1. INTRODUCTION

As a special case of their general theory of record linkage, Fellegi and Sunter (1969) presented a formal model for matching that uses the relative frequency of strings being compared. For instance, a surname that is relatively rare in pairs of records taken from two files has more distinguishing power than a common one. Most applications of frequency-based matching have used close variants of the basic model but have made different simplifying assumptions that reduce computation and facilitate table building.

This paper introduces an extended methodology under weaker assumptions. While the amount of computation is significantly increased (as much as an order of magnitude), the need for expert human intervention is reduced. Most or all of the matching parameters can be automatically computed using file characteristics alone. The methodology does not require calibration data sets on which true match status has been determined. No a priori assumptions about parameters or previously created lookup tables are needed.

Relative frequency tables are more suitable for situations when one list cannot be assumed a near subset of another. When one list is a near subset of the other and a number of other simplifying assumptions are made, the new method yields tables comparable to those obtained via previous methods. If the matching is performed on a subset of pairs (such as those agreeing on Soundex code of surname or on specific geographic identifiers), then adjustments of the parameters and decision rules to the subsets are also automatic.

In the second section of the paper, background on the Fellegi-Sunter model of record linkage is presented. The third section is divided into five parts. The first contains the basic theory for the new frequency-based methods. The theory holds for all pairs in the product space of two files. In the second, a method of adjusting for typographical variation is given. The method partially accounts for the fact that observed frequencies do not necessarily correspond to true frequencies.

The third part shows how matching decision rules can utilize both frequency-based weights and simpler agree/disagree weights obtained via the Expectation-Maximization (EM) Algorithm (Winkler 1988, 1989a,c; Thibaudeau 1989). As the EM-derived weights are sometimes obtained on subsets of pairs such as those agreeing on geographical subregions, two methods for adjusting the frequency-based weights to subsets are given.

The fourth part contains empirical results for a comparison of files having substantial amounts of accurate information. In the fifth part, a comparison of files having greater amounts of missing data and/or typographical variation is presented.

The fourth section contains a five part discussion. In the first, the relationship of the method of this paper to the method of Fellegi and Sunter (1969) is discussed. The second part gives limitations of the adjustment for typographical variation. The third part presents characteristics of the subsets that affect the validity of the adjustment to subsets. In the fourth part, the relationship to other methods is described. The fifth part covers the limitations of the automatic estimation procedures.

The fifth section is a summary.

2. MODEL OF FELLEGI AND SUNTER

The Fellegi-Sunter Model uses a decision-theoretic approach establishing the validity of principles first used in practice by Newcombe (Newcombe, Kennedy, Axford, and James 1959). To give an overview, we describe the model in terms of ordered pairs in a product space. The description closely follows Fellegi and Sunter (1969, pp. 1184-1187).

There are two populations A and B whose elements will be denoted by a and b. We assume that some elements are common to A and B. Consequently the set of ordered pairs

\[ AXB = \{(a,b): a \in A, b \in B\} \]

is the union of two disjoint sets of matches

\[ M = \{(a,b): a = b, a \in A, b \in B\} \]

and nonmatches

\[ U = \{(a,b): a \neq b, a \in A, b \in B\} \]

The records corresponding to members of A and B are denoted by \( o(a) \) and \( r(b) \), respectively. The comparison vector \( \gamma \) associated with the records is defined by:

\[ \gamma = \{\gamma_1(a,b), \gamma_2(a,b), \ldots, \gamma_K(a,b)\} \]

Each of the \( \gamma_i \), i = 1, ..., K, represents a specific comparison. For instance, \( \gamma_1 \) could represent agreement/disagreement on sex. \( \gamma_2 \) could represent the comparison that two surnames agree and take a specific value or that they disagree.

Where confusion does not arise, the function \( g \) on \( AXB \) will be denoted by \( g(a,b) \). The conditional probability of \( \gamma \) given \( (a,b) \in M \) is given by

\[ m(\gamma) = P(\gamma(a,b) = 1, \gamma_2(a,b) = 1, \ldots, \gamma_K(a,b) = 1) \]

The conditional probability of \( \gamma(a,b) \) if \( (a,b) \in M \) is given by

\[ m(\gamma) = \sum_{(a,b) \in M} P(\gamma(a,b)) P((a,b) = M) \]
Similarly we denote the conditional probability of $\gamma$ if $(a,b)\in U$ by $u(\gamma)$.

We observe a vector of information $\chi(a,b)$ associated with pair $(a,b)$ and wish to designate a pair as a link (denote the decision by $A_1$), a possible link (decision $A_2$), or a nonlink (decision $A_3$). A linkage rule $L$ is defined a mapping from $\Gamma$, the comparison space, onto a set of random decision functions $D = \{d(\gamma)\}$ where

$$d(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma)\}; \forall \Gamma$$

and

$$\Sigma \ P(A_i|\gamma) = 1.$$ 

There are two types of error associated with a linkage rule. A Type I error occurs if an unmatched comparison is erroneously linked. It has probability

$$P(A_3|U) = \sum \ u(\gamma) \ P(A_3|\gamma)$$

A Type II error occurs if a matched comparison is erroneously not linked. It has probability

$$P(A_1|U) = \sum \ m(\gamma) \ P(A_1|\gamma)$$

Fellegi and Sunter (1969) define a linkage rule $L_0$ with associated decisions $A_1$, $A_2$, and $A_3$, that is optimal in the following sense:

**Theorem** (Fellegi-Sunter 1969). Let $L'$ be a linkage rule with associated decisions $A_1'$, $A_2'$, and $A_3'$ such that it has the same error probabilities $P(A_1'|M) = P(A_1|M)$ and $P(A_3'|U) = P(A_3|U)$ as $L_0$. Then $L_0$ is optimal in that $P(A_2'|U) \leq P(A_2|U)$ and $P(A_2'|M) \leq P(A_2|M)$.

In other words, if $L'$ is any competitor of $L_0$ having the same Type I and Type II error rates (which are both conditional probabilities), then the conditional probabilities (either on set $U$ or $M$) of not making a decision under rule $L'$ are always greater than under $L_0$.

To describe rule $L_0$, we need the following likelihood ratio

$$R = R(\gamma(a,b)) = m(\gamma)/u(\gamma).$$

We observe that, if $\gamma$ represents a comparison of $K$ fields, then there are at least $2^K$ probabilities of form $m(\gamma)$. If $\gamma$ represents agreements of $K$ fields, we would expect this to occur more often for matches $M$ than for nonmatches $U$. The ratio $R$ would then be large. Alternatively, if $\gamma$ consists of disagreements, the ratio $R$ would be small.

If the numerator is positive and the denominator is zero in (2.1), we assign an arbitrary very large number to the ratio. The Fellegi-Sunter linkage rule $L_0$ takes the form:

- If $R > \text{UPPER}$, then denote $(a,b)$ as a link.
- If $\text{LOWER} \leq R \leq \text{UPPER}$, then denote $(a,b)$ as a possible link.
- If $R < \text{LOWER}$, then denote $(a,b)$ as a nonlink.

The cutoffs $\text{LOWER}$ and $\text{UPPER}$ are determined by the desired error rate bounds.

The Fellegi-Sunter linkage rule is actually optimal with respect to any set $Q$ of ordered pairs in $AXB$ if we define error probabilities $P_a$ and a linkage rule $L_a$ conditional on $Q$. Thus, it may be possible to define subsets of $AXB$ on which we make use of differing amounts and types of available information.

### 3. FREQUENCY-BASED MODEL

#### 3.1. Basic Frequency-Based Parameter Estimation

In this section, we also closely follow the terminology of Fellegi and Sunter (1960, section 3.3.1). Let the true frequencies of occurrence of a specified string in files $A$ and $B$, respectively, be

$$f_1, f_2, \ldots, f_m; \Sigma f_i = N_A$$

and

$$g_1, g_2, \ldots, g_m; \Sigma g_i = N_B.$$

Let the corresponding true frequencies in $A \cap B$ be

$$h_1, h_2, \ldots, h_m; \Sigma h_i = N_{AB}.$$

We note that $h_i \leq \min(f_i, g_i), j = 1, 2, \ldots, m$. For the empirical examples of sections 3.3 and 3.4, for $j = 1, 2, \ldots, m$, we will generally use

$$h_i = \min(f_i, g_i) \quad \text{if } f_i > 1 \text{ or } g_i > 1,$$

$$h_i = 2/3 \quad \text{otherwise.}$$

The latter part of the definition implicitly means that, if we observe only one pair agreeing on a specific string, the pair has 2/3 chance of being a match and 1/3 chance of being a nonmatch.

The following additional notation is needed:

- $e_a$ or $e_b$ the respective probabilities of a name being misreported in $A$ or $B$ (independent of a particular name);
- $e_{ab}$ or $e_{ba}$ the respective probabilities of a name not being reported in $A$ or $B$ (name independent);
- $e_t$ the probability that a name is differently (but correctly) reported in the two files.

Then we have the following representations:

$$P(\text{string agrees} \ & \text{jth string} \mid M) = h_i (1-e_a)(1-e_b)(1-e_t)(1-e_{ab})/N_A,$$

$$P(\text{string disagrees} \mid M) = [1-(1-e_a)(1-e_b)(1-e_t)](1-e_{ab}) = e_a + e_b + e_t;$$

$$P(\text{string missing on either file} \mid M) = 1 - (1-e_a)(1-e_b) = e_{ab} + e_{ba};$$

$$P(\text{string agrees} \ & \text{jth string} \mid U) = (f_i - h_i)(1-e_a)(1-e_b)(1-e_t)(1-e_{ab})/(N_A N_B - N_{AB});$$

$$P(\text{string agrees} \ & \text{jth string} \mid U) = (f_i - h_i)(1-e_a)(1-e_b)(1-e_t)(1-e_{ab})/(N_A N_B - N_{AB});$$
P(string disagrees | U) = \[1 - (1 - e_o)(1 - e_A)(1 - e_T) \sum_j \left( (f_j g_j h_j) / (N_j N_k - N_{jk}) \right) \left( e_A e_T \right) \] and

\[P(string missing on either file | U) = 1 \cdot (1 - e_A)(1 - e_T) = e_A + e_T.\]

If \(f_j > 1\) or \(g_j > 1\) define the weight for agreement on the \(j\)th specific string, \(j = 1, 2, \ldots, m\), by

\[wgt(j) = h_j(N_j N_k - N_{jk}) / (f_j g_j h_j)  \]

and

\[P(\text{string missing on either file | U}) = 1 \cdot (1 - e_A)(1 - e_T) = e_A + e_T.\]

The weight represents the probability of agreement over the entire product space.

We observe that \(e_o\) and \(e_{AB}\) can be estimated directly using file characteristics. To estimate \(e_o\), \(e_A\) and \(e_T\), we need to know \(A \cap B\). They cannot be estimated directly. In practice, guesses based on past experience are often used.

For a production matching system for the 1990 Decennial Census (Winkler 1988, 1989a, c), use of the EM-Algorithm (see e.g., Dempster, Laird and Rubin 1977) allows direct estimation of \(P(\text{string disagrees | M})\) and, thus, approximate estimation of the sum \(e_A + e_T + e_T\). For most matching (and for the methods of the next section), we only need to estimate \(P(\text{string disagrees | M})\).

### 3.2. Weight Adjustment for Typographical Variation

If, say, \(x\) percent of files \(A\) and \(B\) contain typographical errors that are uniformly distributed through the files, then, for \(i = 1, 2, \ldots, n\), the true frequencies \(f_i\) and \(g_i\) will not correspond the observed frequencies \(x^i\) and \(x^g\), and the (average) agreement weight will not be correct.

In particular,

\[P_o(\text{agree | M}) = \frac{\sum_i (x^i \cdot h_i) / (\sum_i x^i)}{\sum_j (x^j \cdot g_j \cdot h_j) / (N_x N_g)}\]

\[P_o(\text{agree | U}) = \frac{\sum_i (x^i \cdot g_i \cdot h_i) / (N_x N_g)}{\sum_j (x^i \cdot g_i \cdot h_i) / (N_x N_g)}\]

\[P_o(\text{agree | M}) = \frac{\sum_i (x^i \cdot g_i \cdot h_i) / (N_x N_g)}{\sum_j (x^j \cdot g_j \cdot h_j) / (N_x N_g)}\]

where \(P_o\) and \(P_U\) depend on observed and true frequencies, respectively.

The adjustment factor (for typographical variation) is the ratio \(\beta\) of the number of pairs agreeing on the string to the number of pairs agreeing on Soundex code of the string. The adjusted weights are obtained by multiplying the existing weights by \(\beta\). The ratio adjusts for the fact that the observed frequencies of strings such as 'Smith' are less than the true frequencies because of typographical variation. The adjustment is always less than one. The adjustment is assumed independent of the subset over which the weights are computed or applied.

To use the adjustment, we must assume that Soundex encoding brings most pairs having typographical variation together. Under the assumptions that typographical errors \(e_o, e_A, \) and \(e_T\) are uniformly distributed on \(M\) and \(U\) and for the original string and the Soundex of the string, \(\beta\) effectively is the ratio

\[(P_o(\text{agree Soundex | M}) / P_o(\text{agree Soundex | U})) / (P_o(\text{agree | M}) / P_o(\text{agree | U})).\]

### 3.3. Combining Frequency-Based and EM-Derived Parameters on Sets of Pairs Obtained by Logical Blocking Criteria

Typically, matching parameters (or weights) and the associated decision rules are applied on small subsets of the entire product space \(AXB\). The subset might consist of pairs agreeing on a character-by-character basis for a specified string. Such a string might be the Soundex code of the surname. Soundex coding can sometimes account for very minor spelling variations. The subset might also consist of those pairs agreeing on a geographical subregion such as a set of Census blocks.

On the subset, we compute simple agree/disagree matching probabilities for each comparison field using the EM Algorithm (Winkler 1988, 1989a, c). The probabilities and resultant matching weights give the relative distinguishing power of various fields with respect to each other. For instance, agreement on first name might have a much larger positive weight than agreement on marital status. Disagreement on first name might have a lesser negative weight than disagreement on marital status.

Adjustments to subsets assure that the frequency-based weights do not overwhelm other weights that are computed over the subset. For instance, the scale, or range, of weights associated with the frequencies of surnames might be too great. Then the designation of a pair as a match or possible match could depend almost solely on the weight associated with surname.

Two different types of adjustments to subsets are needed. The first assures that the average (in a sense to be made clear) frequency-based weight agrees with the agreement weight computed via the EM Algorithm. The second assures that the average frequency-based weight associated with a field for which a weight is not presently available has the proper scale.

In the second case, such a weight might be associated with a field used as logical blocking characteristic used in creating a subset of pairs. The EM Algorithm cannot be used to estimate a simple agree/disagree weight on the subset because all pairs agree on the characteristic.

Let \(m_o\) and \(u_o\) be the respective estimated probabilities of agreement on a string given a match and given a nonmatch relative to a subset \(Q\). The probabilities could be obtained via the EM Algorithm or some other method.

Let

\[\alpha_o = \frac{m_o N_{ab}}{(\sum_i a_i^1)}\]  \hspace{1cm} (3.2a)

\[\alpha_o = \frac{u_o (N_x N_y - N_{ab})}{(\sum_i b_i^1)}\]  \hspace{1cm} (3.2b)

where \(a_i^1 = h_j^1\) if \(f_j > 1\) or \(g_j > 1\) and
\[ a_j = \begin{cases} \frac{2}{3} & \text{otherwise}, \\ g_j - h_j & \text{if } f_j > 1 \text{ or } g_j > 1 \\ \frac{1}{3} & \text{otherwise, and} \end{cases} \]

\[ b_i = \begin{cases} 1 & \text{if the } j \text{th string occurs in } Q \text{ and} \\ 0 & \text{otherwise.} \end{cases} \]

The assumption that \( a_j = \frac{2}{3} \) and \( b_j = \frac{1}{3} \) means that, if we observe only one pair agreeing on a specific string, the pair has \( \frac{2}{3} \) chance of being a match and \( \frac{1}{3} \) chance of being a nonmatch.

We approximate \( P(\text{string agrees } \& \text{ jth string } | M \cap Q) \) by

\[ \frac{\alpha_k h_j}{N_{ab}} \text{ if } f_j > 1 \text{ or } g_j > 1 \]

and

\[ \frac{\alpha_k (2/3)}{N_{ab}} \text{ otherwise}. \]

We approximate \( P(\text{string agrees } \& \text{ jth string } | U \cap Q) \) by

\[ \frac{\alpha_k (g_j - h_j)(N_kN_b - N_{ab})}{(N_kN_b - N_{ab})} \text{ if } f_j > 1 \text{ or } g_j > 1 \]

and

\[ \frac{\alpha_k (1/3)}{N_{ab}} \text{ otherwise}. \]

The adjusted weights associated with the jth specific value of the string are given by

\[ wgt_i(j) = \frac{\alpha_k wgt(i) / \alpha_k}{wgt(j)} \text{ for all } i \text{ and } j, \]

where \( wgt(i) \) is given by (3.1a,b) and \( wgt(j) \) is the quotient of (3.3a) and (3.3b). We observe that

\[ \Sigma P(\text{string agrees } \& \text{ jth string } | M \cap Q) = m' \] (3.4a)

\[ \Sigma P(\text{string agrees } \& \text{ jth string } | U \cap Q) = u' \] (3.4b)

\[ wgt_i(j)/wgt_i(j') = wgt(i)/wgt(j) \text{ for all } i \text{ and } j. \] (3.4c)

Note that \( m'/u' \) is the simple agree/disagree weight. Equation (3.4c) yields the fact that the adjusted frequency-based weights have the same relative distinguishing power with respect to specific strings as the unadjusted frequency-based weights. The sums in (3.4) are over those specific strings that occur in \( Q \).

For strings for which we do not have estimates of the simple agree and disagree probabilities \( m_i \) and \( u_i \), the adjusted weights \( wgt_i(j) \), \( i = 1, 2, \ldots, m \), depend on the number of pairs \( N_k \) in the subset \( Q \) and the number of pairs \( N_{10} \) in \( Q \) that are matches. We estimate, except for typographical variation,

\[ m_i = 1 \] and

\[ u_i = (M_0 - N_{10})/N_0 \]

where \( M_0 \) is the number of pairs in \( Q \) that agree on the string. \( N_{10} \) is estimated like \( N_{0r} \). It is the minimum of the number of strings in \( Q \) that agree on the specific values if the strings occur more than once; \( 2/3 \), otherwise. As with (3.1), (3.2), and (3.3), we implicitly assume that

\[ m_0 = P(\text{agree on string } | M) = 1, \]

except for typographical variation due to errors. (3.5a) is only suitable for use with a string such as last name (or possibly, first name) that we expect virtually all matches to agree on. We also assume that typographical variation effects \( m_i \) and \( u_i \) identically so that they cancel in \( m_i/u_i \).

3.4. Application Using Files with Good Distinguishing Information

The results in this section are from a computer matching application with two files of Los Angeles data. Each file contains 20,000 records. The observed counts for 1024 agree/disagree patterns for ten variables in a set of pairs are used in obtaining most matching parameter estimates. The ten variables are first name, middle initial, house number, street name, unit (apartment #), age, sex, relation, marital status, and race. Frequency-based weights are created for surname and first name.

The set of pairs consists of 249,000 agreeing on geocodes and first character of surnames. The geocode is the Census block number.

As there can be at most 20,000 matches, it is not computationally practicable to consider counts based on all 400 million pairs in the product space. Based on prior experience, it is known that more than 70 percent of the matches will be in the set of 249,000 pairs.

We arbitrarily use 0.00001 as the estimate for the sum \( e_A + e_s + e_r \) associated with the frequency-based weights for both the last name and the first name. Direct estimation of the sum for first name using the EM-Algorithm yields approximately 0.01.

The downward adjustment increases the distinguishing power of the frequency-based weights. The increase occurs because the total weight of pairs for disagreeing on either last name or first name and agreeing randomly on demographic characteristics such as age, marital status, and relationship or address character is decreased.

The typographical adjustment for first name is 0.42 and for last name is 0.63. The adjustment to the set of pairs \( Q \) is 0.30 for first name and 0.16 for last name. The values of 0.30 and 0.16 indicates that the distinguishing powers (i.e., range of weights) of first name and last name are greater on the whole space than on \( Q \). The overall adjustment to the frequency-based weights for first name and last name are 0.12 (=0.42.0.30) and 0.10 (=0.63.0.16), respectively.

For each type of matching, the high cutoff \( \text{UPPER} \) is chosen so that less than one percent of the matches is false. The low cutoff \( \text{LOWER} \) is chosen so that few, if any, matches having total weight less than the cutoff exist. Both determinations are subjective because true match statuses are unknown.

There are several reasons why frequency-based matching performs better than basic matching that uses only agree/disagree weights (Table 1). First, the number of designated matches increases from 12,455 to 13,136...
when compared to a basic matcher. The increase is basically due to pairs having rare surnames and rare first names. Such pairs, if they have a moderate number of disagreements on other characteristics, are designated as possible matches by the basic matcher.

Second, a set of 350 pairs that a nonmatches with the basic matcher are designated possible matches by the agreeing on rare surnames but agreeing on few other characteristics.

3.5. Application Using Files with Poor Distinguishing Information

The results in this section are from a computer matching application with two files of St Louis data. The larger file contains 13,719 records while the smaller 2,777. The smaller data file was obtained from various administrative data sources.

The observed counts for 128 agree/disagree patterns for seven variables in a set of pairs were used for most parameter estimates. The seven variables were first name, middle initial, address, age, sex, telephone, and race. Frequency-based weights were created for surname and first name. The set of pairs consists of 43,577 agreeing on Soundex code of surname.

We arbitrarily use 0.0001 as the estimate for the sum $e_a + e_b + e_t$ associated with the frequency-based weights for both the last name and the first name. Direct estimation of the sum for first name using the EM-Algorithm yields approximately 0.01. The downward adjustment increases the distinguishing power of the frequency-based weights.

Arbitrarily chosen adjustments of 0.25 and 0.0625 for typographical variation in first name and last name were used. As matching was for all pairs, no adjustment for a subset $Q$ was used.

Both basic and frequency-based matching do not perform well (Table 2). Each match only slightly more than 300 of the 2800 records. The file, which is used for obtaining additional information about black males between ages 18 and 44, contains much missing data. For middle initial, telephone, and race, there are 1201, 2153, and 1091 missing data items, respectively. Age and address are also typically inaccurate.

Frequency-based matching designates more pairs as possible matches (269 versus 157). The increased number represents those pairs agreeing on both a relatively rare first and a relatively rare last name while most other characteristics are either missing or disagree.

4. DISCUSSION

4.1. Finite Population Correction

The chief difference between the frequency-based methods of this paper (section 3.1) and those of Fellegi and Sunter (1969) are a type of finite population correction.

In two files $A$ and $B$, for some $j$, we observe $f_j$ and $g_j$ occurrences of the $j$th string. In this paper, if $f_j > 1$ or $g_j > 1$, we take $h_j = \min(f_j, g_j)$ as the number of matches associated with the string and $f_j - f_j \cdot h_j$ as the number of nonmatches. The main reason that the adjustments of this paper do not assure that the sums of the numerators and denominators in the weights equal estimated $m_0$ and $u_0$ as we do.

The main reason that the adjustments of this paper and of Rogot, Sorlie, and Johnson (1986) are taken to the whole space, the finite population correction is ignored in our computation, and the adjustment for deaths of Rogot, Sorlie, and Johnson is ignored (i.e., taken to be uniformly one), then the two approaches coincide.

On the subsets, the methods of getting relative scales according to values of the strings are different. Also, Rogot, Sorlie, and Johnson do not assure that the sums of the numerators and denominators in the weights equal estimated $m_0$ and $u_0$ as we do.

The chief difference between the frequency-based methods of this paper (section 3.1) and those of Fellegi and Sunter (1969) are a type of finite population correction.

In two files $A$ and $B$, for some $j$, we observe $f_j$ and $g_j$ occurrences of the $j$th string. In this paper, if $f_j > 1$ or $g_j > 1$, we take $h_j = \min(f_j, g_j)$ as the number of matches associated with the string and $f_j - f_j \cdot h_j$ as the number of nonmatches. The main reason that the adjustments of this paper do not assure that the sums of the numerators and denominators in the weights equal estimated $m_0$ and $u_0$ as we do.

The main reason that the adjustments of this paper and of Rogot, Sorlie, and Johnson (1986) are taken to the whole space, the finite population correction is ignored in our computation, and the adjustment for deaths of Rogot, Sorlie, and Johnson is ignored (i.e., taken to be uniformly one), then the two approaches coincide.

On the subsets, the methods of getting relative scales according to values of the strings are different. Also, Rogot, Sorlie, and Johnson do not assure that the sums of the numerators and denominators in the weights equal estimated $m_0$ and $u_0$ as we do.
\[ wg_{0}(j) = \left( m_{0} \Sigma \left\{ g_{0}/N_{0} \cdot g_{0}/N_{0} \right\} \right) / \left( u_{0} \cdot g_{0}/N_{0} \right) = \]

\[ m_{0} \cdot \text{freq}(j) / u_{0}, \]

where \( m_{0} \) and \( u_{0} \) are the general agreement probabilities on \( AXB \). The factor \( \text{freq}(j) \) that adjusts the general agreement weight to the specific weight agrees with the standard adjustment factor (see e.g., Newcombe, Fair, and Lalonde 1987 pp. 134-135; 1989 pp. 88-89).

4.5. Limitations

Presently, the only apparent limitations are with the EM-Algorithm-based procedures used to compute the \( m \)-and \( u \)-probabilities representing simple agree/disagree comparisons (Winkler 1989a, section 4.3). For smaller files having less than 2,000 records and exhibiting large amounts of typographical variation, the \( m \)-probabilities associated with key matching fields such as first name have occasionally had to be adjusted upward to improve matching decision rules.

If adjusting parameters improves the rules, then the original unadjusted parameter estimates do not accurately represent the true distributions (e.g., Winkler 1989b, sections 2.3.4, 4.2).

### 4. SUMMARY

The frequency-based matching parameter estimation methodology of this paper extends the methodology of Fellegi and Sunter (1969). If strong simplifying assumptions are made, then the methodology of this paper approximately agrees with methods currently in use. Methods of accounting for certain types of typographical variation and for estimating parameters on subsets of the product space are introduced. No calibration data sets having known true match status are needed.

### REFERENCES


#### Table 1

<table>
<thead>
<tr>
<th></th>
<th>Frequency-Based</th>
<th>Los Angeles Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>12320</td>
<td>128</td>
</tr>
<tr>
<td>Possible</td>
<td>8081</td>
<td>2146</td>
</tr>
<tr>
<td>Nonmatch</td>
<td>8</td>
<td>386</td>
</tr>
<tr>
<td>Total</td>
<td>13136</td>
<td>2660</td>
</tr>
</tbody>
</table>

#### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Frequency-Based</th>
<th>St Louis Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>305</td>
<td>15</td>
</tr>
<tr>
<td>Possible</td>
<td>15</td>
<td>106</td>
</tr>
<tr>
<td>Nonmatch</td>
<td>2</td>
<td>106</td>
</tr>
<tr>
<td>Total</td>
<td>322</td>
<td>269</td>
</tr>
</tbody>
</table>

* This paper reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.