Introduction to Data Linkage with R

Radmila Velichkovich

Business Research Partner

Empowerment Through Knowledge

Supported by:

Data Science Conference 5.0 Serbia

and

Data Science and Marketing Academy, ETK

Belgrade, Serbia

16 November 2019

TABLE OF CONTENT

LET'S GET TO KNOW EACH OTHER

About me

- 2015-2018: MA Survey Methodology and Public Opinion
- 2016-2017: Institute for Marketing and Market Research
- 2017: Data analyst in media industry
- 2018-2019: R Instructor (freelance)
- August 2019- present

ETK

AT A GLANCE

ETK (Empowerment Through Knowledge) is a London based **market research and data science consulting firm**.

We have offices in London, GB (headquarters), Frankfurt, DE, Riyadh, SA and Belgrade, RS.

Our strength lies in both the highest professional standards and profound knowledge of the local market and its consumer.

Our dedicated team believes that our international experience combined with the deep knowledge of the local market and culture makes the competitive edge.

We demonstrate and set high expectations for ourselves in providing clients with the exceptional service. ETK promotes a culture of professionalism by leading by example.

Introduction to Data Linkage

Introduction to Data Linkage
Definition

- **Data linkage** is the task of *identifying, matching, and merging records* that correspond to the same entites (people, families, companies, product items, events etc.) from several databases or combination of different sources.
- By matching we:
- Avoiding new data collection by linking existing datasets (lower financial burden, lower burden for respondents, a potential solution for low response rate)

- Are in position to perform additional analyses

- Can obtain additional (less biased?) insights

Data linkage=record linkage= entity resolution=object identification=field matching

Introduction to Data Linkage

My concerns were…

- What if I didn't have unique ID at all?
- What if the unique ID was unique to one dataset and differs from the ID in the other dataset ?
- Is there a theory-implentation gap?
- What is the data quality of linked dataset? What power do I gain as a researcher?
- Privacy-preserving record linkage

The GDPR does not apply to personal data that has been anonymised. Recital 26 explains that:

> "…The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes."

Introduction to Data Linkage examples

- Swiss Federal Statistical Office: (e.g **Longitudinal linka**

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

- SwissRDL: medical Registries and Data Linkage

SWISSRDL

- SORS: Challenge for 2021; Linking census and geospatial data

Statistical Office of the Republic of Serbia

- SILC (Survey on Income and Living Conditions): In France, households get less questions on their income when they allow INSEE to access their tax returns; linking administrative with survey data

Introduction to Data Linkage
Typology

- **DETERMINISTIC**- The simplest kind of record linkage, generates links based on unique identifier(s) that match among the available data sets or some other rule-based

- **PROBABILISTIC** - takes a different approach to the record linkage problem by considering a wider range of potential identifiers and calculating the probability that two given records refer to the same entity. Record pairs with probabilities above a certain threshold are considered to be matches, while pairs with probabilities below another threshold are considered to be non-matches; pairs that fall between these two thresholds are considered to be "possible matches" and can be dealt with accordingly (e.g., human reviewed, linked, or not linked, depending on the requirements).

- **MACHINE LEARNING APPROACH-** classification algorithms

Introduction to Data Linkage
Challenges

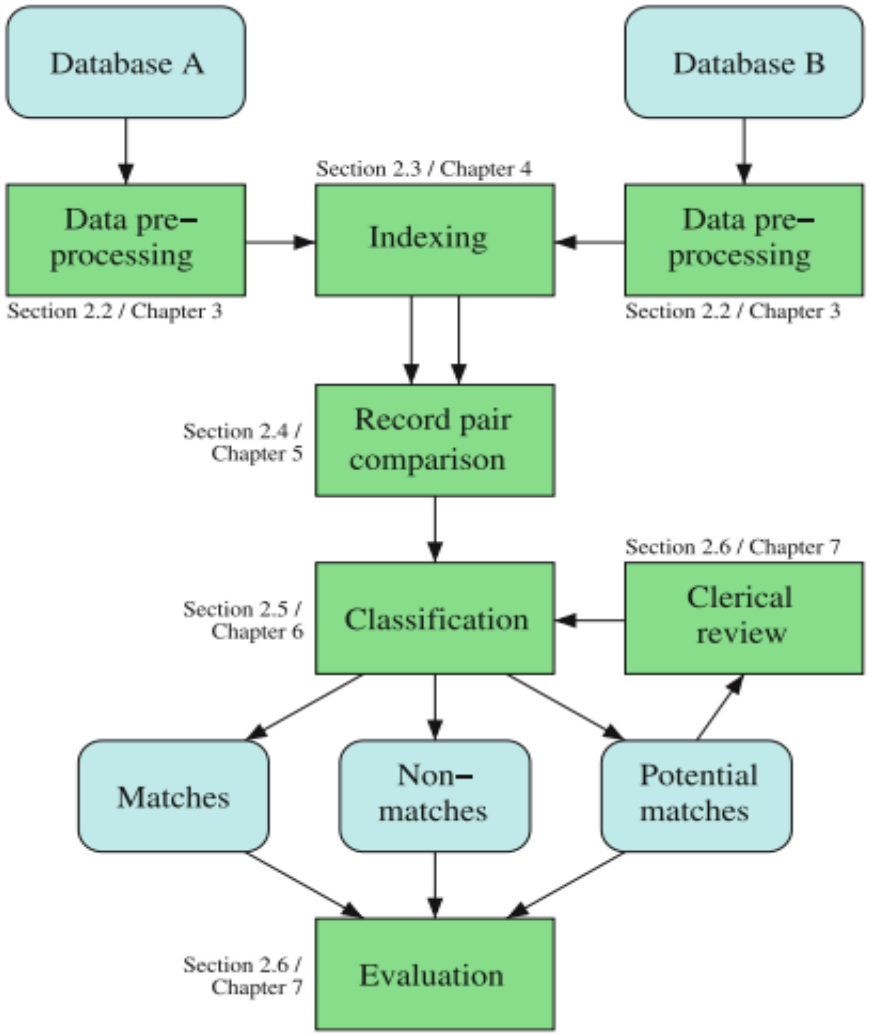
- Identifying duplicates
- As the databases to be matched get larger, the computation complexity of data matching therefore grows
- The true status of two records that are matched across two databases is not known
- Privacy and confidentiality (privacy preserving data linkage)

Introduction to Data Linkage Challenges

- Tables that contain the same type of information can have different names
- Attributes coded/formatted differently (e.g. different date convention)
- Compound information (e.g. Postcode and Location) in the same column
- Attributes can be recorded in different measurements
- Even if there is an ID it is differently named

DATA MATCHING PROCESS

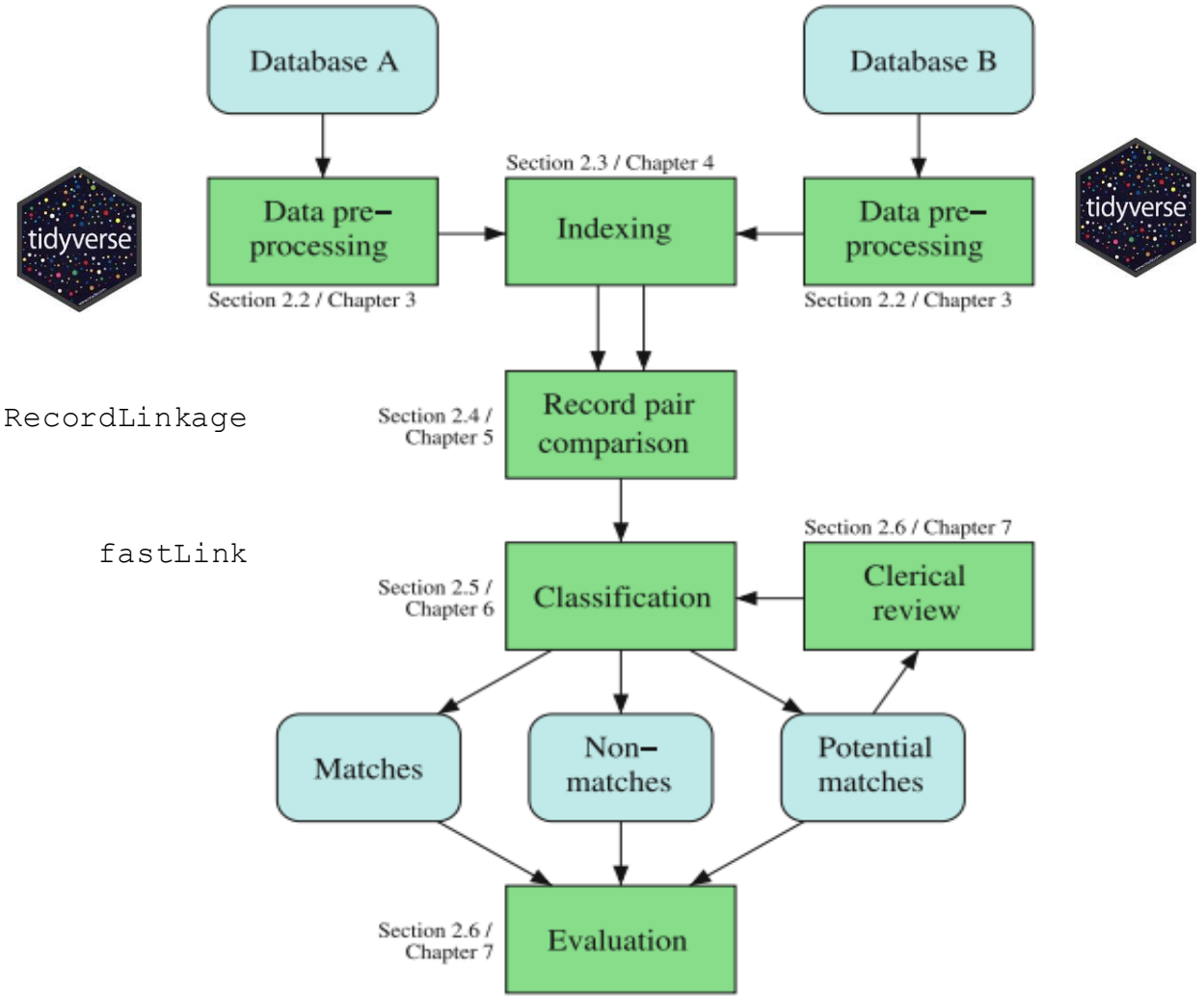
Data Matching Process



Source: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection **Christen**, Peter

Data Matching Process

How does  fit?



Source: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection **Christen**, Peter

Data Matching Process
Data pre-processing

Within one dataset

- Missing values
- Smoothing noisy values
- Identifying and correcting inconsistencies

Data Matching Process
Data pre-processing

- For Data Matching
- Removing unwanted characters and tokens (plus correct misspelling)
- Segmenting attributes into well-defined and consistent output
- Verifying attributes based on existing attributes

Data Matching Process
Indexing

- Record pair comparison process is computationally expensive even for small datasets.
- This is because we need to compare one record with all the remaining ones.
- This results to a total number of record pair comparisons that is quadratic in the size of the datasets to be matched. For example, matching 7x7=49 comparisons.
- Clearly, this naïve comparison of all records does not scale to very large datasets. Matching two datasets with 1 000 000 records each results in 1 000 000 000 000 record pair comparison. This takes time (calculated in days).

Data Matching Process
Indexing

- To reduce the possibly very large number of pairs of records that need to be compared, indexing techniques are commonly applied. They generate candidate record pairs that will be compared in more details in the comparison step of the data matching process to calculate the detailed similarities between the records.

- Various indexing techniques for data matching and deduplication have been developed (e.g. blocking). It splits dataset into smaller blocks according to some blocking criteria (blocking keys).

Data Matching Process
Record pair comparison

- Comparing record pairs with and without indexing.
- However, data used in the matching process can be of low quality (errors, typographical variations, names and address values can change over time)
- Even sophisticated data cleaning and standardisation techniques are not always able to create high quality data that will convert values into the same form for all attributes in pair of records that refer to true matches.
- Rather than comparing attribute values between two records using only an exact comparison function, it is vital for data matching to employ comparison functions that return some indication of how similar two attribute values are.
- There exists various similarity measures (q-gram based string comparison, Jaro and Winkler, Extended Jaccard Comparison).

Data Matching Process
Classification

- Deciding on matching status (match/ non-match/ possible match)

- Classification as a result of determistic/probabilistic/or ML methods

Data Matching Process Evaluation

- True positive - These are record pairs that have been classified as matches and that are true matches. These are the pairs where both records refer to the same entities.

- False positive - These are the records that have been classifed as matches, but they are not true matches. The two records in these pairs refer to two different entities. The classifier has made a wrong decision with these record pairs.
- True negative - These are the record pairs that have been classified as non-matches, and they are true non-matches. The two records in pairs refer to two different real-world entities.
- False negative - These are the record pairs that have been classified as non-matches, but they are true matches.

AVAILABLE TOOLS

Available tools

R

Various packages available

PYTHON

Python Record Linkage Toolkit

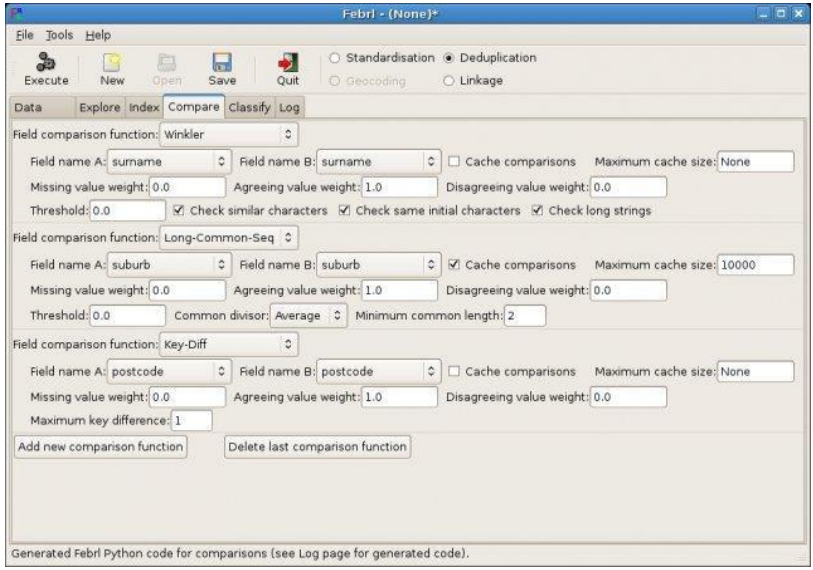
OTHER

Other tools

- **BigMatch**
- Developed and by the US Census Bureau to match very large census data collections
- It is not a full data matching system, rather it is a programme that can be used to extract potential matches from very large files. These plausible matches are saved into smaller files so that they can be individually processed with a proper data matching system later on.

Other tools

- **FEBRL**
- The Freely Extensible Biomedical Record Linkage system is an open source data matching system that has been developed since 2003 at the Australian National University.
- The aim was to improve pre-processing, deduplication and data matching.
- Written in Python



Literature and resources

- Books:
- Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Peter Christen, 2012, Springer
- Data Quality and Record Linkage Techniques, Herzog et al, 2007, Springer
- Methodological Developments in Data Linkage, Harron e al., 2015, Wiley
-

- PPRL R package: https://cran.r-project.org/web/packages/PPRL/PPRL.pdf
- fastLink R package: https://cran.r-project.org/web/packages/fastLink/fastLink.pdf
- Record Linkage R package: https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf
- Privacy Preserving Record Linkage Workshop: https://sites.google.com/view/pprl2019workshop/
-