CURRENT RECORD LINKAGE RESEARCH

Matthew Jaro, U.S. Bureau of the Census

This paper discusses problems involved in the design and implementation of record linkage algorithms for file matching under conditions of uncertainty. Current research activities in this area are summarized, along with a brief survey of some underlying theoretical considerations. This paper stresses techniques that might be used for obtaining confidence in the match decision and algorithm validation. The research being conducted for the 1985 pretest in Tampa, Florida is discussed.

1. SUMMARY OF RESEARCH ACTIVITIES

Record linkage is the process of examining two computer files and locating pairs of records (one from each file) that agree (not necessarily exactly) on some combination of identifiers (or fields). For the Census Bureau this process is typically executed on two files containing individual names, addresses and demographic characteristics. Specifically, record linkage is important for census undercount determination, address list compilation and general census evaluation.

Record linkage research is focused on the development of an algorithm and accompanying manual procedures that will accomplish the above goals in a statistically justifiable manner. To this end the following major activities must be initiated:

A. development of a statistical foundation for the record linkage process;
B. construction of a data base that can be used for calibration, validation and testing of the characteristics of the linkage process;
C. development of methods to obtain information on the discriminating power of the various identifiers and their associated error rates (discriminating power is a measure of an identifier's usefulness in predicting true match pairs); and
D. design and implementation of computer algorithms to perform the actual linking.

The results of this research will be:

A. more accurate undercount determination and coverage analysis;
B. reduction of costly clerical procedures by use of automated methods;
C. a statistically valid process which can replace previous ad hoc techniques; and
D. algorithms that will be useful for overcoverage determination and address list compilation.

2. AREAS OF INVESTIGATION

There are several areas of investigation that must be pursued in order to design and implement a successful matching system. These areas are currently the focus of attention for the Record Linkage Research Staff.

2.1 Blocking and Other Search Restricting Techniques

The set of records that will be examined to find a match for a given record is called a block. Obviously, if an entire file were searched for a match for each record, the probability of finding a true match would be highest, since no records are excluded from consideration. However, the cost of such a process would be prohibitive. As we restrict our search, we exclude records and increase the probability that the "true match" record would be excluded--but the cost of searching decreases.

The ideal blocking identifier would be one which nearly always agrees in "true match" record pairs but nearly always disagrees between pairs which are not valid matches. This ideal blocking identifier must have a large enough number of possible values to insure that the file will be partitioned into many (and therefore smaller) blocks. R. Patrick Kelley of our staff has developed a method for computing an optimal blocking strategy, considering the tradeoffs of computation cost against errors introduced by restricting the search for matches. See [4].

2.2 Weights

The terms "identifier" or "component" represent fields on a computer file (and are used interchangeably). Typical components are street name, street type (e.g., Street, Avenue, etc.) surname, given name, etc. The discriminating power of a component (or identifier) is a measure of how useful that component is in predicting a match. Consider a component such as surname. Common values of surname (such as "Smith") have greater chances of accidental agreement than do rare values (such as "Humperdinck"). Consequently, the frequency of occurrence of a particular value of an identifier is one determinant of the weight or importance of that value as an indicator of matched or unmatched records. Another determinant of the weight is the error rate associated with the value of that component. High error rates diminish the predictive usefulness of an identifier or its values.

Fellegi and Sunter, in [1], presented a general theory of record linkage, including discussions of weight calculations and the development of optimal decision rules. Their basic idea for weighting is summarized below.

The two files (A and B) to be linked consist of a number of components (identifiers) in common. Consider all possible pairs of records. A particular pair is either truly a matched pair (an element in the set M of all matched pairs) or an unmatched pair (an element in the set U of all unmatched pairs).

For all pairs (p) and each component (or component-value state) i let:

$$m_i = \Pr(\text{component agrees} \mid p \in M)$$
$$u_i = \Pr(\text{component agrees} \mid p \in U).$$

Weight for the ith component = $\log_2 (m_i/u_i)$.

The above computation would be the same if we were considering specific values of components (such as "Smith" or "Humperdinck") rather than the component as a whole (surname). Similar weights can be computed for disagreements. m_i is computed by examining all matched pairs; u_i is computed by examining all unmatched pairs. For the two files A and B,

$$\{U\} = \{A \times B\} - \{M\} .$$

Since the cartesian product A x B is $O(n^2)$ and M is O(n) (where n is the number of records in the smaller file), then { U } is much greater than { M } and the u_i can be computed by taking the frequency counts of the components in both files.

The calculation of m requires a prelinked set of records M. This fact presents the greatest practical difficulty because of the large sample size necessary, the cost of producing such samples and the inherent error in manual processes.

Fellegi and Sunter, in [1], suggest a method of weight calculation that does not require prelinked pairs. It uses an assumption of the statistical independence of the components and requires the solution of a non-linear system of equations. We plan to investigate the use of this method, which to our knowledge has never been tested.

Another method of weight calculation that we will consider is that of iterative refinement. We propose this method to avoid the construction of costly samples. If there were no errors in a given component, the value "m" for that component would be 1 and the weight for the component could be calculated from the frequency of occurrence of the component value states.

These initial weights can be refined as follows: Whenever a record pair disagrees on a component, that pair would be presented to an operator by the matching program. The operator can then make a decision as to whether the pair is a match or not. This places the pair in either the set M or U and the weights can now be updated (since m is now less than 1 -- because of the detected error -- if this pair is placed in {M}).

The program can obtain information regarding the error rates of each component in this manner, updating the probability as records are processed. The operator supplies the "truth" regarding each record in question (does this pair belong to set { M } or to set { U } ?). This teaches the program to make similar decisions to those of the operator.

The operator can set the level of errors that will control the display of candidate record pairs. In this way, records can be matched automatically despite small errors in components. As confidence is gained, the thresholds for manual intervention can be moved. After all records have been processed, the entire file can be rematched using the new weights and the process can be continued until consecutive iterations produce small differences.

An investigation into this technique is required to determine whether such iterations will converge to a stable set of weights and to determine the amount of bias introduced by such estimation techniques.

A third method of weight calculation that might be explored would involve automatically making the "M" or "U" decisions, instead of relying on human operators. This would be accomplished by considering pairs of records that match on all fields except a specified number. Those pairs could be assigned a match status if the composite weight (Σw_i) for the pair was sufficiently greater than the cut-off threshold. The distance from the cut-off would leave room for weight estimation error without effecting the "M" or "U" decision, and hence, the "M" decision could be made automatically with some degree of confidence. These cases would be used to tabulate the error rate probabilities.

Since the cut-off threshold for a match decision is dependent upon the weights of each field, this threshold would move as weights are revised. The effect of this concomitant variation on the weight estimation must be investigated.

2.3 Composite Weights

If the components are assumed to be statistically independent, then the composite weight is equal to the sum of the individual component weights. Adding the weights is equivalent to multiplying the conditional probabilities. Weights for disagreements can be computed similarly to weights for agreement. Disagreements are generally given negative weights, whereas agreements receive positive weights.

We know that some dependencies exist (such as sex and given name) but the extent to which dependence changes the matching decision rules must be analyzed. For example, "Robert" is principally a male given name, but "Stacy" could be either male or female. Such dependencies could have an effect on the probabilities of agreement given unmatched pairs. If the errors in the fields are dependent, then the probabilities of agreement given matched pairs could change. The disagreement weights would also change proportionally.

We are currently designing simulation experiments to study the effect of covariance on the decision results. It is hoped that a regression analysis will provide information concerning this relationship after a number of runs with differing covariance configurations.

2.4 Error Rates

If a plot were to be made of numbers of observations versus composite weight, a bi-modal distribution would result. Since most pairs are elements of { U } , the disagreement mode is much larger than that for agreement.

For each pair, one of three decisions is made. The pair is said to match if the weight is greater than a threshold μ, or not to match if the weight is less than a second, lower threshold λ . Pairs having weights between these thresholds are classed in the "don't know" category. These pairs must be followed-up using a computer-assisted manual approach.

Once the thresholds are set, bounds on the probabilities of false matches and false nonmatches can be computed by integrating the portions of the distribution tails lying beyond the threshold values. By tabulating weights of candidate pairs, the matcher program could provide information on the error rates associated with the component values. These error rates are useful for verification. The success of this technique will depend upon our ability to fit a curve to the observed tails of each mode in order to perform the integration.

2.5 Component Values

The matcher algorithm will use a table of weights derived from investigations on weight methodologies (see 2.2). One weight would be associated with each predetermined component or identifier value. The algorithm would store the most frequent values of components from tables prepared by other programs and component values not in this list would be given a relatively high weight. Thus, popular names (which have low discriminating power) would receive lower weights than comparatively rare ones, without requiring the construction of exhaustive lexicons. Value tables would only be used if successful results could not be obtained by considering a component to have a single weight.

The weight tables for the program will include expected frequencies of occurrence of component values, error rate information and number of records processed for past data. Information from the current data could be used to update the weight tables as the program gains experience matching.

2.6 Bayesian Adjustment

In addition to keeping records of expected frequencies (based on earlier observed frequencies), the program will also keep observed frequencies of a block for a specific file. If there is much deviation between observed and expected frequencies, temporary modification to the weights can be considered. For example, in a Spanish-speaking area, the name "GONZALEZ" might occur relatively more frequently than it does on the average for the United States.

Missing data values could also result in the reduction of discriminating power of a field within a block.

We have incorporated a Bayesian adjustment technique into our experimental matcher. We have assumed a Beta prior distribution and are investigating parameter estimation techniques for this distribution.

2.7 Distance Metrics

Simple agreement/disagreement patterns of component pairs are not adequate for character strings and numeric data. We are investigating prorating the weight on the basis of degree of agreement.

A number of character-string comparison routines for component values which do not agree completely are available, including the routine designed by Jaro and Corbett, which has been used for 12 years in the UNIMATCH system [3]. Through the use of such a routine, words can be matched despite spelling errors. The UNIMATCH algorithm is an information-theoretic comparator which takes into account phonetic errors, transpositions of characters and random insertion, replacement and deletion of characters. These approaches will be tested in the matcher algorithm.

2.8 Assignment

After blocking, the program uses the various techniques described above to construct a composite weight for each pair in the block. These weights are stored in a cost matrix and the assignments can be made by solving the problem:

$$\text{Maximize} \quad Z = \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} X_{ij}$$

Subject to

$$\sum_{j=1}^{n} X_{ij} = 1 \qquad i=1,2,\ldots,n$$

$$\sum_{i=1}^{n} X_{ij} = 1 \qquad j=1,2,\ldots,n$$

where C_{ij} is the cost (weight) of matching record i with record j. X is an indicator variable. The matrix is made square by the use of dummy weights.

This problem is the linear sum assignment problem, which is a degenerate transportation problem that can be solved efficiently using only additions and subtractions. Once an optional assignment set is obtained, the Fellegi-Sunter decision procedure is applied to determine whether an assignment represents a match, a clerical review case or a non-match.

3. MATCHER IMPLEMENTATION PLANS

An experimental program has been implemented that incorporates the techniques discussed in this paper so that controlled tests can be conducted without undue difficulty. This program is operating on an IBM Personal Computer.

For production matching it is anticipated that not more than two passes will be required to match nearly all records not requiring professional review. Records failing to match on blocking components in the first pass would have a second chance to match on different blocking components during a second pass. By selecting two high discrimination/low error rate sets for blocking, the probability of intersecting errors is minimized. The high discrimination/low error rate property for a component means there is a high probability that the component can accurately predict a matching record pair. By using two such components, the chance of a successful match is relatively good, since errors on both components would be required to reject a record.

We plan to utilize experience gained by Statistics Canada (the Generalized Iterative Record Linkage System [2]) and others in our investigation into the problems of record linkage. It is our intent to have an operational program for use with the 1985 Census pretest. One of the most important applications will be coverage evaluation for the Decennial Census.

REFERENCES

[1] Fellegi, I.P. and Sunter, A.B., A Theory for Record Linkage, Journal of the American Statistical Association, Vol 64, 1969 pp 1183-1210

[2] Generalized Iterative Record Linkage Systems (GIRLS), Institutional & Agriculture, Survey Methods Division, Statistics Canada, Internal Documentation, Oct. 1978

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

[4] Kelley, R. Patrick. Blocking Considerations for Record Linkage Under Conditions of Uncertainty. Proceedings of the American Statistical Association, Social Statistics Section, Philadelpia, 1984, pp. 602-605. (Sections 3 & 4 were prepared with the assistance of D. Childers.)