History of Matching

Prior to 1994, the education studies used various approaches to conduct matching between the survey data and external publicly available data files. Numerous homegrown methods were developed that attempted to handle the following:

- Variables that are exact matching (such as region, state);
- Variables that are continuous and may be changeable (enrollment, number of teachers);
- Variables with missing data;
- Derived variables;
- Weights

The key to ensuring that data released to the public do not infringe on the privacy of the respondents is to design a statistical approach to compare data from the survey data to the publicly available data to see if any school or institution can be identified. Early methodologies to matching (Euclidean distance, etc.) had limited success in identifying disclosure risk of schools.

In 1995, the IES DRB approved Westat's first use of the probabilistic matching procedures and specific software (AutoMatch) for institution-based surveys when external public files are present for matching. Matching to the external files is conducted to identify potential disclosure risk cases. The focus of the procedures in this process is to determine whether the variables selected for inclusion on the file—namely, the background variables—can be used to reconstruct a school that can be matched back to the CCD or PSS. If schools cannot be identified, then no specific data changes should be required. If schools can be identified through matching (e.g., "Rule of 3"), normally, deterministic swapping would be used to remedy disclosure risk schools but since external files exist, then various data elements would be tested for deletion or further coarsening to ensure that the school could not be identified. The analytical value of the data must be preserved while maintaining data confidentiality.

The IES confidentiality standards generally require both deterministic and, starting around 1999, controlled random data swapping, to prevent the disclosure of participating schools. If a match is identified, a deterministic swapping approach is used to accomplish data confidentiality. A non-targeted approach, such as random swapping, cannot accomplish the assurance that the released data are confidentialized. Deterministic data swapping is used for schools that are identified as high risk when the school data is probabilistically matched to external school files (i.e., the CCD and PSS). Controlled random data swapping of elements within the school (and other background questionnaire data files) adds a measure of uncertainty to school, student (when applicable), and teacher identification. The *DataSwap* software package was reviewed and approved by NCES to satisfy the controlled random data swapping requirement for IES data dissemination of both public- and restricted-use data. The focus of this section is to describe the method for satisfying the

deterministic data swapping requirement. For the IES survey restricted-use release, only controlled random swapping is implemented. No matching procedures are conducted.

Newer, more Robust Probabilistic Matching Software Replaces AutoMatch

The matching software, as most software technology, has improved which makes public-use data even more vulnerable to disclosure risk. In 2014, the Fine-grained Record Integration and Linkage (FRIL) – was examined, and approved by the IES DRB as a suitable software solution for probabilistic matching (replacing the antiquated AutoMatch software). FRIL had been a universally available software package distributed in the public domain by the Centers for Disease Control (CDC). Although it is no longer supported by the CDC, it remains a viable software package. It provides the functionality required, is reasonably simple to use, and provides a useful documentation through both a User's Guide and Tutorial. The comprehensive FRIL tutorial takes the reader step-by-step through the general architecture of the FRIL system as well as how to bring up the application, define data sources, and specify which variables to add to a comparison vector. The probabilistic functionality of FRIL parallels the key procedures used previously in AutoMatch. Probabilistic matching using sophisticated software such as FRIL is better suited to meet the needs of the DRB than various deterministic and/or Euclidean distance approaches.

The FRIL tool offers several features that make it especially appropriate for disclosure-proofing of public use tables, microdata, and other quantitative data releases that carry with them the risk of exposure of confidential or private information. FRIL offers options for linking on numeric as well as character data. The methods for linking numeric data fields support transformations and approximations. In tables that contain counts such as number of students or number of teachers, for example, the Distance Metric options for numeric variables include ranges around a value in levels and percentages. Lower and upper ends of ranges do not have to be the same. This non-exact matching rule is sometimes called "fuzzy" matching. It helps reduce the impact of small variations and errors in the recording of data, improving the sensitivity of linkage. When multiple variables are matched this way, it reduces the chances of a false match by controlling the specificity of linkage. Fuzzy matches of multiple variables offset the tendency of imprecise matches on a single variable to produce false matches. To evaluate matches of character variables, FRIL offers special methods for names, postal codes, and similarity of strings. For example, the Jaro-Winkler function computes a value of a distance metric that approaches 0 when strings have neither content nor order in common, and has a value of 1 for identical string values.

A "probabilistic" linkage approach can be used to calculate the likelihood of a correct match between the school represented on the school records on the IES SURVEY file and school records on the CCD, and PSS files. Computerized probabilistic record linkage methodology was first shown to be feasible in 1959 by Howard B. Newcombe's research at Canadian Atomic Energy Chalk River Laboratories. A decade later, Fellegi and Sunter (1969) developed what has become a widely accepted mathematical theory of record linkage. With this method, the comparison algorithm calculates a weight for each record pair that indicates the likelihood that a record pair relates to the same entity – in this case, e.g., school. The general approach to determine matched pairs is to

calculate the odds of a match based on the reliability and discriminating power of the variables used in the comparison. Then those pairs with odds above a specified critical level are declared matched pairs, and those pairs with odds below are treated as unmatched pairs.

In applying this technology to the present problem, the general approach taken was assuming the role of a public data user who wished to identify schools. In this context, it seems reasonable that a user would match the school file against other publicly available school files, using advanced record-matching software that is readily available. The data user simply runs this software over the paired survey/CCD, and/or survey/PSS data files, adjusting the software parameters on the basis of variable reliability and consistency between the files. FRIL can calculate the likelihood of a linkage while allowing for incomplete and/or error data conditions within the linked records. Each selected variable (identifier) contributes to the estimate of match probability. As determined by matching rates either provided to or generated by the software, some component identifiers may contribute more weight and/or have higher error rates than others.

For each value contained in the variables utilized as linkage identifiers, FRIL evaluates the reliability and/or probability of accidental agreement. Reliability measures the error rate of the identifier. If an identifier has a very high reliability, there is little chance it would disagree in a pair of matched records. If an identifier has a low reliability, then matched records – even a pair of correctly matched records – will often have different values for that identifier. The probability of accidental agreement permits determination of the discriminating power of an identifier.

FRIL performs one-to-one matching between records in two designated files. Its probabilistic methodology algorithm assigns a score (weight) to every record identifier evaluated, then calculates an aggregate score for each record. The aggregate score (weight) represents the statistical probability (justification) of the paired records being the correct match, providing a measure of how good the match is for each set of records.

References

Fellegi, I.P., and Sunter, A.B.. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.