then designate pair as a nonlink.

The cutoff thresholds *UPPER* and *LOWER* are determined by a priori error bounds on false matches and false nonmatches. Rule (2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (1) would be small.

Fellegi and Sunter (1969, Theorem) showed that the decision rule is optimal in the sense that for any pair of fixed upper bounds on the rates of false matches and false nonmatches, the clerical review region is minimized over all decision rules on the same comparison space Γ. The theory holds on any subset such as pairs agreeing on a postal code, on street name, or on part of the name field. Ratio R or any monotonely increasing transformation of it (such as given by a logarithm) is defined as a matching weight or *total agreement weight*. In actual applications, the optimality of the decision rule (2) is heavily dependent on the accuracy of the estimates of the probabilities given in (1). The probabilities in (1) are called *matching parameters*.

3.3. **Basic Parameter Estimation under the Independence Assumption**

Fellegi and Sunter (1969, pp. 1194-1197) were the first to observe that certain parameters needed for decision rule (2) could be obtained directly from observed data if certain simplifying assumptions were made. For each $\gamma \in \Gamma$, they considered

$$P(\gamma) = P(\gamma \mid M)\ P(M) + P(\gamma \mid U)\ P(U) \tag{3}$$

and noted that the proportion of pairs having representation $\gamma \in \Gamma$ could be computed directly from available data. If $\gamma \in \Gamma$ consists of a simple agree/disagree (zero/one) pattern associated with three variables satisfying the conditional independence assumption that there exist vector constants (marginal probabilities) $m \equiv (m_1, m_2, \cdots, m_K)$ and $u \equiv (u_1, u_2, \cdots, u_K)$ such that, for all $\gamma \in \Gamma$,

$$P(\gamma \mid M) = \prod_{i=1}^{K} m_i^{\gamma^i} (1-m_i)^{(1-\gamma^i)}$$

and (4)

$$P(\gamma \mid U) = \prod_{i=1}^{K} u_i^{\gamma^i} (1-u_i)^{(1-\gamma^i)} .$$

then Fellegi and Sunter provide the seven solutions for the seven distinct equations associated with (3).

If $\gamma \in \Gamma$ represents more than three variables, then it is possible to apply general equation-solving techniques such as "the method of moments" (e.g., Hogg and Craig 1973, pp. 205-206). Because the "method of moments" has shown numerical instability in some record linkage applications (Jaro 1989, p. 417) and with general mixture distributions (Titterington *et al.* 1988, p. 71), maximum-likelihood-based methods such as the Expectation-Maximization (EM) algorithm (Dempster *et al.* 1977; also Wu 1983; Meng and Rubin 1993) may be used.

The EM algorithm has been used in a variety of record linkage situations. In each, it converged rapidly to unique limiting solutions over different starting points (Thibaudeau 1989; Winkler 1989a, 1992). The major difficulty with the parameter-estimation techniques (EM or an alternative such as method of moments) is that they may yield solutions that partition the set of pairs into two sets that differ substantially from the desired sets of true matches and true nonmatches. In contrast to other methods, the EM algorithm converges slowly and is very stable numerically (Meng and Rubin 1993).

3.4. **Adjustment for Relative Frequency**

Newcombe introduced methods for using the specific values or relative frequencies of occurrence of fields such as surname (Newcombe *et al.* 1959). The intuitive idea is that if surnames such as "Vijayan" occur less often than surnames such as "Smith", then "Vijayan" should have more distinguishing power. A variant of Newcombe’s ideas was later mathematically formalized by Fellegi and Sunter (1969; also see Winkler 1988, 1989c for extensions). Copas and Hilton (1990) introduced a new theoretical approach that, in special cases, has aspects of the approach of Newcombe but has not yet applied in a record linkage system. While the value-specific approach can be used for any matching field, it necessitates making strong assumptions about the independence between agreement on specific value states of one field and agreements on other fields.

The ideas of Fellegi and Sunter (1969, pp. 1192-1194) help describe the situation more exactly. To simplify the ideas, files A and B are assumed to contain no duplicates. The true frequencies of specific values of a string such as first name in files A and B, respectively, are given by

$$f_1 , f_2 , \cdots, f_m \; ; \;\; \Sigma_{j=1}^{m} \;\; f_j = N_A$$

and

$$g_1 , g_2 , \cdots, g_m \; ; \;\; \Sigma_{j=1}^{m} \;\; g_j = N_B.$$

If the mth string, say "Smith", occurs f_m times in A and g_m times in B, then pairs agree on "Smith" $f_m \, g_m$ times in **A × B**. The corresponding true frequencies in M are given by

$$h_1, h_2, \cdots, h_m ; \ \Sigma_{j=1}^{m} h_j = N_M,$$

and note that $h_j \leq min\ (f_j, g_j), j = 1, 2, \cdots, m$. For some implementations, h_j is assumed to equal the minimum, that *P(agree jth value of string | M)* $= h_j / N_M$, and that *P(agree jth value of string | U)* $= (f_j\, g_j - h_j)/(N_A \cdot N_B - N_M)$. In practice, observed values rather than true values must be used. The variants of how frequencies h_j are computed involve slight differences in how typographical errors are modeled, what simplifying assumptions are made, and how frequency weights are scaled to simple agree/disagree probabilities (Newcombe 1988; Fellegi and Sunter 1969; Winkler 1988, 1989c). As originally shown by Fellegi and Sunter (1969, pp. 1192-1194) the scaling can be thought of as a means of adjusting for typographical error; the scaling is

$$P(agree\ on\ string \mid M) = \Sigma_{j=1}^{m} P(agree\ on\ jth\ value\ of\ string \mid M),$$

where the probability on the left is estimated via the EM algorithm or another method. With minor restrictions, the ideas of Winkler (1989c) include those of Fellegi and Sunter (1969, pp. 1192-1194), Newcombe (1988, pp. 88-89), and Rogot *et al.* (1986) as special cases.

In some situations (Winkler 1989c) the frequency tables are created "on-the-fly" using the files actually being matched; in others, the frequency tables are created a priori using large reference files. The advantage of "on-the-fly" tables is that they can utilize relative frequencies in different geographic regions; for instance, Hispanic surnames in Los Angeles, Houston, or Miami or French surnames in Montreal. The disadvantage of "on-the-fly" tables is that they must be based on files that cover a high percentage of a target population. If the data files contain samples from a population, then the frequency weights should reflect the appropriate frequencies in the population. For instance, if two small lists of companies in a city are used, and "George Jones, Inc" occurs once on each list, then a pair should not be designated as a link using name information only. Corroborating information such as address should also be used because the name "George Jones, Inc" may not uniquely identify the establishment in the city.

3.5. **Jaro String Comparator Metrics for Typographical Error**

Jaro (1989) introduced methods for dealing with typographical error such as "Smith" versus "Smoth". Jaro's procedure consists of two steps. First, a string comparator returns a value based on counting insertions, deletions, transpositions, and string length. Second, the value is used to adjust a total agreement weight downward toward the total disagreement weight. Jaro's string comparator was extended by causing agreement in the first few characters of a string to be more important than agreement on the last few (Winkler 1990b). As Table 5 illustrates, the original Jaro comparator and the Winkler-enhanced comparator yield a more refined scale for describing the effects of typographical error than certain standard methods in computer science such as the Damerau-Levenstein metric (Winkler 1985a, 1990b).

[Table 5 about here.]

Jaro's original weight-adjustment strategy was based on a single adjustment function that he developed via ad hoc methods. Using calibration files having true matching status, I extended Jaro's strategy by applying crude statistical curve fitting techniques to define several adjustment functions. Different curves were developed for first names, last names, street names, and house numbers. When used in actual matching contexts, the new set of curves and enhanced string comparator improved matching efficacy when compared to the original Jaro methods (Winkler 1990b). With general business lists, the same set of curves could be used or new curves could be developed. In a very large experiment using files for which true matching status was known, Belin (1993) examined effects of different parameter-estimation methods, uses of value-specific weights, applications of different blocking criteria, and adjustments using different string comparators. Belin demonstrated that the original Jaro string comparator and the Winkler extensions were the two best ways of improving matching efficacy in files for which identifying fields had significant percentages of minor typographical error.

3.6. **General Parameter Estimation**

Two difficulties arise in applying the EM procedures of Section 3.3. The first is that the independence assumption is often false (Smith and Newcombe 1975; Winkler 1989b). The second is that, due to model misspecification, the EM or other fitting procedures may not naturally partition the set of pairs into the desired sets of matches M and nonmatches U.

To account for dependencies between the agreements of different matching fields, an extension of an EM-type algorithm due to Haberman (1975; see also Winkler 1989a) can be applied. Because there are many more parameters associated with general interaction models than are associated with independence models, it may be possible to fit only a fraction of all interactions. For instance, if there are ten matching variables, there are only sufficient degrees of freedom to fit all 3-way interactions (e.g., Bishop *et al.* 1975; Haberman 1979); with fewer matching variables, it may be necessary to fit various subsets of the 3-way interactions.

To address the natural partitioning problem, **A** × **B** is partitioned into three sets of pairs C_1, C_2, and C_3 using an equation analogous to (3). The EM procedures are then divided into *3-class* or *2-class* procedures. When appropriate, two of the three classes are combined into either a set that represents M or U. The remaining class represents the complement. When information from both names and addresses is used for matching, the 2-class EM tends to divide a set of pairs into those agreeing on address information and those not. If address information associated with many pairs is indeterminate (e.g., Rural Route 1 or Highway 65 West), the 3-class EM can yield a proper partition because it tends to divide the set of pairs into (1) matches at the same address, (2) nonmatches at the same address, and (3) nonmatches at different addresses.

The general EM algorithm is far slower than the independent EM algorithm because the M-step is no longer in closed form. Convergence is speeded up by using variants of the Expectation-Conditional Maximization (ECM) and Multi-Cycle ECM (MCECM) Algorithm (Meng and Rubin 1993; Winkler 1989a). The major difficulty with the general EM procedures is that different starting points will often yield different limiting solutions. It has been observed, however, that if the starting point is relatively close to the solution given by the independent EM algorithm, then the limiting solution is generally unique (Winkler 1992). The independent EM algorithm often provides starting points that are suitable for the general EM algorithm.

The dramatic improvement that the automatic EM-based parameter-estimation procedures can yield is illustrated in Figures 1-8. As there are no available business files for which true matching status is known, files of individuals having name, address, and demographic characteristics such as age, race, and sex are used. Each figure contains a plot of the estimated cumulative distribution curve via equation (2) versus the truth that is given by the 45 degree line. Figures 1-4 for matches and Figures 5-8 for nonmatches successively display, fits according to (1) iterative refinement (e.g., Newcombe 1988, pp. 65-66), (2) 3-class independent EM, (3) 3-class, selected interaction EM, and (4) 3-class 3-way interaction EM with convex constraints. Iterative refinement involves the successive manual review of sets of pairs and the reestimation of probabilities given a match under the independence assumption. Iterative refinement is chosen as a reference point (Figures 1 and 4) because it is known to yield reasonably good matching decision rules (e.g., Newcombe 1988; Winkler 1990b). The algorithm for fitting selected interactions is due to Armstrong (1992). The EM algorithm with convex constraints that predispose a solution to the proper subregion of the parameter space is due to Winkler (1989a, also 1992, 1993b). All 3-way interactions are used in the last model.

[Figures 1-8 about here.]

The basic reason that iterative refinement and 3-class independent EM perform poorly is that independence does not hold. The reason that 3-class independent EM yields results that are closer to the truth is that it divides the set of pairs that agree on address into those agreeing on name and demographic information and those that do not. Thus, nonmatches such as husband-wife and brother-sister pairs are separated from matches such as husband-husband and wife-wife. As shown by Thibaudeau (1993) with this data, departures from independence are moderate among matches while departures from independence among nonmatches (such as the husband-wife and brother-sister pairs at the same address) are quite dramatic.

The reason that the selected interaction EM does well (Figures 3 and 7) is that the true matching status is used to determine the most important set of interactions that must be included. It is unreasonable to expect that true matching status will be available for many matching situations or that the exact set of interactions that were developed for one application will be suitable for use in another. Furthermore, loglinear modelling in latent-class situations is more difficult than with basic loglinear situations where such modelling is known to be difficult (e.g., Bishop *et al.* 1975). To alleviate the situation, it may be suitable to take a model having all 3-way interactions and use convex constraints that bound some of the probabilities. The bounds would be based on similar matching situations. It is known that the all 3-way interaction model without convex constraints does not provide accurate fits (Winkler 1992). If the convex constraints are chosen properly, then the 3-way interaction EM with convex constraints provides fits (Figures 4 and 8) that are nearly as good as those obtained with the selected interaction EM (Winkler 1993b).

4. **EVALUATING THE QUALITY OF LISTS**

The quality of lists is primarily determined by how useful common variables are for matching purposes. If files are large, then the first concern is how effective common identifiers (blocking criteria) are at reducing the set of pairs to a manageable size. The effectiveness of a set of blocking criteria is also determined by the estimated number of missed matches. Applying a greater number of matching variables generally improves matching efficacy. Name information generally provides more distinguishing power than receipts, sales, or address information. Parameter estimates must be as good as possible. Improving parameter estimates can reduce clerical review regions by as much as 90 percent.

4.1. **Quality of Blocking Criteria**

While use of blocking criteria can facilitate the matching process by greatly reducing the number of pairs that are considered, it can raise the number of false nonmatches because some pairs do not agree on the blocking criteria. The following describes an investigation of how well different sets of blocking criteria yielded sets of pairs containing all matches (Winkler 1985b, 1984). The sets of pairs were constructed from 11 Energy Information Administration (EIA) lists and 47 State and industry lists containing 176,000 records. Within the set of pairs from the original set of files, name and address information allowed 110,000 matches to be identified. From the remaining 66,000 records, there were 3,050 matches having somewhat similar names and addresses and 8,510 matches having either a different name or a different address. The remaining 11,560 matches (0.18 of 66,000 records) were identified via intensive manual review and were used in analyzing various blocking criteria. In the subsequent analysis, only the 3,050 matches having similar names and addresses are considered.

In each of the following blocking criteria, NAME represents an unparsed name field. Only the first few characters from different fields are used (Table 6).

[Table 6 about here.]

These criteria were the best subset of several hundred criteria that were considered for blocking a list of sellers of petroleum products (Winkler 1984).

[Table 7 about here.]

The results (Table 7) illustrate that for certain sets of lists it is quite difficult to produce groups of blocking criteria that will give a set of pairs that include all the matches. With the union of pairs based on the best two sets of criteria, 15.1 percent of the matches are dropped from further consideration; with three, 3.7 percent. The last (fifth) criterion is not too useful because it enlarges the set of pairs with 16 additional matches and 4363 additional nonmatches.

4.2. **Estimation of False Nonmatches Not Agreeing on Multiple Blocking Criteria**

If estimates of the numbers of missed matches are needed, then lists can be sampled directly. Even with very large sample sizes, estimated variances of the error rate estimates will often exceed the estimates (Deming and Gleser 1959).

If samples are not used, then, following the suggestion of Scheuren (1983), capture-recapture techniques as in Sekar and Deming (1949; see also Bishop *et al.* 1975, Chapter 6) can be applied to the set of pairs captured by the first four sets of blocking criteria of Section 4.1 (Winkler 1987). The best-fitting loglinear model yields the 95 percent confidence interval (27,160). The interval, which represents between 1 percent and 5 percent of true matches, contains the 50 matches that were known to be missed by the blocking criteria and found via intense clerical review.

4.3. **Number of Matching Variables**

As the number of matching variables increases, the ability to distinguish matches usually increases. For instance, with name information alone, it may only be feasible to create subsets of pairs that are held for clerical review. With name and address information, a substantial number of the matches could be correctly distinguished. With name, address, and financial information (such as receipts or income) it may be possible to distinguish most matches automatically.

The exceptions occur if some of the matching variables have extreme typographical variation and/or are correlated with other matching variables. For instance, consider the following. Two name fields are available for each of the pairs. The first is a general business name that typically agrees among matches. The second name field in one record corresponds to the owner of a particular business license (e.g., in some States, all fuel storage facilities must be licensed) and in the other record, the name field corresponds to the accounting entity that keeps financial records. While the owner of a particular business license will sometimes correspond to the financial person (owner of a gasoline service station), the two names will often disagree among true matches. When both name fields are used in software that assumes that agreements are uncorrelated, contradictory information can cause loss of distinguishing power. Expedient solutions are to drop the contradictory information in the second name field or to alleviate the problem via custom software modifications.

4.4. **Relative Distinguishing Power of Matching Variables**

Without a unique identifier such as a verified U.S. Employer Identification Number (EIN), the name field typically has more distinguishing power than other fields such as address. The ability of name information to distinguish pairs can vary dramatically from one set of pairs to another. For instance, in one situation, properly parsed name information, when combined with other information, may allow good automatic decision rules; in other situations it may not.

As an example of the first situation, consider the 1992 U.S. Census of Agriculture in which name parsing software was optimized to try to find surnames (or suitable surrogates) and first names. Because the overwhelming majority of farming operations have names of the form given in Table 8,

[Table 8 about here.]

the resultant parsed names will likely all have "Smith" as a surname that will yield good distinguishing power when combined with address information. The exception can occur when two names containing "Smith" have the same address. A similar situation occurs with the 1992 match of the Standard Statistical Establishment List (SSEL) of U.S. businesses with a list of small nonemployers from an Internal Revenue Service (IRS) 1040C file of records for which the EIN is not available.

General business lists can signify the second situation because of the way in which the name field can be represented. For instance, the same business entity may appear in the following forms given in Table 9:

[Table 9 about here.]

Even if name parsing software can properly represent subcomponents of the names, it may be difficult to use the subcomponents to distinguish matches. If the name information and clerical-review status were retained, then clerical review could be reduced during future updates. Each business could be represented by a unique record that has pointers to significant name variations of matches and nonmatches along with match status. If a potential update record is initially designated as a possible link because of a name variation, then the associated name variations could be searched to decide whether a record with a name similar to the potential update record had previously been clerically reviewed. If it had, then the prior followup results could be used to determine whether the new record is a match.

4.5. **Good Matching Variables but Unsuitable Parameter Estimates**

Even when name and other matching variables can be properly parsed and have agreeing subcomponents, automatic parameter estimation software may not yield good parameter estimates because lists have little overlap or because model assumptions in parameter-estimation software are incorrect. In either situation, matching parameters are usually estimated via an iterative procedure involving manual review. Generally, matching personnel start with an initial set of parameters. The personnel review a moderately large sample of matching results and estimate new parameters via ad hoc means. The review-reestimation process is repeated until matching personnel are satisfied that parameters and matching results will not improve much.

The most straightforward means of parameter-reestimation is the iterative refinement procedure of Statistics Canada (e.g., Newcombe 1988, pp. 65-66; Statistics Canada 1983; Jaro 1992). After each review and clerical resolution of match results, marginal probabilities given a match are reestimated and matching (under the independence assumption) is repeated. Marginal probabilities given a nonmatch are held as constant because they are approximated by probabilities of random agreement over the entire set of pairs. If the proportion of nonmatches within the set of pairs is very high, then the random-agreement approximation is valid because decision rules using the random agreement probabilities are virtually the same as decision rules using true marginal probabilities given a nonmatch.

For the 1992 U.S. Census of Agriculture, initial estimates obtained via the independent EM algorithm were replaced by refined estimates that accounted somewhat for lack of independence. The refined estimates were determined by reviewing a large sample of pairs, creating adjusted probability estimates, and repeating the process. For instance, if two records simultaneously agreed on surname and first name, their matching weight was adjusted upward from the independent weight.

5. **ESTIMATION OF ERROR RATES AND ADJUSTMENT FOR MATCHING ERROR**

Fellegi and Sunter (1969, pp. 1194-1197) introduced methods for automatically estimating error rates when the conditional independence assumption (4) is valid. Their methods did not involve sampling and, as they indicated, can be extended to more general situations. This section provides methods for estimating of error rates within a set of pairs that are different than the one given in subsection 3.6. Estimation of false nonmatches due to pairs missed because of disagreement on blocking criteria was covered in Section 4. The final part of this section describes new work that investigates how statistical analyses can be adjusted for matching error.

5.1. **Sampling and Clerical Review**

Estimates of the number of false matches and nonmatches can be obtained by reviewing a sample of pairs designated as links and nonlinks. Sample size can be minimized by concentrating the sample in weight ranges in which error is likely to have taken place. With a weighting strategy that yielded good distinguishing power with decision rule (2), most of the error among computer-designated links and nonlinks will occur among weights that are close to the thresholds *UPPER* and *LOWER*. Within the set of possible links that were clerically designated as links and nonlinks, simple random samples can be used. While the amount of manual review needed for confirming or correcting the link-nonlink designations can require substantial resources, reasonable estimates within the fixed set of pairs can be obtained.

An alternative to sampling is to develop effective statistical models that allow automatic estimation of error rates. At present, such methods are the subject of much research and should show improvements in the future.

5.2. **Rubin-Belin Estimation**

Rubin and Belin (1991) developed a method of estimating matching error rates when the curves (ratio *R* versus frequency) for matches and nonmatches are somewhat separated and the failure of the independence assumption is not too severe. Their method is applicable to weighting curves *R* obtained via a 1-1 matching rule (Jaro 1989) and to which a number of ad hoc adjustments are made (Winkler 1990b). The 1-1 matching rule can dramatically improve matching performance because it can eliminate nonmatches such as husband-wife or brother-sister pairs that agree on address information. Without 1-1 matching, such pairs receive sufficiently high weights to be designated as potential links.

To model the shape of the curves of matches and nonmatches, Rubin and Belin require true matching status for a representative set of pairs. For a variety of basic settings, the procedure (Rubin and Belin 1991; Scheuren and Winkler 1993; Winkler 1992) yields reasonably accurate estimates of error rates and is not highly dependent on a priori curve shape parameters. The SEM algorithm of Meng and Rubin (1991) is used to get 95 percent confidence intervals for the estimates.

While the Rubin-Belin procedures were developed using files of individuals (for which true match statuses were known), I expect the procedures also to be applicable to files of businesses. When 1-1 matching is used, the method of Rubin and Belin can give better error rate estimates than a modified version of the method of Winkler given in subsection 3.6 (e.g., Winkler 1992). If 1-1 matching is not used, then the method of Winkler (1992; also 1993b) can yield accurate parameter estimates while the method of Rubin and Belin can not be applied because the curves associated with matches and nonmatches are not sufficiently separated.

5.3. **Scheuren-Winkler Adjustment of Statistical Analyses**

Information that resides in separate files can be very useful for analysis and policy decisions. For instance, an economist might wish to evaluate energy policy decisions by matching a file with fuel and commodity information for a set of companies against a file with the values and types of goods produced by the companies. If the wrong companies are matched, then analyses based on the files can yield erroneous conclusions. Scheuren and Winkler (1993) introduced a method of adjusting statistical analyses for matching error. If the probability distributions for matches and nonmatches are accurately estimated, then the adjustment method is valid in simple cases where one variable is taken from each file. Accurate estimates can sometimes be obtained via the method of Rubin and Belin (1991). Empirical applications have been performed for ordinary linear regression models (Winkler and Scheuren 1991) and for simple loglinear models (Winkler 1991). Extensions to situations of more than one variable from each file are under investigation.

6. **COMPUTING RESOURCES AND AUTOMATION**

Many large record linkage projects require new software or substantial modification of existing software. The chief difficulty with these projects is developing highly skilled programmers. Few programmers have the aptitude and are given the years needed to acquire the proficiency in advanced algorithm development and with the type of multi-language, multi-machine development needed to modify and enhance existing software. For example, a government agency may utilize software that another agency spent several years developing in PL/I because PL/I is the only language their programmers know. Possibly more appropriate software written in C may not be used because the same programmers do not know how to compile and run C programs. The same PL/I programmers may not have the type of skills that allow them to make major modifications in PL/I software that they did not write or port new algorithms in other languages to PL/I.

A secondary concern is lack of appropriate, general-purpose software. In many situations for which name, address, and other directly comparable information are available, some of the existing matching software will work well provided names and addresses can be parsed correctly. Directly comparable information might consist of receipts for comparable time periods. Non-directly comparable information might consist of receipts in one source and sales in another. To use such data, custom software modifications would have to be added to software. The advantage of some of the existing software is that, without modification, it often parses a substantial percentage of the records in files.

6.1. **Need for General Name-Parsing Software and What is Available**

At present, the only general-purpose business-name-parsing software that has been used by an assortment of agencies is the NSKGEN software from Statistics Canada. The software is written in a combination of PL/I and IBM Assembly language. NSKGEN software is primarily intended to create search keys that bring appropriate pairs of records together. Because it does a good job of parsing and standardizing names, it has been used for record linkage (Winkler 1986, 1987). I recently wrote general business-name-parsing software that was used in a match of the U.S. SSEL list of business establishments with the U.S. IRS 1040C list that contains many small establishments (Winkler 1993a). The software achieves better than a 99 percent parsing rate with an error rate of less than 0.2 percent with the aforementioned lists. It has not yet been tested on a variety of general lists. The code is ANSI-standard C and, upon recompilation, runs on a number of computers. While name parsing software is written and used by commercial firms, the associated source code is generally considered proprietary.

6.2. **Need for General Address-Parsing Software and What is Available**

Statistics Canada has the ASKGEN package (again written in PL/I and IBM Assembly language) which does a good job of parsing addresses (Winkler 1986, 1987). ASKGEN has recently been superceded by Postal Address Analysis System (PAAS) software. PAAS has not yet been used at a variety of agencies but, with limitations, has been used in creating an address register for the

1991 Canadian Census. The limitations were that most of the source address lists required special preprocessing to put individual addresses in a form more suitable for input to PAAS software (Swain *et al.* 1992). In addition to working on English types of addresses, the ASKGEN and PAAS software works on French types of addresses such as "16 Rue de la Place".

At the U.S. Census Bureau, address-parsing software has been written in ANSI-standard C and, upon recompilation, currently runs on an assortment of computers. The software has been incorporated in all major U.S. Bureau of the Census geocoding systems, has been used for the 1992 U.S. Census of Agriculture, and was used in several projects involving the 1992 U.S. SSEL. As with name-parsing software, source code for commercial address-parsing software is generally considered proprietary.

6.3. **Matching Software**

At present, I am unaware of any general software packages that have been specifically developed for matching lists of businesses. While the ASKGEN and NSKGEN standardization packages were used with the Canadian Business Register in 1984, associated matching is based on search keys generated through compression and standardization of corporate names. One-to-many matches are reviewed by clerks who select the best match with the help of interactive computer software. At the U.S. Bureau of the Census, I have been involved with the development of software for large projects in which the Fellegi-Sunter model was initially used and a number of ad hoc modifications were made to deal with name-parsing failure, address-parsing failure, sparse and missing data, and data situations that were unique to the sets of files being matched. In every case, the ad hoc modifications improved matching performance substantially over performance that would have been available from the general software. The recent projects were the 1992 U.S. Census of Agriculture, the 1993 match of the SSEL file of U.S. businesses with the IRS 1040C list of nonemployers, and the 1993 matching of successive years’ SSEL files and the unduplication of individual year’s files. The latter two projects used files from 1992. A set of software for agricultural lists and several packages for files of individuals is described below.

The U.S. Department of Agriculture (USDA 1980) has a system for matching lists of agricultural businesses that was written in FORTRAN for IBM mainframes in 1979 and has never been updated. Name parsing software is available as part of the system. The software applies Fellegi-Sunter-type matching to the subsets of pairs corresponding to individuals. The remaining records that are identified as corresponding to partnerships and corporations are matched clerically when an exact character-by-character match fails.

If the pairs of businesses generally have names that allow them to be represented in forms similar to the ways that files of individuals have their names represented, then matching software (or modifications of it) designed for files of individuals can be used.

While the ASKGEN and NSKGEN packages from Statistics Canada have been given out to individuals for use on IBM mainframes, associated documentation does not cover installation or details of the algorithms. To a lesser extent, the lack of detailed documentation is also true for the USDA system. The software packages require systems analysts and matching experts for installation and use.

General matching software has only been used on files of individuals due to the difficulties of name and address standardization and consistency in establishment files. Available systems are Statistics Canada’s GRLS system (Hill 1991), the system for the U.S. Census (Winkler 1990a), Jaro’s commercial system (Jaro 1992), and University of California’s CAMLIS system. None of the systems provides name or address-parsing software. Only the Winkler system is free and, upon recompilation, runs on a large collection of computers. Source code is available with the GRLS system and the Winkler system. The GRLS system has the best documentation.

7. CONCLUDING REMARKS

This chapter provided background on how the Fellegi-Sunter model of record linkage relates to developing general automated software for establishment lists. Much of the presentation showed how a variety of existing techniques have been created to alleviate specific problems due to name and/or address parsing failure or inappropriateness of assumptions used in simplifying computation associated with the Fellegi-Sunter model.

Much research is needed to improve record linkage of establishment lists. The challenges facing agencies and individuals are great because substantial time and resources are needed for

1. creating and enhancing general name and address parsing/software,
2. performing, circulating, and publishing methodological studies, and
3. generalizing and adding features to existing matching software that improve its effectiveness when applied to general establishment lists.

ACKNOWLEDGEMENT AND DISCLAIMER

The author appreciates many useful comments by the section editor and an anonymous reviewer. The opinions expressed are those of the author and not necessarily those of the U.S. Bureau of the Census.

REFERENCES

Armstrong, J. A. (1992) "Error Rate Estimation for Record Linkage: Some Recent Developments," in *Proceedings of the Workshop on Statistical Issues in Public Policy Analysis*, Carleton University.

Belin, T. R. (1993) "Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment", *Survey Methodology*, **19**, pp. 13-29.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Brackstone, G. J. (1987), "Issues in the use of administrative records for administrative purposes," *Survey Methodology*, **13**, pp. 29-43.

Cooper, W. S., and M. E. Maron (1978), "Foundations of Probabilistic and Utility-Theoretic Indexing", *Journal of the Association for Computing Machinery*, **25**, pp. 67-80.

Copas, J. R., and F. J. Hilton (1990), "Record Linkage: Statistical Models for Matching Computer Records," *Journal of the Royal Statistical Society*, **A**, **153**, pp. 287-320.

DeGuire, Y. (1988), "Postal Address Analysis," *Survey Methodology*, **14**, pp. 317-325.

Deming, W. E., and G. J. Gleser (1959), "On the Problem of Matching Lists by Samples," *Journal of the American Statistical Association*, **54**, pp. 403-415.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, **B**, **39**, pp. 1-38.

Federal Committee on Statistical Methodology (1980), Statistical Policy Working Paper 5: Report on Exact and Statistical Matching Techniques," Washington, DC: Office Federal Statistical Policy and Standards, U.S. Department of Commerce.

Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, pp. 1183-1210.

Haberman, S. J. (1975) "Iterative Scaling for Log-Linear Model for Frequency Tables Derived by Indirect Observation," *Proceedings of the Section on Statistical Computing*, *American Statistical Association*, pp. 45-50.

Haberman, S. (1979), *Analysis of Qualitative Data*, New York: Academic Press.

Hill, T. (1991), "GRLS-V2, Release of 22 May 1991," unpublished report, Ottawa, Ontario, Canada: Statistics Canada (Available from Ted Hill, General System, R. H. Coats Bldg., 14-O, Statistics Canada, Ottawa, Ontario K1A 0T6).

Hogg, R. V., and A. T. Craig (1978), *Introduction to Mathematical Statistics*, Fourth Edition, New York, NY: J. Wiley.

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**, pp. 414-420.

Jaro, M. A. (1992), "AUTOMATCH Record Linkage System," unpublished, (Available from Mathew Jaro, 14637 Locustwood Lane, Silver Spring, MD 20905, USA).

Meng, X., and D. B. Rubin (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm," *Journal of the American Statistical Association*, **86**, pp. 899-909.

Meng, X., and D. B. Rubin (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, to appear.

Neter, J., E. S. Maynes, and R. Ramanathan, (1965), "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, **60**, pp. 1005-1027.

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

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic Linkage of Vital Records," *Science*, **130**, pp. 954-959.

Rogot, E., P. Sorlie, and N. Johnson (1986), "Probabilistic methods of matching Census samples to the National Death Index," *Journal of Chronic Disease*, **39**, pp. 719-734.

Rubin, D. B., and T. R. Belin (1991), "Recent Developments in Calibrating Error Rates for Computer Matching," *Proceedings of the 1991 Census Annual Research Conference*, pp. 657-668.

Scheuren, F. (1983), "Design and estimation for large federal surveys using administrative records," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 377-381.

Scheuren, F., and W. E. Winkler (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, **19**, pp. 39-58.

Sekar, C. C., and W. E. Deming (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, **44**, pp. 101-115.

Smith, M. E., and H. B. Newcombe (1975), "Methods of Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories," *Methods of Information in Medicine*, **14**, pp. 118-125.

Statistics Canada (1983), "Generalized Iterative Record Linkage System," unpublished report, Ottawa, Ontario, Canada: Systems Development Division.

Swain, L., J. D. Drew, B. LaFrance, and K. Lance (1992), "The Creation of a Residential Address Register for

Coverage Improvement in the 1991 Canadian Census," *Survey Methodology*, **18**, pp. 127-141.

Thibaudeau, Y. (1989), "Fitting Log-Linear Models When Some Dichotomous Variables are Unobservable," in *Proceedings of the Section on Statistical Computing, American Statistical Association*, pp. 283-288.

Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, **19**, pp. 31-38.

Titterington, D. M., A. F. M. Smith, U. E. Makov (1988), *Statistical Analysis of Finite Mixture Distributions*, New York: J. Wiley.

U.S. Department of Agriculture (1988), "Record Linkage System Documentation," unpublished report, (available from Fred Vogel, National Agricultural Statistical Service, USDA, South Bldg., 14th St. & Independence Ave., SW, Washington, DC 20250, USA).

Van Rijsbergen, C. J., D. J. Harper, and M. F. Porter (1981), "The Selection of Good Search Terms," *Information Processing and Management*, **17**, pp. 77-91.

Winkler, W. E. (1984), "Exact Matching Using Elementary Techniques," unpublished report, Washington DC: Energy Information Administration Technical Report, U. S. Dept. of Energy.

Winkler, W. E. (1985a), "Preprocessing of Lists and String Comparison," in W. Alvey and B. Kilss, (eds.) *Record Linkage Techniques- 1985*, U.S. Internal Revenue Service, Publication 1299 (2-86), pp. 181-187.

Winkler, W. E. (1985b), "Exact Matching Lists of Businesses: Blocking, Subfield Identification, Information Theory," in W. Alvey and B. Kilss (eds.) *Record Linkage Techniques- 1985*, U.S. Internal Revenue Service, Publication 1299 (2-86), pp. 227-241.

Winkler, W. E. (1986), "Record Linkage of Business Lists," unpublished report, Washington DC: Energy Information Administration Technical Report, U. S. Dept. of Energy.

Winkler, W. E. (1987), "An Application of the Fellegi-Sunter Model of Record Linkage of Business Lists," unpublished report, Washington DC: Energy Information Administration Technical Report, U. S. Dept. of Energy.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 667-671.

Winkler, W. E. (1989a), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, pp. 145-155.

Winkler, W. E. (1989b), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, **15**, pp. 101-117.

Winkler, W. E. (1989c), "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statististical Association*, pp. 778-783.

Winkler, W. E. (1990a), "Documentation of record-linkage software," unpublished report, Washington DC: Statistical Research Division, U.S. Bureau of the Census.

Winkler, W. E. (1990b), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Assn.*, pp. 354-359.

Winkler, W. E. (1991), "Error Model for Analysis of Computer Linked Files," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 472-477.

Winkler, W. E. (1992), "Comparative Analysis of Record Linkage Decision Rules," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 829-834.

Winkler, W. E. (1993a) "Business Name Parsing and Standardization Software," unpublished report, Washington, DC: Statistical Research Division, U.S. Bureau of the Census.

Winkler, W. E. (1993b), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

Winkler, W. E., and F. Scheuren (1991), "How Matching Error Effects Regression Analysis: Exploratory and Confirmatory Results," unpublished report, Washington DC: Statistical Research Division Technical Report, U.S. Bureau of the Census.

Wu, C. F. J. (1983), "On the convergence properties of the EM algorithm," *Annals of Statistics*, **11**, pp. 95-103.

Yu, C. T., K. Lam, and G. Salton, "Term Weighting in Information Retrieval Using the Term Precision Model," *Journal of the Association for Computing Machinery*, **29**, pp. 152-170.

Table 1 Examples of Name Parsing

Standardized	Parsed							
	PRE	FIRST	MIDDLE	LAST	POST1	POST2	BUS1	BUS2
1. DR John J Smith MD	DR	John	J	Smith	MD			
2. Smith DRY FRM				Smith			DRY	FRM
3. Smith & Son ENTP				Smith		Son	ENTP	

Table 2 Examples of Address Parsing

Standardized	Parsed									
	Pre2	Hsnm	Stnm	RR	Box	Post1	Post2	Unit1	Unit2	Bldg
1. 16 W Main ST APT 16	W	16	Main			ST		16		
2. RR 2 BX 215				2	215					
3. Fuller BLDG SUITE 405									405	Fuller
4. 14588 HWY 16 W		14588	HWY 16				W			

Table 3 Names Referring to the Same Business Entities

	Name	Reason
1.a. b.	John J Smith ABC Fuel Oil	One list has owner name while the other list has business entity name.
2.a. b.	John J Smith, Inc. J J Smith Enterprises	Either name may be used by the business.
3.a. b.	Four Star Fuel, Exxon Distrib. Four Star Fuel	Independent fuel oil dealer associated with major oil company.
4.a. b.	Peter Knox Dairy Farm Peter J Knox	One list has establishment name while the other has owner name.

Table 4 Names Referring to Different Businesses

	Name	Reason
1.a.	John J Smith	Similar initials or names but different companies.
b.	Smith Fuel	
2.a.	ABC Fuel	same as previous
b.	ABC Plumbing	
3.a.	North Star Fuel, Exxon Distrib.	Independent affiliate and company with which affiliated.
b.	Exxon	

Table 5 Comparison of String Comparators
Rescaled between 0 and 1

Strings		Winkler	Jaro	D-L
billy	billy	1.000	1.000	1.000
billy	bill	0.967	0.933	0.800
billy	blily	0.947	0.933	0.600
massie	massey	0.944	0.889	0.600
yvette	yevett	0.911	0.889	0.600
billy	bolly	0.893	0.867	0.600
dwayne	duane	0.858	0.822	0.400
dixon	dickson	0.853	0.791	0.200
billy	susan	0.000	0.000	0.000

Table 6 Blocking Criteria

1. 3 digits ZIP code, 4 characters NAME
2. 5 digits ZIP code, 6 characters STREET
3. 10 digits TELEPHONE
4. 3 digits ZIP code, 4 characters of largest substring in NAME
5. 10 characters NAME

Table 7 Incremental Decrease in False Nonmatches
Each Set Consists of Pairs in the Union
of Sets Agreeing on Blocking Criteria

Group of Criteria	Rate of False Nonmatches	Matches/ Incremental Increase	Nonmatches/ Incremental Increase
1	45.5	1460/ NA	727/ NA
1-2	15.1	2495/1035	1109/ 289
1-3	3.7	2908/ 413	1233/ 124
1-4	1.3	2991/ 83	1494/ 261
1-5	0.7	3007/ 16	5857/4363

Table 8 Examples of Agricultural Names

John A Smith,
John A and Mary B Smith,
John A Smith and Robert Jones
Smith Dairy Farm

Table 9 Examples of Business Names that are Difficult to Compare

John A Smith and Son Manufacturing Company, Incorporated
John Smith Co
John Smith Manufacturing
J A S Inc.
John Smith and Son

Figure 1 Estimates vs Truth
Cumulative Distribution of Matches
2 – Class, Iterative

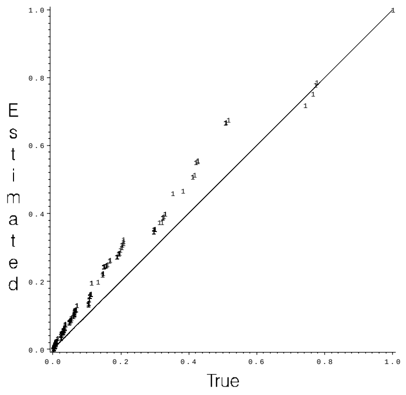


Figure 2 Estimates vs Truth
Cumulative Distribution of Matches
3 – Class, Independent EM

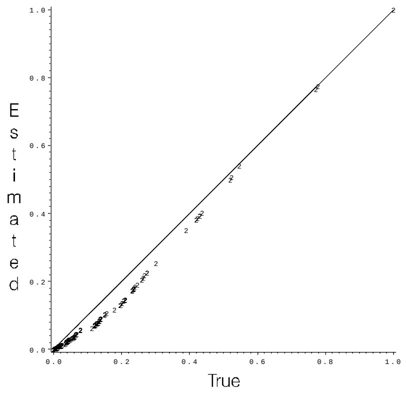


Figure 3 Estimates vs Truth
Cumulative Distribution of Matches
3 – Class, ja Interaction EM

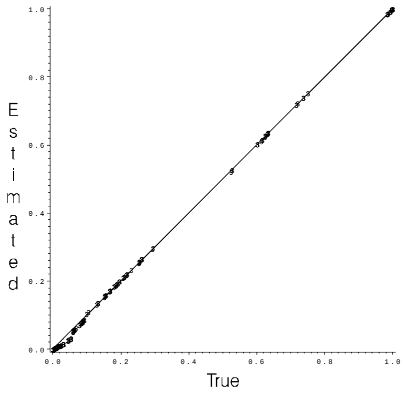


Figure 4 Estimates vs Truth
Cumulative Distribution of Matches
3 – Class, 3 – way Interaction EM, convex

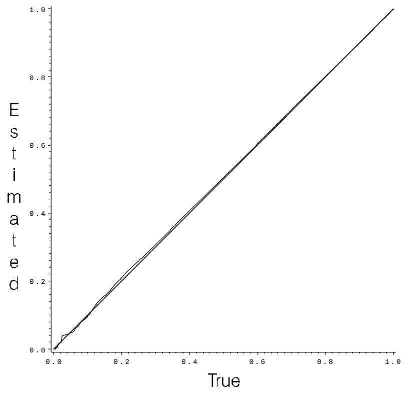


Figure 5 Estimates vs Truth
Cumulative Distribution of Nonmatches
2 – Class, Iterative

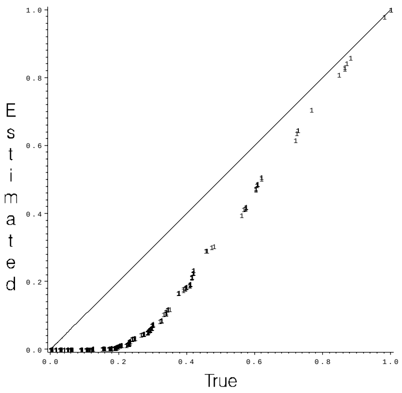


Figure 6 Estimates vs Truth
Cumulative Distribution of Nonmatches
3 – Class, Independent EM

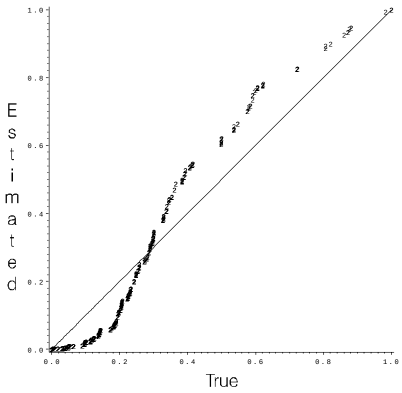


Figure 7 Estimates vs Truth
Cumulative Distribution of Nonmatches
3 – Class, ja Interaction EM

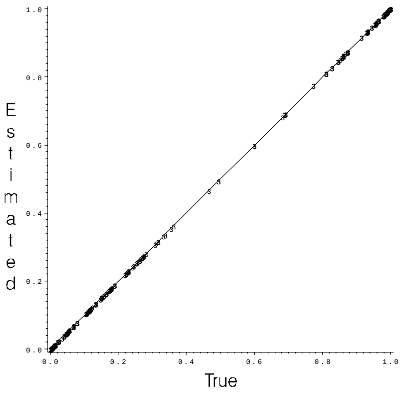
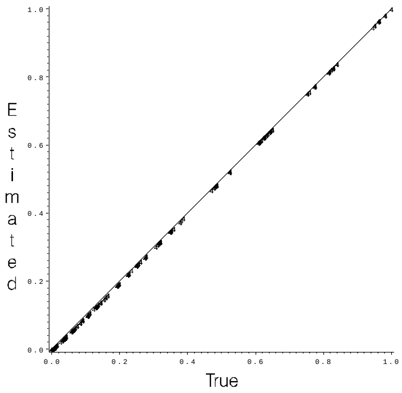


Figure 8 Estimates vs Truth
Cumulative Distribution of Nonmatches
3 – Class, 3 – way Interaction EM, Convex


E s t i m a t e d
True