RESEARCH REPORT SERIES
(*Statistics #2002-05)*

# Methods for Record Linkage and Bayesian Networks

William E. Winkler

Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: November 4, 2002

# Methods for Record Linkage and Bayesian Networks

William E. Winkler[1], william.e.winkler@census.gov 2002Oct10
U.S. Bureau of the Census, Room 3000-4, Washington, DC 20233-9100

Although terminology differs, there is considerable overlap between record linkage methods based on the Fellegi-Sunter model (JASA 1969) and Bayesian networks used in machine learning (Mitchell 1997). Both are based on formal probabilistic models that can be shown to be equivalent in many situations (Winkler 2000). When no missing data are present in identifying fields and training data are available, then both can efficiently estimate parameters of interest. When missing data are present, the EM algorithm can be used for parameter estimation in Bayesian Networks when there are training data (Friedman 1997) and in record linkage when there are no training data (unsupervised learning). EM and MCMC methods can be used for automatically estimating error rates in some of the record linkage situations (Belin and Rubin 1995, Larsen and Rubin 2001).

Keywords: likelihood ratio, Bayesian Nets, EM Algorithm

## 1. INTRODUCTION

*Record linkage* is the science of finding matches or duplicates within or across files. Matches are typically delineated using name, address, and date-of-birth information. Other identifiers such as income, education, and credit information might be used. With a pair of records, identifiers might not correspond exactly. For instance, income in one record might be compared to mortgage payment size using a crude regression function. In the computer science literature, *datacleaning or object identification* often refers to methods of finding duplicates.

In the model of record linkage due to Fellegi and Sunter (1969, hereafter FS), a product space $\mathbf{A} \times \mathbf{B}$ of records from two files A and B is partitioned into two sets *matches* M and *nonmatches* U. Pairs in M typically agree on characteristics such as first name, last name, components of date-of-birth, and address. Pairs in U typically have isolated (random) agreements of the characteristics. We use $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ to denote an arbitrary agreement pattern. For instance, $\gamma$ might be agreement on first name, agreement on last name, and agreement on date-of-birth. In the FS model, obtaining accurate estimates of the probabilities $P(\gamma \mid M)$ and $P(\gamma \mid U)$ are crucial to finding the best possible decision rules for separating matches M and nonmatches U. The conditional independence assumption CI is that $P(\gamma \mid C) = \prod_i P(\gamma_i \mid C)$ where the set C can be either M or U. Under CI, FS showed that it is possible to estimate $P(\gamma \mid M)$ and $P(\gamma \mid U)$ automatically without training data. For situations in which identifying information among matches is reasonably good, Winkler (1988) showed how to estimate $P(\gamma \mid M)$ and $P(\gamma \mid U)$ using the EM algorithm. The EM algorithm can provide optimal separation between M and U because its parameters can correspond to the form needed for the classification rule. If assumption CI is not made, then a general EM (Winkler 1989, 1993, Larsen 1996) can provide parameters yielding better separation between M and U. The theory for the general EM (Winkler 1990, 1993) is more general than the iterative scaling procedure of Della Pietra et al. (1997) in the sense that it allows general convex constraints rather than just linear constraints. The advantage of less general EM under assumption CI is that it yields computational speed-ups of orders between 100 and 10,000 in contrast to methods that use dependencies between variables.

Bayesian Networks are graphical networks (Lauritzen and Spiegelhalter 1989) that are often used in the machine learning literature. A Naïve Bayes Network is a Bayesian Network under assumption CI. Naïve Bayes networks are typically applied in situations in which representative training data are available. Naïve Bayes methods have been extended to situations in which a mixture of labeled training data and unlabeled data are used for text classification (Nigam et al. 2000). Parameter estimation was done using a version of the EM algorithm that is effectively identical to that used by Winkler (2000) and Larsen and Rubin (2001) when training data are not available. In the latter situations, assumption CI was not needed.

In record linkage, it is known that dropping assumption CI can yield better classification rules and automatic estimates of error rates (Winkler 1993, Larsen and Rubin 2001). This is true even in situations where training data are not available (unsupervised learning). Record linkage has the advantage that its natural small dimensionality (six to twenty) makes accounting for dependencies more computationally tractable. Characteristics of the data and of the computational algorithms can reduce the number of computational paths to produce good parameter estimates in unsupervised learning. In text classification and other general applications of Bayes Nets, the large dimensionality (from 1,000 to 200,000) often rules out using methods that account for dependencies between identifying variables. Accounting for all of the two-way dependencies (Sahami 1996, Dumais et al. 1998) did not yield improved text classification rules for Bayesian Networks. Accounting for selected interactions involving two or more interactions did improve text classification rules (Winkler 2000).

Nigam et al. (2000) demonstrated that if a small amount of labeled training data is combined with a moderate or relatively large amount of unlabeled data, then classification rules can be improved. The improvement is in contrast to those methods in which only labeled data are used in training. In general, machine learning, training data provides necessary structure so that parameter estimation can provide classification rules that perform relatively well. In contrast to unsupervised learning methods for which no training data are available, the training data drastically reduces the number of computational paths that are considered by the parameter-estimation algorithms. Unsupervised learning methods have typically performed very poorly for general machine learning classification rules.

The unsupervised learning methods of record linkage (Winkler 1988, 1993) performed relatively well because they were applied in a few situations that were extremely favorable. Five conditions are favorable application of the unsupervised EM methods. The first is that the EM must be applied to sets of pairs in which the proportion of matches M is greater than 0.05 (see Yancey 2002 for related work). The second is that one class (matches) must be relatively well-separated from the other classes. The third is that typographical error must be relatively low. For instance, if twenty percent of matches have first name pairs that are of the form (Robert, brother), (Bob, blank), or (Robert, James) then it may be difficult to separate matches from nonmatches. The fourth is that there must be redundant identifiers that overcome errors in other identifiers. The fifth is that parameters obtained under assumption CI yield good classification rules. Under the five favorable conditions, the number of computational paths considered by the EM algorithm is greatly reduced from the number of computational paths under general EM when the five assumptions may not hold.

This paper will demonstrate how to find situations when the five assumptions can be relaxed. The main intent is to focus on relatively parsimonious computational extensions of the narrowest EM methods. The extensions provide better parameter estimates in more general situations when the five favorable conditions do not necessarily hold. Ng and Jordan (2002) have observed that

Naïve Bayes classifiers can often perform well even though there are strong theoretical reasons why they should perform relatively poorly. We slightly extend those situations when Naïve Bayes (condition CI) is known to work well. The improvements reduce error rates and, where necessary, sizes of clerical review regions. As a long-term goal, our secondary intent to develop exploratory tools for improving matching efficacy in very large administration list situations when each list may contain between 100 million and 1 billion records. BigMatch technology (Yancey and Winkler 2002) that we use for the large matching situations is straightforward to update with the methods described in this paper.

The outline for this paper is as follows. In the second section, we cover background on the Fellegi-Sunter model, Bayes Networks, EM Algorithms, use of training data, and effects of typographical error in identifiers. In the third section, we describe variants of the EM algorithm and the empirical data files. The fourth section provides results. We give some discussion in the fifth section. The final section is concluding remarks.

## 2. **BACKGROUND**

This section gives basic background on record linkage and specific issues that relate to it. In the first subsection, we formally describe the main theory of Fellegi and Sunter. The second subsection covers Bayesian Networks and their relationship to Fellegi-Sunter theory. The third subsection specifically gives insights into the strengths and limitations of training data in the record linkage setting. The fourth subsection give reasons why typographical error and related representational differences affect and limit efficacy of EM methods. The fifth subsection describes how different sets of blocking criteria and specific ways of applying weak classifiers create suitable sets of pairs for later applying stronger classifiers for separating matches from nonmatches. In the sixth subsection, we cover extensions for approximate string comparison and the relative frequency of different value-states of strings (fields).

### 2.1. Fellegi-Sunter Model of Record Linkage

Fellegi and Sunter (1969) provided a formal mathematical model for ideas that had been introduced by Newcombe (1959, 1962, see also 1988). They provided many ways of estimating key parameters. 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:

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

where $\gamma$ is an arbitrary agreement pattern in a comparison space $\Gamma$. For instance, $\Gamma$ 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 $R > T_\mu$, then designate pair as a match.

If $T_\lambda \le R \le T_\mu$, then designate pair as a possible match and hold for clerical review. (2)

If $R < T_\lambda$, then designate pair as a nonmatch.

The cutoff thresholds $T_\mu$ and $T_\lambda$ 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. The region $T_\lambda \le R \le T_\mu$ is referred to 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. Figure 1c shows hypothetical cutoffs weights.

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 $T_\mu = T_\lambda$, then decision rule (1) can be used for separating records (correspondingly pairs) into those that are in one class from those that are not. 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*. A *false match* is a pair that is designated as a match and is truly a nonmatch. A *false nonmatch* is pair designated as a nonmatch and is a truly a match.

## 2.2. Bayesian Networks

Nigam et al. (2000) observed two strengths of Bayesian networks. The first is that the method is based on a formal probabilistic model that lends itself to statistical interpretation. The second is that it provides a straightforward way of combining labeled and unlabeled data during training. In most machine learning applications, only labeled training data for which the true classification status is known is used. Because training data are very expensive and unlabeled data are easy to collect, Nigam et al. (2000) showed how to combine moderate amounts of labeled data with varying amounts of unlabeled data to produce classification decision rules that improved on classification rules that were based on the moderate amounts of labeled data alone. They showed that too small an amount of labeled training would not yield suitable decision rules. Furthermore, they showed that, if too large an amount of unlabeled training data was combined with a moderate amount labeled training data, then decision rules could also be worse than those based on labeled data alone.

Nigam et al. (2000) and others (e.g., Ng and Jordan 2002) have shown that classification decision rules that are based on naïve Bayesian networks (i.e., conditional independence assumption) work well in practice. The conditional independence is useful because it makes computation much more tractable (Nigam et al. 2000, Winkler 1988). Varying authors have

observed that the fields in pairs are quite dependent and that the computed probabilities for pairs do not even remotely correspond to the true underlying probabilities. Winkler (1989, 1993) observed that, if dependencies are dealt with, computed probabilities can somewhat correspond to the true probabilities in a few situations. Dependencies can be computed with conventional hierarchical latent class methods as introduced by Winkler (1989, 1993) when the number of fields is moderate (say, 20 or less).

### 2.3. Use of Training Data

For general machine learning, it is well known that a suitable amount of representative data can yield good classifiers. A suitable amount of training data is generally O(m) where m are the number of parameters in the machine learning model (e.g., Ng and Jordan 2002). When m is moderately large, then the amount of training data O(m) can be quite large and expensive to obtain. The intuition of using a more moderate amount of training data with a relatively large amount of (inexpensive) unlabeled data is two-fold. First, the training data provides more structure for the computational algorithms in the sense that they can severely limit the number of computational paths. Second, when the unlabeled data is combined in a suitable manner with labeled training data, effective size of training data is increased. In some situations, Larsen and Rubin (2001) showed that training with unlabeled data and increasing amounts of labeled training data can significantly improve matching efficacy.

### 2.4. Errors in Identifiers

Record linkage and general text classification have a major difference in how they deal with typographical error. In text classification, most typographical error is in relatively commonly occurring words that can be corrected by spell checkers. In record linkage, many of the identifiers consist of components of the name, address, and date-of-birth for which there are no spell checkers that can be automatically run. There is a strong need for nearly automatic methods of dealing with differing amounts of typographical error in pairs of files. For instance, P(agree characteristic | M) and P(agree characteristic | U) can vary dramatically from one pair of files to another. More surprising is that these probabilities can vary substantially from an urban region to an adjacent suburban region even when identical sets of fields are used in the matching. Fellegi and Sunter (1969, see also Winkler 1994) indicated that part (or most) of the difference is due to differing amounts of typographical error. For instance, Winkler (1989) showed that P(agree first name | M) and P(agree age | M) vary dramatically from one region to the next. By more carefully modeling the effect of string comparators (see section 2.6 for string comparators) on the likelihood ratios (1) and improving the form of the application of EM under condition CI, it was possible to estimate optimal parameters for 450 different regions in the 1990 Decennial Census automatically. The amount of clerical review was reduced by 2/3 in comparison to the 1998 Dress Rehearsal Census.

Instead of just considering yes/no agreement between identifiers, we can consider various degrees of partial agreement. The EM can be extended in a straightforward fashion to deal with more than two value-states (i.e., agree/disagree) with different fields. Furthermore, parameter-estimation methods might account for the relative frequency of occurrence of certain fields. For instance, agreement on a relatively rare name such as Zabrinsky might have more value than agreement on a frequently occurring name such as Smith. The EM might be further extended to account for different relative frequency categories and the typographical errors that occur in them (see Winkler 1988).

## 2.5. Identifying Suitable Sets of Pairs

It is not possible to consider all pairs from two files A and B. In record linkage, we consider only those pairs agreeing on blocking criteria. One blocking pass might consider only those pairs agreeing on a geographic identifier and Soundex of surname. Another blocking pass might only consider pairs agreeing on house number, ZIP code, and the first initials of the first and last name. The idea of blocking is to find a set of pairs in which matches are concentrated. Multiple blocking passes are needed to find duplicates in a subsequent blocking pass that are not found on a prior pass. Due to high typographical error rates in most files (e.g., Winkler 1994, 1995), it is quite unusual to find all matches in just one blocking pass.

Unlike general text classification, in record linkage it is quite feasible to use an initial guess of parameters associated with agreements to get an enriched set of pairs within a blocking criteria. Virtually all matches will be concentrated in those pairs having matching weight (1) above a certain value. Yancey (2002) shows how to improve matching parameters within such classes of pairs via an EM algorithm.

## 2.6. Approximate String Comparison and Frequency

Many matches have typographical error in key identifying fields. For instance, in 1988 Dress Rehearsal Census data among pairs that are true matches, 20% of first names and 15% of last names did not agree on an exact character-by-character basis. Ages differed by more than 1 year with at least 15% of matches. To alleviate some of the effect of typographical error, we use string comparators (e.g., Winkler 1994, 1995) that return values between 1 for exact agreement and 0 for total disagreement. Table 1 provides a comparison of string comparator values. Bigrams are widely used in the computer science literature.

Table 1. Comparison of String Comparators Using Last Names and First Names

| Two strings | | String comparator values | | |
|---|---|---|---|---|
| | | Jaro | Winkler | Bigram |
| SHACKLEFORD | SHACKELFORD | 0.970 | 0.982 | 0.700 |
| DUNNINGHAM | CUNNIGHAM | 0.896 | 0.896 | 0.889 |
| NICHLESON | NICHULSON | 0.926 | 0.956 | 0.625 |
| JONES | JOHNSON | 0.790 | 0.832 | 0.204 |
| MASSEY | MASSIE | 0.889 | 0.933 | 0.600 |
| ABROMS | ABRAMS | 0.889 | 0.922 | 0.600 |
| HARDIN | MARTINEZ | 0.000 | 0.000 | 0.365 |
| ITMAN | SMITH | 0.000 | 0.000 | 0.250 |
| JERALDINE | GERALDINE | 0.926 | 0.926 | 0.875 |
| MARHTA | MARTHA | 0.944 | 0.961 | 0.400 |
| MICHELLE | MICHAEL | 0.869 | 0.921 | 0.617 |
| JULIES | JULIUS | 0.889 | 0.933 | 0.600 |
| TANYA | TONYA | 0.867 | 0.880 | 0.500 |
| DWAYNE | DUANE | 0.822 | 0.840 | 0.200 |
| SEAN | SUSAN | 0.783 | 0.805 | 0.289 |
| JON | JOHN | 0.917 | 0.933 | 0.408 |
| JON | JAN | 0.000 | 0.000 | 0.000 |

To utilize training data more effectively, we divide the agreement values associated with pairs into more value states according to ranges of the string comparators. This may allow us to model the effect of partial agreements more effectively. With strings such as first names and last names, we associate value-states of an identifier about whether the string is very frequently occurring, frequently occurring, approximately average occurring, and two categories of relatively rarely occurring. Winkler (e.g., 1994, 1995) developed relatively crude ways for dealing with the effect of string comparators and frequency that depending on a priori training data to get the shapes of various functions that downweight the likelihood ratio (1). The downweighting takes place as string comparator values decrease from one and the relative frequency of a value-state of string (e.g., Zabrinsky versus Smith) increases.

With training data, we may be able to find additional relationships that have not been previously conceived and modeled. Generally, accounting for partial agreement with string comparators makes dramatic improvements in matching efficacy (Winkler 1990b, 1995). From one pair of files to the next, typographical error rates can dramatically affect the probabilities P(agree field | M). For instance, in an urban area or a rural area, the P(agree first name | M) and P(agree last name | M) may be significantly lower than for a corresponding suburban region. Figure 2 illustrates the situation. The curves of log frequency versus weight for matches can overlap more in an urban region (Figure 2a) and a rural region (Figure 2c) than in a suburban region (Figure 2b). In the overlap regions between the curves of matches and nonmatches, having suitably chosen training data may help in finding better estimates of error rates.

If high quality, current geographic identifiers are associated with records, then accounting for frequency may not help matching (Winkler 1989, Yancey 2000). Across larger geographic regions (e.g., an entire ZIP code or County or State), accounting for frequency may improve matching efficacy. By *frequency*, we mean accounting for specific value-states of a string such as last name. A frequently occurring string such as Smith may have less distinguishing power than a less frequently occurring string such as Zabrinsky.

## 3. METHODS AND DATA

Our main theoretical method is to use the EM algorithm and maximum likelihood to obtain parameters and associated classifiers for separating $\mathbf{A} \times \mathbf{B}$ into matches M and nonmatches U. The data files are Decennial Census files for which the truth of classification is known. The truth is obtained through several levels of clerical review and field followup.

### 3.1. EM Methods

In the models of Nigam et al. (2000) and of this paper, words (comparison fields) are used to classify documents into different classes. Our development is identical theoretically to that of Nigam et al. Our notation differs very slightly because it deals more with the representational framework of record linkage. Let $\gamma_i$ 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 subsets of M and the remaining $C_j$ are subsets of U. Unlike general text classification in which every document may have a unique agreement pattern, in record linkage, some agreement patterns $\gamma_i$ may have many pairs $p_{i(l)}$ associated with them. Here l will run through an appropriate index set. Specifically,

$$P(\gamma_i \mid \Theta) = \sum_i^{|C|} P(\gamma_i \mid C_j ; \Theta) P(C_j ; \Theta) \qquad (4)$$

where $\gamma_i$ is a specific pair, $C_j$ is a specific class, and the sum is over the set of classes. Under the Naïve Bayes or conditional independence, we have

$$P(\gamma_i \mid C_j ; \Theta) = \prod_k P(\gamma_{i,k} \mid C_j ; \Theta) \qquad (5)$$

where the product is over the kth individual field agreement $\gamma_{i,k}$ in pair agreement pattern $\gamma_i$. In some situations, we use a Dirichlet prior

$$P(\Theta) = \prod_j ( \Theta_{Cj} )^{\alpha-1} \quad \prod_k ( \Theta \gamma_{i,k \mid Cj} )^{\alpha-1} \qquad (6)$$

where the first product is over the classes $C_j$ and the second product is over the fields. Nigam et al. (2000) set $\alpha$ equal to two and refer to the effect of the prior as Laplace smoothing. The prior (6) helps keep most of the estimated probabilities away from zero. We use $D^u$ to denote unlabeled pairs and $D^l$ to denote labeled pairs. Given the set D of all labeled and unlabeled pairs, the log likelihood is given by

$$\begin{aligned} l(\Theta \mid D) = & \log ( P(\Theta)) + \\ & \sum_{i \epsilon Du} \log \sum_j P(\gamma_i \mid C_j ; \Theta) \; P(C_j ; \Theta) + \\ & \sum_{i \epsilon Dl} \log \sum_j P(\gamma_i \mid C_j ; \Theta) \; P(C_j ; \Theta) \end{aligned} \qquad (7)$$

where the first sum is over the unlabeled pairs and the second sum is over the labeled pairs. If we let $z_{ij}$ be a missing data indicator that pair i in class j is observed, then we have the complete data equation (CDE)

$$\begin{aligned} l_c(\Theta \mid D; z) = & \log ( P(\Theta)) + \\ & \sum_{i \epsilon D} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta) \; P(C_j ; \Theta)) \end{aligned} \qquad (8)$$

where the first sum is over all pairs and the second sum is over the classes. If labeled and unlabeled pairs are mixed in proportions $\lambda$ and $1-\lambda$, $0< \lambda < 1$,we have

$$\begin{aligned} l_c(\Theta \mid D; z) = & \log ( P(\Theta)) + \\ & (1-\lambda ) \sum_{i \epsilon Du} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta) \; P(C_j ; \Theta)) + \\ & \lambda \sum_{i \epsilon Dl} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta) \; P(C_j ; \Theta)). \end{aligned} \qquad (9)$$

We use the EM algorithm to estimate (9). The specific form of the EM algorithm depends on the exact parametric form that we assume for $P(\gamma_i \mid C_j ; \Theta) \; P(C_j ; \Theta)$. Under condition CI, we let

$$P(\gamma_i \mid C_j ; \Theta) = \prod_k \mu_{jk}^{\gamma k} (1-\mu_{jk})^{(1-\gamma k)} \qquad (10)$$

where the product is over all the comparison fields and $\gamma_k$ is an indicator (or value-state) of the kth field is observed in the pair $p_i$. The starting points for the EM might be the estimates of $\mu_{jk}$ and $P(C_j ; \Theta)$ that are available from the labeled data. Under the conditional independence assumption, if $\Theta^t = (\mu_{jk}^t , P^t (C_j ; \Theta) : j, k)$ is the current estimate of $\Theta$, then

$$\mu_{jk}^{t+1} = [(\alpha-1) + (1-\lambda) \sum_{i \epsilon Du} E(z_{ij} \mid C_j)\, \gamma_k + \lambda \sum_{i \epsilon Dl} E(z_{ij} \mid C_j)\, \gamma_k] / [2(\alpha-1) + (1-\lambda) \sum_{i \epsilon Du} E(z_{ij} \mid C_j)\, 1 + \lambda \sum_{i \epsilon Dl} E(z_{ij} \mid C_j)\, 1] \quad (11)$$

and

$$P^{t+1}(C_j ; \Theta) = [[(\alpha-1) + (1-\lambda) \sum_{i \epsilon Du} E(z_{ij} \mid C_j) + \lambda \sum_{i \epsilon Dl} E(z_{ij} \mid C_j)]/ [\,|C|(\alpha-1) + (1-\lambda) \sum_{i \epsilon Du} 1 + \lambda \sum_{i \epsilon Dl} 1] \quad (12)$$

If expected values $E(z_{ij} \mid Cj)$ are substituted in the (9), then Equation (11) follows by taking partial derivatives and setting the resultant equation equal to zero. Equation (12) follows by standard multinomial reasoning (e.g., McLachlan and Krishnan, pp. 17-19). The parameter $\alpha$ can be varied independently for $\mu_{jk}$ and $P(C_j ; \Theta)$. For the empirical example, we vary $\alpha$ between 1.00001 and 1.001. The smoothing via different values of $\alpha$ in the prior causes the successive estimates $\mu_{jk}^{t+1}$ and $P^{t+1}(C_j ; \Theta)$ to stay away from zero and one. Because of the relatively low dimensionality of record linkage problems, we can consider such small values of $\alpha$. We note that, under condition CI, the maximization step is in closed form.

An alternative form of smoothing is to add a small value $\delta$ to every cell as suggested by Larsen (1996). With a moderate number of fields or with different parametric representations of the fields in the pair, some differing pairs can have the same representation. In some instances, differing pairs may have the same representation in fields. This can happen with a small number of fields and when different string values of agreement are not explicitly included. We use freql(i, j) to represent the frequency of the $j^{th}$ pattern in the $i^{th}$ class of the labeled pairs and frequ(j) be the frequency of the $j^{th}$ pattern in the unlabeled pairs. With a slight abuse of notation, we let the sum over the labeled pairs to be over all of the observed patterns in the labeled and unlabeled data. The understanding is that, for a given j, freql(i,j) is zero for all i if a pattern is observed only in the unlabeled data. Similar to Equation (9) we have,

$$l_c(\Theta \mid D; z) = (1-\lambda) \sum_{i \epsilon Du} \sum_j z_{ij}\, \text{frequ}(j) \log (P(\gamma_i \mid C_j ; \Theta)\; P(C_j ; \Theta)) + \lambda \sum_{i \epsilon Dl} \sum_j z_{ij}\, (\text{freql}(i,j) + \delta) \log (P(d_i \mid C_j ; \Theta)\; P(C_j ; \Theta)). \quad (13)$$

In Equation (13), the value $\delta$ is added to each cell in every class of every observed data pattern from the labeled and unlabeled data. In analogy to Equations (9) and (10), we have estimates at step t of

$$\mu_{jk}^{t+1} = [(1-\lambda) \sum_{i \epsilon Du} \text{frequ}(j)\, E(z_{ij} \mid Cj)\, \gamma_k + \lambda \sum_{i \epsilon Dl} (\text{freql}(i,j) + \delta)\, E(z_{ij} \mid Cj)\, \gamma_k] / [(1-\lambda) \sum_{i \epsilon Du} \text{frequ}(j)\, E(z_{ij} \mid Cj)\, 1 + \lambda \sum_{i \epsilon Dl} (\text{freq}(i,j) + \delta)\; E(z_{ij} \mid Cj)\, 1] \quad (14)$$

and

$$P^{t+1}(C_j ; \Theta) = [[(1-\lambda) \sum_{i \epsilon Du} E(z_{ij} \mid Cj) + \lambda \sum_{i \epsilon Dl} (freql(i,j) + \delta) E(z_{ij} \mid Cj)]/ [(1-\lambda) \sum_{i \epsilon Du} 1 + \lambda \sum_{i \epsilon Dl} (freql(i,j) + \delta) 1] \quad (15)$$

The specific computational procedure can be best understood if the $z_{ij}$ in Equation (13) can be replaced by $E(z_{ij} \mid \Theta^t)$

$$L_e(\Theta^{t+1} \mid D; z) = (1-\lambda) \sum_{i \epsilon Du} \sum_j E(z_{ij} \mid \Theta^t) frequ(j) \log (P(d_i \mid C_j ; \Theta^t) P(C_j ; \Theta^t)) + \lambda \sum_{i \epsilon Dl} \sum_j E(z_{ij} \mid \Theta^t) (freql(i,j) + \delta) \log (P(d_i \mid C_j ; \Theta^t) P(C_j ; \Theta^t)). \quad (16)$$

We can assume that both first summations are over all of the observed patterns in the labeled and unlabeled by setting frequ(j) and freql(j) equal to zero when j is a pattern that is not in the unlabeled data and labeled data, respectively. If we renormalize, the coefficients in front of the logs so that the terms add to one (which does affect the maximization of the likelihood), then we have equations of the following form

$$L_e(\Theta^{t+1} \mid D; z) = \sum_{ij} p_{et}(i,j) \log (p_t(i,j)) \quad (17)$$

where $p_{et}(i,j) = ((1-\lambda) E(z_{ij} \mid \Theta^t) frequ(j) + \lambda E(z_{ij} \mid \Theta^t) (freql(i,j) + \delta))/N_C$, $N_C$ is the normalization constant, and $p_t(i,j) = P(d_i \mid C_j ; \Theta^t) P(C_j ; \Theta^t)$. Let $P_j$ be the interaction patterns that are to be fit in class $C_j$. Each interaction pattern in $P_j$ represents a listing of the fields (terms) that must be summed over. For pattern i in $P_j$, let $I_i$ represent the specific subsets l of fields. For instance, if $P_i$ represents the presence of k specific fields in a pair, then $I_i$ has $2^k$ subsets. The $2^k$ subsets in $I_i$ partition the entire set of pairs. In the following, the notion $i \in l$ means that the pair i has the pattern of fields represented by l. The specific fitting procedure $F_t$ at step t is:

1. For each pattern i in $P_j$ and each l in $I_i$, let $M_{tl} = \sum_{i \in l} p_t(i,j)$ and $E_{tl} = \sum_{i \in l} p_{et}(i,j)$. For each class $k \neq j$, let $M_k = \sum_i p_t(i,k)$ and $E_k = \sum_i p_{et}(i,k)$.
2. If $i \in l$ in $P_j$, then $p_{t+1}(i,j) = p_t(i,j) E_{tl} / M_{tl;}$ and, if $k \neq j$ , $p_{t+1}(i,k) = p_t(i,k) E_k / M_k$.
3. Repeat 1 and 2 for all classes $C_j$ and all patterns i in $P_j$.

Then each $F_t$ is one cycle of iterative proportional fitting (e.g., Winkler 1989, 1993, Meng and Rubin 1993) and increases the likelihood. The last equation in step 2 assures that the new estimates add to a proper probability. If necessary, the procedure can be extended to general I-Projections that also increase the likelihood and have strong constraints for keeping the probability estimates $p_t(i,j)$ from converging to zero or one (e.g., Winkler 1990a). The smoothing with the constant delta in Equation (18) has the effect of assuring that most probability estimates $p_t(i,j)$ do not converge to zero. For a fixed pattern i, some of the probability estimates $p_t(i,j)$, however, may differ by several orders of magnitude across the different classes $C_j$. If necessary, affine constraints may be use to restrict the differing relative sizes of the $p_t(i,j)$ (Winkler 1990a).

We observe that if λ is 1, then we only use training data and our methods correspond to naïve Bayes methods in which training data are available. If λ is 0, then we are in the unsupervised learning situations of Winkler (1993) and Larsen (1996).

3.2. Data Files

Three pairs of files were used in the analyses. The files are from 1990 Decennial Census matching data 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 followup that were not incorporated in computer files available to us. We did not use 2000 Decennial Census data because we received files too late to verify the accuracy of match status codes. The later files will be used in work subsequent to the work of this paper.

A summary of the overall characteristics of the empirical data is in Table 2. We only consider pairs that agree on census block id 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.

Table 2. Summary of Three Pairs of Files

| | Files | | Files | | Files | |
|---|---|---|---|---|---|---|
| | $A_1$ | $A_2$ | $B_1$ | $B_2$ | $C_1$ | $C_2$ |
| Size | 15048 | 12072 | 5022 | 5212 | 4539 | 4851 |
| # pairs | 116305 | | 37327 | | 38795 | |
| # matches | 10096 | | 3623 | | 3490 | |

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 help distinguish individuals within household. Some pairs have both missing first name and age. In the initial results, all comparisons are considered agree/disagree (base 2). This basic situation corresponds to matching comparisons that were used in matching systems in 1990 and 2000 Decennial Censues. The eight data fields yield 256 data patterns for which frequencies (proportions) are calculated. If one or both identifiers of a pair are blank, then the comparison (blank) is considered a disagreement. This only substantially affects age (15% blank) and phone (35% blank). Name and address data are almost never missing.

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]. The first interval is what we refer to as disagreement. We combine the disagreement with the three partial agreements and blank 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.

The pairs naturally divide into three classes: $C_1$- match within household, $C_2$ - nonmatch within household, $C_3$ – nonmatch outside household. Although we considered additional dependency models, we only consider the two models. The first is of Larsen and Rubin (2001), called $g_1$: I,HP,HP, in which we fit a conditional independence model in class $C_1$ and 4-way interaction models in classes $C_2$ and $C_3$. The second is similar to ones considered by Winkler

(1993). It is called $g_3$: HP+,HP+,HP+ in which we fit slightly more interactions than in $g_1$ in all three classes. The analysis framework is quite flexible. It is summarized in Table 2. We refer to the different five components in Table 3 as the metaparameters of the modeling framework. This is reasonably consistent with how Hastie, Thibshirani, and Friedman (2001) describe metaparameters. Metaparameters are generally chosen based on some knowledge of the data and the potential classification rules.

Table 3. Metaparameters of the Modeling

1. Models –
   a. CI – independent – *i1(i0 – 1990 version) I,I,I*
   b. Larsen-Rubin CI in class 1, 4-way person,
      4-way household in classes 2 and 3, *g1 I,HP,HP*
   c. Winkler 4+ way interactions in all classes, *g3(g0 1990 version)*
2. lambda – how much to emphasize training data
3. delta – 0.000001 to 0.001 – smooth out peaks
4. how many iterations
5. number of degrees of partial agreement
   a. agree, disagree (and/or blank) [small base = 2]
   b. very close agree, moderately close agree, somewhat agree, blank, disagree [large base = 5]

We draw relatively small and relatively large samples of training data. The sample sizes are summarized in Table 4.

Table 4. Training Data Counts with Proportions of Matches

| | A | B | C |
|---|---|---|---|
| Large Sample | 7612 (0.26) | 3031 (0.29) | 3287 (0.27) |
| Small Sample | 588 (0.33) | 516 (0.26) | 540 (0.24) |

The overall comparisons are summarized in Table 5.

Table 5. Summary of Comparison Scenarios

| **1990** | **2002** |
|---|---|
| yes/no | 3-level yes, blank, no |
| CI (i0), interact (g0) | CI (i1), interaction (g1, g3) |
| I,I,I ; HP+,HP+,HP+ | I,I,I ; I,HP,HP; HP+,HP+,HP+ |
| 1-1, non-1-1 | 1-1, non-1-1 |
| no delta | delta smoothing |

Under each of the scenarios, we do unsupervised learning ($\lambda <= 0.001$) and supervised learning ($\lambda = 0.9, 0.99$ or $0.999$). In the supervised learning situation, we use both large and small samples.

We have two overall measures of success. The first is applied only when we use 1-1 matching. At a set of fixed error levels (0.002, 0.005, 0.01, and 0.02), we consider the number of pairs

brought together and the proportion of matches that are obtained. This corresponds to production matching systems used in 1990 and 2000 Decennial Censuses. The second is applied only when we use non-1-1 matching. We determine how accurately we can estimate the lower cumulative distributions 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. This is known to be an exceptionally difficult problem (e.g. Vapnik 1999, Hastie, Thibshirani, and Friedman 2001). Our comparisons consist of a set of figures in which we compare a plot of the cumulative distribution of estimates of matches versus the true cumulative distribution with the truth represented by the 45 degree line. We also do this for nonmatches. As the plots get closer to the 45 degree lines, the estimates get closer to the truth.

## 4. **RESULTS**

The results are divided into two subsections. In the first, we consider results from 1-1 matching. In the second, much more difficult situation, we consider the estimation of the tails of distributions.

### 4.1. Results under 1-1 Matching

In Table 5, we provide results from 1-1 matching. At differing error rate levels and in the differing files, the 1990 matching procedures that (also used in 2000) were nearly as effective as the newer procedures. Use of the larger base sometimes improves results by 0.005 and use of interaction models also sometimes improves results by 0.005.

Table 5. Matching efficacy for 1-1 matching

Counts of pairs and proportions of true matches

| Error level | | | |
|---|---|---|---|
| 0.002 | File A | File B | File C |
| g3 | 9780 (0.967) | 3428 (0.944) | 3225 (0.922) |
| g1 | 9741 (0.965) | 3448 (0.950) | 3261 (0.932) |
| i1 | 9640 (0.956) | 3277 (0.903) | 3042 (0.867) |
| i0 | 9701 (0.959) | 3489 (0.961) | 3306 (0.945) |
| g0 | 9649 (0.954) | 3422 (0.943) | 3273 (0.936) |
| 0.005 | | | |
| g3 | 9882 (0.974) | 3547 (0.974) | 3409 (0.972) |
| g1 | 9868 (0.973) | 3523 (0.967) | 3386 (0.965) |
| i1 | 9855 (0.971) | 3513 (0.965) | 3314 (0.945) |
| i0 | 9857 (0.971) | 3540 (0.972) | 3379 (0.963) |
| g0 | 9810 (0.967) | 3511 (0.964) | 3329 (0.949) |
| 0.010 | | | |
| g3 | 9955 (0.976) | 3584 (0.979) | 3452 (0.979) |
| g1 | 9948 (0.976) | 3568 (0.975) | 3441 (0.976) |
| i1 | 9942 (0.975) | 3566 (0.974) | 3414 (0.968) |
| i0 | 9952 (0.976) | 3580 (0.978) | 3431 (0.973) |
| g0 | 9878 (0.969) | 3536 (0.966) | 3372 (0.956) |
| 0.020 | | | |
| g3 | 10062 (0.976) | 3622 (0.980) | 3491 (0.980) |
| g1 | 10057 (0.976) | 3614 (0.978) | 3487 (0.979) |
| i1 | 9942 (0.976) | 3614 (0.978) | 3481 (0.977) |
| i0 | 10065 (0.977) | 3623 (0.980) | 3489 (0.980) |
| g0 | 9998 (0.970) | 3589 (0.971) | 3417 (0.960) |

i0,i1 independence; g0, g1, g3 interaction models

The reason that we do not show results for error rates above two percent (0.02) is that almost all of the pairs brought together are nonmatches below the cutoff associated with the two percent error rate.

4.2. Results under non-1-1 Matching

Figures 3-11 represent results from non-1-1 matching. In Figures 3-5, we use the small base 2 and in Figures 6-11, we use the large base. In Figures 3, 4, and 5, we consider unsupervised learning for the interaction model $g_3$, interaction model $g_1$, and conditional independence model $i_1$, respectively. All results are for base 2. All models perform poorly, particularly in the lowest 10 percent of the cumulative distribution of matches. The independence model, possibly somewhat surprisingly performs, slightly better than the two interaction models. The EM algorithm when applied under independence restraints does better than the two interaction models in estimating an accurate proportion for the class of matches within a household. Interaction model $g_1$ yields estimates of this proportion that are slightly too low; interaction model $g_3$ estimates this proportion that are too slightly high.

In Figures 6 and 7, we consider unsupervised learning for the interaction models $g_3$, and conditional independence model $i_1$, respectively. All results are large base results. All models perform poorly, particularly in the lowest 10 percent of the cumulative distribution of matches. Again, the independence model, possibly somewhat surprisingly performs, slightly better than the interaction model. Interaction model $g_1$ is not shown because it performs slightly more poorly than interaction model $g_3$. Again, interaction model $g_1$ yielded estimates of the proportion of the matches within a household that are slightly too low; interaction model $g_3$ yields estimates of this proportion that are slightly too high.

In Figures 8 and 9, we consider small sample training with lambda mixing proportions of 0.9 and 0.999, respectively. The lambda mixing proportion 0.999 provides excellent accuracy in the crucial range of 0.0-0.1. These results are better than the results with lambda mixing proportion of 0.9. Only the independence model is shown because it provides the best accuracy. The large sample results of Figures 10 and 11 are slightly better than the small sample results given in Figures 8 and 9. The independence model in the large sample situation provides excellent accuracy. Although not shown, the two interaction models in the large sample situation provide even better accuracy than the independence model.

## 5. **DISCUSSION**

From large amounts of previous empirical work, we know that typographical error can vary significantly between two pairs of files. We define typographical error as any difference in two corresponding fields for a character-by-character comparison. The typographical error can be significantly different for two pairs of files representing two adjacent regions such as an urban region and one of its suburban regions. This particularly affects the probability of agreement on first name given a match and the probability of agreement on age given a match. These typographical differences can occur because there are genuine differences in the quality of the files from two different regions. These differences can be compounded if one individual somehow makes significant keypunch errors while others do not. Then one portion of a file may contain substantial typographical error that affects the shape of the curve of matches. These types of typographical variations can occur in administrative lists that are melded from several different sources files. One or more of the sources files may have significant error.

The fact that the (partially) supervised learning illustrated in Figures 8-11 performs much better than the corresponding unsupervised learning results given in Figures 3-7 is not surprising.

We need a representative sample of pairs to determine the shape of the lowest part of the cumulative match curve. In some pairs of files, 0.001 of the pairs for a fixed agreement pattern may be matches. In another pairs of seemingly similar files, 0.02 of the pairs for the same fixed agreement pattern may be matches. For instance, in pairs in which first name and age are missing, the proportions of matches can vary between 0.001 and 0.02.

In the case of unsupervised learning, the interaction models have strong tendencies to put extra pairs in class $C_1$ or fewer pairs in class $C_1$ depending on whether dependencies are fit in class $C_1$ or conditional independence is fit in class $C_1$. We have not yet investigated whether fitting with convex constraints (Winkler 1993) that force certain of the estimated probabilities into narrower ranges may help. Early results are not too promising. Given the extreme differences in typographical error rates for individual agreement patterns over differing, but similar, types of files, we are not optimistic.

We still find the high quality of the estimates of the cumulative distributions (Figures 8-11) under conditional independence to be somewhat surprising. In future work, we will investigate this. We do note that, within the set of matches, 40-45 percent agree on every field almost on a character-by-character basis.

## 6. CONCLUDING REMARKS

This paper examines methods for weakening some of the stringent conditions that were implicitly assumed in earlier applications of EM parameter estimation to record linkage. The EM-based estimation methods potentially yield better parameters for separating matches from nonmatches when they are applied in appropriate situations. The estimates improve over iterative refinement methods that proceed through a series of clerical reviews and expert guessing such as are available in certain commercial record linkage systems. They are also far faster and better use resources than iterative refinement methods.

1/ This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

## REFERENCES


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 (available at http://www.fcsm.gov under methodology reports).

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.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997), “Inducing Features of Random Fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 380-393.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998), “Inductive Learning Algorithms and Representations for Text Categorization,” In *Proceedings of the 7th International Conference on Information and Knowledge Management*.

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

Friedman, J. (2001), “Greedy Function Approximation: A Gradient Boosting Machine,” Annals of Statistics, **29** (5), 1389-1432.

Friedman, N. (1997), “Learning Belief Networks in the Presence of Missing Values and Hidden Variables,” *in D. Fisher, ed., ‘Proceedings of the Fourteenth International Conference on Machine Learning’, Morgan Kaufmann, San Francisco, CA, pp. 125-- 133*.

Friedman, N. (1999), "The Bayesian Structural EM Algorithm," *in G. F. Cooper & S. Moral, eds., 'Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)', Morgan Kaufmann, San Francisco, CA.*

Friedman, N., Geiger, D., and Goldszmidt, M. (1997), "Bayesian Network Classifiers," *Machine Learning*, **29**, 131-163.

Friedman, J., Hastie, T. and Tibshirani, R. (2000), "Additive Logistic Regression: a Statistical View of Boosting," *Annals of Statistics*, **28**, 337-407.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer: New York.

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.

Lauritzen, S. L. (1996), *Graphical Models*, Clarendon Press, Oxford, United Kingdom.

Laurtizen, S. L. and Spiegelhalter, D. J. (1989), "Local Computation with Probabilities on Graphical Structures and Their Application to Expert Systems (with discussion)," *Journal of the Royal Statistical Society, B*, **50**, 157-224.

Larsen, M. D. (1996), "Bayesian Approaches to Finite Mixture Models," Ph.D. Thesis, Harvard University.

Larsen, M. D., and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, **79**, 32-41.

McLachlan, G. J., and T. Krisnan, (1997), *The EM Algorithm and Extensions*, John Wiley: New York.

Meng, X., and Rubin, D. B. (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.

Ng, A. and Jordan, M. (2002), "On Discriminative vs. Generative classifiers: A comparison of logistic regression and naïve Bayes," Neural Information Processing Systems 14, to appear.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000), "Text Classification from Labeled and Unlabelled Documents using EM, *Machine Learning*, **39**, 103-134.

Sahami, M. (1996), "Learning Limited Dependence Bayesian Classifiers," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR: AAAI, 335-338.

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

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. (1989), "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. (1990a), "On Dykstra's Iterative Fitting Procedure," *Annals of Probability*, **18**, 1410-1415.

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 Association*, 354-359.

Winkler, W. E. (1993), "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).

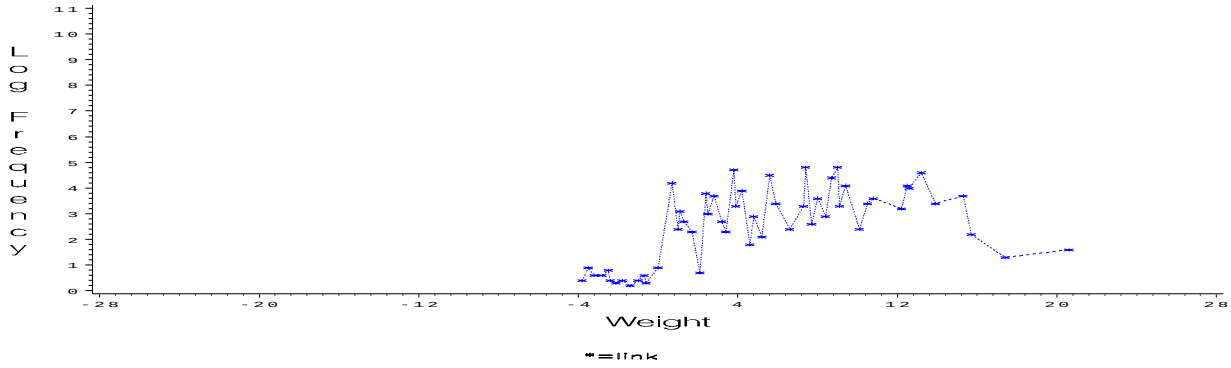
Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al.* (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.

Winkler, W. E. (1999), "The State of Record Linkage and Current Research Problems," *Statistical Society of Canada, Proceedings of the Section on Survey Methods*, 73-79 (longer version report rr99/04 available at http://www.census.gov/srd/www/byyear.html).

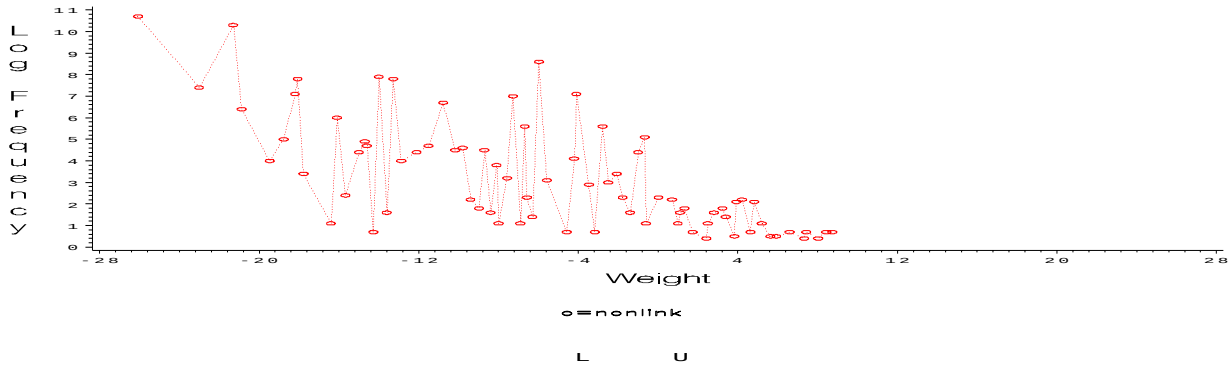
Winkler, W. E. and Thibaudeau, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," U.S. Bureau of the Census, Statistical Research Division Technical report RR91/09 (available at http://www.census.gov/srd/www/byyear.html).

Yancey, W. E. (2000), “Frequency Dependent Probability Measures for Record Linkage,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 752-757.
Yancey, W. E. (2002), “Improving EM Algorithm Estimates for Record Linkage Parameters,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.
Yancey, W. E. and Winkler, W. E. (2002), “BigMatch software,” computer system, documentation is in research report RRC2002/01 at http://www.census.gov/srd/www/byyear.html.

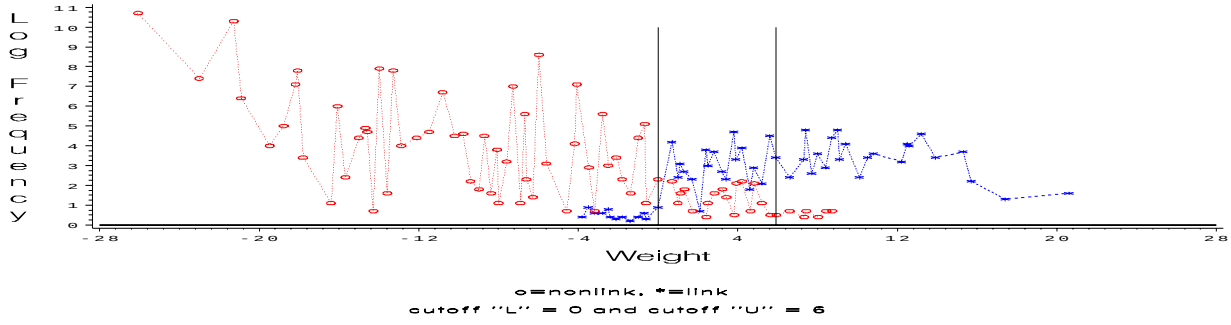
## Figure 1a. Log Frequency vs Weight Links



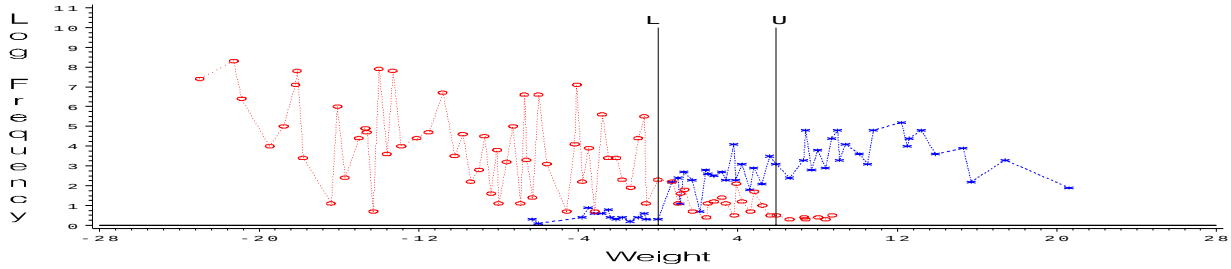

## Figure 1b. Log Frequency vs Weight Nonlinks




## Figure 1c. Log Frequency vs Weight Links and Nonlinks Combined

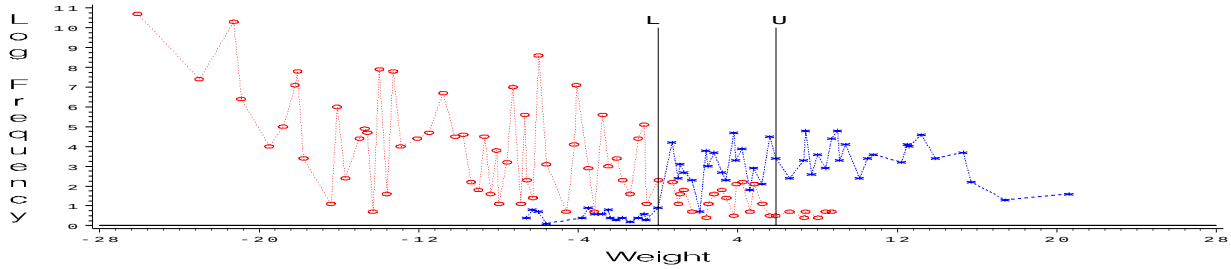


## Figure 2a. Log Frequency vs Weight
## Adjacent Suburban Area



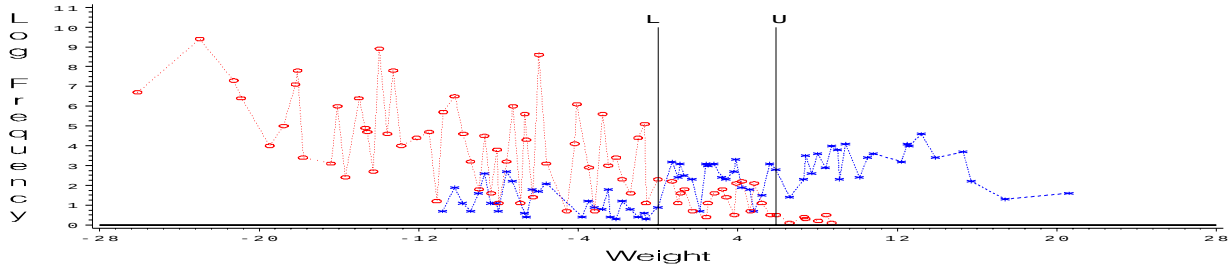

o=nonlink, *=link
cutoff "L" = 0 and cutoff "U" = 6

## Figure 2b. Log Frequency vs Weight
## Adjacent Urban Area




o=nonlink, *=link
cutoff "L" = 0 and cutoff "U" = 6

## Figure 2c. Log Frequency vs Weight
## Very Difficult Rural Area




o=nonlink, *=link
cutoff "L" = 0 and cutoff "U" = 6

Figure 3a. Estimates vs Truth, File A
Cumulative Distribution of Matches
3rd Dependent EM, non−1−1

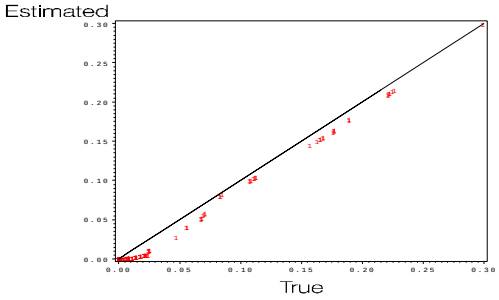



Figure 3b. Estimates vs Truth, File A
Cumulative Distribution of Nonmatches
3rd Dependent EM, non−1−1

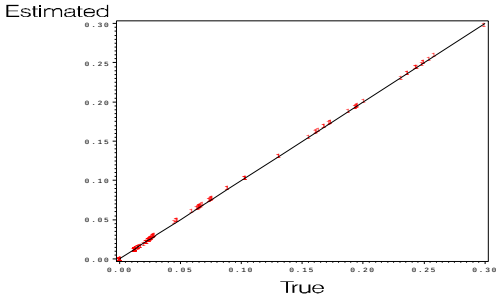



Figure 3c. Estimates vs Truth, File B
Cumulative Distribution of Matches
3rd Dependent EM, non−1−1

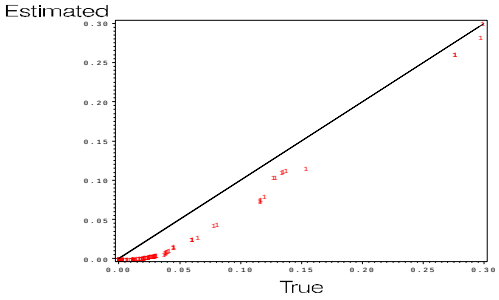



Figure 3d. Estimates vs Truth, File B
Cumulative Distribution of Nonmatches
3rd Dependent EM, non−1−1

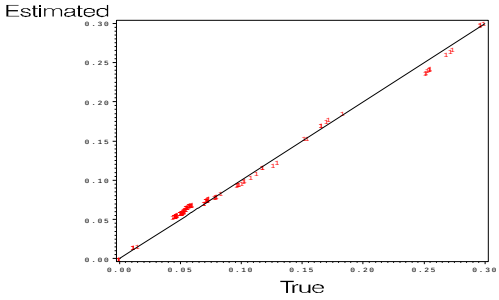



Figure 3e. Estimates vs Truth, File C
Cumulative Distribution of Matches
3rd Dependent EM, non−1−1

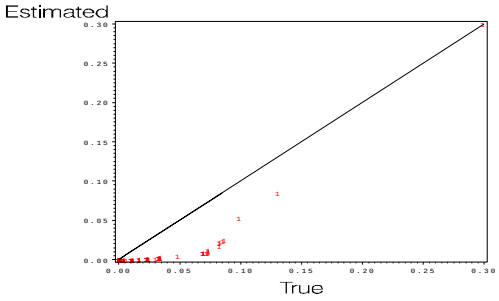



Figure 3f. Estimates vs Truth, File C
Cumulative Distribution of Nonmatches
3rd Dependent EM, non−1−1

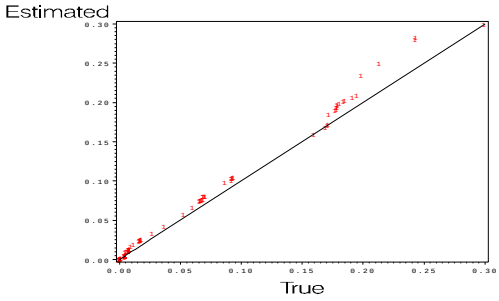


Figure 4a. Estimates vs Truth, File A
Cumulative Distribution of Matches
1st Dependent EM, non−1−1

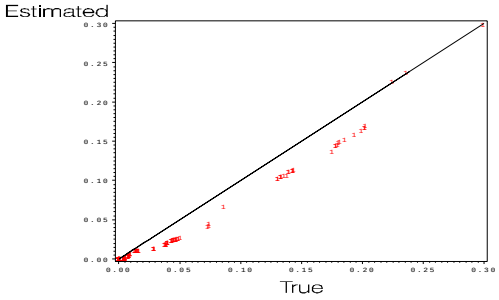



Figure 4b. Estimates vs Truth, File A
Cumulative Distribution of Nonmatches
1st Dependent EM, non−1−1

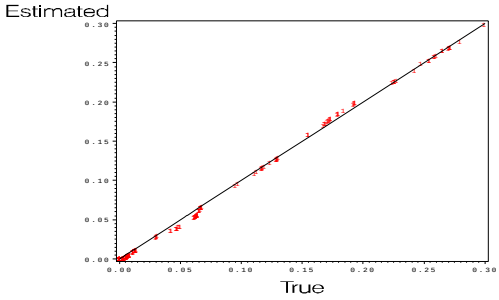



Figure 4c. Estimates vs Truth, File B
Cumulative Distribution of Matches
1st Dependent EM, non−1−1

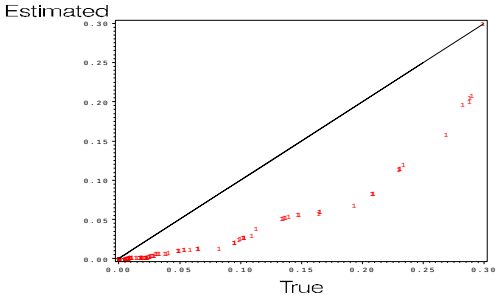



Figure 4d. Estimates vs Truth, File B
Cumulative Distribution of Nonmatches
1st Dependent EM, non−1−1

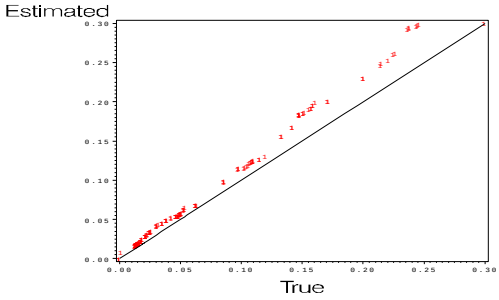



Figure 4e. Estimates vs Truth, File C
Cumulative Distribution of Matches
1st Dependent EM, non−1−1

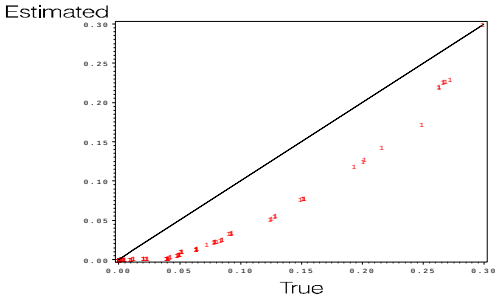



Figure 4f. Estimates vs Truth, File C
Cumulative Distribution of Nonmatches
1st Dependent EM, non−1−1

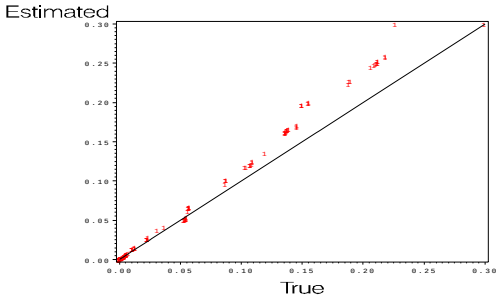


Figure 5a. Estimates vs Truth, File A
Cumulative Distribution of Matches
Independent EM, non−1−1

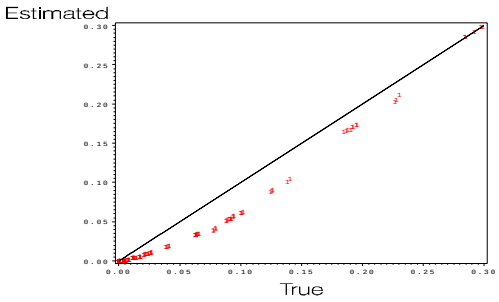



Figure 5b. Estimates vs Truth, File A
Cumulative Distribution of Nonmatches
Independent EM, non−1−1

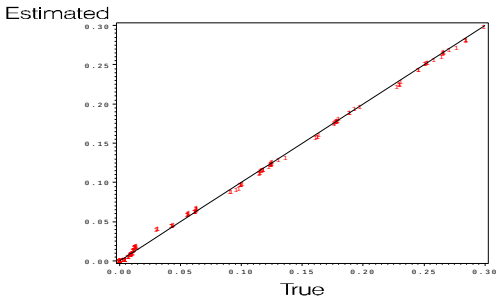



Figure 5c. Estimates vs Truth, File B
Cumulative Distribution of Matches
Independent EM, non−1−1

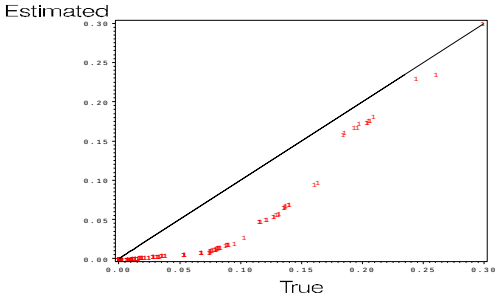



Figure 5d. Estimates vs Truth, File B
Cumulative Distribution of Nonmatches
Independent EM, non−1−1

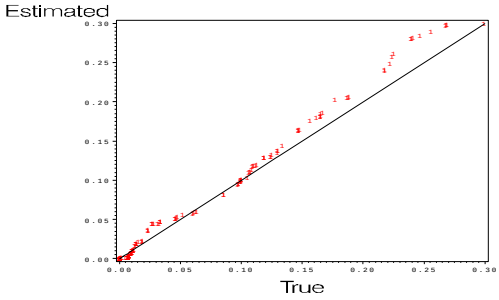



Figure 5e. Estimates vs Truth, File C
Cumulative Distribution of Matches
Independent EM, non−1−1

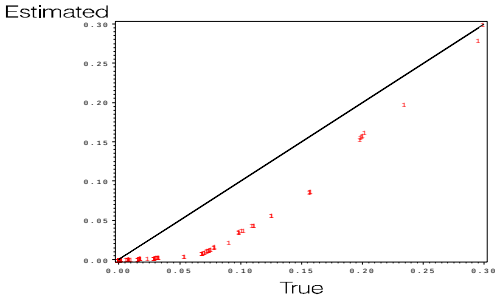



Figure 5f. Estimates vs Truth, File C
Cumulative Distribution of Nonmatches
Independent EM, non−1−1

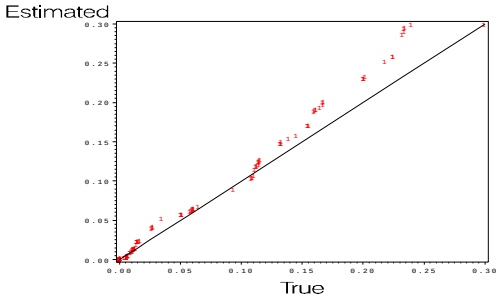


Figure 6a. Estimates vs Truth, File A
Cumulative Distribution of Matches
Large base, 1st Dependent EM, non−1−1

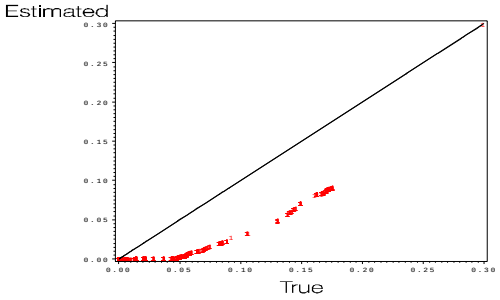



Figure 6b. Estimates vs Truth, File A
Cumulative Distribution of Nonmatches
Large base, 1st Dependent EM, non−1−1

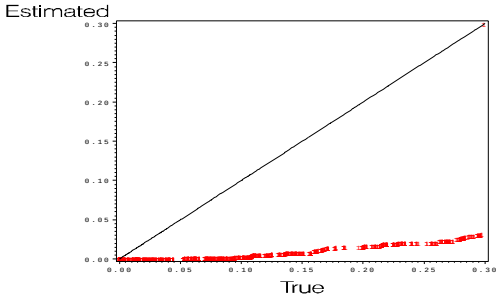



Figure 6c. Estimates vs Truth, File B
Cumulative Distribution of Matches
Large base, 1st Dependent EM, non−1−1

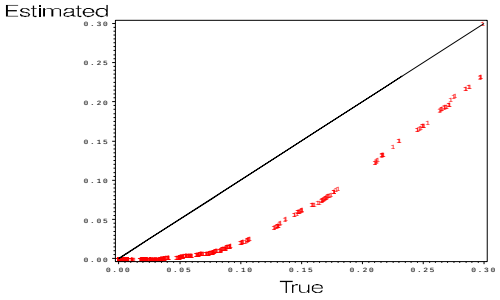



Figure 6d. Estimates vs Truth, File B
Cumulative Distribution of Nonmatches
Large base, 1st Dependent EM, non−1−1

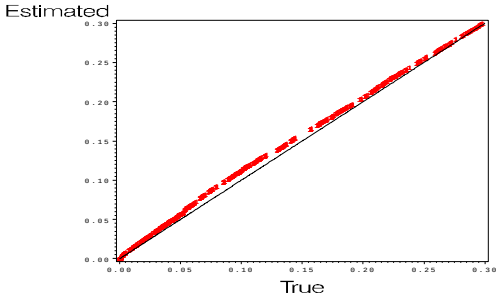



Figure 6e. Estimates vs Truth, File C
Cumulative Distribution of Matches
Large base, 1st Dependent EM, non−1−1

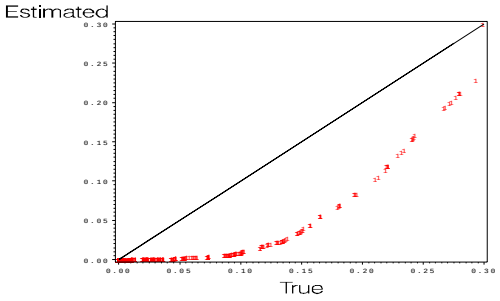



Figure 6f. Estimates vs Truth, File C
Cumulative Distribution of Nonmatches
Large base, 1st Dependent EM, non−1−1

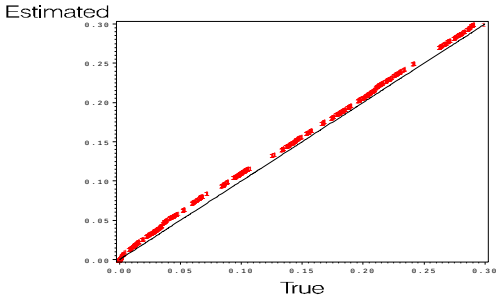


Figure 7a. Estimates vs Truth, File A
Cumulative Distribution of Matches
Large base, Independent EM, non−1−1

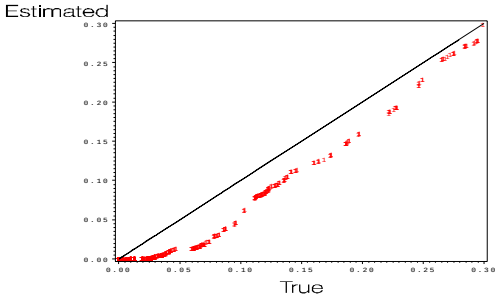



Figure 7b. Estimates vs Truth, File A
Cumulative Distribution of Nonmatches
Large base, Independent EM, non−1−1

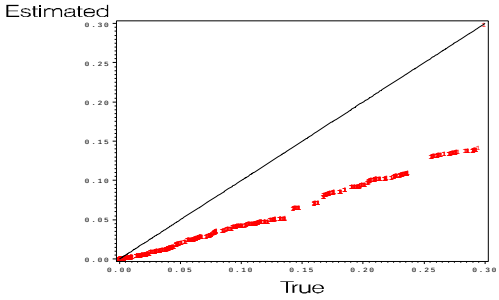



Figure 7c. Estimates vs Truth, File B
Cumulative Distribution of Matches
Large base, Independent EM, non−1−1

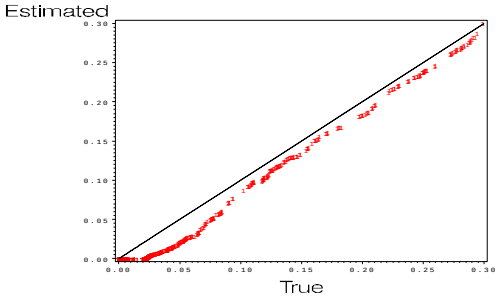



Figure 7d. Estimates vs Truth, File B
Cumulative Distribution of Nonmatches
Large base, Independent EM, non−1−1

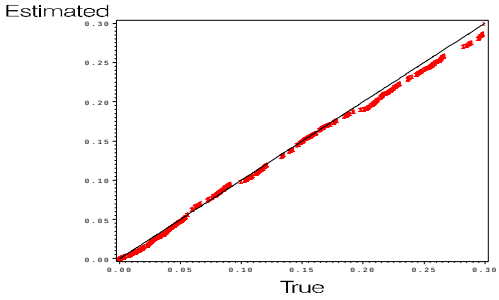



Figure 7e. Estimates vs Truth, File C
Cumulative Distribution of Matches
Large base, Independent EM, non−1−1

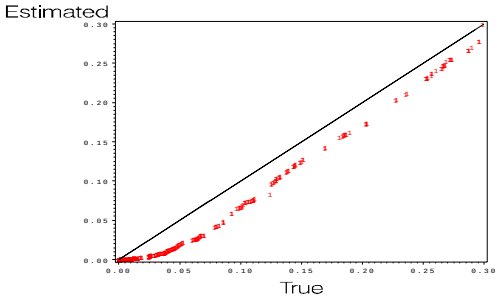



Figure 7f. Estimates vs Truth, File C
Cumulative Distribution of Nonmatches
Large base, Independent EM, non−1−1

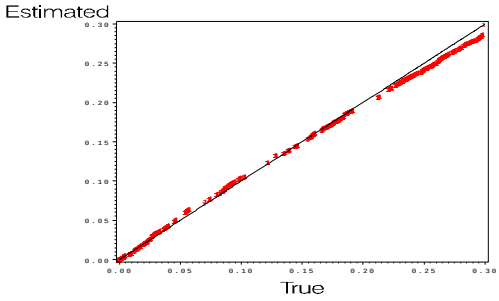


Figure 8a. Estimates vs Truth, File A
Cumulative Matches, Lambda = 0.9
Small Sample, Independent EM, non-1-1

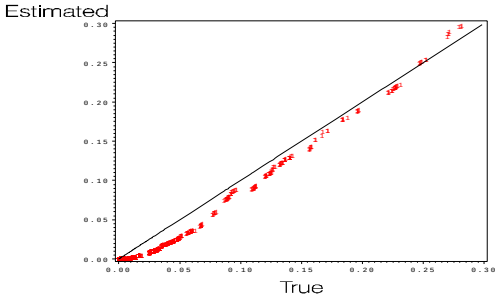



Figure 8b. Estimates vs Truth, File A
Cumulative Nonmatches, Lambda = 0.9
Small Sample, Independent EM, non-1-1

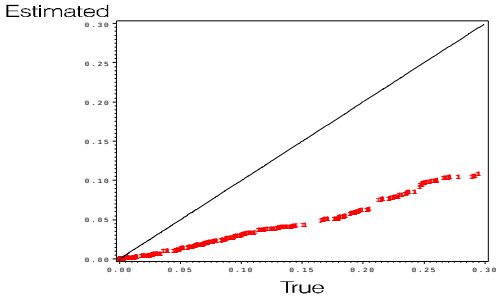



Figure 8c. Estimates vs Truth, File B
Cumulative Matches, Lambda = 0.9
Small Sample, Independent EM, non-1-1

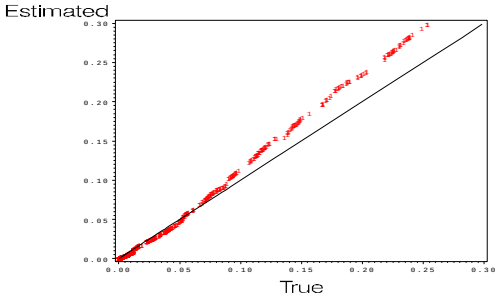



Figure 8d. Estimates vs Truth, File B
Cumulative Nonmatches, Lambda = 0.9
Small Sample, Independent EM, non-1-1

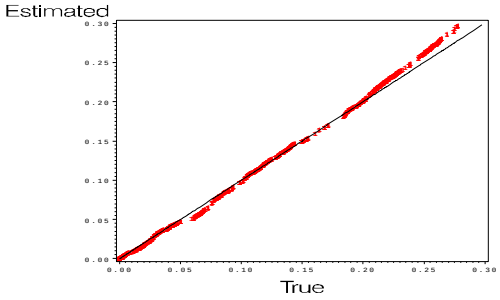



Figure 8e. Estimates vs Truth, File C
Cumulative Matches, Lambda = 0.9
Small Sample, Independent EM, non-1-1

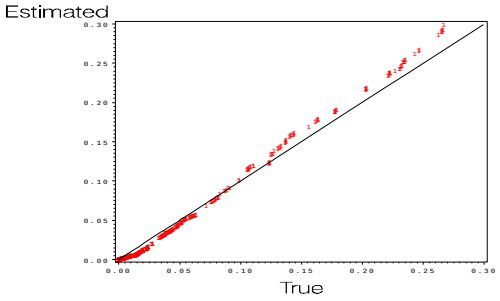



Figure 8f. Estimates vs Truth, File C
Cumulative Nonmatches, Lambda = 0.9
Small Sample, Independent EM, non-1-1

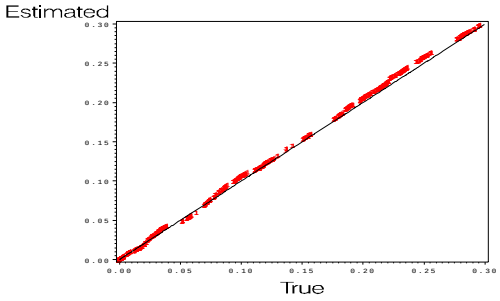


Figure 9a. Estimates vs Truth, File A
Cumulative Matches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

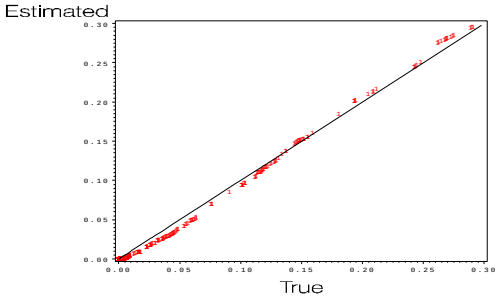



Figure 9b. Estimates vs Truth, File A
Cumulative Nonmatches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

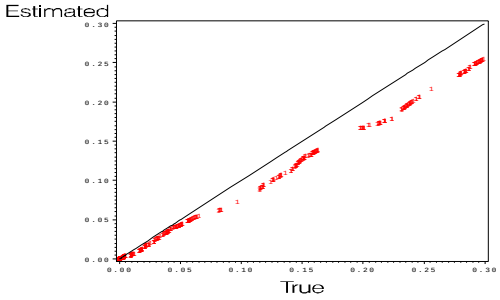



Figure 9c. Estimates vs Truth, File B
Cumulative Matches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

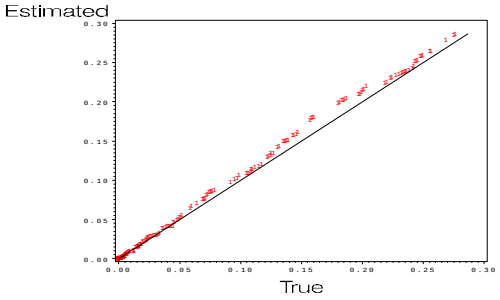



Figure 9d. Estimates vs Truth, File B
Cumulative Nonmatches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

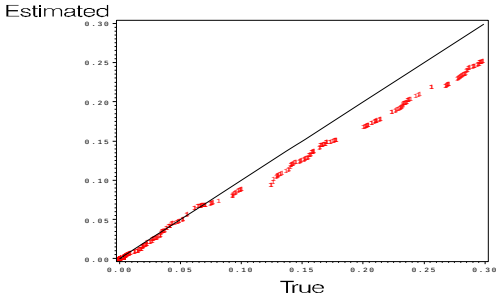



Figure 9e. Estimates vs Truth, File C
Cumulative Distribution of Matches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

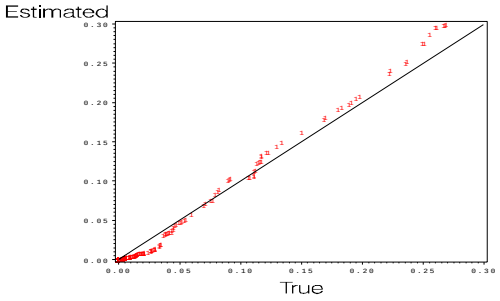



Figure 9f. Estimates vs Truth, File C
Cumulative Nonmatches, Lambda = 0.99
Small Sample, Independent EM, non-1-1

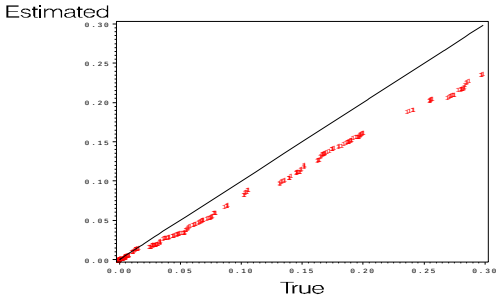


Figure 10a. Estimates vs Truth, File A
Cumulative Matches, Lambda = 0.9
Large Sample, Independent EM, non-1-1

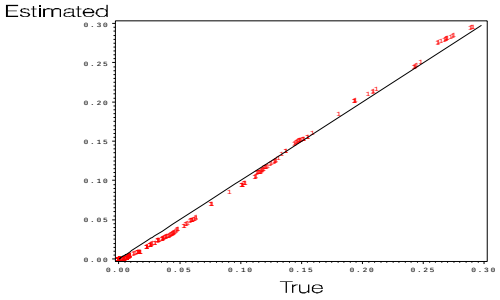



Figure 10b. Estimates vs Truth, File A
Cumulative Nonmatches, Lambda = 0.9
Large Sample, Independent EM, non-1-1

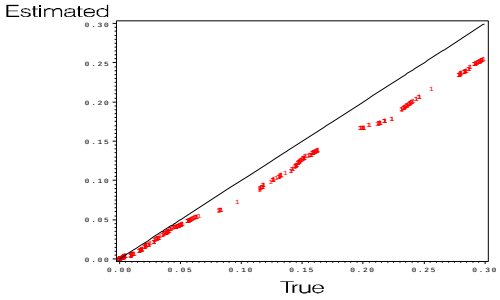



Figure 10c. Estimates vs Truth, File B
Cumulative Matches, Lambda = 0.9
Large Sample, Independent EM, non-1-1

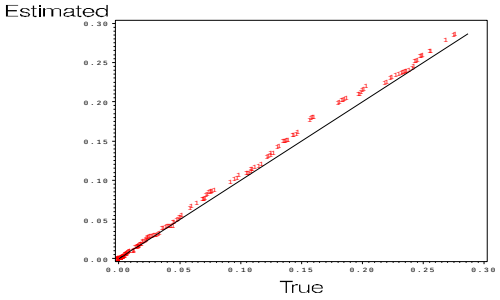



Figure 10d. Estimates vs Truth, File B
Cumulative Nonmatches, Lambda = 0.9
Large Sample, Independent EM, non-1-1

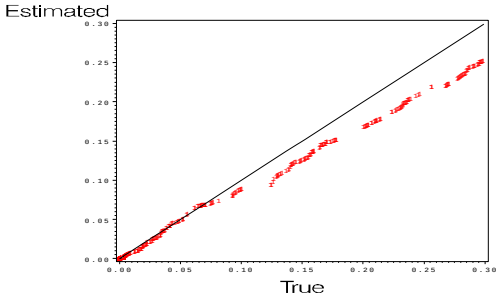



Figure 10e. Estimates vs Truth, File C
Cumulative Matches, Lambda = 0.9
Large Sample, Independent EM, non-1-1

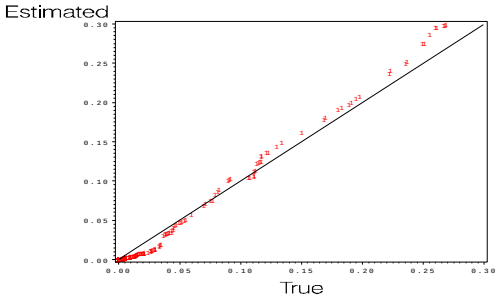



Figure 10f. Estimates vs Truth, File C
Cumulative Nonmatches, Lambda = 0.9
Large Sample, Independent EM, non-1-1

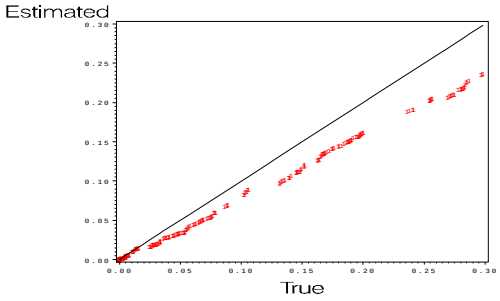


Figure 11a. Estimates vs Truth, File A
Cumulative Matches, Lambda = 0.999
Large Sample, Independent EM, non-1-1

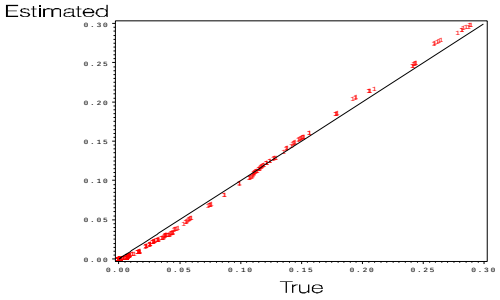



Figure 11b. Estimates vs Truth, File A
Cumulative Nonmatches, Lambda = 0.999
Large Sample, Independent EM, non-1-1

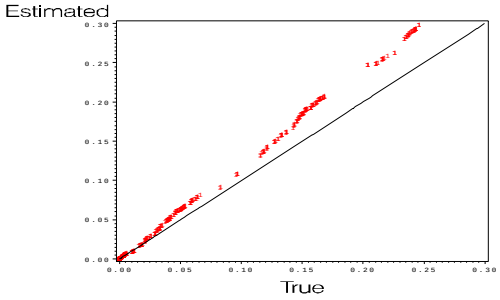



Figure 11c. Estimates vs Truth, File B
Cumulative Matches, Lambda = 0.999
Large Sample, Independent EM, non-1-1

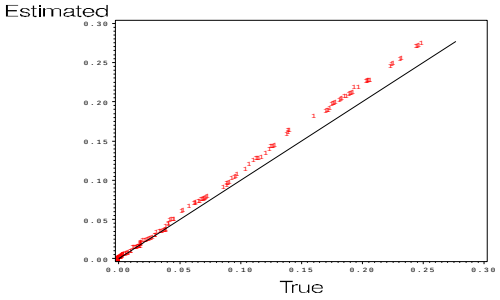



Figure 11d. Estimates vs Truth, File B
Cumulative Nonmatches, Lambda = 0.999
Large Sample, Independent EM, non-1-1

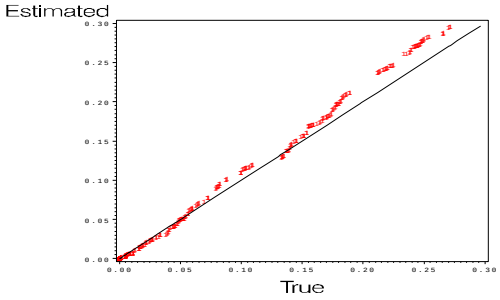



Figure 11e. Estimates vs Truth, File C
Cumulative Matches, Lambda = 0.999
Large Sample, Independent EM, non-1-1

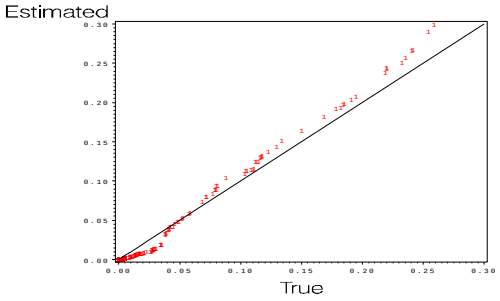



Figure 11f. Estimates vs Truth, File C
Cumulative Nonmatches, Lambda = 0.999
Large Sample, Independent EM, non-1-1