Introduction to record linkage

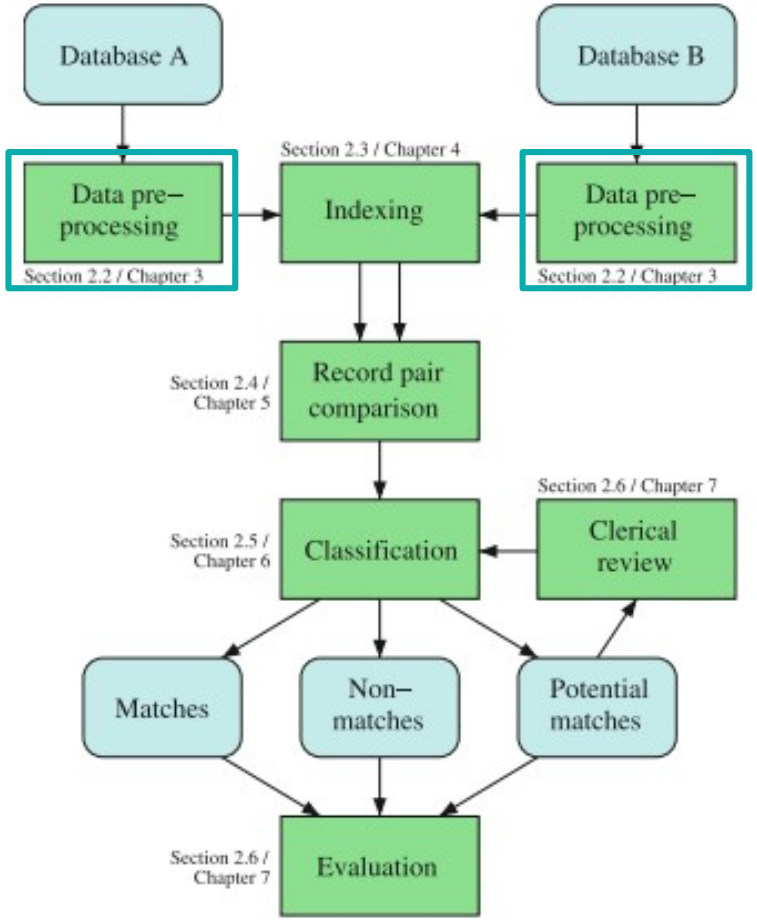
Vocabulary and key concepts

Insee

Measuring, understanding

WHAT IS RECORD LINKAGE?

- Record linkage is the task of identifying records representing the same entity across two data sets.
- The task is straightforward in presence of a common unique identifier or with data of perfect quality, otherwise it becomes complex.
- Different kinds of applications:
  - Enriching a file with an exhaustive database
  - Joining two non-overlapping files
  - Deduplication (~ linking a file with itself)
- Different goals: administrative or statistical?
- Disclaimer: this presentation is purely methodological but the legal question is a major aspect of record linkage for official statistics.



1) Data pre-processing

2) Indexing

3) Pair comparison

4) Classification

5) (Clerical review)

6) Quality evaluation

Source : Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Christen, 2012

CONTEXT

- First step of any treatment on a dataset, pre-processing has a major impact on the results of a record linkage process (garbage in, garbage out).

GOAL

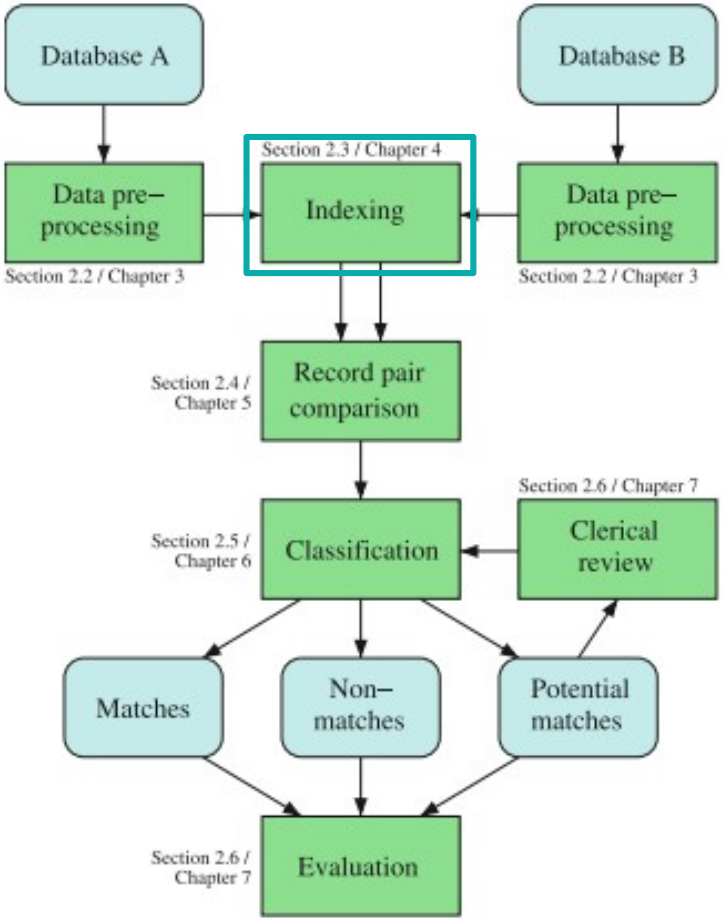
- Clean, standardise and prepare the data for the matching process

PITFALL

- A brutal standardisation leads to a significant drop in variance: information is lost.

EXAMPLES OF TREATMENTS

- **Convert all letters to uppercase**
- **Remove unwanted characters** (JEAN-MICHEL L'HÉRITIER → JEANMICHEL LHERITIER)
- **Remove stop words**
- **Use rules or look-up tables to correct common variations** (av. → avenue)
- **Check for outliers and inconsistent values and correct them** (age > 120, or negative)
- **Verify attribute values with a reference table** (e.g. confront address to a national database)
- **Segment attributes into several fields containing only one piece of information** (dates, addresses, etc.)



1) Data pre-processing

2) Indexing

3) Pair comparison

4) Classification

5) (Clerical review)

6) Quality evaluation

Source : Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Christen, 2012

INDEXATION (1/3)

CONTEXT

- For large files, it is computationally impossible to compare every pair of the Cartesian product ($O(N^2)$ problem).
- Most pairs can be ruled out easily as non-matches.

GOAL

- Reduce dimension by considering only the pairs that possibly correspond to true matches

PITFALL

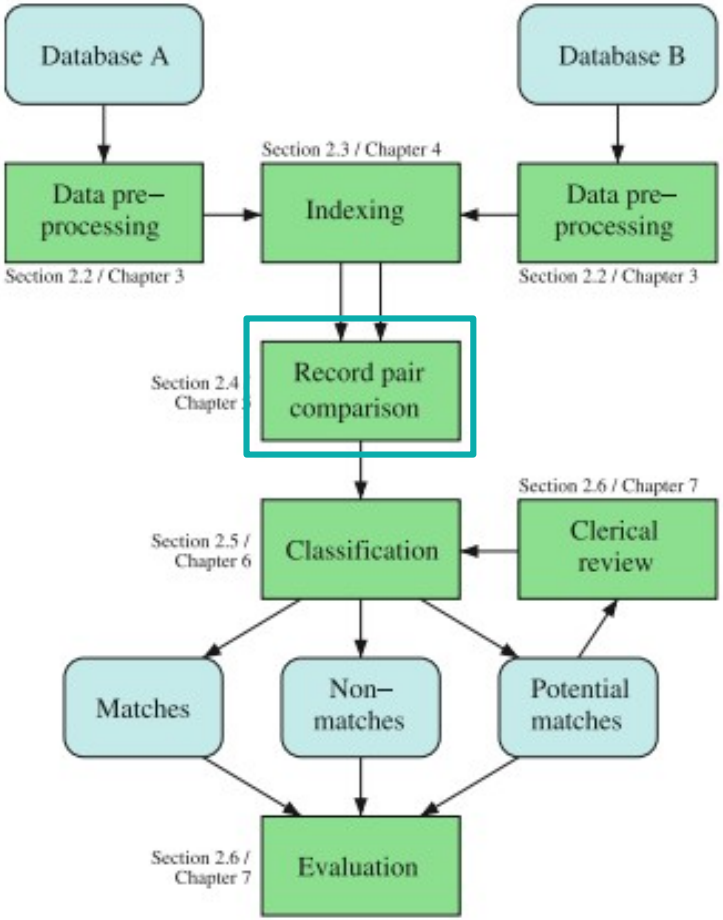
- A drastic indexation may create false negatives i.e. some of the true matches may be missed because of the indexation step. There is a trade-off between dimension reduction and quality of the matching process.

METHODS

- Classic approach: blocking
  - One of the fields is chosen as a blocking key
  - Only the records sharing the same value on the blocking key are kept as potential matches.
  - E.g. if the blocking key is the birth year, only records sharing the same birth year will be compared with each other.
  - The blocking key needs to be of high quality, otherwise the indexation step will create a lot of false negatives.

METHODS

- Variations around blocking:
  - **Use several blocking keys** (e.g. first block on birth year, then on postal code and concatenate all pairs from the two blocking steps)
  - **Build a blocking key with several fields** (e.g. both birth year and postal code must agree for a pair to be considered as a potential match)
  - **Use a distance (e.g. Levenshtein) instead of or in addition to an exact comparison** (e.g. same birth year and maximum Levenshtein distance of 3 for surname)
  - **Compare phonetically (e.g. Soundex for English)**
- Other methods: sorted neighbourhood approach, clustering...



1) Data pre-processing

2) Indexing

3) Pair comparison

4) Classification

5) (Clerical review)

6) Quality evaluation

Source : Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Christen, 2012

CONTEXT

- After the indexing step, the remaining pairs must be compared. Exact matches are not sufficient because most datasets in record linkage problems contain errors in the identifying fields.

GOAL

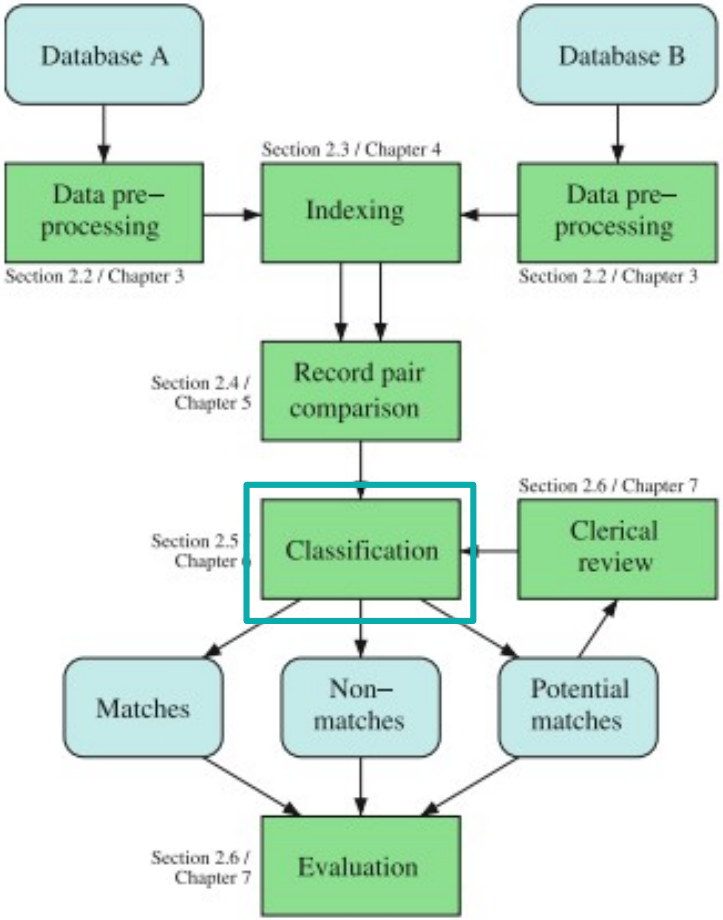
- Compute similarity measures for each identifying field and for each potential pair kept after indexing

PITFALL

- This step is closely related to the next one. Both need to be thought together because some similarity measures work better with certain classification algorithms.

METHODS

- A lot of different measures exist (Levenshtein, Jaro-Winkler, Editex, Q-gram…)
- The choice depends on the field type (string, numeric, date…)
- After normalisation, each similarity measure ranges from 0 to 1 (1 for identical values and 0 for totally different values)



1) Data pre-processing

2) Indexing

3) Pair comparison

4) Classification

5) (Clerical review)

6) Quality evaluation

Source : Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Christen, 2012

CONTEXT

- The aggregation of similarity measures computed in the previous step enables to decide on a match status for each pair.

GOAL

- Classify the remaining pairs into two or three categories: match, non match and possibly potential match

METHODS

- Classification algorithms for record linkage are split into two mains classes: deterministic and probabilistic methods

EXAMPLES OF DETERMINISTIC METHODS

- **Rule-based approach:**
  - Pairs are classified based on a set of rules manually developed
  - If there are several successive steps, records matched at a given point are not considered in the following steps
- **Threshold-based approach:**
  - A weighted sum of similarity measures gives a global similarity at the pair level.
  - If it is above a threshold, it is classified as a match, otherwise as a non-match.
- **Supervised machine learning:**
  - A training sample needs to be manually labelled so that a machine learning algorithm (e.g. SVM or decision tree) learns to classify the pairs based on the similarity measures.

PROBABILISTIC METHODS

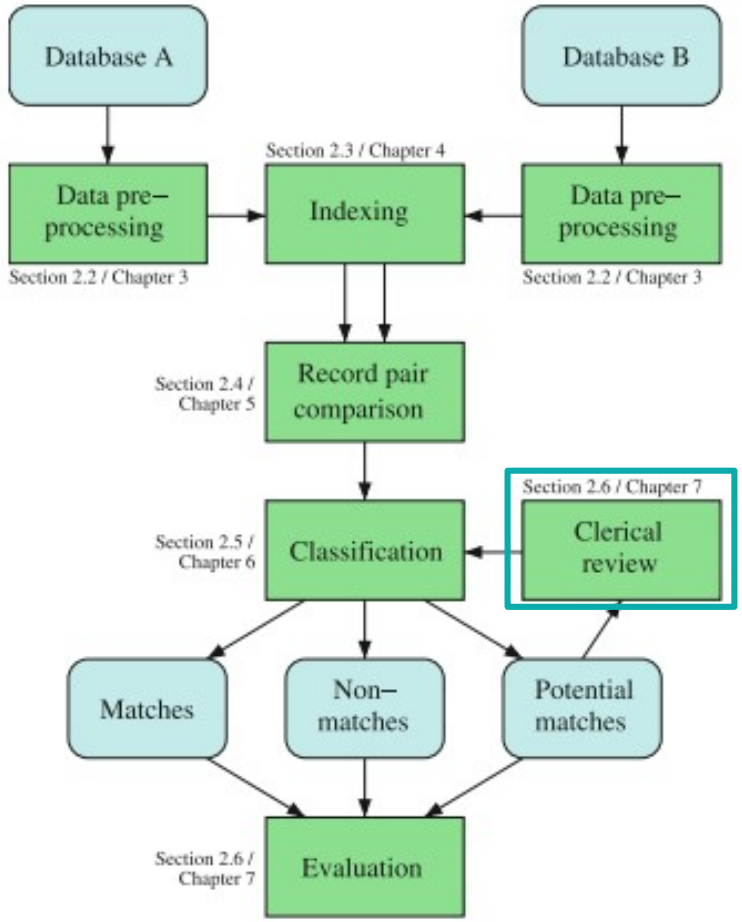
- These methods are based on a framework that was first exposed by Fellegi and Sunter in 1969.
- The difference with deterministic methods lies in the **use of the probability that a pair is a true match**. This probability is used to classify pairs: two cut-off values split the pairs in three classes (match, potential match and non-match).
- The idea is to assign weights to each attribute based on how well an agreement on this attribute predicts a true match (are these variables discriminating enough?).
- These weights are conditional probabilities estimated with external data, a sample of manually labelled pairs or the Expectation-Maximisation algorithm.
- The traditional model relies on an assumption of conditional independence.
- A lot of variations exist and research is still active on this topic.

QUICK COMPARISON OF DETERMINISTIC AND PROBABILISTIC METHODS

- Probabilistic methods allow a better control of error bounds
- They also require less human intervention in the choice of parameters than most deterministic approaches.

BUT

- Deterministic methods are more straightforward and easier to develop ad hoc.
- Probabilistic linkage is more computationally intensive than optimized deterministic classification algorithms, which may be an issue with large datasets.



1) Data pre-processing

2) Indexing

3) Pair comparison

4) Classification

5) (Clerical review)

6) Quality evaluation

Source : Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Christen, 2012

CLERICAL REVIEW (DEPENDING ON THE CLASSIFICATION MODEL)

CONTEXT

- Some models class pairs as potential matches. They require an extra step called clerical review.

GOAL

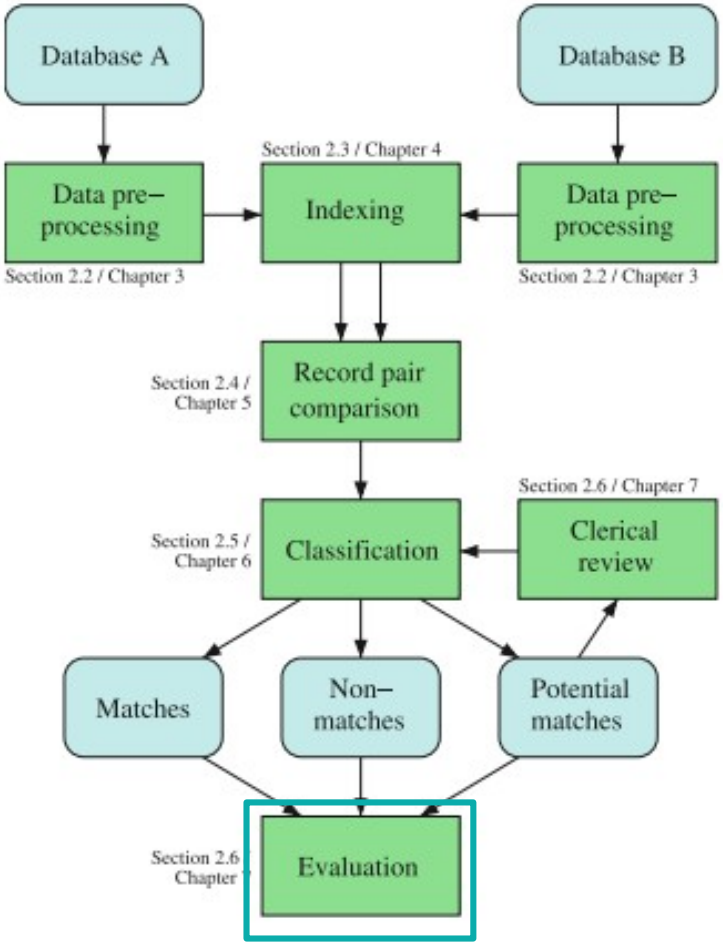
- Call on a human judgement to decide on the cases for which the algorithm was unable to settle.

PITFALL

- This process is expensive, it is important that the number of potential matches classed by the model remains at a reasonable level.

NOTES

- An interface offering a clear visualisation of differences in each field greatly facilitates the task.
- A manual review may be needed at other steps of the process: before the classification for supervised machine learning algorithms and during the evaluation step to compute performance metrics.



1) Data pre-processing

2) Indexing

3) Pair comparison

4) Classification

5) (Clerical review)

6) Quality evaluation

Source : Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Christen, 2012

QUALITY EVALUATION (1/2)

CONTEXT

- Evaluating quality is an essential part of the record linkage process, particularly when studies are based on its output.

GOAL

- Gain as much information as possible on the quality of the matching that was performed

PITFALL

- Most indicators require specific knowledge about the datasets.
  - Gold-standard : representative sample for which the real status of pairs is known
  - Manually labelled sample
  - Statistics and distributions coming from external sources

INDICATORS

- Share of linked records
- Classic indicators for binary classification:
  - true/false positives, true/false negatives
  - precision, recall and F-measure
- Analysis of the distribution of linked pairs, non-linked pairs and pairs with a wrong link status.
- Evaluation of the impact of errors on subsequent studies

NOTE

- Part of the quality evaluation process relies on the users of the linked data.

Join us on

insee.fr

Lucas Malherbe
Data scientist
SSP Lab
lucas.malherbe@insee.fr

Insee
Measuring, understanding