Frequency-Dependent Probability Measures for Record Linkage

William E. Yancey
Statistical Reasearch Division, U.S. Census Bureau

KEY WORDS record linkage, computer matching, value-specific

Abstract

Record linkage procedures based on the Fellegi-Sunter theory (JASA 1969) require the estimation of the conditional probabilities of the agreement patterns. Under the assumption of conditional independence, this reduces to the estimation of the conditional probabilities of the agreement of the individual matching fields. We consider methods for using value-specific, frequency-based methods to modify the agreement probabilities according to the rate of recurrence of the common matching field value in the matching set. We compare and analyze the effects of the methods when applied to Census data sets, and assess their value and usability.

1. Introduction

A record linkage methodology seeks to identify pairs of records from two files that both represent the same entity. We consider two sets A, B representing populations. We may partition the set $A \times B$ into the sets

$$\begin{aligned} M &= \{ (a,b) \in A \times B | \ a = b\} \\ U &= \{ (a,b) \in A \times B | \ a \neq b\} \end{aligned}$$

the set of matched pairs M and unmatched pairs U. For each element of our sets, we have a corresponding data record. Hence we have data from the sets A, B based on functions

$$\begin{aligned} \alpha_A &: A \rightarrow \alpha_A(A) \\ \alpha_B &: B \rightarrow \alpha_B(B) \end{aligned}$$

where $\alpha_A(A)$ and $\alpha_B(B)$ are files containing recorded information about the two populations. We tacitly assume that the data records contain enough information so that the functions α_A, α_B are one-to-one. We suppose that we have a comparison function

$$f : \alpha_A(A) \times \alpha_B(B) \rightarrow \Gamma$$

where Γ is a finite comparison space. We may denote $f(\alpha_A(a), \alpha_B(b)) = \gamma \in \Gamma$ by $\gamma(\alpha_A(a), \alpha_B(b))$ where $\gamma = (\gamma_i)$ is a comparison vector of dimension n where each γ_i takes on finitely many possible values, depending on the agreement of the records $\alpha_A(a), \alpha_B(b)$ on a set of matching fields. We wish to identify which pairs of records $(\alpha_A(a), \alpha_B(b))$ correspond to a matching pair $(a,b) \in M$ based on their corresponding comparison vector $\gamma(\alpha_A(a), \alpha_B(b))$. Usually each comparison field is binary, $\gamma_i \in \{0,1\}$, corresponding to the record pairs either disagreeing or agreeing on a particular field of comparison.

1.1 The Fellegi and Sunter Approach

We base our record linkage methods on the fundamental theoretical development due to Fellegi and Sunter [1]. The basis of record linkage decisions is the conditional probabilities

$$P(\gamma | M) \text{ and } P(\gamma | U)$$

which represent the probability that a record pair exhibits the comparison pattern γ given that the pair represents a match (resp. nonmatch). From these pattern probabilities we can compute a pattern weight

$$w(\gamma) = \log \frac{P(\gamma | M)}{P(\gamma | U)}$$

and we declare record pairs as links when they have a pattern weight above a high cutoff value, as nonlinks when they have a pattern weight below a low cutoff value, and as a clerical pair when they have a pattern weight between the cutoff values. Fellegi and Sunter show that for the false match and false nonmatch error rates for a chosen pair of cutoff values, this linkage method produces the minimal clerical region among record linkage methods with these error rates.

Since the number of comparison patterns γ grows exponentially with the number of comparisons made (i.e. the dimension n of $\gamma = (\gamma_i)$), we generally reduce the probability estimation burden by making the conditional independence assumption

$$\begin{aligned} P(\gamma | M) &= \prod_{i=1}^{n} P(\gamma_i | M) \\ P(\gamma | U) &= \prod_{i=1}^{n} P(\gamma_i | U). \end{aligned}$$

[*] 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.

This generally amounts to estimating the conditional probabilities of agreement and disagreement of a record pair on each particular field of comparison. That is, for each $i = 1, \dots, n$, we estimate $P(\gamma_i = 1 | M)$ (and hence $P(\gamma_i = 0 | M) = 1 - P(\gamma_i = 1 | M)$) and $P(\gamma_i = 1 | U)$, generally using the EM algorithm [4]. Then for any binary pattern vector γ, we compute $P(\gamma | M)$ and $P(\gamma | U)$ using the appropriate product, and thus compute the weight $w(\gamma)$ to determine our record linkage procedure.

We may question whether we could derive a more accurate record linkage procedure if we took into account the actual value in the comparison field in addition to simple agreement/disagreement. If we use more information from the files, can this result in better linkage decisions?

Under the conditional independence assumption, the agreement weight of a pattern γ is determined by the weights associated to the individual components γ_i. For simplicity, we assume that γ_i takes on values depending on the contents of a particular matching field, and let us fix that field and consequently drop the subscript i.

2. The Fellegi and Sunter Approach to Value-Specific Matching

We wish to formulate a model for frequency-based matching. Fellegi and Sunter develop the following approach to estimating the conditional probabilities for the comparison fields based on the properties of the population sets A, B.

For the purposes of frequency-based matching, we assume that the actual populations A, B have unique "true" values corresponding to the given matching field. In other words, our matching field might be surname, and we suppose that everyone in A and B has a true unique surname which we attempt to record in $\alpha_A(A)$ and $\alpha_B(B)$ respectively. Since our populations are finite, there are only finitely many, say m, actual values that this field can take on in $A \cup B$. We enumerate the number of times that each of these values occurs in each population

$$f_{A_1}, f_{A_2}, \dots, f_{A_m}; \quad \sum_{j=1}^{m} f_{A_j} = N_A$$

$$f_{B_1}, f_{B_2}, \dots, f_{B_m}; \quad \sum_{j=1}^{m} f_{B_j} = N_B$$

and on the overlapping population $A \cap B$

$$f_1, f_2, \dots, f_m; \quad \sum_{j=1}^{m} f_j = N_{A \cap B}$$

where N_C denotes the cardinality of the set C.

Fellegi and Sunter then introduce some error terms. We can think of them as corresponding to the following events occurring in the set $\alpha_A(A) \times \alpha_B(B)$.

$$\begin{aligned}
E_A &= \text{value from set } A \text{ misrecorded in set } \alpha_A(A) \\
E_B &= \text{value from set } B \text{ misrecorded in set } \alpha_B(B) \\
E_{A_0} &= \text{value from set } A \text{ missing in set } \alpha_A(A) \\
E_{B_0} &= \text{value from set } B \text{ missing in set } \alpha_B(B) \\
E_T &= \text{value } x \in A \cap B \text{ changed from } A \text{ to } B
\end{aligned}$$

The idea of this last event seems to be to capture the temporal nature of the sets A, B. If the sets are snapshots of a population at specific times, it is possible that a person's actual name or address or phone number may change over time. In any case, we consider the event

$$F = \neg E_A \wedge \neg E_B \wedge \neg E_{A_0} \wedge \neg E_{B_0} \wedge \neg E_T$$

and assume that the above events are independent so that

$$P(F) = \\ (1 - p_{E_A})(1 - p_{E_B})(1 - p_{A_0})(1 - p_{B_0})(1 - p_{E_T}).$$

Whether or not this formulation covers all errors and inconsistencies or not may be a matter of interpretation. For instance, does this cover reporting variations of the true field value, as in variant first name reports for William, Will, Bill, Billy, Willie, etc? In any case, the idea of the event F seems to be that the actual matching field value is correctly and consistently reported in both files $\alpha_A(A)$ and $\alpha_B(B)$, and the probability of event F may depend on the matching field but not on the particular pair $(\alpha_A(a), \alpha_B(b))$.

Using the notation, for $C \subset \alpha_A(A) \times \alpha_B(B)$,

$$m(C) = P(C | M)$$
$$u(C) = P(C | U)$$

if we define the events

$$\begin{aligned}
G &= \alpha_A(a) \text{ and } \alpha_B(b) \text{ agree on value} \\
G_j &= a \text{ and } b \text{ take on the } j^{\text{th}} \text{ value}
\end{aligned}$$

then Fellegi and Sunter state that

$$m(G \wedge G_j) = \frac{f_j}{N_{A \cap B}} P(F)$$

$$u(G \wedge G_j) = \frac{f_{A_j}}{N_A} \frac{f_{B_j}}{N_B} P(F).$$

In the first equation, the first factor $\frac{f_j}{N_{A\cap B}}$ represents the probability that given that $(a,b) \in M$, so that $a = b \in A \cap B$, the true value of the matching field is the j^{th} value. The second factor is $P(F)$, the probability that the matching field value was correctly and consistently recorded in both files. This formula is under the assumption that these two probabilities represent independent events and $P(F|\ M) = P(F)$.

The second equation does seem to be more of an approximation. We are given that $(a,b) \in U$, so that $a \in A, b \in B$, and $a \neq b$. To get a and b both taking on the same j^{th} matching value, we have probability $\frac{f_{A_j}}{N_A}\frac{f_{B_j}}{N_B} = \frac{f_{A_j \times B_j}}{N_{A\times B}}$, but to assure that $a \neq b$, we should subtract to get $\frac{f_{A_j}}{N_A}\frac{f_{B_j}}{N_B} - \frac{f_i}{N_A N_B}$. In general, this is a lower order correction, but for rare values it can be significant, as can be seen when a matching value is unique in each set. As before, multiplying by $P(F)$ indicates that these matching field values are correctly recorded in both files. Again we are ignoring many terms where a and b have different field values in truth, but they are recorded as identical in both files. The total probability of such accidentally agreeing matching fields may not be large for a matching field like surname, but it might be significant for a matching field like sex, where the number of and difference between field entries is small.

3. File-Based Frequency Modification

A slightly different point of view for modeling frequency-based agreement probabilities, based on [3], is to replace $P(F)$, which is based on file error probabilities that are likely difficult to estimate accurately, and instead base all of our probability estimates on the contents of the files $\alpha_A(A), \alpha_B(B)$. Thus as above, we can consider subsets $G, G_j \subset \alpha_A(A) \times \alpha_B(B)$ where

$$G = \{\alpha_A(a), \alpha_B(b) \text{ agree on value}\}$$

(except that we have changed to set notation), but now we will define

$$G_j = \left\{\alpha_A(a), \alpha_B(b) \text{ both take on the } j^{\text{th}} \text{ value}\right\}$$

where the set of possible matching field values has been determined by the contents of the data files $\alpha_A(a), \alpha_B(b)$. Now when we consider

$$\begin{aligned} m(G \cap G_j) &= P(G \cap G_j|\ M) \\ &= P(G_j|\ G, M)\, P(G|\ M), \end{aligned}$$

This probability calculus identity serves to shift the frequency emphasis from the population to the data files. The quantity $P(G|\ M)$ functions somewhat similarly as the above $P(F)$ in the sense of translating from what is true about the actual population to what is actually reflected in the files, in that if the files held completely accurate data from the population, we would expect both $P(F)$ and $P(G|\ M)$ to be nearly equal to 1. However, in the previous model, the frequencies such as f_{Aj}, f_{B_j}, f_j indicated frequencies of occurrence in the actual population sets $A, B, A\cap B$, whereas by being given G, the probability $P(G_j|\ G, M)$ indicates that we are considering the cases of record pairs that both have the j^{th} value given that they both have the same value (and they represent a match). From this viewpoint, however, the second factor is different for the nonmatch case, namely

$$u(G \cap G_j) = P(G \cap G_j|\ U) = P(G_j|\ G, U)\, P(G|\ U).$$

In the case where $P(G|\ M)$ and $P(G|\ U)$ have been estimated separately, as with the EM algorithm [2], we can compute a frequency-based adjustment by estimating $P(G_j|\ G, M)$ and $P(G_j|\ G, U)$.

By definition we have

$$\begin{aligned} P(G_j|\ G, M) &= \frac{P(G_j \cap G \cap M)}{P(G \cap M)} = \frac{P(G_j \cap M)}{P(G \cap M)} \\ P(G_j|\ G, U) &= \frac{P(G_j \cap G \cap U)}{P(G \cap U)} = \frac{P(G_j \cap U)}{P(G \cap U)} \end{aligned}$$

since $G_j \subset G$. For a given blocking subset $S \subset \alpha_A(A) \times \alpha_B(B)$, we can try to estimate the counts

$$\frac{\#(G_j \cap M)}{\#(G \cap M)} \text{ and } \frac{\#(G_j \cap U)}{\#(G \cap U)}$$

as subsets of S.

Remark 1 *Since we will now only be using the data files $\alpha_A(A), \alpha_B(B)$ and not be referring to the original population sets A, B, in the remainder of the paper we drop the α notation and think of our files as sets A and B.*

We may denote the blocking set S is of the form

$$S = \bigcup_{i=1}^{m} A_i \times B_i \subset A \times B$$

where the A_i are pairwise disjoint subsets of A and the B_i are pairwise disjoint subsets of B. Let $s_1, s_2, \ldots, s_n$ be a list of all possible values of the given matching field and let

$$\begin{aligned} A_{ij} &= \{x \in A_i \,|\, x \text{ has matching field value } s_j\} \\ B_{ij} &= \{x \in B_i \,|\, x \text{ has matching field value } s_j\} \\ a_{ij} &= \#(A_{ij}) \\ b_{ij} &= \#(B_{ij}) \\ m_{ij} &= \min(a_{ij}, b_{ij}). \end{aligned}$$

We can say that

$$\#(G_j \cap M) \leq \sum_{i=1}^{m} m_{ij}$$

and

$$\#(G \cap M) \leq \sum_{j=1}^{n} \sum_{i=1}^{m} m_{ij}.$$

One possibility is to estimate the extent of overlap in the sets and assume that a fixed proportion of the records in one set can be correctly assigned a match from the other set. That is, suppose there is a constant $0 < \rho \leq 1$ such that the number of matches brought together by the blocking criterion is a fixed proportion ρ of the maximum possible number of matches, so that the number of matches in $A_{ij} \times B_{ij}$ is ρm_{ij}. For example, if we assume that $A \subset B$, we would expect ρ to be near 1 (and $m_{ij} = a_{ij}$). In this case we would have

$$\begin{aligned} \frac{\#(G_j \cap M)}{\#(G \cap M)} &\doteq \frac{\sum_{i=1}^{m} \rho m_{ij}}{\sum_{j=1}^{n} \sum_{i=1}^{m} \rho m_{ij}} \\ &= \frac{\sum_{i=1}^{m} m_{ij}}{\sum_{j=1}^{n} \sum_{i=1}^{m} m_{ij}} \end{aligned}$$

and

$$\frac{\#(G_j \cap U)}{\#(G \cap U)} \doteq \frac{\sum_{i=1}^{m} (a_{ij} b_{ij} - \rho m_{ij})}{\sum_{j=1}^{n} \sum_{i=1}^{m} (a_{ij} b_{ij} - \rho m_{ij})}.$$

4. Tempered File-Based Frequency Modification

If we allow the binary agreement weight

$$\frac{P(G|\, M)}{P(G|\, U)}$$

to be modified by value-specific factors

$$\frac{P(G_j|\, M)}{P(G_j|\, U)} = \frac{P(G_j|\, G, M)}{P(G_j|\, G, U)} \frac{P(G|\, M)}{P(G|\, U)}$$

where we estimate

$$P(G_j|\, G, M) \doteq \frac{\sum_{i=1}^{m} m_{ij}}{\sum_{j=1}^{n} \sum_{i=1}^{m} m_{ij}} \qquad (1)$$

$$P(G_j|\, G, U) \doteq \frac{\sum_{i=1}^{m} (a_{ij} b_{ij} - \rho m_{ij})}{\sum_{j=1}^{n} \sum_{i=1}^{m} (a_{ij} b_{ij} - \rho m_{ij})} \qquad (2)$$

then the adjusted weight

$$\frac{P(G_j|\, M)}{P(G_j|\, U)}$$

can vary over a wide range.

For the most common matching field values G_j, this adjusted weight can be less than 1. This leads to the somewhat counterintuitive result that two records can have their total matching weight reduced by agreeing on a certain field value. Elsewhere I have shown that a distribution of rare and common values can always exist to product this effect for common values regardless of the average marginal agreement conditional probabilities. However, if the two records have a number of fields of agreement, the total agreement weight should still be fairly high even after such a downweighting. In any case, this effect is a mathematical consequence of this value-specific methodology, and hence one should live with the consequences if one uses the method.

On the other hand, the above formula can result in a large upweighting if the value-specific factor is very large. This can be bad if it dominates mediocre agreement weight factors from other matching fields, resulting in a high agreement rate based mostly on one field value. This is also a problem because the value-specific factor will be large when the matching field value is rare in the set of record pairs. This means that the probability estimates will be based on a small number of sample values, making the estimate more statistically suspect. Also, for a specific j^{th} value, the factor is determined by the ratio

$$\frac{\sum_{i=1}^{m} m_{ij}}{\sum_{i=1}^{m} (a_{ij} b_{ij} - \rho m_{ij})}$$

which for

$$x = \frac{\sum_{i=1}^{m} a_{ij} b_{ij}}{\sum_{i=1}^{m} m_{ij}}$$

is basically of the form

$$\frac{1}{x - \rho}$$

which changes most rapidly for small x. For example, an extreme case is for a rare field value to have either $a_{ij} = 0$ or $b_{ij} = 0$ or both $a_{ij} = 1$ and $b_{ij} = 1$ for all $1 \leq i \leq m$. That is, we have a rare field value such that whenever it appears in both files, it appears just once in each. In this case, we would have

$$\sum_{i=1}^{m} m_{ij} = \sum_{i=1}^{m} a_{ij} b_{ij}$$

and thus $x = 1$, the minimum value. Especially when the match proportion ρ is near 1, the resulting upweighting factor can be quite large. If a few of the a_{ij} or b_{ij} values change slightly, the upweighting factor can decrease substantially. Thus the value-specific factor estimate is more sensitive to errors when the value is rare in the file. For these reasons, it may be preferable to average the value-specific factors for the rate values to try to reduce the error variance and to moderate the size of the upweighting.

The value-specific conditional probabilities partition the binary conditional probabilities in the sense that

$$\sum_j P(G_j|M) = P(G|M)$$
$$\sum_j P(G_j|U) = P(G|U).$$

We can impose a cap on the value-specific factor effect by specifying a factor $\alpha > 0$ such that we only individually compute value-specific effects where

$$\frac{P(G_j|M)}{P(G_j|U)} \leq \alpha \frac{P(G|M)}{P(G|U)}$$

that is,

$$\frac{P(G_j|G,M)}{P(G_j|G,U)} \leq \alpha.$$

Let us denote

$$m_j = \sum_{i=1}^{m} m_{ij}$$
$$(ab)_j = \sum_{i=1}^{m} a_{ij}b_{ij}$$
$$N = \sum_{j=1}^{n}\sum_{i=1}^{m} m_{ij}$$
$$P = \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij}b_{ij}$$

so that our frequency-based estimates for the value-based probabilities are

$$P(G_j|G,M) \doteq \frac{m_j}{N}$$
$$P(G_j|G,U) \doteq \frac{(ab)_j - \rho m_j}{P - \rho N}.$$

Then our estimates will satisfy

$$\frac{P(G_j|G,M)}{P(G_j|G,U)} \leq \alpha \tag{3}$$

when

$$\frac{m_j}{(ab)_j - \rho m_j} \leq \alpha \frac{P - \rho N}{N} \tag{4}$$

So if we let

$$T = \left\{ j \,\middle|\, \frac{m_j}{(ab)_j - \rho m_j} > \alpha \frac{P - \rho N}{N} \right\}$$

then we can estimate the probabilities of the rare values by the average, for $j \in T$

$$P(G_j|G,M) \doteq \frac{1}{\#(T)} \sum_{j \in T} \frac{m_j}{N}$$
$$P(G_j|G,U) \doteq \frac{1}{\#(T)} \sum_{j \in T} \frac{(ab)_j - \rho m_j}{P - \rho N}$$

so that we would still have

$$\sum_j P(G_j|M) = \sum_{j \notin T} P(G_j|M) + \sum_{j \in T} P(G_j|M) = P(G|M)$$

and similarly for $P(G|U)$. The averaged value-specific weight factor $\bar{w}_T$ for $j \in T$ is estimated by

$$\begin{aligned} \bar{w}_T &= \log \frac{P(G_j|G,M)}{P(G_j|G,U)} \\ &\doteq \log \left(\frac{\sum_{j \in T} m_j}{\sum_{j \in T} (ab)_j - \rho m_j} \frac{P - \rho N}{N} \right). \end{aligned} \tag{5}$$

4.1 Tempered Frequency-Based Tables

At the Census Bureau, we have some pairs of population survey files that have been extensively reviewed to determine the record pairs that represent true matches. We can use these to evaluate record linkage method results against the truth. We experimented with this frequency-based modification for some matching fields in our Census test files. For the test files 2021, 3031, and STL, we considered the matching fields last name, first name, and street name, and values of $\alpha = 1, 2, 3$. For each field value for each file, we used the binary matching weight

$$w = \log \frac{P(G|M)}{P(G|U)}$$

based on the EM algorithm estimates. We printed out the matching field values and their frequency-based estimated value-specific weights

$$w_j = \log \frac{P(G_j|M)}{P(G_j|U)}$$

where $j \notin T$ and computed the averaged estimated weight for rare values $j \in T$. When $\alpha = 1$, we are printing out those common names which result in a downweighting factor, but the default weight $\bar{w}_T$ for names not in the list, corresponding to $j \in T$, will have a higher weight than w. As we increase α, we include more names and produce a higher default weight $\bar{w}_T$.

For each of the files and each of the selected fields, we printed out a table of all of the common names with value-based weight $w_j < w + \log \alpha$ as well as the default rate weight $\bar{w}_T$. As one would expect, the number of individual common names and the tempered default weight $\bar{w}_T$ both increase with increasing α, with $\bar{w}_T$ exceeding the overall average weight w.

4.2 Frequency-Based Matching Results

We performed several runs on the sample files using various combinations of value-specific, frequency-based matching. We provide some summary table results of some of the more extreme runs. We consider as a baseline the results of using no frequency-based matching (corresponding to $\alpha = 0$), using frequency-based matching in all three fields only for downweighting with no tempered upweight adjustment, and using frequency-based matching in all three fields using tempered adjustment with $\alpha = 3$. The general result is that the frequency-based techniques do not make much difference overall. Using only downweighting tends to lower the total weights, using tempered adjustment increases the spread by raising the weights, but there does not appear to be a great amount of movement. Somewhat arbitrarily, we grouped matcher output pairs by those with matching weight $w \geq 3$, $-2 \leq w < 3$, and $w < -2$. It seemed fairly consistent across all the data sets and all the runs that almost all pairs with $w \geq 3$ were true matches, almost all pairs with $w < -2$ were true nonmatches, and there were substantial numbers of both in the intermediate group, according to the clerical review records. It would appear that some of the exceptional cases are clerically mislabled. In any case, we see slight drifts across these arbitrary barriers. The examined pairs have been truncated by an output low cutoff of -5. It might be interesting to examine differences in the results of the different methods in terms of individual cases. Can we draw any conclusions about which if any pairs were matched differently, or were the same pairs just given different weight values? If pairs were given different weight values, which pairs had the most dramatic weight reevaluations? By performing multiple matching runs with and without frequency-based matching for different fields and reviewing record pairs that either are inconsistently output or which have large agreement weight changes, we may be able to detect some additional matches or false matches.

5. Summary and Discussion

We have seen that in the context of our record linkage examples, adding the extra refinement of value-specific modifications does not significantly affect the matching results. Frequency-based calculations may improve record linkage results in other data or methodology contexts. We should note that our test examples are pairs of modest sized, relatively clean personal data files which represent substantially overlapping populations from a fairly small geographic region. Our frequency-based adjustments are applied within the methodology context which computes agreement weights for record pairs under the conditional independence model where the individual field agreement conditional probabilities are estimated to represent the average agreement probabilities over all field values.

When matching these kinds of files, some of the matching fields help to distinguish between households and some fields help to individuate persons within households. Fields such as last name and street name along with house number tend to determine household agreement. Adding value-specific adjustments to last name or street name fields may not significantly enhance common household identifiability. The first name field has strong distinguishing power between individuals within a household, but again the employment of statistics about specific first name distributions in the entire population at hand may not have much added effect on distinguishing individuals within a household.

REFERENCES

Fellegi, Ivan P. and Alan B. Sunter. "A Theory for Record Linkage." *JASA*. December, 1969.

Winkler, William E. "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Survey Research Methods Section,* Amer. Stat. Assn, pp.667-671. 1988.

Winkler, William E. "Frequency-Based Matching in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Survey Research Methods Section,* pp.778-783. Amer. Stat. Assn. 1989.

Winkler, William E. "Comparative Analysis of Record Linkage Decision Rules." *Proceedings of the Section on Survey Research Methods,* pp.829-834. Amer. Stat. Assn. 1992.