Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

Probabilistic or deterministic? Linkage methods tested for the Résil program

by Olivier Haag, Heïdi Koumarianos and Lucas Malherbe

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Probabilistic or deterministic? Linkage methods tested for the Résil program

Olivier Haag, Heïdi Koumarianos, Lucas Malherbe[1]

Abstract

The purpose of this article is to compare the linkage results for individuals from French tax sources with those of the 2019 *Enquête Annuelle de Recensement* (EAR), obtained through different methods. Such a comparison will decide whether the Répertoires Statistiques d'Individus et de Logements (Résil) program should be equipped with a probabilistic matching tool for its administrative source identification and matching engine.

Three different methods were implemented:

- Rapsodie: This tool, developed by INSEE, uses a deterministic matching method.
- Relais: This tool, developed by Istat, uses the Fellegi and Sunter's probabilistic matching method.
- R and Python packages that use probabilistic methods like Fellegi and Sunter's method as well as deterministic methods involving machine learning.

Keywords: matching, administrative sources, Fellegi and Sunter

1. Issue (Résil)

The Résil (*Répertoires Statistiques d'Individus et de Logements*) program aims to build a sustainable, scalable system of statistical directories of individuals, households and dwellings that is updated using a variety of administrative sources.

In this context, linkages are essential not only to build the directories, but also because the directory system will provide a framework for the DSDS information system. In fact, it will enable matching with other sources: survey and administrative data, either directly or through pre-identification.

Therefore, in order to define the type of identification proposed by Résil, a decision was made to test various matching methods in order to choose the one or ones that seem most effective not only in terms of statistical quality, but also from an IT performance perspective (essential, given the volumes to process).

2. Results

The quality of the matches depends not just on the matching process itself, but also on the quality of the input files. For this reason, this part will present the two sources used, as well as the quality of its variables used for linkage. These variables are as follows:

- Last names (married and maiden)
- First names
- Date and place of birth
- Home address.

This phase involved measuring the indicators below, for each variable from the *Enquête Annuelle de Recensement* (EAR) file and the *Fichier d'Imposition des Personnes* (FIP) file, which are useful for linkage:

- Partial non-response rate
- Rate of erroneous values (when the variable's criteria belong to a list or interval)

[1]INSEE

- Rate of questionable values[2]

The number of duplicates (identical criteria for each variable) was also calculated.

2.1. Description of the EAR data

The data used for this test are from the 2019 *Enquête Annuelle de Recensement* (EAR) (Godinot, 2005). It includes 5 million dwellings and 9 million individuals. For our study, only data on individuals over 15 years of age living in regular dwellings were used (to be comparable with FIP data).

Only a portion of the territory is surveyed every year (one-fifth of the communes with a population of less than 10,000 and about 8% of the dwellings in communes with a population of more than 10,000). Only the information collected about individuals' marital status and address were used for this test. Note that the "quality" indicators presented above do not identify all problems. For example, they do not identify errors caused by optical scanning of last names and first names in the EAR, which prevent many exact matches. For this reason, it was decided to distinguish three EAR subpopulations in the different analyses:

- Individuals who responded via the Internet (not subject to scanning issues)
- Individuals from the EDP who responded by paper questionnaire (which required a higher scanning quality of the "last name" and "first name" variables)
- Others (who responded by paper questionnaire and outside the EDP).

2.2 Description of the FIP data

The *fichier d'imposition des personnes* (FIP) is used to identify taxpayers. These data are from the information collected by the tax services (for example, from a questionnaire) or from taxpayers' income tax and real estate wealth tax returns. It contains information on the marital status and address of the people who fall within the scope of income tax, housing tax, contribution to public service broadcasting, annual vacant housing tax, or real estate wealth tax. For individuals under age 15, only the year of birth is available, which is why they were excluded from the scope of this study. Ultimately, the chosen scope represents more than 80% of the individuals in the FIP.

This file contains all the members of tax households that paid one of the taxes listed above. For example, for a couple with three children, one of whom is older than 15 years and still a dependent, we will have the full marital status of the parents and the child over age 15 and only the years of birth of the two youngest children. Therefore, these two children are not in the scope of this matching test.

2.3 Quality comparison of the two sources

Table 2.3-1
Summary of indicators in the FIP and EAR input files

		FIP	EAR
Number of individuals selected		55,102,356	7,243,345
Individuals with at least one anomaly	Only one anomaly	0.5%	12.3%
	More than one anomaly	0.3%	0.9%
	Share of people born abroad	48.2%	10.8%
Individuals with an erroneous place of birth[3]		0.2%	6.5%
Individuals with questionable first names[4]		0%	5.2%

[2] For example, last name or first name containing something other than letters; last name or first name with fewer than three characters; last name or first name containing a string of at least three identical letters, or just consonants or just vowels; etc.
[3]Including people born abroad for whom department 99 is expected.
[4]The following are considered questionable: missing values or first names with at least one special character or a number; first names with fewer than three characters; first names with at least three characters, those containing only vowels or consonants, or at least three identical letters in a row.

Most individuals with at least one anomaly are individuals for whom only the place of birth is missing. However, when an individual has multiple errors, it is often missing values combined with the day and month of birth.

Other results lead to the following diagnosis:

- For people born abroad, the January 1 date of birth is overrepresented (three times more people born on January 1 than on other days of the year in the FIP, for example) (undoubtedly associated with vital statistics gaps in some countries) in both the FIP and the EAR.
- When broken down by collection method for the EAR, as expected, Internet responses are slightly better (12% of individuals with at least one error versus 15% for paper).

In conclusion, it can be said that for the variables that are useful for matching, the FIP file is of excellent quality at first glance. However, the EAR file is of slightly lesser quality, but since not many individuals have multiple problems, the impact on the effectiveness of the matching process should be limited. However, it must be remembered that these tables do not factor in potential errors in entering first and last names. For this, a test was performed to measure the rate at which first names from the EAR appear on the list of first names on the Insee.fr website. We note that more than 10% of first names from the EAR are not on the list!

3. Methods

3.1 Rapsodie

Rapsodie is a matching and enhancing tool developed by the *pôle Revenus Fiscaux et Sociaux* of the Department of Rennes (Jabot et al., 2010).

It is based on the principle of successive rounds of matching to optimize processing times. At the outset, there were theoretically more than 750 tera pairs to study.

1. The first-round matches individuals in the two files who have exactly the same last names, first names, sex, date of birth, and commune of residence. This step matches almost 70% of the individuals in the EAR, therefore cutting the number of pairs to examine by a third.

2. The second round matches individuals not matched in the previous round, using a “nearest echo” method with the commune of residence as the blocking key. Therefore, two individuals are matched who live in the same commune in both files and who’s weighted[5] sum of the distances[6] of the differences between each variable (last name, first name, date and department of birth, and the first two keywords in the address) is below a previously defined threshold. Adding management rules[7] makes it possible to consider some individuals as matched even if the sum of their distances exceeds the threshold. Note that if, for a given individual, there are multiple pairs with a distance below the threshold, only the pair with the smallest distance is selected. The use of “blocking keys” limits the number of pairs to analyze by dividing it by more than 5,000.

3. The third-round matches, among individuals not matched in the previous rounds, those who have exactly the same last names, first names, dates of birth and sexes, and different communes of residence.

3.2 Optimizing matching via Rapsodie

As a reminder, the quality of the matching process is not only measured by match rate bias: it is possible to have matched 100% of a file yet made 30% errors. The following populations should also be factored in (Doidge et al., 2021).

[5] The distances of the last names and first names were weighted by 1.5 and the other distances by 1.

[6] The Levenshtein distance is used for wordings (last name, first name, address), while for the other variables (date of birth and geographic code), the distance is 0 if they are equal, and 1 if not.

[7] Example of a management rule: EAR first name = FIP last name, FIP first name = EAR last name and other exact variables (date of birth, address).

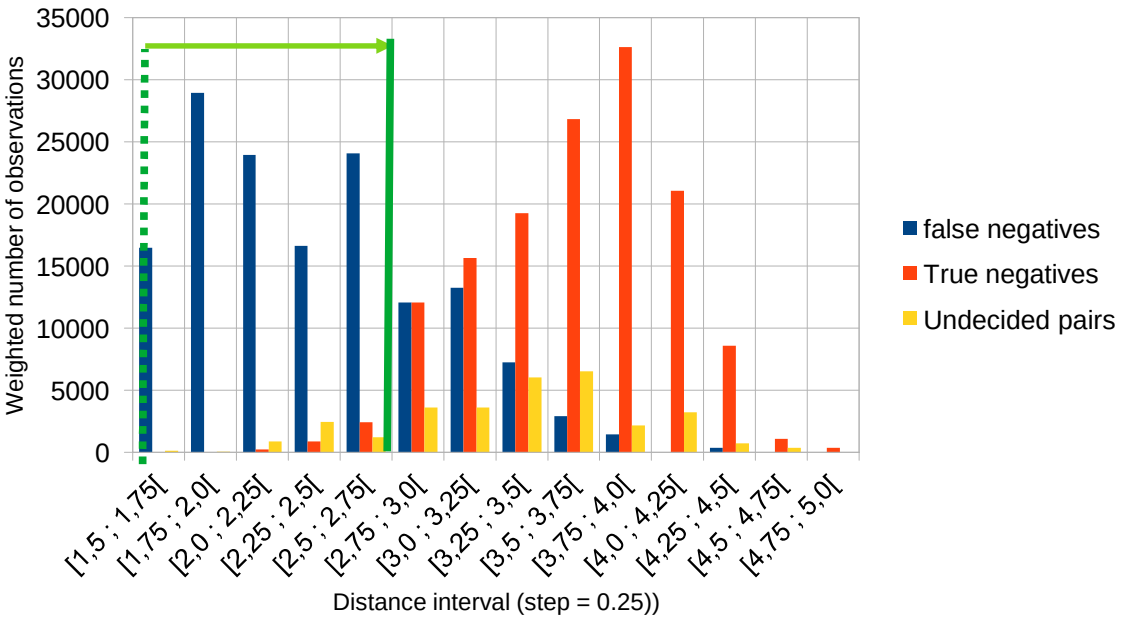
Table 3.2-1
Types of pairs encountered in imperfect matching

	True pairs (same individuals)	False pairs (different individuals)
Accepted in the result file	True positive (around 7 million)	False positive (to be minimized)
Rejected in the result file	False negative (to be minimized)	True negative (around 400 trillion)

A visual analysis of 3,000 pairs (1,000 accepted and 2,000 rejected) was performed not only to estimate the quality of the matches (estimate of false positives and false negatives), but also to optimize it by providing a more effective pair acceptance threshold and new management rules.

The pairs to visually check were selected randomly through stratified sampling based on the distance of the pairs, by overrepresenting the distances in the neighbourhood of the decision thresholds. To facilitate diagnosis, in addition to the characteristics of the individuals in the two sources used for matching, the control file included the household composition of the individuals in the pairs in both sources (list of individuals living in the same dwelling from the EAR and list of individuals from the same tax household from the FIP for the individuals in the pair). This strategy produced more pairs close to the threshold and optimized the latter using a sufficient number of analyzed pairs.
The weighted results are shown below. The false positive rate is zero at the 1.6 threshold and very low (0.3%) at the 2.7 threshold. For false negatives, the results (weighted) of these checks are in the chart below:

Figure 3.2-1: Distribution of the pair type, by distance[8]



Scope: 319,277 rejected pairs for which a calculated distance is available.

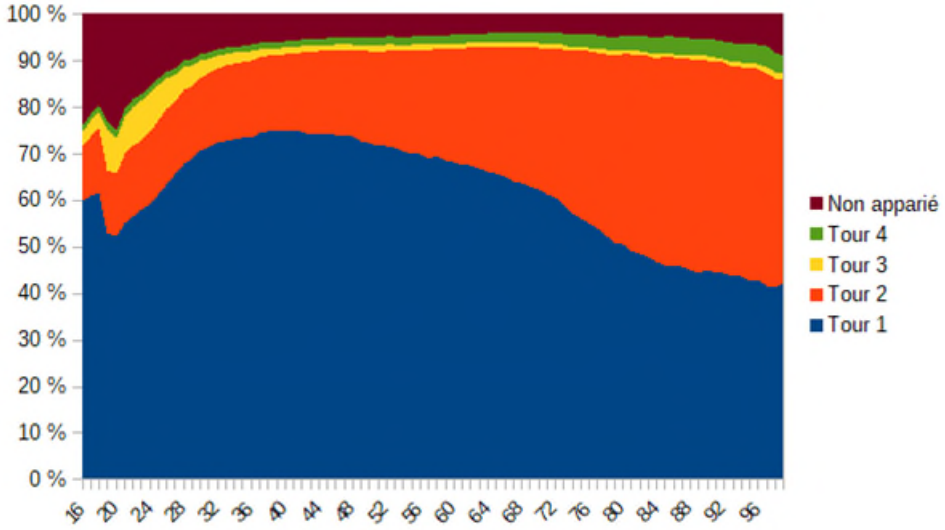
This chart identifies the optimal threshold for pair acceptance. Up to 2.75, we see that the false negative rate is actually higher than the false positive rate and beyond that, the potential matching gain includes a very high number of false positives. Therefore, this analysis helped to set this threshold at 2.75 (when it had originally been set at 1.6); this adds a “fourth round” to the first three. This decision results in a false positive rate of around 0.3%, but also a gain of 2 points in the matching rate. Ultimately, with these parameters, 92.7% of individuals from the EAR (limited to individuals over the age of 15 in regular dwellings) matched with an individual from the FIP.

3.3 Representativeness of the population matched via Rapsodie

[8] The undecided pairs are instances where the visual analysis did not definitively distinguish between false positives and false negatives, despite the additional information.

It is very important to analyze the representativeness of the population resulting from the matching to measure potential biases. That way, they can be corrected before they start being used to produce statistics.

Chart 3.3-1: Match rates between the EAR and FIP, by round and age



Scope: Persons over the age of 15 from the 2019 EAR

For individuals aged 16 in the EAR, 60% were matched with an individual from the FIP in the first round of Rapsodie, 12% more in the second round, 2% more in the third round, and just under 1% in the fourth round. In the end, just under 75% were matched. The first conclusion is that the matched population is biased since it underrepresents youth under 30 years, particularly those between the ages of 18 and 22. If we want to use this population to produce statistics, it would be a good idea, for example, to calibrate it on the age structure of the French population over 15 years.

It was also found that the exact matching (round 1) achieves a match rate of 65%, which is a good result. However, we note that the youngest and oldest individuals have poorer exact match rates. The reasons for this are not exactly the same at first glance.

- For the oldest individuals, this decrease can be due mainly to their higher propensity to respond to the paper questionnaire, which results in more errors in family names, thereby preventing exact matching. This problem is partially rectified in rounds 2 and 4 (close matching rounds, with pair acceptance thresholds of 1.6 and 2.75, respectively), which match individuals with close, but not necessarily identical, characteristics.

- For the youngest individuals, the reasons are twofold:
  - their location can differ from one source to another,[9] which prevents exact matching. Thus, we can see that exact matching for all of France, which removes the commune of residence constraint (round 3 in this chart) retrieves more young people.
  - The coverage bias of the FIP frame of 18- to 20-year-olds, which is therefore found in the matched population.

3.4 Relais

Relais is a matching tool developed by Istat, the Italian National Institute of Statistics. The tool has a graphical user interface where non-expert users can perform matching, using a deterministic or probabilistic method (Cibella et al., 2010). Relais allows users to download data within the tool, to select matching variables, choose a problem-reducing method (such as blocking), and select the record comparison features. The probabilistic method sets the pair retention/rejection thresholds and proposes an estimate of "quality" indicators, such as recall[10] and accuracy.[11]

[9] Young students can be surveyed in their student housing, but are associated with their parent's tax household, therefore located at the parents' address in the FIP.
[10] Recall is the proportion of true pairs correctly identified among all true pairs.
[11] Accuracy is the proportion of true **pairs** among selected pairs.

The method chosen for this study is the probabilistic method, based on the Fellegi-Sunter theory (Fellegi andSunter, 1969). The data are from the EAR and FIP, already standardized by the Rapsodie tool, for the departments of Lozère and Ille-et-Vilaine. Relais is very limited on the volumes of data used. Data on the department of Lozère could be processed during the same matching process.

However, the size constraints weighed more heavily in favour of doing the data matching for Ille-et-Vilaine and resulted in the file being divided into seven parts: five files divided by a commune-based criterion, and two others based on one commune-based criterion and year of birth. This division resulted in a suboptimal process: first, the probability estimate under the Fellegi-Sunter theory is done several times and can differ from one file to another; then, the division by year of birth actually constitutes an additional partial block.

For each of these files, a block was done on the commune of residence. The following comparison criteria were used:
- Levenshtein distance (standardized) greater than 0.9 for birth name;
- Levenshtein distance (standardized) greater than 0.8 for first name;
- equality for the variables year of birth, month of birth, day of birth, and place of birth;
- Levenshtein distance, standardized, greater than 0.8 for the first two keywords of the address.

Note: within Relais, it is not possible to specify a combination of conditions (like equality of names at birth or equality of usual names).
The pair acceptance threshold is set at a probability of 0.9.

3.5 Presenting the Python library *recordLinkage* (called “python” in section 4)

The library *recordLinkage* is a Python library, so it is an open-source tool. It allows the implementation of deterministic matches as probabilistic. In this article, we tested the probabilistic method, resulting from the Fellegi-Sunter framework. The library provides a set of features for implementing the various steps of matching, such as blocking, calculating distances and, of course, classifying pairs.

As with Relais, the data used were those standardized via Rapsodie.

For Lozère, the matching strategy chosen involved three steps. The individuals matched in one step are no longer considered for the following ones.

1. The first step involves exact matching of last name, first name, sex, date of birth, and commune of residence.
2. The second step is probabilistic matching with blocking on commune of residence.
3. The third step is probabilistic matching with blocking on year of birth.

For Ille-et-Villaine, applying the same strategy as for Lozère leads to excessive use of RAM, exceeding 250 GB. Therefore, a step was added after the exact matching to reduce the total number of pairs compared. The strategy was as follows:

1. exact matching;
2. probabilistic matching with blocking on commune of residence and year of birth, where only individuals living in the same commune and born in the same year are compared;
3. probabilistic matching with blocking on commune of residence only;
4. probabilistic matching with blocking on year of birth only.

Even when applying this strategy, RAM usage still approaches 200 GB during the most usage-intensive step, the third one, during which over 400 billion potential pairs are processed. This volume is mainly due to the presence of large communes in the department, especially Rennes.

As with Relais, the probabilistic classification algorithm implemented in the *recordLinkage* library operates with binary variables. This does not require making exact comparisons only; it is still possible to use fuzzy comparisons, as long as a threshold is decided for “binarizing” them later. The comparison rules chosen were as follows:

- last name, first name, and the two basic words of the address: Jaro-Winkler similarity with a threshold of 0.92;
- commune of residence and day, month, year, and commune of birth: exact comparison

The algorithm outputs a probability for each pair from the blocking step, i.e., each pair for which comparisons were made and distances were calculated. The pair acceptance threshold was set at 0.5 here, but this threshold can be adjusted following visual analysis of a sample of pairs.

4. Comparison of Results

4.1 Comparison methodology

The comparison among the three methods was done in a detailed manner for two departments (48 and 35).

Note that, for performance reasons, the matches could not be done the same way in both departments, as mentioned at points 3.2 and 3.3 above.

The comparison principle was as follows:

- Comparison of the match rates for the three methods
- Identification of the different matches
- Selection of a sample of 1,000 pairs matched in different ways
- Measurement of the false positives through visual analysis of the previous sample.

4.2 Main results

Table 4.2-1: Main results from the comparison of the three methods

	Department 48	Department 35
Population over age 15 EAR	10,127	130,950
Population over age 15 FIP	62,823	874,304
Exactly matched	6,239	96,697
Exactly matched outside of the department[12]	622	3,087
“Rapsodie” match rate (%)	91.3	94.8
“Relais” match rate (%)	92.1	93.7
“Python” match rate (%)	92.4	95.1
“Rapsodie” false positive (%)	0.02	0.05
“Relais” false positive (%)	0.3	0.03
“Python” false positive (%)	0.3	0.7
Corrected “Rapsodie” match rate (%)[13]	91.3	94.7
Corrected “Relais” match rate (%)	91.9	93.6
Corrected “Python” match rate (%)	92.2	94.4

These initial results show that the three methods yield very close results with a very good quality level (a low rate of false positives).

At first glance, the probabilistic methods yield better results (instances with department 48 where these methods could be implemented without constraints). However, when it is necessary to constrain them (see points 3.2 and 3.3 above) because of the size of the input files (instance with department 35) the results are found to be at the

[12]These individuals were removed from the match rate calculation because they could not be found by Relais and the Python method, which searched for individuals only in the files of 48 or 35. This is problematic when we see that 10% of Lozère matches were outside the department.

[13]The corrected rate is obtained by removing the false positives.

same level or even somewhat worse than a deterministic method whose parameters were optimized further to the visual checks.

Note that such a threshold optimization of the probabilistic methods was not implemented; thus, it is reasonable to think that the false positive rate of these methods could be reduced, but at the expense of the overall match rate. An initial examination of the “Python” method’s false positives had still made it possible to reduce the number of false positives by increasing the pair acceptance threshold.

4.3 Analysis of false positives

Table 4.3-1: Distribution of the pairs based on status (correct, incorrect, etc.) and method used

Pair selected by			Correct pairs (%)		Incorrect pairs (%)		Undecided pairs (%)		Total	
Rapsodie	Relais	Python	Dept. 48	Dept. 35	Dept. 48	Dept. 35	Dept. 48	Dept. 35	Dept. 48	Dept. 35
Yes	Yes	No	80.0	90.0	20.0	3.0	0.0	7.0	5	227
Yes	No	Yes	100.0	98.7	0.0	0.5	0.0	0.8	44	1,155
Yes	No	No	97.7	84.0	2.3	6.5	0.0	9.5	43	851
No	Yes	Yes	76.2	93.0	11.9	5.0	11.9	2.0	143	424
No	No	Yes	52.6	26.3	31.6	61.0	15.8	12.7	19	1,465
No	Yes	No	72.0	82.3	20.0	9.4	8.0	8.3	25	113
Yes	Yes	Yes	100.0	100.0	0.0	0.0	0.0	0.0	8,591	126,151

Scope: analysis of the samples of different pairs

As might be expected, the false positive rates are the highest for the pairs identified by only one matching method. An analysis of the “Python” method’s false positives (61% false positives for the pairs identified by this method only) shows that they could be partially removed by increasing the pair acceptability threshold. This will, of course, have a parallel impact on the final match rate. In addition, we note that, in general, the false positives have an incorrect year of birth. This finding could also be factored into the specification of the probabilistic matching model.

4.4 Analysis of the false negatives

Of the individuals matched using Python but not by Rapsodie, in two-thirds of the cases for Department 48, they are people living in two different cities between the EAR and the FIP file. This is since the Rapsodie method implemented was using the commune as a blocking key! In addition, even though Rapsodie ultimately performed a final all-France exact matching step, it did not match two individuals living in two different communes in the two sources and having close characteristics (typo in the last name, for example). To rectify this problem, all that would be needed is a second Rapsodie round, using, for example, year of birth as a blocking key. This improvement was simulated for department 48 and would have resulted in a match rate of around 92.2%. Therefore, using this second blocking key apparently made it possible to get closer to the results of the probabilistic methods.

5. Conclusion

This work confirmed several aspects previously observed by other similar studies on other datasets:

- the value of a visual analysis of accepted or rejected pair samples to measure the quality of the matching and improve its quality by optimizing its parameters.

- The difficulties implementing probabilistic methods on huge files. These problems are due not just to the total volume of records in the file, but also to the maximum size of the subpopulations resulting from the blocking, which defines the largest Cartesian product to be implemented.

- A properly parameterized deterministic algorithm yields results close to those of a probabilistic algorithm with no constraints on file size. In addition, its operation is easier to explain in plain language.

This experiment also shows that different methods can be complementary and useful for assessing their respective quality. A pair found by various methods is less likely to be a false positive than a pair found by a single method. This finding leads us, in the context of Résil, to consider putting in place, along with the deterministic matching tool in production, a probabilistic matching tool that could work on a sample and ensure the quality of the results obtained in production.

Finally, the decision was also made, in the context of Résil, to set up an interface for measuring the quality of the matches through pair annotation. Such an interface will have several benefits:

- measuring the quality of the identification (estimate of the accuracy and recall mentioned above)
- optimizing the process (by improving the parameters or proposing new management rules based on the results of the visual checks.
- Having tagged pairs that could be used to train a supervised machine-learning model that could be considered for pair classification (Midy, 2021).

References

Doidge J. et al. (2021), ‘‘Quality assessment in data Linkage’’, unpublished report, London, Great Britain: Office for National Statistics, https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

Fellegi, I. P. and A. B. Sunter (1969), ‘‘A theory for record linkage’’, *Journal of the American Statistical Association*, 64, pp. 1183–1210.

Godinot, A. (2005), “Pour comprendre le recensement de la population”, unpublished report, Paris, France : INSEE, https://www.insee.fr/fr/information/2579979

Jabot P. et al. (2010), “Appariement d’enquêtes avec des données administratives sociales ou fiscales’’, unpublished report, Paris, France : INSEE, http://www.jms-insee.fr/2018/S20_1_ACTEv2_TREYENS_JMS2018.pdf

Midy L. (2021), “Un outil d’appariement sur identifiants indirects : l’exemple du système d’information sur l’insertion des jeunes’’, *Courrier des statistique n°6*, https://www.insee.fr/fr/information/5398689?sommaire=5398695.