This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Probabilistic Record Linkage

Contents

General section

1. Summary

In this section the problem of probabilistic record linkage is explored. It can be also viewed as the weighted matching in case of an explicit use of probabilities. Generally speaking record linkage (or object matching, see also module on object matching) can be defined as the set of methods and practices aiming at accurately and quickly identify if two or more records, stored in sources of various type, represent or not the same real world entity. As usually data sources are hard to integrate due to errors or lacking information in the record identifiers, record linkage can be seen as a complex process consisting of several phases involving different knowledge areas. In research literature a distinction between deterministic (matching identifiers) and probabilistic approaches (matching with matching weights) is often made, where the former is associated with the use of formal decision rules while the latter makes an explicit use of probabilities for deciding when a given pair of records is actually a match but a clear separation between the two approaches is very difficult.

Compared with the deterministic approach, the probabilistic one can solve problems caused by bad quality data and can be helpful when differently spelled, swapped or misreported variables are stored in the two data files; the attention in this section is only devoted to the probabilistic record linkage approach which allows also to evaluate the linkage errors, calculating the likelihood of the correct match.

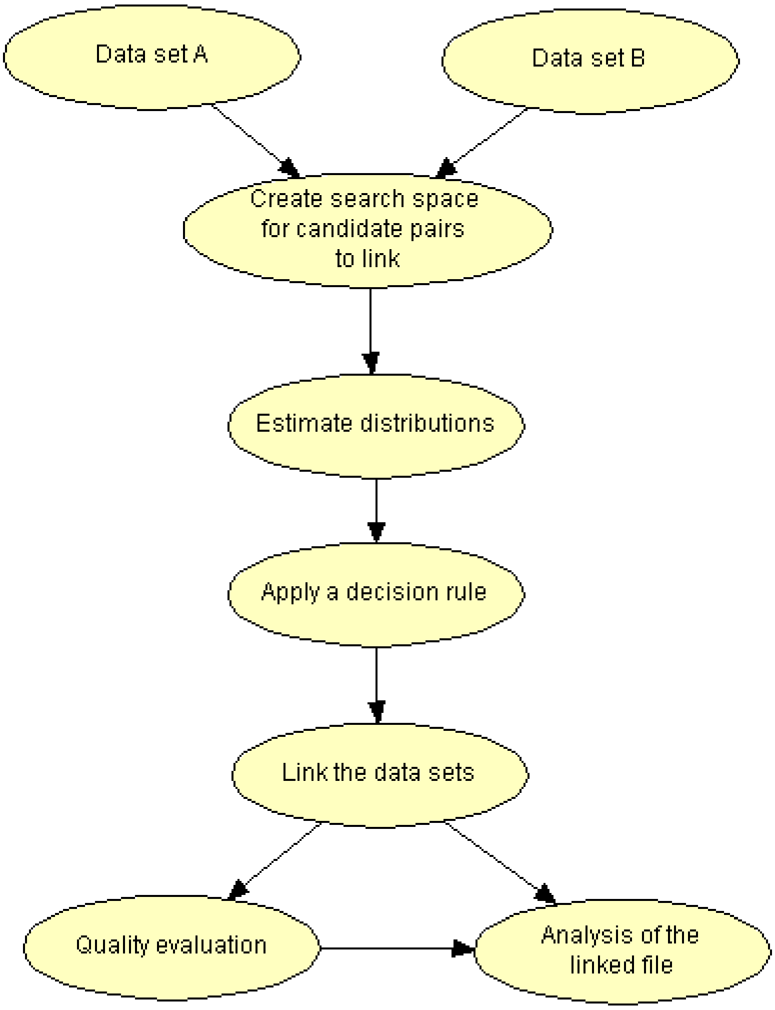
Generally speaking, the deterministic and the probabilistic approaches can be combined in a two-step process: firstly the deterministic method can be performed on the high quality variables then the probabilistic approach can be adopted on the residuals, the units not linked in the first step; however the joint use of the two techniques depends on the aims of the whole linkage project.

2. General description

Record linkage is widely performed in order to enrich, update or improve the information stored in different sources; to create a sampling list; to study the relationship among variables reported in different sources; to eliminate duplicates within a data frame; to assess the disclosure risk when releasing microdata files, etc. In official statistics, the advantages, in terms of quality and costs, due to the combined use of administrative data and sample surveys strongly encourage the researchers to the investigation of new methodologies and instruments to deal with record linkage projects and to identify quickly and accurately units across various sources. Since the earliest contributions to modern record linkage, dated back to Newcombe et al. (1959) and to Fellegi and Sunter (1969) where a more general and formal definition of the problem is given, there has been a proliferation of different approaches, that make use also of techniques based on data mining, machine learning, equational theory.

According to some authors (e.g., Statistic Canada) deterministic record linkage is defined just as the method that detects links if and only if there is a full agreement of unique identifiers or a set of common identifiers, the matching variables. Other authors backed up that in deterministic record linkage a pair is a link also if it satisfied some specific criteria a priori defined; actually not only the matching variables must be chosen and combined but also a threshold has to be fixed in order to establish whether a pair should be considered a link or not. Deterministic record linkage can be adopted, instead of probabilistic method, in presence of error-free unique identifiers (such a fiscal code) or when matching variables with high quality and discriminating power are available and can be combined so as to establish the pairs link status; in this case the deterministic approach is very fast and effective and its adoption is appropriate. From the other side, the rule definition is strictly dependent on the data and on the knowledge of the practitioners. Moreover, due to the importance of the matching variable quality, in the deterministic procedure, some links can be missed due to presence of errors or missing values in the matching variables; so the choice between the deterministic and probabilistic methods must take into account "the availability, the stability and the uniqueness of the variables in the files" (Gill, 2001). It is important also to underline that, in a deterministic context, the linkage quality can be assessed only by means of re-linkage procedures or accurate and expensive clerical reviews.

Probabilistic record linkage is a complex procedure that could be decomposed in different steps. For each step we can adopt different techniques. The following workflow has been taken from the WP1 of the ESSnet on ISAD (integration of surveys and administrative data), Section 1.2 (Cibella et al., 2008a) and represents the whole record linkage process:



In a linking process of two already harmonised data sets, namely A and B of size NA and NB respectively, let us consider the search space $\Omega = \{(a,b),\ a \in A \text{ and } b \in B\}$ of size $N=NA \times NB$. 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 ($a \neq b$)

2.1 Search space reduction

When dealing with large datasets, comparing all the pairs (a; b), a belonging to A and b belonging to B, in the cross product is almost impracticable and this causes computational and statistical problems. To reduce this complexity it is necessary to reduce the number of pairs (a; b) to be compared. There are many different techniques that can be applied to reduce the search space; blocking and sorted neighbourhood are the two main methods. Blocking consists of partitioning the two sets into blocks and of considering linkable only records within each block. The partition is made through blocking keys; two records belong to the same block if all the blocking keys are equal or if a comparison function applied to the blocking keys of the two records gives the same result. Sorted neighbourhood sorts the two input files on a blocking key and searches possible matching records only inside a window of a fixed dimension which slides on the two ordered record sets.

2.2 The matching variables

Starting from the reduced search space, we can apply different decision models that enable to classify pairs into M, the set of matches and U, the set of non-matches.

In this section the probabilistic approach is formalised according to the Fellegi and Sunter theory which is described in details in the module "Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage". The method requires an estimation of the model parameters that can be performed via the EM algorithm, Bayesian methods, etc.

In order to classify the pairs, some k common identifiers, either quantitative or qualitative, called matching variables,

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

have to be chosen so that, for each pairs, a comparison vector $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_K\}$ can be defined, where

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

It is important to choose matching variables that are as suitable as possible for the considered linking process. The matching attributes are generally chosen by a domain expert. If unique identifiers are available in the linkable data sources, the easiest and most efficient way is to use these ones as link variables; but very strict controls need to be made in case of using numeric identifiers alone. Variables like name, surname, address, date of birth, can be used jointly instead of using each of them separately; in such a way, one can overcome problems like the wide variations of the name spelling or the changes in surname depending on the variability of the marital status. It is evident that the more heterogeneous are the items of a variable, the higher is its identification power; moreover, if missing cases are relevant in a field it is not useful to choose it as a matching variable.

2.3 *The comparison functions*

The comparison functions are used to compute the distance between records compared on the values of the chosen matching variables. Some of the most common comparison functions are (for a review, see Koudas and Srivastava, 2005):

a) equality that returns 1 if two strings fully agree, 0 otherwise;

b) edit distance that returns the minimum cost in terms of insertion, deletions and substitutions needed to transform a string of one record into the corresponding string of the compared record;

c) Jaro counts the number of common characters and the number of transpositions of characters (same character with a different position in the string) between two strings;

d) Hamming Distance that computes the number of different digits between two numbers;

e) Smith-Waterman that uses dynamic programming to find the minimum cost to convert one string into the corresponding string of the compared record; the parameters of this algorithm are the insertions cost, deletions cost and transposition cost;

f) TF-IDF that is used to match strings in a document. It assigns high weights to frequent tokens in the document and low weights to tokens that are also frequent in other documents.

2.4 *The decision rule and parameters estimation*

Following Fellegi and Sunter (1969), the ratio

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

between the probabilities of γ given the pair (a,b) membership either to the subset M or U is used so as classifying the pair. Fellegi and Sunter proposed an equation system to achieve the explicit formulas for the estimates of m(g) and u(g) when the matching variables are at most three (see the method module “Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage” for details).

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 a decision rule based on two thresholds Tm and Tu (Tm > Tu)

$$\begin{array}{rcl} r_{(a,b)} > T_m & \Rightarrow & (a,b) \in M^* \\ T_m \geq r_{(a,b)} \geq T_u & \Rightarrow & (a,b) \in Q \\ r_{(a,b)} < T_u & \Rightarrow & (a,b) \in U^* \end{array}$$

- 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 chosen so as to minimise two types of possible errors: false matches (FMR, or mismatch, false positive match, Type I error, see module on object matching) and false non-matches

(FNMR, missed match, false negative match, Type II error) that refers respectively, as stated above, to the matched records which do not represent the same entity and to the unmatched records not correctly classified, that imply truly matched entities were not linked.

The Fellegi and Sunter 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.

Armstrong and Mayda (1993) assume that the frequency distribution of the observed patterns γ is a mixture of the matches m(γ) and non-matches u(γ) distributions

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

where p=P(M).The latent variable C denotes the unknown linkage status and is equal to 1 in case of a match, with the probability p, so the joint distribution of the observations γ and the latent variable C=c (c=(0,1)) is given by:

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

Since vector C is not directly measurable, the maximum likelihood estimates of parameters mk(γ), uk(γ) and p can be obtained through EM algorithm (Dempster et al., 1977) as proposed in Jaro (1989). A simplification of the estimates, which is often made in order to keep easier the parameters estimation, is the so called local independence assumption, where r is written as

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

Even local independency assumption works well in most of the practical application, 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 local independence assumption. In these cases, however, it is not sure if the best model in terms of fitting could be also considered as the most accurate in terms of linkage results and errors.

2.5 *Alternative probabilistic record linkage methods*

Also other approaches could be considered in the estimation of parameters (the following description has been taken from the WP1 of the ESSnet on ISAD, Section 1.5 (Cibella et al., 2008a)):

The Bayesian approaches – Fortini et al. (2001, 2002) look at the status of each pair (match and non-match) as the parameter of interest. For this parameter and for the parameters of the latent variables that generate matches and non-matches they define natural prior distributions. The Bayesian approach consists in marginalising the posterior distribution of all these parameters with respect to the parameters of the comparison variables (nuisance parameters). The result is a function of the status of the different pairs that can be analysed for finding the most probable configuration of matched and unmatched pairs.

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.

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. 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.

2.6 Record linkage quality

As not every record matched in the linkage process refers to the same identity, at the end of the record linkage process is really important to assess the "quality" of the procedure establishing whether a match is a "true one" or not. In other words, during a linkage project is necessary to classify records as true link or true non link, minimising, according to the Fellegi and Sunter theory, the two types of possible errors: false matches and false non-matches that refers respectively, as stated above, to the matched records which do not represent the same entity and to the unmatched records not correctly classified, that imply truly matched entities were not linked. False non-matches of matching cases are the most critical ones because of the difficulty of checking and detecting them. In general, it's not easy to find automatic procedures to estimate these types of errors so as to evaluate the quality of record linkage procedures. The same accuracy indicators are also used in the research field of information retrieval, although they are usually named precision and recall and can be evaluated even if the linkage procedure is performed through techniques different from the probabilistic one, as for instance supervised or unsupervised machine learning (Elfeky et al., 2003).

Errors can also be introduced by the choices that are made in the matching process itself. For instance, an incorrect or overly limited matching key may be used, the way in which the weights are calculated may be incorrect, or the cut-off values against which the weights are set off may lead to matching errors.

Also the time consumed by software programmes and by the number of records that require manual review could be considered additional performance criteria for the process (see the WP1 of the ESSnet on ISAD, Section 1.7 for details (Cibella et al., 2008a)) or also, as stated in the module "Micro-Fusion – Object Matching (Record Linkage)", all the choices that are made in the matching process itself could have an impact on the record linkage quality (e.g., an incorrect or overly limited matching key).

The final step of the whole record linkage process is devoted to the subsequent studies of the linked data set, taking in mind that this file can contain matching errors and all the derived analysis could be affected by the two types of errors: the percentage of incorrect acceptance of false matches and, on the other hand, the incorrect rejection of true matches. Record linkage procedures must deal with the existing trade-off between these two errors and/or measure the effects on the parameter estimates of the models that are associated to the obtained files.

The following desciption has been selected from Section 1.8 of the WP1 of the ESSnet on ISAD (Cibella et al., 2008a): different approaches have tackled the problem, the first due to Neter et al. (1965) that has studied bias in the estimates of response errors when the results of a survey are partially improved through record checks, and raises awareness of substantial effects in the results with relatively small errors in the matching process.

Scheuren and Oh (1975) focus on different problems noticed in a large-scale matching task as a Census - Social Security match through Social Security Number (SSN)1. They focus attention to the impact of different decision rules on mismatching and erroneous no matching. Furthermore they point out the constraints to develop an appropriate comparison vector when statistical purposes differ from administrative aims that generated the file and that regulate its maintenance. Nevertheless their approach does not offer general criteria to estimate the parameters of the distributions, as m(γab) and u(γab). Their approach is to select a sample of records, manually check their status of matched and unmatched pair, and estimate those parameters from the observed proportions.

Some more complete methodologies have been developed by Scheuren and Winkler (1993, 1997) through recursive processes of data editing and imputation.

Larsen (2004) and Lahiri and Larsen (2000 and 2005) have widely discussed the use of the former methodology for mixture models, trying to improve the estimates of the probability that a pair of records is actually a match. Those estimates can be found through maximum likelihood or Bayesian analysis, and then adjust the regression models by an alternative to the bias correction method used in Scheuren and Winkler.

Additionally, Liseo and Tancredi (2004) develop a brief regression analysis based on a Bayesian approach to record linkage while Winkler (2006) suggests that the use of a regression adjustment to improve matching can be done by means of identifying variables that are not strictly the same, but actually include the same information from different points of view.

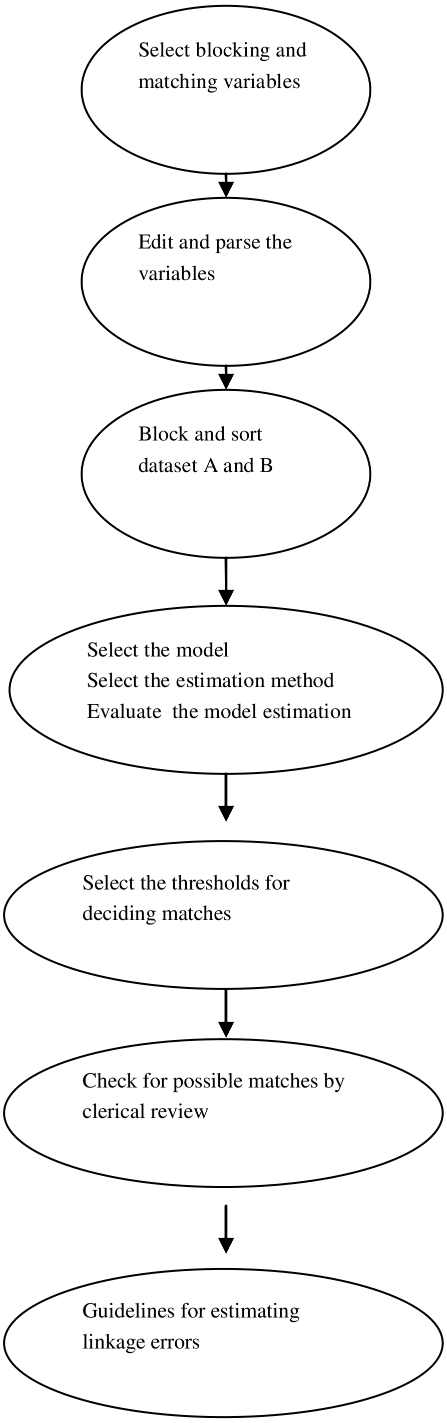
3. Design issues

This present section has been taken from the WP2 of the ESSnet on ISAD (integration of surveys and administrative data), Section 2.1 (Cibella et al., 2008b).

Record linkage is a complex procedure that can be decomposed in many different phases. Each phase implies a decision by a practitioner, which cannot always be justified by theoretical methods. In the following figure, a workflow of the decisions that a practitioner should assume is given. The figure is adapted from a workflow in Gill et al. (2001), p. 33.

---

[1] Although a unique common identifier is used to fuse data from two files, some different problems can arise even when linkage is achieved through some automated process. Scheuren and Oh (1975) report problems related to misprints, absence of SSN in one of the two records that are candidate to be matches, unexplainable changes of SSN in records known to be from the same person, etc.


Select blocking and matching variables
Edit and parse the variables
Block and sort dataset A and B
Select the model
Select the estimation method
Evaluate the model estimation
Select the thresholds for deciding matches
Check for possible matches by clerical review
Guidelines for estimating linkage errors

The workflow describing the practical actions of a practitioner for applying record linkage procedures shows that the actual record linkage problem (as described in WP1 in Section 1, Cibella et al., 2008a) is tackled only in a few steps (the selection of model with the estimation method and the evaluation of the model estimation; the selection of the thresholds for deciding matches).

The steps to be performed are 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 blocking and matching variables should be appropriately harmonised before applying any record linkage procedure.
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, as described above in par.2.
4) After the selection of a comparison function a 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 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. Available software tools

The main use of the record linkage techniques in official statistics produced many software and tools both in the academic and private sectors, like BigMatch (Yancey, 2007), GRLS (Fair, 2001), Febrl (http://www.sourceforge.net/projects/febrl), Link Plus (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm), Tailor (Elfeky et al., 2002), etc.

In the ESSnet on Integration of Surveys and Administrative data (ISAD) the characteristics of some available software tools explicitly developed for record linkage and based on a probabilistic paradigm were analysed (see WP3, Chapter 1,section 1.1 and 1.3, Cibella et al., 2008c).

The probabilistic record linkage tools that have been selected among the most well-known and adopted ones are:

1. AutoMatch, developed at the US Bureau of Census, now under the purview of IBM [Herzog et al., 2007, chap.19].
2. Febrl - Freely Extensible Biomedical Record Linkage, developed at the Australian National University [FEBRL].
3. Generalized Record Linkage System (GRLS), developed at Statistics Canada [Herzog et al., 2007, chap.19].
4. RELAIS, developed at ISTAT [RELAIS].
5. The Link King, commercial software [LINKKING].
6. Link Plus, developed at the U.S. Centre for Disease Control and Prevention (CDC), Cancer Division [LINKPLUS].

An interesting feature of some tools is related to the fact that some record linkage activities are performed "within" other tools. For instance, there are several data cleaning tools that include record linkage but they are mainly dedicated to standardisation, consistency checks etc. A second example is provided by the recent efforts by major database management systems' vendors (like Microsoft and Oracle) that include record linkage functionalities for data stored in relational databases (Koudas et al., 2006).

In the following, two comparison tables are presented and described with the aim of summarising and pointing out the principal features of each tool so far described. In Table 1, the selected values for the characteristics specified above for each of the analysed tools are reported.

Table 1: Main features

	Free/Commercial	Domain Specificity	Level of Adoption
AUTOMATCH	commercial	functionalities for English words	high
FEBRL	free/source code available	no specific domain	medium
GRLS	commercial (government)	functionalities for English words	medium
RELAIS	free/source code available	no specific domain	low
THE LINK KING	free/source code available (SAS licence is needed)	mixed/requires first and last names, date of birth	high
LINK PLUS	free/source code not available	mixed- general features	high

In Table 2 the details on the specific method used for the estimation of the Fellegi and Sunter model parameters are reported.

Table 2: Estimation methods implemented in the record linkage tools

	Fellegi Sunter Estimation Techniques
AUTOMATCH	Parameter estimation via frequency based matching
FEBRL	Parameter estimation via EM algorithm
GRLS	Parameter estimation under agreement/disagreement patterns
RELAIS	EM method Conditional independence assumption of matching variables
THE LINK KING	Ad hoc weight estimation method Not very clear theoretical hypotheses
LINK PLUS	Default M-probabilities + user-defined M-probabilities EM algorithm

The WP3 of the Essnet DI (Data Integration) was focused on the development of common software tools. In particular, as far as record linkage method is concerned, the goal was to improve Relais, the software for record linkage developed by a team of the Italian National Statistical Institute (ISTAT), with pre-processing facilities and a new manual (http://www.essnetportal.eu/sites/default/files/131/Relais2.3Preprocessing.pdf).

5. Decision tree of methods

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. et al. (2008a), Chapter 1. Literature review on probabilistic record linkage. Section 1.2, 1.5 and 1.7 of *WP1 Report of the ESSnet on Integration of Surveys and Administrative Data* (http://www.cros-portal.eu/content/isad-finished).

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

Cibella, N. et al. (2008c), Software tools for record linkage. Chapter 1 of the *Report on WP3 of the ESSnet on Integration of Survey and Administrative Data* (http://www.cros-portal.eu/content/isad-finished).

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.

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.

Elfeky, M., Verykios, V., Elmagarmid, A. K. (2002), A Record Linkage Toolbox. *Proceedings of the 18th International Conference on Data Engineering IEEE Computer Society*, San Jose, CA, USA.

Elfeky, M. G., Verykios, V. S., Elmagarmid, A., Ghanem, M., and Huwait, H. (2003) Record Linkage: A Machine Learning Approach, a Toolbox, and a Digital Government Web Service. Department of Computer Sciences, Purdue University, Technical Report CSD-TR 03-024.

Fair, M. (2001), Recent developments at Statistics Canada in the linking of complex health files. Federal Committee on Statistical Methodology, Washington D.C., USA.

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

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001), On Bayesian record linkage. *Research in Official Statistics* **4**, 185–198. Published also in: E. George (ed.), *Bayesian Methods*, Monographs of Official Statistics, Eurostat, 155–164.

Fortini, M., Nuccitelli, A., Liseo, B., and Scanu, M. (2002), Modelling issues in record linkage: a Bayesian perspective. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1008–1013.

Herzog T. N., Scheuren F. J., and Winkler, W. E. (2007), *Data Quality and Record Linkage Techniques*. Springer Science+Business Media, New York.

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

Koudas, N. and Srivastava, D. (2005), Approximate joins: Concepts and techniques. *Proceedings of VLDB 2005*.

Koudas, N., Sarawagi, S., and Srivastava, D. (2006), Record linkage: similarity measures and algorithms. *SIGMOD Conference 2006*, 802–803.

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

Lahiri, P. and Larsen, M. D. (2000), Model-based analysis of records linked using mixture models. *Proceedings of the Section on Survey Research Methods Section*, American Statistical Association, 11–19.

Lahiri, P. and Larsen, M. D. (2005), Regression Analysis With Linked Data. *Journal of the American Statistical Association* **100**, 222–230.

Larsen, M. D. (2004), Record Linkage of Administrative Files and Analysis of Linked Files. In: *IMS-ASA's SRMS Joint Mini-Meeting on Current Trends in Survey Sampling and Official Statistics*, The Ffort Radisson, Raichak, West Bengal, India.

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

Liseo, B. and Tancredi, A. (2004), Statistical inference for data files that are computer linked. *Proceedings of the International Workshop on Statistical Modelling*, Firenze Univ. Press, 224–228.

Neter, J., Maynes, E. S., and Ramanathan, R. (1965), The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association* **60**, 1005–1027.

Newcombe, H., Kennedy, J., Axford, S., and James, A. (1959), Automatic Linkage of Vital Records. *Science* **130**, 954–959.

Scheuren, F. and Oh, H. L. (1975), Fiddling around with nonmatches and mismatches. *Proceedings of the Social Statistics Section*, American Statistical Association, 627–633.

Scheuren, F. and Winkler, W. E. (1996), Recursive analysis of linked data files. U.S. Bureau of the Census, Statistical Research Division Report Series, n.1996/08.

Scheuren F. and Winkler W. E. (1997), Regression analysis of data files that are computer matched – part II. *Survey Methodology* **23**, 157–165.

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.

Winkler, W. E. (1989), 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.

Winkler, W. E. (2006), Overview of Record Linkage and Current Research Directions. U.S. Bureau of the Census, Statistical Research Division Report Series, n.2006/2.

Yancey, W. (2007), BigMatch: A Program for Extracting Probable Matches from a Large File. Research Report Series Computing, 2007-01, Statistical Research Division, U.S. Bureau of the Census, Washington D.C.

Interconnections with other modules

8. Related themes described in other modules

1. Micro-Fusion – Object Matching (Record Linkage)

9. Methods explicitly referred to in this module

1. Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage
2. Micro-Fusion – Weighted Matching of Object Characteristics

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 – Process

12. Tools explicitly referred to in this module

1. AUTOMATCH
2. Febrl
3. GRLS
4. RELAIS (REcord Linkage At IStat)
5. THE LINK KING
6. LINK PLUS

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.1: Integrate data

Administrative section

14. Module code

Micro-Fusion-T-Probabilistic Record Linkage

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	11-05-2012	first version	Nicoletta Cibella	Istat
0.2	02-10-2012	second version	Nicoletta Cibella	Istat
0.2.1	03-10-2013	preliminary release		
0.3	09-10-2013	EB comments	Nicoletta Cibella	Istat
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:58