Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences

Nicoletta Cibella[1], Gervasio-Luis Fernandez[2], Marco Fortini[1], Miguel Guigò[2], Francisco Hernandez[2], Monica Scannapieco[1], Laura Tosco[1], Tiziana Tuoto[1]

[1]Italian National Statistical Institute – ISTAT – Italy

[2]Spanish National Statistical Institute – INE – Spain

Abstract

In official statistics, the combined use of data from statistical surveys and administrative sources is very common. Indeed, the joint analyses of statistical and administrative data allow to reduce survey costs and response burden: this is an advantage in a context of increasing demands of statistical information on one side and stricter budgetary constraints on the other side. Unfortunately, data sources are often hard to combine since errors or missing information on record identifiers may complicate the integration. Record linkage techniques offer a multidisciplinary set of methods and practices aiming at identifying the same real word entity, which can be differently represented in data sources.

The present paper describes the joint work of ISTAT and INE in order to evaluate the efficiency and effectiveness of the RELAIS toolkit in their own specific context and from their own perspectives. RELAIS (Record Linkage At Istat) is a software for record linkage designed and developed by ISTAT. It permits the dynamic selection of the most appropriate technique for each record linkage phase and the combination of the selected techniques in order to support the definition of the most appropriate strategy on the basis of application and data specific requirements. RELAIS is configured as an open source project, a winning choice for sharing techniques and software.

Some interesting remarks came from the profitable collaboration between ISTAT and INE in exchanging knowledge and solutions related to record linkage, in particular the realized awareness of the common nature of the faced problems and the advantages in prearranging standardized answers to specific but widespread applications.

Keywords: [probabilistic record linkage, open-source software]

1. Introduction

In official statistics, the combined use of data from statistical surveys and administrative sources is very common. Indeed, the joint analyses of statistical and administrative data allow to save time and money for instance by reducing survey costs, response burden, etc.; this is an advantage in a context of increasing demands of statistical information on one side and stricter budgetary constraints on the other side. Unfortunately, data sources are often hard to combine since errors or missing

information on record identifiers may complicate the integration. Record linkage techniques offer a multidisciplinary set of methods and practices aiming at identifying the same real word entity, which can be differently represented in data sources.
Being record linkage a complex process, it has been subject of research for many years and several new methodologies and instruments are currently investigated. From the earliest contributions to modern record linkage, dated back to Newcombe et al. (1959) and to Fellegi and Sunter (1969), a huge number of record linkage solutions have been proposed. However, despite this proliferation, no particular record linkage technique has emerged as the best solution for all cases. We believe that such a solution does not actually exist, and that an alternative strategy should be adopted. Specifically, record linkage can be seen as a complex process consisting of several distinct phases involving different knowledge areas; moreover, for each phase several techniques can be selected. The choice of the most appropriate technique not only depends on the practitioner's skill but it is also application specific. Moreover, in some instances there is no evidence that a given method should be preferred to others or that different choices taken at some linkage stages will conduct to the same results. Furthermore, the overall record linkage workflow could change from user to user, due to different restrictions, such as legal and practical issues, in various fields and countries. Even in a statistical system with shared goals and regulations, as the European Statistical System, different constraints, for instance based on language features, may be present and affect the outcome of the same linkage.

2. RELAIS : the Italian Solution

The absence of a unique solution to record linkage problem led an Italian team of IT researchers and statistical methodologists to design and implement the RELAIS (Record Linkage At Istat) system. The main ambitious goal of this software is to allow the dynamic selection of the most appropriate technique for each of the record linkage phases and the combination of the selected techniques so that the resulting workflow is actually built on the basis of application and data specific requirements.

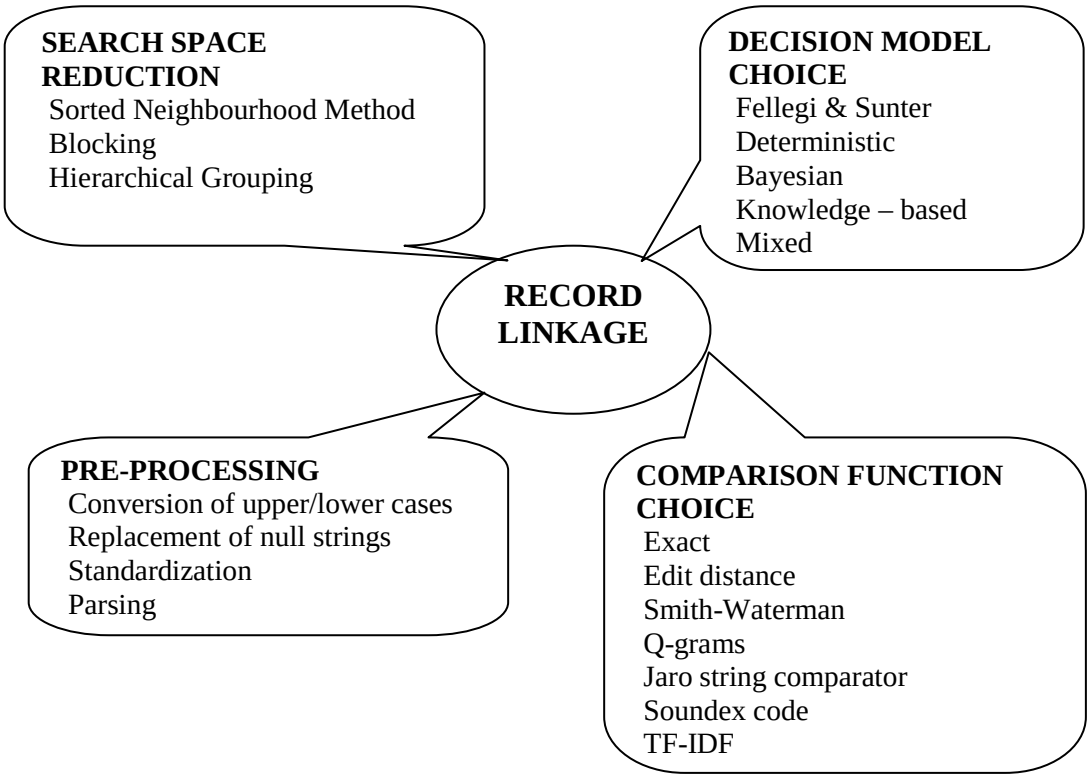
2.1 The RELAIS's Idea

The complexity of the whole linkage process relies on several aspects; for example the lack of unique identifiers requires sophisticated statistical procedures, the huge amount of data to process involves complex IT solutions, constraints related to a specific application may require the solution of difficult linear programming problems. In order to better face with such a complexity, it can be suitable to decompose a record linkage process into some main phases:

1. Pre-processing of the input files
2. Choice of the identifying attributes (matching variables)
3. Choice of the comparison function
4. Creation of the search space of link candidate pairs
5. Choice of the decision model
6. Selection of unique links
7. Record linkage evaluation

The phase decomposition allows to reduce the overall complexity of the linkage process, subdividing the whole problem into sub-problems and searching the more suitable solution for each phase, see Figure 1.

Figure 1: *The record linkage complexity*



In this way the question on which method is better compared to the others is overcome, being convinced that at the moment there is not a unique technique dominating all the others. Moreover, the approach of splitting the overall problem into sub-problems permits to select different methods or techniques among those proposed and available for each one of the linkage phase in order to achieve the definition of the most appropriate overall strategy on the basis of application and data specific requirements.

Keeping in mind this approach to record linkage, the RELAIS toolkit is composed by a collection of techniques for each record linkage phase that can be dynamically combined in order to build the *best record linkage strategy*, given a set of application constraints and data features provided as input. As an example, if it is known that the datasets to compare have poor quality, it is suitable the usage of comparison functions ensuring high precision; as a further example, if no specific error-rates are required by the application, it can be appropriate the usage of an empirical decision model. Also choosing which decision model to apply is not immediate: the usage of a probabilistic decision model can be more appropriate for some applications but it can be less appropriate for others, for which an empirical decision model could prove more successful. Furthermore, even using the same decision model in different application scenarios, a comparison function could fit better than others. Some phases of the record linkage process can be missing: for instance the search space reduction phase makes sense only for huge data volumes, or for applications that have time constraints. In addition, RELAIS exploits at the best the statistical and computational essences of the matching issue.

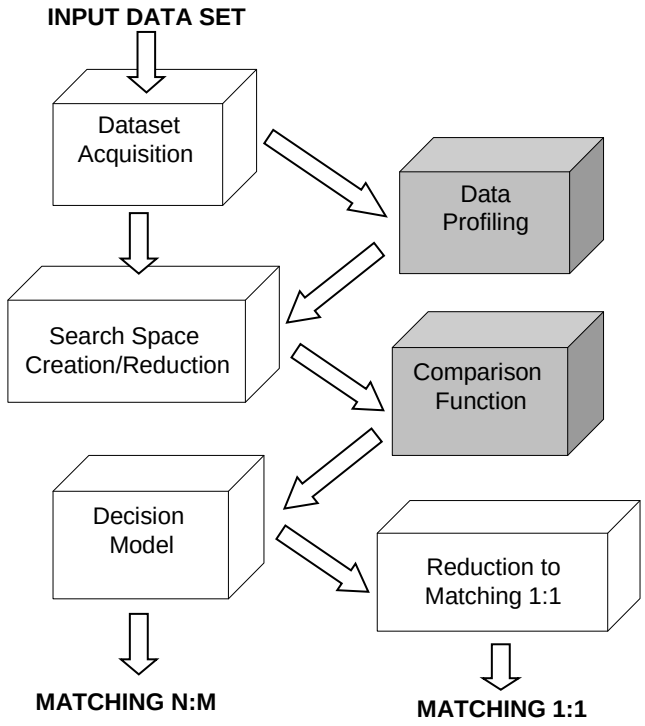
The strength of RELAIS consists of considering alternative techniques for the different phases composing the record linkage process. Therefore, we claim that no record linkage strategy, deriving from the choice and combination of a specific technique for each phase, is the best for all applications. RELAIS wants to help and guide users in defining their specific linkage strategy, supporting the practitioner's skill, due to the fact that most of the available techniques are inherently complex, thus requiring not trivial knowledge in order to be appropriately combined.
RELAIS is proposed also as a toolkit for researchers: in fact, it gives the possibility to experiment alternative criteria and parameters in the same application scenario, that's really important from the analyst's point of view. However, the relevance of an instrument like RELAIS is mainly appreciable in a statistical system with shared goals and regulations, as the European Statistical System: the availability of different methods and techniques and the possibility of selecting the most appropriate strategy with respect to the specific problem considered using a unique software, guarantee the harmonization of tools and methodologies and the comparability of outcomes.

2.2 Main features of RELAIS

RELAIS has been designed with a modular structure. The modules implement distinct record linkage techniques and each one has a well defined interface towards other modules. In this way it is possible to have a parallel development of the different modules, and to easily include new ones in the system. Moreover, the overall record linkage process can be designed according to specific application requirements, combining the available modules. A user interface guides the design of the record linkage workflow in order to help the user to easily combine the available modules in a meaningful and controlled process.
Moreover, RELAIS is configured as an open source project, released under the EU Public Licence. There are at least two reasons for this choice. First, there are many possible techniques that can be implemented in parallel for each record linkage phase: relying on a community of developers such set can be increased and maintained very rapidly. Second, in the last years there have been several independent efforts towards the resolution of record linkage problems but such efforts have not led to the best solution. An open source project could instead give the possibility of gathering together the contributions already done in order to make them available to the community for the most appropriate usage; in this way it could be possible to reach the goal of providing, in the shortest possible time, a generalized toolkit for building dynamic record linkage workflows.
From an implementation perspective, RELAIS is written in Java and R languages. Both languages are open source and can be used on different technological platforms. The combined usage of these two languages permits to rely on the best features of each of them. Specifically, Java is used for the data-oriented tasks and for the development of the user interface, while R is used for the computational-oriented ones. Some R packages have been used to solve specific problems as a clear example of the possible re-use allowed by open source projects. R code is embedded in Java code, so that the calls to R software are completely transparent to the final users. The currently available version of RELAIS is the 1.0 and it is a beta version. It has a file based architecture, meaning that all data reside on text files, both input and output data as well as the intermediately produced data. Such a file based architecture is evolving into a relational database architecture that permits to manage huger amounts of data in a more efficient way.

Figure 2 shows the different record linkage phases that are implemented in the RELAIS system. The grey blocks are under development and will be available in the next release of the software.

Figure 2: *Phases implemented in RELAIS*



The dataset acquisition phase permits to read two input datasets from text files. The datasets must have the same names for the common variables that are the ones considered by the system in the subsequent phases.
From the acquisition phase, it is possible to pass directly to the search space creation/reduction phase or to the data profiling phase.
The data profiling phase permits to characterize available variables with respect to some quality features that can be used to support two tasks, that is blocking variables choice and matching variables selection. The quality features currently under development are: accuracy, consistency, identification power, completeness, correlation and entropy.
The search space creation/reduction phase allows to build the set of the candidate pairs to be linked. Two methods for space reduction have been implemented, namely blocking and sorted neighborhoood method (SNM).
A set of comparison function is available in order to compare strings according to an exact or an approximate procedure.
The most important phase is the choice of the decision model to apply for taking the matching decision on candidate pairs. The currently implemented model is the Fellegi and Sunter one. In the next version of RELAIS also exact and deterministic models will be available.
After the decision model choice it is possible to produce an N:M matching result or a 1:1 matching result, applying a dedicated reduction phase. This phase has been implemented by resolving a linear programming problem on the N:M output.

The output of the linkage process consists of three disjoint datasets: match, non-match and possible match. Possible matches need to be processed by clerical review. Further analyses can be performed starting from such an output, that is a new record linkage process can be initiated with the residual non-matched records.

3. Sharing Solutions: the Spanish and the Italian Experiences with RELAIS

The Italian idea and RELAIS's objectives have been immediately shared with some other European statistical offices (by exploiting the ESSnet on ISAD – Integration of Statistical and Administrative Data, formerly CENEX). The Spanish National Statistical Institute - INE – has been testing the toolkit in order to evaluate it in their own specific context and from their own perspectives. The following paragraphs describe both the Italian and the Spanish experiences.

3.1 The Italian experimentations

The Italian tests reported below refers to 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, which are the smallest territorial level considered by the Census. The size of the PES's sample was about 70 000 households and 180 000 individuals and the variables stored in the files are name, surname, gender, date and place of birth, marital status, etc. The estimates of the Census coverage rate through capture-recapture model (Wolter, 1986) 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 it is possible to evaluate the performances of the linkage implemented in RELAIS.

The RELAIS performances were deeply tested on different subset of data, in particular the effectiveness performances of the overall strategy (mainly applying blocking procedure, 1:1 matching, probabilistic model) presented below, were evaluated on a 8 000 record size subset. The effectiveness of the linkage performances were evaluated in terms of match rate, false match rate and false non-match rate. The match rate is defined as the number of linked record pairs divided by the total number of true match record pairs. The false match rate and the false non-match rate correspond to the well-known type II and type I errors in a one-tail hypothesis test context. The false non-match rate indicates the ratio between the number of incorrectly non matched records and the whole number of the true matched records. The false match rate denotes the ratio between the records incorrectly matched and the whole number of matched pairs.

The efficacy performances were tested using the RELAIS software, ignoring the known true matching status. As matching variables all the strongest identifiers were used: name and surname, gender, day, month, and year of birth. The equality was 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 Matches. The upper threshold was fixed assigning to the set of Matches all the pairs with the composed matching weights correspondent to estimated matching probability higher than 0.99; the set of the possible links were created fixing the lower threshold level with the composite matching weight correspondent to the estimated matching probability lower than 0.50. The pairs falling into the set of the Possible Matches were assigned to the set of Matches without a clerical supervision of the results.
A blocking phase was performed considering as a blocking variable the month of birth of the household header. 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 sizes 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 related to record linkage procedure performed in analogous conditions in terms of identification variables, number of matched records and 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 finalized 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 minimize the risk of loosing matches.
RELAIS also allows to analyze the results of the linkage procedure in more detail. For each block the amount of pairs linked by the procedure is reported together with the number of pairs that the procedure identifies as possible links and for which a manual review or a more in-depth analysis is suggested. In this way, knowing both true matches and false matches it's possible to evaluate the true match rate and the false match rate for each block. This is particularly important when one or more specific block displays anomalous results with respect to the other blocks, suggesting to adopt different strategy for such a particular subset. Moreover an analysis of the blocking procedure that considers each block allows to evaluate how much of the missed matches is introduced by the blocking procedure itself, because true links cannot be individuated due to the fact that the records do not agree on the blocking variable. In particular, the analysis has to concentrate on the categories of the blocking variable affected by errors that cause a higher amount of false non-matches.
Another relevant point regards the time and the efforts consumed in performing the linkage. With respect to the data considered in this experiment, the complex linkage procedure applied for obtaining the Post Enumeration Survey estimates required several days of work and more than one dedicated person. On the contrary, the linkage performed by RELAIS was obtained in less than one day by only one person.

3.2 The Spanish Experiences

The Spanish experience refers to a record linkage operation also performed in a real-world environment, integrating data from the Living Conditions Survey (LCS) and the Central Population Register (CPR) by means of RELAIS. This was done in order to obtain from the CPR the ID number of the polled individuals, which was not available in the LCS survey. The absence of ID number is not unusual in personal surveys, and it is due either to legal issues or to non-response problems that make advisable to skip that question. Since other variables as name, address, or date of birth are, though,

easily collectable, the chance of matching them with the corresponding ID number through record linkage methods, in order to use additional information that is available in different administrative records and then enrich the original dataset, must be strongly taken into account.
Besides this objective, the Spanish application had a twofold aim: the first was to asses the capabilities of the various functionalities included in the RELAIS toolkit, e.g. the use of the EM algorithm for record linkage purposes; the second was to compare the results achieved by the software with those obtained throughout some alternative ad hoc techniques. Furthermore, the Spanish tests focused on using blocking methods in order to reduce the space search, given the high amount of registers to be compared.
In this case, the task had to face additional challenges, in the sense that name was the only data available to compare, since no harmonisation procedure for postal addresses was possible; each character string corresponding to a name was split into separate entities, and then the first, second, last but one and last items were selected for comparisons.
Even though it is still in a development stage, the output of the record linkage action via RELAIS is not only significantly better than the one produced by *ad hoc* techniques but it is also highly satisfactory in the sense that an initial 90 percent of records could be matched just through equality functions and a standard blocking method based on geographic areas.
The Spanish tests highlight some strengths and weaknesses. The record linkage results seem really startling taking into account that just an equality function on parsed names was used. An advantage is also given by a noticeable flexibility: for instance, in the open-source philosophy, it has been possible to modify a specific pass of the reduction from M:N matching to 1:1 matching and to modify some of the implemented choices, achieving solutions more suitable for the considered application. On the other hand, the current release shows difficulties when handling very large amounts of data and choosing the blocking alternative, since linking routines must be repeated for each block one after another, although modifying the set of linking variables. This could result in a high time consumption for large number of blocks. Nevertheless, it must be emphasized that in some special cases, different scenarios for each subset can be obtained once the sets have been split into blocks, so different models with different variables could really apply; for example, some geographic areas could include a much higher proportion of foreigners and force to include –or discard- some other information such as date of birth or first name. Therefore, it is not strictly a design weakness and can be easily tackled once the application includes an option to solve the blocks using the same model on an automated basis.
Finally, the whole cross-product in absence of blocks could possibly cause overflow problems in the writing phase for extremely large datasets.

4. Concluding Remarks

As a consequence of the feedbacks provided by the Spanish group, the Italian team is working on improvements and new functionalities are in order to overcome the limitations of the 1.0 version of RELAIS.
More specifically, the new features include:

- A relational database architecture in order to optimize the performances with respect to the management of huge amount of data through the whole record linkage process (input, intermediate phase and output).
- Several distance functions for approximate string comparisons. Specifically, matching variables will be compared not only by means of the equality function but also via several other suitable distances (both for numerical and string variables).
- Exact and deterministic decision models will be available, to be used either as alternatives or in conjunction with the probabilistic model.
- A data profiling phase in which a set of quality metadata are calculated starting from real data; these metadata help the user in the critical phases of choosing the best blocking or matching variables.

Moreover, the collaboration with the Spanish group also resulted in an enriched planning of RELAIS's future work. Specifically, the need for further decision models, alternative to the already included ones, has emerged as an important issue. For example, record linkage algorithms based on evolutionary computation appears as promising. These methods can be very useful due to their approximate (rather than exact) nature in looking for a solution to the record linkage problem. Especially in the presence of large amount of data, they can lead to an approximated solution in an efficient way. Moreover, algorithms that improve the performance of the 1:1 reduction phase will be investigated. Another aspect that deserves further study is related to methods able to evaluate the quality of the record linkage process.

Finally, some remarks immediately come from this profitable experience in exchanging knowledge and solutions among different national institutes and countries in dealing with 'real-world' tasks: first of all, the realized awareness of the common nature of the faced problems; then, the advantages in designing standardized answers to specific but widespread applications; finally, the winning choice of the open-source solution for sharing techniques and software.

References

Cibella N, Fortini M, Scannapieco M, Tosco L, Tuoto T (2008), "Theory and practice of developing a record linkage software, in Proceeding of the Workshop "Combination of surveys and administrative data" of the CENEX Statistical Methodology Project Area "Integration of survey and administrative data"- Vienna 29-30 May 2008

Fellegi I. and Sunter A. (1969) A Theory for Record Linkage. Journal of the American Statistical Association, 64.

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

Relais 1.0. User's guide, Istat, http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/

Tuoto T, Cibella N, Fortini M, Scannapieco M and Tosco L (2007), "RELAIS: Don't Get Lost in a Record Linkage Project", in Proceeding of the Federal Committee on Statistical Methodologies (FCSM 2007) Research Conference, Arlington, VA, USA

Wolter K. (1986) Some coverage error models for census data. Journal of the American Statistical Association, 81:338-346.