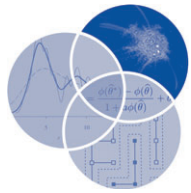
Overview



Matching and record linkage

William E. Winkler*

This overview gives background on a number of statistical methods that have been proven effective for record linkage. To prepare data for the main computational algorithms, we need parsing/standardization that allows us to structure the free-form names, addresses, and other fields into corresponding components. The main parameter-estimation methods are unsupervised methods that yield 'optimal' record linkage parameters. Extended methods provide estimates of false match rates in both unsupervised and, with greater accuracy, in semi-supervised situations. Finally, the paper describes ongoing research for adjusting standard statistical analyses for linkage error.

How to cite this article:
WIREs Comput Stat 2014, 6:313–325. doi: 10.1002/wics.1317

Keywords: entity resolution; classification rules; string comparison; false match and nonmatch rates; unsupervised learning

INTRODUCTION

Record linkage or computerized matching refers to the science (and art) of matching the same entities (person, business, etc.) using quasi-identifiers such as name, address, date of birth, etc. Quasi-identifiers do not uniquely identify by themselves but may, in combination, uniquely identify an entity. For instance, name plus date of birth may not uniquely identify an individual due to individuals with the same name having coincidental agreement on date of birth. The addition of a current address when combined with name plus date of birth will usually allow unique identification.

In this overview, we describe the statistical underpinnings of record linkage (Fellegi and Sunter[1] and Newcombe et al.[2,3]) and the methods (preprocessing, standardization, and parsing) developed for putting the raw data into a form suitable for effectively applying the theoretical models. In the first following section, we provide background on the preprocessing needed for files and the need for understanding of how the files will be used during matching and subsequent analyses. This preprocessing typically has a larger effect of matching efficacy (lower false match and lower false nonmatch rates or higher precision and higher recall) than improved models or better parameters associated with the improved models. In the second following section, we describe nuances of the theoretical models including the methods of supervised learning, semi-supervised learning [MCMC (Markov chain Monte Carlo) or EM (expectation–maximization)], and unsupervised learning (EM or MCMC) plus the methods of approximate string comparison that successfully allow the use of corresponding fields with minor typographical error. In the third following section, we review current research of methods of adjusting statistical analyses for matching error. Although promising in very small situations and theoretically correct, current research on the methods relates to realistic and computationally tractable extensions with much larger and more realistic real data.

*Correspondence to: william.e.winkler@census.gov
CSRM, US Census Bureau, Suitland, MD, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

BACKGROUND ON RECORD LINKAGE METHODS

Record linkage (entity resolution, computer matching) are methods for linking two files A and B using quasi-identifiers such as name, address, date of birth, and other fields. For instance, 'Marcus Van Huizen'

may not uniquely identify a person but in combination with a current address or accurate date of birth may uniquely identify. A relatively commonly occurring name such as 'Robert Smith' even when combined with a date of birth may not uniquely identify because several 'Robert Smith's may have the same date of birth.

Before proceeding to mathematical models for record linkage, we will describe methods of preprocessing and standardizing records. The preprocessing/standardization can have a larger effect on record linkage than improved models and parameter estimation in the models. We will provide a description of the main three components of preprocessing/standardization prior to giving a summary of some of the deleterious effects of failures of the preprocessing/standardization to put corresponding fields in records into forms that we can compare. The three components are (1) reformatting, (2) free-form name standardization, and (3) free-form address standardization. We describe examples from U.S. lists, but the methods for Western Europe, most areas of the Western hemisphere, and Australia/New Zealand are virtually identical. Randall et al.[4] provide an overview of the effects of the clean-up of data on record linkage. Herzog et al.[5] give more examples.

Reformatting into a Common Form

The first obvious need is for putting dates of birth into comparable forms with a formatting convention of MMDDYYYY (in Europe DDMMYYYY). If we have the form 'October 22, 1974', then we reformat into the form '10221974' for each of the lists. This allows us to compare the day of birth, month of birth, and year of birth across different lists. In some lists, the SEX code may be given by 2-female, 1-male, and 0-missing. We convert these to F-female, M-male, and ' '-missing.

Free-Form Name Standardization

A free-form name in the United States may be given by 'John A. Smith, M.D.' (first-name first) or 'Dr. Smith Junior, John A.'. Each of these forms needs to be converted into forms for which we can compare the first name John, the middle initial A, and the last name Smith with the corresponding fields in other records. We typically remove punctuation. We convert different spellings to a common form (M.D. ->MD, Junior to JR). Even with successful standardization, we cannot accurately match a name such as 'Mrs. J. K. Smith' (common in Southern United States with husband' initials until the 1960s) with 'Margaret Smith' (Margaret is first name on birth certificate) or 'Susan K. Fields' with 'Karen Jones' (individual goes by middle name and has married with change in last name).

Free-Form Address Standardization

A free-form address of the form '16 West Maple Street' may have words such as 'West' converted to a common symbol such as 'W' and 'Street' to a common abbreviation 'St'. After the conversion to common spellings, we parse the address into components so that a street name 'Maple' and a house number '16' can be compared to corresponding street names and house numbers. Addresses such as 'Box 16', 'Rural Route 1, or Box 12' are usually straightforward to standardize/parse but may not have as much distinguishing power as a house-number street-name type of address. In the situation of apartments, we need to separate out unit identifiers such as 'Apartment 23', 'Apartment 1A', 'Suite 334', and 'Trailer 66'.

We do most preprocessing/standardization using rule-based systems (if-then-else rules that first decide on a 'pattern' for the address prior to parsing). Although hidden Markov models (Christen et al.[6]) have also been shown to work well (and have a potential to be more easily adaptable to new situations), they presently do not outperform the best name standardizers (SAS Dataflux, IBM, Census—there are numerous others) or the best address standardizers (SAS Dataflux, IBM, Census, others). The hidden Markov-based methods outperform rule-based methods with South Asian and Southeast Asian addresses (Christen et al.[6]) and are known to work very well in more difficult situations (Cohen and Sarawagi[7] and Agichtein and Ganti[8]).

Generally, if a record fails reformatting or is reformatted into inconsistent forms, fails name, address, or other preprocessing/standardization, then we cannot compare the components via automated methods based on machine learning models (described later). For instance, with two files A and B, if 5% of names in each fail name standardization and 5% of addresses fail address standardization, then there is a potential to miss as many as 20% (5% + 5% + 5% + 5%) of 'true' matches. Even at a superficial level, we can assume that the errors/standardization failure can be additive in terms of decreasing matching rates.

Typical Additional Situations that Prevent Automated Matching

Missing data in fields such as first name, parts of the dates of birth, components of the address such as the unit identifiers can prevent matching or greatly reduce the accuracy of the matching. Even if we have

successfully standardized/parsed a name or address, we need to deal with typographical error such as comparing a first name such as 'Robert' with 'Robrrt' or a street name such as 'Maple' with 'Mapel' with the latter two versions contain transcription/keying error. In some situations, it can be helpful to deal with nicknames (William vs Bill, Robert vs Bob) but we must take care because 'Bobbie' may be a woman's real name from her birth certification and not a nickname for 'Roberta'. String comparators can deal with minor typographical errors (e.g. Winkler[9] and Cohen et al.[10,11]).

Summarizing Comment

Failures in file preprocessing/standardization can reduce the number of matches by 20–30% or more (see above paragraphs on name or address standardization). In some situations, it may be impossible to use a list in matching (such as the situation where one keypuncher erroneously keyed almost 100% of first names, last names and dates of birth and another miskeyed close to 0% of the records). In the next section, we will see that with the same quality of lists and preprocessing, the differing record linkage models and parameter-estimation methods will only make a few percentage points difference in matching efficacy.

RECORD LINKAGE MODEL OF FELLEGI AND SUNTER

The basic ideas of computerized record linkage were introduced by geneticist Howard Newcombe et al.[2,3] and were later mathematically formalized by Fellegi and Sunter[1] (hereafter FS). Additionally, FS provided methods of estimating crucial parameters in the model without training data (i.e., unsupervised learning). Cooper and Maron[12] in the information retrieval literature rediscovered the mathematical model but they did not provide proofs of the optimality of the classification (decision) rules.

To begin, we need notation. We match two files A and B. 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. FS considered ratios of probabilities of the form:

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

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. We refer to the ratio R or any monotonely increasing function of it such as the natural log as a *matching weight* (or score).

The decision rule is given by:

$$\begin{aligned} &\text{If } R > T_\mu, \text{ then designate pair as a match.} \\ &\text{If } T_\lambda \le R \le T_\mu, \text{ then designate pair as a possible} \\ &\qquad \text{matchand hold for clerical review.} \\ &\text{If } R < T_\lambda, \text{ then designate pair as a nonmatch.} \end{aligned} \qquad (2)$$

The cutoff thresholds T_μ and T_λ 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. Rule (2) partitions the set $\gamma \in \Gamma$ into three disjoint subregions. We refer to the region $T_\lambda \le R \le T_\mu$ as the no-decision region or clerical review region. In some situations, resources are available to review pairs clerically.

Figure 1 provides an illustration of the curves of log frequency versus log weight for matches and nonmatches, respectively. The two vertical lines represent the lower and upper cutoffs thresholds T_λ and T_μ, respectively. The *x*-axis is the log of the likelihood ratio *R* given by Eq. (1). The *y*-axis is the log of the frequency counts of the pairs associated with the given likelihood ratio. The plot uses pairs of records from a contiguous geographic region that we matched in the 1990 Decennial Census. The clerical review region between the two cutoffs primarily consists of pairs within the same household that are missing both first name and age, the only two fields we had for distinguishing individuals within a household.

The Methods of Fellegi and Sunter

FS[1] (Theorem 1) proved the optimality of the classification rule given by Eq. (2). Their proof is very general in the sense that it holds for any representations $\gamma \in \Gamma$ over the set of pairs in the product space $\mathbf{A} \times \mathbf{B}$ from two files. As they observed, the quality of the results from classification rule (2) was dependent on the accuracy of the estimates of $P(\gamma \in \Gamma | M)$ and $P(\gamma \in \Gamma | U)$.

FS were the first to give very general methods for computing these probabilities in situations that differ from the situations of Newcombe in the previous section. As the methods are useful, we describe what they introduced and then show how the ideas led into more general methods that can be used for

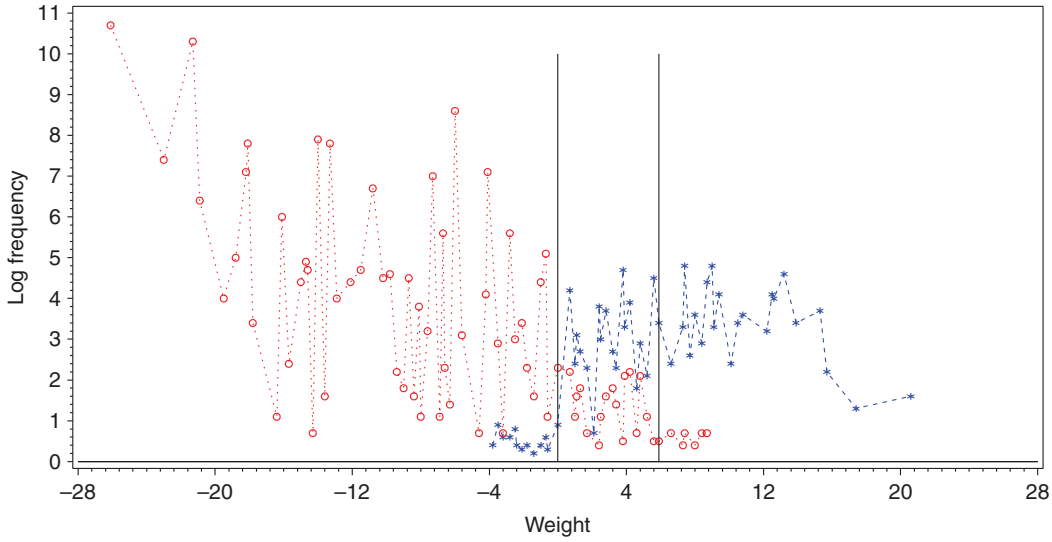


FIGURE 1 | Plots of log frequencies versus matching weight of nonmatches and matches. INLINE GRAPHICS CIRCLE = nonmatch, * = match, cutoff 'L' = 0 and cutoff 'U' = 6.

unsupervised learning (i.e., without training data) in a large number of situations.

FS observed several things. First,

$$P(S) = P(S \mid M)\, P(M) + P(S \mid U)\, P(U) \qquad (3)$$

for any set S of pairs in $\mathbf{A} \times \mathbf{B}$. We can compute the probability on the left directly from the set of pairs. In the following, we will use Ag_i^x to represent either agreement or disagreement on the ith field but not both. If sets S represent simple agreement/disagreement associated with three fields, under conditional independence (CI), we obtain

$$P\left(Ag_1^x \cap Ag_2^x \cap Ag_3^x \mid D\right) = P\left(Ag_1^x|D\right) P\left(Ag_2^x|D\right) P\left(Ag_3^x|D\right) \qquad (4)$$

then Eqs (3) and (4) provide seven equations and seven unknowns (as x represents agree or disagree) that yield quadratic equations that they solved. Here D is either M or U. Equation (or set of equations) (4) can be expanded to k fields. Although there are eight patterns associated with the equations of the form (4), we eliminate one because the probabilities must add to one. In general, with more fields but still simple agreement/disagreement between fields, we can solve the equations via the EM (expectation–maximization) algorithm in the section The Model for Semi-Supervised Learning. Probabilities of the form $P\left(Ag_i^x \mid D\right)$ are referred to as m-probabilities if $D = M$ and u-probabilities if $D = U$.

There are a number of implicit assumptions when matching two files and computing probabilities using Eqs (3) and (4). The first is that there is significant overlap between two files A and B. This essentially means that $A \cap B$ is either most of A or most of B. If this assumption is not true, then the probabilities obtained via Newcombe's methods[2,3]) or the FS methods may not work well. The second assumption is that neither file A nor file B can simultaneously be sampled from two larger files A_2 and B_2. Deming and Gleser[13] provided theory demonstrating the unreliability of determining the sampling overlap (i.e., number of duplicates) from two sample files. As a case in point, if $A_2 = B_2$ each contains 1000 records on which 1% have the last name of 'Smith', among the matches M between A_2 and B_2, there is a 1% probability of being a pair agreeing on 'Smith' actually being a match. If A and B are 10% samples of A_2 and B_2, respectively, then among matches between A and B, there is a 0.1% probability of a pair agreeing on 'Smith' actually being a match. The third assumption is that the typographical error rates are quite low so the frequency-based computations based on the different observed values of the fields are valid. If a relatively rare value of last name such as 'Zabrinsky' has six different spellings in the six records in which it appeared, then it is not possible to compute accurate frequency-based probabilities (also referred to as value-specific probabilities) directly from the file.

In practice, it is necessary to perform *blocking* of two files that affect how pairs are brought together. If two files A and B each contains 10,000 records, then there are 10^8 pairs in the product $\mathbf{A} \times \mathbf{B}$. Until very recently, we could not do the detailed computation of 10^8 pairs. In *blocking*, we only consider pairs that agree on certain characteristics. For instance, we may

only consider pairs that agree on first initial of first name, last name, and date of birth. If we believe (possibly based on prior experience) that we are not getting a sufficiently large proportion of matches with a first blocking criteria, we may try the second. For instance, we may only consider pairs that agree on first initial of first name, first initial of last name, and the ZIP + 4 U.S. postal code (that represents approximately 50 households). FS gave the straightforward theoretical extensions for blocking. In performing computation over pairs P_1 in $\mathbf{A} \times \mathbf{B}$ obtained via blocking, there is a fourth implicit assumption: that the pairs in P_1 contain a moderately high proportion of matches (say 3 + % of P_1 consists of matches). In the next section, we return to the minimal needed proportion of pairs needing to be matches in more general situations. The methods of obtaining the probabilities given by Eqs (3) and (4) break down when the proportion of matches from M in the set of pairs P_1 is too low. The computations also break down if we do the computation over all 10^8 pairs in $\mathbf{A} \times \mathbf{B}$. In $\mathbf{A} \times \mathbf{B}$, if all records in A can be matched to records in B, then at most 0.01% of the pairs are matches ($10^4/10^8$). In the next section, we will show how we can effectively find reasonable estimates of false match probabilities in a variety of situations (without training data).

Michelson and Knoblock[14] and Bilenko et al.[15] provided methods for obtaining sets of blocking criteria. Christen[16] provided an excellent survey/review of methods for blocking. Kenig and Gal[17] provided a rather extensive methodological development. Winkler et al.[18] give an application of four blocking criteria in the production Decennial Census Matching system that reduced the amount of computation from 10^{17} pairs (300 million × 300 million) to 10^{12} pairs taking less than 30 h using 40 cpus on an SGI Linux machine. Winkler[19] provides a method for estimating the number of matches missed by a set of blocking criteria. Contrary to many individuals[14,15,17] developing sets of blocking criteria can be straightforward in many situations. Three individuals each independently developed methods that improved over methods given by Winkler.[19] These three individuals subsequently developed the four blocking criteria in the methods by Winkler et al.[18]

Applying Extensions of Fellegi–Sunter Methods—The EMH Algorithm

The natural ways of extending the parameter-estimation methods of FS to k fields with simple agree–disagree comparisons are via the EM algorithm, methods of moments, and least squares. The EM algorithm is greatly preferred (Titterington et al.[20]) because of its numerical stability in computation and ability to extend it to models involving interactions between fields and other generalizations such as approximate string comparison for accounting for typographical error (Winkler[9] and Cohen et al.[10,11]. A general EM algorithm, referred to as the EMH algorithm (Winkler[21,22]) extends the Expectation-Conditional Maximization (ECM) algorithm (Meng and Rubin[23]) from linear to convex constraints. Linear constraints are the standard constraints of loglinear modeling of contingency tables.[24,25] To improve possibly the classification, convex constraints[21,22] can predispose certain convex sums of the probabilities associated with appropriate groups of cells by putting upper bounds on the sums or the complements of the sums because the total sum over all cells must add to 1.0. The EMH algorithm[22] is suitable for all models in this paper.

Herzog et al.[26] provide more detail about how the EM is applied in production matching for the Decennial Census where 'optimal' parameters differ in each of the ~500 contiguous regions into which the United States is divided and which must be estimated without training data. The 'optimal' parameters reduced the clerical review region by approximately two thirds. The highly structured EM showed its strength. In three regions among a number of regions processed in 1 week, clerical review became much larger with the EM parameters than was expected. Quick review showed that two keypunchers had managed to bypass edits on the year of birth. All records from these keypunchers disagreed on the computed age. The clerical review became much larger because first name and the age were the main fields for separating persons within a household. It further provided parameters and classification rules (Herzog et al.[26]) that improved on the type of active learning (a type of semi-supervised learning) in which subsets of pairs are reviewed and m-probabilities are reestimated with an iterative procedure until they stabilize in each region.

In this section, we provide a summary of current extensions of the EM procedures for estimating false match rates. With any matching project, we are concerned with false match rates among the set of pairs among designated matches above the cutoff score T_μ in Eq. (2) and the false nonmatch rates among designated nonmatches below the cutoff score T_λ in Eq. (2). Very few matching projects estimate these rates although valid estimates are crucial to understanding the usefulness of any files obtained via the record linkage procedures. We can sometimes provide reasonable upper bounds for the estimated error rates via experienced practitioners and the error rates are validated during follow-up studies (Winkler[27]).

If a small amount of training data is available, then it may be possible to improve record linkage and good estimates of error rates. Larsen and Rubin[28] combined small amounts of (labeled) training data with large amounts of unlabeled data to estimate error rates using an MCMC procedure. In machine learning (Winkler[29]), the procedures are referred to as *semi-supervised learning*. In ordinary machine learning, the procedures to get parameters are 'supervised' by the training data that is labeled with the true classes into which later records (or pairs) will be classified. Winkler[30] also used semi-supervised learning with a variant of the general EM algorithm. Both the Larsen and Rubin[28] and Winkler[29,30] methods were effective because they accounted for interactions between the fields and were able to use labeled training data that was concentrated between the lower cutoff T_λ and the upper cutoff T_μ of Eq. (2).

Belin and Rubin[31] were the first to provide an unsupervised method for obtaining estimates of false match rates. The method proceeded by estimating Box–Cox transforms that would cause a mixture of two transformed normal distributions to closely approximate two well-separated curves such as given in Figure 1. They cautioned that their methods might not be robust to matching situations with considerably different types of data. Winkler[27] observed that their algorithms would typically not work with business lists, agriculture lists, and low quality person lists where the curves of nonmatches were not well separated from the curves of matches. Scheuren and Winkler,[32] who used the Belin–Rubin EM-based fitting software, observed that the Belin–Rubin method did work reasonably well with a number of well-separated person lists.

Because the EM-based methods of this section serve as a template for other EM-based methods, we provide details of the unsupervised learning methods of Winkler[33] that are used for estimating false match rates. The basic model is that of semi-supervised learning in which we combine a small proportion of labeled (true or pseudo-true matching status) pairs of records with a very large amount of unlabeled data. The CI model corresponds to the naïve Bayesian network formalization of Nigam et al.[34] The more general formalization of Winkler[29,30] allows interactions between agreements (but we do not use the general formalization in this paper).

The Model for Semi-Supervised Learning

Our development is similar theoretically to that of Nigam et al.[34] The notation differs very slightly because it deals more with the representational framework of record linkage. Let γ_i be the agreement pattern associated with pair p_i. Classes C_j are an arbitrary partition of the set of pairs D in $\mathbf{A} \times \mathbf{B}$. Later, we will assume that some of the C_j will be the subsets of M and the remaining C_j are the subsets of U. For coherence and clarity Eqs (5) and (6) repeat earlier Eqs (3) and (4) but use slightly different notation that brings in the parameter Θ that we are estimating and which was implicit in Eqs (3) and (4). The parameter Θ might refer to a specific type of model such as multinomial or Dirichlet-multinomial (see e.g. Agresti[24] and Bishop et al.[25]). Unlike general text classification in which every document may have a unique agreement pattern, in record linkage, some agreement patterns γ_i may have many pairs $p_{i(\mathrm{l})}$ associated with them. Specifically,

$$P\left(\gamma_i|\Theta\right) = \sum_{i}^{|C|} P\left(\gamma_i|\ C_j;\Theta\right)\ P\left(C_j;\Theta\right) \tag{5}$$

where γ_i is a specific pair, C_j is a specific class, and the sum is over the set of classes. Under the Naïve Bayes or CI, we have

$$P\left(\gamma_i|\ C_j;\Theta\right) = \Pi_k P\left(\gamma_{ik}|\ C_j;\Theta\right) \tag{6}$$

where the product is over the kth individual field agreement γ_{ik} in the pair with agreement pattern γ_i. In some situations, we use a Dirichlet prior

$$P\left(\Theta\right) = \Pi_j \left(\Theta_{Cj}\right)^{\alpha-1} \Pi_k \left(\Theta\gamma_{i,k|\ Cj}\right)^{\alpha-1} \tag{7}$$

where the first product is over the classes C_j and the second product is over the fields. The Dirichlet prior can assure that maximum likelihood estimates usually exist and that most cells in the contingency table estimated via the EMH algorithm have positive probability. We use D_u to denote unlabeled pairs and D_l to denote labeled pairs. Given the set $D = D_\mathrm{u} \cup D_\mathrm{l}$ of all labeled and unlabeled pairs, the log likelihood is given by

$$\begin{aligned} l_c\left(\Theta|D,\ z\right) = \log\left(P(\Theta)\right) + \ (1-\lambda) \sum_{i\in D_\mathrm{u}} \sum_j z_{ij} \\ \times \log(P\left(\gamma_i\ |\ C_j\ ;\Theta\right) P\left(C_j\ ;\Theta\right)) \\ + \sum_{i\in D_\mathrm{l}} \sum_j z_{ij} \log\left(P(\gamma_i\ |\ C_j\ ;\Theta)\ P\left(C_j\ ;\Theta\right)\right). \end{aligned} \tag{8}$$

where $0 \le \lambda \le 1$. The first sum is over the unlabeled pairs and the second sum is over the labeled pairs. In the third terms of the right-hand-size of Eq. (8), we sum over the observed z_{ij}. In the second term, we put in expected values for the z_{ij} based on the initial estimates $P(\gamma_i|C_j;\ \Theta)$ and $P(C_j;\ \Theta)$. After reestimating the

parameters $P(\gamma_i|C_j;\Theta)$ and $P(C_j;\Theta)$) during the M-step [that is in closed form under condition (CI)], we put in new expected values and repeat the M-step. The computer algorithms are easily monitored by checking that the likelihood increases after each combination of E- and M-steps and by checking that the sum of the probabilities add to 1.0. We observe that if λ is 1, then we only use training data and our methods correspond to naïve Bayes methods for which training data are available. If λ is 0, then we are in the unsupervised learning situations of Winkler.[21,33] Winkler[29,30] provides more details of the computational algorithms.

The Data Files Used in the Analyses

Three representative pairs of files were used in the analyses. The three pairs were chosen from 457 pairs of files used in evaluating undercount/overcount in the 1990 Census. The files are from 1990 Decennial Census matching data (more details in Refs 26 and 30) in which the entire set of 1–2% of the matching status codes that were believed to have been in error for these analyses have been corrected. The corrections reflect clerical review and field follow-up that were not in computer files originally available to us. In each pair of files, one file consists of a large sample of blocks from the Census and the other file represents a semi-independent re-enumeration of the same sample of blocks that were performed 6 weeks after the Census.

A summary of the overall characteristics of the empirical data is in Table 1. We only consider pairs that agree on census block id (small geographic area representing approximately 50 households) and on the first character of surname. Less than 1–2% of the matches are missed using this set of blocking criteria. They are not considered in the analysis of this paper.

The matching fields that are:

Person characteristics: first name, age, marital status, sex.

Household characteristics: last name, house number, street name, phone.

Typically, everyone in a household will agree on the household characteristics. Person characteristics such as first name and age help distinguish individuals within household. Some pairs (including true matches) have both missing first name and age.

We also consider partial levels of agreement in which the string comparator values are broken out as [0, 0.66], [0.66, 0.88], [0.88, 0.94], and [0.94, 1.0]. The intervals were based on knowledge of how string comparators were initially modeled (Winkler[9]) in terms of their effects of likelihood ratios (1). The first interval is what we refer to as disagreement. We combine the disagreement with the three partial agreements and blank (missing) to get five value states (base 5). The large base analyses consider five states for all characteristics except sex and marital status for which we consider three (agree/blank/disagree). The total number of agreement patterns is 140,625. In the earlier work (Winkler[30]), the five levels of agreement worked consistently better than two levels (agree/disagree) or three levels (agree/blank/disagree).

The pairs naturally divide into three classes: C_1—match within household, C_2—nonmatch within household, i_3—nonmatch outside household. In the earlier work (Winkler[30]), we considered two dependency models in addition to the CI model. In that work small amounts of labeled training data were combined with unlabeled data, the CI model worked well and the dependency models worked slightly better.

We create 'pseudo-truth' data sets in which matches are those unlabeled pairs above a certain high cutoff and nonmatches are those unlabeled pairs below a certain low cutoff. Figure (1) illustrates the situation using actual 1990 Decennial Census data in which we plot the log of the probability ratio (1) against the log of frequency. With the data sets of this paper, we choose high and low cutoffs in a similar manner so that we do not include in-between pairs in our designated 'pseudo-truth' data sets. We use these 'designated' pseudo-truth data sets in a semi-supervised learning procedure that is nearly identical to the semi-supervised procedure where we have actual truth data. A key difference from the corresponding procedure with actual truth data is that the sample of labeled pairs is concentrated in the difficult-to-classify in-between region (see Figure 1), where, in the 'pseudo-truth' situation, we have no way to designate comparable labeled pairs. The sizes of the 'pseudo-truth' data are given in Table 2. The error rates associated with the artificial 'pseudo-truth' are in parentheses following the counts. The *Other* class gives counts of the pairs and proportions of true matches that are not included in the 'pseudo-truth' set of pairs. In the *Other* class, the proportions of matches vary somewhat and would be difficult to determine without training data.

TABLE 1 | Summary of Three Pairs of Files

	Files		Files		Files	
	A1	A2	B1	B2	C1	C2
Size	15,048	12,072	4539	4851	5022	5212
Number of pairs	116,305		38,795		37,327	
Number of matches	10,096		3490		3623	

TABLE 2 | 'Pseudo-Truth' Data with Actual Error Rates Corresponding to Table 1

	Matches	Nonmatches	Other
A pairs	8817 (0.008)	98,257 (0.001)	9231 (0.136)
B pairs	2674 (0.010)	27,744 (0.0004)	8377 (0.138)
C pairs	2492 (0.010)	31,266 (0.002)	3569 (0.369)

The Results of the Modeling

We determine how accurately we can estimate the lower cumulative distributions of matches and the upper cumulative distribution of nonmatches. This corresponds to the overlap region of the curves of matches and nonmatches. If we can accurately estimate these two tails of distributions, then we can accurately estimate error rates at differing levels. Our comparisons consist of a set of figures in which we compare a plot of the cumulative distribution of estimates of matches divided by the true cumulative distribution versus the true cumulative distribution with the truth represented by the 45° line. As the plots get closer to the 45° lines, the estimates get closer to the truth. Our plotting is only for the bottom 30% of the curves given in Figures 2 and 3 of Belin and Rubin[31]. Generally, we are only interested in the bottom 10% of the curves for the purpose of estimating false match rates. Because of the different representation with the 45% curve, we can much better compare three different methods of estimation for false match rates than with the type of scaling in the Figures of Belin and Rubin.[31]

Our primary results are from using the CI model and 'pseudo-semi-supervised' methods of this section with the conditional independence model and actual semi-supervised methods of Winkler.[33] With our 'pseudo-truth' data, we obtain the best sets of estimates of the bottom 30% tails of the curve of matches with CI and $\lambda = 0.2$ (Eq. 8). Figure 2(a)–(c) illustrates the set of curves that provide quite accurate fits for the estimated values of matches versus the truth. The 45° line represents the truth whereas the curve represents the cumulative estimates of false match rates for the right tails of the distribution. The plots are for the estimates of the false match probabilities divided by the true false match probabilities. Although we looked at results for $\lambda = 0.1$, 0.5, and 0.8 and various interactions models, the results under CI were the best with $\lambda = 0.2$ (Eq. 8). We also looked at several different ways of constructing the 'pseudo-truth' data. Additionally, we considered other pairs of files in which all of the error-rate estimates were better (closer to the 45° line) than those for the pair of files given in Figure 2.

Figure 2(d)–(f) are the corresponding curves using the methods of Belin and Rubin.[31] The curves are substantially farther from the 45° lines because they are only using the distributions of weights (natural logs of likelihood ratio (1)). Using the detailed breakout of the string comparator values, the fact that the three-class EMH algorithm (Winkler[21]) provides much better estimates, and the breakout available from Eq. (8), we use more information that allows the estimated curves in Figure 2(a)–(c) to be closer to the truth than the corresponding curves in Figure 2(d)–(f).

The final sets of curves (Figure 2(g)–(i)) are similar to the semi-supervised learning of Winkler[30] that achieved results only very slightly worse than Larsen and Rubin[24] but for which the EM computational speeds (10 min in each of ~500 regions) were at least 100 times as the MCMC methods of Larsen and Rubin. Although Larsen and Rubin[28] did not give detailed timing results, they used the same input files as Winkler[30] on the exact same machine. It is difficult for the unsupervised methods to perform as well as the semi-supervised methods because the relatively small sample can be concentrated in the clerical review region between the lower cutoff T_λ and the upper cutoff T_μ. Because we had underlying truth data, we knew that in some regions only 1/40 of the 'truth' sample was truly a match whereas in other regions 1/10 of the 'truth' sample was truly a match. In the 1990 Decennial Census, the clerical region consisted almost entirely of individuals within the same household who were missing both first name and age (the only two fields for distinguishing within the household). Because we needed to match all 457 regions of the United States in 3–6 weeks to provide estimates required by law, we could not perform clerical review in each region or use approximations across certain regions because the optimal parameters vary significantly from region to region (Winkler[35]).

Summarizing Comments

Some groups have had difficulty in applying the EM algorithm effectively (based on private communications with several universities and national statistical agencies over 20 plus years). The first difficulty is typically in performing preprocessing/standardization effectively so that corresponding values of fields can be compared across records. The second is having an effective blocking strategy where a suitable subset S of pairs from $\mathbf{A} \times \mathbf{B}$ contains most true matches and the proportion of true matches within S is between 1 and 5% or even higher. In situations where there is moderate typographical error, the third difficulty is making better use of differing ranges of the string comparator

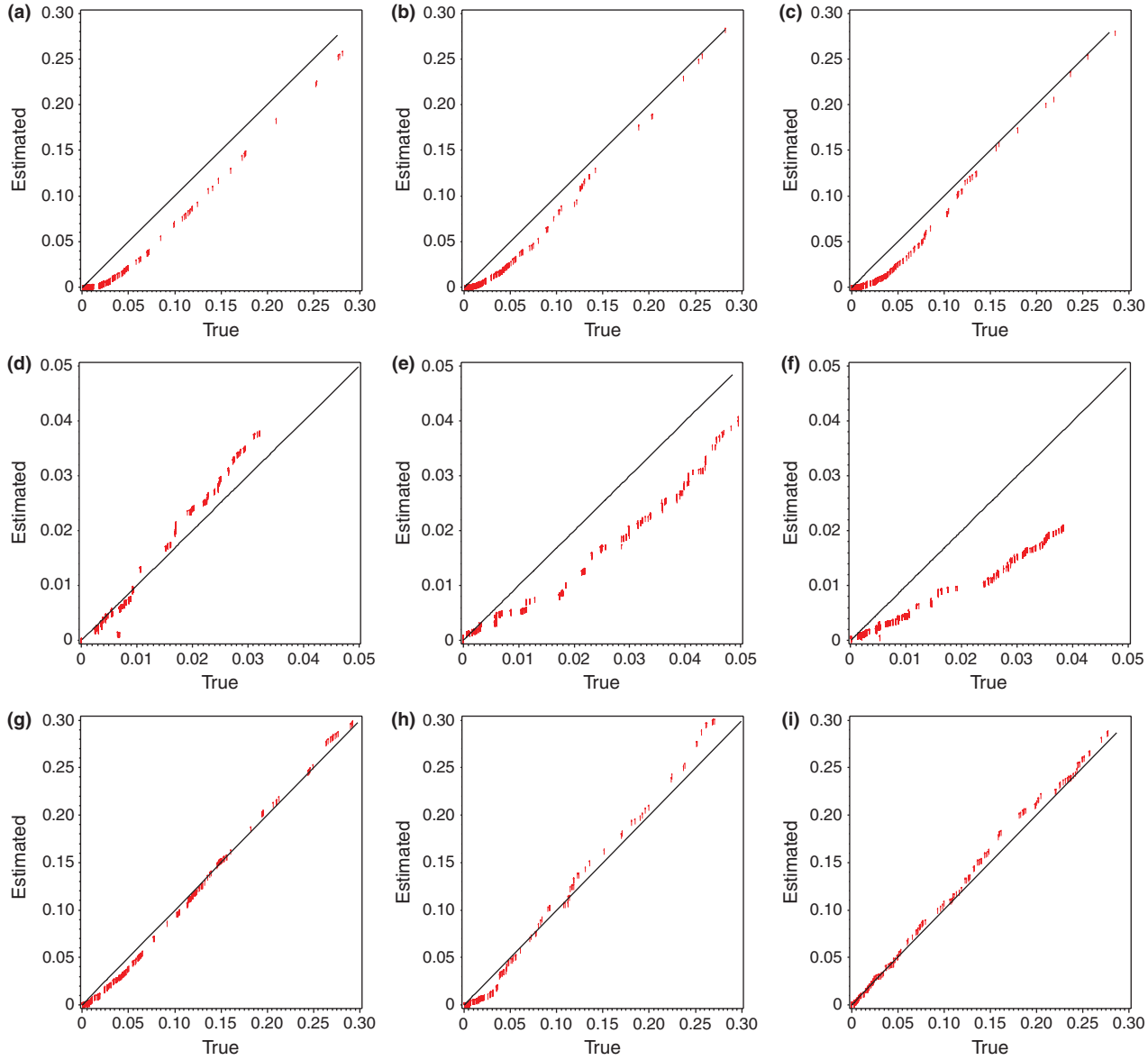


FIGURE 2 | False match rate estimates from three methods applied to three pairs of files. (a) Estimates versus truth, file A cumulative false matches, unsupervised independent EM, $\lambda = 0.2$. (b) Estimates versus truth file B cumulative false matches, unsupervised independent EM, $\lambda = 0.2$. (c) Estimates versus truth file C cumulative false matches, unsupervised independent EM, $\lambda = 0.2$. (d) Estimates versus truth, file A cumulative false match rates, unsupervised Belin-Rubin procedure. (e) Estimates versus truth, file B cumulative false match rates, unsupervised Belin-Rubin procedure. (f) Estimates versus truth, file C cumulative false match rates, unsupervised Belin-Rubin procedure. (g) Estimates versus truth file A cumulative false matches, semi-supervised small sample, independent EM, $\lambda = 0.99$. (h) Estimates versus truth file B cumulative false matches, semi-supervised small sample, independent EM, $\lambda = 0.99$. (I) Estimates versus truth file C cumulative false matches, semi-supervised small sample, independent EM, $\lambda = 0.99$.

values. Effective application of these three steps can allow the EM algorithm to yield very good parameters and make a substantial improvement in overall matching efficacy.

Cohen et al.[10,11] further developed string comparators. They observed that the Jaro-Winkler string comparator (Winkler[9]) works somewhat better than variants of the edit (Damerau-Levenstein) distance with Census-type lists but not necessarily with other types of lists. They also noted that the Jaro-Winkler string comparator is 10 times as fast to compute as edit distance.

In addition to the monograph of Christen,[36] general overviews in the computer science literature on record linkage (entity resolution) are by Elmagarmid et al.[37] and Naumann and Hershel.[38]

Alternative models for record linkage (in the supervised learning situation) are support vector machines (Vapnik[39]) and certain tree-based methods (Hastie et al.[40]). Christen[36] has applied the alternate models for basic record linkage. Presently, researchers have not applied SVM, tree-based methods and other methods for false match error-rate estimation (known as the *regression problem* in statistical learning theory when there is substantial training data). Virtually in all situations, training data are unavailable for record linkage.

CURRENT RESEARCH IN ADJUSTING STATISTICAL ANALYSES FOR LINKAGE ERROR

In this section, we describe research into methods for adjusting statistical analyses for linkage error. We will primarily be concerned with more straightforward regression-type models because that is the area with much more historical work. Later work also deals with the discrete data as in contingency tables (Chipperfield et al.,[41] Goldstein et al.,[42] and Winkler[43]). The discrete-data situation much more clearly demonstrates how difficult the research problem is with realistic situations that approximate the use of a pair of national files.

Unlike the much more mature methods in the previous two sections, there are substantial research problems. Scheuren and Winkler[32] extended methods of Neter et al.[44] to more realistic record linkage situations in the simple analyses of a regression of the form $y = \beta x$, where y is taken from one file A and x is taken from another file B. Because the notation of Lahiri and Larsen[45] is more useful in describing extensions and limitations, we use their notation.

Consider the regression model $\mathbf{y} = (y_1, \ldots, y_n)'$:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, n \tag{9}$$

where $x_i = (x_{i1}, \ldots, x_{ip})$ is a column vector of p known covariates $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2$, $\text{covariance}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, $i, j = 1, \ldots, n$. Scheuren and Winkler[32] considered the following missing-data model for $\mathbf{z} = (z_1, \ldots, z_n)$ given y:

$$z_i = \begin{cases} y_j \text{with probability } q_{ij} \\ y_j \text{with probability } q_{ii} \text{ for } i \neq j, i, j = 1, \ldots, n \end{cases} \tag{10}$$

where $\sum_{j=1}^{n} q_{ij} = 1$ for $i = 1, \ldots, n$. Define $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{in})'$, $j = 1, \ldots, n$, and $\mathbf{Q} = (\mathbf{q}_i, \ldots, \mathbf{q}_n)'$. The naïve least squares estimator of $\boldsymbol{\beta}$, which ignores mismatch errors, is given by

$$\hat{\boldsymbol{\beta}}_N = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{z},$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is an $n \times p$ matrix.

Under the model described by Eqs (9) and (10)

$$E(z_i) = \mathbf{w}'\boldsymbol{\beta}$$

where $\mathbf{w}_i = \mathbf{q}_i'\,\mathrm{X} = \sum_{j=1}^{n} q_{ij}\,\mathbf{x}_j'$ is a $p \times 1$ column matrix. The bias of the naïve estimator $\hat{\boldsymbol{\beta}}_N$ is

$$\begin{aligned} \text{bias}\left(\hat{\boldsymbol{\beta}}_N\right) &= E\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right) = \left[\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathrm{W} - \mathrm{I}\right]\boldsymbol{\beta} \\ &= \left[\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathrm{Q}\mathrm{X} - \mathrm{I}\right]\boldsymbol{\beta}. \end{aligned} \tag{11}$$

If an estimator of $\mathbf{B}$ is available where $\mathbf{B} = (B_1, \ldots, B_n)'$ and $B_i = (q_{ii} - 1)\ y_i + \sum_{j \neq 1} q_{ij}\ y_j$, then the Scheuren–Winkler estimator is given by

$$\hat{\boldsymbol{\beta}}_{SW} = \hat{\boldsymbol{\beta}}_N - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{12}$$

If q_{ij1} and q_{ij2} denote the first and second highest elements of the vector $\mathbf{q}_i$ and z_{j1} and z_{j2} denote the elements of the vector $\mathbf{z}$, then a truncated estimator of B is

$$\hat{\mathbf{B}}i^{TR} = \left(q_{ij1} - 1\right)\ z_{j1} + q_{V2}z_{j2}. \tag{13}$$

Scheuren and Winkler[32] used estimates of q_{ij1} and q_{ij2} based on software/methods from Belin and Rubin.[31] Lahiri and Larsen[45] improve the estimator (12) (sometimes significantly) by using the unbiased estimator

$$\hat{\boldsymbol{\beta}}_U = \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathrm{W}\prime\mathbf{z}. \tag{14}$$

The issues are whether it is possible to obtain reasonable estimates of $\mathbf{q}_i$ or whether the crude approximation given by Eq. (13) is suitable in many situations.

Under a significantly simplified record linkage model where each q_{ij} take the same value for $i \neq j$, $1, \ldots, n$, Chambers[46] (also Kim and Chambers[47,48]) provides an estimator approximately of the following form

$$\hat{\boldsymbol{\beta}}_U = \left(\mathbf{W}'\mathbf{Cov}_\mathbf{z}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}\prime\mathbf{Cov}_\mathbf{z}^{-1}\mathbf{z} \tag{15}$$

that has lower variance than the estimator of Lahiri and Larsen.[45] The matrix $\mathbf{Cov}_\mathbf{z}$ is the variance–covariance matrix associated with $\mathbf{z}$. The estimator in Eq. (15) is the best linear unbiased estimator (BLUE) using standard methods that improve over the unbiased estimator (14). Chambers[46] further provides an iterative method for obtaining an empirical BLUE using the observed data.

The issue with the Chambers' estimator is whether the drastically simplified record linkage model is a suitable approximation of the realistic model used by Lahiri and Larsen.[45] The issue with both the models of Chambers[46] and Lahiri and Larsen[45] is that they need both a method of estimating q_{ij} for all i, j with all pairs of records and a method of designating which of the q_{ij} is associated with the true match. Scheuren and Winkler[32] provided a much more *ad hoc* adjustment with the somewhat crude estimates of the q_{ij} obtained from the model of Belin and Rubin.[31] Lahiri and Larsen[45] demonstrated that the Scheuren–Winkler procedure was inferior for adjustment purposes when the true q_{ij} were known. Winkler and Scheuren,[49] however, were able to determine that their adjustment worked well in a large number of empirical scenarios (approximately one hundred) because the bias of the q_{ij}-estimation procedure and adjustment partially compensated for the increased bias of the Scheuren–Winkler procedure. Further, Winkler[30] provided a 'generalization' of the Belin–Rubin estimation procedure that provides somewhat more accurate estimates of the q_{ij} and holds in a moderately larger number of situations.

There have been no direct comparisons of the methods of Scheuren and Winkler,[32] Lahiri and Larsen,[45] and Chambers.[46] On theoretical grounds, Lahiri and Larsen[45] have the most complete model that should always outperform Scheuren and Winkler[32] if highly accurate estimates of the q_{ij} matching parameters are available. The models of Chambers et al.[46–48] have not been extended to the realistic record linkage model of Lahiri and Larsen.[45]

Summarizing Comments

Regression problems of the type described have initially provided the most progress on how matching error affects statistical analyses because of the relatively strong assumptions about the relationships of most variables. Ultimately, the methods for dealing with discrete data with general loglinear models (Chipperfield et al.[41] and Goldstein et al.[42]) may provide additional insights that are needed for completely general data situations.[49] The models of Chipperfield et al.[41] and Goldstein et al.[42] provide theoretically correct, computationally tractable solutions in relatively small situations but do not readily extend to (**X**, **Y**) situations where vector **X** in one file and vector **Y** from another file each take hundreds or thousands of distinct values.

Bhattacharya and Getoor[50] have applied MCMC methods for estimating basic record linkage parameters in the unsupervised situation, whereas Larsen and Rubin have applied MCMC in the semi-supervised situation. Tancredi and Liseo[51] and Liseo and Tancredi[52] have applied MCMC methods for situations of adjusting undercount estimation and regression analysis for linkage error. Goldstein et al.[42] have applied modified MCMC methods originally developed for imputation to a range of general statistical analyses that adjust for linkage error.

CONCLUDING REMARKS

This paper provides some background on the FS model of record linkage, how 'optimal' parameters are estimated with the EM algorithm without training data, how false match rates are estimated with an unsupervised learning method that mimics ideas from semi-supervised learning, and current research in adjusting statistical analyses for linkage error. The paper does not cover the extensive computer science literature in entity resolution (record linkage).

ACKNOWLEDGMENTS

The views in this paper are those of the author and not necessarily those of the Census Bureau. The author thanks two referees for insightful and helpful comments.

REFERENCES

1. Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc* 1969, 64:1183–1210.
2. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. *Science* 1959, 130:954–959.
3. Newcombe HB, Kennedy JM. Record linkage: making maximum Use of the discriminating power of identifying information. *Commun ACM* 1962, 5: 563–567.
4. Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. *BMC Med Inform Decis Mak* 2013, 13:64. http://www.biomedicalcentral.com/1472-6947/13/64. (Accessed June 4, 2014).

5. Herzog TN, Scheuren F, Winkler WE. *Data Quality and Record Linkage Techniques*. New York: Springer; 2007.
6. Christen P, Churches T, Zhu JX. Probabilistic name and address cleaning and standardization, In: *The Australian Data Mining Workshop*, November, 2002. Available at: http://datamining.anu.edu.au/projects/linkage.html. (Accessed June 4, 2014).
7. Cohen WW, Sarawagi S. Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. In: *Proceedings of the ACM Knowledge Discovery and Data Mining Conference*, 2004, 89–98.
8. Agichtein E, Ganti V. Mining reference tables for automatic text segmentation. In: *ACM Knowledge Discovery and Data Mining Conference*, 2004, 20–29.
9. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 1990, 354–359. Available at: www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf. (Accessed June 4, 2014).
10. Cohen WW, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and addresses. In: *International Joint Conference on Artificial Intelligence*. *Proceedings of the Workshop on Information Integration on the Web*, Acapulco, Mexico, August 2003.
11. Cohen WW, Ravikumar P, Fienberg SE. A comparison of string distance metrics for name-matching tasks. In: *Proceedings of the ACM Workshop on Data Cleaning*, *Record Linkage and Object Identification*, Washington DC, August 2003.
12. Cooper WS, Maron ME. Foundations of probabilistic and utility-theoretic indexing. *J Assoc Comp Mach* 1978, 25:67–80.
13. Deming WE, Gleser GJ. On the problem of matching lists by samples. *J Am Stat Assoc* 1959, 54:403–415.
14. Michelson M, Knoblock CA. Learning blocking schemes for record linkage. In: *Proceedings of AAAI*, 2006.
15. Bilenko M, Kamath B, Mooney RJ. Adaptive blocking: learning to scale up record linkage. In: *Proceedings of the 6th IEEE International Conference on Data Mining*, 2006, 87–96.
16. Christen P. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Trans Knowl Data Eng* 2012, 24:1537–1555.
17. Kenig B, Gal A. Efficient entity resolution with MFI blocks. In: *Proceedings of VLDB*, Lyon, France, August 2009.
18. Winkler WE, Yancey WE, Porter EH. Fast record linkage of very large files in support of decennial and administrative records projects. In: *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 2010, CD-ROM. Available at: http://www.amstat.org/sections/srms/proceedings/y2010/Files/307067_57754.pdf. (Accessed June 4, 2014).
19. Winkler WE. Approximate string comparator search strategies for very large administrative lists. In: *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 2004, CD-ROM. Available at: http://www.census.gov/srd/papers/pdf/rrs2005-02.pdf. (Accessed June 4, 2014).
20. Titterington DM, Smith AFM, Makov UE. *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons; 1988.
21. Winkler WE. Improved decision rules in the Fellegi-Sunter model of record linkage. In: *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 1993, 274–279. Available at: http://www.census.gov/srd/papers/pdf/rr93-12.pdf. (Accessed June 4, 2014).
22. Winkler WE. On Dykstra's iterative fitting procedure. *Ann Probab* 1990, 18:1410–1415.
23. Meng X, Rubin DB. Maximum likelihood via the ECM algorithm: a general framework. *Biometrika* 1993, 80:267–278.
24. Agresti A. *An Introduction to Categorical Data Analysis*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2007.
25. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press; 1975.
26. Herzog TN, Scheuren F, Winkler WE. Record linkage. *WIREs: Comput Stat* 2010, 2:535–543.
27. Winkler WE. Matching and record linkage. In: Cox BG, Binder DA, Chinnappa BN, Christianson A, Colledge MA, Kott PS, eds. *Business Survey Methods*. New York: Wiley; 1995, 355–384. Available at: http://www.fcsm.gov/working-papers/wwinkler.pdf. (Accessed June 4, 2014).
28. Larsen MD, Rubin DB. Alterative automated record linkage using mixture models. *J Am Stat Assoc* 2001, 79:32–41.
29. Winkler WE. Machine learning, information retrieval, and record linkage. In: *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 2000, 20–29. Available at: http://nisla05.niss.org/affiliates/dqworkshop/papers/winkler.pdf. (Accessed June 4, 2014).
30. Winkler WE. Record linkage and Bayesian networks. In: *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 2002, CD-ROM. Available at: http://www.census.gov/srd/papers/pdf/rrs2002-05.pdf. (Accessed June 4, 2014).
31. Belin TR, Rubin DB. A method for calibrating false-match rates in record linkage. *J Am Stat Assoc* 1995, 90:694–707.
32. Scheuren F, Winkler WE. Regression analysis of data files that are computer matched. *Surv Methodol*

1993, 19:39–58. http://www.fcsm.gov/working-papers/scheuren_part1.pdf. (Accessed June 4, 2014).

33. Winkler WE. Automatic estimation of record linkage false match rates. In: *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 2006, CD-ROM. Available at: http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf. (Accessed June 4, 2014).
34. Nigam KA, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. *Mach Learn* 2000, 39:103–134.
35. Winkler WE. Near automatic weight computation in the Fellegi-Sunter model of record linkage. In: *Proceedings of the Fifth Census Bureau Annual Research Conference*, 1989, 145–155.
36. Christen P. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. New York: Springer; 2012.
37. Elmagarmid AK, Ipeirotis PG, Verykios V. Duplicate record detection: a survey. *IEEE Trans Knowl Data Eng* 2007, 19:1–16.
38. Naumann F, Hershel M. An introduction to duplicate detection. *Synth Lect Data Manage* 2010, 2:1–87.
39. Vapnik V. *The Nature of Statistical Learning Theory*. 2nd ed. Berlin: Springer; 2000.
40. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer; 2001.
41. Chipperfield JO, Bishop GR, Campbell P. Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data Survey. *Methodology* 2011, 37:13–24.
42. Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. *Stat Med* 2011, 31:3481–3493. doi: 10.1002/sim.5508.
43. Winkler WE. Cleanup and statistical analysis of sets of national files. In: *Proceedings of the Federal Committee on Statistical Methodology*. 2013. http://www.fcsm.gov/13papers/J1_Winkler_2013FCSM.pdf. (Accessed June 4, 2014).
44. Neter J, Maynes ES, Ramanathan R. The effect of mismatching on the measurement of response errors. *J Am Stat Assoc* 1965, 60:1005–1027.
45. Lahiri PA, Larsen MD. Regression analysis with linked data. *J Am Stat Assoc* 2005, 100:222–230.
46. Chambers R. Regression analysis of probability-linked data. *Statisphere* 2009, 4. Available at: http://www.statisphere.govt.nz/further-resources-and-info/official-statistics-research/series/volume-4-2009.aspx. (Accessed June 4, 2014).
47. Kim G, Chambers R. Regression analysis under incomplete linkage. *Comput Stat Data Anal* 2012, 56:2756–2770.
48. Kim G, Chambers R. Regression analysis under probabilistic multi-linkage. *Stat Neerl* 2012, 66:64–79.
49. Winkler WE, Scheuren F. How computer matching error effects regression analysis: exploratory and confirmatory analysis. *U.S. Bureau of the Census*, *Statistical Research Division Technical Report*, 1991.
50. Bhattacharya I, Getoor L. A latent Dirichlet allocation model for entity resolution. In: *Proceedings of the 6th SIAM Conference on Data Mining (SDM '06)*, 2006, 47–58 – best paper.
51. Tancredi A, Liseo B. A hierarchical Bayesian approach to matching and size population problems. *Ann Appl Stat* 2011, 5:1553–1585.
52. Liseo B, Tancredi A. Bayesian estimation of population size via linkage of multivariate normal data sets. *Surv Methodol* 2011, 27:491–505.

FURTHER READING

Winkler WE. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. In: *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 1988, 667–671. Available at: http://www.census.gov/srd/papers/pdf/rr2000-05.pdf.