Probabilistic Matching Software – Data Analyst Description

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). There are comparable matching software available – whether commercial or home-grown. They should provide comparable tools and capabilities as FRIL. The description here centers on FRIL since the IES DRB is familiar with, and approved the software for disclosure analysis.

Please note, however, that both the positive power and potential weakness of FRIL (and comparable matching software) is the ability of the user to maximize matches by properly implementing the parameters and identification of the variables. Thus, improperly implemented by not incorporating the full set of matching variables and/or by not factoring the reliability of the variables, disclosure risk schools may be missed in running the software.

FRIL is free open source tool that enables fast and easy record linkage. The tool extends traditional record linkage tools with a richer set of parameters. Users may systematically and iteratively explore the optimal combination of parameter values to enhance linking performance and accuracy. Although the software is no longer supported, it remains a viable software solution.

Key features of FRIL include:

- Rich set of user-tunable parameters
- Advanced features of schema/data reconciliation
- User-tunable search methods (e.g., sorted neighborhood method, blocking method, nested loop join)
- Transparent support for multi-core systems
- Support for parameters configuration
- Dynamic analysis of parameters

FRIL is 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. Currently, FRIL provides the functionality required for optimal matching, 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.

The IES DRB disclosure requirements accept a "probabilistic" linkage approach that is used to calculate the likelihood of a correct match between the schools represented on the school level (or derived school) records on the IES study file and school level records on the CCD, IPEDS, and PSS files.

Historically, 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 IES Study schools. In this context, it seems reasonable that a user would match the IES Study 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 IES Study/CCD, and/or IES Study/PSS data files, adjusting the software parameters on the basis of variable reliability and consistency between the files. FRIL is a generalized, multi-discipline software package that can calculate the likelihood of a linkage while allowing for incomplete and/or error data conditions within the linked records. Each selected variable (identifier) record component 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.