The State of Record Linkage and Current Research Problems

William E. Winkler, U. S. Bureau of the Census[1]

ABSTRACT

This paper provides an overview of methods and systems developed for record linkage. Modern record linkage begins with the pioneering work of Newcombe and is especially based on the formal mathematical model of Fellegi and Sunter. In their seminal work, Fellegi and Sunter introduced many powerful ideas for estimating record linkage parameters and other ideas that still influence record linkage today. Record linkage research is characterized by its synergism of statistics, computer science, and operations research. Many difficult algorithms have been developed and put in software systems. Record linkage practice is still very limited. Some limits are due to existing software. Other limits are due to the difficulty in automatically estimating matching parameters and error rates, with current research highlighted by the work of Larsen and Rubin.

Keywords: computer matching, modeling, iterative fitting, string comparison, optimization

RÉSUMÉ

Cet article donne une vue d'ensemble sur les méthodes et les systèmes qui ont été mis en place pour le couplage d'enregistrements. Newcombe, qui développe une aproche nouvelle, et Fellegi et Sunter avec leur model mathématique, nous ont laisse les bases nécessaires pour un traitement moderne de la discipline du couplage d'enregistrement. Dans leur travail fondamental, Fellegi et Sunter ont introduit des méthodes puissantes pour l'estimation des paramètres sous-jacents, ainsi que des idées qui continuent d'influencer la pratique du couplage d'enregistrement. La recherche sur le couplage d'enregistrement se charactérise par une synergie de la statistique, de l'informatique, et de la recherche opérationnelle. Malgré l'intégragion sous formes de logiciels de plusieurs algorithmes difficiles, la pratique du couplage d'enregistrement n'en reste pas moins limitée. Cette limitation est due en partie aux defauts des logiciels eux-mêmes, mais aussi aux difficultés à estimer de facon systématique les paramètres sous-jacents ainsi que les taux d'erreurs encourues. Le problème de l'estimation automatique des taux d'erreurs encourues font l'object d'une recherche récente par Larsen et Rubin.

Mots Clés: couplage d'enregistrements, modeling, comparaison de chaîne de caractères, optimisation

1. INTRODUCTION

Record linkage is the methodology of bringing together corresponding records from two or more files or finding duplicates within files. The term record linkage originated in the public health area when files of individual patients were brought together using name, date-of-birth and other information. In recent years, advances have yielded computer systems that incorporate sophisticated ideas from computer science, statistics, and operations research. Some of the work originated in epidemiological and survey applications. Very recent work is in the related areas of information retrieval and data mining.

The ideas of modern record linkage originated with geneticist Howard Newcombe (Newcombe et al. 1959, 1962) who introduced odds ratios of frequencies and the decision rules for delineating

[1] William E. Winkler, Statistical Research Division, Room 3000-4, Bureau of the Census, Washington, DC, 20233-9100 USA, bwinkler@census.gov

matches and nonmatches. Newcombe's ideas have been implemented in software that is used in many epidemiological applications and often rely on odds-ratios of frequencies that have been computed a priori using large national health files. Fellegi and Sunter (1969) provided the formal mathematical foundations of record linkage. Their theory demonstrated the optimality of the decision rules used by Newcombe and introduced a variety of ways of estimating crucial matching probabilities (parameters) directly from the files being matched.

The outline of this paper is as follows. The second section provides more details on intuition about and the theoretical model for record linkage. Ideas of Newcombe have had the most important application in the development of national health files of individuals. The more general ideas of Fellegi and Sunter have been instrumental in estimating crucial matching parameters and estimating error rates for wide classes of lists. Methods for overcoming messy-data problems are described systematically in relation to the formal model of Fellegi and Sunter. In the third section, some of the basic research problems are covered. Although some of the problems have been (partially) solved for high quality pairs of lists, the solution methods do not easily extend to most matching situations. The fourth section describes three research areas that have arisen in recent years and depend heavily on record linkage ideas. The first is microdata confidentiality and associated re-identification methods. The second is analytic linking as introduced by Scheuren and Winkler (1993, 1997). *Analytic linking* refers to the merging and proper analysis of data (quantitative and discrete) taken from two or more files. The analysis is intended to adjust for the biases due to linkage error. The third presents some of the methods of information retrieval and machine learning as used by computer scientists in web search engines and data mining applications. Concluding remarks are given in the final section.

2. BACKGROUND ON RECORD LINKAGE

Howard Newcombe had crucial insights that led to computerized approaches for record linkage. The first was that the relative frequency of the occurrence of a value of a string such as a surname among matches and nonmatches could be used in computing a binit weight (score) associated with the matching of two records. The second was the scores over different fields such as surname, first name, age, etc. could be added to obtain an overall matching score. More specifically, he considered odds ratios

$$\log_2(p_L) - \log_2(p_F) \qquad (1)$$

where p_L is the relative frequency among links and p_F is the relative frequency among nonlinks. Since the true matching status is often not known, he suggested approximating the above odds ratio with the following ratio

$$\log_2(p_R) - \log_2(p_R)^2 \qquad (2)$$

where p_R is the frequency of a particular string (first, initial, birthplace, etc). If one matches a large universe file with itself, then the second ratio is a good approximation of the first ratio. Newcombe's ideas have been extended in a variety of ways (e.g., Newcombe et al., 1988, 1992, Gill 1999)

Fellegi and Sunter (1969) introduced a formal mathematical foundation for record linkage. To begin, notation is needed. Two files **A** and **B** are matched. The idea is to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true matches, and U, the set of true nonmatches. Fellegi and Sunter, making rigorous concepts introduced by Newcombe (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U) \quad (3)$$

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. The ratio R or any monotonely increasing function of it such as the natural log is referred to as a matching weight (or score).

The decision rule is given by:

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

If $LOWER \leq R \leq UPPER$, then designate pair as a possible match and hold for clerical review. (4)

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

The cutoff thresholds *UPPER* and *LOWER* are determined by a priori error bounds on false matches and false nonmatches. Rule (4) 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 (3) would be small.

Pairs with weights above the upper cut-off are referred to as *designated matches* (or links). Pairs below the lower cut-off are referred to as *designated nonmatches* (or nonlinks). The remaining pairs are referred to as *designated potential matches* (or potential links).

If one considers a situation where there are three matching fields and only simple agree/disagree weights are considered, then a conditional independence assumption can be made to simplify computation.

P(agree first, agree last, agree age | M)

= P(agree first | M) P(agree last | M) P(agree age | M) (5a)

Similarly,

P(agree first, agree last, agree age = | U)

= P(agree first | U) P(agree last | U) P(agree age | U) (5b)

This conditional independence assumption must hold on all combinations of fields (variables) that are used in matching. The probabilities P(agree first | M), P(agree last | M), P(agree age | M), P(agree first | U), P(agree last | U), and P(agree age | U) are called *marginal probabilities*. P(| M) & P(| U) are called the m- and u-probabilities, respectively. The natural logarithm of the ratio R of the probabilities is called the *matching weight or total agreement weight*. The logarithms of the ratios of probabilities associated with individual fields (marginal probabilities)

are called the *individual agreement weights*. The m- and u-probabilities are also referred to as *matching parameters*.

Fellegi and Sunter showed that it is possible to compute the unknown m- and u- probabilities directly in the 3-variable, conditional independence case. More generally, in the conditional independence situation, the parameters can be computed via a straightforward application of the EM algorithm (Winkler 1988). If the conditional independence assumption does not hold, then the parameters can be computed by generalized EM methods (Winkler 1988, 1989a, 1993b, Armstrong and Mayda 1993, see also Meng and Rubin 1993), by scoring (Thibaudeau 1993), and by Gibbs sampling (Larsen 1996, Larsen and Rubin 1999). The methods of Larsen and Rubin (1999) are the most general. These methods can yield more accurate matching parameters and better decision rules. These parameter-estimation methods do not always yield sufficiently accurate probability estimates for estimating record linkage error rates. An error-rate estimation method that is somewhat supplemental to these is due to Belin and Rubin (1995). Although the method of Belin and Rubin requires calibration data, it is known to work well in a narrow range of situations (Winkler and Thibaudeau, 1991; Scheuren and Winkler, 1993). The situations are those in which there is substantial separation of the curves of log frequency versus matching weight for matches and nonmatches.

Generally, good separation of curves occurs with high-quality lists of individuals containing only moderate amounts of typographical error and reasonable amounts of homogeneity in the characteristics respectively used in classifying pairs as matches and nonmatches. With some administrative lists and most agricultural and business lists, such homogeneity does not occur. For instance, if names or address do not standardize, then it is unlikely that true matches having nonstandardized names or addresses can be identified. If homogeneity holds, then most matches have similar characteristics within the group of matches. Most nonmatches have similar characteristics within the group of nonmatches. In some situations, difficulties with business lists can be dealt with via software loops that deal with list-specific nonhomogeneity. Each of the major departures from homogeneity due to severe typographical error must be dealt with via a separate software loop. Other departures from nonhomogeneity occur when either the class of matches or the class of nonmatches naturally divide into subclasses. For instance, when matching persons within household, the class of nonmatches naturally divides into those that agree on address (household characteristics) and those that do not. Some of the general methods for dealing with nonhomogeneity of identifying characteristics are described in Winkler (1993b). EM methods and ideas for dealing with one major type of nonhomogeneity similar to Winkler (1988, 1989, 1993b) have recently been applied to the general problem of text classification in machine learning and data mining by Nigam et al. (1999). The methods of Winkler are more general because they allow for dependencies of fields and convex constraints on probabilities (either class or marginal) that predispose estimates to subregions of the parameter based on prior knowledge from similar matching situations.

2.1 String Comparators

In many matching situations, it is not possible to compare two strings exactly (character-by-character) because of typographical error. Dealing with typographical error via approximate string comparison has been a major research project in computer science (see e.g., Hall and Dowling, 1980). In record linkage, one needs to have a function that represents approximate agreement, with agreement being represented by 1 and degrees of partial agreement being represented by numbers between 0 and 1. One also needs to adjust the likelihood ratios (3) according to the partial agreement values. Having such methods is crucial to matching. For instance, in a major census application for measuring undercount, more than 25% of matches

would not have been found via exact character-by-character matching. Three geographic regions are considered in Table 1. The function Φ_n represents exact agreement when it takes value one and represents partial agreement when it takes values less than one. In the St Louis region, for instance, 25% of first names and 15% of last names did not agree character-by-character among pairs that are matches.

Table 1 Proportional Agreement by
String Comparator Values
Among Matches
Key Fields by Geography

	StL	Col	Wash
First			
$\Phi_n=1.0$	0.75	0.82	0.75
$\Phi_n \geq 0.6$	0.93	0.94	0.93
Last			
$\Phi_n=1.0$	0.85	0.88	0.86
$\Phi_n \geq 0.6$	0.95	0.96	0.96

Jaro (1976, see also 1989) introduced a string comparator that accounts for insertions, deletions, and transpositions. The basic Jaro algorithm has three components: (1) compute the string lengths, (2) find the number of common characters in the two strings, and (3) find the number of transpositions. The definition of common is that the agreeing character must be within half the length of the shorter string. The definition of transposition is that the character from one string is out of order with the corresponding common character from the other string. The string comparator value (rescaled for consistency with the practice in computer science) is:

$$\Phi_j(s1,s2) = 1/3(\ \#common/str\_len1 + \#common/str\_len2 + 0.5\ \#transpositions/\#common),$$

where s1 and s2 are the strings with lengths str_len1 and str_len2, respectively.

Using truth data sets, Winkler (1990b) introduced crude methods for modeling how the different values of the string comparator affect the likelihood in the Fellegi-Sunter decision rule. Winkler also showed how a variant of the Jaro string comparator Φ_n dramatically improves matching efficacy in comparison to situations when string comparators are not used. The variant employs some ideas of Pollock and Zamora (1984) in a large study for the Chemical Abstracts Service. They provided empirical evidence about how the probability of keypunch errors increased as the character position in a string moved to the right. Budzinsky (1993) in a review of twenty string comparators concluded that the methods of Jaro and Winkler worked second best and best, respectively. The Winkler string comparator Φ_n is used in the Generalized Record Linkage System software of Statistics Canada.

2.2 Heuristic Improvement by Forcing 1-1 Matching

Jaro (1989) introduced a linear sum assignment procedure (lsap) to force 1-1 matching because he observed that greedy algorithms often made erroneous assignments. A greedy algorithm is one in

which a record is always associated with the corresponding available record having the highest agreement weight. Subsequent records are only compared with available remaining records that have not been assigned. In the following, the two households are assumed to be the same, individuals have substantial identifying information, and the ordering is as shown. A lsap algorithm causes the wife-wife, son-son, and daughter-daughter assignments correctly because it optimizes the set of assignments globally over the household. Other algorithms such as greedy algorithms can make erroneous assignments such as husband-wife, wife-daughter, and daughter-son.

```
HouseH1     HouseH2

husband
wife        wife
daughter    daughter
son         son
```

c_{11} c_{12} c_{13}
c_{21} c_{22} c_{23}
c_{31} c_{32} c_{33}
c_{41} c_{42} c_{43}

4 rows, 3 columns
Take at most one in each row and column

c_{ij} is the (total agreement) weight from matching the ith person from the first file with the jth person in the second file. Winkler (1994) introduced a modified assignment algorithm that uses 1/500 as much storage as the original algorithm and is of equivalent speed. The modified assignment algorithm does not induce a very small proportion of matching error (0.1-0.2%) that is caused by the original assignment algorithm.

2.3 Why the methods do not always work well.

The record linkage methods described above can perform well when there is little typographical variation and other forms of nonhomogeneity in the identifying characteristics of lists. The methods may not work well due to failures of the assumptions used in the models, lack of sufficient variables for matching, sampling or lack of overlap in lists, and extreme variations in the messiness of data. The idiosyncrasies of messy data are most easily described. Each of the following types of errors provides examples of situations where pairs of records will not have homogeneous identifying characteristics.

1. Records that do not address standardize.
2. Records that do not name standardize.
3. Records that have more information or missing matching variables.
4. Records that do not have easily comparable fields.

Name	Ralph Smith	R J Smith
Address	123 Main St	PO Box 9128
Age	54	50

If the PO Box address in the right-most column were replaced by a street address that corresponds almost exactly to the street address given in the second column, then it might be possible to accurately match. If R J Smith is actually Roberta Joan Smith, then the match would be in error. Inconsistencies of name and address information are typically even greater with agriculture and business lists. During name and address standardization, commonly occurring words such as

Mister, Road, Post Office Box, etc. are replaced by standardized spellings and the components of the names and addresses are placed in fixed locations. If standardization fails for a record, then automatic matching in software may be impossible. This is due to specific information needed for comparison and computing weights that is missing. If two lists of individuals are small samples, then we may not be able to match on certain commonly occurring names such as John Smith without substantial corroborating information. The difficulty of estimating the overlap of samples has most effectively been dealt with by Deming and Gleser (1959) in situations where there is no matching error. When there is matching error, the estimation can be more difficult.

3. BASIC RESEARCH PROBLEMS

The basic research problems have been open since the work of Newcombe et al. (1959) and Fellegi and Sunter (1969). Partial progress in solving the problems has occurred. The major difficulties in all situations have been determining how identifying information can be used and what the relative value of a field is in matching in comparison with other fields.

3.1 When can frequency-based matching improve over simple agree/disagree matching?

The ideas of frequency-based (value-specific) matching were introduced by Newcombe et al. (1959). Fellegi and Sunter (1969) gave two methods for computing frequency-based weights in the context of their formal model that have been extended by Winkler (1988, 1989). The basic idea is that agreements on rarely occurring values of a field (variable) are better at distinguishing matches than agreements on commonly occurring values of a field. The agreement on a rare value is also better than the general yes/no agreement (i.e., non-value-specific) on a field. For instance,

P(agree last name = ‘Zabrinsky’, agree first name ‘Zbigniew’ | M) >

P(agree last name, agree first name | M) > (6)

P(agree last name = ‘Smith’, agree first name ‘James’ | M) .

Reasonably correct frequencies are computed and used in matching. The intuition is that frequency-based weights given by the first and third probabilities in (6) are better able to delineate matches and nonmatches than the simple agree/disagree probabilities given in the second probability in (6). Names by themselves are seldom effectively used in matching. Additional fields such as components of the address, age or full date-of-birth, maiden name, sex, and race are also needed to reduce error rates to acceptable levels. In some early experiments, frequency-based matching often did better than simple agree/disagree matching. With the development of more sophisticated models for estimating agree/disagree matching parameters via the EM algorithm, simple agree/disagree weights sometimes performed better. The reason is due to the fact that, in many files, a moderate number of false matches agree on relatively rarely occurring names. In those situations, pairs that might be in the clerical review region given in (4) might move upward to the designated match region. If there is a substantial number of fields available for matching, then the redundancy provided by the extra fields can reduce matching error. If redundancy is sufficient to reduce matching error, then it seems likely that frequency-based matching is not needed. Raising the total agreement weights for pairs associated with less frequently values of a variable will not improve matching.

There are, nevertheless, a number of important situations when it is likely that frequency-based matching may be demonstrated to work at least as well as simple agree/disagree matching. The

major situations all involve large national health files that have been significantly cleaned for typographical error and for which accurate probabilities can be computed a priori using true population counts. The research question is "Are there situations for which it can be shown that frequency-based matching improves over simple agree/disagree matching?" It seems that with many business lists, agriculture lists, and general administrative lists that frequency-based matching may not yield improvements because of the large amounts of typographical variation. These lists often have moderate to large proportions of records that fail standardization, have excessively high typographical error rates, and have only moderate overlap. If any one of these three situations occurs, then frequency-based matching may be seriously compromised.

3.2 What is the best method for estimating parameters under conditional independence when non-1-1 (or 1-1) matching is done?

Parameter estimates obtained under the conditional independence EM can be superior to other parameter estimates (Winkler, 1990b) and can be obtained more easily. The conventional methods estimate the marginal probabilities P(agree field | M) and P(agree field | U) directly using samples for which truth has been obtained via possibly time-consuming manual review. The estimates are obtained more easily because the known truth of matches on subsets is not needed (Winkler, 1988). The reason that the EM parameters work better is that they effectively represent the conditional probabilities such as the following

$$P(\text{agree field 1, agree field 2, agree field 3} \mid M) = \tag{7}$$

$$P(\text{agree field 1} \mid M)\, P(\text{agree field 2} \mid \text{field 1}, M)\, P(\text{agree field 3} \mid \text{field 1, field 2}, M).$$

The EM algorithm decides what ordering of the fields in (7) is optimal in estimating the likelihoods. These probabilities implicitly perform a minor automatic adjustment for the lack of conditional independence. The EM algorithm still makes a *homogeneity* assumption because it assumes that the same ordering can be applied to all pairs conditional on whether they are a match or nonmatch. Because the EM-parameters are designed to maximize the likelihood, they produce better decision rules than the probabilities estimating under the conventional methods. The conventional parameters do not maximize the likelihood because of the strong conditional independence assumption that is made. Winkler (1990b) provided an exact comparison of decision rules using parameters obtained by the two estimation techniques. Caution in the automatic use of the EM-probabilities is needed because the EM may not exactly divide the set of pairs into two classes that correspond exactly to matches and nonmatches. The difficulty of having EM-determined classes that correspond to true matching classes has been addressed by Winkler (1993b) and by Nigam et al. (1999). The caution may not apply to conventionally estimated parameters because the clerical review can better assure that estimated parameters are consistent with model assumptions.

The EM probabilities are estimated using all pairs and often used in matching software that forces 1-1 matching. Although the mechanisms for forcing 1-1 matching are not explicitly accounted for, the probabilities are known to work well in those situations. The research question is "When can the EM-probabilities estimated under conditional independence be effectively used in 1-1 matching decision rules?" If marginal probabilities are conventionally estimated via samples, when can they be effectively used in 1-1 matching?

3.3 When does accounting for dependencies help in matching?

If conditional independence does not hold, then

P(agree first name, agree last name | M) ≠ P(agree first | M) P(agree last | M) .

Decision rules that apply probabilities estimated under the conditional independence assumption may be suboptimal. Smith and Newcombe (1975 gave a modified decision rule that adjusts for the lack of dependence that have been effectively extended and applied by others (Gill, 1999). The modified decision rules are heavily dependent on the assumption that the adjustments based on a sample for which truth is known can be used in a variety of matching situations. The assumption is likely to be reasonable in situations of large national health files for which truth is known on a large subset. Winkler (1989a), Thibaudeau (1993), Armstrong and Mayda (1993), and Larsen and Rubin (1999) have all given formal models for estimating the record linkage parameters (probabilities) under general dependence models. Winkler (1989a) also showed that the values of matching parameters vary significantly from one list to another. The variation occurs even when the lists have the same matching variables and the same amount of overlap but represent different geographic regions. All of the authors have shown that the development of appropriate dependence models takes considerable skill and suitable software. They have also shown that probabilities estimated under dependence are more accurate. None of the authors has been able to show whether the new parameter-estimation method can be assured to yield appropriately good decision rules in actual record linkage software on a day-to-day basis. A basic research question is "For what types of files and matching situations can general dependence-based probabilities and decision rules improve matching?" There is still considerable empirical evidence that matching under the conditional independence assumption is effective in practice. Winkler (1993b, 1994) demonstrated that matching under the conditional independence assumption worked nearly as well as matching under more general dependency models in certain situations. The situations included population files having multiple individuals per household in which 1-1 matching was forced. Winkler (1994) did suggest accounting for dependencies might yield better automatic estimates of error rates.

3.4 What are (suitable) ways of estimating error rates?

The method of Belin and Rubin (1995) is currently the only method for automatically estimating record linkage error rates. Belin and Rubin were able to achieve highly accurate estimates (Winkler and Thibaudeau 1991, Scheuren and Winkler 1993) in a narrow range of situations. The situations generally involved population files where there was good separation between the matching weights associated with nonmatches and matches. If there is not good separation, then methods that use more information from the matching process may ultimately yield suitable estimates in a larger range of situations as suggested by Winkler (1994) and Larsen and Rubin (1999). The estimation methods and the means of evaluating the fits of the latent class models are quite difficult because the usual Chi-square methods do not work (Rubin and Stern, 1993). The basic research question is "How does one automatically estimate error rates?"

4. ADVANCED RESEARCH PROBLEMS

Three areas use methods and underlying models that are closely related to the basic ideas of record linkage. Confidentiality of microdata is most closely related because record linkage methods can be used for evaluating the re-identification risk in public-use files. Since the quantitative data in a public-use file are typically masked, new metrics for comparing quantitative data can yield higher re-identification rates. Analytic Linking is the methodology (Scheuren and Winkler 1997) for using not directly comparable data items to improve matching and to account for the effect of matching error in analyses. For instance, if one administrative file has receipts and another has income, an additional variable, predicted income, can be added to the first file to

improve matching. The matching can also be improved by targeting outliers and systematic errors in the merged files in a manner that identifies likely false matches. Data mining and some models for information retrieval in computer science use Bayesian networks for classifying documents using free-form textual information. The representations in Bayesian networks from machine learning can be viewed as a special case of representations in the Fellegi-Sunter model. Recent advances in applying the EM algorithm in machine learning settings give insight into how to better use training data (if available), how to better structure the models, and how to use free-form text in a rigorous model.

4.1 Confidentiality

There is substantially increased need to supply researchers with large, general-purpose public-use files that can be used for a variety of analyses. Balancing the analytic needs are the requirements that agencies not release individually identifiable data. If a public-use file is created, then agencies must determine if the file meets analytic needs and is confidential. Record linkage methods (Winkler 1998) that employ new metrics for comparing somewhat related quantitative data provide a useful enhancement and yield higher re-identification rates than less sophisticated methods. If an agency can effectively determine that a small percentage of records might be re-identified, it can take additional precautions.

Methods for masking data are intended to make re-identification more difficult. Existing masking methods cover a variety of areas. Global recoding and local suppression (DeWaal and Willenborg 1996, 1998; Sweeney, 1999) have been successfully used to create public-uses files and other security procedures. The advantage of the methods is that available general software is often straightforward to apply. The associated research problems relate to how seriously analytic properties are compromised. Additive noise is known to preserve some of the analytic properties of files (Kim 1986, 1989; Fuller 1993). Research problems are whether general software can be developed and whether files are free of disclosures. Combinations of additive noise and limited swapping have been used by Kim and Winkler (1995) and Winkler (1998). Data perturbation methods (Tendick and Matloff 1994) are closely related to additive noise. The methods are good at preserving confidentiality and yielding totals on a number of subdomains that are consistent with unreleased confidential data. The basic research problems are whether the methods can be extended to preserved second order and higher statistics as the additive noise methods do. Camouflage (Gopal, Goes, and Garfinkel 1998) is a sophisticated method that returns intervals rather than point estimates for large classes of functions on arbitrary subdomains. A basic research question is whether these methods can produce the types of information that users of the public-use files need. Microaggregation (see e.g., Mateo-Sanz and Domingo-Ferrer 1998) is a method of replacing values of individual variables in ranges with means. The algorithms can be quite sophisticated. The research questions are: “Do these methods compromise analytic validity seriously?” and “What are re-identification rates with certain classes of files?” The most sophisticated methods involve models for re-identification risk and analytic properties of files. Fienberg, Makov and Sanil (1997) and Fienberg, Makov, and Steele (1998) have introduced some promising ideas that need extension to encompass different classes of data and to achieve computational tractability. With all of these methods the basic research question is “If analytic validity is preserved, then what is the re-identification rate?” If good source files for matching and suitable re-identification software are available to an intruder, then what is the re-identification rate?

4.2 Analytic Linking

Researchers often have the need to analyze large amounts of data that result from the merger of two or more administrative files in which unique identifiers are unavailable. Scheuren and Winkler (1993) showed how regression analyses might be adjusted for biases due to linkage errors. In the simplest situation of two variables, the dependent variable might be taken from one file and the independent variable from another file. If there is matching error, then the dependent and independent variables associated with false matches generally will not correspond as closely as those associated with true matches. The adjustments were highly dependent on accurate probabilities obtained by the methods of Belin and Rubin (1995). If error-rates are estimated accurately, then the bias-adjustments for matching error were reasonably accurate.

One administrative file may have a number of data fields (variables) that are correlated or otherwise related to a number of data fields in another administrative file. Scheuren and Winkler (1997) introduced analytic linking methods that place predictors in one file that can be used to improve matching with another file. After each matching pass, data are again modeled to refine the predictors. Through a series of iterations in which predictors and matching are improved, Scheuren and Winkler showed how matching could be performed in situations that were previously considered impossible. If matching error is low, then adjustment methods may not be needed (Scheuren and Winkler 1993). If matching error is moderate, then the adjustment method of Scheuren and Winkler (1993) may help. The basic research problems are "What are more generally applicable adjustments methods for matching error?" How can all of the information in two files be used? Scheuren and Winkler (1997) used simple predicted values that may not account for many types of matching error and may not be suitable as a global set of predictions. The work of Scheuren and Winkler has a strong visual component. Summary representations in graphs (images) are successively improved as erroneous data due to false matches are eliminated. Much of the erroneous data shows up as outliers that detract from the graph that would be obtained from the true model having no noise. If the underlying analytic model and the effects of some of the matching error are effectively modeled (i.e., learned), then the images associated with the process also improve. Improving the methods may involve advanced image resolution ideas (Besag et al. 1974, 1986, 1995; Geman and Geman 1984). Other improvements may be due to better modeling of the components of the iterative analytic linking process. Van Dyk (1999) has recently introduced methods for speeding up EM-type computations associated with hierarchical models that contain ideas that might improve the methods of Winkler and Scheuren. Although the specific types of speed-ups may not be needed, the insight that Van Dyk offered into modeling a large number of components of a process seems to be needed.

Winkler (1999) also indicated how large *bridging* files can be used to improve matching with two smaller files. A *bridging* file is a large universe file that approximately contains the two smaller files. Bridging files might be a large file such as the main Social Security Administration file of the U.S. population or a large credit database with associated information. Although the large bridging file does not generally have sufficient information for matching all records in the smaller files, it has sufficient information for reducing the set of potential matches to small subsets. Additional matching runs on the smaller data files can then yield higher proportions of matches. The research question is "What are effective ways of using bridging files?" Bridging files also should have significant power for improving re-identification experiments.

4.3 Data Mining

Machine learning algorithms that employ Bayesian networks are tools being applied to classify text into different groups. Bayesian networks are one of the standard tools in data mining. They are also used for information retrieval methods such as used in some of the web search engines. The latest algorithms (Nigam et al., 1999) utilize EM-based methods that are closely related to

methods used by Winkler (1988, 1989, 1993b) and Larsen and Rubin (1999). The EM-based algorithms for finding maximum likelihood estimates in the latent classes models of record linkage are a direct generalization of ideas for automatically estimating parameters given in Fellegi and Sunter (1969). The basic research problems are quite difficult. The first is how to automatically obtain parameters and latent classes that allow automatic accurate determination of error rates. The second is how to effectively use combinations of training data for which true classification is known and general data for which true classification is unknown. Presently, some of the examples in machine learning suggest that appropriate training data – often obtained via very expensive clerical review – can be useful in some situations. Because of the additional structure available in record linkage, some authors (Winkler 1993b, 1994; Larsen and Rubin 1999) have been able to obtain good matching results without subsets of training data. The advantage of training data is that it implicitly imposes additional structure on the learning with general text. With record linkage, the additional structure is due to knowing that fields such as first name, last name, house number, and date-of-birth need to be compared. With general text, the algorithms of machine learning must create a structure for comparing that is facilitated by the training data. The machine learning algorithms are useful in record linkage situations when free-form names or addresses cannot be parsed into components. Winkler (1993b) and Nigam et al. (1999) have shown that each of the latent classes may be best estimated as a further mixture of latent classes. A third research problem emphasized by Nigam et al. (1999) is when the classes obtained under the theoretical latent class models correspond to true classes into which individuals might want to classify the data. Winkler (1989) showed that the parameters of the latent classes sometimes yield very poor matching performance if the latent classes do not correspond to the true classes of matches and nonmatches. Winkler (1993b) showed that dramatic improvements in matching can occur if the class of nonmatches is estimated as a mixture of two subclasses. To better make use of a priori information, Winkler (1988, 1989, 1993b) showed how convex constraints such as P(disagree first | M) < a, 0 < a < 1, or P(M) < b, 0 < b < 1, could be used to force estimates obtained via versions of the EM algorithm into regions of the subspace of parameters.

5. CONCLUDING REMARKS

This paper describes current research problems in record linkage and some related research in microdata confidentiality, information retrieval and data mining.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion. A shorter version appeared in the 1999 Statistical Society of Canada Proceedings of Survey Methods. The translation of the abstract to French was facilitated by Dr. Yves Thibaudeau.

REFERENCES

Aarts, E. H. L., and J. K. Lenstra (eds.) (1997), *Local Search in Combinatorial Optimization*, New York, NY: Wiley-Interscience.

Alvey, W. and Jamerson, B. (eds.) (1997), *Record Linkage Techniques -- 1997* (Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997, in Arlington VA), Washington, DC: Federal Committee on Statistical Methodology.

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.

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

Besag, J. (1974), “Spatial Interaction and the Statistical Analysis of Lattice Systems,” *Journal of the Royal Statistical Society, B*, **34**, 192-236.

Besag, J. (1986), “On the Statistical Analysis of Dirty Pictures (with discussion),” *Journal of the Royal Statistical Society, B*, **46**, 25-37.
Besag, J., P. J. Green, D. Higdon, and Mengersen, K. (1995), “Bayesian Computation and Stochastic Systems,” *Statistical Science*, **10**, 3-41.
Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.
Budzinsky, C. D. (1991), "Automated Spelling Correction," Statistics Canada.
Copas, J. R., and F. J. Hilton (1990), "Record Linkage: Statistical Models for Matching Computer Records,"*Journal of the Royal Statistical Society*, **A**, **153**, 287-320.
DeGuire, Y. (1988), "Postal Address Analysis," *Survey Methodology*, **14**, 317-325.
De Waal, A.G. and L.C.R. J. Willenborg (1996), “A View on Statistical Disclosure Control of Microdata,” *Survey Methodology*, **2**2, 95-103.
De Waal, A.G. and L.C.R. J. Willenborg (1998), “Optimal Local Suppression in Microdata,” *Journal of Official Statistics*, **14**, 421-435.
Deming, W. E., and G. J. Gleser (1959), "On the Problem of Matching Lists by Samples," *Journal of the American Statistical Association*, **54**, 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**, 1-38.
Fayad U., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.) (1996), *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: The MIT Press.
Fellegi, I. P. (1999), “Record Linkage and Public Policy: A Dynamic Evolution” in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 3-12.
Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, 1183-1210.
Fienberg, S. E., U. E. Makov, and A. P. Sanil (1997), “A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data,” *Journal of Official Statistics*, **13**, 75-89.
Fienberg, S. E., U. E. Makov, and R. J. Steele (1998), “Disclosure Limitation and Related Methods for Categorical Data,” *Journal of Official Statistics*, **14**, 485-502.
Frakes, W. B., and Baeza-Yates, R. (ed.) (1992), *Information Retrieval: Data Structures & Algorithms*, Upper Saddle River, NJ: Prentice-Hall PTR.
Friedman, J., T. Hastie, R. Tibshirani (1998), “Additive Logistic Regression: a Statistical View of Boosting,” Stanford University, Statistics Department Technical Reprort.
Fuller, W. A. (1993), “Masking Procedures for Microdata Disclosure Limitation,” *Journal of Official Statistics*, **9,** 383-406.
Geman, S. and D. Geman (1984), “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images,” IEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, 721-741.
Gill, L. (1999), “OX-LINK: The Oxford Medical Record Linkage System,” in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 15-33.
Gopal, R., P. Goes, and R. Garfinkel, “Confidentiality Via Camouflage: The CVC Approach to Database Query Management,” in *Statistical Data Protection ’98*, Eurostat, Brussels, Belgium, 1-8.
Hall, P. A. V., and Dowling, G. R. (1980), "Approximate String Comparison," *Computing Surveys*, **12**, 381-402.
Heckerman, D. (1996), “A Tutorial on Learning with Bayesian Networks,” Microsoft Research, Technical Report MSR-TR-95-06.
Jaro, M. A. (1976), "UNIMATCH," Software system (no longer available).
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.
Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, *Proceedings of the Section on Survey*

Research Methods, 303-308.
Kim, J. J. (1990), "Subdomain Estimation for the Masked Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 456-461.
Larsen, M. D. (1996), "Bayesian Approaches to Finite Mixture Models," Ph.D. Thesis, Harvard University.
Larsen, M. D., and D. B. Rubin (1999), "Iterative Automated Record Linkage Using Mixture Models," Statistics Department Technical Report, Harvard University.
Mateo-Sanz, J. M. and J. Domingo-Ferrer (1998), "A method for Data-Oriented Multivariate Microaggregation,"in *Statistical Data Protection '98*, Eurostat, Brussels, Belgium, section 1.
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**, 899-909.
Meng, X., and D. B. Rubin (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, **80**, 267-278.
Mitchell, T. M. (1997), *Machine Learning*, New York, NY: McGraw-Hill.
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**, 1005-1027.
Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press (out of print).
Newcombe, H. B., M. E. Fair, and P. Lalonde (1992), "The Use of Names for Linking Personal Records (with discussion), Journal of the American Statistical Association, **87**, 1193-1208.
Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.
Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell (1999), "Text Classification from Labeled and Unlabelled Documents using EM, *Machine Learning*, to appear.
Pollock, J. and Zamora, A. (1984), "Automatic Spelling Correction in Scientific and Scholarly Text," *Communications of the ACM*, **27**, 358-368.
Porter, E. H., and W. E. Winkler (1999), "Approximate String Comparison and its Effect in an Advanced Record Linkage System," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 190-199.
Rogot, E., Sorlie, P., and Johnson, N. J. (1986), "Probabilistic Methods in Matching Census Samples to the National Death Index," *Journal Chronological Disease*, **39**, 719-734.
Rubin, D. B. and H. S. Stern (1993), "Testing in latent class models using a posterior predictive check distribution," in *Analysis of latent variables in developmental research*, (eds., Clogg, C. and von Eye, A.).
Scheuren, F., and W. E. Winkler (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, **19**, 39-58.
Scheuren, F., and W. E. Winkler (1997), "Regression analysis of data files that are computer matched, II," *Survey Methodology*, **23**, 157-165.
Sekar, C. C., and W. E. Deming (1959), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, **44**, 101-115.
Smith, M. E. and H. B. Newcombe (1975), "Methods for Computer Linkages of Hospital Admission-Separation Records into Cumulative Health Histories," *Meth. Inform. Medicine*, 18-25.
Sweeney, L. (1999), "Computational Disclosure Control for Medical Microdata: The Datafly System" in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 442-453.
Tendick, P. and N. Matloff (1994), "A Modified Random Perturbation Method for Database Security," *ACM Transactions on Database Systems*, **19**, 47-63.
Thibaudeau, Y. (1989), "Fitting Log-Linear Models When Some Dichotomous Variables are Unobservable," in *Proceedings of the Section on Statistical Computing, American Statistical*

Association, 283-288.
Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, **19**, 31-38.
Titterington, D. M., A. F. M. Smith, U. E. Makov (1988), *Statistical Analysis of Finite Mixture Distributions*, NewYork: J. Wiley.
Van Dyk, D. A. (1999), "Nesting EM Algorithms for Computational Efficiency," *Statistica Sinica*, to appear.
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*, 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*, 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**, 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 Statistical Association*, 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.*, 354-359.
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*, 274-279.
Winkler, W. E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472 (longer version report 94/05 available at http://www.census.gov/srd/www/byyear.html).
Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al*. (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.
Winkler, W. E. and Scheuren, F. (1995), "Linking Data to Create Information," *Proceedings of Symposium 95, From Data to Information - Methods and Systems*, Statistics Canada, 29-37.
Winkler, W. E. and Scheuren, F. (1996), "Recursive Analysis of Linked Data Files," *Proceedings of the 1996 Census Bureau Annual Research Conference*, 920-935.
Winkler, W. E. (1997), "Producing Public-Use Microdata That are Analytically Valid and Confidential," Proceedings *of the Section on Survey Research Methods, American Statistical Association*, 41-50.
Winkler, W. E. (1998), "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," *Research in Official Statistics*, **1**, 87-104.
Winkler, W. E. (1999), "Issues with Linking Files and Performing Analyses on the Merged Files," *Proceedings of the Section on Social Statistics, American Statistical Association*, to appear.
Winkler, W. E. and Scheuren, F. (1991), "How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis," U.S. Bureau of the Census, Statistical Research Division Technical Report.
Winkler, W. E. and Thibaudeau, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," U.S. Bureau of the Census, Statistical Research Division Technical report RR91/09.