A MODEL FOR OPTIMUM LINKAGE OF RECORDS*

BENJAMIN J. TEPPING
Bureau of the Census

A model is presented for the frequently recurring problem of linking records from two lists. The criterion for an optimum decision rule is taken to be the minimization of the expected total costs associated with the various actions that may be taken for each pair of records that may be compared. A procedure is described for estimating parameters of the model and for successively improving the decision rule. Illustrative results for an application to a file maintenance problem are given.

1. INTRODUCTION

THE problem of record linkage arises in many contexts. A typical example is that of file maintenance. In this example there is a file, which we shall call the master file, whose constitution is to be changed from time to time, by adding or deleting records or by altering specific records. Notice of these required changes is given by means of another file of records, which we shall call the transaction file. Presumably, each transaction record specifies the addition of a new master file record, or the deletion of an existing master file record, or the alteration of an existing master file record. It may not be known whether there exists a master file record that corresponds to a given transaction record so that the determination of whether a master file record is to be changed or a new master file record added must wait until it is found whether a corresponding master file record exists. Thus, the fundamental problem is to determine, for each transaction record, which master file record corresponds to it or that no master file record corresponds to it.

If each master file record and each transaction record carried a unique and error-free identification code, the problem would reduce to one of finding an optimum search sequence that would minimize the total number of comparisons. In most cases encountered in practice, the identification of the record is neither unique nor error-free. Thus it becomes necessary to make a decision as to whether or not a given transaction record ought to be treated as though it corresponded to a given master file record. The evidence presented by the identification codes of the two records in question may possibly be quite clear that the records correspond or that they do not correspond. On the other hand, the evidence may not clearly point to one or the other of these two decisions. Thus it may be reasonable to treat the records temporarily as if they corresponded or to treat them temporarily as if they did not correspond, but to seek further information. Or it may be reasonable in a particular case to take no overt action until further information has been obtained. The amount of effort that it is reasonable to expend in resolving a particular problem is also a variable. Thus it is clear that in making the decision on the correspondence between a transaction record and a master file record, there are available at least two and perhaps more possible decisions. If one considers now the costs of the various actions that might be taken and the utilities associated with their pos-

*Reprinted with permission from the Journal of the American Statistical Association, American Statistical Association, December 1968, Vol. 63, pp. 1321-1332.

sible outcomes, it appears to be desirable to choose decision rules that will in some sense minimize the costs of the operation.

There are many other contexts in which record linkage takes place. One example is that in which two files are to be consolidated. Information about some individuals may be contained in one or another of the two files, while for other individuals some information may be in one file and some in the other. Another example is that of multi-frame sample surveys in which it may be necessary to determine which of the sampling units in one frame are also included in the other frame. A third example is that of geographic coding in which the master file consists of a street address guide and the transaction records are particular addresses; the problem here is to assign to each address a geographic code as given by the street address guide. The reader can doubtless supply many other examples.

The literature on this subject is replete with descriptions of actual matching operations ([2], [3], [4], [7], [8], [10], [11], [12], [13], [17], [18]). Several also deal with principles for the design of matching operations ([4], [7], [8], [9], [11], [12]). Some formulate mathematical models to serve as a basis for the design of a matching process that will be optimum in some sense. Thus, in analogy to the Neyman-Pearson theory of testing statistical hypotheses, Sunter and Fellegi [14][1] fix the probabilities of erroneous matches and erroneous non-matches and minimize the probability of cases for which no decision is made. Nathan ([5], [6]) proposes a model that involves minimization of a cost function, but restricts detailed discussion to cases in which the information used for matching appears in precisely the same form whenever the item exists in either list. Du Bois' [1] approach is to attempt to maximize the set of correct matches while minimizing the set of erroneous matches.

This paper proposes a mathematical model of the record linkage problem and a decision rule which minimizes the cost. The implementation of this model in practice depends upon the estimation of the parameters of the model. These parameters are costs and certain probabilities. The parameters may be difficult to determine. Also, it will be seen, the mathematical model (as usual) is not an exact representation of the real world. Nevertheless, the model provides useful guides for the construction of efficient linkage rules, as will be illustrated in the sequel.

2. A MATHEMATICAL MODEL

There are given two lists: a list A (the master file, say) which consists of a set of labels $\{\alpha\}$ and a list B (the transaction file, say) consisting of a set of labels $\{\beta\}$. (See Section 6 for a simple example.) Each label α is to be compared with each label β and an action taken on the basis of that comparison. The action taken must be one of a list of possible actions exemplified by, but not confined to, the following:

1. Treat the labels α and β as if they designated the same individual of some population. We shall say that the pair (α, β) is a "link".

[1] The notation and terminology used here follow, generally, those of the Sunter-Fellegi paper.

2. Temporarily treat the labels α and β as a link but obtain additional information before classifying the pair as a link or a non-link.
3. Take no action immediately but obtain additional information before classifying the pair as a link or non-link.
4. Temporarily treat the labels α and β as if they were associated with different individuals of the population, but obtain additional information before classifying the pair as link or non-link.
5. Treat the labels α and β as if they were associated with different individuals of the population (non-link).

Other actions may be added to the list, including for example the use of a randomizing device to determine the treatment of the pair (α, β). Each pair (α, β) will be called a "comparison pair." It is assumed that each pair (α, β) is either a "match" (the labels α and β are associated with the same individual of the population) or a "nonmatch" (the labels α and β are associated with different individuals of the population). Thus the set of all comparison pairs is the sum of mutually exclusive sets M (the "match" pairs) and U (the "nonmatch" pairs).

It should be noted that the labels α and β are, in general, vector-valued. Thus a label may contain, for example, a name, address, age, and other characteristics of a person.

Theoretically, any comparison of the label α with the label β consists of constructing a vector-valued function γ of the comparison pair (α, β). (See Section 6 for a simple example of a comparison function.) The comparison function γ serves to classify all pairs into classes: (α_1, β_1) and (α_2, β_2) are members of the same class if and only if $\gamma(\alpha_1, \beta_1) = \gamma(\alpha_2, \beta_2)$. The comparison pairs in each given class are to be subjected to exactly one of s possible "actions" $a_1, a_2, \cdots, a_s$. (Examples of five possible actions were given above.) A "linkage rule" consists of the assignment of an action to each class.

Let a label α be selected at random from list A and a label β from list B, and let a non-negative loss $g(a_i; \alpha, \beta)$ be associated with taking action a_i on a pair (α, β). Let

$$P[M \mid \gamma] \equiv \text{Prob}[(\alpha, \beta) \epsilon M \mid \gamma(\alpha, \beta)]$$

denote the conditional probability that the pair (α, β) is a match, given the value of γ.

We assume here that G, the expected value of $g(a_i; \alpha, \beta)$, is a function only of a_i and $P[M \mid \gamma]$. (This assumption is discussed below, in Section 4.) Thus

$$G = \mathcal{E}\{g(a_i; \alpha, \beta) \mid a_i, P[M \mid \gamma]\} = G(a_i, P[M \mid \gamma]).$$

Given a linkage rule, the total expected loss of the rule is

$$\sum P(\gamma) \times G(a_i, P[M \mid \gamma])$$

where a_i is the action specified for γ by the linkage rule, and the summation extends over all γ. To minimize the total loss, we need only minimize each term of the sum, each term being non-negative.

A special case of the above is that in which there is a loss G_{i1} associated

with taking action a_i on a pair (α, β) when in fact that pair is a match, and a loss G_{i2} when in fact the pair is a nonmatch. In this case G, the expected value of the loss, can easily be seen to be a linear function of the conditional probability that the comparison pair is a match, given γ, for each action a_i.

If the functions G are linear in $P(M|\gamma)$, the interval (0, 1) for the probability of a match is divided into at most s "action intervals" each of which corresponds to one of the possible s actions. The action interval for a given action is the interval in which the cost function G for that action is less than the cost function for any other action.

Figure 1 illustrates a case in which $G(a_i, P[M|\gamma])$ is a linear function of

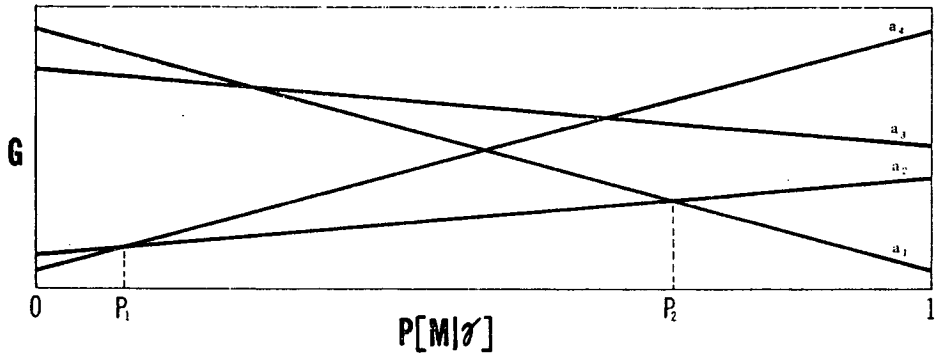

FIG. 1.

$P[M|\gamma]$ for each a_i. In this illustration, the optimum linkage rule specifies:

$$\begin{aligned} &\text{Take action } a_4 \quad \text{if} \quad 0 \leqq P[M|\gamma] \leqq P_1 \\ &\text{Take action } a_2 \quad \text{if} \quad P_1 < P[M|\gamma] \leqq P_2 \\ &\text{Take action } a_1 \quad \text{if} \quad P_2 < P[M|\gamma] \leqq 1 \end{aligned}$$

If the functions G are not linear in $P[M|\gamma]$, an "action set" of points of the interval (0, 1) that correspond to one of the possible actions will not be an interval in general. The treatment of the nonlinear case, however, proceeds along the same lines.

The conditional probability that a comparison pair is a match, given that the comparison function γ has a stated value depends upon the prior definition of the comparison function γ or, equivalently, upon the definition of the corresponding classification of comparison pairs.

As noted above, any comparison function γ defines a classification of the pairs (α, β). Let γ' be any other comparison function, which therefore defines another classification. It is possible to pass from the classification γ to the classification γ' by a sequence of steps, each of which consists either of splitting a class into two classes or of combining two classes into a single class. Therefore, if we begin with a tentative comparison function γ, we may seek ways of splitting some classes or combining some classes in such a way as to reduce the contribution of the classes involved to the loss function.

Consider the case of splitting a class γ into two classes γ_1 and γ_2. Without

loss of generality, we may assume that

$$P(M \mid \gamma_1) \leqq P(M \mid \gamma_2).$$

But then, clearly,

$$P(M \mid \gamma_1) \leqq P(M \mid \gamma) \leqq P(M \mid \gamma_2).$$

If $P(M|\gamma_1)$ and $P(M|\gamma_2)$ are in the same action set as $P(M|\gamma)$, there is no gain in making the split. But if either $P(M|\gamma_1)$ or $P(M|\gamma_2)$ falls into a different action set, the loss is necessarily (and sometimes materially) reduced.

To determine for which classes splits should be considered, one may first calculate the expected loss contribution for each class. It is evident that if the expected loss for a class is a small proportion of the total, little can be gained by splitting that class. Therefore, attention should be given first to classes whose expected loss contribution is a substantial proportion of the total. The illustration given subsequently shows that large reductions in the total expected cost can be attained by this technique.

With regard to the combining of classes, it is clear that this cannot result in reducing the expected cost. But if the classes to be combined are in the same action set, no increase in the cost will be sustained while the combination may reduce somewhat the operational costs of implementing the linkage rule. The combining of classes is useful also as an initial step, for the purpose of reducing the number of classes for which estimates need to be made, as detailed in Section 3, below.

3. ESTIMATION PROBLEMS

The application of the mathematical model involves estimating the cost function for each action as a function of the probability of a match, and estimating the probability that a comparison pair is a match.

The estimation of the cost function is often extremely difficult. Usually the cost consists of two classes of components, one class consisting of the cost of actual operations that may be involved and the other of the less tangible losses associated with the occurrence of errors of matching. The former can often be estimated very well, but estimates of the latter may depend upon judgment in large part. Despite the possible dependence on judgment, in the framework of the mathematical model even rough guesses at the cost function are extremely useful.

It may be noted that the first class of components of the cost function usually contains some components that are functions of the linkage rule (specifically, of the classification imposed). This is not reflected in the model, which only defines an optimum linkage rule for a fixed classification or comparison function.

It should be noted in connection with the estimation of the probabilities that it is necessary only to determine in which of the action sets a given probability falls. Ordinarily the probabilities will be estimated by selecting a sample in each comparison class. The sampling designs used should be chosen with the whole problem in mind, so that unnecessary sampling costs are avoided when, for example, the probability being estimated is near the center of an

action interval or when an error in the estimate of the probability will have little effect on the total cost. The latter may occur if the frequency of the given comparison class is small or if the alternative actions in the neighborhood of a given probability lead to costs which are only slightly different.

The successive steps in the application of the mathematical model may be described as follows:

1. The possible actions that may be taken on a comparison pair are listed.
2. For each action, the mathematical expectation of the cost as a function of the probability of a match is estimated.
3. An initial comparison function, i.e., an initial classification of comparison pairs into comparison classes, is determined on the basis of judgment or past experience (see, for example, [2], [3], [4], [7], [8], [9], [10], [11], [12], [15], [17], [18]), or on the basis of mathematical conclusions following from specified assumptions[2] about the interaction of the components of the labels α and β. The more nearly the initial classification resembles the optimum classification, the less is the amount of subsequent work required to attain the classification that will finally be used.
4. Samples are selected from each comparison class and the probability of a match estimated for each comparison class. This determines the optimum action pattern for the given classification.
5. The contributions of the several comparison classes to the total cost is now analyzed, and the classes that provide large contributions to that total cost are identified.
6. On the basis of that analysis, the classification is revised by splitting and recombining classes.
7. Steps 4 to 6 are repeated until step 6 indicates that no substantial additional reduction of cost can be made.

4. SOME COMMENTS ON THE MODEL

As is usually the case with a mathematical model, the model does not, in every respect, faithfully represent the real world that it is intended to describe.

The model assumes that every possible comparison pair will actually be examined. With large files, this would involve an inordinate number of comparisons. In practice, comparisons would be confined to specified subsets of the master file, and corresponding subsets of the transaction file. From the point of view of the mathematical model, the comparisons not actually made are being treated as non-links.

A limitation of the model is that it permits a given element of the transaction file to be treated as a link with more than one element of the master file. In many situations, this treatment may be intolerable. The difficulty can be handled by subjecting all such multiple-link cases to a subsequent stage in

[2] Thus Sunter and Fellegi [14] suggest that the components of the comparison vector may be grouped into sub-vectors which are statistically independent on each of the sets M and U. They then show how the value of a parameter equivalent to $P[M \mid \gamma]$ may be estimated on the basis of a knowledge of the frequency distribution of γ. This would serve to define an initial comparison function, even if the assumption of independence is not a satisfactory one.

which the transaction record is linked with at most one of the master file records associated with it in the first stage. If the cost or frequency of such cases is small, the mathematical model described in this paper remains a useful one for guiding the design of the linkage rule.

Similarly, there exist situations in which the linkage of a master file record with more than one transaction record is not tolerated.

There are some situations in which the cost is not only a function of the probability of a match but also of some other characteristic of the comparison pair. Thus, there may be two types of master file records, with the cost of an erroneous link being different for the two types. In such a situation, the comparison pairs may be classified in such a way that the characteristic is constant within each class and then the problem of optimum linkage may be treated as a separate problem in each of these classes.

The model is applicable also to cases in which the master file is not fixed but changes from one time period to another. Each transaction record is to be compared with the master file as it exists at the time period when the transaction record enters the system. We may consider the sequence of master files as constituting list A and a corresponding sequence of transaction files as constituting list B. The identity of the particular file becomes a component of the comparison vector γ, and we may define (α, β) to be a member of U if α and β are not from corresponding files. In this manner, this situation is covered by the model.

Some comments on the characteristics of useful comparison function are in order. Typically, the cost function

$$G(P) = \min_{a_i} G(a_i, P[M \mid \gamma])$$

is a concave function of P, with $G(0) = G(1) = 0$. Thus, the ideal comparison function is one for which $P[M \mid \gamma]$ is either 0 or 1 for every value of γ that may be observed. This ideal is usually not attained. However, one can usually find an initial comparison function such that the distribution of $P[M \mid \gamma]$ over the set of all comparison pairs is U-shaped, with low frequency where the cost function is high and high frequency where the cost function is low. Carrying through the steps given in Section 3 will often result in revising the comparison function γ so that the distribution of $P[M \mid \gamma]$ is shifted nearer the endpoints of the interval (0, 1).

Finally, it should be noted that the successive steps listed in Section 3 do not necessarily converge to the optimum decision rule. The procedure does provide an effective means of reducing the cost, as illustrated in Section 5.

5. AN ILLUSTRATION

The model described above was developed in connection with a file maintenance application, the master files being the lists of subscribers of two large magazine publishers ([15], [16]). In connection with the development of a system employing a large-scale electronic computer for the maintenance of the files of subscribers, it was necessary to develop explicit rules for matching the transaction file with the master file of subscribers. Initially, matching rules were developed on an intuitive basis, but the subsequent development of the

mathematical model indicated ways in which the matching rules could be substantially improved. The illustration presented here is confined to transactions which are subscription orders. (Other types of transactions included changes of address, complaints of non-delivery, subscription cancellations, and so forth. Separate linkage rules should be established for each type.)

TABLE 1. TENTATIVE UNIT COSTS

Action	True Status	
	Match	Non-match
1	\$0.00	\$6.01
2	.41	1.13
3	.77	.77
4	.82	.41
5	2.59	.00

Table 1 shows tentative unit costs developed by the staff of one of the publishers on the basis of consideration of the character of the actions and the consequences of these actions. The actions listed are roughly the same as those given above as examples in the description of the model. Computation from these unit costs would indicate that the optimum action intervals are as follows:

Action	Probability of a Match
1	$P > .92$
2	$.64 < P < .92$
3	—
4	$.19 < P < .64$
5	$P < .19$

Figure 2 shows the cost function for each of the possible actions. Note that action 3 is never used, since its cost function lies everywhere above some other cost function.

A systematic sample of approximately 10,000 subscription orders during a period of four months was selected. The portion of the master file used for this study consisted of those records for which the post office and the first four letters of the surname were the same as some record in the sample of transactions. Thus, comparison pairs to be examined were confined to those in which the post office and the first four letters in the surname were the same in the two members of the pair. (This is consonant with the comment made above in Section 4 that, in practice, comparisons are usually confined to specified subsets of the master file and the transaction file. This procedure adds, to the cost of any of the alternative linkage rules considered, the contribution from linking errors made for pairs (α, β) that are not actually examined.) To reduce the size of the master file for the purpose of this study, a subsample of one in ten of the master file records not matching a transaction record was selected from those sets that contained 100 or more records, a set here being defined as

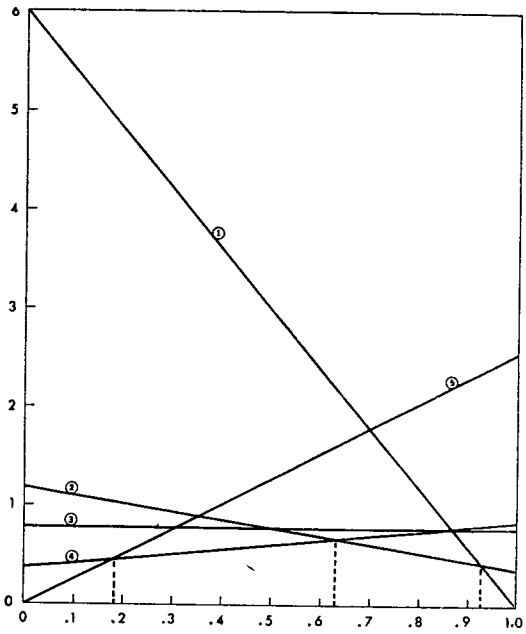

FIG. 2. Cost function for each of five actions, and the optimum action intervals.

a group of master file records having the same post office and first four letters of surname. The number of master file records in the final sample was about 83,000 and the number of comparison pairs about 192,000.

The comparison pairs in the sample were then classified into comparison classes that corresponded to the initial intuitive rule already being employed in the system. The probability of a match in each comparison class was estimated as the proportion of the comparison pairs in that class that were judged to correspond to each other. The determination as to whether a given comparison pair was or was not a match cannot be regarded as definitive since that determination was based upon judgment. However, there were at least two independent judgments for each case, and discrepancies between the judgments were resolved by further review and judgments. It was planned, but never carried out, that results should be refined by selecting a subsample of comparison pairs from the classes defined and then making more intensive investigations of each of the subsample pairs in an effort to determine definitively whether or not the pair was a match. However, it is suggestive to consider some of the consequences if the match status assigned is assumed to be correct. For example, it is interesting to consider the difference in the cost of the initial intuitive rule and the optimum rule based upon the assumed cost system.

Table 2 lists the 52 classes of comparison pairs with the size of each class and the estimated probability of a match in each class. For the initial intuitive rule and for the optimum rule, the table shows the action to be taken for each class, the expected cost for this sample, and the percentage of the total cost. Thus, it is estimated that the expected cost using the initial rule would have been $1,800 for this sample while the cost using the optimum rule was reduced

TABLE 2. COSTS FOR THE SAMPLE, FOR TWO MATCHING RULES, ASSUMING THE TENTATIVE UNIT COSTS

Comparison class	Total pairs	Estimated percent match	Estimated Expected Costs					
			Initial Rule			Optimum Rule		
			Act	$	% of total	Act	$	% of total
1	1,496	99.5	1	42.07	2.3	1	42.07	4.4
2	17	47.1	1	54.09	3.0	4	13.53	1.4
3	544	87.5	1	408.68	22.7	2	272.00	28.7
4	31	96.8	1	6.01	.3	1	6.01	.6
5	38	97.4	1	6.01	.3	1	6.01	.6
6	59	100.0	1	0.00	.0	1	0.00	.0
7	4	100.0	1	0.00	.0	1	0.00	.0
8	63	98.4	1	6.01	.3	1	6.01	.6
9	16	50.0	1	48.08	2.7	4	9.84	1.0
10	14	100.0	1	0.00	.0	1	0.00	.0
11	13	92.3	1	6.01	.3	1	6.01	.6
12	84	94.0	1	30.05	1.7	1	30.05	3.2
13	17	94.1	1	6.01	.3	1	6.01	.6
14	13	53.8	1	36.06	2.0	4	8.20	.9
15	10	70.0	1	18.03	1.0	2	6.26	.7
16	93	86.0	1	84.14	4.7	2	48.21	5.1
17	56	46.4	1	180.30	10.0	4	33.62	3.6
18	56	98.2	2	23.68	1.3	1	6.01	.6
19	26	0	2	29.38	1.6	5	0.00	.0
20	161	8.1	2	172.57	9.6	5	33.67	3.6
21	53	100.0	2	21.73	1.2	1	0.00	.0
22	17	0	2	19.21	1.1	5	0.00	.0
23	77	19.5	2	76.21	4.2	4	37.72	4.0
24	66	54.5	2	48.66	2.7	4	31.47	3.3
25	11	90.9	4	8.61	.5	2	5.23	.6
26	44	0	4	18.04	1.0	5	0.00	.0
27	97	3.1	4	41.00	2.3	5	7.77	.8
28	17	94.1	4	13.53	.8	1	6.01	.6
29	6	0	4	2.46	.1	5	0.00	.0
30	52	7.7	4	22.96	1.3	5	10.36	1.1
31	30	6.7	4	13.12	.7	5	4.10	.4
32	101	9.9	4	45.51	2.5	5	23.90	2.5
33	36	8.3	4	15.99	.9	5	7.77	.8
34	24	29.2	4	18.31	1.0	4	12.71	1.3
35	163	0	5	0.00	.0	5	0.00	.0
36	454	0.2	5	2.59	.1	5	2.59	.3
37	62	0	5	0.00	.0	5	0.00	.0
38	2,822	1.1	5	77.70	4.3	5	77.70	8.2
39	43,678	0	5	0.00	.0	5	0.00	.0
40	129,936	0.005	5	15.54	.9	5	15.54	1.6
41	265	2.3	5	15.54	.9	5	15.54	1.6
42	30	16.7	5	12.95	.7	5	12.95	1.4
43	646	0	5	0.00	.0	5	0.00	.0
44	1,709	0	5	0.00	.0	5	0.00	.0
45	74	0	5	0.00	.0	5	0.00	.0
46	62	0	5	0.00	.0	5	0.00	.0
47	25	8.0	5	5.18	.3	5	5.18	.5
48	8	37.5	5	7.77	.4	4	4.51	.5
49	491	1.2	5	15.54	.9	5	15.54	1.6
50	1	100.0	5	2.59	.1	1	0.00	.0
51	168	20.2	5	88.06	4.9	4	82.82	8.7
52	8,089	0.2	5	33.67	1.9	5	33.67	3.6
Totals	192,125			$1,799.65	99.8%		$946.59	99.6%

to about $950, or about one-half. The estimated standard error of the estimated percentage reduction in cost is approximately 2 percentage points. It is also suggestive to note that 4 of these comparison classes account for more than half of the expected cost of the optimum rule but involve fewer than 2 per cent of all comparison pairs. There is a distinct possibility that an intensive investigation of these 4 comparison classes could markedly reduce the cost of the optimum rule by subdividing these comparison classes.

6. A SIMPLE EXAMPLE OF A COMPARISON FUNCTION

To clarify the notion of a comparison function, the following simple example is given. The example is given for illustration only and bears no direct relationship to the numerical illustration given above, in which the comparison classes are defined in a more complex way.

Let each label α or β consist of the following components, a "blank" being an admissible entry for a component:

1. Surname
2. Given name
3. House number
4. Street name
5. Post office zip code

Then $\gamma(\alpha, \beta)$ may be defined as a vector $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ where

$\gamma_1 = 0$ if the surname is blank in either α or β.
1 if the surname is the same in α and β, and is a member of a specified list of common surnames.
2 if the surname is the same in α and β, and is not a member of the specified list of common surnames.
3 if the surname is different in α and β, and at least one of them is a member of the specified list of common surnames.
4 if the surname is different in α and β, and neither is a member of the specified list of common surnames.

$\gamma_2 = 0$ if the given name is blank in either α or β.
1 if the given name is the same in α and β.
2 if the given name is different in α and β.

$\gamma_3 = 0$ if the house number is blank in either α or β.
1 if the house number is the same in α and β.
2 if the house numbers are different in α and β, but one is a permutation of the other.
3 if the house numbers are different in α and β, and one is not a permutation of the other.

$\gamma_4 = 0$ if the street name is blank in either α or β.
1 if the street names are the same in α and β.
2 if the street names are different in α and β.

$\gamma_5 = 1$ if the zip codes are the same in α and β.
2 if the zip codes are different in α and β.

(It is assumed that the zip code is always present or can be supplied.) Thus the function γ may have up to 360 distinct values in this example.

It should be noted that the number of distinct values of the comparison function may be reduced by a process of combination. That is, we may define another comparison function γ' in terms of *sets* of values γ. Let the 360 possible values of γ be classified into sets S_i. Then $\gamma'(\alpha, \beta) = \gamma'_{(i)}$ if and only if $\gamma(\alpha, \beta) \epsilon S_i$.

I thank the referees for their helpful comments.

REFERENCES

[1] Du Bois, N. S. D'Andrea. "On the problem of matching documents with missing and inaccurately recorded items (Preliminary report)." *Annals of Mathematical Statistics*, 35 (1964), p. 1404 (Abstract).

[2] Fasteau, Herman H. and Minton, George. *Automated Geographic Coding System*. 1963 Economic Census: Research Report No. 1, U. S. Bureau of the Census (unpublished). (1965).

[3] Kennedy, J. M. *Linkage of Birth and Marriage Records Using a Digital Computer*. Document No. A.E.C.L.-1258, Atomic Energy of Canada Limited, Chalk River, Ontario. (1961).

[4] Kennedy, J. M. "The use of a digital computer for record linkage." *The Use of Vital and Health Statistics for Genetic and Radiation Studies*, United Nations, New York, (1962), pp. 155–60.

[5] Nathan, Gad. *On Optimal Matching Processes*. Doctoral Dissertation, Case Institute of Technology, Cleveland, Ohio (1964).

[6] Nathan, Gad. "Outcome probabilities for a record matching process with complete invariant information." *Journal of the American Statistical Association*, 62 (1967), pp. 454–69.

[7] Newcombe, H. B. "The study of mutation and selection in human populations." *The Genetics Review*, 57 (1965), pp. 109–25.

[8] Newcombe, H. B. and Kennedy, J. M. "Record linkage: Making maximum use of the discriminating power of identifying information." *Communications of the Association for Computing Machinery*, 5 (1962), pp. 563–66.

[9] Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. "Automatic linkage of vital records." *Science*, 130 (1959), pp. 954–9.

[10] Newcombe, H. B. and Rhynas, P. O. W. "Child spacing following stillbirth and infant death." *Eugenics Quarterly*, 9 (1962), pp. 25–35.

[11] Nitzberg, David M. and Sardy, Hyman. "The methodology of computer linkage of health and vital records." *Proceedings, Social Statistics Section, American Statistical Association*. (1965), pp. 100–6.

[12] Perkins, Walter M. and Jones, Charles D. "Matching for Census Coverage Checks." *Proceedings, Social Statistics Section, American Statistical Association*. (1965), pp. 122–39.

[13] Phillips, William and Bahn, Anita K. "Experience with matching of names." *Proceedings, Social Statistics Section, American Statistical Association*. (1963), pp. 26–9.

[14] Sunter, A. B. and Fellegi, I. P. *An Optimal Theory of Record Linkage*. Unpublished paper presented at the 36 Session of the International Statistical Institute, Sydney, Australia (1967).

[15] Tepping, Benjamin J. *Progress Report on the 1959 Matching Study*. National Analysts, Inc., Philadelphia, Pa. (1960).

[16] Tepping, Benjamin J. and Chu, John T. *A Report on Matching Rules*. National Analysts, Inc., Philadelphia, Pa. (1958).

[17] U.S. Bureau of the Census. *Evaluation and Research Program of the U. S. Censuses of Population and Housing, 1960: Record Check Studies of Population Coverage*. Series ER 60, No. 2. U. S. Government Printing Office, Washington, D. C. (1964).

[18] U.S. Bureau of the Census. *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics as Measured by CPS-Census Match*. Series ER 60, No. 5. U. S. Government Printing Office, Washington, D. C. (1964).