ADVANCES IN RECORD LINKAGE METHODOLOGY: A METHOD FOR DETERMINING THE BEST BLOCKING STRATEGY

R. Patrick Kelley, Bureau of the Census

I. INTRODUCTION

The term record linkage, as it will be used in this paper, is a generic term for any process by which the set of reporting units common to two or more files of data is determined.

Historically, government agencies have been the primary users of record linkage techniques. The reasons such agencies carry out record linkage projects are as varied as the purpose and scope of the agencies themselves. Consider the following examples:

a) The United States Department of Agriculture uses record linkage to update mailing lists (see Coulter and Mergerson, 1977).
b) Statistics Canada uses record linkage as a tool in epidemological research (see Smith, 1982).
c) The United States Census Bureau uses record linkage as a tool in coverage and content evaluation (see Bailar, 1983).

For a more detailed discussion of the history and and use of record linkage by United States government agencies see U.S. Department of Commerce (1980).

As an area of study, Record Linkage, with its associated statistical problems, is a special case of a larger area of concern. This area makes use of various mathematical and statistical techniques to study the problems involved in the classification of observed phenomena.

Discriminant analysis, discrete discriminant analysis, pattern recognition, cluster analysis and mathematical taxonomy are some of the specific fields which study various aspects of the classification problem. While record linkage contains its own specific set of problems it also has a great deal in common with these other fields.

The basic unit of study in the linking of two files F1 and F2 is F1XF2, the set of ordered pairs from F1 and F2. Given F1XF2, our job is to classify each pair as either matched or unmatched. This decision will be based on measurements taken on the record pairs. For example, if we are linking person records, a possible measurement would be to compare surnames on the two records, and assign the value 1 for those pairs where there is agreement and 0 for those pairs where there is disagreement. These measurements will yield a vector, Γ, of observations on each record pair.

The key fact which will allow us to link the two files is that Γ behaves differently for matched and unmatched pairs. Statistically we model this by assuming that Γ is a random vector generated by $P(\cdot \mid M)$ on matched pairs and $P(\cdot \mid U)$ on unmatched pairs. Thus, the Γ value for a single randomly selected record pair is generated by $pP(\cdot \mid M)+(1-p)\,P(\cdot \mid U)$ where p is the proportion of matched records.

This model for the record linkage problem is the same as the one used in discriminant analysis. In particular, as Γ is almost always discrete, the literature on discrete discriminant analysis is extremely useful (see for example Goldstein and Dillon, 1978). There are, however, several areas of concern that seem to be a great deal more important for record linkage than for the other classification techniques.

Our topic of discussion in this paper, blocking, arises from consideration of one of these problem areas. That area concerns the extreme size of the data sets involved for even a relatively small record linkage project. The size problem precludes our being able to study all possible record pairs. So, we must determine some rule which will automatically remove a large portion of record pairs from consideration. Such a rule is referred to as a blocking scheme since the resulting subset of record pairs often forms rectangular blocks in F1XF2.

The literature on the blocking problem is not extensive. Brounstein (1969), Coulter and Mergerson (1977) and U.S. Department of Commerce (1977) contain discussions of the practical aspects of choosing a blocking scheme; however, they provide no general framework within which to make such a selection. Jaro (1972) provides a framework for the selection of a blocking scheme but doesn't discuss the errors induced by blocking. Many other papers, particularly those on clerical matching, contain implicit information on blocking. But so far there has been no systematic study of this area.

To provide such a study we begin with the following three questions:

1) What are the benefits and costs involved in blocking and how do we measure them?
2) How do we select between competing blocking schemes? Is there a best scheme?
3) How do the various computing restrictions effect our blocking scheme selection?

These three questions will serve as a guideline for our investigation of the blocking problem. But, before we begin this investigation, we need to consider some background material on record linkage.

II. BACKGROUND

Again, our job in linking the two files F1 and F2 is to classify each record pair as either matched or unmatched. In practice, however, we usually include a clerical review decision for tricky cases. So, our set of possible decisions is

A1: the pair is a match
A2: no determination made - clerical review
A3: the pair is not a match.

Now, consider the class of decision functions $D(\cdot)$ which transform our space of comparison vector values, elements of which we will denote by γ, to the set of decisions $\{A1,A2,A3\}$. Given

two or more decision functions in this class, what criterion will we use to choose between them?

In Fellegi and Sunter (1969) the argument is put forward that, as decision A2 will require costly and error prone clerical review, we should pick a decision procedure which will minimize the expected number of A2 decisions while keeping a bound on the expected number of pairs which are classified in error. Since the unconditional distribution of the comparison vector is the same for any randomly chosen pair, this reduces to picking that decision procedure which will minimize P(A2) subject to $P(A1|U) <= \mu$ and $P(A3|M) <= \lambda$.

Given that you know $P(\cdot\,|M)$ and $P(\cdot\,|U)$, Fellegi and Sunter prove that the decision procedure which solves this problem is of the form

$$(1)\quad D(\gamma) = \begin{cases} A3 \text{ if } \ell(\gamma) <= t1 \\ A2 \text{ if } t1 < \ell(\gamma) < t2 \\ A1 \text{ if } \ell(\gamma) >= t2 \end{cases}$$

where $\ell(\gamma) = P(\gamma\,|M)/P(\gamma\,|U)$, t1 is the largest value in the range of $\ell(\cdot)$ for which $P(A3|M) <= \lambda$ and t2 is the smallest value in the range of $\ell(\cdot)$ for which $P(A1/U) <= \mu$.

It is this decision procedure that forms the basis for our study of the blocking problem.

III. MEASUREMENT OF THE COST AND BENEFIT OF BLOCKING

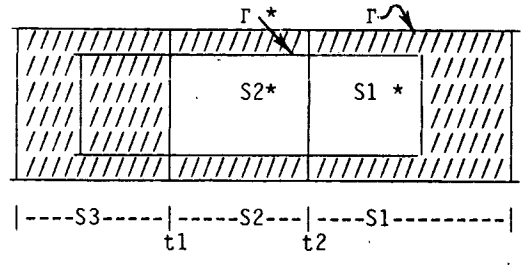
In the past sections we have outlined the more general aspects of record linkage and defined the blocking problem. In this section we will discuss blocking in the context of the decision procedure given in section II.

We base our general blocking strategy on the fact that the proportion of matched pairs in F1XF2 is small. So we will concentrate on blocking rules in which the pairs removed by the rule will be assigned the status of unmatched.

Fellegi-Sunter (1969) provides a formal model for blocking. This model defines a blocking scheme to be a subspace, say Γ^*, of the comparison space. Kelley (1984) provides a preliminary study of selected methods of measuring cost and benefit. The method found to have the most intuitive appeal is one that is based on the following amended decision procedure:

$$(2)\quad D'(\gamma) = \begin{cases} A3 \text{ if } \ell(\gamma) <= t1 \text{ or } \gamma \in \Gamma^{*c} \\ A2 \text{ if } t1 < \ell(\gamma) < t2 \text{ and } \gamma \in \Gamma^* \\ A1 \text{ if } \ell(\gamma) >= t2 \text{ and } \gamma \in \Gamma^* \end{cases}$$

A Venn diagram of this situation is given by



where S3* is represented by the shaded region.

In this design Si and Si* are the regions of Γ values for which we make decision Ai under decision functions given by (1) and (2), respectively.

The error levels for this amended decision rule are given by

$$P(S3^* \mid M) = P(S3 \mid M) + P(S3^* - S3 \mid M)$$

$$= \lambda + P(S3^* - S3 \mid M).$$

and

$$P(S1^* \mid U) = P(S1 \mid U) - P(S1 \cap S3^* \mid U)$$

$$= \mu - P(S1 \cap S3^* \mid U).$$

These equations give us a means to compute a cost incurred by blocking on the subspace Γ^*, namely, $P(S3^* - S3 \mid M)$, the increase in probability of a false nonmatch. The benefit gained from blocking on Γ^* takes the form of a decrease in the number of pairs which will have to be processed. We will measure this benefit by the unconditional probability that a randomly chosen record pair yields a Γ vector in the block.

Now, given two blocking schemes which both have cost less than or equal to a fixed amount, the preferred scheme is the one with greatest benefit. Thus, we define the best blocking scheme to be that scheme which minimizes $P(\Gamma^*)$ subject to $P(S3^*-S3|M) <= w$, where w is an independently determined upper bound on blocking costs.

IV. COMPUTING THE BEST BLOCKING SCHEME - THE ADMISSIBILITY CONCEPT

Since the comparison vector is discrete, the computation of the best blocking scheme will require a comparison of all competing schemes. So, it's in our best interest to reduce the number of competing schemes. To make this reduction we note that if $\Gamma1^*$ and $\Gamma2^*$ are two competing schemes such that $\Gamma1^*$ is a subset of $\Gamma2^*$ then $\Gamma1^*$ is uniformly better than $\Gamma2^*$. So, we can remove $\Gamma2^*$ from the set of competing blocking schemes. The following definition formalizes this example:

> Γ^* will be said to be an admissible blocking scheme at w = w0 if
> a) $P(S3^* - S3 \mid M) <= w0$ and
> b) for every Γ^{**} that is a subset of Γ^* $P(S3^{**} - S3 \mid M) > w0$.

The concept of an admissible blocking scheme given by this definition is analogous to the concept of an admissible decision procedure. It serves to reduce, hopefully to a reasonable size, the number of blocking schemes competing for best. But, unfortunately, when actually applied to the task of computing the set of admissible blocking schemes, this definition is very cumbersome. The following lemma gives necessary and sufficient conditions for admissibility which are more favorable to algorithm development:

Lemma 1:

Γ^* is admissible at $w = w0$ if and only if $\Gamma^* \cap S3 = \emptyset$ and
$P(\gamma|M) > w0 - P(S3^*-S3|M) \geq 0$ for all γ in Γ^*.

Proof:

If Γ^* is admissible then $P(S3^*-S3|M) <= w0$. Further, for $\Gamma^{**}= \Gamma^* - \{\gamma\}$ we have $P(S3^{**} - S3|M) > w0$. But $S3^{**} - S3 = (S3^* - S3) \cup (\{\gamma\}-S3)$. So, $P(\{\gamma\}-S3|M) + P(S3^* - S3|M) > w0$.

From this relationship we see that if γ is in S3 then $P(S3^*-S3|M) > w0$; thus, $\Gamma^* \cap S3 = \emptyset$. So we have $P(\gamma|M) > w0 - P(S3^*- S3|M)$ for all γ in Γ^*.

Conversely, we first note that $P(S3^*-S3|M) <= w0$. Next, let Γ' be a proper subset of Γ^* then Γ' is a subset of $\Gamma^*- \{\gamma\}$ for some γ. So, $P(S3'-S3|M) >= P(S3^*-S3|M) + P(\{\gamma\}-S3|M)$. Thus, we have $P(S3'-S3|M) >= P(S3^*-S3|M) + P(\gamma|M) > w0$. Hence, Γ^* is admissible.

Now, in theory, we can use the result of lemma 1 to compute all admissible schemes. However, since the minimum number of dimensional Γ vector values is 2**n, we would have to generate and classify on the order of 2**(2**n) subsets.

For n=5 this yields 4,294,967,300 subsets, which is clearly too large for practical consideration. So, while the admissibility concept is helpful in reducing the number of competing schemes, it hasn't served to provide us with a practical algorithm for the computation of the best blocking scheme. In the next section, we will give more attention to the development of such an algorithm.

V. IMPLEMENTATION CONSIDERATIONS

The previous section provides a general framework for studying blocking; however, it doesn't give us much insight into the practical side of determining a block of records for possible linkage. If we keep in mind that I/O and computing the comparison vector are the biggest consumers of time in the linkage operation we see that admissible blocking schemes that require the computation of a Γ vector value for each record pair are not practical. Thus, though a scheme might be theoretically admissible it might not be feasible.

One solution for this problem is to block by using certain fields on the record (such as soundex code of surname or address range) as sort keys. The blocks would be determined by those record pairs with equal keys. Thus, the match status of unmatched pairs would be implicitly assigned to all record pairs with unequal keys.

Restricting our study to blocking schemes which are determined by sort keys implies that the comparison vector we want to use will consist of dichotomous components measuring agreement on the record identifier fields. We will further assume that the components of the comparison vector are stochastically independent for both matched and unmatched record pairs.

Now, letting $m_i = P(\Gamma_i=1|M)$, $u_i=P(\Gamma_i=1|U)$ and Γ^* be the blocking scheme determined by sorting on components $i1,\ldots,ik$ we have the following result:

Lemma 2:

Suppose that $m_i>1/2$ and $u_i<m_i$ for all i then Γ^* is admissible at w0 if and only if

a) $w0 - P(S3^*-S3|M) >= 0$

b) $P(\gamma^*|M) > \max\{t1P(\gamma^*|U), w0 - P(S3^*-S3|M)\}$,

where γ^* is such that $\gamma i1^* = 1,\ldots, \gamma ik^* = 1$, $\gamma ik+1^* = 0, \ldots, \gamma ip^* = 0$.

Proof:

First suppose that Γ^* is admissible at w0 then conditions a) and b) follow directly from lemma 1 and the fact that $P(\gamma|M) > t1\, P(\gamma|U)$ for all γ in $S3^c$.

Now, to establish the converse we first note that, since $m_i > 1/2$ for all i, $P(\gamma^*|M) = \min_{\gamma\in\Gamma^*} P(\gamma|M)$. So $P(\gamma|M) > w0 - P(S3^*-S3|M) >= 0$ for all γ in Γ^*. Next we need to prove that $\Gamma^* \cap S3 = \emptyset$. To prove this we note that $u_i < m_i$ implies that $m_i/u_i > (1-m_i)/(1-u_i)$. So, $P(\gamma|M)/P(\gamma|U) > P(\gamma^*|M)/P(\gamma^*|U)$ for all γ in Γ^*. Thus, $\Gamma^* \cap S3 = \emptyset$. The converse follows from lemma 1.

In comparing lemma 2 with lemma 1, we see that lemma 2 has a definite computational advantage above and beyond the reduction in competing schemes gained by restricting attention to those schemes based on sorting. That advantage lies in the requirement to check for admissibility at only one point in the blocking scheme, namely γ^*. This results in tremendous savings in computing time and simplifies algorithm construction and coding considerably. In the next section we apply lemma 2 to a simple numeric example.

VI. AN EXAMPLE

As an example, let's consider matching two files of records based on the identifiers surname, first name, and sex.

Suppose we have determined beforehand that,

for surname m1 = .90 and u1 = .05,
for first name m2 = .85 and u2 = .10,
and for sex m3 = .95 and u3 = .45.

Retaining the assumption of the previous section our discriminant function is given by

$$L(\gamma)= \ln2(l(\gamma)) = \sum_{i=1}^{3} [\gamma i \ln2 (m_i/u_i) + (1-\gamma i) \ln2 ((1-m_i)/(1-u_i))].$$

To compute the Fellegi-Sunter decision procedure we first compute L for each agreement pattern and then we order the patterns on increasing L. The following table gives the results of this operation:

Pattern	Sum of $P(\cdot\|M)$	One minus sum of $P(\cdot\|U)$	L
(0,0,0)	.00075	.52975	-9.29
(0,0,1)	.01500	.14500	-4.76
(0,1,0)	.01925	.09275	-3.62
(1,0,0)	.02600	.06800	-1.87
(0,1,1)	.10675	.02525	.92
(1,0,1)	.23500	.00500	2.67
(1,1,0)	.27325	.00225	3.79
(1,1,1)	1.00000	0.00000	8.34

Using this table it is clear how one would compute t1 and t2 for given λ and μ.

For example, if we let λ = .05 and μ = .05 then t1 = -1.87 and t2 = 2.67. The actual values of λ and μ are .026 and .02525, respectively. We will use this decision procedure to discuss the blocking problem.

Consider our space of admissible blocking schemes based on sorting. We note that since no single component blocking scheme is admissible, we have a total of four schemes to test. Now, for convenience let B1 denote blocking on surname and first name, B2 denote blocking on surname and sex, B3 denote blocking on first name and sex, and B4 denote blocking on all components.

The following table gives the information necessary to determine the admissibility of Bi:

Bi	P(S3*-S3|M)	P(γ*|M)	values of w0 for which Bi is admissible
B1	.209	.03825	.209 $\leq$ w0 < .24725
B2	.119	.12825	.119 $\leq$ w0 < .24725
B3	.1665	.08075	.1665 $\leq$ w0 < .24725
B4	.24725	.72675	.24725 $\leq$ w0 < .974

Before we go on it is interesting to note that the minimum w0 value for which any of the B_i is admissible is .119. Thus, the minimum loss we can incur by blocking is an increase in false non-match probability of .119.

Looking at the admissible blocking schemes as a function of w0, we have the following:

1. For .119 $\leq$ w0 < .1665 B2 is admissible.
2. For .1665 $\leq$ w0 < .209 B2 and B3 are admissible.
3. For .209 $\leq$ w0 < .24725 B1, B2, B3 are admissible.
4. For .24725 $\leq$ w0 < .974 B4 is admissible.

Now, to compute the best admissible blocking scheme we must determine which of the competing schemes has the smallest probability of occurrence. The probability of occurrence of schemes Bi, say P(Bi), is given by pP(Bi|M)+(1-p)P(Bi|u), where p is the proportion of matched record pairs. Thus, in general, the best admissible scheme will be a function of p.

To compute the best blocking scheme for cases 2 and 3 consider the following table:

	P(Bi|M)	P(Bi|U)
B1	.765	.005
B2	.855	.0225
B3	.8075	.045

So, for case 2, B2 is the best blocking scheme for values of p <= .3214 and B3 is the best blocking scheme for p > .3214. For case 3, B1 is uniformly the best blocking scheme.

At this point, we have demonstrated how to select the best blocking scheme for a fixed value of w0. But it still is unclear how one would use this information to actually make a decision about which scheme to use. To study this question let's consider the nature of such a decision. To select a blocking scheme we need to balance the cost with the overall benefit. Let's redo our example this time for several different values of w0 and compare the benefits for the resulting schemes.

The following is the first part of the list of the best blocking schemes for all values of w0. This list is presented in increasing order of w0. The expected benefit, in terms of the percent of F1XF2 that would be examined, is given for each scheme. To compute this benefit the approximate sizes of F1 and F2 are required. We used F1 size = 200,000 and F2 size = 100,000 in this example.

1. Admissible blocking schemes at w0=0.0492501 are as follows:
   The scheme determined by sorting on sex.
   The expected percent of the cross product of this blocking scheme would examine is bounded above by 45.00005%.
2. Admissible blocking schemes at w0=0.0992500 are as follows:
   The scheme determined by sorting on surname.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 5.00009%.
3. Admissible blocking schemes at w0=0.1442501 are as follows:
   The scheme determined by sorting on surname and sex.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 2.25008%.
4. Admissible blocking schemes at w0=0.149250 are as follows:
   The scheme determined by sorting on first name.
   The scheme determined by sorting on surname and sex.
   Of these, the best blocking strategy, as a function of the proportion of matched pairs, is as follows:

   For p=0.000000000 to p=0.939394700 sort on components surname and sex.
   For p=0.939394700 to p=1.000000000 sort on components first name.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 2.25008%.

To use this list for decision-making purposes one would have to have some idea about how much data they can afford to look at and how large a false non-match rate they could tolerate. For example, in looking at the scheme determined by sorting on sex, we have a small (though maybe not small enough) w0 value but the number of record pairs we would have to look at would be around 9x10**10, which is clearly not feasible. Sorting on surname has a slightly higher w0 value, but reduces the number of records to 10**10. If we are willing to accept an even higher w0, then we can sort on surname and sex, which further reduces the number of record pairs to 4.5x10**9.

Another important piece of information that we shouldn't overlook is the number of record pairs we can hold in memory at any one time. We don't want to select a blocking scheme for which the individual block sizes are too large. So not only is the total number of pairs in the block important but so is the number of states of the sorting variable and the distribution of that

variable over those states.

VII. SUMMARY

The blocking problem is intrinsic to record linkage. As such, before a link between files is attempted a decision must be made concerning the appropriate blocking method.

In this paper we study this decision, along with its costs and benefits, through the record linkage methodology developed in Fellegi and Sunter (1969). This methodology applies classic decision theory techniques to the record linkage problem, constructing the optimum classifer under a loss function analogous to that of hypothesis testing.

The result of our study is a method which can be used to balance the cost and benefit of blocking. This method involves maximizing benefit subject to an upper bound on cost. The measurement of cost and benefit is based on the Fellegi-Sunter method and, as such, makes use of a similar loss function.

NOTES AND REFERENCES

Bailar, Barbara A. (1983), Counting or Estimation in a Census -- A Difficult Decision, Proceedings of the American Statistical Association, Social Statistics Section, pp. 42-49.

Brounstein, S. H. (1969), Data Record Linkage Under Conditions of Uncertainty, delivered at the Seventh Annual Conference of the Urban and Regional Information Systems Association.

Coulter, Richard W. and Mergerson, James, W. (1977), An Application of a Record Linkage Theory in Constructing a List Sampling Frame. List Sampling Frame Section, Sample Survey Research Branch, Statistical Reporting Service, U.S. Department of Agriculture.

Fellegi, Ivan and Sunter, Alan (1969), A Theory for Record Linkage, Journal of the American Statistical Association, vol. 64, pp. 1183-1210.

Goldstein, Matthew and Dillon, William (1978), Discrete Discriminant Analysis, Wiley.

Jaro, M. A. (1972), UNIMATCH - A Computer System for Generalized Record Linkage Under Conditions of Uncertainty, AFIPS - Conference Proceedings, vol. 40, pp. 523-530.

Kelley, Robert Patrick (1984), Blocking Consideration for Record Linkage Under Conditions of Uncertainty, Statistics of Income and Related Administrative Record Research: 1984, Department of the Treasury, Internal Revenue Service, pp. 163-165.

Smith, Martha E. (1982), Value of Record Linkage Studies in Identifying Population at Genetic Risk and Relating Risk to Exposures. Progress in Mutation Research, vol. 3, pp. 85-98.

U. S. Department of Commerce, National Bureau of Standards (1977), Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers.

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards (1980), Statistical Policy Working Paper 5 - Report on Exact and Statistical Matching Techniques.

DISCUSSION

Eli S. Marks, Consultant

WINKLER

This paper discusses Bill Winkler's presentation on "Preprocessing of Lists and String Comparison."

Key factors in "Preprocessing of Lists" are:

1. The objectives of the system and the costs of various levels and types of matching error.
2. Costs of attaining a given matching accuracy level by preprocessing vs. other alternatives (e.g., suitably tailored "tolerances").
3. The nature of the matching system--manual, computerized, "mixed," etc.
4. How preprocessing is performed.

1. Objectives

The objectives of the system and the costs of matching error are intimately related. For example, if the objective is to estimate undercoverage of the U.S. census in each state, city, county, township, place, etc. for purposes of allocation of representation in Congress and state legislatures, city/county councils, etc. and for allocating federal and state funds to state and local jurisdictions, a uniform level of matching error everywhere is more important than the absolute level of matching error. Thus, preprocessing may have little value if its effect is to reduce the different types of matching errors by the same percentages in all jurisdictions. On the other hand, if preprocessing reduces urban matching error more than rural, it may be desirable or undesirable, depending upon whether the level of urban matching error without preprocessing is greater or less than the level of rural matching error without preprocessing.

2. Alternative Techniques

The objective of preprocessing (i.e., reduction of matching errors) can be attained by other means (e.g., the prescription of matching "tolerances"); and these techniques may cost less than preprocessing. For example, soundex coding is a form of "matching tolerance." That is, all disagreements of vowels and some disagreements of consonants are ignored in determining whether a pair of records match on the soundexed "identifier." One can, in fact, combine some preprocessing with tolerances (and, perhaps, other error-reducing techniques) to get a more efficient matching system than either can give alone. For example, one can prescribe standard abbreviations for the address suffixes "Avenue," "Street," "Road," "Drive," "Place," "Boulevard," etc., but also provide that an address match where the suffixes differ will be accepted unless there is another address match where the suffixes agree. For example, "Sutton Drive" would match "Sutton Road" unless either file contains both "Sutton Road" and "Sutton Drive."

Standard spelling of name and address may be achieved more accurately and more cheaply by controlling data collection, recording and "keying" (to put the data in machine readable form) than by preprocessing. This would, for example, avoid most of the errors of preprocessing by ZIPSTAN exhibited by the examples shown in the paper. Preprocessing errors can also be reduced or eliminated by other means, such as the clerical insertion of distinctive symbols to designate components of name and address, as outlined in Section 4 below.

It should be noted that selection of an "optimum matching strategy" is heavily dependent upon the type(s) of matching system(s) considered and that the choice of type of matching system is a vital part of the determination of "optimum matching strategy."

3. Kind of Matching System

The paper by Winkler notes that matching systems can be manual or computerized and implies that preprocessing is largely unnecessary for manual matching systems. I think his suggestion that individuals can usually determine accurately whether a pair of name and address records is actually a match or nonmatch is somewhat optimistic. Individuals can make this determination (so can a computer system), but how accurately depends on the kind of system. The great advantage of a competent human matcher operating in a properly designed matching system is the use of judgmental flexibility, provided, of course, he or she has good judgment and the matching rules permit him (her) to use that judgment (and I have seen many sets of matching instructions which do not). The great disadvantage of a well-designed manual matching system with competent matchers is the human matcher's slowness and the inevitable drop in efficiency in operating in a system which requires examining large masses of records; and not in lack of clear decision rules, inconsistency of application of decision rules, and nonreproducibility of results. All of the latter do occur, but can be adequately controlled in a well-designed matching system (although it is not easy!). However, humans cannot match the forte of the computer--its speed in examining large masses of data.

The solution to this problem is to let the computer do what it does well and let humans do what they do well. That is, design a mixed computer-human system, in which the computer handles the large mass of cases which can be classified as positive links or positive nonlinks, on a mechanical, routine basis. Carefully trained and well-motivated humans could then try to match the remaining cases,