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

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

A proposal for the problem of matching probabilities estimation in record linkage

by Mauro Scanu, Tiziana Tuoto, Marco Fortini and Sara Piombo

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

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

A proposal for the problem of matching probabilities estimation in record linkage

Mauro Scanu, Tiziana Tuoto, Marco Fortini and Sara Piombo[1]

Abstract

Record linkage aims at identifying record pairs related to the same unit and observed in two different data sets, say A and B. Fellegi and Sunter (1969) suggest each record pair is tested whether generated from the set of matched or unmatched pairs. The decision function consists of the ratio between $m(\gamma)$ and $u(\gamma)$,probabilities of observing a comparison γ of a set of k>3 key identifying variables in a record pair under the assumptions that the pair is a match or a non-match, respectively. These parameters are usually estimated by means of the EM algorithm using as data the comparisons on all the pairs of the Cartesian product Ω=A×B. These observations (on the comparisons and on the pairs status as match or non-match) are assumed as generated independently of other pairs, assumption characterizing most of the literature on record linkage and implemented in software tools (e.g. RELAIS, Cibella et al. 2012). On the contrary, comparisons γ and matching status in Ω are deterministically dependent. As a result, estimates on $m(\gamma)$ and $u(\gamma)$ based on the EM algorithm are usually bad. This fact jeopardizes the effective application of the Fellegi-Sunter method, as well as automatic computation of quality measures and possibility to apply efficient methods for model estimation on linked data (e.g. regression functions), as in Chambers et al. (2015). We propose to explore Ω by a set of samples, each one drawn so to preserve independence of comparisons among the selected record pairs. Simulations are encouraging.

Key Words: Probabilities of error in record linkage; matching variable quality; sampling pairs of units.

1. The record linkage problem and the role of independence

1.1 Description and notation

Assume A and B are two data sets whose list of observed units is overlapping, at least partially. Record linkage aims at identifying which records in the two data sets could be linked (and the corresponding variables juxtaposed) in order to get a new data (sub)set of linked units with a richer set of information that can be statistically studied and analyzed.

Fellegi and Sunter (1969) suggest to base the decision on whether a pair is a match or not by comparing a set of k variables (k>3) that are jointly able to identify units (usually named *matching variables*) but with the problem that they can be affected by quality issues or stability over time. Hence, it is not ensured that the same unit reports the same matching variables values in the two data sets. Assuming that the status of a pair of records from the two data sets is represented by the random variable C (with values 1 if the pair is a match and 0 otherwise) while the comparison between the k matching variables is represented by a k-valued random vector Γ where each component is either 1 (if the corresponding variable is the same on the two units) or 0 (otherwise), Fellegi and Sunter consider the following probability distributions:

$$m(\gamma) = P(\Gamma = \gamma | C = 1), u(\gamma) = P(\Gamma = \gamma | C = 0), \quad \forall \gamma. \tag{1}$$

Furthermore, they adopt the likelihood ratio

$$r(\gamma) = \frac{m(\gamma)}{u(\gamma)}, \forall \gamma,$$

as the decision function for declaring each pair as a match (if $r(\gamma)$ is larger that a fixed threshold) or not. Fellegi and Sunter show how to estimate the distributions in (1) in a simple case by the method of moments. More complex

[1]Mauro Scanu, Tiziana Tuoto, Sara Piombo, Marco Fortini, Istituto Nazionale di Statistica ISTAT, via Cesare Balbo 16, 00184 Roma, Italy (e-mail: scanu@istat.it)

situations have been solved by means of the EM algorithm applied on the Cartesian product **Ω**=A×B (Winkler, 1988). Anyway, the EM algorithm is applied over all the pairs in **Ω**. Hence, the observed likelihood function takes the form:

$$\prod_{ab\in\Omega}(m(\gamma_{ab})p)^{c_{ab}}\big(u(\gamma_{ab})(1-p)\big)^{1-c_{ab}} \quad (2)$$

where p is the probability that *C=1* and the status of each pair c_{ab} is not observed. The product implies that the bivariate replicates (*C*, Γ) on the pairs in **Ω** are independent. Anyway, if c_{ab} is equal to 1 for the pairs *a'b'*, *a'b''* and *a''b'*, then the pair *a''b''* is not free to assume any value but 1, so that *C* is not independent for all pairs. The same can be stated for Γ. Hence, the product in the observed likelihood is wrong and the good quality characteristics of the parameter estimators based on maximizing (2) are jeopardized.

The consequences are different. First of all, the decision function r_{ab} depends on the estimates of the distributions *m* and *u*, consequently it can be questioned whether the decisions taken with a wrongly estimated decision function are correct. There are results (Tuoto, 2016) that state that decisions are not much dependent on the parameter estimates obtained by means of the EM algorithm, anyway improvements can be expected. Secondly, given the inaccuracy of the estimates obtained by the EM algorithm, it is impossible to assume as a measure of record linkage accuracy the probability α and β of wrong decisions, where α and β are equal to:

$$\alpha = \sum_{ab:\, r_{ab}>\lambda} u(\gamma_{ab}),\ \beta = \sum_{ab:\, r_{ab}<\mu} m(\gamma_{ab}),$$

where λ is a threshold over which all the pairs with $r_{ab} > \lambda$ are declared as matches and μ is the threshold under which the pairs with $r_{ab} < \mu$ are declared non matches. Finally, the linked data set can be used for further statistical analyses whose objective is to estimate model parameters of variables observed distinctly in the two files A and B. Chambers and Kim (2015) show that knowledge of the distributions *m* and *u* allow the definition of more efficient estimators than the traditional ones that do not assume the existence of linkage errors.

For these reasons, the estimation of the record linkage parameters is an issue that, in our opinion, needs to be addressed.

2. The proposal

2.1 How to preserve independence between pairs (as much as possible)

As already written in Section 1, the problem we are tackling is in the lack of independence of C for the pairs in **Ω**, as well as Γ for the pairs in **Ω**, due to the fact that we are considering observations based on *pairs* of units. Indeed, observations on units, e.g. for the key variables $X_1,\ldots X_k$, can be considered as generated independently on different units. Hence, we propose to explore **Ω** (for the distribution *m* estimation) sampling pairs in **Ω** so that independence between pairs can be preserved as much as possible. Sampling from **Ω** has already been proposed by Yancey (2004) in order to increase the percentage of matches in **Ω** when these are just a few, and in Fortini (2020), in order to reduce the comparison space without imposing constraints as blocks or filters.

A very simple idea is to take the data set with smallest size, say A, and follow these steps:

1. For each unit in A select randomly a unit in B, until all records in A create a pair with a corresponding record in B. If selection on B is performed without repetitions, the set of variables (*C*, Γ) on the selected pairs will be independent because pairs are formed by distinct units.
2. Iterate steps 1 for a number *T* of times.

Selection can be in many different ways:

a. As already stated, records in B can be selected with or without repetition.
b. Selection on B can follow a uniform distribution on the units in B, or privilege the selection of units so that *m* can be explored in the quicker and most efficient way.

Finally, the *T* samples can be either analyzed separately in order to draw estimates of *m* that should finally be aggregated, or the *T* samples can be composed in just one big sample to be analyzed. The idea is to use the same procedure as described in Section 1, with the difference that the observed likelihood (2) is computed over the pairs in: *i)* each selected sample *t, t=1,…,T*; *ii)* the overall sample obtained composing the *T* samples.

Each approach has its pros and cons.

Selection without repetition preserves independence and the form of the likelihood in each sample *t*, while independence is further allowed in the composition of the *T* samples if already selected pairs are not allowed to get in

a new sample; anyway, it is computationally more cumbersome given the necessary restrictions to introduce. Selection with repetition from B may introduce some dependences among pairs, anyway they are far less than in **Ω**. Hence, the use of the product over all the pairs in the selected samples in the observed likelihood function is a better approximation of the true likelihood.

As already said, independence on *C* and on Γ is ensured for the pairs in each sample *t*, *t*=1,…,*T*: anyway, we already know that the number of pairs in **Ω** is not so large and expected number of matches in the samples is very low. This fact may jeopardize the efficiency of the EM algorithm in finding the components of the mixture between the distributions *m* and *u*, in each sample *t*. Given that *m* is the most difficult distribution to estimate, the proposal is to inflate the possibility to include matches in the selection procedure. One possibility is to select the units in B in step 1 according to a distribution that depends on the comparisons Γ: given *a* in A, the higher the number of equal key variables, the higher the probability to select *b* in B. Our simulations take into account all these situations

2.2 Estimation of the other parameters

The distribution *u* can be computed from the distinct data sets A and B. In general, considering just one of the key variables and assuming it is categorical with *k* categories, it is possible to say:

$$u_j(1) = P\left(\Gamma_{X_j}(a,b) = 1 | C = 0\right) = \sum_{k=1}^{K} P_A(X_j = k) P_B(X_j = k)$$

for the independence between units (a is not b). Assuming independence between the key variables on distinct units (non matches) allows us to estimate the overall *u* for the whole vector of *K* key variables.
As far as *p* is concerned, this is the proportion of expected matches in the *T* samples as a whole. This percentage may be inflated if selection of *b* privileges the domain of Γ with a large number of concordances between key variables. Anyway, it can still be computed, by means of the Bayes theorem, the probability of having a match given a specific γ:

$$P(C = 1 | \Gamma = \gamma) = \frac{pm(\gamma)}{pm(\gamma) + (1-p)u(\gamma)}$$

This probability can be applied on the set of records with each specific γ in order to derive the matches in A and B.

As a matter of fact, estimation of *m* and *p* by means of the EM algorithm on the selected pairs as suggested in Section 2.1 and of *u* on the key variables as estimated in A and B respectively completes the estimation of the probability of errors and the posterior probability of being a match given a comparison by a simple plug in.

2.3 Interaction with other problems

As already remarked elsewhere in the literature (Yancey (2004), Fortini (2020)), one prominent problem in parameter estimation is the fact that **Ω** consists of an asymmetric partition in matched and non-matched pairs: when a record can be matched with at most one record of the other file, the relative frequency of matched pairs in **Ω** cannot be more than the inverse of the largest file size. Hence, matched pairs become very rare the larger are the files, if compared to the size of **Ω**. When there is such a disproportion in the partition of **Ω** between matches and non-matches, estimation of the distribution *m* can be extremely inaccurate. How this problem interacts with the use of the likelihood function (2) affected by lack of independence in some of the pairs?

3. Exercise on real data

3.1 Description of the data

We have considered two files for which we already know the real exact matches: the size of the two files are 25343 and 24613. The common variables are of different nature: there are high quality variables with a scarce discriminant power (gender) and others with less quality but high discriminant power (surname, year of birth, day of birth) and variables with intermediate discriminant power (month of birth).

3.2 Description of the simulation

We have selected files A and B of size 50 (for both) and 200 (for both), with an overlap of matching units equal to 20% for both sizes or 80%. Hence, we have investigated a total of 4 different combinations between file size and file overlap. The selected matching variables follow this schema:

- Schema 1: surname, sex, month and year of birth.
- Schema 2: name, surname, sex and month of birth.
- Schema 3: name, surname, sex and year of birth.
- Schema 4: name, surname, month and year of birth.

Hence, the 4 schema can be considered as ordered from the less (schema 1) to the most (schema 4) discriminating set of matching variables. This means that schema 1 can admit more equalities between the matching variables, that induce a larger number of dependencies in **Ω**.

3.2.1 Experimental outcomes

Table 3.2.1-1 shows the actual *m* distribution, known in advance, for Schema 2 in the different scenarios. The EM algorithm applied on **Ω** gives results far from the truth, especially in the case of files sized 200 and 20% of overlap (this is the case when matches are rare in **Ω**, resulting in a very poor performance of the EM algorithm).

Table 3.2.1-1
Distribution of $m(\gamma)$ in the 4 simulated scenarios and corresponding estimates by the EM algorithm on Ω, Schema 2

Files size	**50**				**200**			
Overlap	**20%**		**80%**		**20%**		**80%**	
γ	$m(\gamma)$	$\widehat{m_\Omega}(\gamma)$	$m(\gamma)$	$\widehat{m_\Omega}(\gamma)$	$m(\gamma)$	$\widehat{m_\Omega}(\gamma)$	$m(\gamma)$	$\widehat{m_\Omega}(\gamma)$
(0 0 0 0)	0	0,00000	0	0,00001	0	0,00679	0	0,00006
(0 0 0 1)	0	0,00191	0	0,00006	0	0,00157	0	0,00043
(0 0 1 0)	0,05	0,00000	0	0,00035	0,025	0,26015	0,00625	0,00480
(0 0 1 1)	0	0,02493	0	0,00264	0	0,05999	0,0125	0,03222
(0 1 0 0)	0	0,00000	0	0,00007	0	0,00098	0	0,00024
(0 1 0 1)	0	0,00953	0	0,00052	0,05	0,00023	0,00625	0,00164
(0 1 1 0)	0	0,00000	0	0,00283	0	0,03746	0,03125	0,01839
(0 1 1 1)	0,15	0,12425	0,05	0,02136	0,15	0,00864	0,11875	0,12358
(1 0 0 0)	0	0,00000	0	0,00030	0	0,01128	0	0,00029
(1 0 0 1)	0	0,00999	0	0,00223	0	0,00260	0	0,00192
(1 0 1 0)	0	0,00000	0	0,01222	0,025	0,43212	0	0,02165
(1 0 1 1)	0,1	0,13027	0,025	0,09225	0,05	0,09965	0,06875	0,14544
(1 1 0 0)	0	0,00000	0	0,00239	0	0,00162	0	0,00110
(1 1 0 1)	0,05	0,04979	0,025	0,01801	0,025	0,00037	0,0125	0,00738
(1 1 1 0)	0	0,00000	0,1	0,09883	0,05	0,06221	0,075	0,08303
(1 1 1 1)	0,65	0,64932	0,8	0,74594	0,625	0,01435	0,66875	0,55783

The sampling strategy on **Ω**, where the unit in B to be attached to a unit in A is selected according to a distribution whose probability increases with the number of 1s in γ, and unit from B can be selected with replacement, gives the estimates in Table 3.2.1-2 and 3.2.1-3. The EM algorithm applied on the whole sample obtained with the union of all the performed iterations gives satisfactorily results gives satisfactorily results when the overlap between A and B is 80%, while confirms to be poor when the overlap is 20%, no matter the number of iterations. This figures can suggest the hypothesis that the rareness of matches in **Ω** is more important than the lack of independence for C and Γ.

As an overall measure of divergence between the estimates and known *m* distribution, the chi-square indicator has been computed. Under schema 2, for both file sizes, the estimator on a sample of Ω pairs seems to be better than the one computed on the overall Ω, with the exception of a situation where matches are rare in Ω.

This approach has been replicated 8 times, in order to investigate whether the additional variability due to sampling from Ω affects results. A chi-square distance between the true and estimated *m* distributions for all schemas, when file size is 200 and 80% of the records in A or B forms a match, is represented in Table 3.2.1-5. This example shows that the estimation on the whole Ω is generally good. Indeed, the one based on a sample of independent pairs needs a larger number of iterations in order to beat the one computed on the whole Ω.

Table 3.2.1-2
Estimates by the EM algorithm on a sample of independent pairs, files size 50, under schema 2

Overlap	20%				80%			
Number T of samples	30	50	80	$m(\gamma)$	30	50	80	$m(\gamma)$
γ	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$		$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	
(0 0 0 0)	0,00000	0,00000	0,00132	0	0,00000	0,00009	0,00002	0
(0 0 0 1)	0,00519	0,00000	0,00248	0	0,00000	0,00065	0,00013	0
(0 0 1 0)	0,00000	0,00000	0,01603	0,05	0,00008	0,00284	0,00034	0,00625
(0 0 1 1)	0,02158	0,00000	0,03009	0	0,00176	0,02017	0,00273	0,0125
(0 1 0 0)	0,00000	0,00000	0,00099	0	0,00000	0,00034	0,00047	0
(0 1 0 1)	0,04630	0,04178	0,00185	0	0,00000	0,00241	0,00380	0,00625
(0 1 1 0)	0,00000	0,00000	0,01198	0	0,00101	0,01048	0,00970	0,03125
(0 1 1 1)	0,19261	0,15062	0,02248	0,15	0,02266	0,07452	0,07788	0,11875
(1 0 0 0)	0,00000	0,00000	0,01384	0	0,00000	0,00073	0,00016	0
(1 0 0 1)	0,01434	0,00000	0,02597	0	0,00000	0,00521	0,00127	0
(1 0 1 0)	0,00000	0,00000	0,16775	0	0,00301	0,02259	0,00323	0
(1 0 1 1)	0,05966	0,00000	0,31482	0,1	0,06731	0,16074	0,02597	0,06875
(1 1 0 0)	0,00000	0,00000	0,01034	0	0,00000	0,00270	0,00451	0
(1 1 0 1)	0,12796	0,17538	0,01941	0,05	0,00000	0,01924	0,03620	0,0125
(1 1 1 0)	0,00000	0,00000	0,12536	0	0,03869	0,08347	0,09230	0,075
(1 1 1 1)	0,53236	0,63222	0,23527	0,65	0,86548	0,59382	0,74129	0,66875

Table 3.2.1-3
Estimates by the EM algorithm on a sample of independent pairs, files size 200, under schema 2

Overlap	20%				80%			
Number T of samples	50	100	300	$m(\gamma)$	50	100	300	$m(\gamma)$
γ	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$		$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	
(0 0 0 0)	0,00000	0,00000	0,00000	0	0,00018	0,00006	0,00006	0
(0 0 0 1)	0,00000	0,00000	0,00000	0	0,00061	0,00043	0,00066	0
(0 0 1 0)	0,54281	0,36497	0,32086	0,025	0,00831	0,00397	0,00199	0,00625
(0 0 1 1)	0,14359	0,07971	0,06513	0	0,02809	0,02814	0,02244	0,0125
(0 1 0 0)	0,00000	0,00000	0,00000	0	0,00052	0,00031	0,00033	0
(0 1 0 1)	0,00000	0,00000	0,00000	0,05	0,00174	0,00217	0,00369	0,00625
(0 1 1 0)	0,05219	0,07755	0,04227	0	0,02385	0,02008	0,01111	0,03125
(0 1 1 1)	0,01381	0,01694	0,00858	0,15	0,08063	0,14220	0,12522	0,11875
(1 0 0 0)	0,00000	0,00000	0,00000	0	0,00107	0,00025	0,00030	0
(1 0 0 1)	0,00000	0,00000	0,00000	0	0,00361	0,00174	0,00334	0
(1 0 1 0)	0,17864	0,31194	0,41363	0,025	0,04942	0,01616	0,01004	0
(1 0 1 1)	0,04725	0,06813	0,08396	0,05	0,16708	0,11444	0,11314	0,06875
(1 1 0 0)	0,00000	0,00000	0,00000	0	0,00306	0,00124	0,00165	0
(1 1 0 1)	0,00000	0,00000	0,00000	0,025	0,01036	0,00881	0,01861	0,0125
(1 1 1 0)	0,01718	0,06628	0,05450	0,05	0,14186	0,08168	0,05602	0,075
(1 1 1 1)	0,00454	0,01448	0,01106	0,625	0,47963	0,57832	0,63140	0,66875

Table 3.2.1-4
Divergence between the true and the estimated *m* distributions under schema 2, when file sizes are 50 (number of drawn samples 80 for $\widehat{m_{\mathrm{I}}}(\gamma)$) and 200 (number of iteration 300 for $\widehat{m_{\mathrm{I}}}(\gamma)$)

Files size	**50**				**200**			
Overlap	**20%**		**80%**		**20%**		**80%**	
Estimator	$\widehat{m_{\Omega}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\Omega}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\Omega}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\Omega}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$
χ^2	0,058732	1,70963	0,11769	0,04513	100,6169	37,17019	0,12716	0,09658

In order to have a preliminary idea of the variance induced by pairs selection in $\widehat{m_{\mathrm{I}}}(\gamma)$ we have also performed 8 different replicates of the estimator when file size is 200, under schema 2 for different number of iterations: 50, 100, 200, 300, 400, 450. It seems that, increasing the number of iterations, the variability attenuates. Anyway, the range of values of the chi-square distance with respect to the true distribution for $\widehat{m_{\mathrm{I}}}(\gamma)$ is generally quite large, and includes the distance of the true distribution with the estimator based on Ω. The estimator based on all pairs in Ω seems to be subject to improvements, even if the strategies adopted so far are not able to constantly make a better work.

Table 3.2.1-5
Divergence between the true and the estimated *m* distributions, when files size is 200 and the overlap between A and B is 80%, under schema 2, for different replicates of $\widehat{m_{\mathrm{I}}}(\gamma)$ estimated with 50 and 300 draws of samples from Ω (divergence for $\widehat{m_{\Omega}}(\gamma)$ is 0,12716037)

T	χ^2							
	Replicates of the method based on draws from Ω							
	n° 1	n° 2	n° 3	n° 4	n° 5	n° 6	n° 7	n° 8
50	0,135127	0,264034	0,278951	0,424961	0,149647	0,106711	0,046630	0,152528
300	0,096584	0,136924	0,163529	0,162267	0,112410	0,079939	0,137170	0,100193

4. Comments and open issues

This paper deals with the problem of parameter estimation in a record linkage problem. Pairs of records for two files can be either a match or a non match. Knowledge is restricted to a set of matching variables in the two files that can be subject to errors or mistakes. The comparison of the matching variables is assumed to be generated by a random variable (r.v.) whose distribution is *m* for matches and *u* for non-matches. Pairs are generally assumed to be independent, and the likelihood function is built accordingly. Anyway, independence between pairs is generally not true. For this reason, given that the Cartesian product of the two files Ω is the only source of information on the status of the pairs, we propose to estimate *m* sampling from Ω so that the pairs in the sample can be independent.

The first question is consequently straightforward: is there a formal way to represent the actual likelihood based on the whole set of pairs in Ω?
The examples represented in this paper show that estimates obtained through the EM algorithm on the whole Ω can be quite distant from the actual ones. Anyway this does not seem to be always true. A first question is to investigate under what conditions the traditional estimator $\widehat{m_{\Omega}}(\gamma)$ is reliable (following also indications in Yancey (2004) and Fortini (2020) for its improvements).

The experiments suggest some aspects that need to be further investigated in order to be confirmed. These are the sentences we wish to investigate in the future.

- The more identifiable are the variables used as matching variables and the higher is the overlap between the two files, the best is $\widehat{m_{\Omega}}(\gamma)$.
- The less the matching variable are identifiable, the more the estimator $\widehat{m_{\mathrm{I}}}(\gamma)$ (as for schema 1) gives less distant results from the truth. Possibly this could be the effect of a larger number of equivalences between the matching variables, that induce more dependence between pairs observations.
- The larger the number of iterations in the estimator based on sampling from Ω, $\widehat{m_{\mathrm{I}}}(\gamma)$, the less distant seem to be the estimate from the truth. Is it possible to have results constantly better than $\widehat{m_{\Omega}}(\gamma)$ for any replicate of the same number of iterations for $\widehat{m_{\mathrm{I}}}(\gamma)$?

Furthermore, there are other aspects that deserve to be tackled.

- Sampling from Ω has been performed in just one situation: selection of independent pairs for each *a* in A, iteration of this selection and joint analyses of all the obtained pairs. It would be important to assess what happens if estimates are performed in each sample, and then a unique estimate is given.
- Perform simulation with different number of iterations, also much larger than the ones already computed in this paper.
- Perform simulations with different number of file size for A and B.
- It was assumed that the comparisons between the matching variables is independent for both matches and non-matches. Given the presence of zeros in the *m* distribution shown in the different tables, this is generally not true. Alternative models can be considered.

This is just a starting point: any comment and additional ideas are very welcome.

References

Chambers R, Kim G. (2015), “Secondary analysis on linked data”, in K. Harron, H. Goldstein, C. Dibben (eds.) *Methodological developments in record linkage*, New York: Wiley, pp 83-108.

Cibella N., Scannapieco M., Tosco L., Tuoto T., and Valentino L. (2012), “Record Linkage with RELAIS: Experiences and Challenges”. *Revista Estadistica Española*, 179, pp 311-328.

Fellegi, I. P., and Sunter A. B. (1969), "A Theory for Record Linkage". *Journal of the American Statistical Association*, 64, pp. 1183-1210.

Fortini, M. (2020), “An Improved Fellegi-Sunter Framework for Probabilistic Record Linkage Between Large Data Sets”, *Journal of Official Statistics*, vol. 36(4), pp. 803-825

Tuoto T. (2016), “New proposal for linkage error estimation”, *Statistical Journal of the IAOS* 32, pp. 413–420.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 667-671.

Yancey, W.E. (2004), “Improving EM Algorithm Estimates for Record Linkage Parameters”, Research Report Series, Statistics #2004-01, U.S. Census Bureau, Washington, U.S.A.