Bayesian Record Linkage and Linkage Imputation Using CODES2000 and LINKSOLV Software

Michael H. McGlincy

Strategic Matching, Inc.
PO Box 334, Morrisonville, NY 12962, mcglincym@strategicmatching.com

Abstract

Probabilistic record linkage is a powerful research technique for building rich datasets by linking related information stored in separate computer files. The technique can be effective even if the files lack common unique personal identifiers or if identifying fields contain errors and omissions. CODES2000 and LINKSOLV are commercial software products that improve upon earlier implementations of the Fellegi and Sunter theory of probabilistic record linkage. The software first estimates a Bayesian posterior probability that each possible record pair is a true match given all observed agreements and disagreements of field values. Then the Bayesian posterior probabilities are used either to select only high probability matches or to multiply impute complete sets of linked record pairs that include both high and low probability matches. Multiple linkage imputations can be analyzed by standard techniques. NHTSA grantees use CODES2000 software to link police reports about motor vehicle crashes to resulting medical treatment records in order to create Crash Outcome Data Evaluation Systems (CODES). LINKSOLV is available to other researchers. This software demonstration illustrates the process by linking two datasets often of interest to crash outcome researchers.