BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR2000/06

Frequency-Based Matching in Fellegi-Sunter Model of Record Linkage

William E. Winkler
Statistical Research Division
Methodology and Standards Directorate
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: October 4, 2000

FREQUENCY-BASED MATCHING IN FELLEGI-SUNTER MODEL OF RECORD LINKAGE

William E. Winkler, william.e.winkler@census.gov
Bureau of the Census

ABSTRACT

This paper extends techniques for frequency-based matching (see e.g., Fellegi and Sunter 1969). The extended techniques allow table-building under weaker assumptions than those typically used in practice. Although CPU requirements can increase, human intervention can be reduced in some situations.

Keywords: Decision rule, error rate.

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.

The outline of this paper is as follows. 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, 1989c; 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 effect 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

$$\mathbf{A}\times\mathbf{B} = \{(a,b): a\varepsilon\mathbf{A}, b\varepsilon\mathbf{B}\}$$

is the union of two disjoint sets of *matches*

$$M = \{(a,b): a=b, a\varepsilon\mathbf{A}, b\varepsilon\mathbf{B}\}$$

and *nonmatches*

$$U = \{(a,b): a\neq b, a\varepsilon\mathbf{A}, b\varepsilon\mathbf{B}\}.$$

The records corresponding to members of **A** and **B** are denoted by $\alpha(a)$ and $ß(b)$, respectively. The *comparison vector* γ associated with the records is defined by:

$$\gamma[\alpha(a),ß(b)] \equiv \{\gamma^1[\alpha(a),ß(b)],\gamma^2[\alpha(a),ß(b)],\cdots,\gamma^K[\alpha(a),ß(b)]\}.$$

Each of the γ^i, $i = 1, \cdots, K$, represents a specific comparison. For instance, γ^1 could represent agreement/disagreement on sex. γ^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 γ on **AXB** will be denoted by $\gamma(\alpha,ß)$, $\gamma(a,b)$, or γ. The set of all possible realizations of γ is denoted by Γ. The conditional probability of $\gamma(a,b)$ if $(a,b)\varepsilon M$ is given by

$$m(\gamma) \equiv P\{\gamma[\alpha(a),ß(b)]|(a,b)\varepsilon M\}$$

$$= \sum_{(a,b)\varepsilon M} P\{\gamma[\alpha(a),ß(b)]\}\cdot P[(a,b)|M].$$

Similarly we denote the conditional probability of γ if $(a,b)\varepsilon U$ by $u(\gamma)$. We observe a vector of information $\gamma(a,b)$ 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 Γ, the comparison space, onto a set of random decision functions $D = \{d(\gamma)\}$ where

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

and

$$\Sigma_{i=1}^{3} P(A_i \mid \gamma) = 1.$$

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

$$P(A_1 \mid U) = \sum_{\gamma \varepsilon \Gamma} u(\gamma) \cdot P(A_1 \mid \gamma)$$

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

$$P(A_3 \mid U) = \sum_{\gamma \varepsilon \Gamma} m(\gamma) \cdot P(A_3 \mid \gamma)$$

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 $P(A_3'|M) = P(A_3|M)$ and $P(A_1'|U) = P(A_1|U)$ 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

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

We observe that, if γ represents a comparison of K fields, then there are at least 2^K probabilities of form $m(\gamma)$. If γ 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 γ 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 $R > T_\mu$, then denote (a,b) as a link.

If $T_\lambda \le R \le T_\mu$, then denote (a,b) as a possible link. (2.2)

If $R < T_\lambda$, then denote (a,b) as a nonlink.

The cutoffs T_λ and T_μ are determined by the desired error rate bounds μ and λ on the false match rates and false nonmatch rates, respectively. The Fellegi-Sunter linkage rule is actually optimal with respect to any set Q of ordered pairs in $\mathbf{A}\times\mathbf{B}$ if we define error probabilities P_Q and a linkage rule L_Q conditional on Q. Thus, it may be possible to define subsets of $\mathbf{A}\times\mathbf{B}$ 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

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

and

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

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

$$h_1, h_2, \cdots, h_m ; \quad \Sigma_{j=1}^{m} h_j = N_{AB}.$$

We note that $h_j \le \min(f_j,g_j)$, $j = 1, 2, \cdots, m$. For the empirical examples of sections 3.3 and 3.4, for $j = 1, 2, \cdots, m$, we will generally use

$h_j = \min(f_j,g_j)$ if $f_j>1$ or $g_j>1$

$h_j = 2/3$ 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** (assumed independent of a particular name);

e_{A0} or e_{B0} 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(string agrees & jth string| M)
$= h_j (1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0})/N_{AB}$
$\approx h_j (1-e_A-e_B-e_T-e_{A0}-e_{B0})/N_{AB}$;

P(string disagrees | M)
$= [1-(1-e_A)(1-e_B)(1-e_T)](1-e_{A0})(1-e_{B0}) \approx e_A + e_B + e_T$;

P(string missing on either file | M)
$= 1 - (1-e_{A0})(1-e_{B0}) \approx e_{A0} + e_{B0}$;

P(string agrees & jth string| U)
$= (f_j \cdot g_j - h_j)(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0})/(N_A \cdot N_B - N_{AB})$
$\approx (f_j \cdot g_j - h_j)(1-e_A-e_B-e_T-e_{A0}-e_{B0})/(N_A \cdot N_B - N_{AB})$;

P(string disagrees | U)
$= [1-(1-e_A)(1-e_B)(1-e_T) \Sigma_{j=1}^{m} (f_j \cdot g_j - h_j)/(N_A \cdot N_B - N_{AB})](1-e_{A0})(1-e_{B0})$
$\approx [1-(1-e_A-e_B-e_T)\Sigma_{j=1}^{m} (f_j \cdot g_j - h_j)/(N_A \cdot N_B - N_{AB})](1-e_{A0}-e_{B0})$; and

P(string missing on either file | U)
$= 1 - (1-e_{A0})(1-e_{B0}) \approx e_{A0} + e_{B0}$.

We define the weight for agreement on the jth specific string, j = 1, 2, ···, m, by

$$wgt(j) = h_j \cdot (N_A \cdot N_B - N_{AB}) / ((f_j \cdot g_j - h_j) \cdot N_{AB}), \quad (3.1a)$$

if either $f_j>1$ or $g_j>1$ and by

$$wgt(j) = 2 \cdot (N_A \cdot N_B - N_{AB})/N_{AB} \quad (3.1b)$$

if $f_j=1$ and $g_j=1$. The weight represents the probability of agreement over the entire product space. We observe that e_{A0} and e_{B0} can be estimated directly using file characteristics. To estimate e_T, e_A and e_B, we need to know $\mathbf{A} \cap \mathbf{B}$. 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, 1989c), use of the EM-Algorithm (see e.g., Dempster, Laird and Rubin 1977) allows direct estimation of P(string disagrees|M) and, thus, approximate estimation of the sum $e_A+e_B+e_T$. For most matching (and for the methods of the next section), we only need to estimate P(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 $i = 1, 2, \cdots, n$, the true frequencies f_i and g_i will not correspond the observed frequencies $x \cdot f_i$ and $x \cdot g_i$ and the (average) agreement weight will not be correct.

In particular,

$$\frac{P_O(\text{agree}|M)}{P_O(\text{agree}|U)} \equiv \frac{(\Sigma_i\, x \cdot h_i)/(\Sigma_i\, x \cdot h_i)}{(\Sigma_i\, (x \cdot f_i \cdot x \cdot g_i - x \cdot h_i)/N_A \cdot N_{B)}} \approx \frac{P_T(\text{agree}|M)}{x \cdot x \cdot P_T(\text{agree}|U)},$$

where P_O and P_T depend on observed and true frequencies, respectively.

The adjustment factor (for typographical variation) is the ratio ß 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 ß. 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_A, e_B, and e_T are uniformly distributed on M and U and for the original string and the Soundex of the string, ß 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

Typically, matching parameters (or weights) and the associated decision rules are applied on small subsets of the entire product space **A**×**B**. 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, 1989c). 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 can not be used to estimate a simple agree/disagree weight on the subset because all pairs agree on the characteristic.

Let m_0 and u_0 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

$$\alpha_1 = m_0 \cdot N_{AB} / (\Sigma_{j=1}^{m} a_j \cdot l_j) \quad \text{and} \tag{3.2a}$$

$$\alpha_2 = u_0 \cdot (N_A \cdot N_B - N_{AB}) / (\Sigma_{j=1}^{m} b_j \cdot l_j), \tag{3.2b}$$

where $a_j = h_j$ if $f_j>1$ or $g_j>1$ and

$aj = 2/3$ otherwise,

$b_j = f_j \cdot g_j - h_j$ if $f_j>1$ or $g_j>1$ and

$b_j = 1/3$ otherwise, and

$l_j = 1$ if the jth string occurs in Q and

$l_j = 0$ otherwise.

The assumption that $a_j=2/3$ and $b_j=1/3$ 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.

We approximate P(string agrees & jth string| M∩Q) by

$$\approx \alpha_1 \cdot h_j / N_{AB} \text{ if } f_j>1 \text{ or } g_j>1$$

and (3.3a)

$$\approx \alpha_1 \cdot (2/3)/N_{AB} \text{ otherwise.}$$

We approximate P(string agrees & jth string| U∩Q) by

$$\approx \alpha_2 \cdot (f_j \cdot g_j - h_j)/(N_A \cdot N_B - N_{AB}) \text{ if } f_j>1 \text{ or } g_j>1$$

and (3.3b)

$$\approx \alpha_2 \cdot (1/3)/(N_A \cdot N_B - N_{AB}) \text{ otherwise.}$$

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

$wgt_Q(j) = \alpha_1 \cdot wgt(j) / \alpha_2, \qquad j = 1, 2, \cdots, m,$

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

$$\Sigma_{j=1}^{m} P(\text{string agrees \& jth string} \mid M \cap Q) = m_0, \tag{3.4a}$$

$$\Sigma_{j=1}^{m} P(\text{string agrees \& jth string} \mid U \cap Q) = u_0, \text{ and} \tag{3.4b}$$

$$wgt_Q(i)/wgt_Q(j) = wgt(i)/wgt(j) \quad \text{for all } i \text{ and } j. \tag{3.4c}$$

Note that m_0/u_0 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_0 and u_0, the adjusted weights $wgt2_Q(i)$, $i = 1, 2, \cdots, m$, depend on the number of pairs N_Q in the subset Q and the number of pairs $N1_Q$ in Q that are matches. We estimate

$$m_0 \approx 1 \quad \text{except for typographical variation and} \tag{3.5a}$$

$$u_0 \approx (M_Q - N1_Q)/N_Q \quad \text{except for typographical variation,} \tag{3.5b}$$

where M_Q is the number of pairs in Q that agree on the string. $N1_Q$ is estimated like N_{AB}. 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

$$m_0 = P(\text{agree on string} \mid M) \approx 1,$$

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_0 and u_0 identically so that they cancel in m_0/u_0.

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 $e_A+e_B+e_T$ 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. 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 T_μ is chosen so that less than one percent of the matches is false. The low cutoff T_λ 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 net of 350 pairs that a nonmatches with the basic matcher are designated possible matches by the frequency-based matcher. The set consists of those pairs 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 43,377 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). The set of matches consists of 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 jth 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 $f_j \cdot g_j - h_j$ as the number of nonmatches. Fellegi and Sunter take the number of matches to be f_j (if **A** is the target file) and the number of nonmatches to be $f_j \cdot g_j$. If both f_j and g_j are 1, then we assume that the number of matches is 2/3 and the number of nonmatches is 1/3. Fellegi and Sunter assume both the number of matches and number of nonmatches are 1.

If both files **A** and **B** are assumed to be subsets of the same population which is basically the union of the two files and both **A** and **B** are approximately equal to the union, then could use $h_j=\max(f_j,g_j)$ as the estimated number of matches and $h_j \cdot h_j - h_j$ as the number of matches. This method of computation may be preferred if one file is known to contain strings subject to substantial typographical variation.

4.2. Subsets Used for Weight Computation

Rogot, Sorlie, and Johnson (1986) use a different approach for adjusting frequency-based weights to subsets. They use both observed frequencies on the whole space and on the subset. If the methods 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 main reason that the adjustments of this paper and of Rogot, Sorlie, and Johnson are useful is that, on the whole space, the relative distinguishing power of uncommon strings as compared with common strings can be preserved. For instance, on the whole space, assume 100 pairs have 'Smith' and one has 'Zabrinsky' while, on the subset, assume that 10 sets, (say, depending on a geocode) of 10 'Smith' occur while one 'Zabrinsky' still occurs. Then, on the subset, the perceived relative distinguishing power of the geocode-'Zabrinsky' relative to geocode-'Smith' is less. If agreement on a blocking characteristic, say geocode, could also be accounted for, then distinguishing power could conceivably increase. The subset adjustment method of this paper has a computational advantage. Lookup tables and adjustments based only on surnames are more easily computed than those based on both geocodes and surnames.

4.3. Adjustment for Typographical Variation

The adjustment ß for typographical variation is intended to account partially for the fact that we cannot observe the true frequencies (independent of typographical variation) in the files being used. For instance, if files **A** and **B** each contain 100 'Smiths' but each file only contains 90 'Smiths' because of typographical variation, then we would calculate the ratio $h_j/(f_j \cdot g_j - h_j)$ based on file characteristics to be 0.0112 ($\approx$90/(90·90-90)) rather than 0.0101 ($\approx$100/(100·100-100)). If perceived and true frequencies are 6 and 9, respectively, then the ratios are 0.125 and 0.2, respectively. Using the files with typographical errors yields weights that are perceived to have more distinguishing power than they actually have.

Because the adjustment ß is based on Soundex code of a string and Soundex is not able to account for most typographical variation, ß is likely to be too large. The adjustment ß performs similarly to the transmission weight adjustment (Howe and Lindsay 1981, Newcombe 1988).

The transmission weight is intended to be a rather precise adjustment for typographical variation that depends on knowledge of true match status. The main advantages of ß is that it is easily computed and does not depend on true match status.

4.4. Comparison with Other Frequency-Based Methods

If we assume that file **A** is a subset of file **B**, that the finite population correction can be ignored (i.e., $f_j \cdot g_j = f_j \cdot g_j - h_j$), that the subset Q is the whole product space **A**×**B**, then the adjusted weights take the form, j = 1, ···, m,

$$wgt_Q(j) = (m_0 \cdot \Sigma_{j=1}^{m} f_j \cdot g_j / N_A \cdot N_B)/(u_0 \cdot g_j / N_B) \equiv m_0 \cdot freq(j)/u_0,$$

where m_0 and u_0 are the general agreement probabilities on **A**×**B**. The factor 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). The ratio m_0/u_0 implicitly accounts for some of the typographical variation. Applying it uniformly to all frequency-based weights is consistent with the explicit assumptions of this paper and of Fellegi and Sunter (1969) and with the implicit assumption of Newcombe, Fair, and Lalonde (1987, 1989).

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).

5.**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.

Author's note: A version of this paper appeared under the same title in the American Statistical Association Proceedings of the Section on Survey Research Methods in 1989 on pages 778-783. The results have not been superseded.

REFERENCES

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) "Maximum Likelihood from Incomplete

Data via the EM Algorithm," *Journal of the Royal Statistical Society B*, **39**, 1-38.
Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **40**, 1183-1210.
Howe, G. R., and Lindsay, J. (1981), "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies," *Computers and Biomedical Research*, 14, 327-340.
Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.
Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.
Newcombe, H. B., Fair, M. E., Lalonde, P. (1987), "Concepts and Practices that Improve Probabilistic Record Linkage," in *Statistical Uses of Administrative Data*, Edited by Coombs, J. W. and Singh, M.P. 127-138.
Newcombe, H. B., Fair, M. E., Lalonde, P. (1989), "Discriminating Powers of Partial Agreements of Names for Linking Personal Records," *Methods of Information in Medicine,* **8**, 86-91.
Rogot, E., Sorlie, P., and Johnson, N. J. (1986), "Probabilistic Methods in Matching Census Samples to the National Death Index," *Journal of Chronic Disease*, **39**, 719-734.
Thibaudeau, Y. (1989), "Fitting Log-Linear Models When Some Dichotomous Variables are Unobservable," American Statistical Association, *Proceedings of the Section on Statistical Computing*, 283-288.
Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 667-671.
Winkler, W. E. (1989a), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, 145-155.
Winkler, W. E. (1989b), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, 101-117.
Winkler, W. E. (1989c), "Maximum Likelihood Estimates for Restrained Mixtures of Multinomial Distributions," *Proceedings of the 21st Symposium on the Interface: Computing Science and Statistics*, 515-519.

Table 1. Comparison of Matching Results from Basic Matcher and Frequency-Based Matcher, Los Angeles Files

		Frequency-Based			
		Match	Possible	Nonmatch	Total
Basic	Match	12320	128	7	12455
	Possible	808	2146	58	3012
	Nonmatch	8	386	3821	4215
	Total	13136	2660	3886	19682

Table 2. Comparison of Matching Results from Basic Matcher and Frequency-Based Matcher, St Louis Files

		Frequency-Based			
		Match	Possible	Nonmatch	Total
Basic	Match	305	21	2	328
	Possible	15	142	0	157
	Nonmatch	2	106	2184	2292
	Total	322	269	2186	2777