

A Review of the Statistics of Record Linkage for Genealogical Research

As Used for the Family History Library,
Church of Jesus Christ of Latter-Day Saints

*David White, Utah State University and
Church of Jesus Christ of Latter-Day Saints*

Introduction

The Church of Jesus Christ of Latter Day Saints maintains massive genealogical files, which consist of millions of names. The two largest files are the International Genealogical Index (IGI) and the Ancestral File.

The IGI contains over 200 million individual vital records and, because of its size, is divided into geographical subfiles. The Ancestral File contains over 21 million names arranged in family groups and pedigrees. These files are growing, and one of the major challenges is to be able to query these files in such a way that correct records are retrieved for various genealogical purposes and adding duplicates to these files is avoided. Record linkage is used for this purpose.

For this paper, a *record* will be defined as the collection of items that refer to a specific event, such as a birth or christening. Each item for the event, such as the day, month, year of birth, surname, and given names of the father and mother, is stored in what are called *fields* in computing terminology. Two records are defined as “linked” if the odds are high that they represent the same person. One of the first challenges for record linkage is finding those fields that are useful for calculating these odds. Although all the records in the IGI are birth and marriage events, they come from various sources. The same is true for the Ancestral File. For example, a birth record for the same individual may come from a civil record, an ecclesiastical record, or a family source. This may result in multiple records in the file for the same person when the available information is sparse or varies.

Comparing a Pair of Records (Calculating the Odds)

We begin with statements about the comparison of field entries coming from two records when it is known that these records refer to the same person. Such records are termed “matched” or “duplicates” by researchers. The records may or may not be from the same source. An example of different sources containing records about the same person would be births coming from civil records and ecclesiastical records. An example where only one source is involved would be ecclesiastical records about the same person who has moved from one jurisdiction to another within the same denomination, and the record keeping agency includes both jurisdictions.

Next, consider birth records and, within a birth record, the field containing the given name of the mother. We consider a pair of birth records and desire evidence to either confirm or deny that these records represent the same person. Suppose the given name of the mother shows up in both records, and we have n pairs of such records, which are matched (i.e., each pair is known to refer to the same person). Further, suppose that in k instances, the mother’s given name for one record of the pair is the same as that for the other record. Then, the probability that the given names are the same when the records are matched is estimated by k/n , and we use the equation

$$P(S|M) \cong k/n \quad (1)$$

where \cong means that the two sides of the equation are “close,” although they may not be exactly equal. We read $P(S|M)$ as the probability that an entry in a specific field is the same for both members of a pair, given that we

have a matched pair.

Next, consider the probability that the given names of the mother are the same when the two records are randomly paired. Such pairs of records are termed “unlinkable” by Newcombe. As an example, suppose we have a file with a total of m different given names appearing for the mothers of the child. Then, a typical event describing a pair with the same given names is:

Both given names are Dorothy, or
 Both given names are Phyllis, or

 Both given names are Agnes,

where Agnes completes the total of the m given names appearing in the records.

First, assume that the records come from the same file and that there are N_1 records with the name Dorothy, N_2 records with the name Phyllis, and so on, to N_m for Agnes. If $N_1 + N_2 + \dots + N_m = N$, then the probability that one element of a pair is Dorothy will be estimated by N_1/N and that both elements are Dorothy will be estimated by $(N_1/N)^2$ -- we multiply the probabilities for the two elements together since the events are independent (any two records were randomly paired). Since we allow that both elements having the same given name can happen with any one of the m alternatives, we add the probabilities for the m possible given names together, to get

$$\begin{aligned}
 P(S) &\cong (N_1/N)^2 + (N_2/N)^2 + \dots + (N_m/N)^2 \\
 &\cong \sum_{j=1}^m (N_j/N)^2 .
 \end{aligned}
 \tag{2}$$

We read $P(S)$ as the probability that two corresponding elements of a pair are the same when the records have been selected at random.

Now, assume that the records come from different files, and that there are m first given names for mothers in common between the two files. Assume that there are L_1 records in the first file with the name Dorothy, and N_1 corresponding records in the second file, or in general, L_j and N_j records with the j^{th} given name.

Then, the probability that the given name of the mother will be the same is estimated by

$$\begin{aligned}
 P(S) &\cong (L_1N_1)/LN + (L_2N_2)/LN + \dots + (L_mN_m)/LN \\
 &\cong \sum_{j=1}^m (L_jN_j)/LN,
 \end{aligned}
 \tag{3}$$

where now,

$$L = \sum_j L_j \quad \text{and} \quad N = \sum_j N_j
 \tag{4}$$

and the summation of the subscript j is not limited to the m alternatives in common, since each file may have alternatives not in common with the other file.

We now have $P(S|M)$, the probability that two elements in a pair are the same, given that the pair is matched, and $P(S)$, the probability that they are the same when they have been paired randomly. For the rest of the paper, we will deal only with the case where the two elements come from the same file; the other case

corresponds in the same way as described above.

Another way to estimate $P(S)$ is to actually create a set of randomly matched pairs and calculate the proportion of matches obtained. This way is computationally less intensive and may be a practical alternative for people with more meager computational resources.

We next consider the probability law:

$$P(M|S)P(M) = P(S|M)P(S). \tag{5}$$

What we want is $P(M|S)$, which is the probability that two records of a pair do, in fact, represent the same person when the first given names of the mothers are the same. $P(S|M)$, which we have, is the probability that the elements of a pair are the same when they are matched. Using equation (5), we get

$$P(M|S) = [P(S|M)P(M)]/P(S). \tag{6}$$

This is an application of what is sometimes called “Bayes’ Rule,” being used more often in recent years and which has caused a good deal of controversy in the statistical community; it has been used successfully in Record Linkage.

Perhaps the pair of records does, in fact, represent the same person, even though the records of birth give a different given name for the mother. We then want $P(M|S^c)$, where S^c means the two elements in a pair do not agree. (S^c is read as the “complement” of S .) Then, the analogue of equation (6) gives

$$P(M|S^c) = [P(S^c|M)P(M)]/P(S^c). \quad \text{Further,} \tag{7}$$

$$P(S^c) = 1 - P(S) = 1 - [(N_1/N)^2 + (N_2/N)^2 + \dots + (N_m/N)^2], \text{ and} \tag{8}$$

$$P(S^c|M) = 1 - P(S|M) = 1 - k/n, \tag{9}$$

so that we get for the probability of a match when the elements are not the same,

$$P(M|S^c) = [(1 - k/n)P(M)]/\{1 - [(N_1/N)^2 + (N_2/N)^2 + \dots + (N_m/N)^2]\}. \tag{10}$$

Note that k and n are different from N_1, N_2, \dots, N_m or N . This is because they come from a sample of duplicates of size n , whereas N_1, N_2, \dots, N_m are the total numbers of records in the file for each of the names. Recall that k is the number of pairs of records in the set of duplicates (or matches) for which the given name of the mother is the same. $P(M)$ in equations (5), (6), (7), and (10) is the probability that two records “match” (represent the same person) when they have been paired at random. It will be very small.

Next, let E be the event describing whether the mothers’ given names are the same, not the same, or missing. Then,

$$P(M|E) \cong P(M) \text{ times } (k/n) / [(N_1/N)^2 + (N_2/N)^2 + \dots + (N_m/N)^2] \tag{11}$$

if the names are the same

and $P(M|E) \cong P(M) \text{ times } (1 - k/n) / \{1 - [(N_1/N)^2 + (N_2/N)^2 + \dots + (N_m/N)^2]\}$

if they are different.

We further define $P(M|E) = P(M)$ if one or both elements are missing in the record pair. This makes sense, since E tells us nothing new about the match when the information is missing.

Since virtually all records contain more than one element or “field,” we must allow for this in our formulas. We let Q be the number of elements or fields common to both records and consider the i^{th} field, where i ranges from 1 to Q . Then, let n_i be the number of elements with both entries present for the i^{th} field in the sample of duplicates, and k_i be the number of element pairs in the i^{th} field which are the same. Letting E_i be the event for the i^{th} field (same, not same, or missing), we get the following:

$$P(M|E_i) \cong P(M)P(E_i/M)/P(E_i) = P(M) \text{ times } (k_i/n_i)/[\sum_j(N_{ij}/N_i)^2], \tag{12}$$

when the i^{th} elements are the same and where the summation \sum_j is over all possible values of j in N_{ij}/N_i for the i^{th} element or field. Note that N_{ij} now has two subscripts, the first subscript (i) to account for the field and the second (j) to account for the alternative values for the i^{th} field. The range of j is from 1 to J_i because there are a different number of alternatives in each field. For example, there are two alternatives for gender and, in our case, m alternatives for the given name of the mother.

If the elements for the i^{th} field are not the same, the formula is:

$$P(M|E_i) \cong P(M)P(E_i/M)/P(E_i) = P(M) \text{ times } (1 - k_i/n_i)/\{1 - [\sum_j(N_{ij}/N_i)^2]\} \text{ and} \tag{13}$$

$P(M|E_i) = P(M)$ when one or both of the i^{th} elements are missing.

$$P(E_i|M)/P(E_i) \tag{14}$$

can be referred to as the “odds” in favor of a match, given the event E_i with respect to the i^{th} field. Note that $P(M)$ does not appear in (14). It does appear in (12) and (13), however, which are the probabilities that two records refer to the same person, given the event E_i for the i^{th} field. There are Q events for each pair of records (an event for each of the fields). Next, we consider

$$P(M|E_1, E_2, \dots, E_Q), \tag{15}$$

which is the probability that both members of the pair represent the same person when events E_1 and E_2 and ... E_Q have occurred. If most of the paired fields are the same, this probability will be close to 1, and we should conclude the pair is “linked,” as distinguished from the cases where they are known to have been matched by prior identification of duplicates. If most of the paired fields are not the same, (15) will be close to zero, and we conclude the records are not a match. There is a gray area in between where the evidence is not conclusive. We assume that the events E_1, \dots, E_Q are independent (that is, that one pair of fields being the same tells us nothing about the sameness of any other pair).

If we assume this, we get the formula:

$$\begin{aligned} P(M|E_1, E_2, \dots, E_Q) &= P(M|E_1)P(M|E_2) \dots P(M|E_Q) = \prod_{i=1}^Q P(M|E_i) \\ (16) \qquad \qquad \qquad &= \prod_{i=1}^Q P(E_i|M)P(M)/P(E_i) \end{aligned}$$

$$= P(M)[\prod_{i=1}^Q P(E_i|M)]/[\prod_{i=1}^Q P(E_i)].$$

(Note that $\prod_{i=1}^Q P(E_i|M)$ means to take the product of the $P(E_i|M)$ as the subscript j ranges from 1 to Q ; similarly for $\prod_{i=1}^Q P(E_i)$ in the above expression.)

The “odds” in favor of the records representing the same person are calculated as the probability of the above events when the records are matched, divided by the same probabilities when the records are randomly paired -- which is the last expression of (16), except that $P(M)$ would be dropped. Referring to the second page of NeSmith’s paper, the probability of the two names “matching by chance” is

$$\prod_{j=1}^Q P(E_j), \tag{17}$$

while the probability of the names being the same in the “truly linked records” is

$$\prod_{i=1}^Q P(E_i|M). \tag{18}$$

These are the two statistics used to calculate the odds. As before stated, this is (18) divided by (17).

Blocking

Finding the matches or duplicates (those pairs which are known to represent the same person) involves the time of experienced researchers who must consider a large sample of record pairs and find those which will be identified as duplicates. If all possible pairs are to be considered for the cases of interest, we will have an impossible task before us, with literally billions of pairs to evaluate. To cut down on the enormity of this task, we attempt to gather together records, which are likely to be matches, by sorting on fields, which will put potential matches close to each other in a listing of available records. Such fields usually include a surname code (such as Soundex), a given name code, and possibly a range for birthdates, and a county identification of some kind. If these four fields were used, a listing of records would put people together if they had the same surname code, given name within the surname code, birthdate range within the names, and the same county.

A *block* is defined as the set of records whose pairs are the same with respect to a set of fields, such as the above four fields. Each distinct set of fields used for this purpose is called a *blocking scheme*. The records whose blocking fields match will be adjacent to each other in a file, which has been indexed on the basis of these fields. Such a list can be constructed with any good data base management system. A block may, and probably will, contain a number of records, which are not duplicates; but a qualified researcher can browse the list and determine which of the pairs within the block should be considered as representing the same person. The size of the block should be modest -- not more than 10 to 20 records, so that the worker can compare them on a monitor screen. In order to find as many duplicates as possible, this process must be repeated for several blocking schemes. Even then, the number of blocks for a data bank may be too numerous to make searching all of them for duplicates feasible. Then, a subset is used, such as some representative date ranges. One of the problems of interest is how large the sample of duplicates obtained by the workers should be. Current practice is to find about 1,000 to 1,500 duplicates -- a substantial amount of work.

One blocking scheme will often have better properties than another. Measures of how good a scheme is include:

- **Blocking Recall.** -- It often happens that when a second blocking scheme is used, there will be a few of the duplicate pairs found in the first scheme, which will now be separated; that is, the two members of the pair will not show up in the same block. Since our searching procedures only look for record matches within the same block, such duplicates will not be detected using the second scheme. If we

now consider several blocking schemes, the percentage of known duplicate pairs, which are picked up with any one of the schemes, may well be less than 100%. Hopefully, we will find one of them, which picks up a higher percentage of duplicate pairs than do the others. *Blocking recall* is defined as the percentage of known duplicates, which are identified with a particular blocking scheme.

- **Block Noise.** -- This is the number of non-duplicates in the blocks divided by the total of the block sizes in the blocking scheme. Greater block noise requires more computing time for a search, but recall for the scheme is usually better.
- **Block Precision.** -- For a blocking scheme, this is the ratio of the number of duplicates to the number of non-duplicates in the blocks, multiplied by 100. A blocking scheme with high precision has mostly duplicate pairs within the block. There are not many non-duplicates.

The greater the precision, the less recall, as a general rule, as indicated on the third page of NeSmith (1994). Her comment about increasing recall without seriously reducing precision relates to the use of a name code, such as Soundex, and a place code, which different versions of place names are tied to. This can be considered as a partial agreement for the fields concerned, and the blocks that use these fields will be somewhat larger, including proper and/or place names, which are “close” to each other. We decrease the block noise and increase the block precision by increasing the number of fields used for blocking. Fewer fields, conversely, increase both noise and recall.

Calculating the Weights

The weights are obtained from the odds by taking logarithms. Using equation (16), we take logarithms, to get

$$\log P(M|E_1, E_2, \dots, E_Q) = \log P(M) + \sum_{i=1}^Q \log \{P(E_i|M)/P(E_i)\}. \quad (19)$$

Note that $P(M)$ is a constant term, which factors out of (16), and is simply an additive constant in (19). Such a constant does not influence the results. We can drop this constant and simply consider the term

$$L = \sum_{i=1}^Q \log \{P(E_i|M)/P(E_i)\}. \quad (20)$$

The w_i weights referred to in the NeSmith paper are the individual terms

$$w_i = \log \{P(E_i|M)/P(E_i)\}. \quad (21)$$

For each field, there are three weights -- one for when the two field entries are the same, one for when they are not, and zero for when one or both entries are missing. This weight will be a positive value when the two fields are the same; it will tend to be negative if the entries for the two fields are different.

If the probability is high that the records are for the same person, then most of the E_i will be in agreement (the i^{th} elements are the same for most of the i), and the sum of the weights (sum of the “log-odds”) will be high, usually positive. If the probability is low, then most of the E_i will not be in agreement, and the sum of the weights will be low, usually negative. If data are missing in the i^{th} field, so that $P(M|E_i) = P(M)$, then from (12), $P(E_i|M)/P(E_i)=1$ and $\log[P(E_i|M)/P(E_i)]=0$. That is, the weight is zero if one or both field elements are missing. Fields with missing elements, therefore, neither add to nor subtract from the evidence we are interested in.

The researcher must identify each of the pairs in the subset of blocks as either a duplicate or a non-duplicate, and the weights are calculated from the duplicate pairs in the blocks, plus a set of counts (the

N_{ij}) for each field. The counts do not come from the blocks but from the complete set of records to be linked – that is, from the entire file. Note that as per the comment in NeSmith on the fifth page, **the weights are not calculated for the fields used as blocks because those fields always are the same for the records in the blocks, whether duplicates or not, and thus have little or no discriminating power.** For the same kind of reason, some fields have poor discriminating power because they do not change a great deal in some files – such as a geographical area in which a few family names predominate. A small amount of variability in a field reduces its usefulness for linkage algorithms. Algebraically, this shows up in the denominator of (18)/(17), above, because the chance of the fields being the same with random pairing becomes larger. But this is (17), which thus decreases the odds for a link when the field entries are the same.

Thresholds

Many genealogical tasks involve a search for someone in a large data bank. This search is usually termed a “query,” and the framework for this is the set of linkage algorithms described above. One uses the fields available for the person to be searched for, chooses a blocking scheme employing some of those fields, and then searches the block into which the person being searched for fits. If a record in the block, when paired with the query, has the sum of the weights higher than a value called the *threshold*, a link has been found for the query. High recall for a blocking scheme means that there is a good chance of finding this link if it exists---but since high recall goes with more “noise,” more computing time is involved. If a large number of queries are involved, the computing time may become an important issue.

The *threshold* is simply a constant value, C , which is a cutoff point for L , the sum of the log-odds for a pair. We consider the pair as representing the same person if L is greater than or equal to C ; that is, the pair is “linked.” The pair is not linked (i.e., considered as representing different people) if L is less than C . As an illustration for thresholds, we consider five sets of date ranges, which were used with Norway data (1736- 1755), (1781-1794), (1805-1814), (1836-1845), (1866-1875). These subsets of the complete set of data were used because finding duplicates for the complete set would have been too time-consuming. Several blocking schemes were used for identifying duplicates. For each blocking scheme and for each block in the scheme, all possible pairs were obtained, and the worker identified each pair as either a match (i.e., a duplicate) or a non-match. The scheme chosen as best for linking on the basis of precision and recall used the fields: birth year, birth county code (to standardize county names), the given name code for the principal (whose birth is recorded), and the father’s given name code. A block consisted of all records, which were the same for all four of the above fields. The fields used for weighting purposes were:

- the latitude minutes of the birth town
- the birth day
- the birth month
- the death day
- the death month
- the death year
- the mother’s given name code, and
- the mother’s surname code.

The N_{ij} were obtained with the computer from all records in the Norway File. The weights were then calculated with computing facilities for each of the eight fields, according to the formulas described in the preceding section and using the duplicates identified by the researchers. Next, **for each pair of records within each block (duplicates or not) and for each field in the pair**, the weights are then used. The total of the weights is obtained to get the value for the sum of the log-odds (see equation (17)). We now have a set of weight totals for the duplicates (matched pairs) and another set for the non-duplicates (unmatched pairs). A frequency histogram was obtained for both the matched and the unmatched pairs; these appear in Table 1.

Table 1. -- Frequency Distributions for Matched and Unmatched Pairs

Class Limits for Weight Totals	Unmatched Pairs	Matched Pairs
-34.35 to 0	-27.56	0
-27.55 to 0	-20.76	6
-20.75 to -13.96	257	0
-13.95 to -7.16	602	1
-7.15 to -0.36	381	33
-0.35 to 6.44	58	255
6.45 to 13.24	2	540
13.25 to 20.04	0	344
20.05 to 26.84	0	15
26.85 to 33.64	0	19
33.65 to 40.44	0	13
40.45 to 47.24	0	0

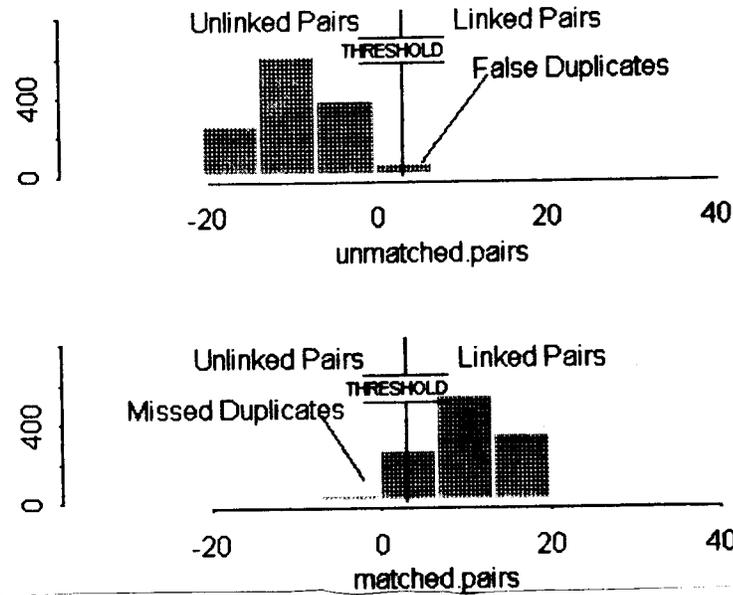
Minimizing False Duplicates

It will be noted that the scores for the unmatched pairs are consistently lower than for the matched pairs, but that occasionally, the scores for the unmatched pairs will be higher than some of the scores for the matched pairs. Figure 1, below, provides graphs for both histograms. The vertical line for both distributions represents the threshold, or value above which a pair will be linked (i.e., considered as representing the same person). In the top graph, consisting of the unmatched pairs, there are a few weight sums for pairs, which fall above the threshold and thus will be “linked” (i.e., considered as representing the same person, even though the pair was judged by the worker to represent different people). These are the “false duplicates.”

Minimizing Missed Duplicates

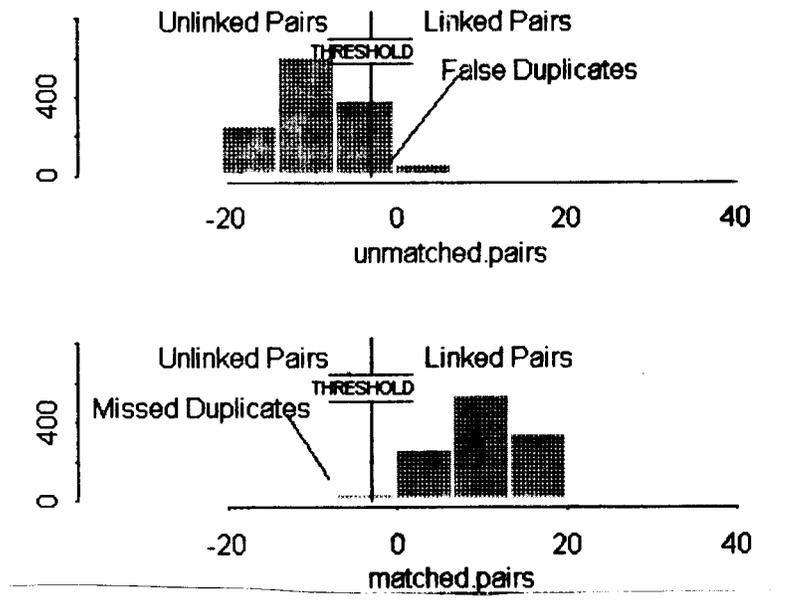
Now, consider the distribution of duplicates or Amatched pairs@ in the lower graph of Figure 1. Notice that with the threshold illustrated, there is a substantial proportion of duplicates that will be unlinked (i.e., considered as representing different people). Rather few of the unmatched pairs will be considered as matched (i.e., will be linked); but substantially more of the duplicates (matched pairs) will fail to be linked. These are the “missed duplicates.” The higher proportion of these is due to minimizing the false duplicate error.

Figure 1. – Threshold which Minimizes False Duplicates



It may be that we consider the “missed duplicate” problem as more serious than the “false duplicate” issue. In this case, we can minimize the missed duplicates by moving the threshold to the left, as in Figure 2. Here, the false duplicate rate (the proportion of non-duplicates which are linked) is now larger than that for the missed duplicates.

Figure 2. – Threshold which Minimizes Missed Duplicates



We have now considered two kinds of errors:

- We can fail to identify a genuine match because our “linking” algorithm did not give the sum of the

weights above the threshold (missed duplicates).

- We can “link” a non-duplicate because our algorithm gave the sum of the weights above the threshold (false duplicates).

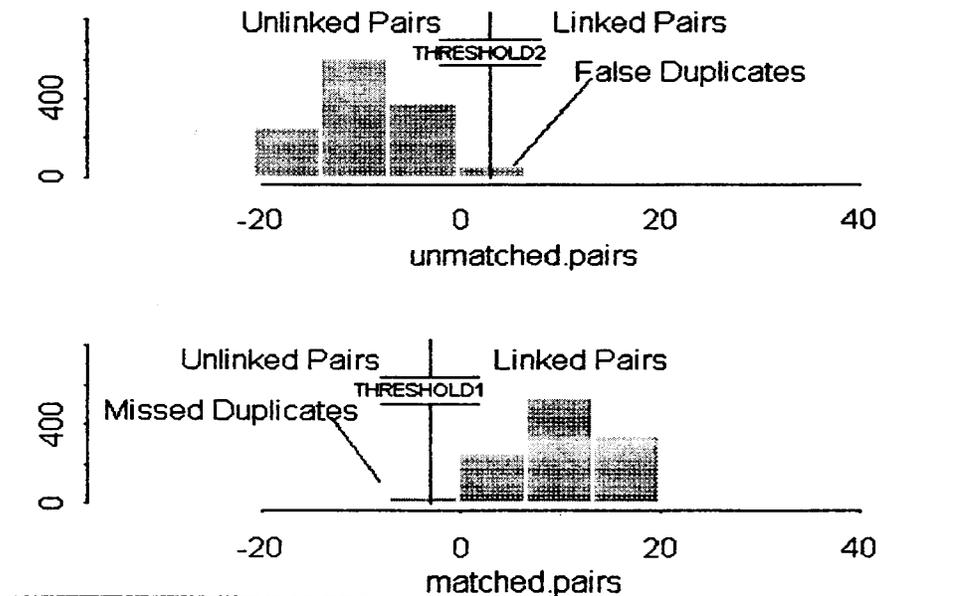
Table 2 gives both errors for ten alternative threshold values. Note that increasing the threshold value decreases false duplicates but increases the percentage of missed duplicates.

Threshold Value for Sum of Log=Odds	% False Duplicates in Nonmatched Sample	% Missed Duplicates in Matched Sample
-8.81	26.85	0.00
-7.03	26.52	0.08
-5.25	26.38	0.08
-3.48	7.66	2.21
-1.70	5.11	2.62
0.06	4.66	2.78
1.84	4.66	2.78
3.61	1.54	16.47
5.39	0.21	23.60
7.16 7.16	0.21	23.93

Now consider Figure 3. If, now, we identify *non-links* as those to the left of the lower threshold (THRESHOLD 1) and *links* as those to the right of the upper threshold (THRESHOLD 2), we have a small error rate for both decisions -- but now, we have a new problem.

There is a “gray” area between THRESHOLD 1 and THRESHOLD 2 where there is no rule on how to make a decision. If the pairs in the gray area need to be inspected manually in order to make a decision, this becomes a task of prohibitive magnitude with large files. If one does not need to make a decision with every pair, the use of two thresholds may be the best alternative. NeSmith notes on the final page of her paper that the purpose of linking needs to be considered when setting the threshold. If missed duplicates are the most serious risk, then a lower threshold as in Figure 2 would be preferred. This would make sense for genealogical queries where the failure to find a genuine link could not be compensated for, while a false duplicate would ordinarily be easy to detect on examination. If a large file is to be cleaned up, however, a false duplicate might be more serious, since merging two individuals would then lose information on one of them. A duplicate of an individual would also be a problem, but possibly less serious, and the higher threshold of Figure 1 might be better. If the file were small enough, then cleaning it up might be best with the two thresholds of Figure 3, with manual inspection of the pairs whose linkage scores fell in the gray area.

Figure 3. – Using Two Thresholds to Control Both Kinds of Errors



Summary

The procedure has several main phases:

- Select a file (or set of files) in which to identify duplicates.
- Pick fields for ordering the records to put likely duplicates close together, using a data base management system with “browsing” capacity (blocking).
- Manually identify between 1,000 and 1,500 duplicate pairs.
- Use the duplicate pairs and the preceding formulas to construct weights for all fields, except those used for blocking.
- Select one or two thresholds to use for “linking” pairs of records as estimated duplicates. The position of the threshold or thresholds depends on the desired type and size of the error rates (see Figures 1-3).
- Merge records which have been linked, allowing storage space for possible conflicts. If the entries for a specific field do not match, both entries should be stored, so that a genealogical researcher using the data bank can evaluate them both.
- Use these algorithms to identify duplicates when records are added to the file, and when queries are being made.

This type of project can be repeated with many different geographical areas: the problems and sets of

weights appropriate for use with patronymics will be much different than those associated with the U.S. and Canada. There are many refinements, which need to be investigated, including the use of value-specific techniques, partial agreements, lack of independence between field entries, and the use of other statistical procedures to enhance current techniques.

Note

David White is professor emeritus at Utah State University, Department of Mathematics, Logan, Utah 84322 and has been statistical consultant to the Record Linkage Team, Family History Department, Church of Jesus Christ of Latter-Day Saints.

References

- Baldwin, J. A.; Acheson, E. D.; and Graham, W.J. (Eds.) (1987). *Textbook of Medical Record Linkage*, New York: Oxford University Press.
- NeSmith, Nancy P. (1992). Record Linkage and Genealogical Files, *Genealogical Journal*, 20, 3-4, 113-119.
- Newcombe, H. B. (1988). *Handbook of Record Linkage*, New York: Oxford University Press.

This paper is designed as a companion to "Record Linkage and Genealogical Files," by Nancy P. NeSmith (in this volume) and parallels as much as possible the description of record linkage given there with the formulas used to put the theory into practice. The material included here is not in any sense original but derives from the work of H. B. Newcombe (1988) and researchers, such as those included in Baldwin, Acheson, and Graham (1987), who have been working in this area, primarily from the decade of the 1960's and after.

If there are any questions relative to the material of these papers, please contact Ms. NeSmith at the address given in her paper, or Dr. White, with respect to the statistics.