BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-86/21

# USE OF FIELD INFORMATION TO MATCH THE RECORDS IN TWO FILES

by

Beverley Causey
Statistical Research Division
Bureau of the Census



Recommended by: Kirk M. Wolter

Report completed: October 16, 1986

Report issued: October 16, 1986

DRAFT
Bev Causey:jo:4C
June 30, 1986

**Use of Field Information to Match the Records in Two Files**

Consider two files of records. Within each file, each record corresponds to a different population unit; but the two files correspond to the same general population. We want to identify "matches," i.e., pairs of records (from the two files) that each correspond to the same population unit.

Each record contains data in K fields which correspond to characteristics such as age, race, etc. We may observe patterns of agreement/disagreement among the fields, for each pair of records. Using this information, we want as best as possible to identify matches. The problem of how best to use the field information has been addressed for K=3, under assumption that the events "agreement in field i," i=1, ..., K are stochastically mutually independent -- for true matches and likewise for true nonmatches. We address the problem for K>3, and avoid reliance on the assumption of independence by fitting interaction terms which reflect stochastic positive dependences.

Key words: record linkage, field dependence, convex programming, log-linear model.

1. Problem Definition. Suppose we have two files "A" and "B" containing $N_A$ records and $N_B$ records respectively. Each A-record corresponds to a different population unit, likewise each B-record. The two files correspond to the same general population; and our goal is to identify "linkages": a linkage is a coupling of an A-record and a B-record that correspond to the same population unit. In other words, we want to decide which pairings correspond to the same population unit, and which do not. We refer to Fellegi and Sunter (1969) for a lengthy discussion of many aspects of this general, frequently arising situation, including their "method of moments," which provides a starting point for the method of this paper. (Because of the directions that this paper takes, and the resulting complications, we depart from the notation of that paper. No disrespect is intended.) As that paper indicates, these files may represent samples from larger populations. We presume that, by whatever means available, we have "unduplicated" each file so that, in essence, no two records within a file correspond to the same population unit. Coulter (1985) suggests one scheme for this "unduplication."

As in our reference list, the Coulter paper appeared in the Proceedings of a May 1985 workshop on exact matching methodologies, co-sponsored by the Internal Revenue Service, the Washington Statistical Society, and the Federal Committee on Statistical Methodology. All our references on record linkage appeared in these lengthy proceedings and, in some cases, appeared earlier elsewhere.

There are $N_A N_B$ possible pairings of A-records and B-records. Out of these, the maximum possible number of "true" matches (same population unit) is $\min(N_A, N_B)$. Hence we introduce

$$F_O = \min(N_A, N_B)/N_A N_B \qquad (1)$$

as the maximum possible fraction of pairings that are true matches; we do not know how many pairings are true matches, but

$F_0$ gives an upper bound. Typically, however, our files, and the population that underlies them, will be "blocked" into disjoint subsets in a parallel fashion so as to guarantee that, if a population unit appears in both files A and B, it will with virtual certainty appear in the same block for both. (If not, we will not be able to identify the true match for this population unit.) Kelley (1985) and Winkler (1985) discuss strategies for blocking; later in this paper we discuss the subject. For now, suppose that the files contain $N_{hA}$ and $N_{hB}$ records in block h. We consider the possible pairings of A and B within each block. Summing over all blocks, we get $\sum N_{hA}N_{hB}$ and $\sum \min(N_{hA},N_{hB})$ as number of possible pairs and maximum possible number of true matches. As maximum possible fraction we now have

$$F_0 = [\sum_h \min(N_{hA},N_{hB})]/(\sum_h N_{hA}N_{hB}) \ . \qquad (2)$$

Typically, $F_0$ will be close to 0. (If external information permits us to determine an upper bound which is even smaller, we could use that in place of (2). The actual fraction may in practice be considerably less than $F_0$. For this reason one ought to carry out analyses of this paper using a smaller $F_0$, as a sensitivity check.)

For each pair we compare K data fields that appear in both the A- and B-records; these fields typically correspond to such characteristics as age, race, or street name. There are $2^K$ cells corresponding to patterns of field agreement/disagreement; each pair falls into one and only one of the cells. We will number the fields from 1 to K. Let S denote a subset of these fields. We designate "cell S" as that cell for which there is agreement for the fields in S, and disagreement for the other fields. Note the terminology: agree/disagree pertains to fields, and match/nonmatch to population units. Thus cell S corresponds to the vector outcome $\gamma$ , associated with a pair, that appears in many papers on record linkage. As an example let K=5 and $S = \{2,3,5\}$. The cell S consists of those pairs for which there

is agreement in fields 2, 3 and 5, and disagreement in fields 1 and 4.

Let $P_S$ denote the observed proportion of all pairs that land in cell S. Let F denote the fraction of all pairs that are true matches. Let $Q_{1S}$ denote the proportion of all true match pairs that fall into cell S, and $Q_{2S}$ denote the same for true non-matches. We have

$$P_S = FQ_{1S} + (1-F)Q_{2S} . \qquad (3)$$

Although we observe $P_S$, we cannot observe its components F, $Q_{1S}$ and $Q_{2S}$; we only know $F \leq F_0$. Using only the known $P_S$ and $F_0$, we will fit values for F, $Q_{1S}$ and $Q_{2S}$.

Once these values are fitted, Fellegi and Sunter (in their Section 2) use $Q_{1S}$ and $Q_{2S}$ to form a rule for deciding whether a pair represents a true match. The procedure corresponds to a hypothesis test in its formation of critical regions based on largest and smallest values of the likelihood ratio - here, $Q_{1S}/Q_{2S}$ - and in its consideration of Type I and Type II error probabilities. Hence this is one major use of $Q_{1S}$ and $Q_{2S}$. Henceforth we let

$U_S$ denote $Q_{1S}/Q_{2S}$.

Instead of the hypothesis test, in this paper we use $Q_{1S}$ and $Q_{2S}$ to develop a posterior probability of true match: that is, the probability of a true match, given that a pair is in cell S. From this point one can make an assignment of file-A records to file-B records by solving a transportation-type linear-programming problem so as to maximize the posterior expected number of true matches, i.e., the sum of posterior probabilities of match, in the assignment. But instead, we might simply specify a cutoff decision rule that a posterior probability exceeding, say, .90, corresponds to true match (and less than .90 does not). Interpretation is simplified by the fact that posterior probabilities largely appear to gravitate toward either

0 or 1. We are inclined to favor investigation of the cutoff decision rule, and thus will not deal here with the intricacies of making an assignment based on the transportation problem.

Hence we have discussed uses of $Q_{1S}$ and $Q_{2S}$ once they are fitted. For the most part this paper considers how best to fit $Q_{1S}$ and $Q_{2S}$, in advance of using them. We will consider fits which give an end result of form

$$\log U_S = \text{constant plus } \sum_{i \in S} w_i \qquad (4)$$

with $w_i > 0$; this model appears in many of the Proceedings papers on record linkage. A (relative) weight $w_i$ is being assigned to field i. By using (4) we maintain a proper hierarchy among likelihood ratios: if (the set of agreeing fields for) S is strictly contained in (same for) cell T, we want $U_S < U_T$. In such an instance, T represents a clear-cut advance beyond S with respect to field agreements; and we want $U_S < U_T$ so that for T the choice "match" is favored over the choice "non-match" more than it is for S. This hierarchy must be maintained; it is not enough merely to fit $Q_{1S}$ and $Q_{2S}$ which are positive.

The relationship in (4) is maintained under the "independence" model which we present at the start of the next section. Such a model has been used in other papers on record linkage. However, we do not wish to operate under this assumption of independence among field agreements. (This assumption states that, for true-match pairs, the events "agreement in field i" and "agreement in field j" are, for $i \neq j$, stochastically independent, and that for true-nonmatch pairs the same thing holds. But typically, for match or nonmatch or both, there will be positive dependence between these events for at least some (i,j). Rather than consider the possible robustness of results based on the independence assumption against positive dependence, we want to obtain results which take the positive dependence into account.) Thus we will develop a log-linear model which, as far as possible, accounts for positive dependences among field agreements - while preserving the

relationship in (4), and the accompanying hierarchy for $U_S$. The development and use of this latter model is the primary idea of the paper.

In regard to our model and method, we stress that the only needed inputs are the observable cell proportions $P_S$. We do not rely on a sample of pairs which are known to be true matches (or true nonmatches) (e.g., Tepping (1968)) for information. (With $F_0$ small, true matches are very scanty.) Nor do we make use of the transmission rates of Howe and Lindsay (1981): probabilities (known) that a recorded field entry is the true entry.

We will use the following example as illustration. We considered files for (A) the 1980 Decennial Census, and (B) the Post-Enumeration Survey for that census. Each record corresponds to a person. We use K = 5 fields:

(1) A combination of race and ethnicity
(2) A combination of gender and marital status
(3) Date of birth (month and year)
(4) Surname
(5) A combination of street name and location within community.

We considered only those records for which all fields were filled. With "blocking" based on "Census Bureau numbering area," there were 49 blocks. The number of total pairs was 1321100, the number of possible true matches 4465, and the value of $F_0$ .003380. The 32 cells, with accompanying $P_S$, are in Table 1; the 5 integers preceding $P_S$ indicate field agreement (1) or field disagreement (0) for the 5 fields. That is, 11000

**Table 1. Values of $P_S$**

| | | | |
|---|---|---|---|
| 00000 .334568 | 00010 .000724 | 00001 .046498 | 00011 .000751 |
| 10000 .353175 | 10010 .001824 | 10001 .081199 | 10011 .002542 |
| 01000 .063819 | 01010 .000121 | 01001 .010829 | 01011 .000172 |
| 11000 .080692 | 11010 .000522 | 11001 .018664 | 11011 .000709 |
| 00100 .000318 | 00110 .000000 | 00101 .000098 | 00111 .000016 |
| 10100 .000372 | 10110 .000017 | 10101 .000179 | 10111 .000100 |
| 01100 .000129 | 01110 .000005 | 01101 .000052 | 01111 .000202 |
| 11100 .000173 | 11110 .000125 | 11101 .000142 | 11111 .001266 |

indicates agreement on fields 1 and 2, and disagreement for 3-4-5. As we see, slightly more than 2/3 of all pairs disagree on all of fields 2-3-4-5.

2. Outline of Method. This section attempts an overview of the paper. We first address the assumption of independence: consider a model of form

$$Q_{1S} = [\prod_{i \in S} y_i][\prod_{i \notin S}(1-y_i)] \ . \tag{5}$$

We express $Q_{2S}$ likewise with $u_i$ replacing $y_i$. Let $E_i$ denote the event "agreement in field i." Then $y_i$ corresponds to $P(E_i)$, conditional, given that there is a true match. The model (5) states that the events $E_1$, ..., $E_K$ are mutually independent, given that there is a true match. Similar mutual independence holds, given that there is a true non-match, with $u_i$ corresponding to conditional $P(E_i)$. Although we later seek to avoid this assumption of mutual independence, we use the above independence model to develop preliminary results. These initial results, in Sec. 3, will in essence extend the 3-field results of Fellegi and Sunter (their Sec. 3.32 and Appendix 2). From (5), the resulting value of $w_i$ in (4) is seen to be

$$w_i = \log(y_i(1-u_i)/u_i(1-y_i)) \tag{6}$$

under this independence model. We will fit $u_i < y_i$, so that $w_i > 0$.

Let $p_i$ denote the fraction of all pairs for which there is agreement in field i, $p_{ij}$ denote the fraction for which there is agreement in fields i and j, and so forth. As a needed step in developing our method, we must consider

$$R_{ij} = p_{ij} - p_i p_j, \text{ for } i \neq j. \tag{7}$$

Incidentally, in our example the values of $p_i$ are

$$.5417,\ .1776,\ .00319,\ .0091,\ .1634\ . \tag{8}$$

For the smallest values of $p_i$, agreement (for field i) is the rarest - and, plausibly, the corresponding fields i discriminate the most usefully between matches and non-matches.

In practice we will typically have all $R_{ij} > 0$; if not, we cannot proceed further. With

$$p_i = Fy_i + (1-F)u_i \tag{9}$$

$$p_{ij} = Fy_iy_j + (1-F)u_iu_j \tag{10}$$

under independence as in (5) and with $y_i > u_i$ presumed, the measured positivity of $R_{ij}$ is due to a combination of three possible sources:

(a) Departure of F from 0.

(b) Excess of $y_i$ over $u_i$ and of $y_j$ over $u_j$.

(c) Positive dependence between $E_i$ and $E_j$ that the independence model does not reflect (that is, (10) understates the value of $p_{ij}$).

(To see how (a) and (b) contribute to the positivity of $R_{ij}$, one substitutes (9) and (10) into (7) and simplifies:

$$R_{ij} = F(1-F)(y_i-u_i)(y_j-u_j).)$$

In Sec. 3 we will fit a value for F, which we denote by $F^*$, under the assumption of independence. Accompanying this value for F will be a fit of values for $y_i$ and $u_i$, which we denote by $y_i^*$ and $u_i^*$, for i=1,...,K. As the Sec. 3 formulas (11), (12), and (18) will indicate, these fitted values depend heavily on the values $R_{ij}$. In effect, we will partition the positivity of the $R_{ij}$'s into the above three sources (a), (b) and (c), in doing this fit.

The fitted value $F^*$ will typically exceed $F_0$, the (logical) maximum possible F. It would be wrong to assign this excess to

effect (a): departure of F from 0. Thus we will presume that any excess of fitted $F^*$ over maximal $F_0$ is, instead, attributable to effect (c): positive dependence. Accordingly, in preparation for Sec. 4 where (corresponding to the value of $F_0$) we measure positive dependence, at the end of Sec. 3 we replace $y_i^*$ and $u_i^*$ by a pair of values $y_i'$ and $u_i'$ which correspond to $F_0$. We choose $y_i'$ and $u_i'$ so as to hold effect (b) (excess of y over u) constant: we maintain the spacing between $y_i^*$ and $u_i^*$ in the sense that the value of $w_i$ in (6) is preserved.

Then, in Sec. 4 we use $F_0$, $y_i'$ and $u_i'$, and a log-linear model, to fit a set of interaction terms that measure the positive dependence of (c). Having fitted these terms, we adjust the table proportions $P_S$ to compensate for interaction effects. We obtain a new "de-interactionized" table for which, with $F = F_0$, the assumption of independence could more nearly plausibly hold. Accordingly, as in Sec. 5, we fit a new set of values $y_i'$ and $u_i'$, and fit additional interaction effects. As explained there, we continue to repeat these steps alternately as long as we continue to obtain an improved fit. (We always have $y_i^* > u_i^*$ and $y_i' > u_i'$. But if either $u_i^* \leq 0$ or $y_i^* \geq 1$ is obtained, we cannot proceed further, in these iterations.)

Thus in summary, we fit results under the independence assumption (Sec. 3), fit interaction terms and de-interactionize (Sec. 4), and repeat these two steps (Sec. 5) as long as forward progress is obtained. Then, in Sec. 6, we discuss final results and their interpretation; we also consider different possible values for the true F. Sec. 7 discusses how best to form fields from a given data base, and Sec. 8 discusses computation issues.

3. Results under Independence. This section fits $Q_{1S}$ and $Q_{2S}$ for K fields under the independence assumption - and then makes the adjustment needed for Sec. 4. As a start, consider three fields i, j and k. For 3 fields under the independence assumption Fellegi and Sunter solve 7 equations (corresponding to $2^3-1$ cell probabilities $P_S$) in 7 unknowns ($F, y_i, u_i$) to obtain fitted $F^*$ given by:

$$C_{ijk} = (R_{ij}R_{ik}R_{jk})^{1/2}; \tag{11}$$

$$D_{ijk} = p_{ijk} - p_i p_j p_k - p_i R_{jk} - p_j R_{ik} - p_k R_{ij}; \tag{12}$$

X the positive root of the equation

$$C_{ijk}(X^2 - 1) - D_{ijk}X = 0; \tag{13}$$

$$F^* = 1/(X^2 + 1). \tag{14}$$

However, we have K>3 fields to consider. There are $\binom{K}{3}$ choices of 3 fields each yielding X in (13). Accordingly,as a composite value for X we use the positive root of

$$(\sum C_{ijk})(X^2-1) - (\sum D_{ijk})X = 0, \tag{15}$$

with summation over the $\binom{K}{3}$ choices, instead of (13). This step yields $F^*$. Notice that $F^*$ (i.e.,X) is expressible in terms of the ratio D/C in (13), and the ratio $(\sum D_{ijk})/(\sum C_{ijk})$ in (15).

We also define $F_h^*$ based on (15) with summation restricted to triplets (i,j,k) for which h = i, j or k. In other words, $F_i^*$ is a composite F corresponding to field i specifically. In our example we have $F^* = .0503$, and values of $F_i^*$ equal to

$$.1620,\ .0615,\ .0099,\ .0230,\ .0796\ . \tag{16}$$

With $F_0$ = only .00338 we regard these values as evidence of positive field dependence.

Given the value of F, under independence we may show

$$R_{ij} = F(1-F)x_i x_j\ . \tag{17}$$

with

$$y_i = p_i + (1-F)x_i \text{ and } u_i = p_i - Fx_i. \tag{18}$$

Hence we now want to fit the values $x_i$ and $x_j$. In (18) there are $\binom{K}{2}$ equations in K unknowns; for K = 3 Fellegi and Sunter in effect solve the system of 3 equations in 3 unknowns. For K>3, from (17) we may form the following composite equation:

$$\log x_i = Z_i/(K-2) - Z/(K-1)(K-2) - .5 \log F_i^*(1-F_i^*) \quad (19)$$

with $Z_i$ the sum of $\log R_{jk}$ over the K - 1 pairs (j,k) for which j = i or k = i, and Z the sum of $\log R_{jk}$ over all $\binom{K}{2}$ pairs, and with $F_i^*$ substituted for F specifically for field i.

Thus we use composite $x_i$ (>0) based on (19). Then we get, from (18),

$$y_i^* = p_i + (1-F_i^*)x_i \text{ and } u_i^* = p_i - F_i^* x_i. \quad (20)$$

These $y_i^*$ and $u_i^*$, like $F_i^*$, are in essence just averages of Fellegi-Sunter 3-field results for the $\binom{K-1}{2}$ triplets which include i. In our example we get

| i | 1 | 2 | 3 | 4 | 5 | (21) |
|---|---|---|---|---|---|---|
| y* | .7134 | .3760 | .2029 | .2829 | .5033 | |
| u* | .5085 | .1646 | .001195 | .002646 | .1340 . | |

Along with $F^*$ and $F_i^*$ too large, we obtain $y_i^*$ too small (it should be much closer to 1) under the assumption of independence. (Yet it would be especially difficult to get the true value of $y_i^*$ by an external sample of the true matches, because with true $F \leq F_0$, i.e., F less than .00338, these true matches are so scanty.)

We have based the development of (20) on the largest aggregates of agreement fractions: $p_i, p_{ij}, p_{ijk}$. This fact is important because for many fields even $p_i$ is much closer to 0 than to 1, as our example indicated in (7); an aggregate such as $p_{ijkmnh}$ may be so close to 0 as to be quite uninformative. In our example, Table 1 shows that $p_{12345}$ is only .001266.

Thus we have fitted results for K fields under the assumption of independence. We now adjust $y_i^*$ and $u_i^*$ in preparation for the

next section: there we will be using the fraction $F_0$, as a device which permits us to fit interaction terms that reflect positive dependence. Let $y_i'$ and $u_i'$ be given by the 2x2 system of equations

$$y_i'(1-u_i')/u_i'(1-y_i') = y_i^*(1-u_i^*)/u_i^*(1-y_i^*) \tag{22}$$

$$F_0 y_i' + (1-F_0)u_i' = p_i \ . \tag{23}$$

The requirement (22) preserves, up to a constant factor, the values of all the likelihood ratios $U_S$: that is, for two different cells S and T the ratio $U_S/U_T$ is unchanged. (The hierarchical relationship among likelihood ratios is thus maintained.) We thus preserve the values of the weights $w_i$ that appear in the constant-plus-sum relationship of (4) (although the value of the constant is altered). In this sense we maintain the spacing between y and u that was obtained from (16): that is, in this sense we hold effect (b) of Sec. 3 constant. We also ensure $0<u_i'<y_i'<1$. The requirement (23) ensures that, in Sec. 4, the fitted overall proportion of agreement for field i equals the observed $p_i$.

One easily sets up and solves a quadratic equation to obtain $y_i'$ and $u_i'$ based on (22) and (23). In our example we get

| i | 1 | 2 | 3 | 4 | 5 | (24) |
|---|---|---|---|---|---|---|
| y' | .7393 | .3965 | .3133 | .5241 | .5587 | |
| u' | .5410 | .1769 | .002141 | .007349 | .1621 | |
| w | .8781 | 1.1177 | 5.3597 | 5.0021 | 1.8789 . | |

4. Interactions. This section fits parameters to reflect positive dependence among field agreements, while preserving the values of $U_S$ obtained in Sec. 3. Consider the fitted probability

$$P_S' = F_0[\textstyle\prod_{i\in S} y_i'][\prod_{i\notin S}(1-y_i')]+(1-F_0)[\prod_{i\in S} u_i'][\prod_{i\notin S}(1-u_i')]. \tag{25}$$

With $P_S'$ as starting point, consider $G_S$ and $P_S'$ given by

$$\log G_S = \log P'_S + \sum_{i<j} b_{Sij} B_{ij} \tag{26}$$

$$P''_S = G_S/G \text{ with } G = \sum G_S. \tag{27}$$

For $P'_S = 0$ we set $P''_S = 0$. Here, $P''_S$ is an update of $P'_S$ (equal to $P'_S$ if all $B_{ij}$ are 0). The quantities $B_{ij}$ are interaction parameters, to be fitted to each cell as a whole. (Here we are thinking of $y'_i$ and $u'_i$ as fitted "main effects," which in this section are fixed.) For $i<j$, the constant $b_{Sij}$ equals 1 for either (a) $i \varepsilon S$ and $j \varepsilon S$ or (b) $i \notin S$ and $j \notin S$; it equals -1 for either (a) $i \varepsilon S$ and $j \notin S$ or (b) $i \notin S$ and $j \varepsilon S$.

(Along with $P''_S$ we also are fitting $Q''_{1S}$ and $Q''_{2S}$ according to (26) and (27), based on replacement of $P'_S$, in (26) by $Q'_{1S}$ and $Q'_{2S}$ respectively. The scaling in (27) forces $Q''_{1S}$ and $Q''_{2S}$, like $P''_S$, to sum to 1 over S. Because we fit interaction parameters for cells as wholes rather than for matched and unmatched pairs separately, the values of all ratios $U_S$ are preserved.)

Values $B_{ij} = 0$ correspond to independence of events $E_i$ and $E_j$, values $B_{ij} > (<) \ 0$ to positive (negative) dependence. To correspond to the typical pattern of nonnegative dependence, we impose the restrictions $B_{ij} \geq 0$. Subject to these restrictions, we fit $B_{ij}$ so as to maximize the likelihood function $\sum P_S \log P''_S$, i.e., to

$$\text{minimize } f = -\sum P_S \log P''_S \tag{28}$$

as a function of the variables $B_{ij}$. As a measure of closeness of fit of (the minimizing) $P''_S$ to the original $P_S$, we use

$$f^* = -\sum P_S \log P''_S + \sum P_S \log P_S, \tag{29}$$

which is always nonnegative. (Here 0 log 0 is 0.) We would obtain $f^* = 0$ only for a perfect fit: $P''_S = P_S$ for all S. (We thus use the maximum likelihood criterion to measure closeness of fitted table proportions to given proportions. We could just as well use minimum Pearson chi-square, with results not greatly

different, with somewhat more complexity in the required computations.)

It is straightforward to show (Fiacco and McCormick 1968) that in (28) we have a convex programming problem. Suppose that we find a set of values $B_{ij}$ ($\geq 0$) such that we have for $i<j$, always, either (a) $df/dB_{ij} = 0$ or (b) $B_{ij} = 0$ and $df/dB_{ij} > 0$. Then for this set of values, we have obtained a solution to the minimization problem in (28); and this solution is unique. A stopping rule may be based on these conditions (a) and (b); other computation issues are discussed in Sec. 8.

For us, a unique solution will routinely exist if the number of cells with $P_S > 0$ comfortably exceeds the number of parameters (in all, $2K + K(K-1)/2$) to be fitted. This requirement needs to be met, furthermore, for proper identifiability of parameters. We will not single out cases where there is no solution; but we would at least require $K=5$. Sec. 7 further discusses the choice of (fields and a value for) K.

In our example the difference $f^*$ in (29) was .007143. Positive interactions were:

$$B_{23} = .0073,\ B_{34} = .1028,\ B_{15} = .0647,\ B_{25} = .0004,\ B_{45} = .0270. \quad (30)$$

With $B_{ij}$ applied to each cell as a whole in (26) and (27) we can preserve the value of $U_S$ and $w_i$, and thus maintain the relationship in (4) and the hierarchical relationship discussed in Sec. 1. We use only $y_i^*$ and $u_i^*$ as determined in Sec. 3 to determine the likelihood ratio $U_S$ mainly because, in trying to minimize f, we would no longer have a convex problem with its corresponding unique global minimum, if we tried to take into account variation in $U_S$. Accordingly, we use the scheme of Sec. 3 to fit $U_S$.

However, we want to be able to apply this scheme of Sec. 3 realistically, under the assumption of mutual independence. Thus having fitted $B_{ij}$, our next step is, in effect, to remove interaction effects from our original table frequencies $P_S$ - before going back to Sec. 3. Our new set of frequencies will

provide an analogue to a deseasonalized time series from which one has removed seasonal effects. For values $B_{ij} \geq 0$ obtained from the solution to the convex problem, we consider $G_S$ and $P_S^*$ given by

$$\log G_S = \log P_S - \sum_{i<j} b_{Sij} B_{ij} \tag{31}$$

$$P_S^* = G_S/G \text{ with } G = \sum G_S, \tag{32}$$

in analogy with (26) and (27). For $P_S = 0$ we set $P_S^* = 0$. Then, $P_S^*$ is our de-interactionized set of cell frequencies and hopefully tends to fit the model for $P_S'$ in (25) (without any interaction terms).

5. Recomputation. This section considers iterative alternation between the procedures in Sec. 3 and Sec. 4. The just-obtained table $P_S^*$ might behave as though (with $F = F_0$) the assumption of mutual independence in fact held. Using $P_S^*$ (rather than $P_S$), we go back and recompute $y_i'$ and $u_i'$ in (22), at the end of Sec. 3. We also recompute $F^*$ as in (15); if the new $F^*$ (while $> F_0$) is not less than the previous $F^*$, we would regard the latest attempt to fit interaction terms, and remove additional positive dependence, as unsuccessful. We also view it as unsuccessful if we should obtain $y_i^*$ or $u_i^*$ outside the range $\langle 0,1 \rangle$, or $R_{ij} \leq 0$ (in which case we can do little but stop).

If we are successful thus far, we fit a new set of interaction terms as follows. From the new $y_i'$ and $u_i'$ we form a new $P_S'$ as in (25), except that to provide correspondence to $P_S$ we add back the already obtained interaction terms - in the manner of (26) and (27), where $P_S''$ was formed. We then have a new starting point $P_S'$, from which to try to remove additional positive dependence. In other words, we are now at (25), at the start of Sec. 4.

New, additional interaction terms, and new $P_S''$, are then fitted using (26) and (27). We get a new minimal value of f as in (28), and a difference as in (29), measuring discrepancy between $P_S''$ and $P_S$. If the value of (29) is greater than the

previous value, however, we regard the whole attempt to re-fit $y_i'$, $u_i'$ and $P_S''$ as unsuccessful. At this point, if we have been unsuccessful in any sense, we go back to the previous set of $y_i'$ and $u_i'$, and use the accompanying $w_i$ and $P_S''$ as final values.

However, if we are successful, we give it another try. Based on the accumulated sums of interaction terms fitted thus far, we de-interactionize, re-fit $P_S'$, re-compute f (and an additional set of interaction terms), etc. We can keep going until we are unsuccessful, and then revert to results for the previous fit. Thus we consider results based on the best available fit for our 2nd-order model: $y_i'$, $u_i'$ and cumulative $B_{ij}$ based on them.

(Quite possibly, straightforward variants of some of the above-discussed computational steps will be helpful for particular sets of data, in getting $f^*$ as small as possible, subject to the model in (26) and (27). The exact computational steps are thus somewhat open to investigation. However, we have found that some of the more obvious variants do not seem helpful. In any event we would preserve the essential, alternating calculations: in Sec. 3 extend the Fellegi-Sunter method of moments, and in Sec. 4 fit second-order nonnegative interaction terms based on a convex programming problem, so as to maintain the likelihood-ratio relationship in (4).)

In our example we were successful until on the 14th set of computations for $y_i^*$ and $u_i^*$ we obtained an $R_{ij} \leq 0$. Hence we consider the results from the 13th set of calculations. We got 13th $F^* = .0147$, vs. initial $F^* = .0503$, and vs. $F_0 = .00338$; hence most of the discrepancy from $F_0$ has been removed, and most of the positive dependence between fields may have been removed. Values of $F_i^*$ also decreased similarly. The value of the difference $f^*$ in (29) was reduced to only .002440, vs. .007143 after the 1st cycle; hence in a sense our iterations have removed most of the initial discrepancy between observed and fitted cell proportions. Accumulated interactions after the 13th cycle were (cf. initial values in (30))

$$B_{12}= .0351,\ B_{23}= .1111,\ B_{14}= .0644,\ B_{34}= .2556, \qquad (33)$$
$$B_{15}= .1172,\ B_{25}= .0082,\ B_{45}= .2589\ .$$

For the 13th cycle we also obtain (compare to initial values in (24) which correspond to the assumption of independence)

| i | 1 | 2 | 3 | 4 | 5 | (34) |
|---|---|---|---|---|---|---|
| y' | .6372 | .4192 | .5476 | .9626 | .4203 | |
| u' | .6199 | .2104 | .0025 | .0120 | .2344 | |
| w | .0740 | .9966 | 6.1913 | 7.6619 | .8623 . | |

In (34), field 1 (especially) and fields 2 and 5 are de-emphasized, while fields 3 and (especially) 4 are given more emphasis, than in (24). Moreover, we now have $w_4>w_3$ in spite of the facts (a) $p_3<p_4$ in (8), indicating that field 3 might be a more useful discriminator than field 4, and (b) initially $w_3>w_4$.

We would use these results for the data of Table 1. In this example a better fit was prevented only by $R_{ij}$ becoming nonpositive.

Iteration in this example was beneficial, with 13 productive cycles. In other examples there will be fewer (or more) such cycles, perhaps only 1; but even a single productive cycle with fitted interactions is an advance from the independence model of Sec. 3.

6. Final Results. At this point we have a clear choice for $w_i$ and $U_S$: the last set of values for these, for which success was achieved. However, we want also to fit the best legitimately constructed final values for $Q_{1S}$ and $Q_{2S}$. In this section we obtain these values, and then consider posterior probabilities based on them: for each cell S, a conditional probability that (a pair belonging to) the cell represents a true match.

We consider a 2-dimensional table. There will be 2 rows, corresponding to match and non-match, and $2^K$ columns, corresponding to the $2^K$ cells S. As in Sec. 4 we form $Q''_{1S}$ in the manner of $P''_S$ except that instead of log $P'_S$ in (26) we use log $Q'_{1S}$; we form $Q''_{2S}$ similarly. Then we put $F_0Q''_{1S}$ in row 1, column S and $(1-F_0)Q''_{2S}$ in row 2, column S. But we want, along

with row sums equal to $F_0$ and $1-F_0$, to have column sums equal to the observed $P_S$: having completed the fit of interactions, we now make full use of the observed proportions, $P_S$. We use iterative proportional fitting (Ireland and Kullback 1968) or "raking" to produce a new table which is minimally distorted from the original $2x2^K$ table, subject to: (a) row and column sums having the above totals; and (b) in the manner of (22) toward the end of Sec. 3, preservation of the constant-plus-sum relationship in (4) and the values of $w_i$ in this relationship. When the raking is completed, we have table entries of form (row 1) $F_0Q^*_{1S}$ and (row 2) $(1-F_0)Q^*_{2S}$ with $Q^*_{1S}$ and $Q^*_{2S}$, each summing to 1 over S, representing final fitted probabilities.

Let M denote the event "match" and $U^*_S = Q^*_{1S}/Q^*_{2S}$. Given that prior $P(M)$ is $F_0$, and $P(S|M)$ is $Q^*_{1S}$, with $P(S) = P_S$, we have the posterior probability

$$P(M|S) = F_0Q^*_{1S}/P_S = F_0U^*_S/(F_0U^*_S + 1-F_0). \qquad (35)$$

(Note that $P_S$ equals $F_0Q^*_{1S} + (1-F_0)\,Q^*_{2S}$, in this development.) Thus we merely divide the fitted cell entry in row 1 and column S, by the column total $P_S$, to obtain $P(M|S)$. As indicated in Sec. 1, we would make use of these posterior probabilities $P(M|S)$. Note that if $P_S$ is 0 we do not have a meaningful $P(M|S)$, but that an empty cell does not matter anyway. Table 2 gives values of $P(M|S)$, in the fashion of Table 1, for our example. Wherever there is agreement for fields 3 and 4, we have $P(M|S)>.98$; otherwise it is $<.5$. Using a cutoff value of .90 or .95, we would for this example be inclined to view agreement in these two fields as evidence of true match, and disagreement in either of them as lack of evidence of it.

**Table 2. Values of P(M|S)**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 00000 | .000050 | 00010 | .096580 | 00001 | .000119 | 00011 | .202047 |
| 10000 | .000054 | 10010 | .103233 | 10001 | .000128 | 10011 | .214242 |
| 01000 | .000136 | 01010 | .224579 | 01001 | .000323 | 01011 | .406871 |
| 11000 | .000147 | 11010 | .237729 | 11001 | .000347 | 11011 | .424847 |
| 00100 | .023975 | 00110 | empty | 00101 | .054982 | 00111 | .991980 |
| 10100 | .025770 | 10110 | .982527 | 10101 | .058957 | 10111 | .992547 |
| 01100 | .062396 | 01110 | .992981 | 01101 | .136160 | 01111 | .997024 |
| 11100 | .066869 | 11110 | .993478 | 11101 | .145101 | 11111 | .997236 |

The purpose of this paper, however, is more than to reach such a conclusion for our example. As we indicated in Sec. 1, our chief goal is to provide the best possible values for $Q_{1S}$ and $Q_{2S}$. These values may then be put to use in a hypothesis-test framework, for the cost-utility approach of Tepping (1968), or in other ways. Or one could solve a transportation problem based on $P(M|S)$ as in Sec. 1.

As indicated in Sec. 1, the value of $F_0$ in (2) represents an upper bound on the true value of F rather than the true value itself; we would carry out our calculations for smaller values of F, to see how the results change. We have done this for $P(M|S)$; but one can, of course, do similar investigations for the hypothesis-test framework. One can surmise from (35), that use of a smaller value for $F_0$ and thus a smaller prior $P(M)$ should yield smaller values for $P(M|S)$; and in our investigation this was the case.

For this investigation we replaced $F_0$ in (2) by $p_{34}$, the proportion of all pairs with agreement in fields 3 and 4, in view of the Table 2 results which suggested the decision "match" if and only if there is agreement for both these fields. The value of $p_{34}$ is .001731, based on numerator 2287 (vs. 4465 for $F_0$). In the sense of Sec. 5 we were unsuccessful ($F^*$ failed to decrease) on the 7th cycle, so we used 6th-cycle results as given in Table 3. There is still a marked dichotomy between cells S for which fields 3 and 4 agree, and all other cells; but Table 3 provides a reminder that $P(M|S)$ in Table 2 is overstated and that agreement in fields 3 and 4 does not guarantee that we have a true match.

**Table 3. Values of $P(M|S)$**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 00000 | .000016 | 00010 | .003472 | 00001 | .000068 | 00011 | .014637 |
| 10000 | .000022 | 10010 | .004787 | 10001 | .000094 | 10011 | .020095 |
| 01000 | .000051 | 01010 | .011011 | 01001 | .000218 | 01011 | .045314 |
| 11000 | .000071 | 11010 | .015137 | 11001 | .000301 | 11011 | .061495 |
| 00100 | .003115 | 00110 | empty | 00101 | .013145 | 00111 | .743760 |
| 10100 | .004295 | 10110 | .484505 | 10101 | .018056 | 10111 | .800279 |
| 01100 | .009884 | 01110 | .685084 | 01101 | .040823 | 01111 | .902673 |
| 11100 | .013594 | 11110 | .750199 | 11101 | .055494 | 11111 | .927555 |

One might wonder how to choose a "best" value for F that is $\leq F_0$. One could consider the final value of $F^*$ in (29): choose the F that minimizes this objective function. But because our model does not fully permit us to account for all positive dependence (source (c) of Sec. 2), it seems that we will fit a value of F which tends to be too large (source (a) of Sec. 2) and are apt to overstate the value of F. Moreover, just as for fitting best values for ratios $U_S$, we encounter difficulty because we no longer have a convex programming problem. (Whereas for $F = F_0$ we obtained, as in Sec. 5, a final $f^* = .002440$, we get $f^* = .004220$ for $F = p_{34}$. That is, $F_0$ corresponds to a better fit than does $p_{34}$.)

A final note: with $E_i$ denoting the event "agreement for field i," we may sum over $Q^*_{1S}$ for $i \in S$, to get final fitted values of $P(E_i|M)$ for i=1, ..., K. In our example these are

$$.8168,\ .6491,\ .5342,\ .9474,\ .8037. \tag{36}$$

These are not as close to 1 as they realistically ought to be - at least partially because the maximal $F_0$ exceeds the true fraction of matches. When we replace $F_0$ by $p_{34}$ as above the fitted values of $P(E_i|M)$ increase to

$$.8719,\ .8928,\ .9098,\ 9752,\ 9217. \tag{36a}$$

7. Choice of Fields. In this section we consider how best to form fields. From Sec. 1, recall that field 1 in our example is based on 2 data items (race and ethnicity). Likewise fields 2, 3 and 4 are each based on 2 items, while field 5 is based on 3 items. In all, we have 11 "items" although we have been only using 5 fields. One might ask, why not form and use 11 fields? (Moreover, by using binary digits for each item, we could form many more than 11 fields.)

Accordingly, we now consider the general issue of how best to form fields from a given set of data items. As a first principle, we would include every item in one (and only one) data

field, in order to make full use of available information concerning agreement/disagreement patterns. But for the following reasons we would not necessarily form a field from each single item. (That is, we would not necessarily form 11 fields in our example.)

(a) As the number of fields (K) increases, there is increasing room for complicated interactions among the fields. Departures from our second-order model in Sec. 4 may be exacerbated.

(b) By increasing K, we do increase the number of fitted parameters $w_i$, and $B_{ij}$. However, we do not necessarily get more information. For example, if in our illustration we split race-ethnicity into two separate fields, we get two separate weights $w_i$, and a suggestion of more information. However, we do not get a single weight for the combination race-ethnicity, that reflects the joint behavior of the two items. Nor do we get a pairwise interaction, $B_{ij}$, which corresponds to this combination, as a single field, in conjunction with a second field such as birthdate.

(c) (less vital) As in the next section, the required order of computations blows up as K increases.

We would prefer K=5, provided that $P_S>0$ for all or nearly all of the 32 cells S, in order to permit comfortable fitting of 20 parameters as in Sec. 4. There are, however, two further possible objections to small K. We state and deal with these as follows.

(a) For K (and $2^K$, the number of cells) small, there can be awkwardness, because of discreteness, in constructing a decision rule analogous to the classical hypothesis test (Sec. 1). Cells are few in number with large probabilities $Q_{1S}$ and $Q_{2S}$ associated with them; as a result one typically cannot come close to attaining the desired Type I and Type II error probabilities. But as Fellegi and Sunter indicate, one can introduce an artifical randomizing device to attain these. Moreover, in Sec. 1 we indicated how to use posterior probabilities either to solve

a transportation problem or to construct decision rules based on cutoffs instead of the hypothesis-test framework. Hence this objection can be overcome.

(b) The proportions of agreement, $p_i$ (as in (8)), are smallest for the smallest K and the most aggregation of items. Small values of $p_i$ can lead to difficulties when the denominator ("D") of $p_i$ is small: that is, when $\sum N_{hA} N_{hB}$ in Sec. 1 is small. For this reason one might thus anticipate difficulty for small K. However, the rest of this section indicates that such difficulty can be avoided for K as small as 5.

For D = 10 million and $p_i$ = .0002 we have a numerator 2000. If numerators of $p_i$ are, in this fashion, in the thousands, then we have relatively small sampling error and may usefully compare $p_i$ to each other - and likewise utilize $p_{ij}$ and $p_{ijk}$ as we have done for our example. On the other hand, suppose D is 10,000; for $p_i$ = .0002 we are dealing with a numerator 2. Sampling error, extreme discreteness, etc., creates an intractable situation. For sufficiently large D, with numerators large enough, there should be no objection to having a small K and small values of $p_i$. (For given files A and B, the numerators will increase as D increases, although they will not proportionately keep pace - because the overall number of true matches, which relatively account for more agreements than nonmatches, stays the same while the overall number of nonmatches increases. In any event, it seems best to choose fields so that the $p_i$'s are equalized as nearly as possible, as a "maximin" strategy.)

Thus we next must consider how big a D we want to deal with, and thus how many blocks we want to form: the fewer the blocks, the bigger D will be, for given files A and B. For K fields we need to make DK field comparisons as to field agreement/disagreement, in order to obtain the quantities $P_S$. These comparisons can be streamlined, and based on just integers. Thus computationally it becomes quite feasible to consider D equal to, say, 10 million (at least) in conjunction with K=5, in striving to make the numerators for $p_i$, $p_{ij}$, $p_{ijk}$ acceptably large.

Another reason for forming a few large blocks, as opposed to many small blocks, is that likelihood of error is reduced, in the following sense. Recall from Sec. 1 that blocking is to be done so that: if a pair of records from A and B constitute a true match, they belong to the same block. Blocking leads to "error" if it assigns two such records to different blocks. By having blocks as few in number and as large as possible, we reduce the possibilities for such error.

8. Computation. This section pertains strictly to computational issues for the convex programming problem of Sec. 4. For this general problem, of course, numerous algorithms have been developed; but we would like to add the following.

We start each (convex programming) problem with $B_{ij}$, corresponding to additional positive interaction, set to 0. To compute $\nabla f$, the vector of 1st derivatives, and $\nabla^2 f$, the matrix of 2nd derivatives, for f in (28) with respect to $B_{ij}$, we need, in the notation for (26), the sums $\sum_S b_{Sij} G_S$. We also need, for $i \neq j \neq k \neq i$, analogous sums $\sum_S c_{Sijk} G_S$ which correspond to 3rd-order interaction; likewise we need sums which correspond to 4th-order interaction. In all, the number of required additions involving $G_S$ is seen to be of order

$$2^{K-1} \sum_{j=2}^{4} \binom{K}{j}, \tag{37}$$

which increases dramatically as K increases: important because of the need to compute these derivatives repeatedly. Thus, as indicated in Sec. 7, there can be computational problems for large K.

A final computational aid for persons familiar with convex programming algorithms: suppose that the direction vector,

$$(\nabla^2 f)^{-1} (\nabla f), \tag{38}$$

has a positive component corresponding to a $B_{ij}$ which is 0. Minimizing f along this direction vector, we would be led to make $B_{ij}$ negative, contrary to the constraints $B_{ij} \geq 0$. Under these

conditions we have recomputed the direction vector with such $B_{ij}$ excluded (temporarily) from the set of variables. If necessary, we repeat to make further exclusions. This approach permits us neatly to bound $B_{ij} \geq 0$ without the use of penalty functions; and we find that we quickly get to the solution. For a similar reason, when there are positive components for $B_{ij} > 0$, we multiply by a positive scalar small enough to keep all $B_{ij} \geq 0$.

## REFERENCES

Coulter, R.W. (1985), "An Application of a Theory for Record Linkage," Record Linkage Techniques-1985, Internal Revenue Service, 89-96.

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

Fiacco, A.V., and McCormick, G.P. (1968), Nonlinear Programming: Sequential Unconstrained Minimization Techniques, John Wiley, New York.

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, Academic Press.

Ireland, C.T., and Kullback, S. (1968), "Contingency Tables with Given Marginals," Biometrika, 55, 179-188.

Kelley, R.P. (1985), "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," Record Linkage Techniques - 1985, Internal Revenue Service, 199-203.

Tepping, B.J. (1968), "A Model for Optimum Linkage of Records," Journal of the American Statistical Association, 63, 1321-1332.

Winkler, W.E. (1985), "Exact Matching Lists of Businesses: Blocking, Subfield Identification, and Information Theory," Record Linkage Techniques - 1985, Internal Revenue Service, 227-241.