BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR2000/05

Using the EM Algorithm for Weight Computation in the 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

USING THE EM ALGORITHM FOR WEIGHT COMPUTATION IN THE FELLEGI-SUNTER MODEL OF RECORD LINKAGE

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

ABSTRACT

Let **A**×**B** be the product space of two sets **A** and **B** which is divided into a (pairs representing the same entity) and *nonmatches* (pairs representing different entities). Linkage rules are those that divide **A**×**B** into *links* (designated matches), *possible links* (pairs for which we delay a decision), and *nonlinks* (designated nonmatches). Under fixed bounds on the error rates, Fellegi and Sunter (1969) provided a linkage rule that is optimal in the sense that it minimizes the set of possible links. The optimality is dependent on knowledge of certain joint inclusion probabilities that are used in a crucial likelihood ratio. In applying the record linkage model, assumptions are often made that allow estimation of weights that are a function of the joint inclusion probabilities. If the assumptions are not met, then the linkage procedure using estimates computed under the assumptions may not be optimal. This paper describes a method for estimating weights using the EM Algorithm under less restrictive assumptions. The weight computation automatically incorporates a Bayesian adjustment based on file characteristics.

Keywords and phrases. Decision rule, error rate.

1. INTRODUCTION

The paper describes a method for using the EM Algorithm (Dempster, Laird, and Rubin 1977, Wu 1983) to improve computational procedures in applications of the Fellegi-Sunter model of record linkage.

Let **A**×**B** be the product space of two sets **A** and **B** which is divided into *matches* (pairs representing the same entity) and *nonmatches* (pairs representing different entities). Linkage rules are those that divide **A**×**B** into *links* (designated matches), possible links (pairs for which we delay a decision), and *nonlinks* (designated nonmatches). Under fixed bounds on the error rates, Fellegi and Sunter (1969, hereafter denoted FS) provided a linkage rule that is optimal in the sense that it minimizes the set of possible links. The optimality is dependent on knowledge of certain joint inclusion probabilities that are used in a crucial likelihood ratio.

In applications, an independence assumption is made that allows estimation of joint inclusion probabilities. If the independence assumption is not valid (Winkler 1985, Kelley 1986), then linkage rules based on the estimated probabilities may not be optimal.

The remainder of this paper contains a methodology for estimating weights under a less restrictive assumption. Section two is divided into four parts. The first part provides a summary of the FS Model of record linkage. The second describes the Conditional Independence Assumption and how computation is simplified under it. The third part introduces a more general class of distributions than those satisfying the Conditional Independence Assumption. For the new class it shows how simple agreement/disagreement weights are computed using the EM Algorithm. The fourth part presents a procedure for deriving frequency-based weights when an additional assumption is met.

The assumption is weaker than the assumption of FS (pp. 1207-1210).

The discussion in the third section comprises four components. The first describes convergence properties of the EM Algorithm. The second describes how the EM Algorithm automatically makes as Bayesian adjustment based on file characteristics. The third discusses the computation of frequency-based weights. The fourth points out how the EM Algorithm, with possible enormous increase in computation, can be extended to parameter estimation for a reasonably general class of distributions. The final section is a summary.

2. MODEL AND COMPUTATIONAL PROCEDURES

2.1. Fellegi-Sunter Model

The FS Model uses an decision-theoretic approach embodying principles first used in practice by Newcombe (Newcombe et al. 1959). To give an overview, we describe the model in terms of ordered pairs in a product space. The presentation closely follows FS (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 **A** and **B** are denoted by $\alpha(a)$ and $ß(b)$, respectively. The *comparison vector* τ associated with the records is defined by:

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

Where confusion does not arise, the function τ on $\mathbf{A}\times\mathbf{B}$ will be denoted by $\tau(\alpha,ß)$, $\tau(a,b)$, or τ. The set of all possible realizations of τ is denoted by Γ.

The conditional probability of $\tau(a,b)$ if $(a,b)\varepsilon M$ is given by

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

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

Similarly we denote the conditional probability of τ if $(a,b)\varepsilon U$ by $u(\tau)$.

We observe a vector of information $\tau(a,b)$ associated with pair (a,b) and wish to designate a pair as a link (in set A_1), a possible link (in set A_2), or a nonlink (in set A_3). We let L denote a linkage rule that divides $\mathbf{A}\times\mathbf{B}$ into A_1, A_2, and A_3. We say that a *Type I error* has occurred if rule L places $m\varepsilon M$ in A_3,

$$P(A_3|M) = \sum_{\tau \varepsilon \Gamma} m(\tau) \cdot P(A_3|\tau),$$

and a *Type II error* if L places $u \varepsilon U$ in A_1,

$$P(A_1|U) = \sum_{\tau \varepsilon \Gamma} u(\tau) \cdot P(A_1|\tau).$$

FS define a linkage rule L_0 with associated sets A_1, A_2, and A_3 that is optimal in the following sense:

THEOREM (Fellegi and Sunter 1969). Let L' be a linkage rule with associated sets A_1', A_2', and A_3' such that $P(A_3'|M) = P(A_3|M)$ and $P(A_1'|U) = P(A_1|U)$. Then $P(A_2|U) \leq P(A_2'|U)$ 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' is always greater than under L_0. The FS 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.

2.2. Computational Procedures

The subsection is divided into four parts. The first describes the general form of the linkage rule. The second presents a simplification of the computational procedures under the Conditional Independence Assumption. The third contains a weaker assumption and the estimation of parameters using the EM Algorithm. The fourth extends parameter estimation to frequency-based weights.

2.2.1. General Form of Linkage Rule

To provide a background for understanding why specific computational procedures are used, we consider the following likelihood ratio

$$R \equiv R[\tau(a,b)] = m(\tau)/u(\tau). \tag{2.1}$$

If the numerator is positive and the denominator is zero in (2.1), we assign a fixed very large number to the ratio. The FS linkage rule takes the form:

If $R > T_\mu$, then denote (a,b) as a link.

If $T_\lambda \leq R \leq T_\mu$, 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.

2.2.2. Simplification Under Conditional Independence Assumption

In practice, computation is simplified two ways. The first is by the Conditional Independence Assumption of FS:

$$m(\tau) = m_1(\tau^1)\cdot m_2(\tau^2)\cdots m_K(\tau^K) \text{ and}$$

$$u(\tau) = u_1(\tau^1)\cdot u_2(\tau^2)\cdots u_K(\tau^K)$$

where for $i = 1, 2, \cdots, K$

$$m_i(\tau^i) = P(\tau^i \mid (a,b)\varepsilon M) \text{ and}$$

$$u_i(\tau^i) = P(\tau^i \mid (a,b)\varepsilon U).$$

The second is to use a computationally convenient function of the ratio in (2.1). Log_2 is used. We then have

$$W \equiv W(\tau) = \text{Log}_2[m(\tau)/u(\tau)] = W^1 + W^2 + \cdots + W^K, \tag{2.3}$$

where $W^i \equiv \text{Log}_2[m_i(\tau^i)/u_i(\tau^i)]$ for $i = 1, 2, \cdots, K$. We call W the *total comparison weight* associated with a pair and W^i, $i = 1, 2, \cdots, K$, the *individual comparison weights*.

2.2.3. Weaker Assumption and EM Algorithm

To describe the computational assumption that is weaker than the Conditional Independence Assumption, we need some additional background. For the remainder of the paper, unless otherwise stated, we will assume that each component τ^i, $i = 1, 2, \cdots K$, in τ represents a two state comparison (e.g., agree/disagree) and define the marginal comparison events by

$$B^i \equiv \{(a,b) \mid \tau^i(a,b) = \tau^i_o\}$$

for one fixed state of τ^i_o.

Let $C_1, C_2, \cdots, C_K$ be any reordering of $B_1, B_2, \cdots, B_K$. On any set of pairs Q in $\mathbf{A}\times\mathbf{B}$

$$P(\tau\varepsilon C_1\cap C_2\cap\cdots\cap C_K \mid Q) =$$

$$P(\tau\varepsilon C_1 \mid Q)\cdot P(\tau\varepsilon C_2 \mid C_1, Q)\cdots P(\tau\varepsilon C_K \mid C_1,\cdots,C_{K-1},Q). \tag{2.4}$$

For $\tau\varepsilon\Gamma$ we can consider each $P(\tau\varepsilon C_i|C_1,\cdots,C_{i-1},Q)$ as the successive incremental discriminating power of C_i in Q, $i = 1, 2, \cdots, K$. The discriminating power is dependent on the ordering $C_1, C_2, \cdots, C_K$, and the pairs in $C_1\cap C_2\cap\cdots\cap C_K\cap Q$. For record pairs r_j, $j = 1, 2, \cdots, N$, from Q, index the comparison vectors τ_j^i as follows:

$$\tau_j^i = \begin{cases} = 1 & \text{if field } i \text{ agrees for record pair } r_j \\ = 0 & \text{if field } i \text{ disagrees for record pair } r_j. \end{cases}$$

The elements in $Q = (Q\cap M)\cup(A\cap U)$ are distributed according to a finite mixture with the

unknown parameters $\Phi = (m, u, p)$ where p is the proportion of matched pairs in Q. Let $\mathbf{x}$ be the complete data vector $g = \langle \tau_j, g_j \rangle$ where

$g_j = (1,0)$ if $r_j \in M \cap Q$ and

$g_j = (0,1)$ if $r_j \in U \cap Q$.

Then the complete data log-likelihood (Dempster, Laird, and Rubin 1977, pp. 15-16) is given by

$$\ln f(\mathbf{x} \mid \Phi) = \sum_{j=1}^{N} g_j \cdot \langle \ln P(\tau_j \mid M \cap Q), \ln P(\tau_j \mid U \cap Q) \rangle$$

$$+ \sum_{j=1}^{N} g_j \cdot \langle \ln p, \ln(1-p) \rangle.$$

Fitting using the EM Algorithm will be performed under the following assumption: There exist vector constants $m \equiv (m_1, m_2, \cdots, m_K)$ and $u \equiv (u_1, u_2, \cdots, u_K)$ such that, for all $\tau \in \Gamma$,

$$P(\tau \mid M \cap Q) = \prod_{i=1}^{K} m_i^{\tau^i} (1-m_i)^{(1-\tau^i)}$$

and (2.5)

$$P(\tau \mid U \cap Q) = \prod_{i=1}^{K} u_i^{\tau^i} (1-u_i)^{(1-\tau^i)}.$$

Probabilities m_i and u_i, $i = 1, 2, \cdots, K$, are constant for all representations τ of pairs in Q. The set of probabilities of form (2.5) includes those obtained by reorderings as in (2.4), provided the reordering is fixed for all $\tau \in \Gamma$. To avoid trivialities, we assume that $0 < m_i, u_i < 1$, $i = 1, 2, \cdots, K$.

If the Conditional Independence Assumption holds, then the m_i and u_i, $i = 1, 2, \cdots, K$, are the usual marginal probabilities as given in FS (pp. 1194-1195). They necessarily are independent of any reordering of $B_1, B_2, \cdots, B_K$. When the Conditional Independence Assumption does not hold probability distributions of form (2.5) constitute a more general class than those obtained under the Conditional Independence Assumption.

We begin the EM Algorithm with estimates of the unknown parameter $\langle \hat{m}, \hat{u}, \hat{p} \rangle$. For the E-step under (2.5), replace g_j with

$\langle P(M \cap Q \mid \tau_j), P(U \cap Q \mid \tau_j) \rangle$ where

$$P(M\cap Q|\tau_j) \equiv \frac{\hat{p}\prod_{i=1}^{K} \hat{m}_i^{\tau_j^i}(1-\hat{m}_i)^{(1-\tau_j^i)}}{\hat{p}\prod_{i=1}^{K} \hat{m}_i^{\tau_j^i}(1-\hat{m}_i)^{(1-\tau_j^i)} + (1-\hat{p})\prod_{i=1}^{K} \hat{u}_i^{\tau_j^i}(1-\hat{u}_i)^{(1-\tau_j^i)}}$$

and (2.6)

$$P(U\cap Q|\tau_j) \equiv \frac{(1-\hat{p})\prod_{i=1}^{K} \hat{u}_i^{\tau_j^i}(1-\hat{u}_i)^{(1-\tau_j^i)}}{\hat{p}\prod_{i=1}^{K} \hat{m}_i^{\tau_j^i}(1-\hat{m}_i)^{(1-\tau_j^i)} + (1-\hat{p})\prod_{i=1}^{K} \hat{u}_i^{\tau_j^i}(1-\hat{u}_i)^{(1-\tau_j^i)}}.$$

For the M step, the complete data log-likelihood can be separated into three maximization problems. Setting the partial derivatives equal to zero and solving for $\hat{m}_i$, i = 1, 2, ···, K, yields:

$$\hat{m}_i = \frac{\sum_{j=1}^{N} P(M\cap Q \mid \tau_j)\cdot \tau_j^i}{\sum_{j=1}^{N} P(M\cap Q \mid \tau_j)}. \quad (2.7)$$

Estimates $\hat{u}_i$, i = 1, 2, ···, K, are derived in a similar manner. The matrix of second partial derivatives can be shown to be negative-definite. The estimate of the proportion of matched pairs is given by

$$\hat{p} = \frac{\sum_{j=1}^{N} P(M\cap Q \mid \tau^j)}{N}.$$

2.2.4. Extension to Frequency-Based Weights

This section considers a procedure for extending simple agreement/disagreement weights to weights that account for frequency. We call such a procedure a *dispersion*. When the more stringent assumptions of FS (pp. 1207-1210) are satisfied our dispersion procedure agrees with theirs. If the agreement/disagreement weights found via the EM Algorithm coincide with the agreement/disagreement weights found via the FS procedures, then the frequency-based weights also coincide.

Frequency-based weights are useful because they can account for the fact that a specific surname pair such as (Zabrinksy, Zabrinsky) occurs less often than a surname pair such as (Smith, Smith).

We need some background material before presenting the computational procedures for frequency-based weights.

We observe that if, for some i and k,

$$m_i = P(\tau^k = 1 \mid M \cap Q)$$

and (2.8)

$$u_i = P(\tau^k = 1 \mid U \cap Q),$$

then the kth comparison is independent of the other $K-1$ comparisons. The right hand sides of (2.8) are just the appropriate marginal inclusion probabilities. Note that m_i and u_i, $i = 1, 2, \cdots, K$, of this paper generally differ from the m_1, m_2, m_3, u_1, u_2, and u_3 in FS (pp. 1194-1195, 1207-1210).

We define a random variable $\tau^{\wedge k}$ by

$$\tau^{\wedge k} = \mu_j^k \text{ if the } k\text{th comparison pair takes value } \mu_j^k$$

where μ_j^k, $j = 1, \cdots, L^k$, is an enumeration of the specific values of the kth comparison. We make two assumptions:

A1. Agreement/disagreement in the kth comparison is independent of the other $K-1$ comparisons.

A2. There exists a comparison k' such that the specific realizations of $\tau^{\wedge k}$ are pairwise independent of agreement/disagreement in the k'th comparison.

If we consider one comparison, say of agreement/disagreement in surname, then we can perform EM fitting under a restricted version of (2.5) by specifying that one of the (m_i, u_i) must converge to the marginal probabilities (as in (2.8)) associated with surname. We can, thus, always find a comparison satisfying assumption A1 for the restricted class of distributions.

Assumption A2 is a weaker form of independence assumption than the one considered by FS (p. 1208). It allows dispersion of the agreement/disagreement weight obtained under assumption A1 to frequency-based weights.

In a manner similar to the dispersion of FS (pp. 1207-1210), we define

$$N_k(\mu_i^k) = P(\tau^{\wedge k} = \mu_i^k, \tau^{k'}=1),$$

$$V_k(\mu_i^k) = P(\tau^{\wedge k} = \mu_i^k),$$

c = # pairs in Q, and

N = # pairs in $M \cap Q$.

Then, for $i = 1, 2, \cdots L^k$,

$$c \cdot N_k(\mu_i^k) = N \cdot P(\tau^{\wedge k} = \mu_i^k \mid M \cap Q) \cdot P(\tau^{k'}=1 \mid M \cap Q) + \quad (2.9)$$

$$(c - N) \cdot P(\tau^{\wedge k} = \mu_i^k \mid U \cap Q) \cdot P(\tau^{k'}=1 \mid U \cap Q)$$

and

$$c \cdot V_k(\mu_i^k) = N \cdot P(\tau^{\wedge k} = \mu_i^k \mid M \cap Q) + \quad (2.10)$$

$$(c - N) \cdot P(\tau^{\wedge k} = \mu_i^k \mid U \cap Q).$$

In (2.9) and (2.10) c, $N_k(\mu_i^k)$, $V_k(\mu_i^k)$, $i = 1, 2, \cdots L^k$, can be computed directly because they are based on observed file characteristics. The marginal probabilities $P(\tau^k=1 \mid M \cap Q)$ and $P(\tau^k=1 \mid U \cap Q)$ and the number of matches N in $M \cap Q$ can be computed using the estimated parameters of (2.4) that are obtained by the EM Algorithm. Equations (2.9) and (2.10), thus, consist of two equations to be solved for the two unknowns $P(\tau^{\wedge k} = \mu_i^k \mid M \cap Q)$ and $P(\tau^{\wedge k} = \mu_i^k \mid U \cap Q)$, $i = 1, 2, \cdots L^k$.

3. **DISCUSSION**

This section is divided into four parts. The first discusses the convergence properties of the EM Algorithm. The second describes how the EM Algorithm automatically incorporates a Bayesian adjustment of weights for file characteristics. The third considers the extension to frequency-based weights. The fourth deals with the extension of the EM Algorithm to parameter estimation for a reasonably general class of distributions.

3.1. Convergence Properties of EM Algorithm

This paper's application of the EM Algorithm most closely resembles the more general approach of Hasselblad (1969). Hasselblad noted the increase in likelihood on successive steps but was unable to prove convergence. Haberman (1977) proved convergence in a substantially more general setting. He observed that the limiting value was dependent of the initial values of the parameters and, thus, not necessarily unique.

Wu (1983) noted that limiting values of the EM Algorithm are stationary points that can either be saddle points or local maxima. He made the conjecture that there is unlikely to be any general condition that assures convergence to a unique maximum. Wu did observe, however, that if the likelihood is unimodal, if the estimated parameters have at most one stationary limiting point, and if a technical condition holds (which it does for the distributions of this paper), then the estimated parameters converge to the unique maximizer of the likelihood. The implication is that, while the EM algorithm of this paper is of value in accounting for failures of the Conditional Independence Assumption, several starting points for the EM Algorithm should be used. The estimated parameters associated with the largest local maximum are the ones that are used. If we can show that there is at most one stationary limiting point, then the parameter estimates will necessarily converge to it.

3.2. Bayesian Adjustment of Weights

If weights are computed via Method II of FS (p. 1194), then they correspond to observed characteristics of the files. Similarly, weights computed using the application of the EM Algorithm of this paper will also correspond to file characteristics.

If, during a file updating project, we use as initial weights those weights obtained from an earlier project using different files, then the EM Algorithm will automatically adjust the weights to the characteristics of the new files. 3.3. Extension to Frequency-Based Weights

Under the more stringent assumptions of FS (pp. 1207-1210) frequency-based weights computed using the techniques of this paper agree with those computed using Method II of FS. This follows because the dispersion method of this paper is identical to the dispersion method of FS and the there can exist at most one local maximum of the likelihood.

The chief value of assumptions like Assumptions A1 and A2 of this paper is that they allow dispersal of agreement/disagreement weights with little increase in computation. Although the EM Algorithm might be extended to allow direct computation of frequency-based weights, such an extension will generally require enormous increases in computation.

3.4. Extension of Computational Procedures

Distributions of form (2.5) are not sufficiently general to deal with the joint relationship between A_1 and A_2. Form (2.5) yields distributions having essentially independent effects. Generally, use of such distributions does not allow effective modeling of events of the form

$$P(\tau\varepsilon A_1^x\cap A_2^x|\ A_3\cap M\cap Q) \text{ and } P(\tau\varepsilon A_1^x\cap A_2^x|\ A_3\cap U\cap Q)$$

where A_i^x represents the comparison event or its complement. A_3 represents any other comparison or set of comparisons.

Rather than give a fully rigorous development, we will consider a relatively simple case. There are two comparisons that are independent of other comparisons but may not be independent of each other. Due to the form of the independence, we can suppress any parameters associated with comparisons other than the first two.

There exist vector constants $m \equiv (m_1, m_2, m_{12})$ and $u \equiv (u_1, u_2, u_{12})$ such that, for all $\tau\varepsilon\Gamma$

$$P(\tau \mid M\cap Q) = m_1^{\tau^1\cdot\tau^2}\cdot(1-m_1)^{(1-\tau^1)\cdot\tau^2}\cdot m_2^{\tau^2}\cdot(1-m_2)^{(1-\tau^2)}\cdot m_{12}^{\tau^1\cdot(1-\tau^2)}\cdot(1-m_{12})^{(1-\tau^1)\cdot(1-\tau^2)}$$

and (3.1)

$$P(\tau \mid U\cap Q) = u_1^{\tau^1\cdot\tau^2}\cdot(1-u_1)^{(1-\tau^1)\cdot\tau^2}\cdot u_2^{\tau^2}\cdot(1-u_2)^{(1-\tau^2)}\cdot u_{12}^{\tau^1\cdot(1-\tau^2)}\cdot(1-u_{12})^{(1-\tau^1)\cdot(1-\tau^2)}$$

Representations of form (3.1) with additional parameters m_{12} and u_{12} can account for the interaction of events A_1 and A_2.

In special cases, form (2.5) might allow modelling such pairwise interactions. If, in (3.1), m_1 =

m_{12} and $u_1 = u_{12}$, we have independence and form (3.1) agrees with form (2.5). Formal extension to general cases is straightforward.

Use of the EM Algorithm with distributions of form (3.1) is straightforward. Probability distributions of form (2.5) in the E step (2.6) are replaced with distributions of form (3.1). The M-step estimates still take form (2.7). We note that two additional parameters m_{12} and u_{12} must be estimated. The reason the extension is straightforward is that the complete data log-likelihood of the M step takes essentially the same form it did for the simpler class of distributions (2.5).

To create classes of distributions that more effectively model multiple interactions of events necessitates straightforward extension of the probability distributions of form (3.1). In practice, we would likely select just a few interactions to model in order to minimize the increase in computation involved with the EM Algorithm.

4. SUMMARY

This paper provides a method of estimating matching weights in the Fellegi-Sunter model of record linkage. Under an assumption weaker than the usual Conditional Independence Assumption, the estimates are obtained via the EM Algorithm. The procedure automatically incorporates a Bayesian adjustment for file characteristics.

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 The shorter version of this paper was presented at the American Statistical Association Annual Meeting and appeared in the 1988 Proceedings of the Section on Survey Research Methods (pp. 667-671).

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 Society*, B, **39**, 1-38.

Haberman, S. J. (1977), "Product Models for Frequency Tables Involving Indirect Observation: Maximum Likelihood Equations," *Annals of Statistics*, **5**, 1124-1147.

Haselblad, V. (1969), "Estimation of Finite Mixtures of Distributions from Exponential Family," *Journal of the American Statistical Association*, **64**,1459-1471.

Kelley, R. P. (1986), "Robustness of the Census Bureau's Record Linkage System, American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 620-624.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.

Winkler, W. E. (1987a), "An Application of the Fellegi-Sunter Model of Record Linkage to Lists of Businesses," Energy Information Administration Technical Report.

Winkler, W. E. (1987b), "Computational Aspects of Applying the Fellegi-Sunter Model of Record Linkage to Lists of Businesses," paper presented at the Symposium on Statistical Uses of Administrative Data.

Wu, C. F. J. (1983) "On the Convergence Properties of the EM Algorithm," Annals of Statistics, **11**, 95-103.