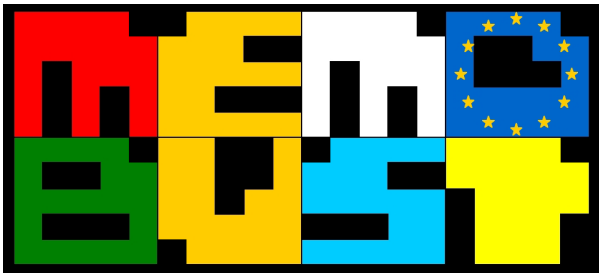

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Fellegi-Sunter and Jaro Approach to Record Linkage

Contents

General section

1. Summary

The Fellegi and Sunter method is a probabilistic approach to solve record linkage problem based on decision model. Records in data sources are assumed to represent observations of entities taken from a particular population (individuals, companies, enterprises, farms, geographic region, families, households…). The records are assumed to contain some attributes identifying an individual entity. Examples of identifying attributes are name, address, age and gender when dealing with people; style (or name) of a firm, legal form, address, number of local units, number of employees, turnover value when dealing with businesses. According to the method, given two (or more) sources of data, all pairs coming from the Cartesian product of the two sources has to be classified in three independent and mutually exclusive subsets: the set of matches, the set of non-matches and the set of pairs requiring manual review. In order to classify the pairs, the comparisons on common attributes are used to estimate for each pair the probabilities to belong to both the set of matches and the set of non-matches. The pair classification criteria is based on the ratio between such conditional probabilities. The decision model aims to minimise both the misclassification errors and the probability of classifying a pair as belonging to the subset of pairs requiring manual review.

2. General description of the method

Record linkage consists in matching the records belonging to different data sets when they correspond to the same unit. Records in data sources are assumed to represent observations of entities taken from a particular population (individuals, companies, enterprises, farms, geographic region, families, households…). The records are assumed to contain some attributes (variables) identifying an individual entity. Examples of identifying attributes are name, address, age and gender. Let A and B be two data sets, partially overlapping and containing the same type of units, of size N_A and N_B respectively. Suppose also that the two files consist of vectors of variables (X_A,Z_A) and (X_B,U_B), either quantitative or qualitative, assuming that X_A and X_B are sub-vectors of common attributes, called key variables or matching variables in what follows, so that any single unit is univocally identified by an observation x. The goal of record linkage is to find all the pairs of units $(a,b) \in \Omega=\{(a,b): a \in A, b \in B\}$, such that a and b refer actually to the same unit (a=b). Hence, a record linkage procedure can be considered as a decision model based on the comparison of the key variables; for each single pair of records either one of the following decisions can be taken: link, possible link and non-link. Since the key variables can be prone both to measurement errors and misreporting, the record linkage problem is far from being a trivial one and Fellegi and Sunter (1969) propose an approach to the probabilistic record linkage based on decision model to minimise the incidence of both the non-decision area and false and missed links.

Let us consider $\Omega = \{(a,b), a \in A \text{ and } b \in B\}$ of size $N=N_A \times N_B$. This method considers all the pairs as a sample of $N_A \times N_B$ records independently generated by a mixture of two distributions: one for the matched pairs and the other for the unmatched ones. The linkage between A and B can be defined as the problem of classifying the pairs that belong to Ω in two subsets M and U independent and mutually exclusive, such that:

M is the set of matches (a=b)

U is the set of non-matches (a≠b)

Actually, the model assumption fails to be true for the sample defined by the set of $N_A \times N_B$ records for the two data sets to link. In that case, it is not possible to state that comparison variables are independently generated by appropriate distributions. For more details about this weakness, see Kelley (1984). It is not yet clear how the failure of this independence hypothesis affects the record linkage results.

In order to classify the pairs, K common identifiers (called matching variables)

$$\mathbf{X}_1^A \quad \mathbf{X}_2^A \quad ... \quad \mathbf{X}_K^A ; \qquad \mathbf{X}_1^B \quad \mathbf{X}_2^B \quad ... \quad \mathbf{X}_K^B$$

have to be chosen (the variables with the same subindex are comparable). So, for each pair, a comparison vector $\gamma = \{\gamma_1, \gamma_2, ..., \gamma_K\}$ can be defined, by means of distance functions applied to matching variables for each pair. For instance, Fellegi and Sunter consider the binary comparison vector

$$_{(a,b)}\gamma_k = \begin{cases} 1 & \text{if } X_k^A = X_k^B \\ 0 & \text{otherwise} \end{cases}$$

For an observed comparison vector γ in Γ, the space of all comparison vectors, $m(\gamma)$ is defined to be the conditional probability of observing γ given that the record pair is a true match: in formula $m(\gamma) = P\left(\gamma \middle| (a,b) \in M\right)$. Similarly, $u(\gamma) = P\left(\gamma \middle| (a,b) \in U\right)$ denotes the conditional probability of observing γ given that the record pair is a true non-match.

There are two kinds of possible misclassification errors: false matches and false non-matches. The probability of false matches is:

$$\mu = P\left(M^* \middle| U\right) = \sum_{\gamma \in \Gamma} u(\gamma) P(M^* \mid \gamma)$$

and the probability of a false non-matches is:

$$\lambda = P\left(U^* \middle| M\right) = \sum_{\gamma \in \Gamma} m(\gamma) P(U^* \mid \gamma)$$

where M* and U* are the sets of estimated matches and estimated non-matches, respectively. For fixed values of μ and λ, Fellegi and Sunter define the optimal linkage rule as the rule that minimises the probability of assigning a pair in the set of no-decision Q, that is the set of pairs requiring clerical review so to be solved. The optimal rule is a function of the probability ratio

$$r = \frac{P\left(\gamma \middle| (a,b) \in M\right)}{P\left(\gamma \middle| (a,b) \in U\right)} = \frac{m(\gamma)}{u(\gamma)}.$$

In practice, once the probabilities m and u are estimated, all the pairs can be ranked according to their ratio $r=m/u$ in order to detect which pairs are to be matched by means of this classification criterion based on the two thresholds T_m and T_u ($T_m > T_u$)

$$r_{(a,b)} > T_m \quad \Rightarrow \quad (a,b) \in M^*$$
$$T_m \geq r_{(a,b)} \geq T_u \quad \Rightarrow \quad (a,b) \in Q$$
$$r_{(a,b)} < T_u \quad \Rightarrow \quad (a,b) \in U^*$$

- those pairs for which *r* is greater than the upper threshold value can be considered as linked

- those pairs for which *r* is smaller than the lower threshold value can be considered as not-linked

The thresholds are assigned solving equations that minimise both the size of the set Q and the false match rate (FMR) and false non-match rate (FNMR).

$$FMR = \sum_{\gamma \in \Gamma} u(\gamma) P(M^* \mid \gamma) = \sum_{\gamma \in \Gamma_{M^*}} u(\gamma) \quad \text{where} \quad \Gamma_{M^*} = \{\gamma : T_m \leq m(\gamma)/u(\gamma)\}$$
$$FNMR = \sum_{\gamma \in \Gamma} m(\gamma) P(U^* \mid \gamma) = \sum_{\gamma \in \Gamma_{U^*}} m(\gamma) \quad \text{where} \quad \Gamma_{U^*} = \{\gamma : T_u \geq m(\gamma)/u(\gamma)\}$$

2.1 Estimation of matching probabilities

In order to apply the model for record linkage described in the previous section, a method for estimating the likelihood ratio *r=m/u* is required. In their seminal paper, Fellegi and Sunter define a system of equations for estimating the parameters of the distributions for matched and unmatched pairs, based on the method of moments; it gives estimates in closed form when the comparison variables are at least three. Currently, the most widespread method for estimating the conditional probabilities *m* and *u* is the expectation-maximisation (EM) algorithm (Dempster et al., 1977), in the record linkage field first used by Jaro (1989). This is why the presented method is called the Fellegi-Sunter and Jaro one. According to this approach, the frequency distribution of the observed patterns γ is viewed as a mixture of the matches m(γ) and non-matches u(γ) distributions

$$\begin{aligned} P(\gamma) &= P(\gamma \mid (a,b) \in M) P((a,b) \in M) + P(\gamma \mid (a,b) \in U) P((a,b) \in U) \\ &= m(\gamma) \cdot p + u(\gamma) \cdot (1-p) \end{aligned}$$

where p=P(M). This means to consider a latent variable C, indicating the actual unknown matching status of the record pair, that takes value 1 corresponding to a match with probability *p* and value 0 corresponding to non-match with probability 1-*p*.

The joint distribution of the observations γ and the latent variable C is given by:

$$\mathrm{P}(\mathrm{C} = \mathrm{c}, \gamma) = \left[\mathrm{p} m(\gamma)\right]^c \left[(1 - \mathrm{p}) u(\gamma)\right]^{1-c} \quad .$$

Jaro restricts to 0/1 values the possible outcomes for the comparison vector γ, as in the previous Fellegi and Sunter model, and assumes conditional independence of the γ_k . These assumptions are currently often made in order to simplify the parameter estimation; in this case the likelihood function for $m_k(\gamma)$, $u_k(\gamma)$ (k=1,…,K) and p can be written as:

$$L = \prod_{(a,b)} \left[\mathrm{p} m(\gamma^{(a,b)})\right]^{c^{(a,b)}} \left[(1 - \mathrm{p}) u(\gamma^{(a,b)})\right]^{1-c^{(a,b)}} \quad .$$

The EM algorithm uses maximum likelihood estimates of $m_k(\gamma)$, $u_k(\gamma)$ and p to estimate the unobserved c. The EM algorithm needs initial estimates of $m_k(\gamma)$, $u_k(\gamma)$ and p and then iterates. Generally, the EM algorithm solutions don't depend on the initial values.

Under the conditional independence assumption the likelihood ratio r is given by:

$$r = \frac{P(\gamma|M)}{P(\gamma|U)} = \prod_{k=1}^{K} \frac{P(\gamma_k|M)}{P(\gamma_k|U)} = \prod_{k=1}^{K} \frac{m_k}{u_k}.$$

Even conditional independence assumption works well in most of the practical applications, it cannot be sure that this hypothesis is automatically satisfied. Some authors (Winkler 1989, and Thibaudeau 1989) extend the standard approach by means of log-linear models with latent variable by introducing appropriate constraints on parameters so to overcome to some extent conditional independence assumption. In these cases, however, it is not sure if the best model in term of fitting could be also considered as the most accurate in terms of linkage results and errors. See item 2 of the following section 11 (Variants of the method) for more details.

The Fellegi–Sunter and Jaro approach is heavily dependent on the accuracy of m(γ) and u(γ) estimates. Misspecifications in the model assumptions, lack of information and other problems can cause a loss of accuracy in the estimates and, as a consequence, an increase of both false matches and non-matches.

For this reason the appropriate thresholds are often identified mainly through empirical methods which need of scrutiny by experts, such as a diagram of the weights distribution as the one showed in the figure below.

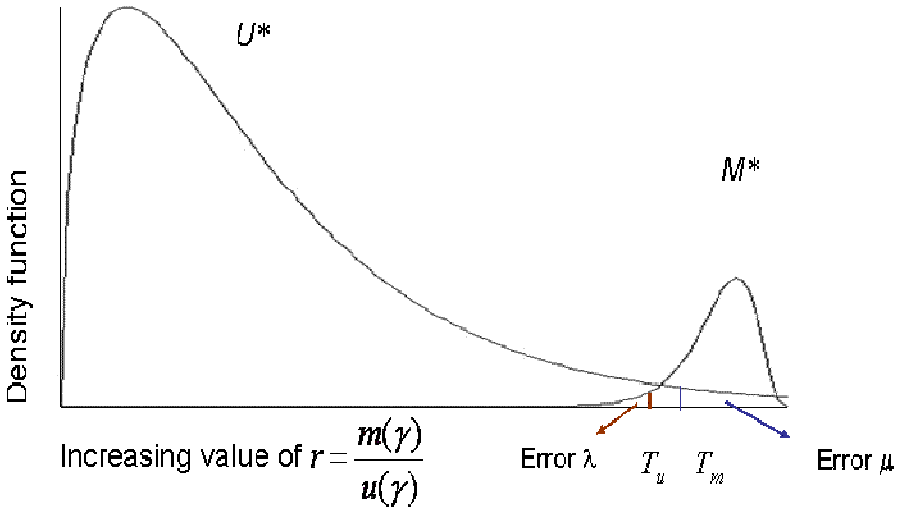

Figure 1. The mixture model for m- and u-distributions

3. Preparatory phase

Probabilistic record linkage, as proposed by Fellegi–Sunter and Jaro, is a complex procedure that can be decomposed in many different phases. The actual probabilistic record linkage model, as described in the previous section, is tackled only in few steps, but, as a matter of fact, all the previous steps are necessary when considering the use of the Fellegi–Sunter and Jaro method in practical situations.

The main points faced in this section are treated in depth in the WP2 of the ESSnet on ISAD (integration of surveys and administrative data), Section 2.1 (Cibella et al., 2008a) and in the theme

module “Micro-Fusion – Probabilistic Record Linkage”. In these papers, recommendations and suggestions are proposed as well, on the basis of the requirements of specific application.

The steps to be performed to apply the method can be summarised in the following list.

1) At first, a practitioner should decide which are the variables of interest available distinctly in the two files. To the purpose of linking the files, the practitioner should understand which variables are able to identify the correct matched pairs among all the common variables. These variables will be used as either matching or blocking variables.
2) The matching variables should be appropriately harmonised before applying any record linkage procedure. Harmonisation is in terms of variable definition, classification, codification, categorisation and so on.
3) When the files A and B are too large (as usually happens) it is appropriate to reduce the search space from the Cartesian product of the files A and B to a smaller set of pairs.
4) For probabilistic record linkage, after the selection of a comparison function, the suitable model should be chosen. This should be complemented by the selection of an estimation method, and possibly an evaluation of the obtained results. After this step, the application of a decision procedure needs the definition of the cut-off thresholds.
5) There is the possibility of different outputs, logically dependent on the aims of the match. The output can take the form of a one-to-one, one-to-many or many-to-many links.
6) The output of a record linkage procedure is composed of three sets of pairs: the links, the non-links, and the possible links. This last set of pairs should be analysed by trained clerks.
7) The final decision that a practitioner should consider consists in deciding how to estimate the linkage errors and how to include this evaluation in the analyses of linkage files.

4. Examples – not tool specific

4.1 Linkage between survey business data and administrative data

The following example is summarised from the paper Ichim et al. (2009). It is regarding the exploitation of administrative data in the NSIs. The administrative data may be useful in the sample design stage or in the estimation phase. Auxiliary information may also be useful in data validation and editing. The original paper reports several experiments undertaken to identify an optimal matching strategy for business data. The problem of linking survey and administrative data is addressed. The business survey data is the Small and Medium Enterprises (SME) survey while the administrative data source is the Balance Sheets (BIL). Small and Medium Enterprises sample survey (SME) is carried out annually by sending a postal questionnaire with the purpose of investigating profit-and-loss account of enterprises with less than 100 persons employed, as requested by SBS EU Council Regulation n. 58/97. The main variables of interest are Turnover, Value added at factor cost, Employment, Total purchases of goods and services, Personnel costs, Wages and salaries, Production value, etc. The frame for the SME survey is the Italian Statistical Business Register (ASIA). ASIA results from the logical and physical combination of data from both statistical and administrative sources. The Business Fiscal Turnover is provided from the Fiscal Register, this variable being a good proxy of the Turnover collected in SME. SME sample survey population of interest is about 4 millions of active

enterprises. Both the selection and estimation phases are based on the information available in ASIA, but a time lag exists between the reference years of SME and BR. The sample size is about 120.000 units. On the other side, the Italian limited enterprises are obliged to fill their financial statements according to the standards specified in the EEC fourth Directive and to transmit them to the Chambers of Commerce. The resulting database is called Balance sheets (BIL). This data source is actually the most used in the production of SBS estimates. In industry and services sectors there are about 500,000 limited enterprises which account for one half of the total employment. BIL data's coverage is 11.3% among 1-19 persons employed size class, it reaches 80.7% in the size class 20-99 and it is 96.2% among larger enterprises. The main aims of the linkage between SME and BIL are related to

- check and validation of survey results;
- obtain auxiliary information to deal with survey non-responses.
- update the frame of the survey reference population
- supply information to plan future survey wages.

The linkage procedures, described in the paper, mainly stress the efforts devoted to the pre-processing phase, the blocking step and the choice of the matching variables and the corresponding distances. Standardisation were applied to the streets typology and the types of the enterprises in both datasets. The unusual characters were deleted (®,¯, =, $, # , &, double spaces). The most frequent strings in the name of the enterprise were standardised, too. In order to select the matching variables, some descriptive statistics and correlations between the numerical variables were calculated. For the identified matching variables, different combinations of the several distance functions were tested to identify the best setting of the linkage experiment. Two reduction methods were applied in these experiments: blocking and sorted neighbourhood. The applied probabilistic model follows the Fellegi-Sunter and Jaro approach. In this work, the thresholds were derived from the probabilities of false nonmatch (0.90) and false match (0.95). Finally, the reduction one-to-one was solved as a linear programming problem. In both BIL and SME datasets, there is a unique identifier, namely the fiscal code. Even if the fiscal code may be subject to some errors, it was used for evaluating the quality of the record linkage through *precision* and *recall* (see section 21 below for these quality measures). Details on the several tests and results can be achieved in the full paper.

4.2 Estimating number of units in a population amount by capture-recapture method

The following example is summarised from the paper Cibella et al. (2008c). It involves data from the 2001 Italian Population Census and its Post Enumeration Survey (PES). The main goal of the Census was to enumerate the resident population at the Census date, 21/10/2001. The PES instead had the objective of estimating the coverage rate of the Census; it was carried out on a sample of enumeration areas (called EA in the following), which are the smallest territorial level considered by the Census. The size of the PES's sample was about 70000 households and 180000 individuals while the variables stored in the files are name, surname, gender, date and place of birth, marital status, etc. Correspondingly, comparable amounts of households and people were selected from the Census database with respect to the same EAs. The PES was based on the replication of the Census process inside the sampled EAs and on the use of a capture-recapture model (Wolter , 1986) for estimating the hidden amount of the population. In order to apply the capture-recapture model, after the PES enumeration of the statistical units (households and people), a record linkage between the two lists of

people built up by the Census and the PES was performed. In this way the rate of coverage, consisting of the ratio between the people enumerated at the Census day and the hidden amount of the population, was obtained.

The Fellegi–Sunter and Jaro linkage procedure, as described in previous sections, is applied on two sub- sets of size 8000 records, corresponding to the EAs of Rome. As matching variables all the strongest identifiers were used: name and surname, gender, day, month, and year of birth. Even if, generally, string variables as name and surname can complicate the linkage process due to diminutives or synonyms, in this example they didn't need further work due to their high quality level. So, the equality were applied as comparison function. The parameters of the Fellegi-Sunter probabilistic model were estimated via the EM algorithm. Two thresholds were fixed in order to individuate the three sets of Matches, of Unmatches and of Possible Links. The upper threshold was fixed assigning to the set of Matches all the pairs with the likelihood ratio corresponding to estimated matching probabilities higher than 0.99; the set of the possible links was created fixing the lower threshold level with the likelihood ratio corresponding to the estimated matching probability lower than 0.50. The pairs falling into the set of the Possible Links were assigned to the set of Matches without clerical supervision of the results.

A blocking phase was used considering as blocking variable the month of birth of the household head. In this way 12 blocks were created, plus a residual block formed by the units with missing information about the month of birth of the household header. The resulting blocking size are quite similar and homogeneous. The overall match rate is equal to 88%, the false match rate is 0.5% and the false non-match rate is 12%. Those results are comfortable and quite optimistic if compared with those coming from the scientific community, when a record linkage is performed in analogous conditions in terms of identification variables, number of matched records, kind of matched units. The results have to be regarded also more optimistic considering the unsupervised possible link data processing. Anyway, when the linkage is finalised to evaluate coverage rate, as in Census Post Enumeration Survey, the value of the false non-match rate has to be as small as possible and the resulting 12% false non-match rate is too high. In this situation, a further linkage procedure should be applied to the records non-linked at the first time, if it is possible without using blocking phase, so to minimise the risk of losing matches. The estimates of the Census coverage rate through capture-recapture model has required to match Census and PES records, assuming no errors in matching operations. Therefore the linkage between the two sources was both deterministic and probabilistic and the results was checked manually; all the linkage operations lasted several working days. Due to the accuracy of the matching procedures adopted, we know the true linkage status of all candidate pairs. In this way we can evaluate the effectiveness of the Fellegi-Sunter and Jaro linkage method in terms of match rate, false match rate and false non-match rate.

4.3 Estimating number of under coverage farms in Agricultural Census

The capture-recapture model introduced in the previous example has been also applied for the estimation of the unknown true number of farms, or, equivalently, for the estimate of the under-coverage rate of the Agricultural census. With respect to the previous example, the general workflow of the linkage procedure is the same, but different problem arise in comparing farms rather than people. In this case, the matching variables are the name (of the company name) of the farm or the name of its owner, the legal form, the utilised agricultural area, the address of the farm or the address

of its owner. Dealing with farms, the pre-processing procedures for name and address (in rural area) standardisation are very important and time-consuming.

4.4 Enriching and updating the information stored in different sources

The following example is summarised from the paper Cibella and Tuoto 2012. A record linkage is applied in order to study the fecundity of married foreign-women with residence in Italy. The Fellegi-Sunter and Jaro linkage method regards data referred to marriages with almost one of the married couple foreign and resident and data referred to babies born in the same Region in 2005-2006, from the registers of births. The size of each file is about 30000 records. The common variables are: fiscal code of the bride/mother, the 3-digit-standardised name and surname of both spouses/parents, the day/month/year of birth of the bridegroom/father and of the bride/mother, the municipality of the event (marriage/birth). Due to the data size, a data reduction method is needed, avoiding to deal with 900 millions of candidate pairs; analyses on the accuracy and of the frequency distribution of the available variables has limited the choice to the 3-digit-standardised name and surname of the bride/mother as blocking keys. The adopted blocking strategy is based on sorted neighbourhood method using as order variable the 6-digit-string of name and surname (composed from joining the 3-digit-standardised name and surname) over a window of size 15. The Fellegi-Sunter and Jaro method has been applied on the about 400000 candidate pairs produced by the sorted neighbourhood reduction, considering as matching variables: the 3-digit-standardised name of the mother and her day/month/year of birth. Equality function was used to compare the variables. The two thresholds to identify the tree sets of Matches, of Unmatches and of Possible Links were fixed in the following way: the upper threshold assigns to the set of Matches all the pairs with the likelihood ratio correspondent to estimated matching probability higher than 0.95; the lower threshold assigns to the set of the possible links all the pairs with the likelihood ratio correspondent to the estimated matching probability lower than 0.80. The procedure identified 567 matches and 457 possible matches. Among the matches, even 499 pairs have the same fiscal code or agree in all the bridegroom/father variables, while, among the possible matches, the concordance in the pairs is 25; so, totally, 592 true matches are identified by this procedure. This result can be compared with the total amount of pairs with common fiscal code in the files (they are 517 records).

5. Examples – tool specific

The examples reported in the previous section were carried out by using the RELAIS tool. It implements the Fellegi and Sunter method for record linkage, using the EM algorithm for the estimation of the conditional probabilities For the EM algorithm, the initial values of the parameters are m(g)=0.8 , u(g)=0.2 and p=0.1; the maximum number of iteration is 5.000 and the stop criterion is achieved when the difference between the estimates of two iterations is 0.000001.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Armstrong, J. and Mayda, J. E. (1993), Model-based estimation of record linkage error rates. *Survey Methodology* **19**, 137–147.

Cibella, N., Scanu, M., and Tuoto, T. (2008a), The practical aspects to be considered for record linkage. Section 2.1 of the *Report on WP2 of the ESSnet on Integration of Survey and Administrative Data* (http://www.cros-portal.eu/content/isad-finished).

Cibella, N. and Tuoto, T. (2008b), Quality assessments. Section 1.7 of the *Report on WP1 of the ESSnet on Integration of Survey and Administrative Data* (http://www.cros-portal.eu/content/isad-finished).

Cibella, N., Fortini, M., Scannapieco, M., Tosco, L., and Tuoto, T. (2008c) Theory and practice of developing a record linkage software. In: *Proceedings of the International Workshop "Combination of surveys and administrative data", 29-30 May, Vienna, Austria*.

Cibella, N. and Tuoto, T. (2012), Statistical perspectives on blocking methods when linking large data-sets. In A. Di Ciaccio et al. (eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets*, ISBN 978-3-642-21036-5, Springer-Verlag, Berlin Heidelberg.

Copas, J. R. and Hilton, F. J. (1990), Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A* **153**, 287–320.

Da Silva, A. D., Martins Romeo, O. S., Soares, T. S., and Xavier, V. L. (2011), Study of Record Linkage Software for the 2010 Brazilian Census Post Enumeration Survey. *Proceedings of the 58th ISI congress, 21-26 August, Dublin*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

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

Gill, L. (2001), Methods for automatic record matching and linkage and their use in national statistics. National Statistics Methodological Series No. 25, London (HMSO).

Heasman, D., Bailliel, M., Danielis, J., McLeod, P., and Elkin, M. (2011), Applications of record linkage to population statistics in the UK. Workshop of the ESSnet Data Integration, 24-25 November, Madrid, Spain.

Ichim, D., Casciano, C., and Seri, G. (2009), A linkage experiment between survey business data and administrative data. Workshop "2009 European Establishment Statistics", September 7-9, Stockholm, Sweden.

Jaro, M. A. (1989), Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* **84**, 414–420.

Kelley, R. B. (1984), Blocking considerations for record linkage under conditions of uncertainty. Statistical Research Division Report Series, SRD Research Report No. RR-84/19, Bureau of the Census, Washington, D.C.

Larsen, M. D. and Rubin, D. B. (2001), Iterative automated record linkage using mixture models. *Journal of the American Statistical Association* **96**, 32–41.

Scanu, M. (2008), Estimation of the distributions of matches and nonmatches. Section 1.5 of the *Report on WP1 of the ESSnet on Integration of Survey and Administrative Data* (http://www.cros-portal.eu/content/isad-finished).

Thibaudeau, Y. (1989), Fitting log-linear models when some dichotomous variables are unobservable. *Proceedings of the Section on statistical computing*, American Statistical Association, 283–288.

Thibaudeau, Y. (1993), The discrimination power of dependency structures in record linkage. *Survey Methodology* **19**, 31–38.

Thompson, G. (2011), Linking Information to the Australian Bureau of Statistics Census of Population and Housing in 2011. Workshop of the ESSnet Data Integration, 24-25 November, Madrid, Spain.

Winkler, W. E. (1989a), Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington D.C., 145–155.

Winkler, W.E. (1989b), Frequency-based matching in Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778–783 (longer version report rr00/06 at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1993), Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 274–279.

Wolter, K. (1986), Some coverage error models for Census data. *Journal of the American Statistical Association* **81**, 338–346.

Specific section

8. Purpose of the method

The purpose of Fellegi-Sunter and Jaro record linkage procedure is to identify the same real world entity that can be differently represented in data sources, even when unique identifiers are not available or are affected by errors. This operation is suitable when two or more partially or completely overlapping sets of data have to be integrated at micro level so as the information available in one frame for a unit can be linked to the information related to exactly the same unit stored in the other frame. The different frame can be statistical or coming from administrative data.

9. Recommended use of the method

1. The Fellegi-Sunter and Jaro method is recommended when unique identifiers are not available for all the units or when they are affected by errors. Regardless of the record linkage purposes, the following logic is adopted in extreme cases: when a pair of records is in complete disagreement on some *key* issues it will be almost certainly composed of different entities; conversely, a perfect agreement will indicate an almost certain match. All the intermediate cases, whether a partial agreement between two different units is achieved by chance or a partial disagreement between a couple of records relating to the same entity is caused by errors in the comparison variables, have to be properly resolved. This method, under the suitable conditions, solve these ambiguous situations.

10. Possible disadvantages of the method

1. The Fellegi-Sunter and Jaro approach is heavily dependent on the accuracy of $m(\gamma)$ and $u(\gamma)$ estimates. Misspecifications in the model assumptions, lack of information, inappropriate choices in the previous steps of the whole record linkage process and so on can cause a loss of accuracy in the estimates. Generally speaking, estimation cannot be reliable when one of the categories of the latent variable (the matches) is too rare. In general, the set of the matched pairs M should be large enough (generally, more than 5% of the overall set of $N_A \times N_B$ pairs). For instance, this is one of the motivations for the application of blocking procedures. However, in most practical cases, even when the parameter estimates are not very reliable, the linkage procedure is robust with respect to the identification of the matches, while it does not allow a reliable estimation of the matching errors.

11. Variants of the method

This section has been taken from the WP1 of the ESSnet on ISAD (integration of surveys and administrative data), Section 1.5 (Scanu 2008).

1. Independence between the comparison variables – This assumption is usually called the Conditional Independence Assumption (CIA), i.e., the assumption of independence between the comparison variables γ_j^{ab}, $j=1,\ldots,k$, given the match status of each pair (matched or unmatched pair). Fellegi and Sunter define a system of equations for estimating the parameters of the distributions for matched and unmatched pairs, based on the method of moments which gives estimates in closed form when the comparison variables are at least three. Jaro (1989)

solves this problem for a general number of comparison variables with the use of the EM algorithm (Dempster et al., 1977).

2. Dependence of comparison and latent variable defined by means of loglinear models – Thibaudeau (1989, 1993) and Armstrong and Mayda (1993) have estimated the distributions of the comparison variables under appropriate loglinear models of the comparison variables. They found out that these models are more suitable than the CIA. The problem is estimating the appropriate loglinear model. Winkler (1989, 1993) underlines that it is better to avoid estimating the appropriate model, because tests are usually unreliable when there is a latent variable. He suggests using a sufficiently general model, as the loglinear model with interactions larger than three set to zero, and incorporating appropriate constraints during the estimation process. For instance, an always valid constraint states that the probability of having a matched pair is always smaller than the probability of having a nonmatch. A more refined constraint is obviously the following:

$$p \leq \frac{n_A}{n_B \cdot n_A} = \frac{1}{n_B}.$$

Estimation of model parameters under these constraints may be performed by means of appropriate modifications of the EM algorithm, see Winkler (1993).

3. Iterative approaches – Larsen and Rubin (2001) define an iterative approach which alternates a model based approach and clerical review for lowering as much as possible the number of records whose status is uncertain. Usually, models are estimated among the set of fixed loglinear models, through parameter estimation computed with the EM algorithm and comparisons with "semi-empirical" probabilities by means of the Kullback-Leibler distance.

4. Other approaches – Different papers do not estimate the distributions of the comparison variables on the data sets to link. In fact, they use ad hoc data sets or training sets. These variants simplify the estimation procedure and can be applied in particular when the linkage is done for files that become available regularly and don't change too much in time. In this last case, it is possible to use comparison variables more informative than the traditional dichotomous ones. For instance, a remarkable approach is considered in Copas and Hilton (1990), where comparison variables are defined as the pair of categories of each key variable observed in two files to match for matched pairs (i.e., comparison variables report possible classification errors in one of the two files to match). Unmatched pairs are such that each component of the pair is independent of the other. In order to estimate the distribution of comparison variables for matched pairs, Copas and Hilton need a training set. They estimate model parameters for different models, corresponding to different classification error models.

12. Input data

1. Input data for the Fellegi-Sunter and Jaro method for record linkage are two or more microdata files referred, partially or completely, to the same units.

2. The input datasets have to contain three or more matching variables, with high level of identification power and quality (few errors, few missing data). Note that the number of

matching variables and some of their characteristics (as the number of categories and their rarity) influence the identification of links.

3. Another type of input of the method is the distance function used to compare each pair of records. This function must be appropriate for reporting the characteristics of the selected matching variables. The equality function is the most widespread. Distance functions based on string comparators (as Levenstein, Jaro, Jaro-Winkler, Soundex, 3grams) can be useful applied when the matching variables are names and are affected by typos or other kind of errors.
4. A further input of the method is the level of acceptable error rates.

13. Logical preconditions

1. Missing values
   1.
2. Erroneous values
   1.
3. Other quality related preconditions
   1.
4. Other types of preconditions
   1.

14. Tuning parameters

1. The acceptable levels of error rates are user-defined. These levels serve to assign the threshold values of the decision rule. Sometimes, due to the poor accuracy of the $m(\gamma)$ and $u(\gamma)$ estimates, the appropriate thresholds are often identified mainly through empirical methods which need scrutiny by experts.

15. Recommended use of the individual variants of the method

1. The model under the conditional independence assumption (CIA) has to be preferred if there is no evidence of marginal dependency among the matching variables and the linkage status, as usual.
2. When training set of data with the true matching status is available, for instance because an error-free identification code is available for a sub-set of records, the Copas and Hilton variant can be applied in order to improve the accuracy of the estimates.

16. Output data

1. The Fellegi and Sunter method produces a single set of data collecting the pairs in common in the two input datasets, i.e., the set of matches. In this dataset, for all matched pairs, all the original variables are available and more an output variable reporting the matching probability.

2. The method generally produces a file of possible links, i.e., pairs that need a manual review or further analyses in order to be assigned to the match set or to be discarded as non-matches.
3. The method also allows to create residual files, i.e., from the original datasets can be created reduced dataset composed of the records that haven't been linked.
4. Finally, the method allows to create the set of non-matched pairs, i.e., the file composed of the pairs that, according to the decision rules, are declared as non-matches. This file can be useful in order to investigate the false non-matches.

17. Properties of the output data

1. The main advantage in using Fellegi-Sunter and Jaro method to solve record linkage problem is the availability of the linkage probability for each pair assigned to the set of matches. This probability allows to evaluate the quality of the linkage and it has to be taken into account in the following phase of the whole process.

18. Unit of input data suitable for the method

Processing full data sets

19. User interaction - not tool specific

1.

20. Logging indicators

1. Number of records in Dataset1
2. Number of records in Dataset2
3. Number of matching variables considered in the model
4. Comparison function used for each variable
5. Error levels considered acceptable

21. Quality indicators of the output data

This section has been taken from the WP1 of the ESSnet on ISAD (integration of surveys and administrative data), Section 1.7 (Cibella and Tuoto, 2008).

1. The first indicator of the output data is the match rate, i.e., the total number of linked record pairs divided by the total number of true match record pairs. In order to compute the match rate, the total number of true matches has to be known. In alternative, when the total number of true matches in unknown and it is not possible to achieve it in different way, a maximum value of the indicator can be calculated as the ratio between the total number of linked record pairs and the number of records of the smallest of the two input datasets.
2. Another indicator is the false match rate is defined the number of incorrectly linked record pairs divided by the total number of linked record pairs. The false match rate corresponds to the well-known 1-α error in a one-tail hypothesis test. The estimate of such indicator is an output of the estimation step of the Fellegi-Sunter and Jaro method. In the epidemiological

field, instead of the false match rate, it is largely used the positive predictive value, defined as one minus the false match rate and corresponding to the number of correctly linked record pairs divided by the total number of linked record pairs.

3. One more indicator is the false non-match rate is defined as the number of incorrectly unlinked record pairs divided by the total number of true match record pairs. The false non-match rate corresponds to the β error in a one-tail hypothesis test. The estimate of such indicator is an output of the estimation step of the Fellegi-Sunter and Jaro method. In the epidemiological field, the sensitivity indicator is defined as the number of correctly linked record pairs divided by the total number of true match record pairs. It can be easily obtained from the false non-match rate.
4. A different performance measure is specificity, defined as the number of correctly unlinked record pairs divided by the total number of true non-match record pairs. The difference between sensitivity and specificity is that sensitivity measures the percentage of correctly classified record matches, while specificity measures the percentage of correctly classified non-matches.
5. In information retrieval the previous accuracy measures take the name of precision and recall. Precision measures the purity of search results, or how well a search avoids returning results that are not relevant. Recall refers to completeness of retrieval of relevant items. Hence, precision can be defined as the number of correctly linked record pairs divided by the total number of linked record pairs, i.e., it coincides with the positive predicted value. Similarly, recall is defined as the number of correctly linked record pairs divided by the total number of true match record pairs, i.e., recall is equivalent to sensitivity. As a matter of fact, precision and recall can also be defined in terms of non-matches.

22. Actual use of the method

1. The method is used in the linkage steps of the Post Enumeration Surveys of Agricultural Census in several countries (for instance in USA since 1985, in Italy since 2011).
2. The method is used in the linkage steps of the Post Enumeration Surveys of Population Census in several countries (in Italy, since 2011).
3. The method is used in linking information to the ABS Census of Population and Housing in 2011 (see Thompson, 2011).
4. The method is used in applications of record linkage to population statistics in the UK (see Heasman et al., 2011).
5. The method is used in the linkage steps of the 2010 Brazilian Census Post Enumeration Survey (see da Silva et al., 2011).
6. The method is used for building preparatory lists for the 2011 Population Census in Italy.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Data Fusion at Micro Level

2. Micro-Fusion – Object Matching (Record Linkage)
3. Micro-Fusion – Probabilistic Record Linkage

24. Related methods described in other modules

1.

25. Mathematical techniques used by the method described in this module

1.

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

1. RELAIS (Record linkage at Istat) is a toolkit providing a set of techniques for dealing with record linkage projects. It allows to dynamically select the most appropriate solution for each phase of record linkage and to combine different techniques for building a record linkage workflow of a given application. It is developed as an open source project. It is released under the EUPL license (European Union Public License) and it can be downloaded for free at http://www.istat.it/it/strumenti/metodi-e-software/software/relais with its User Guide, as well. It has been implemented by using two languages based on different paradigms: Java, an object oriented language, and R, a functional language. It is based on relational database architecture, mySql environment. The RELAIS project aims to provide record linkage techniques easily accessible to non-expert users. Indeed, the developed system has a GUI (Graphical User Interface) that on the one hand permits to build record linkage work-flows with a good flexibility. On the other hand it checks the execution order among the different provided techniques whereas precedence rules must be controlled. The current version of RELAIS provides several techniques to execute record linkage applications, in particular it allows to perform the Fellegi–Sunter and Jaro method for probabilistic record linkage, estimating the conditional matching probabilities via the EM algorithm. Moreover it provides different methods for search space reduction, several comparison functions, some metadata on the common variables in order to select them as matching or blocking variables. It runs under Windows and Linux environments.

28. Process step performed by the method

GSBPM Sub-process 5.1: Integrate data

Administrative section

29. Module code

Micro-Fusion-M-Fellegi-Sunter and Jaro Approach

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	11-05-2012	first version	Tiziana Tuoto	Istat
0.2	01-10-2012	second version	Tiziana Tuoto	Istat
0.2.1	04-10-2013	preliminary release		
0.3	09-10-2013	EB comments	Tiziana Tuoto	Istat
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:59