CompStats 2024

Computational statistics and applications

A stochastic expectation maximisation approach to record linkage

A new method is introduced to combine observations from overlapping data sets without a unique identifier, commonly known as record linkage. This task holds potential importance in healthcare longitudinal studies where one has to rely on partial information to monitor the data, and more broadly, it offers the opportunity to augment data with external sources, circumventing costly data collection. As the main innovations, we address time-varying variables like place of residence and develop an efficient algorithm that can be used on large data sources. The complexity of the record linkage task stems from the sub-par reliability of the partially identifying variables (e.g. initials, birth year, zip code) used to link records and their limited number of unique values. Furthermore, because everyone is often uniquely represented in each file, records from one file can maximally be linked with one record in the other file, making the linkage decisions interdependent. Our new approach uses a Stochastic Expectation Maximisation based on a latent variable model to accommodate registration errors and changes in the identifying information over time. Our model provides a probabilistic estimate of the common set of records that allows for inference. We implement our methodology in an R package, investigate its properties with a simulation study, and apply it to two large surveys.