NCSC
Data Dives

Episode #5: Record Linkage —
The matchmaking tool for data records

By Andre Assumpcao | April 2024

The Problem

You have court records with repeated entities (or individuals) but they lack a unique record ID for you to track the multiple entries of each record. These records are spread across multiple databases (e.g., criminal case and pre-trial data tables). You need to find the same records but don't know how to start.

The Solution

Record Linkage! A systematic way to match records based on their information.

The Method

Record linkage programs use information that can be matched across records (and that indicate that the entities are the same) to suggest records matches. An example is using first and last name to find an individual across datasets – e.g. John Doe and Jon Doe. Though the records differ, record linkage tools use contextual knowledge about how people are named to point to a likely match between John Doe and Jon Doe. A special version of record linkage is de-duplication – a process deployed to find duplicate records in the same database and remove them.

Components

When thinking of a record linkage application, one needs to consider:

THE INFORMATION USED FOR MATCHING

Usual candidates are **names**, **addresses**, **date of birth**, **race and ethnicity** information, other **external IDs** (driver's license IDs, tax IDs). Less common but other good candidates are **professional occupation**, **educational background**, **zip codes**, **emails**, **IP addresses**, etc. Special attention should be given to data quality, such as accuracy, completeness, and consistency of data. Poor data can affect the accuracy of record linkage results as one might miss matching records because data are incomplete, for instance.

THE HIERARCHY BETWEEN INFORMATION USED FOR MATCHING

One may want to **weight one information more heavily than others** given different degrees of contribution of the information for matching. For instance, last names are less commonly shared across people than first names, so it makes sense that last names receive a heavier weight when trying to match records, e.g. a Jeanette Doe and a Janette Doe (last name match) pair should be ranked higher as a potential match than Jane Rivera and Jane Rios (first name match).

WHETHER INEXACT INFORMATION MATCHES SHOULD ALSO BE USED FOR MATCHING

In the examples above, Jeanette Doe and Janette Doe match on last name and are phonetically similar on first name. When doing record linkage, one needs to consider whether they will only use **exact matches** (last name) for linking records or whether **inexact, partial matches** (first name) should also be used.

THE ACCEPTABLE LEVEL OF MATCHING ERRORS

No record linkage tool is perfect and often one will mistakenly attribute two or more records as being the same entity. Thus, it is important to **define one's level of comfort with errors**. If two records match on 7 out of 10 fields in a data table, is that good enough or does one need 9 out of 10 matches? Or even a perfect score of 10 matches? Being comfortable with error rates also means **manually reviewing a random sample of matches** and verifying whether the records are true matches – often human review picks up mistakes that go unnoticed by computers. Would one be comfortable with a 9 out of 10 correct paired records? Or does one need all 10 pairs to be correctly classified?

Applications

Deterministic Linkage vs. Probabilistic Linkage

DETERMINISTIC LINKAGE

Deterministic linkage is a record linkage method that relies on **exact matching of specified information** or fields between records to establish the matches. In this approach, predefined deterministic rules are used to identify and link records that share identical values in key fields, such as full name and date of birth. The process is deterministic because it produces a clear, unambiguous link between records when the specified criteria are met. While this method is straightforward and efficient for exact matches, it may be less effective when dealing with data that contain errors, variations, or lack unique identifiers.

PROBABILISTIC LINKAGE

Probabilistic linkage, on the other hand, is a more flexible approach that **assigns probabilities to potential matches** based on the similarity of fields across records. Instead of requiring exact matches, probabilistic linkage evaluates the likelihood that records refer to the same entity by considering the overall similarity of multiple attributes. This method is particularly useful when dealing with data that may contain errors or variations. Probabilistic linkage employs statistical models, often using machine learning techniques, to calculate the probability of a match. The outcome is a measure of confidence in the linkage, allowing for a nuanced approach to handling uncertainty and reducing the risk of false positives or false negatives in the record linkage process.

Resources

There are many record linkage tools available out there, both proprietary and open-source software. Two open-source options are:

1. **fastLink**
   A program written in programming language R (open-source) and validated in a series of scientific papers by political scientists at Harvard and Princeton.

2. **recordlinkage**
   A program written in programming language Python (also open-source) implementing almost all industry-leading record linkage algorithms.

Last but not least!

All other data governance and analysis best practices apply when matching records across datasets.
Set up processes to preserve data or user privacy (when applicable), to handle inconsistencies, to respond to entities matched in case they question your matching process, to monitor and update your matching algorithm as needed. The success of complex statistical processes is contingent on your ability to communicate them!

NCSC Data Dives is a forum designed to help the court community dive deeper into data analytics and gain valuable insights to help solve data problems. Through a combination of both interactive group and targeted individualized sessions, Data Dives focuses on past projects, innovations, and topics of interest. Visit **ncsc.org/datadives** to learn more.

National Center for State Courts