Proceedings of Q2008
European Conference on Quality in Official Statistics

# The relationship between error rates and parameter estimation in the probabilistic record linkage context

Nicoletta Cibella, Marco Fortini, Tiziana Tuoto
Italian National Statistical Institute ISTAT, Rome, Italy

## 1. Introduction

Nowadays data integration procedures are becoming extremely important in official statistical institutes. In particular, record linkage procedures, aiming at matching records referring to the same entities, both within a dataset and from two or more different data sources, improves the quality of information collected and make possible more detailed analysis. A large number of linkage techniques are available and commonly used; in the field of the official statistics, the quality of the implemented procedures is crucial, not only because it needs to be established the criterion to estimate the accuracy of the procedures but also because has to be evaluated the error match rates. The aim of this paper is assessing the probabilistic record linkage process quality by means of alternatives methods to estimate the parameters of the probability model; in other words, we aim to evaluate how the procedure accuracy is related and dependent on the choices adopted in the parameters estimation phase.

For the parameters estimation phase, alternative methods have been analysed, by performing both the EM algorithm in the probabilistic method, and the deterministic approach, in order to evaluate the improvement due to the increase of the distinguishing power of some variables.

The results of the comparisons are evaluated and synthesized in terms of matching proportion, false match and false not-match rates. Generally, it's not easy to find automatic procedures to estimate these two types of errors so as to evaluate the quality of record linkage procedures. So, finally also those errors are calculated via different methods, firstly starting from the known true matches status, but also through functions of the parameters themselves (Belin and Rubin, 1995; Torelli and Paggiaro, 1999).

The present study tests the alternative choices above described, exploiting the great amount of real data referred to the 2001 Italian Population Census and the related Post Enumeration Survey (PES). Actually, the Census coverage rate is usually estimated on the basis of a post enumeration survey which needs to be linked with the Census data itself in order to estimate the unknown amount of the population, via dual system estimation model. This model assumes the linkage procedure was nearly free of errors, given that a combination of different procedures (deterministic, probabilistic and clerical) was performed so to reach highly accurate results; for this reason, it's possible to consider as known the true match status of each unit so as to evaluate the quality of alternatives procedures in a simulation context.

## 2. Record linkage and Quality2008

The record linkage has the purpose to identify, quickly and accurately, the same real world entity, which can be differently represented in one or more data sources. A record linkage project can be performed for different purposes and this richness of possible applications makes it a powerful instrument to support decisions in large commercial organizations and government institutions. In the official statistics context, the combined use of statistical survey and administrative data is largely widespread and strongly stimulates the investigation of new methodologies and instruments to deal with record linkage projects and to accurately identify units across different data sources. Actually, many potential advantages in using administrative data for statistical purposes are known and shared by the various national statistical institutes. In fact, administrative sources usually contain large amounts of data, often very accurate, due to improvements made over time For this reason in most situations the joint analysis of statistical and administrative sources allows to save time and money, reduce survey costs and response burden, etc. Indeed, cooperation among different public agencies or institutes is actually based on common data sharing, that prevents from recollecting data from citizens or enterprises, if such data are already available at some of the public subjects.

Generally, there are various examples of application of linkage procedures that make use also of techniques based on data mining, machine learning, soft computing and others. Just mentioning only a few of the most important: the update and the de-duplication of frame, when multiple records referring to the same real world entity are stored within one single data source; data integration, across multiple data sources in order to provide a reconciled global record; correction across multiple data sources, performed when one source is known to have higher quality data that can be used for improving the others; measure of a population by capture-recapture for instance on the occasion of the Census post-enumeration survey; confidentiality check of public-use microdata, through re-identification experiments.

However, data sources are often hard to combine since errors or lacking information in the record identifiers may complicate the joint use of information. In order to overcome such difficulties, record linkage techniques provide a multidisciplinary set of methods and practices.

Record linkage procedures substantially improve the quality and the quantity of the available information by integrating different frames. Actually, identifying pairs of records coming from either the same or different data files, can help in evaluating the accuracy of the information coming from a given source, when some of the same variables are collected across different data files and can overcome lack of information; moreover the linkage enables in carrying out more detailed analysis. It needs also to be remarked how important is to evaluate the quality of the linkage output, so reducing as much as possible the matching errors, particularly when further analysis are based on previously linked data .

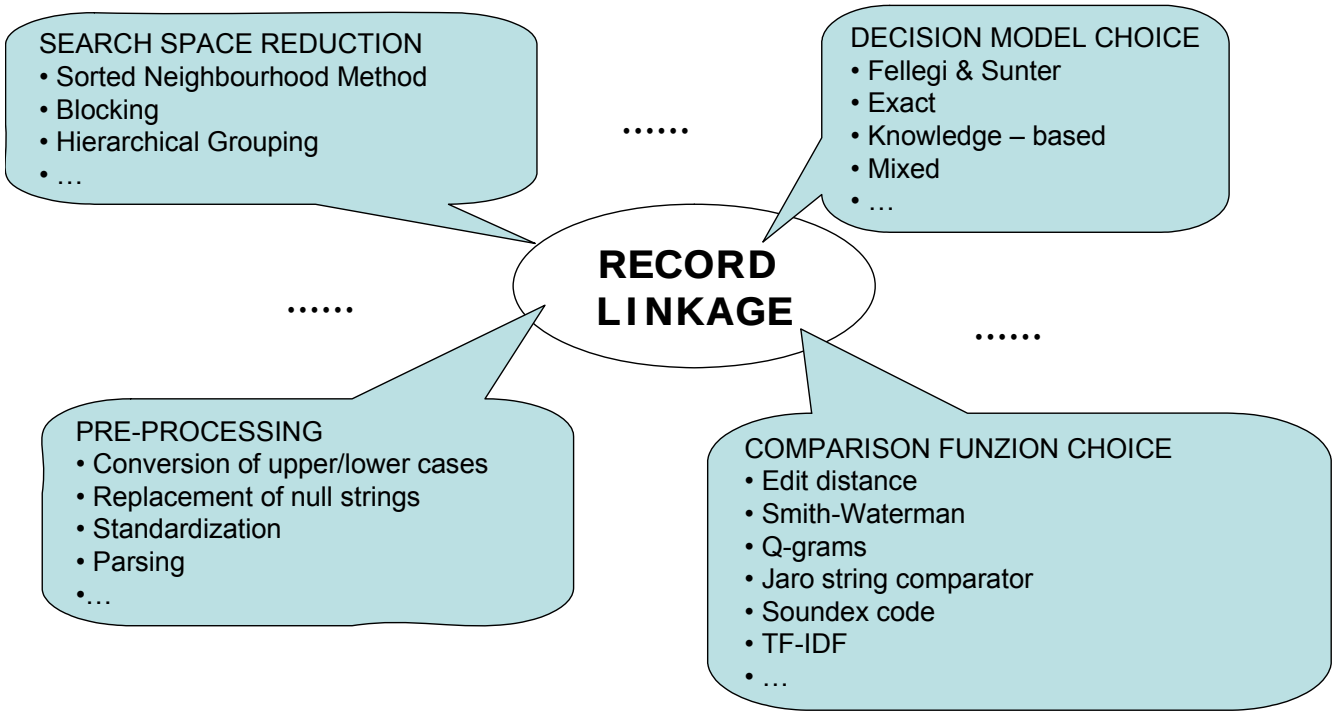
## 3. Linkage Framework

Since procedures aiming to single out the same individual from different sources are very complex, a decomposition of the whole linkage process into its constituting phases is necessary so to highlight the various knowledge areas which are involved and the several techniques that can be adopted in each phase. We can distinguish the whole procedure in at least 7 main steps, namely :

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

For each of these phases we can adopt different techquines (e.g.selecting a different comparison function for each of the choosen matching variables ) and it could be reasonable to combine the selected techniques for building a record linkage workflow of a given application (Figure 1).
We do believe that the choice of the most appropriate technique not only depends on the practitioner's skill but especially it is application specific. Moreover, in some instances there is not evidence that a given method should be preferred to others or that different choices taken at some linkage stage will conduct to the same results. In this case, from the analyst's point of view, it is important to be able trying different alternatives which application criteria and parameters should be properly tuned.

Figure 1. Record linkage phases



Moreover, the complexity of the whole linkage process relies on several aspects; for example the lack of an unique identifier requires the application of 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.

## 4. Record linkage: formalization

Given two data sets *A* and *B of* size $N_A$ and $N_B$ respectively, let us consider $\Omega = \{(a,b),\ a \in A \text{ and } b \in B\}$ of size $N = N_A \times N_B$. 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)

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

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

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

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

Following Fellegi and Sunter (1969), the ratio

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

between the probabilities of $\gamma$ given the pair (a,b) membership either to the subset *M* or *U* is used so as classifying the pair. 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 a classification criterion based on two thresholds $T_m$ and $T_u$ ($T_m > T_u$)

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

- 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 to minimize false match rate (FMR) and false non-match rate (FNMR)

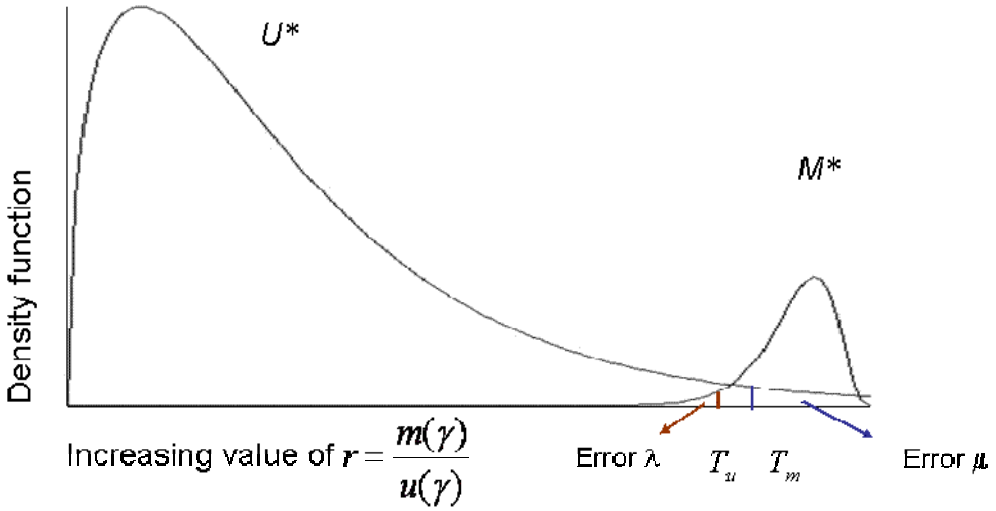
$$FMR = \sum_{\gamma \in \Gamma} u(\gamma) P(M^* \mid \gamma) = \sum_{\gamma \in \Gamma_{M^*}} u(\gamma) \quad \text{where} \quad \Gamma_{M^*} = \{\gamma : T_m \le m(\gamma)/u(\gamma)\}$$

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

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.

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 2. The mixture model for m- and u-distributions




Armstrong and Mayda (1995) assume that the frequency distribution of the observed patterns γ is a mixture of the matches m(γ) and non-matches u(γ) distributions

$$P(\gamma) = P(\gamma|(a,b) \in M)P((a,b) \in M) + P(\gamma|(a,b) \in U)P((a,b) \in U)$$
$$= m(\gamma) \cdot p + u(\gamma) \cdot (1-p)$$

where *p=P(M)*.

This assumption can be viewed as in the figure below where manifest data can be divided into as many sub-tables as the latent classes defined by the linkage problem.

Figure 3. Tables representing the mixture model

| $X_1$ | $X_2$ | ... | $X_k$ | freq |
|---|---|---|---|---|
| 0 | 0 | ... | 0 | $P(\gamma_1)N$ |
| ... | ... | ... | ... | ... |
| 1 | 1 | ... | 0 | $P(\gamma_{2^k-1})N$ |
| 1 | 1 | ... | 1 | $P(\gamma_{2^k})N$ |

| $X_1$ | $X_2$ | ... | $X_k$ | freq_match |
|---|---|---|---|---|
| 0 | 0 | ... | 0 | $P(\gamma_1\|M)P(M)N$ |
| ... | ... | ... | ... | ... |
| 1 | 1 | ... | 0 | $P(\gamma_{2^k-1}\|M)P(M)N$ |
| 1 | 1 | ... | 1 | $P(\gamma_{2^k}\|M)P(M)N$ |

| $X_1$ | $X_2$ | ... | $X_k$ | freq_unmatch |
|---|---|---|---|---|
| 0 | 0 | ... | 0 | $P(\gamma_1\|U)P(U)N$ |
| ... | ... | ... | ... | ... |
| 1 | 1 | ... | 0 | $P(\gamma_{2^k-1}\|U)P(U)N$ |
| 1 | 1 | ... | 1 | $P(\gamma_{2^k}\|U)P(U)N$ |

The joint distribution of the observations $\gamma$ and the latent variable C=c (c=(0,1)) is given by:

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

Under the local independency assumption, we are in fact dealing with a latent class analysis where the likelihood function for $mk(\gamma)$, $uk(\gamma)$ ($k$=1,…,$K$) and $p$ is given by:

$$L=\prod_{(a,b)}\left[\mathrm{p}m(\gamma^{(a,b)})\right]^{c^{(a,b)}}\left[(1-\mathrm{p})u(\gamma^{(a,b)})\right]^{1-c^{(a,b)}}. \quad (2)$$

Since vector ***C*** is not directly measurable, the estimation of parameters $m_k(\gamma)$, $u_k(\gamma)$ and $p$ can be obtained through EM algorithm (Dempster, Laird, Rubin, 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 independency 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 independency 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.

## 5. Some results on real-world data

In this study, the alternative model specifications described above are tested by exploiting the great amount of real-world data referred to the XIV Population Census and the related PES. The main goal of the Census was to enumerate the resident population at the Census reference date, 21/10/2001. The PES instead had the objective of estimating the Census coverage rate; it was carried out on a sample of enumeration areas (*EA*), which are the smallest territorial level considered by the Census. The PES sample size was about 70,000 households and 180,000 individuals while the main variables stored in the files were name, surname, gender, date and place of birth, marital status, education attainment and occupation. 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 K., 2006) 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 coming from 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 estimates of the Census coverage rate through capture-recapture model have 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 were 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 quality of alternatives procedures by using real-world data.
Two files of about 650 records each were drawn out from Census and PES, respectively, with the purpose of testing the procedure. The most powerful variables, in terms of identification power, were selected, namely: name, surname, day and year of birth. All the records in the two files also agree with respect to the month of birth and to geographic codes, due to the fact that such variables were already used for blocking when the files were created. For each couple of records generated by the Cartesian product of the two files, the key variables were compared in terms of their equality-inequality; in table 1 the frequencies of each comparison pattern are reported.

Table1. Frequencies of the comparison patterns.

| Surname | Name | Day of Birth | Year of Birth | Freq |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 414138 |
| 0 | 0 | 0 | 1 | 5321 |
| 0 | 0 | 1 | 0 | 14004 |
| 0 | 0 | 1 | 1 | 168 |
| 0 | 1 | 0 | 0 | 3090 |
| 0 | 1 | 0 | 1 | 43 |
| 0 | 1 | 1 | 0 | 102 |
| 0 | 1 | 1 | 1 | 9 |
| 1 | 0 | 0 | 0 | 969 |
| 1 | 0 | 0 | 1 | 17 |
| 1 | 0 | 1 | 0 | 22 |
| 1 | 0 | 1 | 1 | 19 |
| 1 | 1 | 0 | 0 | 14 |
| 1 | 1 | 0 | 1 | 9 |
| 1 | 1 | 1 | 0 | 6 |
| 1 | 1 | 1 | 1 | 513 |

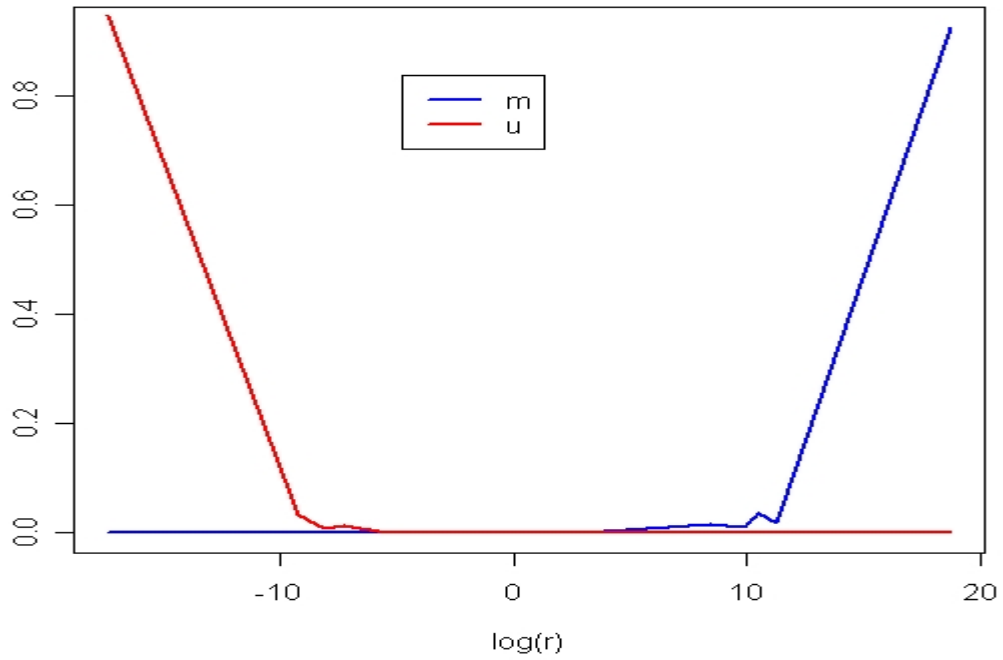
A latent class model was applied to these data, according to the theory briefly recalled in the previous paragraph, and its parameters were estimated through the EM algorithm. Under the hypothesis of local independency, the estimate of the overall matching probability was 0.0013 ; the estimated probabilities of key variables agreement given the pair membership either to set M or U are reported in table 2.

Table 2. Estimates of the agreement probabilities

| Variables | P(X=1\|M) | P(X=1\|U) |
|---|---|---|
| Surname | 0.9853 | 0.0023 |
| Name | 0.9650 | 0.0074 |
| Day of birth | 0.9825 | 0.0327 |
| Year of birth | 0.9889 | 0.0127 |

Instead of recurring to a couple of threshold as in Fellegy and Sunter (1969) only one threshold was fixed so as to assign pairs to sets M and U without any manual revision. This threshold was fixed at a value corresponding to the expected false match error FMR=0.001. In so doing, the expected false non-match rate remains assigned at a value of FNMR= 0.0001. In figure 1, the curves representing the estimates of the m- and u-distributions are drawn against the value of the log (r).

Figure 4. Estimates of m- and u-distributions




The comparison between the results of the linkage procedure and the true linkage status is reported in table 3. Looking at table 3, it is quite evident that the probabilistic procedure is really effective, given that 567 matches are achieved out of the 573 true ones. Only 6 true matches are loosed while 10 pairs are matched even being non-matches.

Table 3. Comparison between the true linkage status and the results of the linkage procedure

| | | ***True Linkage Status*** | | |
|---|---|---|---|---|
| | | ***Matched*** | ***Not Matched*** | ***Total*** |
| ***Results of the Linkage Procedure*** | ***Matched*** | **567** | **10** | **577** |
| | ***Not Matched*** | **6** | | |
| | ***Total*** | **573** | | |

Nevertheless, the probabilistic procedure does not provide a good self-evaluation of the quality. In fact, while the expected FMR and FNMR are 0.001 and 0.0001, respectively, the ‘true’ ones coming from the more accurate procedures, carried out during the PES, are equal to 0.017 and 0.010, respectively. In other words, the linkage procedure

achieves “appreciable” results in terms of linked pairs but the linkage errors connected to these results are underestimated.
Before trying to improve the quality evaluation of the probabilistic procedure by considering more appropriate estimation model, the results of the probabilistic approach have been compared with the outcomes of a deterministic procedure, developed according to the following three steps. First of all, records were considered matched when reporting the same values for all the four matching variables considered also in the probabilistic case; secondly, the less discriminating variable, as identified by the P(X=1|M) reported in table 2, was removed and all the pairs not agreeing only for this variable at the first step, were paired. In other words, the records agreeing on surname, day and year of birth but not on name were added to the set of matches. Finally, the residual records of the previous steps were matched if they report the same values on surname, name and year of birth, i.e. relaxing the equality on the second less powerful variable (i.e. day of birth). Table 4 reports the comparison between the results of deterministic procedure and the true linkage status. Comparing tables 3 and 4, it is clear that the probabilistic procedure is more effective than the deterministic one; the FMR and the FNMR of the latter are 0.005 and 0.06, respectively. Moreover, the deterministic approach cannot provide a direct measure of matching errors when their evaluation is not available in advance, as in this example.

Table 4. Comparison between the true linkage status and the results of the linkage procedure

| | | ***True Linkage Status*** | | |
|---|---|---|---|---|
| | | ***Matched*** | ***Not Matched*** | ***Total*** |
| ***Results of the Linkage Procedure*** | ***Matched*** | **538** | **3** | **541** |
| | ***Not Matched*** | **35** | | |
| | ***Total*** | **573** | | |

In order to try to achieve a better evaluation of the linkage errors in the probabilistic context, we defined an alternative model for the m- and u-probabilities, i.e. we relaxed the conditional independence assumption, considering also interactions between matching variables, given the latent one. This choice was supported by the analyses of some tests of association between variables, such as chi-squared values, that point out a certain degree of association between variables. Models with all the simple effects and almost a two-way association between matching variables were compared each other in terms of both log-Likelihood maximum values and Bayesian Index Criterion (BIC). Differences in log-Likelihood maximum among models is not significant, as shown in table 5, although the best value is achieved by the model which considers the single effects together with interaction between surname and name. The slight preference towards this model which rises from this indicator is definitively reversed looking at the

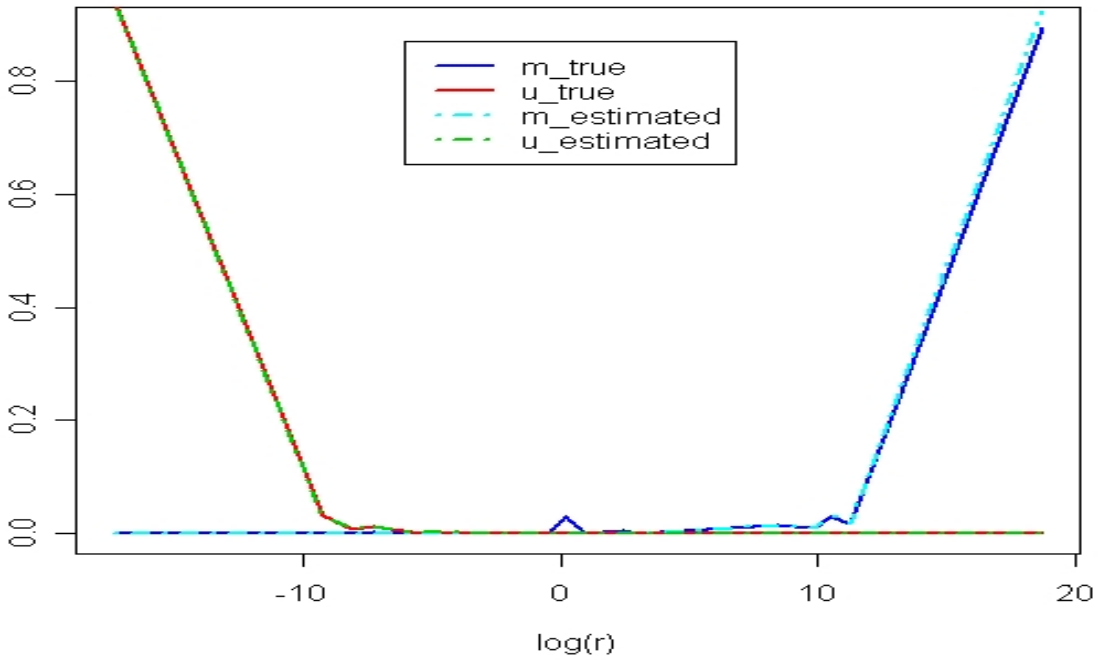
BIC: this index is built starting from the log-Likelihood value and introducing a penalty related to the number of parameters of the model. The best value (the lowest one) of the BIC corresponds to the model based on the local independence assumption, i.e. the one with less parameters. Moreover, as shown in table 5, models considering more than one interaction cannot be estimated because of their non-identifiability.
The local independence assumption model and its alternatives which admit interaction between variables were compared also in terms of error rates, FMR and FNMR. In this sense, all the models turn out to be completely equivalent, since they find out the same pairs of records.

Table 5. Indicators for model selection

| Model | logLikelihood | BIC | npar |
|---|---|---|---|
| Local Indep | -123 586 | 247 289 | 9 |
| Interation Surname and Name | -123 583 | 247 297 | 10 |
| Interation Surname and Day of Birth | -123 584 | 247 298 | 10 |
| Interation Surname and Year of Birth | -123 585 | 247 301 | 10 |
| Interation Name and Day of Birth | -123 586 | 247 302 | 10 |
| Interation Name and Year of Birth | Not identifiable | | 10 |
| Interation Day and Year of Birth | Not identifiable | | 10 |

The reason why matching errors were not correctly evaluated by the local independence assumption model was finally investigated looking at the actual m- and u-probabilities distribution. In figure 5, the two lines corresponding to the estimated m- and u-distributions and the two ones concerning the true values for m- and u- distributions against the logarithm of the estimated r ratio, are drawn together. The estimated lines are very close to the true lines for most of the values of r. It is however evident the anomalous peak of the true m-distribution corresponding to the zero value of the log(r). This peak is determined by few pairs which in fact are the same unit but have missing values in the most discriminating variables, i.e. surname and name. This analysis suggests that the models compared in this study are not able to disclose the particular pattern assumed by these data. For instance, it should be appropriate explicitly introduce some patterns for missing data in case practitioner wants to provide a good evaluation of the linkage quality together with good linkage results.

Figure 5. Estimated and true m- and u- distributions




## 6. Further analyses

In this paper an exercise on real data was shown so to emphasize the statistical nature of record linkage problems and their strong dependence by the accuracy of the underlying statistical model. We have seen that, even in presence of data scarcely affected by errors, few pairs which not conform to the underlying statistical model can result in inaccuracies and linkage errors.
Better parameter estimates could be also achieved by means of different operation on other phases of the linkage process as for instance: the identification of appropriate subsets of all the pairs candidate to the linkage (Yancey, 2004); the use of alternative techniques for reducing the search space Ω, more effective than the conventional blocking criteria (Baxter, 200 ); the application of the recent mapping algorithms, that allow to map objects preserving the similarities and dissimilarities between them (Faloutsos C et al, 1995).
Further analyses can also follow from our initial study. First of all, validity of the local independence assumption among key variables can be tested and properly taken into account by relaxing constraints on the latent class parameters. In our example, key variables were not enough affected by dependencies in order to evaluate if alternative models are able to improve both identification of pairs and linkage errors estimation. Replication of the same exercise including other key variables on the same data or even on a different dataset could allow to evaluate this issue.

As an alternative, perturbation of real data following some errors generation model can be considered. Doing so, associated errors into key variables could be induced and a quasi-experimental situation could be considered.
Many interesting details can be investigated. How much model fitting can be improved by considering more parameters in the latent class model and in what degree this improvement results in a better discriminating power are two examples of these interesting issues. Moreover, it is attractive to investigate how recognition of decision thresholds can be refined by a better model fitting. Finally, the effectiveness of the prediction probabilities of linkage error is another detail which could be enhanced by an improved model definition. Robustness of the local independence model in case of departure of this hypothesis and the use of more parameters in log-linear model with latent variable in order to outperform the simple model will be the main goal of our further studies.

References

Baxter R., Christen P., Churches T., (200 ),"A Comparison of Fast Blocking Methods for Record Linkage" ACM SIGKMOG '03 Workshop on Data Cleaning, Record Linkage and Object Consolidation, August 27, 2003, Washington DC.
Belin T.R., Rubin B. (1995), "A method for calibrating false-match rates in record linkage", Journal of the American Statistical Association, vol.90, n.430.
Faloutsos C., Lin K.-I. (1995) "Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets". In M. J. Carey and D. A. Schneider, editors, SIGMOD, pages 163-174.
Fellegi I.P., Sunter A.B. (1969) "A Theory for record linkage", Journal of the American Statistical Association, 64, 1183-1210.
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, 89, 414-420.
Torelli N., Paggiaro A. (1999) "Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forse di lavoro" Working paper n.15 del progetto di ricerca cofinanziato MURST "Lavoro e disoccupazione: questioni di misura e di analisi", Dipartimento di Scienze Statistiche, Università di Padova.
Yancey W (2004), "Improving algorithm estimates for record linkage parameters", Research report series U.S.Census Bureau,
http://www.census.gov/srd/papers/pdf/rrs2004-01.pdf
Winkler W.E. (1995), "Matching and Record Linkage", in Cox, Binder, Chinnappa, Christianson, Colledge, Kott (a cura di), Businness Survey Methods, Wiley & Sons, pp. 355-384.
Wolter K. (1986) "Some coverage error models for Census data", Journal of the American Statistical Association, 81, 338-346.