WEIGHTS IN COMPUTER MATCHING: APPLICATIONS AND AN INFORMATION THEORETIC POINT OF VIEW

Nancy J. Kirkendall, Energy Information Administration

This paper summarizes the historical development of computerized match/merge procedures and describes the test statistic used to classify record pairs as a match or nonmatch in terms of its information theoretic interpretation. Current match/merge software procedures are compared and contrasted based on their differing approaches to estimation.

INTRODUCTION

The match/merge procedures discussed in this paper are those which are intended to perform exact matching. Exact matching has been defined (U.S. Department of Commerce, 1980) as the linkage of records from two or more files containing units from the same population. The intention of exact matching is to link data for the same unit (e.g., person) from different files. If units which do not represent the same individual are linked, the result is a false match or type 2 error. If units which do represent the same unit are not linked, the result is a missed match, or type 1 error.

There are many different purposes in exact matching. Examples range from obtaining more data elements for an individual by merging information from different surveys, to creating a more comprehensive name and address list by merging the names and addresses from many sources. In the first case, it is important to make sure that matching is done accurately so that the merged data constitute a multivariate observation from a single individual (see Kelley, 1983). In the second case, the merging is intended to ensure as complete a list as possible while eliminating duplication.

The most significant paper on the theory and practice of matching is by Fellegi and Sunter (1969). Their paper documents the derivation of a test statistic and a critical region for deciding whether or not a pair of records is a match. In addition, it discusses some of the assumptions necessary for practical application and describes approaches for estimating the probabilities which are used to calculate the test statistic. Most of the probabilistic match/merge procedures in use today are based on an application of the techniques described in the Fellegi-Sunter paper.

Although the Fellegi-Sunter paper was the first publication of the theoretical background for match/merge procedures, many of the ideas and techniques embodied in the methodology had been used since the late 1950's by Howard Newcombe et al. Newcombe's papers from that time period describe the use of the test statistic for which the derivation was later presented by Fellegi and Sunter. (See Newcombe et al., 1959 and Newcombe and Kennedy, 1962.)

THEORETICAL BACKGROUND

Assume that two files, A and B, are to be merged. Each file contains at least one record for each unit (person or establishment) in the file. Each record contains a set of attributes for that unit. These attributes may include: numerical identifiers with very good identifying characteristics such as the social security number; standard identifiers such as name and address; characteristic information such as sex or date of birth; or any other data which might be available on survey files or administrative record files.

In the matching process, each record in file A can be compared to each record in file B. The comparison of any such pair of records can be viewed as a set of outcomes, each of which is the result of comparing a specific attribute from the record in file A with the same attribute in the record from file B. Outcomes may be defined as specifically as desired. For example, one might define an outcome of a comparison to be simply that the attributes agree or that they disagree. Or, one might define the agreement outcome more specifically, based on the possible values that attribute can take. For example, one outcome might be that the surnames agree and equal "Smith," while another might be that the surnames agree and equal "Zebra," etc.

"Comparison of attributes" is usually interpreted to mean that the same attribute is recorded on each record and that they can be compared directly. However, it is possible to "compare" different attributes which are known to be correlated or to use information from only one record in conjunction with general information from the other file. An example is given in Smith, Newcombe, and Dewar (1983). In their application, records from a file of patients diagnosed as having cancer are linked with records in a death file. The variable "cause of death" in the death file is used in conjunction with general statistics concerning the cause of death among cancer patients and the cause of death among the general population to provide a different sort of "comparison of attributes."

In the above, it was implied that every record from file A is compared to every record from file B. In practice, with large files this would require an extremely large number of comparisons, the vast majority of which would not be matches. To make the size of the problem more manageable, files are generally "blocked" using one or more of the available attributes, and record pairs are assumed to be a possible match and subject to the detailed attribute comparison only if they agree on the blocking attribute. In using a blocking procedure, there is necessarily a higher rate of unmatched

duplicates (type 2 error) because records which do represent the same unit, but disagree on the blocking attribute, are automatically rejected as possible matches. However, the gains in the form of reduced processing are significant. See Kelley (1985) for a probabilistic approach to selecting blocking strategies.

THE PROBLEM

Probabilistic test procedures are based on evaluating record pairs one at a time and subjecting each pair to a decision as to its match status. The procedure does not consider the expected number of matches or nonmatches in a merging of two files, and does not make use of the result of the classification of any previous record pairs.

In this section the test statistic and the critical region are described based on an information theoretic argument. Details of the derivation are presented in the Appendix. The resulting test statistic and critical region are exactly the same as those derived by Fellegi and Sunter. One advantage of the information theoretic approach is that the inclusion of the log of the prior odds of a match, as described by Howe and Lindsay (1981) and by Newcombe and Abbatt (1983) can be directly related to the methodology. Calculation of this test statistic yields a value which is commonly referred to as the "weight" for or against a match.

Given any pair of records, we want to make a decision as to whether they match (H_o -- the null hypothesis) or do not match (H_1 -- the alternative hypothesis). This decision will be based on the observed comparison of the attribute items on the two records. The set of all outcomes resulting from this comparison is the random variable, x_i, which takes values according to the outcomes which were specified for all of the attributes.

The discrete random variable, x_i, can take any of n different values. The number n can be very large, either because a large number of attributes are compared, or because a large number of outcomes are possible for any one attribute comparison. The probabilities with which x_i takes any of the n values under both H_o and H_1 are assumed to be known. The question of estimating these probabilities is addressed later. The decision process is formalized by considering the following two hypotheses:

H_o: The event that two records represent the same unit (i.e., a match). Under H_o, the frequency function of the random variable, x_i, is denoted $P(x_i/H_o) = p_{oi}$ for i=1, ... n.

H_1: The event that the two records represent different units (i.e., a nonmatch.) Under H_1, the frequency function of the random variable, x_i, is denoted $P(x_i/H_1) = p_{1i}$ for i=1, ... n.

AN EXAMPLE OF A COMPARISON VARIABLE

Assume that two records are being compared and that a decision will be made as to their match status based on a comparison of three attributes: surname, first name, and sex. For each attribute there will be two possible outcomes: either they agree or they do not agree. Thus, the comparison set can take any of 2**3 = 8 (n=8) possible values. For simplicity we also assume that the probabilities of agreement or disagreement of the attributes are independent under both H_o and H_1. Thus, given the following table of probabilities, the frequency function of the comparison vector can be calculated under both hypotheses.

TABLE I
PROBABILITIES OF AGREEMENT

Attribute	Under H_o	Under H_1
Surname	.90	.05
First name	.85	.10
Sex	.95	.45

In the following let $x=(a_1,a_2,a_3)$, where $a_i = 0$ if item i disagrees, and $a_i=1$ if item i agrees. The comparison of surname is represented by a_1, the comparison of first name by a_2, and the comparison of sex by a_3. Thus, the random variable, x_i, has the frequency functions given by p_{oi} (under H_o) and p_{1i} (under H_1) in the following table.

TABLE II
PROBABILITIES FOR COMPARISON VARIABLE

i	x_i	p_{oi}	p_{1i}
1	(0,0,0)	.0008	.4703
2	(1,0,0)	.0068	.0248
3	(0,1,0)	.0043	.0523
4	(0,0,1)	.0143	.3848
5	(1,1,0)	.0383	.0028
6	(1,0,1)	.1283	.0203
7	(0,1,1)	.0808	.0428
8	(1,1,1)	.7268	.0023

THE TEST STATISTIC

As shown in the Appendix, the test statistic

$$T(x_i) = \log(p_{oi}/p_{1i}) = I(o:1;x_i). \qquad (1)$$

is a sufficient statistic for discriminating between H_o and H_1. The number $\log (p_{oi}/p_{1i})$ is an information number. It provides a measure of

the information for discriminating for H_o and against H_1 which was gained by observing the random variable, x_i.

$T(x_i)$ is the log of the ratio of the probability of the outcomes, denoted by x_i, under H_o to the probability of the same set of outcomes under H_1 (the log of the likelihood ratio.) Note that if these probabilities are the same then $T(x_i)=0$, and this set of outcomes has no discriminating power for identifying whether records represent the same unit. If p_{oi} is larger than p_{1i}, then $T(x_i)$ will be positive for that category. The larger $T(x_i)$, the stronger is the possibility that observation of this set of outcomes indicates that the records represent the same unit. If p_{oi} is smaller than p_{1i}, then $T(x_i)$ is negative. The smaller $T(x_i)$, the stronger is the possibility that this set of outcomes indicates that the records do not represent the same unit.

DETERMINING THE CRITICAL REGION

The final part of the matching problem is to determine cut-off values, c_1 and c_2, so that H_1 is rejected if $T(x_i)$ is greater than c_2 and H_o is rejected if $T(x_i)$ is less than c_1. If $T(x_i)$ falls between these two values, the test is inconclusive and the record pair may be subject to manual follow up.

In standard applications of testing simple hypotheses, there are only two outcomes: accept the null hypothesis or reject it. Here, the three region test comes from the union of two tests. First, consider a test of H_o vs. H_1. For a test with significance level alpha, this leads to the critical region defined by c_1. Next, consider the test of H_1 vs. H_o with significance level beta. This leads to a critical region defined by c_2. Individually, according to the Neyman-Pearson Lemma, these tests are the best tests at their respective significance levels. The first test rejects H_o if $T(x_i)$ is less than c_1. The second test rejects H_1 if $T(x_i)$ is greater than c_2. Since c_1 is generally less than c_2, the union of these two tests yields the three region test described above.

This is illustrated below with our previous example. In Table III the column labeled $T(x_j)$ is the log of the ratio of p_{oj} and p_{1j} from Table II, but here the table is arranged so that the $T(x_j)$ are in ascending order. The next to last column presents the cumulative probability under H_o of observing $T(x_i)$ less than or equal to the given $T(x_j)$. It is used to specify c_1. In this example, if alpha is equal to .05, then c_1 is equal to -1.9. The last column is the cumulative probability under H_1 of observing $T(x_i)$ greater than or equal to the given $T(x_j)$. It is used to specify c_2. In this example, if beta is equal to .05 then c_2 is equal to 2.7.

TABLE III

THE DISTRIBUTION OF THE TEST STATISTIC

j	x_j	$T(x_j)$	p_{oj}	p_{1j}	$\sum_{k=1}^{j} p_{ok}$	$\sum_{k=j}^{n} p_{1k}$
1	(0,0,0)	-9.2	.0008	.4703	.0008	1.0004
2	(0,0,1)	-4.8	.0143	.3848	.0151	.5301
3	(0,1,0)	-3.6	.0043	.0523	.0194	.1453
4	(1,0,0)	-1.9	.0068	.0248	.0262	.0930
5	(0,1,1)	.9	.0808	.0428	.1070	.0682
6	(1,0,1)	2.7	.1283	.0203	.2353	.0254
7	(1,1,0)	3.8	.0383	.0028	.2736	.0051
8	(1,1,1)	8.3	.7268	.0023	1.0004	.0023

Thus, if alpha and beta both equal .05, we would classify a pair as a match if we observe vectors (1,0,1), (1,1,0), or (1,1,1). We would classify pairs as a nonmatch if we observe (0,0,0), (0,0,1), (0,1,0), or (1,0,0). If we observed (0,1,1): agreement on sex and first name, but disagreement on surname, we would be unable to classify the pair as either a match or a nonmatch.

The test statistic and critical region defined in this way are the same as those developed by Fellegi and Sunter (1969), although that paper also included a discussion of randomization to achieve the type 1 and type 2 error levels exactly. They develop the decision rule for accepting H_o or H_1 based on minimizing the probability of not making a decision. That is: minimizing the probability that $T(x_i)$ falls between c_1 and c_2 for a given alpha and beta.

THE POSTERIOR ODDS RATIO

The development presented here and in Fellegi-Sunter (1969) use the test statistic defined in equation (1). However, equation (A2) can be rewritten as

$$\log P(H_o/x_i)/P(H_1/x_i) = \log p_{oi}/p_{1i} + \log P(H_o)/P(H_1). \quad (2)$$

Here the log of the posterior odds ratio is written as the sum of the information number and the log of the prior odds ratio. Howe and Lindsay (1981) call equation (2) the "total weight" for a match, but acknowledge that the prior odds ratio is difficult to evaluate. The most recent papers by Newcombe and Smith include

procedures for estimating the prior odds ratio in some unique situations (see Newcombe and Abbatt, 1983 and Smith, Newcombe, and Dewar, 1983). Note that the prior odds ratio reflects any information available regarding the match status of a given record pair before the attribute comparison. If the prior odds of a match were the same for each record pair then the test statistic and critical region for the comparison of attributes would both be shifted by the same value. In such a case the inclusion of the prior odds ratio would not change the outcome of the statistical test. However, the posterior odds ratio has the advantage that it can be interpreted directly as the odds that the record pair matches.

In the Smith, Newcombe, and Dewar paper, the prior odds ratio is calculated based on a life table analysis of the severity of cancer diagnosed, an attribute available in the search file, and the year of the death file being searched. In their example, the prior probability of a match is different for each individual in the search file and instead of applying specifically to a record pair, it applies to the individual record initiating the search and to an entire one year death file.

INDEPENDENCE OF ATTRIBUTES -- A SIMPLIFYING ASSUMPTION

In the original pages of this discussion, x_i was defined to be a discrete random variable which was the intersection of m attribute comparisons. If the result of each attribute comparison is denoted as t_j for j=1, ..., m, then x_i can be written as the intersection of the t_j:

$$x_i = t_1 \cap t_2 \cap \ldots \cap t_m.$$

If t_1, ..., t_m are statistically independent, then equation (1) can be written as:

$$I(o:1;x_1) = \sum_{j=1}^{m} I(o:1;t_j).$$

Thus, if the set of attribute variables, t_j, are statistically independent, the weights (i.e., the information) for each t_j can be calculated separately, and the overall weight (the information contained in the intersection of the t_j) is just the sum of the weights for each t_j.

In the previous example, the three attributes were assumed to be independent. Hence, the weight for any observed vector can be calculated as the sum of the information associated with agreement or disagreement on each attribute. For example, for x_i=(0,1,1) the weight can be calculated as the sum of the information associated with disagreement on surname,

$$T(a_1=0) = \log (.1/.95) = -3.25;$$

the information associated with agreement on first name,

$$T(a_2=1) = \log (.85/.1) = 3.09;$$

and the information associated with agreement on sex,

$$T(a_3=1) = \log (.95/.45) = 1.08.$$

The sum of these weights is .92, as shown in Table III for the weight (the value of $T(x_j)$) associated with the observation (0,1,1). Thus, if it is reasonable to assume that the outcomes of attribute comparisons for different attributes are statistically independent, then the calculation of the test statistic is simplified because the weights can be calculated separately and summed.

In this example, it is reasonable to assume that agreement on surname is independent of agreement on either first name or sex. However, if there is agreement on first name, it is likely that there will be agreement on sex. Hence, in this example, the assumption of independence does not really hold. To incorporate this dependence, one would need to consider the probabilities associated with the bivariate random variable.

AN EXAMPLE OF A MULTIPLE OUTCOME COMPARISON

The following is a vastly simplified example of defining the specific outcomes of attribute comparison by making use of the values they can assume. This type of "frequency" argument results in lower weights for agreement on common items and higher weights for agreement on rare items. It is a simplified version of the treatment of frequencies and error structures presented in the Fellegi-Sunter paper, pages 1192 and 1193 (pp. 60 and 61 in this volume).

Here, assume that surnames are being compared in a pair of records. Assume that there are only two frequently occurring names in the file, "Smith" and "Jones"; the other names (m of them) all occurring with roughly the same low frequency. Thus, we define the following set of outcomes of the comparison of surname:

$$x = \begin{cases} \text{"Smith"} & \text{if the two variables agree and both equal "Smith,"} \\ \text{"Jones"} & \text{if the two variables agree and both equal "Jones,"} \\ \text{"other"} & \text{if both variables agree but do not equal either "Smith" or "Jones,"} \\ \text{"disagree"} & \text{if the items disagree.} \end{cases}$$

(Note that the set of outcomes defined for item comparison must specify a partition of the set of all possible results into mutually exclusive and exhaustive subsets.)

Further assume that: 1) surnames in the two files under consideration are both random samples from the same population, and that in this population, "Smith" occurs with probability p_a, "Jones" occurs with probability p_b, and each

of the other m error-free names in the file occurs with probability p_o; and 2) the only errors in the name fields are keypunch errors, which occur at the same rate, 1%, in both files, independent of the particular name.

Under H : A pair of records is a match. Names agree unless there is a keypunch error. Thus, the probability of agreement on Smith is $p_{o1} = p_a*(.99)**2$ (the probability of observing "Smith" times the probability that the value was keypunched correctly on both files). Similarly, the probability of agreement on Jones $p_{o2} = p_b*(.99)**2$, and the probability of agreement on one of the other names is $p_{o3}=p_o*(.99)**2$. The probability of disagreement on name when the record pairs represent the same individual is $p_{o4}= 1-p_{o1}-p_{o2}-m*p_{o3}$ $= (1-(.99)**2)*(p_a+p_b+m*p_o)$ $= 1-(.99)**2=.02$.

Under H_1: The records do not represent the same individual and any agreement on name occurs at random. The probability of agreement with name "Smith" is $(.99*p_a)**2$; the probability of agreement with name "Jones" is $(.99*p_b)**2$; the probability of agreement with some other name is $(.99*p_o)**2$; and the probability of disagreement on name is $1-.99**2*(p_a**2+p_b**2+m*p_o**2)$. (We have assumed that the probability that a keypunch error results in some valid name is negligible.)

Thus, from equation (1) the weight for the various outcomes is:

If x*=Smith,
$T(x*)=\log(.99**2*p_a/.99**2*p_a**2)=\log(1/p_a)$.

x*=Jones,
$T(x*)=\log(.99**2*p_b/.99**2*p_b**2)=\log(1/p_b)$.

x*=other,
$T(x*)=\log(.99**2*p_o/.99**2*p_o**2)=\log(1/p_o)$.

x*=disagree,
$T(x*)=\log(.02/(1-*.99**2*(p_a**2+p_b**2+m*p_o**2)))$.

Newcombe, Kennedy, Axford, and James (1959) noted that in frequency based matching, if an item, a, is found in a master file with probability p_a, and if the two files being matched can be viewed as a sample from that master file, then, when a record pair is a match, the probability that the items agree and equal "a" is proportional to p_a. When the record pair is a nonmatch the probability is proportional to p_a**2 with the same constant of proportionality. Thus, the weight for a match when item a is observed is $\log(p_a/p_a**2) = \log(1/p_a)$. This is illustrated in the example above. Most of the Smith and Newcombe papers describe calculation of the weights for agreement on a particular item as the log of the inverse of the frequency of occurrence of that item.

The Fellegi-Sunter paper presents a derivation of the frequency based weights for specific agreement in the presence of several types of errors. Their procedure still leads to weights for agreement of $\log(1/p_a)$ because, as in the above example, the error terms impact the probability of agreement under H_o and the probability of agreement under H_1 in the same way.

VARIATIONS IN PRACTICE

Probabilistic matching techniques (based on the Fellegi-Sunter paper) have been implemented in many software systems, including the Generalized Iterative Record Linkage System (GIRLS) from Statistics Canada (see Smith and Silins, 1984) which is now called the Canadian Linkage System (CANLINK); UNIMATCH from the U.S. Bureau of the Census (see Jaro, 1972); the Statistical Reporting Service's (SRS) Record Linkage System from the U. S. Department of Agriculture (USDA); and the California Automated Mortality Linkage System (CAMLIS) from the University of California at San Francisco. Work by Rogot et al. (1983) at the National Center for Health Statistics has also used probabilistic matching techniques.

The two major references for this section are a paper by Howe and Lindsay (1981), which describes a version of the GIRLS system, and a number of unpublished papers by Richard Coulter, Max Arellano, William Arends, Billy Lynch, and James Mergerson dated 1976 and 1977, which describe the SRS Record Linkage System. These two systems were included in this review because they are applications of a modified Fellegi-Sunter approach and because the available documentation was thorough.

The GIRLS system was developed to support epidemiological research. Thus, it is primarily intended to link records for a cohort group to morbidity or mortality data. Attributes available for comparison usually include first name, surname, middle initial, sex, date of birth, place of birth, parents' names and places of birth. Some of the application-specific items, such as blocking attribute and definition of outcomes for attribute comparison, are not fixed in the system. They can be specified by the user. In the following, the specific applications by Howe and Lindsay are described.

The SRS record linkage system is intended to support development and maintenance of state-level sampling frames for agricultural surveys. Here, the primary intent of the linkage system is to unduplicate a list created by merging

multiple lists. The most commonly available attributes are surname, first name, and address. In addition to the probabilistic matching procedure, record pairs which have identical address fields are reviewed manually to identify matches. This system is not a general-purpose matching system. It was developed and is used solely to maintain the USDA frames.

Blocking

In these applications, both systems block first on surname code -- a variation of the New York State Identification and Intelligence System (NYSIIS) code. A surname code is an alphabetic code designed so that the most similar names and the names with the most frequently encountered errors of misreporting will have the same code. See Lynch and Arends (1977) for a description of surname codes and the rationale used by SRS to select the NYSIIS code for their system. If the resultant block size is too big, SRS uses secondary blocking on first initial and tertiary blocking on location code. The Howe and Lindsay application blocks first on NYSIIS code, then on sex. In neither case are the weights changed to reflect the impact of blocking.

Weights for Agreement

Both systems make extensive use of frequency-based weights, and both systems use the files being matched to calculate the frequencies. Both systems also assume that these frequencies include keypunch errors, recording errors, and legitimate name changes. This is different from the Fellegi-Sunter approach, which assumed that the frequencies were based on an error-free name file.

The SRS approach handles partial agreements by calculating a weight for agreement on specific surname and a weight for agreement on specific NYSIIS code with disagreement on surname. The Howe-Lindsay paper extends the accounting for partial agreement by specifying agreement on specific first seven characters of surname; agreement on specific first four characters with disagreement on the next three characters; and agreement on specific NYSIIS code with disagreement on the first four characters of surname. In both systems, pairs with disagreement on NYSIIS code will never be considered because of the blocking.

Estimation of Error Rates

Both systems use an iteration scheme to provide final estimates for the required error rates. First, initial estimates are provided, a sample of records is processed through the matching algorithm, and a preliminary set of matched record pairs is identified. These pairs are assumed to be true matches and are used to estimate the error rates, as discussed below. These revised estimates for the error rates are input to the system; the sample is processed again and the newly matched pairs are used to reestimate the error rates. The iteration is continued until the estimates for the error rates converge.

The errors are handled in the Howe-Lindsay paper as transmission rates:

t_1 = the probability that the first seven characters of surname are equal to the "true" value;

t_2 = the probability that the first four characters are equal to the "true" value but the next three characters are different; and

t_3 = the probability that the surname code is equal to the "true" surname code, but that the surnames disagree in the first four characters.

These transmission rates can be estimated from a sufficiently large set of pairs which represent true matches by using the following counts: the number of pairs which agree on the first seven characters; the number of pairs which agree on the first four characters not on the next three, and the number which do not agree on the first four characters. The assumption is made that this set of matched pairs is representative of all possible matched pairs. Note that t_3 will be underestimated because of the blocking.

In the SRS system, the error rates used are:

e = the probability that a name is misreported or misrecorded

e_T = the probability that in a record pair which does represent the same unit, the names are correct but different.

These definitions of the error rates are the same as those used in the Fellegi-Sunter paper. The overall weights for specific agreement are different because the frequencies themselves are derived under different assumptions, as mentioned above. In the SRS system, the error rates are estimated from the set of pairs which represent true matches by using: the number of pairs which have the same name; the number which have different names; and the number which have similar names (where "similar" was not defined). Here, e_T will necessarily be underestimated because the blocking procedure assures that records will be compared only if they agree on NYSIIS code.

The Critical Region

Both systems use an empirical procedure to determine the critical region. That is, a frequency distribution of the weights for a sample of record pairs is plotted, and the critical values are selected based on the shape of the curve. As an alternative, the SRS system also calculates an initial lower critical region as the sum of the weights for agreement of the most common surname, first name, and location. The initial upper critical region is estimated as the initial lower critical region plus the weights for agreement on the most common middle name, route and box number. These calculated upper and lower regions are used during the

iteration to estimate error rates. They are conservative since both are positive.

System Considerations

In the Howe-Lindsay approach, an initial blocking and comparison are done before the frequency based agreement weights are calculated. At this stage, only weights for disagreement are summed and as the accumulated weight becomes too negative, the record pair can be rejected as a possible match before all attributes have been compared. With this approach the order of adding in attributes is important, with those having the greatest negative weight for disagreement entering first. If the total disagreement weight is above the threshold, the record pair is a possible match. A separate file is created containing those possibly matched pairs. For each such pair, this file contains one record with the identification numbers of the two records, the results of the comparison of attributes, and the values taken (if needed for the weight calculation). This potential linked file is then sent to a separate subroutine for calculation of the weights.

Grouping

Both systems create groups consisting of all records which have been linked with each other. (Here linked means that the calculated test statistic is above the upper critical value.) As described in the Howe and Lindsay paper, the group is formed by first taking a single record and adding to the group any records which have been linked to it, then adding all records which were linked to those records, and so on. Additional subgroupings are considered when two records from different groups have a weight between the two critical values.

Interpretation of the groups depends on the application. In the SRS application, members of a group could all be duplicates to each other. In the SRS system, subgroups are analyzed manually. In some of the applications described by Howe and Lindsay, neither input file has any duplication, and there is at most one matched record for a given record in the search file. In this case the groups are analyzed to pick the pair which represents the most likely match, usually the pair with the highest weight.

SUMMARY

This paper has described the probabilistic matching procedures discussed by Fellegi and Sunter (1969) from an information theoretic point of view. This approach gives additional insight into the calculation of the posterior odds ratio as mentioned by Howe and Lindsay, and as implemented in the recent work of Newcombe and Smith. Additionally, it has described some of the differences between two of the major systems which have been implemented based on the Fellegi-Sunter paper. Major differences between systems are in accounting for partial matches, the definition of the error rates, and in the handling of groups of record pairs which are all linked to each other. The major differences between these systems and the Fellegi-Sunter approach are 1) that these systems base their frequency counts on files which are acknowledged to contain errors, and 2) that they use an empirical procedure to determine the critical region for the statistical test.

REFERENCES

Arellano, Max and Arends, William, "The Estimation of Component Error Probabilities for Record Linkage Purposes," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished, May 1976.

Arellano, Max, "Optimum Utilization of the Social Security Number for Matching Purposes," "Weight Calculation for the Place Name Comparison," "Calculation of Weights for Partitioned Variable Comparisons (Trailing Blanks Case)," "Estimation of Component Error Probabilities for Record Linkage Purposes," "Development of A Linkage Rule for Unduplicating Agricultural Lists," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished papers, 1976 and 1977.

Arellano, Max and Coulter, Richard, "Weight Calculation for the Given Name Comparison," "Weight Calculation for the Middle Name Comparison," "Weight Calculation for the Surname Comparison," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished papers, 1976 and 1977.

Coulter, Richard, "An Application of a Theory for Record Linkage," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished, March 1977.

Coulter, Richard and Mergerson, James, "An Application of a Record Linkage Theory in Construction of a List Sampling Frame," Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., April 1977.

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

Howe, G. R. and Lindsay, J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies," Computers and Biomedical Research 14, Academic Press, 1981, pp. 327-340.

Jaro, Matthew, "UNIMATCH--A Computer System for Generalized Record Linkage Under Conditions of Uncertainty," AFIPS Conference Proceedings, Vol. 40 for Spring Joint Computer Conference, May 1972, pp. 523-530.

Jaro, Matthew, "UNIMATCH--Generalized Record Linkage Applied to Urban Data Files," Proceedings of the American Statistical Association.

Kelley, Patrick, "A Preliminary Study of the Error Structure of Statistical Matching," Proceedings of the American Statistical Association, Social Statistics Section, 1983.

Kelley, Patrick, "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," Record Linkage Techniques--1985, Internal Revenue Service.

Kullback, Soloman, Information Theory and Statistics, Dover Publications, Inc., New York, New York, copyright 1968.

Lynch, Billy and Arends, Williams, "Selection of a Surname Coding Procedure for the SRS Record Linkage System," Sample Survey Research Branch, Research Division, Statistical Reporting Service, U. S. Department of Agriculture, Feb 1977.

Newcombe, Howard, "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," American Journal of Human Genetics, Vol 19, No. 3, Part I, (May) 1967.

Newcombe, Howard, and Kennedy, James, "Record Linkage: Making Maximum Use of Discriminating Power of Identifying Information," Communications of the Association for Computing Machinery 5, 1962, pp. 563-566.

Newcombe, H., Kennedy, J., Axford, S., and James, A.,"Automatic Linkage of Vital Records," Science, 130, 1959, pp. 954-959.

Newcombe, H., and Abbatt, J., "Probabilistic Record Linkage in Epidemiology," Report Prepared for Eldorado Resources, Ltd., Oct. 1983.

Rogot, Eugene, Schwartz, Sidney, O'Connor, Karen and Olsen, Christina, "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," Proceedings of the American Statistical Association, Section on Business and Economic Statistics, 1983.

Smith, Martha, Newcombe, Howard, and Dewar, Ron, "Proposed Procedure for the Alberta Cancer Registry Death Clearance," Health Division, Statistics Canada, (OEHRU-No. 1), March 1983.

Smith, Martha, Newcombe, Howard, and Dewar, Ron, "The Use of Diagnosis in Cancer Registry Death Clearance," Health Division, Statistics Canada, (OEHRU-No. 2), April 1983.

Smith, Martha and Silins, John, "Generalized Iterative Record Linkage System," (An excerpt), Statistical Uses of Administrative Records: Recent Research and Present Prospects, Department of the Treasury, Internal Revenue Service, July 1984.

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

APPENDIX

This appendix presents a derivation of the test statistic for determining whether a record pair is a match or a nonmatch using an information theoretic approach (see Kullback, 1968).

WHAT IS AN INFORMATION NUMBER?

Given the prior probabilities associated with a match and a nonmatch, $P(H_o)$ and $P(H_1)$, we use Bayes theorem to calculate the posterior probabilities of H_o and H_1 based on the observed attribute comparison, x_i:

$$P(H_o/x_i) = P(H_o)*p_{oi}/(P(H_o)*p_{oi} + P(H_1)*p_{1i})$$

$$P(H_1/x_i) = P(H_1)*p_{1i}/(P(H_o)*p_{oi} + P(H_1)*p_{1i}).$$

Dividing these gives the posterior odds ratio:

$$P(H_o/x_i)/P(H_1/x_i) = P(H_o)*p_{oi}/(P(H_1)*p_{1i}),$$

and taking the logarithm (to any base) gives:

$$\log P(H_o/x_i)/P(H_1/x_i) = \log p_{oi}/p_{1i} + \log P(H_o)/P(H_1). \qquad (A1)$$

This is the log of the posterior odds ratio or equivalently, the log of the posterior likelihood ratio. It can be rearranged to get:

$$\log p_{oi}/p_{1i} = \log P(H_o/x_i)/P(H_1/x_i) - \log P(H_o)/P(H_1). \qquad (A2)$$

This number is the difference between the log of the posterior odds ratio and the log of the prior odds ratio. Thus, it provides a measure of the information for discriminating in favor of H_o against H_1 which was gained by observing the random variable x_i.

For this reason, the information gained by the set of outcomes of the attribute comparison, x_i, is defined to be:

$$I(o:1;x_i) = \log p_{oi}/p_{1i}. \qquad (A3)$$

THE MEAN INFORMATION

The mean information for discriminating in favor of H_o against H_1 is the expected value of $I(o:1;x_1)$ under H_o, or

$$I(0:1) = E_o(\log p_{oi}/p_{1i})$$

$$= \sum_{i=1}^{n} p_{oi} * \log p_{oi}/p_{1i}. \qquad (A4)$$

Here E_o represents the expectation under H_o. Note that the mean information is simply the expected value of the log of the likelihood ratio under H_o.

One useful mathematical fact is that I(o:1) is always greater than or equal to zero, with equality only when $p_{oi} = p_{1i}$ for all i = 1, ..., n. This gives an approach to selecting between the two hypotheses. Given any sample, it is possible to evaluate the sampling distribution under both hypotheses, and to calculate the mean information between the sampling distribution and the hypothesized distribution. The hypothesized distribution which was closer to the sampling distribution, as measured by the mean information, would be preferred.

THE TEST STATISTIC

When we compare the attributes associated with any two records, the result is one of the n possible values taken by x_i. We denote this observed random variable as x^*. The probability of observing $x^*=x_i$ is p_{oi} under H_o and p_{1i} under H_1. Thus, the sampling distribution of x^* is simply;

$p_i=1$ if $x^* = x_i$, $p_i=0$ if x^* ne x_i.

We can write the mean information between the sampling distribution and H_o as

$$I(x^*:H_o) = \log(1/p_{oi}) \text{ for } x^*=x_i,$$

and the mean information between the sampling distribution and H_1 as

$$I(x^*:H_1) = \log(1/p_{1i}) \text{ for } x^*=x_i.$$

The decision rule, as described in Kullback (1968, chapter 5), is to pick the hypothesis which has the smallest mean information relative to the sampling distribution. That is, we accept the hypothesized distribution which is closest to the sampling distribution.

Thus, the procedure would be to accept H_o if $I(x^*:H_1)-I(x^*:H_o)$ is positive (or "sufficiently large.") and accept H_1 if it is negative (or "sufficiently small.")

This yields the test statistic, $T(x^*)$, where

$$T(x^*) = I(x^*:H_1)-I(x^*:H_o)$$

$$= \log(p_{oi}/p_{1i}) \text{ for } x^*=x_i. \qquad \text{(A5)}$$

$T(x^*)$ is the log of the ratio of the probability of the set of outcomes, x^*, under H_o to the probability of x^* under H_1. Note that if these probabilities are the same then $T(x^*)=0$, and this set of outcomes has no discriminating power for identifying whether records represent the same unit. If p_{oi} is larger than p_{1i}, then $T(x^*)$ will be positive for that category. The larger $T(x^*)$, the stronger is the possibility that observation of this set of outcomes indicates that the records represent the same unit. If p_{oi} is smaller than p_{1i}, then $T(x^*)$ is negative. The smaller $T(x^*)$, the stronger is the possibility that this set of outcomes indicates that the records do not represent the same unit.

Since $T(x^*) = \log(p_{oi}/p_{1i})$ with probability p_{oi} under H_o, and with probability p_{1i} under H_1, the ratio of the probability that $x^*=x_i$ and the probability that $T(x^*) = T(x_i)$ is equal to 1.

Since the ratio of the probability function of x_i and the probability function of $T(x_i)$ does not depend on the p_{oi} or p_{1i}, $T(x_i)$ is a sufficient statistic for discriminating between H_o and H_1.